GI-Edition

Lecture Notes in Informatics

E. Plödereder, L. Grunske, E. Schneider,
D. Ull (Hrsg.)

INFORMATIK 2014

Big Data – Komplexität meistern

22. – 26. September 2014
Stuttgart

Proceedings

232

E. Plödereder, L. Grunske, E. Schneider, D. Ull (Hrsg.): INFORMATIK 2014

E. Plödereder, L. Grunske, E. Schneider, D. Ull (Hrsg.)

44. Jahrestagung der Gesellschaft für Informatik INFORMATIK 2014

Big Data – Komplexität meistern

22. - 26. September 2014
in Stuttgart, Deutschland

Gesellschaft für Informatik e.V. (GI)

Lecture Notes in Informatics (LNI) - Proceedings
Series of the Gesellschaft für Informatik (GI)

Volume P-232

ISBN 978-3-88579-626-8
ISSN 1617-5468

Volume Editors
Prof. Dr. Erhard Plödereder
Prof. Dr. Lars Grunske
Universität Stuttgart, Institut für Softwaretechnologie
Universitätsstr. 38, 70569 Stuttgart, Germany
E-Mail: erhard.ploedereder@informatik.uni-stuttgart.de
lars.grunske@informatik.uni-stuttgart.de
Eric Schneider
Dominik Ull
Universität Stuttgart, Institut für Technische Informatik
Pfaffenwaldring 47, 70569 Stuttgart, Germany
E-Mail: eric.schneider@informatik.uni-stuttgart.de
dominik.ull@informatik.uni-stuttgart.de

Series Editorial Board
Heinrich C. Mayr, Alpen-Adria-Universität Klagenfurt, Austria
(Chairman, mayr@ifit.uni-klu.ac.at)
Dieter Fellner, Technische Universität Darmstadt, Germany
Ulrich Flegel, Hochschule für Technik, Stuttgart, Germany
Ulrich Frank, Universität Duisburg-Essen, Germany
Johann-Christoph Freytag, Humboldt-Universität zu Berlin, Germany
Michael Goedicke, Universität Duisburg-Essen, Germany
Ralf Hofestädt, Universität Bielefeld, Germany
Michael Koch, Universität der Bundeswehr München, Germany
Axel Lehmann, Universität der Bundeswehr München, Germany
Peter Sanders, Karlsruher Institut für Technologie (KIT), Germany
Sigrid Schubert, Universität Siegen, Germany
Ingo Timm, Universität Trier, Germany
Karin Vosseberg, Hochschule Bremerhaven, Germany
Maria Wimmer, Universität Koblenz-Landau, Germany

Dissertations
Steffen Hölldobler, Technische Universität Dresden, Germany
Seminars
Reinhard Wilhelm, Universität des Saarlandes, Germany
Thematics
Andreas Oberweis, Karlsruher Institut für Technologie (KIT), Germany

© Gesellschaft für Informatik, Bonn 2014
printed by Köllen Druck+Verlag GmbH, Bonn

Vorwort der Herausgeber

Die 44. Jahrestagung des Gesellschaft für Informatik (GI), die INFORMATIK 2014, findet vom 22. bis 26. September 2014 auf dem Vaihinger Campus der Universität Stuttgart statt. Zwei Thementage, vier assoziierte Konferenzen, rund dreißig Workshops, acht Tutorien und ein interessantes Rahmenprogramm sorgen für ein breit gefächertes Angebot an Vorträgen, Diskussionen und Impulsreferaten. Die Konferenzregistrierung ermöglicht die individuelle Programmkombination aus allen Konferenzangeboten. Damit kann jeder Teilnehmer seine individuelle Wahl aus dem großen Spektrum an Fortbildungsmöglichkeiten treffen. Nicht ohne Grund ist die Jahrestagung der GI die größte jährliche deutschsprachige Informatikkonferenz.

Der Tagungsband enthält auf mehr als 2500 Seiten die dauerhafte Dokumentation von über 250 Fachbeiträgen der diesjährigen Konferenz. Noch nie haben wir als Herausgeber so umfangreiche Proceedings bearbeitet. Der Tagungsband wird in Form eines E-Books an alle Teilnehmer verteilt.

Mit dem dualen Motto der Konferenz „Big Data – Komplexität meistern“ haben wir zwei Themen aufgegriffen, die uns in der Planung vor zwei Jahren als die aktuell wichtigsten, aber durchaus getrennten Themen der Informatik erschienen. Seitdem gab es eine Flut an „Big Data“-Konferenzen, die die wachsende Bedeutung dieses Themas unmittelbar zum Ausdruck brachte. Mit der zu meisternden Komplexität wollen wir die explosiv zunehmenden Anforderungen an die digitalen Systeme in unserem täglichen Leben thematisieren: Die Fahrerassistenzsysteme in Autos, die wachsende Vernetzung der täglich verwendeten Geräte wie Smartphones, Tablets oder Verkaufsautomaten, die Robotik in der industriellen Produktion, die Vernetzung der Börsen, der Internethandel, die Werbeeinblendungen der Suchmaschinen usw.; die Liste ließe sich beliebig fortsetzen. Hier zeigen die industrielle Realität und unsere Nutzererfahrung, wie wichtig es geworden ist, die Komplexität und damit die Zuverlässigkeit der Systeme zu beherrschen. „Industrie 4.0“ oder „Cyber-Physical Systems“ sind Schlagworte, die diesen Bereich abstecken. Seit den Planungstagen der Konferenz konnten wir beobachten, wie diese beiden Themenbereiche mit „Big Data“ zusammenwuchsen. Als übergreifende und in der Vergangenheit oft vernachlässigte Fragen kristallisierten sich die Sicherheit und der Schutz der Daten und Systeme sowohl bei „Big Data“ als auch bei den komplexen Systemen der Industrie und Wirtschaft heraus. Das Augenmerk der Öffentlichkeit hat sich auf diese Themen gerichtet, so dass unsere Themenschwerpunkte nicht nur für Fachleute aktuell und interessant sind.

Erstmals findet auf der Jahrestagung neben dem Tag der Informatik ein zweiter Thementag statt, der Tag der Wirtschaft, mit dem wir demonstrieren, dass die Informatik der Industrie und der Wirtschaft ganz wesentliche Impulse gibt. Und natürlich braucht auch die Informatik die Impulse aus Industrie und Wirtschaft.

Eine Konferenz dieser Dimension und Vielfalt zu planen und durchzuführen, erfordert ein großes, auch örtlich verteiltes Team an Helfern und Entscheidungsträgern, die ihre jeweilige Kompetenz und Arbeitskraft einbringen. Konferenz- und Workshopchairs, Programmkomitees und Tutoren sorgen für die Qualität des wissenschaftlichen Programms.

Ein industrieller Beirat unterstützt uns tatkräftig bei der Gestaltung der Konferenz und insbesondere des Tags der Wirtschaft. Viele Sponsoren ermöglichen ein ansprechendes Ambiente der Konferenz. Autoren stellen ihre Resultate zur Publikation bereit. Eingeladene Redner tragen mit ihren richtungsweisenden Impulsvorträgen wesentlich zur Attraktivität der Konferenz bei. Auch wenn wir hier auf die Nennung einzelner Namen verzichten, sollten Sie sich als Mitglied einer dieser Gruppen doch unseres herzlichen und persönlichen Danks gewiss sein. Ohne Sie wäre eine Tagung dieser Art nicht möglich!

Nicht zuletzt danken wir auch den Teilnehmerinnen und Teilnehmern der Konferenz für ihr Kommen. Wir wünschen allen einen hohen Erkenntnisgewinn, interessante Gespräche, wichtige Kontakte und einen erinnerungswürdigen Aufenthalt in Stuttgart.

Stuttgart, im August 2014

Erhard Plödereder
Lars Grunske
Eric Schneider
Dominik Ull

Grußwort der Bundesministerin für Bildung und Forschung, Prof. Dr. Johanna Wanka – Schirmherrin der INFORMATIK 2014 –

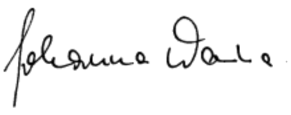
Die Digitalisierung beeinflusst unseren Alltag in zunehmendem Maße. Sei es beim Lesen von E-Mails auf dem Tablet-PC oder beim Fahrkartenkauf über das Smartphone: Digitale Technologien durchdringen fast alle gesellschaftlichen Bereiche.

Das Wissenschaftsjahr 2014 – Die digitale Gesellschaft befasst sich mit der Frage, was die Digitalisierung für unser Leben bedeutet. Dabei setzen wir uns sowohl mit den Chancen als auch mit den Herausforderungen auseinander, die mit der digitalen Entwicklung verbunden sind. Wichtig dabei ist, über den Umgang mit immer größer werdenden Datenmengen, bekannt als „Big Data“, nachzudenken.

Die Gesellschaft für Informatik greift dieses Thema in ihrer Jahrestagung „Big Data – Komplexität meistern“ auf und bietet eine Plattform zum Austausch. Ihren Mitgliedern kommt dabei eine wichtige Rolle zu. Mit ihrer Forschung legen sie die Grundlage für die Zukunft der digitalen Entwicklung und „Big Data“ in Deutschland. Gleichzeitig liefern sie fundamentale Beiträge zur gesellschaftlichen Debatte um die Digitalisierung. Dazu möchte ich auch weiterhin alle Mitglieder ausdrücklich ermuntern.

Wir benötigen Wissenschaft und Forschung, um auf die Existenz großer Datenmengen angemessen reagieren zu können. Die riesigen Datenschätze müssen verantwortungsvoll genutzt werden. Intelligente Methoden und Werkzeuge zu Datenmanagement und -analyse verschaffen uns in vielen Bereichen strategisch klare Vorteile und stärken unsere Wettbewerbsfähigkeit. Dies gilt jedoch nur, wenn wir gleichzeitig den Missbrauch der Daten und Möglichkeiten von „Big Data“ verhindern. Deshalb sind Datensicherheit, Datenschutz und Privatheit zentrale Themen, sie sind untrennbar mit „Big Data“ verbunden.

Allen Teilnehmerinnen und Teilnehmern der Jahrestagung wünsche ich spannende und erkenntnisreiche Vorträge und Workshops sowie lebhafte und kontroverse Diskussionen.



Prof. Dr. Johanna Wanka
Bundesministerin für Bildung und Forschung

Grußwort des Bundesministers für Wirtschaft und Energie, Sigmar Gabriel

Die riesigen Datenmengen, die Tag für Tag erzeugt und verarbeitet werden, haben einen hohen Wert. Sie sind einerseits ein sensibles und hochgradig schützenswertes Gut und andererseits ein immer wichtigerer Rohstoff für unsere Wirtschaft. Die intelligente Erhebung und Nutzung von Daten in industriellen Prozessen wird Deutschlands starke Branchen wie den Automobilbau, die Elektrotechnik, den Maschinen- und Anlagenbau oder die Energiewirtschaft modernisieren und ihre Wettbewerbsfähigkeit verbessern. Der globale Wachstumsmarkt für Datentechnologien bietet zudem zusätzliche ökonomische Chancen für etablierte IKT-Unternehmen und schafft Raum für neue Geschäftsmodelle unserer Start-ups.

Wir stehen aber erst am Beginn dieser Entwicklung und müssen alle Potenziale nutzen. Mit der Digitalen Agenda wird die Bundesregierung die Verzahnung von traditionell starken Anwenderindustrien mit der Digitalen Wirtschaft vorantreiben. Die besondere Herausforderung ist dabei nicht mehr, große Mengen an Daten sammeln und speichern zu können. Vielmehr geht es darum, aus dem Meer an Informationen die relevanten Daten herauszufiltern und dabei gleichzeitig höchsten Datenschutzstandards zu genügen. „Big Data wird Smart Data“ war deshalb auch das Leitthema der „Big Data Days“, die im November 2013 gemeinsam vom Bundeswirtschaftsministerium und der Gesellschaft für Informatik veranstaltet wurden.

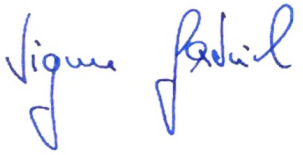
Gleichzeitig haben wir Ende vergangenen Jahres den Technologiewettbewerb „Smart Data – Innovationen aus Daten“ gestartet. Er bildet eine gute Grundlage, um Forschung und Entwicklung für Big Data-Lösungen „Made in Germany“ zu fördern. Ziel ist es, die Schlüsselbranchen der deutschen Wirtschaft und insbesondere unseren Mittelstand bei der Digitalisierung zu unterstützen. Der wichtigste Aspekt in diesem „Smart Data“-Wettbewerb ist konsequenterweise Klasse statt Masse: Es geht vor allem um Datenschutz und Urheberrecht sowie Sicherheit und Qualität der Daten.

Die durch das Bundeswirtschaftsministerium ausgewählten Smart Data-Projekte können sich nun hier auf der Jahrestagung der Gesellschaft für Informatik am Tag der Wirtschaft erstmals öffentlich vorstellen. An den 13 Projekten sind innovative Unternehmen unterschiedlichster Branchen – über die Hälfte aus dem Mittelstand – sowie zahlreiche Forschungsinstitute beteiligt. Die Projektideen greifen ein breites Spektrum an Forschungsfragen auf und versprechen ein großes Marktpotenzial. Ich wünsche den Forschern und Entwicklern viel Erfolg und vor allem eine schnelle Umsetzung in marktgerechte Produkte und Dienstleistungen.

Unser Ziel ist es, technische Innovationen und unternehmerisches Handeln in Deutschland deutlich stärker zu fördern. Nur so können wir den großen Herausforderungen unserer Zeit erfolgreich begegnen. Deshalb habe ich auch gerne die Schirmherrschaft für den Innovations- und Entrepreneurpreis der Gesellschaft für Informatik übernommen. Ich wünsche den Preisträgern, die hier auf der Jahrestagung ausgezeichnet werden, viel Erfolg bei der Umsetzung ihrer Ideen und gute Geschäftsabschlüsse.

Ich hoffe, dass der Kongress „Big Data – Komplexität meistern“ Ihnen viele nützliche Informationen über die Fortschritte der Wissenschaft und die neusten Entwicklungen aus den Unternehmen liefern kann. Nutzen auch Sie die Chancen der Digitalisierung!

Ihr



Sigmar Gabriel
Bundesminister für Wirtschaft und Energie

Grußwort des Ministerpräsidenten des Landes Baden-Württemberg, Winfried Kretschmann

Zur Jahreskonferenz der Gesellschaft für Informatik heiße ich alle Teilnehmerinnen und Teilnehmer herzlich in unserer Landeshauptstadt willkommen und wünsche insbesondere den auswärtigen Gästen einen angenehmen Aufenthalt in Baden-Württemberg.

Mit dem Tagungsmotto „Big Data – Komplexität meistern“ haben die Organisatorinnen und Organisatoren ein hochaktuelles Thema gewählt. Als maßgebliche Querschnittstechnologie des 21. Jahrhunderts ist die Informations- und Kommunikationstechnologie (IKT) aus dem Leben der Menschen in Deutschland und Baden-Württemberg nicht mehr wegzudenken. Sie beeinflusst alle Bereiche des täglichen Lebens, sei es Medizin, Kommunikation oder die Unterhaltungs- und Freizeitbranche. Die Fülle an verfügbaren Daten und Informationen verändert die wissenschaftliche Arbeit und vermag unternehmerisches Handeln zu beeinflussen: Trendforschung allgemein, aber auch Prognosen für wirtschaftliche Entscheidungen bekommen durch die Analyse großer Datenmengen eine neue Qualität und Aussagekraft. Big Data und Smart Data erhalten für Wissenschaft und Wirtschaft damit einen neuen Stellenwert.

Bei der Lösung wirtschaftlicher und gesellschaftlicher Probleme werden die Informations- und Kommunikationstechnologien zukünftig eine noch größere Rolle einnehmen als bislang. Der Bedarf an intelligenten und vernetzten Konzepten und Systemen wird zweifelsohne steigen – ebenso wie das Datenaufkommen. Diesem steigenden Bedarf muss allerdings gleichzeitig durch höhere Anforderungen an die Verlässlichkeit und Sicherheit der Konzepte und Systeme Rechnung getragen werden. Von der Jahreskonferenz der Gesellschaft für Informatik werden zu diesem sowie zu weiteren Themen sicherlich wichtige Impulse ausgehen.

Die baden-württembergische Landesregierung ist sich der großen Bedeutung der Informations- und Kommunikationstechnologien bewusst. Wir arbeiten deshalb mit dem Strategiekonzept „Forward IT“ und dem „E-Science“-Konzept daran, die Innovationskraft des Landes durch die Bereitstellung einer leistungsfähigen, effizienten und innovativen Rechnerinfrastruktur, durch die Vernetzung von Wissenschaft und Wirtschaft und durch die Ausbildung von wissenschaftlichem Nachwuchs und qualifizierten Fachkräften nachhaltig zu stärken und weiter voranzubringen.

Der Veranstaltung wünsche ich einen erfolgreichen Verlauf sowie allen Teilnehmerinnen und Teilnehmern interessante und anregende Begegnungen und Gespräche.

Winfried Kretschmann
Ministerpräsident des Landes Baden-Württemberg

Grußwort des Präsidenten der Gesellschaft für Informatik, Prof. Dr.-Ing. Peter Liggesmeyer

Sehr geehrte Damen und Herren, liebe GI-Mitglieder,

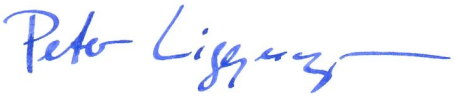
herzlich willkommen in Stuttgart zur INFORMATIK 2014, der 44. Jahrestagung der Gesellschaft für Informatik e.V. (GI). Ich freue mich sehr, mit Ihnen spannende Tage zum Motto unserer Tagung „Big Data – Komplexität meistern“ zu verbringen.

„Big Data“ – dieses Thema bewegt uns schon eine Zeit lang, und hat dennoch nichts von seiner Attraktivität, aber auch seiner Brisanz verloren. Unsere vernetze Gesellschaft ist auf den Zugang zu Daten mit den entsprechenden Chancen und Risiken angewiesen – *Social Media* und *Snowden* sind nur zwei Stichworte. Natürlich gilt es, die Chancen von Big Data zu nutzen und die damit verbundenen Herausforderungen zu meistern. Informatik bietet dafür die Lösungen. An einem eigenen „Tag der Wirtschaft“ stellen Referenten das Thema „Big Data“ und den Umgang damit deshalb explizit auch aus Unternehmenssicht dar. Big Data bietet bemerkenswerte Chancen – und die wollen wir ergreifen, ohne unakzeptabel große Risiken eingehen zu müssen.

Ganz besonders freue ich mich, dass Bundesbildungsministerin Frau Prof. Dr. Johanna Wanka die Schirmherrschaft über unsere Tagung übernommen hat – zeigt dies doch sehr deutlich, welche Bedeutung auch in der Politik unserer Disziplin beigemessen wird.

Neben dem wissenschaftlichen Programm – bestehend aus Tutorien, Workshops, einem Studierenden- und einem Doktorandenprogramm und dem Tag der Informatik – hat das Tagungsteam unter Leitung von Erhard Plödereder ein attraktives Rahmenprogramm auf die Beine gestellt, unter anderem einen Empfang im Rathaus auf Einladung des Bürgermeisters der Stadt Stuttgart, Fritz Kuhn, einen Festabend im Mercedes-Benz-Museum, einen Science Slam und einen Abend bei der IBM. Dem gesamten Team, dem Verein Informatik-Forum Stuttgart infos e.V. und auch den beteiligten Partnern und Sponsoren gebührt mein Dank. Ohne dieses große, von vielen Schultern getragene Engagement könnten wir Ihnen keine so attraktive und vielseitige Tagung bieten.

Ich bin sicher, mit Ihnen interessante, lehrreiche und schöne Tage zu erleben!



Prof. Dr.-Ing. Peter Liggesmeyer
Präsident der Gesellschaft für Informatik e.V. (GI)

Inhalt

Tag der Informatik

Manfred Broy
Cyber-Physical Systems - Konsequenzen für eine Informatik mit Nähe zur physikalischen Wirklichkeit 41

Karl-Heinz Streibich
Smart Big Data 49

Peter Schaar
Big Data mit Datenschutz - Mission impossible? 51

Stefan Wrobel
Big Data Analytics - Vom Maschinellen Lernen zur DataScience 53

Gerhard Weikum
Big Text: von Sprache zu Wissen 55

Peter Sanders
Algorithm Engineering for Big Data 57

Simone Rehm
Vorstellung ausgewählter „Grand Challenges" der Informatik 59

Tag der Wirtschaft

Lothar Baum, Wolfgang Blochinger, Michael Peters, Jörg Sommer
Big Data im Internet der Bosch-Dinge und -Dienste 63

Klaus Bauer
Industrie 4.0: Potenziale intelligenter Vernetzung 67

Bernhard Neumair, Achim Streit
Unterstützung datenintensiver Forschung am KIT - Aktivitäten, Dienste und Erfahrungen 69

Thomas Liedtke, Bernhard Hohlfeld
Gibt es sichere Software? 71

Andreas Tönne
„Eventually Consistent" - Neubetrachtung von Enterprise-Anwendungsarchitekturen unter BigData 75

Benjamin Klöpper, Jan Christoph Schlake
Aufbrechen der Datensilos - Big Data Forschungsfragen aus dem Bereich Industrial Analytics 79

Holger Petersen
Professionelles Testmanagement in Datenreinigungsprozessen 83

Simulation Technology: Systems for Data Intensive Simulations

Vorwort: Simulation Technology: Systems for Data Intensive Simulations (SimTech@GI) 87

Christoph Dibak, Boris Koldehofe
Towards Quality-aware Simulations on Mobile Devices 89

Andreas Weiß, Dimka Karastoyanova, David Molnar, Siegfried Schmauder
Coupling of Existing Simulations using Bottom-up Modeling of Choreographies 101

Benjamin Uekermann, Bernhard Gatzhammer, Miriam Mehl
Coupling Algorithms for Partitioned Multi-Physics Simulations 113

Michael Hahn, Dimka Karastoyanova
Configurable and Collaborative Scientific Workflows 125

Patrick Boba, Stefan Gries, Kay Hamacher
R as an Integration Tool in High Performance Computing - Lessons Learned 137

Tanja Blascheck, Karolina Vukojevic-Haupt, Dominik Weber, Dimka Karastoyanova, Thomas Ertl
Towards Automated Analysis of Eye Tracking Studies using the Workflow Technology 149

Rodrigo Iza Teran, Jochen Garcke
Data Analytics for Simulation Repositories in Industry 161

Big Data in Mobility and Logistics

Vorwort: Big Data in Mobility and Logistics (BDMobiLog 2014) 171

Volker Frehe, Thorsten Kleinschmidt, Frank Teuteberg
Big Data in Logistics - Identifying Potentials through Literature, Case Study and Expert Interview Analyses 173

Deniz Özcan, Michael Fellmann, Oliver Thomas
Towards a Big Data-based Technical Customer Service Management 187

Jürgen Antes, Robert T. Firla, Martin Thust
Insights, Analytics & Modelling - Wie Reisende den ICE nutzen und nutzen werden .. 199

Bereitstellung und Verwaltung von portablen und sicheren Cloud-Diensten

Vorwort: Bereitstellung und Verwaltung von portablen und sicheren Cloud-Diensten (CloudCycle14) 209

Christian Jung, Andreas Eitel, Reinhard Schwarz
Enhancing Cloud Security with Context-aware Usage Control Policies 211

Kevin Körner, Thomas Walter
Sichere und benutzerfreundliche Schlusselverteilung auf Basis von QR-Codes 223

Antonio Brogi, Jacopo Soldani
Reusing cloud-based services with TOSCA 235

Pascal Hirmer, Uwe Breitenbücher, Tobias Binz, Frank Leymann
Automatic Topology Completion of TOSCA-based Cloud Applications 247

Simon Harrer, Jörg Lenhard, Guido Wirtz, Tammo van Lessen
Towards Uniform BPEL Engine Management in the Cloud 259

Guntram Flach, Michael Kubach, Eray Oezmue, Immo Wehrenberg
Realisierung von vertrauenswürdigem Cloud-Computing mit SkIDentity: Der Anwendungsfall in Wertschöpfungsketten der Automobilindustrie 271

Alexander Blehm, Volha Kalach, Alexander Kicherer, Gustav Murawski, Tim Waizenegger, Matthias Wieland
Policy-Framework - Eine Methode zur Umsetzung von Sicherheits-Policies im Cloud-Computing 277

Thomas Kunz, Annika Selzer, Ulrich Waldmann
On the Measurement of Data Protection Compliance of Cloud Services 289

Alexander Grzesik, Torsten Frank
Sicheres Cloud Computing im Gesundheitswesen 297

Gamification und Virtuelle Welten

Vorwort: Gamification und Virtuelle Welten (GameWorlds) 307

Christian W. Scheiner
Online Ideation Games as Means to Change 309

Thomas Wiegand, Stefan Stieglitz
Serious Fun - Effects of Gamification on Knowledge Exchange in Enterprises 321

Benjamin Heilbrunn, Philipp Herzig, Alexander Schill
Towards Gamification Analytics - Requirements for Monitoring and Adapting Gamification Designs 333

Nicolai Hanner, Thorsten Pröhl, Rüdiger Zarnekow
Smarter Games durch Big Data Analysen: Potentiale und Handlungsfelder 345

Maike Müller, Laura Lenz, Martin R. Wolf
Steigerung der Effektivität durch Prozessmodellierungs-Tools - Änderbarkeit und Transparenz von Prozessen am Beispiel der Business Simulation Apollo 13 357

Stine Labes
A game-based evaluation model for a successful cooperation in cloud computing 371

Silke Plennert, Susanne Robra-Bissantz
Ein Serious Game als neue qualitative Erhebungsmethode für die Customer Journey beim Automobilkauf 383

Enterprise Architekturen mit Big Data & Cloud

Vorwort: Enterprise Architekturen für Big Data & Cloud 401

Rainer Schmidt, Michael Möhring, Ralf-Christian Härting, Alfred Zimmermann, Andreas Graule, Marius Schwarzer, Kristian Goßen
Datenzentrierte Unternehmensarchitekturen im Bekleidungseinzelhandel 405

Alfred Zimmermann, Kurt Sandkuhl, Rainer Schmidt, Dierk Jugel, Matthias Wisotzki, Michael Möhring
Adaptive Digitale Enterprise Architekturen für Big Data und Cloud-Systeme 417

Kurt Sandkuhl, Matthias Wißotzki
Towards Data Supply Chains in Enterprise Architecture Management 429

Santiago Gómez Sáez, Vasilios Andrikopoulos, Frank Leymann
Performance-aware Application Distribution in the Cloud 443

Oliver Bossert, Michael Falkenthal, Mark Mauerwerk, Gunther Piller
Einfluss von Cloud Services auf Auswahlverfahren für IT-Lösungen 453

Andreas Jede, Frank Teuteberg
Cloud-gestützte Supply-Chain-Architekturen - Entwicklung eines Referenzmodells auf Basis einer systematischen Literaturanalyse 463

Applications of Semantic Technologies

Vorwort: Applications of Semantic Technologies (AST 2014) 477

Timo Stegemann, Jürgen Ziegler
SemwidgJS: A Semantic Widget Library for the Rapid Development of User Interfaces for Linked Open Data 479

Eric Forkel, Janek Götze, Michael Clauß, Christan-Andreas Schumann
Facility Management in the Context of Self-Organisation 491

Nils Masuch, Christian Kuster, Sahin Albayrak
Semantic Service Manager - Enabling Semantic Web Technologies in Multi-Agent Systems 499

Recht und Technik: Datenschutz und Informationsmanagement im Diskurs

Vorwort: Recht und Technik - Datenschutz und Informationsmanagement im Diskurs (RuT) 513

Thomas Bräuchle
Die datenschutzrechtliche Einwilligung in Smart Metering Systemen - Kollisionslagen zwischen Datenschutz- und Energiewirtschaftsrecht 515

Nils Gruschka, Meiko Jensen
Aligning User Consent Management and Service Process Modeling 527

Anke Brummund
Smartphones und Apps: Datenschutzrechtliche Risiken und deren Begrenzung 539

Markus Kring
Big Data und der Grundsatz der Zweckbindung 551

Matthias Huber, Jörn Müller-Quade, Tobias Nilges, Carolin Thal
A Provably Privacy Preserving Video Surveillance Architecture for an Assisted Living Community 563

Matthias Rodeck, Caroline Voigt, Arndt Schnütgen, Ina Schiering, René Decker
Toolgestützte Assessments zu Datenschutz und Datensicherheit in kleinen und mittelständischen Unternehmen 575

Second European STAMP Workshop

Vorwort: Second European STAMP Workshop (ESW 2014) 589

Rui Wang, Wei Zheng, Ci Liang, Gilles Motet
An Integrated Hazard Identification Method for Socio-technical Systems based on STPA 591

Nancy G. Leveson
Using STAMP to Develop Leading Indicators 597

Cody H. Fleming
Model-based Concept Development and Safety Driven Design 601

Robert J. de Boer, Raymon van der Maarel
Application of STAMP to Improve the Evaluation of Safety Management Systems 607

Brian Edwards
Utilising STAMP to define the First-into-Man system for medicines 613

Mitka Eleftheria
Applying STAMP to safety standards of mowing robots 619

Yasuhiko Kawabe, Tatsuya Yanagisawa
Applying STAMP/STPA to Human Safety System for Four Wheel Drive Power-train .. 625

Aubrey Samost
Evaluating Systems with Multiple Processes Using STPA: A Case Study in a Medical Intensive Care Unit 637

John Thomas
Developing systematic procedures to identify causal factors in Systems Theoretic Process Analysis 641

Kip Johnson, Nancy Leveson
Investigating Safety and Cybersecurity Design Tradespace for Manned-Unmanned Aerial Systems Integration Using Systems Theoretic Process Analysis 643

Rejzek Martin, Christian Hilbes
Use of STPA in digital instrumentation and control systems of nuclear power plants . 649

Future Search Engines - Big Data und semantische Technologien in Backend und Frontend von Suchmaschinen

Vorwort: Future Search Engines - Big Data und semantische Technologien in Backend und Frontend von Suchmaschinen (FoRESEE) 653

Florian Haag, Qi Han, Markus John, Thomas Ertl
Aspect Grid: A Visualization for Iteratively Refining Aspect-Based Queries on Document Collections 655

Roger Schaer, Dimitrios Markonis, Henning Müller
Architecture and applications of the Parallel Distributed Image Search Engine (ParaDISE) 661

Alexandr Uciteli, Christoph Goller, Patryk Burek, Sebastian Siemoleit, Breno Faria, Halyna Galanzina, Timo Weiland, Doreen Drechsler-Hake, Wolfram Bartussek, Heinrich Herre
Search Ontology, a new approach towards Semantic Search 667

Martin Voigt, Daniel Hladky
Semantik-gestutzte Analyse von und Suche in Kundenspezifikationen im Maschinenbau 673

Mandy Keck, Martin Herrmann, Dana Henkens, Severin Taranko, Viet Nguyen, Fred Funke, Steffi Schattenberg, Andreas Both, Rainer Groh
Visual Innovations for Product Search Interfaces 679

Data Management in the Cloud

Vorwort: Data Management in the Cloud (DMC2014) 691

Steffen Friedrich, Wolfram Wingerath, Felix Gessert, Norbert Ritter
NoSQL OLTP Benchmarking: A Survey 693

Andreas Meister, Sebastian Breß, Gunter Saake
Cost-Aware Query Optimization during Cloud-Based Complex Event Processing 705

Erhard Rahm, Wolfgang E. Nagel
ScaDS Dresden/Leipzig: Ein serviceorientiertes Kompetenzzentrum für Big Data 717

Volker Markl
Breaking the Chains: On Declarative Analysis and Independence in the Big Data Era 719

Harald Schöning
Smart Data Innovation Lab - eine Forschungsplattform mit realen Daten 721

Felix Gessert, Steffen Friedrich, Wolfram Wingerath, Michael Schaarschmidt, Norbert Ritter
Towards a Scalable and Unified REST API for Cloud Data Stores 723

Peter Reimann, Tim Waizenegger, Matthias Wieland, Holger Schwarz
Datenmanagement in der Cloud fur den Bereich Simulationen und Wissenschaftliches Rechnen 735

Stefanie Scherzinger
Datenmodellierung in der Anwendungsentwicklung mit NoSQL-Datenbanken 747

Torsten Steinbach, Michael Wurst
IBM's Analytic Warehouse Cloud Service 751

Andreas Müller, Sebastian Hudert, Victor Fäßler
Privacy-protecting Data Management in distributed Cloud Infrastructures 753

Dynamik und Compliance in IT-gestützten Geschäftsprozessen

Vorwort: Dynamik und Compliance in IT-gestützten Geschäftsprozessen (DyCo 2014) 757

Daniel Burgwinkel
Unterstützung von dynamischen Compliance Prozessen durch Business Process Management Plattformen 759

Monika Kochanowski, Christoph Fehling, Falko Koetter, Frank Leymann, Anette Weisbecker
Compliance in BPM today - an insight into experts' views and industry challenges ... 769

Jens Drawehn, Dominic Walter
Soziale Software und Dynamik im Geschäftsprozessmanagement 781

Michael Becker, Stephan Klingner, Michael Sonnenberg
Anbieterübergreifende Konfiguration komplexer Dienstleistungsportfolios: Ein Anwendungsfall im Bereich erneuerbarer Energien 793

David Hopfmüller, Sigrid Schefer-Wenzl, Mark Strembeck
Kaskadierender Widerruf von Delegationen in prozessbasierten Informationssystemen 805

Stefan T. Ruehl, Urs Andelfinger, Andreas Rausch, Stephan A.W. Verclas
Who may share with Whom?! Mixed-Tenancy: Conditional Sharing of Cloud Resources 817

Management komplexer IT-Systeme und Anwendungen

Vorwort: Management komplexer IT-Systeme und Anwendungen (MITA 2014) 831

Wolfgang Johannsen
Konzepte der Information-Governance 833

Andreas Textor
Ein Ansatz fur eine Ontologie-basierte Verbindung von IT Monitoring und IT Governance 845

Peter Klotz
Disaster-Readiness als Beispiel für Ontologie-basiertes IT-Management im Geschäftskontext 857

Fabian Meyer
Ontologie-basiertes Monitoring von IT-Systemen 861

Cataldo Mega, Christoph Lange
Optimizing resource topologies of workload in the cloud by minimizing consumption and maximizing utilization while still meeting service level agreements 873

Michael Pisula
Agile Migration 883

Frank Termer, Volker Nissen, Nathalie Wessels
Geschäftsprozessflexibilität durch Service-orientierte Architekturen - grundlegende Zusammenhänge und weiterer Forschungsbedarf 889

Andreas Kiesow, Novica Zarvic, Oliver Thomas
Continuous Auditing in Big Data Computing Environments: Towards an Integrated Audit Approach by Using CAATTs 901

Maher Salem, Ulrich Buehler
A Comprehensive Model for Revealing Anomaly in Network Data Flow 913

Jan Schaefer
Semantic Self-Management and Mediation for Home Service Platforms 925

IT-Unterstützung in Emergency Management & Response

Vorwort: IT-Unterstützung in Emergency Management & Response (IT-Rettung) 939

Thomas Ludwig, Christian Reuter
Entwicklung einer mobilen Reporting-Applikation zur Artikulation entscheidungsrelevanter Informationsbedarfe im Katastrophenschutz 941

Silvia Krug, Maik Debes, Jochen Seitz
Framework zur Simulation von realistischen Einsatzszenarien 953

Johannes Sautter, Manuel Habermann, Sandra Frings, Friederike Schneider, Bernhard Schneider, Holger Bracker
Übungsunterstützung für Einsatztrainings des Massenanfalls von Verletzten (MANV) 965

Christian Reuter, Michael Ritzkatis, Thomas Ludwig
Entwicklung eines SOA-basierten und anpassbaren Bewertungsdienstes für Inhalte aus sozialen Medien 977

Sebastian Alsbach, Maria A. Wimmer
Brand- und Katastrophenschutzportal Rheinland-Pfalz als TETRA-Statusübermittler für die Feuerwehreinsatzzentralen des Bundeslandes 989

Datenmanagement und Management elektronischer Ressourcen in Bibliotheken

Vorwort: Datenmanagement und Management elektronischer Ressourcen in Bibliotheken (DERM 2014) 1003

Martin Czygan
Design and Implementation of a Library Metadata Management Framework and its Application in Fuzzy Data Deduplication and Data Reconciliation with Authority Data 1005

Andreas Nareike, Natanael Arndt, Norman Radtke, Sebastian Nuck, Leander Seige, Thomas Riechert
AMSL - Managing Electronic Resources for Libraries Based on Semantic Web 1017

Thomas Efer, Ninja Steinbach-Hüther
Quantitative Analyses in Global and Area Studies using Graph-based Filtering of Heterogeneous Catalogue Data 1027

BPM im Öffentlichen Sektor

Vorwort: Workshop zum „BPM im Öffentlichen Sektor“ 1041

René Schumann, Serge Delafontaine, Caroline Taramarcaz, Florian Evéquoz
Effective Business process documentation in federal structures 1043

Daniel Hadrian, Florian Evequoz
CARES: Requirements Specification with BPMN 2.0 in WTO Procurement 1059

Walter Berli, Daniel Lübke, Werner Möckli
Terravis - Large Scale Business Process Integration between Public and Private Partners 1075

André Göbel, Emanuel Hesse
Geschäftsprozessmanagement in der Wirtschaftsförderung. Status Quo zu Notwendigkeiten, Anwendungen und Hemmnissen. 1091

Eliane Maalouf, Maria Sokhn
Semantic based auto-completion of business process modelling in eGovernment 1107

Stefan Gröger, Matthias Schumann
Drittmittelprojekte in Hochschulen - Entwicklung eines fallstudienbasierten Referenzprozessmodells 1119

Oliver Arnold, Ralf Laue
Untersuchungen zur Beherrschbarkeit von Variabilität in Prüfungsprozessmodellen deutscher Hochschulen 1135

Dirk Frosch-Wilke, Rudolf Gaspary
Mitarbeiterwiderstände und -partizipation bei der Einführung von BPM - Ein Multi-Level-Ansatz am Beispiel des Landespolizeiamtes Schleswig-Holstein 1149

Konrad Walser
Das BPM-Standardisierungssystem in der Öffentlichen Verwaltung der Schweiz und dessen Theorie-basierte Evaluierung 1161

Big Data Technologien zur Analyse und Steuerung komplexer cyber-physischer Systeme

Vorwort: Big Data Technologien zur Analyse und Steuerung komplexer cyber-physischer Systeme 1177

Ilias Gerostathopoulos, Jaroslav Keznikl, Tomas Bures, Michal Kit, Frantisek Plasil
Software Engineering for Software-Intensive Cyber-Physical Systems 1179

Benjamin Beichler, Alexander Nitsch, Frank Golatowski, Christian Haubelt
Ein abstraktes SystemC-Modell zur Analyse und Leistungsabschätzung des europäischen Zugsicherungssystems ETCS 1191

Martin Kasparick, Frank Golatowski, Dirk Timmermann
Cyber-physische Systeme im OP-Saal - Ein Machbarkeitsnachweis 1203

Tobias Hoppe, Harald Eisenmann, Alexander Viehl, Oliver Bringmann, Christian Hennig
Sicherstellen semantischer Integrität im Big Data Umfeld - Problemstellung und Lösungsansatz 1215

Sebastian Schmitz, Julian Krenge
Big Data im Störmanagement 1229

Sebastian von Mammen, Carsten Grenz, Jörg Hähner, Sabine Timpf, Daniel Loebenberger, Stefan Mandl, Oleksandr Kozachuk
Organic Data: Ein sicheres, dezentralisiertes Big Data Konzept 1241

Julian Krumeich, Jens Schimmelpfennig, Dirk Werth, Peter Loos
Realizing the Predictive Enterprise through Intelligent Process Predictions based on Big Data Analytics: A Case Study and Architecture Proposal 1253

David Schumm, Anna Julia Spanke, Kirsten Dreier, Victor Fäßler, Birger Becker, Hartmut Schmeck
Konzept und Methodik zur Entwicklung einer mobilen digitalen Persönlichkeit für nutzergerechte Anwendungen 1265

Roboterkontrollarchitekturen

Vorwort: Roboterkontrollarchitekturen (ROBOCONTROL) 1281

Nikolaus Vahrenkamp, Mirko Wächter, Manfred Kröhnert, Peter Kaiser, Kai Welke, Tamim Asfour
High-Level Robot Control with ArmarX 1283

Matthias Lutz, Dennis Stampfer, Alex Lotz, Christian Schlegel
Service Robot Control Architectures for Flexible and Robust Real-World Task Execution: Best Practices and Patterns 1295

Steffen Schütz, Max Reichardt, Michael Arndt, Karsten Berns
Seamless Extension of a Robot Control Framework to Bare Metal Embedded Nodes 1307

Andreas Bihlmaier, Heinz Wörn
Increasing ROS Reliability and Safety through Advanced Introspection Capabilities 1319

Raphael Maas, Erik Maehle
Health Signal Generation in the ORCA (Organic Robot Control Architecture) Framework 1327

Elektronische Wahlen: Unterstützung der Wahlprozesse mittels Technik

Vorwort: Elektronische Wahlen: Unterstützung der Wahlprozesse mittels Technik 1341

Daniela Simic-Draws, Katharina Bräunlich
Prozessmodell und -analyse für die Stadtrats- und Kommunalwahl 2014 in Koblenz 1343

Ali Fawzi Najm Al-Shammari, Adolfo Villafiorita
Iraqi Elections in 2014: a Privacy Requirement Evaluation Based on a Polling Place Experience 1359

Robert Stein, Gregor Wenda
Online-Partizipation in der EU? - Erste Erfahrungen mit Europäischen Bürgerinitiativen aus Sicht Österreichs 1371

Robert Krimmer
Identifying Building Blocks of Internet Voting: Preliminary Findings 1381

Philipp Locher, Rolf Haenni
A Lightweight Implementation of a Shuffle Proof for Electronic Voting Systems 1391

Christian Feier, Stephan Neumann, Melanie Volkamer
Coercion-Resistant Internet Voting in Practice 1401

Jurlind Budurushi, Stephan Neumann, Genc Shala, Melanie Volkamer
Entwicklung eines Common Criteria Schutzprofils für elektronische Wahlgeräte mit Paper Audit Trail 1415

Robert Stein, Gregor Wenda
Das Zentrale Wählerregister – Ein skalierbares Instrument zur Bürgerbeteiligung mit 1:1-Verifikation 1427

Informationssysteme mit Open Source

Vorwort: Informationssysteme mit Open Source (ISOS 2014) 1439

Herbert Neuendorf
Potenziale von Server Push-Mechanismen im Kontext von Informationssystemen 1441

Martin Helmich
Systemarchitektur eines verteilen, ereignisgetriebenen Konfigurationsmanagement-Frameworks auf Basis von Open-Source-Komponenten 1453

Philipp Köppen, Stefan Langenberg
Erfahrungsbericht uber den Einsatz des open source Kommunikationsservers Mirth Connect am Universitätsklinikum Bonn 1465

Michael Hitz, Thomas Kessel
Erfolg innovativer Open-Source-Ansätze am Beispiel asynchroner Servertechnologien 1475

IT-Governance und Strategisches Informationsmanagement

Vorwort: IT-Governance und Strategisches Informationsmanagement (ITG-SIM) 1489

Franziska Jahn, Michael Schaaf, Barbara Paech, Alfred Winter
Ein semantisches Netz des Informationsmanagements im Krankenhaus 1491

Norbert Ahrend, Petra Wolf, Michael Räckers, Alexander Dentschev, Marlen Jurisch
Federal Information Management - Context and Effects 1499

Tobias Grönert, Jens Pöppelbuß, Andreas Breiter
Reifegradbestimmung der IT-Governance: Eine Fallstudie zur Anwendbarkeit des COBIT 5 PAM in der öffentlichen Verwaltung 1513

Konrad Walser, Matthias Goeken, Klaus Schoo, Can Adam Albayrak, Jörg Puchan, Donat Sterren, Jan Thielscher, Raymond Tischendorf, Matthias Knoll, Hans-Jürgen Ley, Jens Bartenschlager, Fritz Pörtig
StratIT - Ein Rahmenwerk zu den Inhalten von IT-Strategien. Grundlegung und methodische Fundierung 1527

Thomas Widjaja
Der kostenoptimale Grad funktionaler Redundanz in IT-Anwendungslandschaften .. 1539

Wiebke Kappenberg, Paul Drews
Softwarealterung aus Sicht des IT-Managements - Ergebnisse einer qualitativ-empirischen Analyse in der Finanzindustrie 1553

Thomas Bauer, Stephan Buchwald, Julian Tiedeken, Manfred Reichert
Konzeption eines SOA-Repository mit Analysefähigkeiten 1565

Rainer Schmidt, Michael Möhring, Robert Rieg, Stefan Maier, Gianfranco Walsh, Ralf-Christian Härting
Forensic Accounting - Potenziale von Big Data und Predictive Analytics im Rahmen der IT Governance 1577

Elektromobilität, Agenten und Smart Grids

Vorwort: Workshop „Elektromobilität, Agenten und Smart Grids“ (EASG) 1591

Thomas Görzen, Christian Meier, Dennis Kundisch
Ein Total Value-of-Ownership Modell zur Bewertung elektrifizierter Fahrzeugflotten 1593

Andreas Helferich, Norman Pelzl, Sixten Schockert
Unterstützung von KMU bei der Erbringung komplexer Mobilitätsservices in Koopkurrenznetzwerken 1605

Carola Gerwig, Dennis Behrens, Ralf Knackstedt, Helmut Lessing
Innovatives eCarSharing-Konzept am Beispiel des Hildesheimer Tandem-Modells .. 1617

Benjamin Klör, Sebastian Bräuer, Daniel Beverungen
A Business Process Model for the Reverse Logistics of Used Electric Vehicle Batteries 1631

Armin Wolf, Steffen Unger, Marc Schultz
Eigenverbrauchs- und Kostenoptimierung zur Untersuchung des Einflusses dynamischer Stromtarife auf die Netznutzung 1645

Beyond Data: Mastering Management of Complex Research Data and Processing Environments

Vorwort: Beyond Data: Mastering Management of Complex Research Data and Processing Environments (BeyondData2014) 1659

Rami Akkad, Henning Schmitz, André Deienno Pansani
Operating The HPI Future SOC Lab - An Infrastructure Laboratory For Researchers - A Practitioners' Report 1661

Dennis Zielke
LAUDATIO-Repository.org - A Modular Approach of an Open Access Research Data Repository for a discipline 1667

Frank Tristram, Dennis Wehrle, Uğur Çayoğlu, Jessica Rex, Dirk von Suchodoletz
Status Report of bwFDM-Communities - A State Wide Research Data Management Initiative 1669

Dennis Wehrle, Wiebke van Ekeris, Uli Hahn
Looking Beyond Data Through the Ages 1675

Hashim Iqbal Chunpir, Thomas Ludwig
Reviewing the Governance Structure of End-User Support in e-Science Infrastructures 1687

Volker Frehe, Vilius Rugaitis, Frank Teuteberg
Scientometrics: How to perform a Big Data Trend Analysis with ScienceMiner 1699

Michael Diepenbroek, Frank Oliver Glöckner, Peter Grobe, Anton Güntsch, Robert Huber, Birgitta König-Ries, Ivaylo Kostadinov, Jens Nieschulze, Bernhard Seeger, Robert Tolksdorf, Dagmar Triebel
Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio) 1711

Modeling and Optimization of high-Complex systems

Vorwort: Modeling and Optimization of high-Complex systems (MOC 2014) 1725

Anna Prenzel, Georg Ringwelski
Nine Interaction Guidelines for the Design of Human Computer Interaction in Detailed Scheduling Systems 1727

Studentische Ausbildung und berufliche Weiterbildung in Angewandter Softwarequalitätssicherung

Vorwort: SAG WAS – Studentische Ausbildung und berufliche Weiterbildung in Angewandter Softwarequalitätssicherung 1741

Axel Böttcher, Veronika Thurner, Daniela Zehetmeier
Der frühe Tester fängt den Bug 1743

Barbara Dörsam
Softwaretests- und Software-Qualitätsthemen in der Hochschullehre 1745

Karin Vosseberg
Studienprojekte als Experimentierraum für QS 1747

Andreas Spillner
Certified Tester: Ein Lehrplan für berufliche Weiterbildung und studentische Ausbildung 1749

System Software Support for Big Data

Vorwort: System Software Support for Big Data (BigSys) 1753

Edward Ching, Norbert Egi, Masood Mortazavi, Vincent Cheung, Guangyu Shi
Unleashing the Hidden Power of Integrated-GPUs for Database Co-Processing 1755

Fahad Khalid, Frank Feinbube, Andreas Polze
Pattern-guided Big Data Processing on Hybrid Parallel Architectures 1767

Alkida Balliu, Dennis Olivetti, Ozalp Babaoglu, Moreno Marzolla, Alina Sîrbu
BiDAl: Big Data Analyzer for Cluster Traces 1781

Kevin Beineke, Florian Klein, Michael Schöttner
Asynchronous Logging and Fast Recovery for a Large-Scale Distributed In-Memory Storage 1797

Hochschule 2020

Vorwort: Hochschule 2020 1813

Markus von der Heyde
CIO-Modelle an deutschen Hochschulen - Netzwerk der entscheidenden Faktoren .. 1815

Steven Strehl, Andreas König
Hochschulrechenzentren als IT-Stadtwerke - Konzeption und kritische Diskussion innovativer IT-Dienstangebote von Hochschulen 1825

Andreas Hartmann, Wilhelm Rossak
Eine IT-Referenzarchitektur für Hochschulen: Vom generischen Modell zur spezifischen Lösung 1837

Dietmar Zoerner, Florian Gößler, Ulrike Lucke
Hochschul-Apps im Überblick 1849

Patrick Bittner, Christopher Ritter, Thomas Hildmann
Get Your Study Plan 1861

Raphael Zender, Vyacheslav Rogozhkin, Ulrike Lucke
Einsatz von MultiTouch-Tischen zur kooperativen Semesterorganisation 1873

Umweltinformatik zwischen Nachhaltigkeit und Wandel

Vorwort: Umweltinformatik zwischen Nachhaltigkeit und Wandel (UINW) 1887

Maximilian Schneider, Kurt Weissenbach, Volker Wohlgemuth
EnergieNetz - Eine webbasierte, erweiterbare Open-Source-Software für das Energiemanagement in lernenden Energieeffizienz-Netzwerken 1889

Thomas Gutzke, Frank Reußner
Herausforderungen bei der Redigierung und Versionierung fortlaufender Massendaten 1901

Florian Stiel, Frank Teuteberg
Environmental-Oriented Information Systems Design - The Concept of Life Cycle Impact Modelling and its Application to Cloud Computing 1911

Eva Kern, Bianca Luther-Klee, Stefan Naumann, Andreas Filler
Kriterien für Webportale zur Unterstützung nachhaltiger Regionalentwicklung am Fallbeispiel "vitaminBIR" 1923

Benno Schmidt, Andreas Wytzisk
Software Engineering und Integrative Nachhaltigkeit 1935

Hans-Knud Arndt
Big Data oder Grand Management Information Design? 1947

The 4th International Workshop on Advanced Design Concepts and Practice

Vorwort: The 4th International Workshop on Advanced Design Concepts and Practice (ADCP2014) 1959

Laura T. Beilstein, Stephan Rudolph
Optimum function for minimum weight evaluation of structural joints 1961

Akram Chamakh, Dieter Roller
Intelligent requirement analyzis and detection for product development 1967

Chong Li, Zoltán Rusák, Imre Horváth, Yuemin Hou, Linhong Ji
Optimizing patients' engagement by a cyber-physical rehabilitation system 1971

Julian R. Eichhoff, Dieter Roller
Approach to Learning Production Rules for Grammar-Based Functional Design 1977

Dieter Roller, Erik Engesser
BPMN Process Design for Complex Product Development and Production 1979

Stefano Gagliardo, Franca Giannini, Marina Monti, Stefano Mottura, Marios Pitikakis
The Pipeline from CAD to Virtual Reality in VISIONAIR Project 1985

Fariz Muharram Hasby, Dieter Roller
Cooperative CAD System for Conceptual-Embodiment Design Phase of Product Design 1989

Imre Horváth, Zoltán Rusák, Yuemin Hou, Lihong Ji
On some theoretical issues of interaction with socialized and personalized cyber-physical systems 1995

Shahab Pourtalebi, Imre Horváth, Eliab Z. Opiyo
First steps towards a mereo-operandi theory for a system feature-based architecting of cyber-physical systems 2001

Zoltán Rusák, Imre Horváth, Yuemin Hou, Ji Lihong
Towards affordance based human-system interaction based on cyber-physical systems 2007

Hou Yuemin, Michel van Tooren, Ji Linhong
Algorithms for Computational Mechanism Synthesis and Analysis 2013

Luo Xin, Liang Chengyuan, Chen Xuedong
Process Modeling of Dynamics-Driven Structural Design of High-Tech Semiconductor Fabrication Equipment 2019

Jianfeng Sui, Linhong Ji
Intelligent Control based on Location-Force Model 2025

Fan Wu, Yixu Song, Nan Wu, Hongjun Yang, Xiaomin Sun
A GPU Based Parallel Computing Mechanism Of an Etching Profile Evolution Model 2031

Huang Zun-di, Yang Tie-niu, Li Chang-ming, Huang Hui, Zhang Qi-li, Hou Shu-guang
Design of Processing Chamber Based on the Flow Uniformity 2037

Yuan Fang, Wang Shibo, Zeng lizhan, Jin jianxin
Design and Analysis of a Lorentz Linear Motor for Precision Vibration-Isolation ... 2043

Wu Xiaojing, Cheng Jia, Ji Linhong, Hou Yuemin, Lu Yijia
A Kind of Multi-disciplinary Simulation and Design Platform for IC Chamber based on Commercial Solver 2049

Xinyu Guan, Jingyan Qin, Yuemin Hou, Ji Linhong
Study of Design Method for Tangible User Interface 2055

Jia Lin, Jia Cheng, Hou Yuemin, Ji Linhong
The Prototype of Design Product Knowledge Representing System 2057

Bernadetta Kwintiana Ane
Cellular Genetic Algorithm Feature-based Modelling in Product Data Exchange 2059

Risikokommunikation im Kontext von IT-Sicherheit und Safety

Vorwort: Risikokommunikation im Kontext von IT-Sicherheit und Safety (RiskKom 2014) 2067

Steffen Bartsch, Carina Boos, Gamze Canova, Dominic Dyck, Birgit Henhapl, Michael Schultheis, Melanie Volkamer
Interdisziplinäres Bewertungskonzept für Risiken auf Webseiten 2069

Jana Fruth, Matthias Thimm, Sven Kuhlmann, Jana Dittmann
Ein erster Prototyp: Sicherheitsguide für Grundschulkinder beim Umgang mit dem Internet 2081

Thomas Pfeiffer, Michaela Kauer and Julia Röth
"A Bank Would Never Write That!" A Qualitative Study on E-Mail Trust Decisions 2093

Automotive Software Engineering

Vorwort: Automotive Software Engineering (ASE) 2107

Sebastian Mattheis, Kazi Khaled Al-Zahid, Birgit Engelmann, Andreas Hildisch, Stefan Holder, Olexiy Lazarevych, Daniel Mohr, Felix Sedlmeier, Richard Zinck
Putting the car on the map: A scalable map matching system for the Open Source Community 2109

Thomas Stone, Olga Birth, André Gensler, Andreas Huber, Martin Jänicke, Bernhard Sick
Location based learning of user behavior for proactive recommender systems in car comfort functions 2121

Dominik Bauch, David Lenz, Pascal Minnerup, Andrei Avram
Interface Definition and Code Generation in heterogeneous Development Environments from a Single-Source 2133

Georg Macher, Eric Armengaud, Christian Kreiner
Automated Generation of AUTOSAR Description File for Safety-Critical Software Architectures 2145

Christian Berger, Michael Dukaczewski
Comparison of Architectural Design Decisions for Resource-Constrained Self-Driving Cars - A Multiple Case-Study 2157

Jörg Holtmann, Dimitar Shipchanov
Considering Architectural Properties in Real-time Play-out 2169

Remo Lachmann, Ina Schaefer
Towards Efficient and Effective Testing in Automotive Software Development 2181

Thomas Kuhn
Frühzeitiges Absichern von Entwurfsentscheidungen durch Virtuelle Deployments .. 2193

Didaktik und Grundlagen der Informatik, Festveranstaltung zum 70. Geburtstag von Volker Claus

Vorwort: Didaktik und Grundlagen der Informatik 2207

Ernst-Rüdiger Olderog
Von Petri-Spielen zu endlichen Automaten 2209

Karsten Weicker, Martin Waßmann
Evolution von LEGO-Bauwerken 2211

Oliver Vornberger
So mooc wi dat - Von der Vorlesung zum Massive Open Online Course 2213

Ralf Romeike
Wie Kinder und Jugendliche lernen, die digitale Gesellschaft mitzugestalten 2215

Doktoranden-Programm

Vorwort: Doktorandenworkshop 2014 2219

Dilshodbek Kuryazov
Delta Operation Language for Model Difference Representation 2221

Andreas Sailer
Towards an Automated Reverse Engineering of Design Models from Trace Recordings 2233

Joerg Schoenfisch
Querying Probabilistic Ontologies with SPARQL 2245

Mathias Schubanz
Towards Enabling Design Rationale Capture by Asking the Right Questions 2257

Tom Thaler
Towards Usability Mining 2269

Sebastian Bittmann
Cooperative-Intrinsic Planning and Model-Driven Design of Business Information Systems 2281

Rafiaa Boujbel
Channel classification with hidden Markov models in mobile networks 2287

Yihan Deng
SurgLink: Cross-media Linking in the Operating Room 2293

Hannes Grunert
Distributed Denial of Privacy 2299

Mirco Josefiok
Measuring and Monitoring the Rehabilitation of Patients on Monitoring Stations via the Analyses of poly-structured Data 2305

Wiltrud Kessler
Improving the Detection of Comparison Arguments in Product Reviews 2311

David Kuhlen
Wertanalyseverfahren für Kundenanforderungen 2317

Deniz Özcan
Towards a holistic Approach for Service Management in the Technical Customer Service Domain 2323

David Georg Reichelt
Berücksichtigung von Softwareperformanz im Entwicklungsprozess 2329

Studierenden Konferenz Informatik 2014

Vorwort: Studierendenkonferenz Informatik 2014 (SKILL 2014) 2337

Juliane Bredack
Automatische Extraktion von Fachterminologie aus kunst-historischen Volltexten ... 2341

Friedrich Pawelka, Thomas Wollmann, Jakob Stöber, Tommy Vinh Lam
Erfolgreiches Lernen durch gamifiziertes E-Learning 2353

Daniel Steiman
Context-Awareness in Multiagentensimulationen auf Basis von Complex Event Processing 2365

Christoph Winter
Entwicklung algorithmischer Skelette für CUDA am Beispiel von Affintiy Propagation 2377

Fritjof Bornebusch, Glaucia Cancino, Alvine Nzeungang Fanseu, Maryam Farajzadeh Jalali, Jamal Mohsen, Max Nitze, Christina Plump, Ronald Smith Djomkam Yotedje, Hubert Fred Tchambo, Toni, Henning Ziegler
Aspekt-basierte Sentiment Analysis 2389

Robert Fruth, Henry Moews, David Silva Ribeaux, Tom Tschernack, Norbert Wall, Patrick Wolfien
Campus-Navigations-Applikation für Studierende mit Beeinträchtigungen 2401

Ionut-Gabriel Farcas
Comparison of Numerical Methods in Uncertainty Quantification 2413

Steffen Hirschmann
GPU-Based Regression Analysis on Sparse Grids 2425

Johannes Obermaier, Tobias Laas, Markus Roner
Timing Attack on a Modified Dynamic S-box Implementation of the AES InvSubBytes Operation 2437

Anna-Luisa Schwartz
Analysis of Crash Simulation Data using Spectral Embedding with Histogram Distances 2449

Helena Glatzel
Schema-Management ohne Schema? - Schema-Verwaltung in NoSQL-Datenbanksystemen 2461

Markus Paatsch
Konzeption und Entwicklung eines Informationssystems zur Unterstützung der Vertriebsplanung im Lösungsvertrieb 2473

Fabian Trautsch
Evaluating the Combination of two User-oriented Usability Evaluation Methods: Thinking-aloud and Questionnaires 2485

Emanuel Jöbstl, Jérôme Urhausen, Elisaweta Masserova, Ainara Askar
sGrid: Centralized management of BOINC-based research projects - Summary 2487

Julius Höfler
Drohnensteuerung im Programmierunterricht 2489

Markus Schmidt
Statische Informations- und Datenflussanalyse von Android-Apps 2491

Christian Meltebrink
Algorithmenbeschleunigung durch FPGAs zur Massendatenverarbeitung 2493

Sarah Seifert
Intent-basierte Kommunikation im Web: Übersicht und Vergleich 2495

Hendrik Geßner, Matthias Weise
3D-Rekonstruktion mit der AR.Drone 2.0 2497

Michael Lettrich
Efficient Regression for Big Data Problems using Adaptive Sparse Grids 2499

Integrata Kongress 2014: Privatheit in der E-Society

Hans-Peter Fröschle
Vorwort 2503

Constanze Kurz
Profilierung menschlichen Verhaltens 2505

Christof Leng
Heartbleed: Lessons from the System Failure 2507

Gottfried Vossen
Bankbook - Sicherung der Privatheit in einer experimentellen Bankfiliale bei Facebook 2521

Christiane Eckardt, Danilo Kardel, Christof Leng, Jan Marco Leimeister, Michael Mörike, Welf Schröter, Hans-Dieter Zimmermann
„Privatheit in der E-Society - Zwischenbilanz eines Diskussionsprozesses“ 2523

Welf Schröter
Vorbemerkung 2525

Danilo Kardel
Schutz der Privatheit 2527

Christof Leng
Privatheit und Sicherheit 2531

Hans-Dieter Zimmermann
Smart Cities und Privatheit 2535

Michael Mörike
Exkurs 1: Schutz der Privatheit bei ePartizipation 2541

Christiane Eckardt
Exkurs 2: Privatheit in der eSociety - Innerer Dialog eines Nutzers 2545

Welf Schröter
Privatheit und virtuelle Arbeit in Industrie 4.0 2549

Jan Marco Leimeister
Exkurs 3: Arbeitgeber und Gewerkschaften im Vergleich - wie gut sind diese auf Crowdsourcing vorbereitet? 2553

Bertram Lohmüller, Rolf Pfeiffer
Vorstellung des Qualitätssiegels „Lebensqualität durch IT” der Integrata-Stiftung .. 2557

Wolfgang Heilmann
Der Preis für humane Nutzung der IT - heute und in Zukunft 2559

Tag der Informatik

(24.09.2014)

Cyber-Physical Systems – digital vernetzt in die physikalische Wirklichkeit

Erweiterte Kurzfassung

Manfred Broy

Institut für Informatik
Technische Universität München
D-80290 München Germany
broy@in.tum.de

Abstract: Cyber-Physical Systems im Sinne der acatech Studie (vgl. [Geisberger, Broy 12]) sind software-intensive Systeme, die unmittelbar mit der physikalischen Wirklichkeit verbunden sind, aber im Gegensatz zu klassischen eingebetteten Systemen tief und umfassend über globale Netzstrukturen, wie das Internet und damit verfügbare Daten und Dienste, vernetzt sind. Dadurch entstehen Systeme, die eng in die physikalische Wirklichkeit integriert sind, diese auch in das Internet transportieren und umgekehrt Daten und Dienste aus dem Internet nutzen. Informatiksysteme werden Teil der Wirklichkeit und spiegeln diese in ihren Konzepten unmittelbar wider. Sie bilden die technische Basis für das „Internet der Daten, Dienste und Dinge", „Big Data" und „Context Aware Cloud".

Neben die diskreten Konzepte der Informatik und qualitative Modelle auf Basis einer zweiwertigen Logik (tertium non datur) treten in Cyber-Physical Systems Unschärfe (engl. „Uncertainty") und quantitative Modelle. Dies hat drastische Auswirkungen auf die Modelle, Methodik und Vorgehensweisen der Informatik bei der Gestaltung solcher Systeme. Bisher zielten die Modelle der Informatik eher auf den Lösungsraum (die abstrakte Rechenmaschine) und weniger auf den Problemraum (die Welt der organisatorischen oder physikalischen Prozesse).

1 Konnektivität

Ein Phänomen ist die zunehmende Vernetzung unterschiedlicher Systeme. Anders als in den ersten Jahren der Informatik, als die Verbindung von Programmen zu ihrer Umgebung als so unwesentlich erachtet wurde, dass in der ansonsten richtungsweisenden Sprache Algol 60 darauf verzichtet wurde, die Ein/Ausgabe mit zu standardisieren, ist heute der interaktive Umgang mit Programmen unverzichtbar. Interaktion zwischen Programmen über Kommunikationsverbindungen ist wesentliches Merkmal heutiger Informatiksysteme.

Es erfordert spezifische Konzepte, um Phänomene der Kommunikation und Interaktion zu spezifizieren, zu modellieren und zu realisieren. Die alte Vorstellung, dass ein Programm, das in einem bestimmten Zustand gestartet wird, so lange rechnet, bis ein Endzustand erreicht ist, der das Ergebnis repräsentiert, entspricht dieser Sichtweise längst nicht mehr. Allerdings ist zu beobachten, dass selbst neuere Programmiersprachen nicht wirklich konsequent auf den Umstand ausgerichtet sind, dass Programme heute interaktiv und verteilt ablaufen und ein Großteil ihrer Funktionalität sich aus der Interaktion – mit Nutzern, mit der physikalischen Umgebung und mit anderen Systemen – ergibt. Dies erfordert konsequente Ausrichtung der Systeme auf Interoperabilität und auch eine Erweiterung theoretischer Konzepte wie des Begriffs der Berechenbarkeit (vgl. [Broy 07]).

Weitere Phänomene heutiger Systeme sind Multifunktionalität, Nebenläufigkeit, hohe Sicherheitsanforderungen und Themen der Safety und Security. Betrachtet man diese essentiellen Merkmale und Anforderungen heutiger Software, so stellt man fest, dass keiner dieser Aspekte in heutigen Programmiersprachen und methodischen Programmieransätzen in dem Umfang berücksichtigt ist, wie das erforderlich scheint.

2 Von Software zum System

Software ist immer mehr Teil umfassender Systeme (Ausführungsplattform und -umgebung). Systeme sind selbst wieder in einen operationellen Kontext (Nutzer, physikalische Umgebung, technische oder organisatorische Prozesse, weitere Software-intensive Systeme) eingebettet, mit dem sie interagieren. Dies erfordert in den Anforderungen und in der Architektur eine Berücksichtigung und systematischen Erfassung der Interaktion mit dem operationellem Kontext (Schnittstellenverhalten der Systeme und der Software) und der Annahmen über den operationellen Kontext (insbesondere über sein „Verhalten“, siehe auch Stichwort „Domänenmodellierung“, vgl. [Jackson 10]).

Software Engineering ist dann Teil eines Systems Engineering. Systeme sind eingebettet in umfassendere Systeme, abgegrenzt durch präzise festlegte Systemgrenzen, über die sie im Rahmen ihres Schnittstellenverhaltens mit ihrem operationellen Kontext interagieren.

3 Konsequenzen stärkerer „Physikalität“

Software Engineering war in seinen Anfängen vor mehr als 50 Jahren zunächst bestrebt physikalische Aspekte soweit möglich aus seinen Modellen, Spezifikationen und letztlich Programmiersprachen, abgesehen von einigen Spezialsprachen, zu verbannen, so dass etwa Fragen der Zeit bewusst weg abstrahiert wurden. Ziel war, Programmierparadigmen zu schaffen, die abstrakt, von der Ausführungsumgebung losgelöst Algorithmen als mehr oder weniger rein mathematische Gebilde erfassten. Die ersten Jahre der Informatik waren somit durch die Suche nach umfassenden Abstraktionen gekennzeichnet, die sich von den technischen Details einer Ausführungsumgebung lösen und möglichst abstrakte Programmierkonzepte anbieten. Nahezu alle herkömmlichen Programmiersprachen klammern bis heute das Thema Zeit konzeptionell aus und stellen Abstraktionen bereit, für die Zeit in Berechnungen keine explizite Rolle spielt. Dies reflektieren auch die Methoden und Theorien der Programmiersprachen und ihrer Semantik. Ziel ist von Zeit- und Ausführungsdetails zu abstrahieren und ein möglichst abstraktes „maschinenunabhängiges“ Konzept anzubieten, um über geeignete Abstraktion Portierbarkeit zu garantieren. Dies ändert sich mit der Entwicklung von Cyber-Physical Systems, in denen physikalischen Gesichtspunkten, unter anderem Zeit (vgl. [Broy 06]), ein dominierende Rolle zukommt.

Für die Realisierung zeitkritischer Systeme ist bisher eine ganze Reihe spezieller Modelle, Programmiersprachen und Paradigmen entwickelt worden, allerdings alle mit einer oft eigenartigen Vermischung aus Ausführungszeit und funktionalen Anforderungen in Hinblick auf die physikalische Zeit in dem operationellem Kontext, auf den das System wirkt. Zeitkonzepte werden bisher in Programme nicht direkt eingebracht, sondern eher implizit über Ausführungsplattformen und Laufzeitumgebungen. Das Zeitverhalten ist dann für Echtzeitanwendungen durch komplizierte Verfahren („Worst Case Execution Time“ WCET vgl. [Wilhelm, Grund 14]) bezogen auf die Ausführungsplattform zu ermitteln. Programmiersprachen werden um spezifische Konstrukte erweitert, die es erlauben auf Ausführungs- oder Systemzeit zuzugreifen, allerdings nur mit indirektem Bezug zu physikalischer Zeit im Sinn der Anwendungsdomäne.

Es stellt sich die Frage, ob es nicht konsequenter wäre, Modellierungskonzepte zu verwenden und Sprachen zu schaffen, die Zeit als explizites Konzept (vgl. [Broy 09]) enthalten und sich so von der Vorstellung lösen, dass das Zeitverhalten eines Programms Ergebnis der Ausführungsgeschwindigkeit der entsprechenden Hardware ist. Im Gegensatz dazu sind Programmiersprachen denkbar, die es erlauben, explizit funktionale Zeitanforderungen zu formulieren. Solche Programme können natürlich auf Hardware nur korrekt ausgeführt werden, wenn die Ausführungsgeschwindigkeit der Hardware ausreicht. Dieses Phänomen ist aber nicht neu. Gleiches gilt heute für Algorithmen allgemein. Denn wenn der Speicherplatz einer Maschine nicht ausreicht, ist das idealisierte Konzept eines Programms, das über Speicherplatzbeschränkungen ja keine Aussagen macht, auch nur korrekt ausführbar, wenn letztendlich genügend Speicherplatz zur Verfügung steht.

Die immer stärkere Präsenz von Physikalität im Kontext der Programme hat zur unvermeidlichen Konsequenz, Konzepte in der Modellierung und letztlich auch Modelle, Spezifikation und Programme zu entwickeln, die physikalische Realität direkt widerspiegeln. Das gilt nicht nur für die Modellierung von Zeit, sondern auch für weitere Gesichtspunkte wie Kausalität, Raum und Geometrie.

4 Diskretes und kontinuierliches Verhalten

Ein wesentlicher Unterschied vieler Modelle der physikalischen Realität zu den Modellen der Informatik besteht darin, dass Modelle der Informatik inhärent diskret und abstrakt sind und sich stark auf klassische Logikkonzepte stützen, wohingegen physikalische Modelle in aller Regel kontinuierliche Funktionen als Basis haben, wie sie unter anderem in der Regelungstechnik verwendet werden, um unmittelbar gängige Phänomene der physikalischen Realität zu modellieren. Heute gilt, dass kontinuierliche Funktionen auf digitalen Maschinen nicht analog umgesetzt werden, sondern über numerische Methoden diskretisiert approximiert, also doch wieder diskret ausgeführt werden. In den Modellierungsansätzen werden aber diese Funktionen völlig zurecht über kontinuierliche Modelle der Mathematik behandelt, da eine ganze Reihe typischer Methoden durch diese Modellbildung zur Verfügung stehen.

Allerdings ist bisher nicht konsequent verstanden und untersucht worden, inwieweit eine stärker logische Sicht, also nicht eine numerische approximierende Sicht, sondern eine diskrete Sicht, die Logik eines Systems, das über kontinuierliche Funktionen modelliert ist, beschreibt. Auch in kontinuierlichen Systemen sind logische Ereignisse identifizierbar (vgl. [Broy 12]). Aus dem kontinuierlichen Verhalten eines Systems lässt sich durch die Einführung entsprechender prägnanter Ereignisse ein diskretes Modell, etwa in Form einer Zustandsmaschine, gewinnen, das bei geschickter Wahl die Logik und damit die Funktionalität des Systems erfasst.

5 Unschärfe und unsicheres Wissen

Klassische Informatik und Algorithmik leben – abgesehen von der Numerik – in einer diskreten Welt. Algorithmen sind digital. Antworten auf Fragen sind eindeutig. Eine Bedingung ist wahr oder falsch. Ein Programm ist korrekt oder inkorrekt.

Aus den bisher beschriebenen Phänomenen bei der Modellierung physikalischer Wirklichkeit ergeben sich Unschärfe und unsicheres Wissen. Die Gründe sind vielfältig (vgl. [Fantl, McGrath 12]). Fehler, Unschärfe und Ungenauigkeiten in den Modellen („wann ist eine Verkehrssituation kritisch, so dass eine automatische Notbremsung ausgelöst wird“), in der Erfassung von Werten durch Sensoren („ist das Werkstück passgenau“), ihrer Interpretation („ist auf dem Videobild ein Kind zu erkennen“), in den Berechnungen mit approximierten Werten (Probleme der Fehlerfortpflanzung) sowie in der Hardware („liegt ein Übertragungs- Speicher- oder Berechnungsfehler vor“) und Software („ist die Anwendung korrekt modelliert“) führen zu Unsicherheit über die

Zuverlässigkeit von Systemen, die Gültigkeit von berechneten Werten und in Konsequenz zu unsicherem Wissen. Diese Unschärfe und Ungenauigkeiten akkumulieren sich in Systemen über die Zeit durch die Alterung von ohnehin unsicheren Daten (Berechnung des Standorts eines Quadrocopters in einem Gebäude ohne GPS-Zugriff unter fortschreibender Berechnung des Standorts aus Sensordaten und gespeicherten Daten).

Bedingungen über kontinuierlichen Funktionen sind immer kritisch, wenn sie etwa in Kontrollstrukturen abgefragt werden und nur kleine Unterschiede in den Werten für das Ergebnis der Bedingungsabfrage entscheidend sind. Durch Ungenauigkeit in der Repräsentation kontinuierlicher Funktionen und Berechnungen kann dies – wie aus der Numerik bekannt – zu verfälschten Ergebnissen führen. Anders ausgedrückt, wenn kleine Unterschiede in den Funktionswerten zu unterschiedlichen Ergebnissen bei Abfragen führen, haben wir es mit chaotischen Verhalten und damit mit Unsicherheit zu tun, inwieweit die Ergebnisse der Abfragen tatsächlich korrekt das widerspiegeln, was sie modellieren sollen, wenn mit zwangsläufig ungenauen Approximationen gearbeitet wird.

Schwerer wiegt, dass in vielen Fällen physikalische Modelle probabilistische Sichtweisen erfordern. Wahrscheinlichkeitskonzepte sind für viele Situationen der Modellierung technischer und physikalischer Phänomene unverzichtbar – gleichermaßen aleatorische Wahrscheinlichkeit als Folge eines stochastischen physikalischen Prozesses wie auch epistemische Wahrscheinlichkeit als Folge fehlenden Wissens über kausale Zusammenhänge und Hintergründe. Es wird notwendig, probabilistische Konzepte (vgl. [Neubeck 12]) in die Modellierung mit aufzunehmen, und zwar sowohl in die Domänenmodellierung, in die Spezifikation, aber auch in die Betrachtung von Programmverhalten mit probabilistischen Phänomenen ihres operationellen Kontexts. Eine besondere Herausforderung liegt in der Verwaltung von Daten unter Unschärfe, Unsicherheit und alternder Aktualität sowie der Errechnung von Prognosen unter Unschärfe und unsicherem Wissen als Voraussetzung für Reaktionen autonomer Systeme.

6 Quantitative und qualitative Modellierung

Logische und algebraische Ansätzen zur Spezifikation und diskreten Modellierung von Software führen auf logische, zweiwertige Konzepte etwa von Korrektheit. Ein Programm ist in einer qualitativen Modellierung entweder korrekt oder fehlerhaft – tertium non datur.

Software mit enger Interaktion mit der physikalischen Realität und mit menschlichen Nutzern erfordert aber „weichere“ Begriffe einer graduellen Korrektheit (vgl. [Chatterjee et al. 08]). Software ist bis zu einem bestimmten Grad korrekt, zuverlässig, sicher. Sind Maßzahlen oder Abstandsfunktionen (vgl. [Henzinger 10]) gegeben, dann kann ein Programm für eine gegebene Spezifikation „korrekter“ sein als ein anderes Programm. Probabilistische Spezifikationen fordern ein bestimmtes „korrektes“ Systemverhalten nur mit einer spezifizierten Wahrscheinlichkeit.

Dies erfordert neben den logischen, qualitativen Ansätzen quantitative Ansätze (vgl. [Broy 14]). Die Korrektheit von Software ist dann graduell (Software ist „in einem bestimmten Maß korrekt“). Praktikern ist das längst vertraut. Software ist korrekt unter bestimmten Annahmen, in einem bestimmten Grad, über einen bestimmten Zeitraum, in einem bestimmten Sinn, mit einer bestimmten Wahrscheinlichkeit.

Dies erfordert besondere Modellierungsansätze wie Probabilistik, Fuzzy Logics, Metriken oder Maßtheorie. Eine besondere Herausforderung ist die Präzisierung unscharfer Begriffe (wie Qualität, Privatheit, Nutzfreundlichkeit etc.) durch Maßzahlen und die Bewertung ihrer Aussagekraft für die Umsetzung in die Funktionalität von Software-Systemen.

7 Konsequenzen für Software, Systems und Requirements Engineering

Um Software Engineering, eingebettet in Systems Engineering, auf die Cyber-Physical Systems der Zukunft auszurichten, ist eine konsequente Erweiterung der Modelle notwendig. Das bedeutet keineswegs, dass wir uns von diskreten Modellen abwenden, denn die diskrete Sicht auf die Realität ist wesentlicher Teil der Stärke der Informatik und bringt völlig neue Erkenntnisse und Möglichkeiten, wie es die letzten Jahre eindrucksvoll gezeigt haben. Jedoch muss eine viel engere Verzahnung stattfinden zwischen den klassischen Modellen der Informatik, den diskreten Systemen, Schnittstellenkonzepten, den Zustandsmaschinen, den diskreten Abläufen, den Prozessbegriffen und den Modellen, die durch die Mathematik für die Physik geschaffen worden sind, wie kontinuierliche Funktionen und Probabilistik. Von besonderer Bedeutung sind letztlich Modelle von Zeit und Raum. Eine Informatik, die diese Konzepte harmonisch integriert und in einen Engineering-Ansatz umsetzt, ist das geeignete Vehikel, um die Cyber-Physischen Systeme der Zukunft zu erschließen.

Die Einführung von Zeit, Raum und Wahrscheinlichkeit von Unsicherheit und Unschärfe in die Modellierung Software-intensiver Systeme hat weitreichende Konsequenzen für den gesamten Begriffsapparat der Informatik und des Software Engineering sowie insbesondere des Requirements Engineering. Darüberhinaus sind das Grundlagen für zentrale Themen der Informatik wie „Human Centric Engineering“ (vgl. [Broy, Schmidt 14]), Big Data, verteilte, mobile, adaptive und autonome Systeme.

Danksagung

Ich danke meinen Kollegen Herrn Bernhard Schätz, Erhard Plödereder, Alexander Pretschner für inspirierende Diskussionen und kritische Anmerkungen.

Literaturverzeichnis

[Broy 06] M. Broy: The ‚Grand Challenge' in Informatics: Engineering Software-Intensive Systems. IEEE Computer, Oktober 2006, 72 – 80

[Broy 07] M. Broy: Interaction and Realizability, In: Jan van Leeuwen, Giuseppe F. Italiona, Wiebe van der Hoek, Christoph Meinel, Harald Sack, Frantisek Plasil (eds.): SOFSEM 2007: Theory and Practice of Computer Science, Lecture Notes in Computer Science 4362, Springer 2007, S. 29–50

[Broy 09] M. Broy: Relating Time and Causality in Interactive Distributed Systems. In: M. Broy, W. Sitou, T. Hoare (eds): Engineering Methods and Tools for Software Safety and Security, NATO Science for Peace and Security Systems, D: Information and Communication Security, Vol. 22, IOS Press 2009, 75-130

[Broy 12] M. Broy: System Behaviour Models with Discrete and Dense Time. S. Chakraborty, J. Eberspächer (eds.): Advances in Real-Time Systems, Springer Berlin Heidelberg, 2012, 3-25

[Broy 14] M. Broy: Formalizing and Relating System and Software Properties: Their Dependency and Independency. Unpulished Manuscript

[Broy, Schmidt 14] M. Broy, Albrecht Schmidt: Challenges in Engineering Cyber-Physical Systems. IEEE Computer Society, Ausgabe Februar 2014, S. 70-72

[Chatterjee et al. 14] K. Chatterjee, L. Doyen, Th. A. Henzinger. Quantitative languages. Proceedings of the 17th International Conference on Computer Science Logic (CSL), Lecture Notes in Computer Science 5213, Springer, 2008, 385-400

[Fantl, McGrath 08] J. Fantl, M. McGrath: Knowledge in an Uncertain World. Oxford University Press, 2012

[Geisberger, Broy 12] E. Geisberger, M. Broy: agendaCPS – Integrierte Forschungsagenda Cyber-Physical Systems (acatech STUDIE), Heidelberg u.a.: Springer Verlag 2012

[Henzinger 10] Th. A. Henzinger: From Boolean to Quantitative Notions of Correctness. POPL '10 Proceedings of the 37th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages, 2010, 157-158

[Jackson 10] M. Jackson: The Operational Principle and Problem Frames. In: C. B. Jones, A. W. Roscoe, K. R. Wood (eds.): Reflections on the Work of C. A. R. Hoare, Springer Verlag, London, 2010

[Neubeck 12] P. R. Neubeck: A Probabilitistic Theory of Interactive Systems. Technische Universität Münschen, Fakultät für Informatik, Dissertation 2012

[Wilhelm, Grund 14] R. Wilhelm, D. Grund: Computation Takes Time, but how much? Commun. ACM 57(2), 2014, 94-103

Smart Big Data

Karl-Heinz Streibich

CEO Software AG

Abstract: Die intelligente Nutzung von digitalen Daten spielt eine zentrale Rolle, wenn es darum geht, Antworten auf zukünftige wirtschaftliche und gesellschaftliche Herausforderungen zu geben. Software ist hierbei der fundamentale Werkstoff für innovative Produkte und Dienstleistungen. Die Wettbewerbsfähigkeit der deutschen Wirtschaft wird entscheidend von der Fähigkeit abhängen, softwarebasierte Produkte und Dienstleistungen mit höchster Qualität zu erstellen. "Smart Data" ist für die Industrie eine technologische Jahrhundertchance und wird in Zukunft der wichtigste Rohstoff für alle Bereiche der Unternehmenswelt.

Big Data mit Datenschutz - Mission impossible?

Peter Schaar

Vorsitzender, Europäische Akademie für Informationsfreiheit und Datenschutz

Abstract: Big Data stellt wesentliche Datenschutzprinzipien auf den Kopf: Die massenhafte, ergebnisoffene und auf Dauer angelegte Sammlung und Auswertung von Daten scheint in einem unlösbaren Widerspruch zu den Anforderungen der Erforderlichkeit, Zweckbindung und der Datensparsamkeit zu stehen. Dabei gibt es durchaus Ansätze, wie sich der Konflikt entschärfen lässt: Die frühzeitige Berücksichtigung datenschutzrechtlicher Fragestellungen, Vorkehrungen zur Anonymisierung und Pseudonymisierung und ein Höchstmaß an Transparenz können verhindern, dass sinnvolle Big Data-Ansätze am Datenschutz scheitern.

Big Data Analytics - Vom Maschinellen Lernen zur Data Science

Prof. Dr. Stefan Wrobel[1,2]

[1] Universität Bonn
Institut für Informatik III
Römerstr. 164
53117 Bonn
wrobel@cs.uni-bonn.de

[2] Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS
Schloß Birlinghoven
53757 Sankt Augustin
stefan.wrobel@iais.fraunhofer.de

Abstract: Die Informatik-Teilgebiete Maschinelles Lernen und Data Mining beschäftigen sich schon lange auf statistischer Basis mit der algorithmischen Analyse von Daten. Im Zeitalter von Big Data kommen zu den ursprünglichen analytischen Zielen jedoch ein Vielzahl von Anforderungen, die sich aus der zentralen Rolle solcher Verfahren in Unternehmensprozessen und Gesellschaft ergeben. Der Begriff Data Science umfasst daher einen breiten Bereich, von Datenauswahl und -verknüpfung über Analytik, technische Realisierung, Unternehmensbezug bis hin zu ethischen Fragen. Der Vortrag diskutiert diese unterschiedlichen Aspekte und macht deutlich, wie sich daraus neue Forschungsfragen und Forschungsfelder ergeben. Wir illustrieren diese Themen mit aktuellen Ergebnissen und Anwendungsbeispielen insbesondere hinsichtlich der Analyse unter Wahrung der Privatsphäre, der Steuerung von Analyseverfahren und der visuellen Analytik.

Big Text: von Sprache zu Wissen

Gerhard Weikum
Max-Planck-Institut für Informatik &
Universität des Saarlandes
D-66123 Saarbrücken
weikum@mpi-inf.mpg.de

Abstract: Nachrichten, soziale Medien, Webseiten und Dokumente in Unternehmen beinhalten eine Fülle wertvoller Informationen in textueller oder gesprochener Form. Ein entscheidender Schritt, um diese Form von Big Data für Analytik und Entscheidungsfindung zu nutzen, ist die Identifikation von Entitäten wie Personen, Organisationen oder Produkten sowie von Beziehungen zwischen Entitäten. Dieser Schritt ermöglicht auch das Verknüpfen von strukturierten und unstrukturierten Daten für Big-Data-Anwendungen. Allerdings steht dieser Schritt vor dem herausfordernden Problem, dass Namen von Entitäten und relationale Phrasen oft in hohem Maße mehrdeutig sind. Man benötigt eine skalierbare, hochwertige Lösung für die Disambiguierung von Namen in Texten und heterogenen Tabellen. Dieser Vortrag diskutiert den Stand der Kunst, Anwendungen und offene Probleme zu diesem Thema. Er diskutiert außerdem die daraus resultierenden Perspektiven für das umfassendere Thema der Big-Data-Analytik.

Algorithm Engineering for Big Data

Peter Sanders

sanders@kit.edu

Abstract: Perhaps the most fundamental challenge implied by advanced applications of big data sets is how to perform the vast amount of required computations sufficiently efficiently. Efficient algorithms are at the heart of this question. But how can we obtain innovative algorithmic solutions for demanding application problems with exploding input sizes using complex modern hardware and advanced algorithmic techniques? This tutorial gives examples how the methodology of algorithm engineering can be applied here. Examples include sorting, main memory based data bases, communication efficient algorithms, particle tracking at CERN LHC, 4D image processing, parallel graph algorithms, and full text indexing. Compared to a previous tutorial in Koblenz 2013 with the same title, this tutorial talks less about methodology and more about actual algorithms and applications. For further reading refer to [San13] and, for selected individual results to [DS03, KS07, SSP07, MS08, San09, RSS10, SS12, DS13].

Literatur

[DS03] R. Dementiev und P. Sanders. Asynchronous Parallel Disk Sorting. In *15th ACM Symposium on Parallelism in Algorithms and Architectures*, Seiten 138–148, San Diego, 2003.

[DS13] Jonathan Dees und Peter Sanders. Efficient Many-Core Query Execution in Main Memory Column-Stores. In *29th IEEE Conference on Data Engineering*, 2013.

[KS07] F. Kulla und P. Sanders. Scalable Parallel Suffix Array Construction. *Parallel Computing*, 33:605–612, 2007. Special issue on Euro PVM/MPI 2006, distinguished paper.

[MS08] K. Mehlhorn und P. Sanders. *Algorithms and Data Structures — The Basic Toolbox*. Springer, 2008.

[RSS10] M. Rahn, P. Sanders und J. Singler. Scalable Distributed-Memory External Sorting. In *26th IEEE International Conference on Data Engineering*, Seiten 685–688, 2010.

[San09] P. Sanders. Algorithm Engineering – An Attempt at a Definition. In *Efficient Algorithms*, Jgg. 5760 of *LNCS*, Seiten 321–340. Springer, 2009.

[San13] Peter Sanders. Engineering Algorithms for Large Data Sets. In *39th Conf. on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, Jgg. 7741 of *LNCS*, Seiten 29–32. Springer, 2013. invited talk.

[SS12] Peter Sanders und Christian Schulz. Distributed Evolutionary Graph Partitioning. In *ALENEX 2012*, Seiten 16–29. SIAM, 2012.

[SSP07] J. Singler, P. Sanders und F. Putze. MCSTL: The Multi-core Standard Template Library. In *13th Euro-Par*, Jgg. 4641 of *LNCS*, Seiten 682–694. Springer, 2007.

Vorstellung ausgewählter „Grand Challenges“ der Informatik

Dr. Simone Rehm

GI-Vizepräsidentin
simone.rehm@gi.de

Abstract: Innerhalb der Gesellschaft für Informatik e.V. (GI) hat sich eine Gruppe auf die Suche gemacht nach den „Grand Challenges“ der Informatik. Eine "Grand Challenge" - so der Versuch einer Definition - stellt ein grundsätzliches Problem in der Informatik oder für die Informatik dar, dessen Lösung mit Informatik-Hilfsmitteln einen deutlich spürbaren Fortschritt in ökonomischer, sozialer oder gesellschaftlicher Hinsicht für unser aller Leben bedeutet.

GI-Fachgruppen wie auch die Mitglieder der GI waren aufgerufen, Vorschläge einzureichen. Manche ähnlich lautende Einreichungen wurden zu Themenblöcken zusammengefasst und inhaltlich geschärft, andere blieben einzeln als Grand Challenge bestehen. Am Ende beschloss das Präsidium der GI fünf „Grand Challenges“, die nun als wesentliche aktuelle Herausforderungen für die Informatik angesehen werden:

1. Verlässlichkeit von Software
2. Allgegenwärtige Mensch-Maschine-Interaktion
3. Erhalt des digitalen Kulturerbes
4. Internet der Zukunft – Sicher, schnell, vertrauenswürdig
5. Beherrschung systemischer Risiken in IT-unterstützten weltweiten Netzen

Im Rahmen der INFORMATIK 2014 wird einer der Einreicher, Prof. Michael Koch, eine dieser „Grand Challenges“ ausführlicher präsentieren. Abgerundet wird die Vorstellung von einer Erläuterung der Ziele, die mit dieser Initiative verbunden sind. Im Grundsatz geht es darum, der Informatik als Disziplin ein spannendes und lebendiges Gesicht zu geben und den jungen Menschen, die sich für Informatik begeistern, Anreize für Forschungsthemen zu vermitteln.

Tag der Wirtschaft

(23.09.2014)

Big Data im Internet der Bosch-Dinge und -Dienste

Lothar Baum[1], Wolfgang Blochinger[2], Michael Peters[3], Jörg Sommer[4]

[1] Corporate Research, [2-4] Corporate Sector Information Systems and Services
Robert Bosch GmbH
Postfach 30 02 20, 70442 Stuttgart
{Lothar.Baum | Wolfgang.Blochinger | Michael.Peters | Joerg.Sommer}@de.bosch.com

1 Motivation und Anwendungen bei Bosch

Bosch ist heute noch überwiegend ein Hersteller von *Dingen* der materiellen Welt. Was 1886 mit einfachen elektrotechnischen Apparaten begann, erstreckt sich heute über verschiedenste Anwendungsdomänen und Geschäftsbereiche. Kfz-Komponenten, Heizungsanlagen, Haushaltsgeräte und Industriesteuerungen gehören ebenso zum Produktportfolio wie Sicherheitssysteme, Beschallungsanlagen oder auch medizintechnische Produkte. Gut drei Viertel aller von Bosch hergestellten Produkte werden bereits heute bzw. in naher Zukunft durch embedded Software gesteuert. Dies erlaubt nicht nur eine effizientere und flexiblere Steuerung, sondern auch eine Vernetzung dieser Systeme. Auf diese Weise ist es erstmals möglich, genaue Informationen über Einsatz und *Leben* der Produkte im Feld zu sammeln und diese Informationen zur Produktoptimierung oder auch für neue Dienstleistungen zu nutzen. Bosch hat die Chancen des *Internets der Dinge und Dienste* früh erkannt und stellt sich in vielen Bereichen entsprechend strategisch auf.

2 Big Data und Data Mining

Ein Internet der Dinge und Dienste eröffnet den Zugang zu einer dramatisch wachsenden Zahl von Datenquellen – im Jahr 2020 werden 50 Milliarden [Ba13] *Connected Devices* weltweit aktiv sein, die potenziell Sensor- und Zustandsinformationen verfügbar machen. Auch wenn in konkreten Anwendungen nur ein Teil dieser Datenquellen zur Verfügung steht oder genutzt werden kann bzw. relevant ist, so sind die Datenmengen dennoch beachtlich. So generieren alleine die Logging-Informationen der ESP-Steuergeräte von ca. 400 000 BMW 5er der E60-Baureihe jährlich mehrere Gigabytes. Bei der Überwachung von Fertigungsparametern fallen in Bosch-Werken für einzelne Produktgruppen Daten in der Größenordnung von 50 Terabyte und mehr an. Der Wert dieser Daten ergibt sich jedoch erst durch eine systematische Analyse und den daraus ableitbaren Erkenntnissen – kurz: dem Data Mining. Während die mathematischen Grundlagen (Statistik, maschinelles Lernen) lange bekannt sind, besteht die besondere Herausforderung heute im Umgang mit extrem großen Datenmengen, in der teilweise hohen Geschwindigkeit, in der neue Daten anfallen sowie in der Heterogenität der Datenquellen.

3 Data Mining Plattform

Bosch stellt sich diesen Herausforderungen mit einer Reihe von Aktivitäten und Projekten, allen voran einem großen, geschäftsbereichsübergreifenden Data Mining Projekt. Eines der Hauptziele dieses Projekts ist der Aufbau und die Bereitstellung einer geeigneten Big Data Plattform. So entstand im vergangenen Jahr ein High-Performance Computing Cluster als Referenz-Umgebung speziell für Boschs Data Mining Fragestellungen. Diese Umgebung setzt auf bereits etablierte Standards und Open-Source-Lösungen wie dem Hadoop Ökosystem (HDFS, MapReduce, Hive, Hbase, und Mahout) [Ap14][Sa12][Wh12]. Im Rahmen des Data Mining Projekts wurde auch eine Vielzahl kommerzieller Lösungen analysiert und auf Eignung für Bosch-Anwendungen geprüft. Die sechs am weitesten verbreiteten Produkte Knime, RapidMiner, IBM SPSS, Revolution R, Statistica und SAS sowie drei weitere ausgewählte Tools (Alpine, HP Vertica und EMC Pivotal) wurden schließlich einem Benchmark auf einem einheitlichen Datensatz unterzogen. Es stellte sich heraus, dass dieser Satz an (nicht Bosch-spezifischen) echten Anwendungsdaten bereits ausnahmslos allen Tools Probleme bereitet. Während sich die meisten Tools dank guter Bedienoberflächen und breiter Datenquellenanbindung gut zum interaktiven Ausprobieren und Prototyping eignen, setzt Bosch aktuell für produktive Lösungen vorranging auf Eigenentwicklungen, etwa auf Basis von Mahout.

4 Herausforderungen aus industrieller Sicht

4.1 Nutzbarkeit der Daten hinsichtlich Qualität

Bevor Daten mit Analyse-Tools ausgewertet werden können, müssen sie in aller Regel zunächst geeignet aufbereitet werden. Datensätze sind im *echten Leben* nie perfekt: mal fehlen einzelne Datenpunkte, mal sind (Mess-)Fehler enthalten, mal wechseln Messverfahren oder Bezugswerte und machen Datenpunkte somit schwer vergleichbar. Je größer die Datenmengen werden, umso aufwendiger werden auch die notwendigen Bereinigungen. So erklärt sich, dass in typischen Data-Mining-Anwendungen über 80% des Aufwandes für das Sichten, Verstehen, Hinterfragen und Aufbereiten der Daten anfällt. Auch läßt sich im Vorhinein kaum genau sagen, ob bestimmte Data-Mining-Fragestellungen mit ausreichender Genauigkeit oder gar überhaupt mit den vorhandenen Daten beantwortet werden können. Oft stellt sich somit nicht die Technik (sprich: Analytik) als der Flaschenhals heraus, sondern vielmehr die Menge und Qualität der verfügbaren Daten.

4.2 Sicherheit

Große Datenmengen, insbesondere solche, die sich zur Analyse und zum Data Mining eignen, stellen einen enormen Wert dar. Damit besteht das Risiko, dass diese zum Ziel von externen Angriffen werden. Eine besondere Herausforderung beim Umgang mit Big Data ist daher der Schutz der Daten. Technische Ansätze wie Zugriffskontrolle,

Verschlüsselung oder auch Anonymisierung sind zwar bekannt, führen im Alltag allerdings auch zu Einschränkungen oder offenbaren Unzulänglichkeiten. Wie lassen sich beispielsweise Daten analysieren, ohne den gesamten Datenbestand offenlegen (entschlüsseln) zu müssen? Hier arbeitet Bosch z.B. an eigenen Verfahren zum Schutz der Privatsphäre, sogenannten *Privacy-enhancing Technologies*.

4.3 Zentrale vs. dezentrale Infrastruktur

Seit einigen Jahren werden bei Bosch zur Nutzung von Skaleneffekten und Erhöhung der Sicherheit IT-Ressourcen in wenigen Rechenzentren konsolidiert. Die damit einhergehende Automatisierung und Virtualisierung ermöglichen zusätzlich eine höhere Auslastung und eine höhere Flexibilität. Aus Sicht der IT-Infrastruktur ergibt sich nun die Frage, ob die in den (Produktions-) Werken anfallenden, großen Datenmengen von mehreren Terabytes wirtschaftlich über das globale Unternehmensnetz in den zentralen Rechenzentren gespeichert und verarbeitet werden können oder ob ein dezentraler Ansatz von Rechenleistung und Speicherkapazität für Big Data und Data Mining effizienter und effektiver ist. Dies hätte eine Umkehrung des Trends und eine De-Zentralisierung zur Folge.

4.4 Global Deployment

Eine weitere signfikante Herausforderung stellt die Etablierung einer unternehmensweiten, standardisierten und integrierten IT-Infrastruktur für die Untestützung des gesamten Data Mining Prozesses dar. Diese muss neben der Erstellung und Validierung von Modellen mittels leistungsfähiger *High Performance Computing* (HPC) Plattformen sowohl Aspekte der Datenaggregation und Aufbereitung als insbesondere auch Aspekte des Modell-Deployments in verschiedensten Produktionsumgebungen und Geschäftsprozessen möglichst nahtlos integrieren. Als wesentlichen Bestandteil für das Modell-Deployment setzt Bosch die *Predictive Model Markup Language* (PMML) zur einheitlichen Beschreibung von Modellen ein. Die Modell-Ausführung kann dann mittels einer in Produktionsumgebungen eingebetteten Modell Execution Server Infrastruktur, zum Beispiel auf Basis der Scoring Engine Adapa von Zementis, bewerkstelligt werden.

Literaturverzeichnis

[Ap14] Apache Software Foundation: Apache Hadoop Project Website, http://hadoop.apache.org. Zugriff am 07.03.2014.

[Ba13] Bassi, A. et al.: Enabling Things to Talk – Designing IoT solutions with the IoT Architectural Reference Model, Springer, Berlin, 2013.

[Sa12] Sammer, E.: Hadoop Operations. O'Reilly, 1st Edition, 2012.

[Wh12] White, T.: Hadoop – The Definitive Guide, O'Reilly, 3rd Edition, 2012.

Industrie 4.0: Potenziale intelligenter Vernetzung

Klaus Bauer

TRUMPF Werkzeugmaschinen GmbH + Co.
Klaus.bauer@de.trumpf.com

Abstract: Nach Mechanisierung, Elektrifizierung und Informatisierung der Industrie läutet der Einzug des „Internets der Dinge und Dienste“ in die Fabrik eine 4. Industrielle Revolution ein. Das Potential durch eine umfassende datentechnische Vernetzung ist immens. Neben einer deutlichen Verbesserung der Ressourcen – und Energieeffizienz werden vor allem Wertschöpfungsprozesse effizienter und flexibler.

In dem von der Acatech geleiteten Arbeitskreis zum Zukunftsprojekt „Industrie 4.0“ wurden wesentliche Kennzeichen der Fabrik der Zukunft, der sogenannten Smart Factory definiert. Die Maschinen sind als Social Machines intelligent vernetzt und reagieren flexibel, eigenständig und situationsabhängig. Sie gleichen so Belastungen aus und kompensieren Ausfälle. Gleichzeitig kommunizieren sie mit Systemen von Zulieferern und Kunden und reagieren auf mögliche Abweichungen sofort, um das Gesamtoptimum zu erreichen. Das Smart Product, das durch die Fertigung läuft, kommuniziert ebenfalls beispielsweise über einen integrierten Chip mit seinem Umfeld. So kann der Rohling, der seinen Status und seine Historie selbst kennt, die Maschine anweisen, wie er bearbeitet werden muss. Zudem ist der Status des entstehenden Produkts jederzeit und von überall einsehbar. Auf den ersten Blick scheint eine so vernetzte Produktion den Menschen zu ersetzen. Doch die Mitarbeiter bekommen eine wichtige Rolle. Sie überwachen und steuern die Fertigungsprozesse dezentral mithilfe von neuartigen IT-basierten Assistenzsystemen. Hierbei werden verstärkt mobile Geräte zum Einsatz kommen

Der Weg vom Internet der Dinge zum Internet der Dienste erfolgt schrittweise. Zunächst wird die physische Welt mit der virtuellen IT-Welt vernetzt. Intelligente Sensoren liefern dabei nicht nur die notwendige Basis für robuste und effiziente Produktionsprozesse sondern auch für Condition Monitoring. Der Anlagenstatus wird über mobile Assistenzsysteme gesteuert und überwacht.

Digitale Daten stellen auch die Grundlage für weitere innerbetriebliche Prozessoptimierungen dar und ermöglichen Dienstleistungen wie z.B. Remote Support, automatisierte Ersatzteilbestellung und Technologie-Datenbereitstellung über Web-Shops.

Offene Datenschnittstellen und Standards erleichtern die firmengrenzüberschreitende Kommunikation in neuartigen Partnernetzwerken. Der Handel mit Daten über Cloud-Marktplätze führt zu einer noch nicht dagewesenen Optimierung von gesamten Wertschöpfungsketten.

Unterstützung datenintensiver Forschung am KIT - Aktivitäten, Dienste und Erfahrungen

Bernhard Neumair, Achim Streit

Karlsruhe Institute of Technology (KIT)
Steinbuch Centre for Computing
Hermann-von-Helmholtz-Platz 1
76344 Eggenstein-Leopoldshafen
bernhard.neumair@kit.edu
achim.streit@kit.edu

Datenintensive Forschung oder auch *Big Data in Science* adressiert nicht nur die Herausforderungen, die durch die schiere Größe und Menge an produzierten Daten (Volume) entstehen, sondern auch die Fragestellungen rund um die Geschwindigkeit der Entstehung der Daten (Velocity), der Vielfalt der Daten (Variety), der Wahrhaftigkeit und Authentizität der Daten (Veracity) und – nicht zu vergessen – den Wert der Daten (Value). Dabei stellt die *Exploration der Daten* eine Revolution dar, wie wissenschaftliche Erkenntnisse gewonnen werden können; dies bildet die *4. Säule der modernen Wissenschaft* neben Theorie, Experiment und Simulation. Die hohe Bedeutung der Exploration großer Datenmengen zeigt das Beispiel des Large Hadron Colliders (LHC) am CERN. Neben dem Beschleuniger und den Experimenten bildet das groß-skalige, verteilte Management von Daten und deren Analyse die Grundlage für die Entdeckung des Higgs-Teilchens.

Wissenschaftliche Fragestellungen rund um die Unterstützung datenintensiver Forschung am KIT sind: Wie kann die langfristige Erhaltung von Daten und die Bestandspflege mit Werkzeugen und Prozessen unterstützt werden? Wie können wertvolle und personenbezogene Daten geschützt werden? Wie können Daten auf einfache Art und Weise in globalen Kooperationen gemeinsam benutzt werden? Wie kann der Umgang mit Meta-Daten optimiert werden? Wie können Daten archiviert und für lange Zeit aufbewahrt werden? Wie können Erkenntnisse aus Daten mit neuartigen Methoden gewonnen werden? Und wie müssen Curricula zur Ausbildung zukünftiger Datenwissenschaftler, Dateningenieure und Datenanalysten aussehen?

Zur Adressierung der o.g. Fragestellungen und zur optimalen Unterstützung datenintensiver Forschung sind Kooperationen ein elementarer Faktor. Am KIT arbeiten wir sehr eng in gemeinsamer Forschung mit zahlreichen *Wissenschaftscommunities* zusammen und kooperieren auf *nationaler und internationaler Ebene* mit anderen Wissenschaftseinrichtungen und der Industrie.

Im Vortrag wird auf diese Beispiele, die gemeinsamen Aktivitäten, die darin entwickelten Dienste und die gewonnen Erfahrungen ausführlich eingegangen. Unter anderem werden vorgestellt: *GridKa als deutsches Tier-1 Daten- und Rechenzentrum*, das alle 4

Experimente des LHC-Beschleunigers am CERN unterstützt; die *Large Scale Data Facility (LSDF)* zur Bereitstellung von Speicher- und Analysekapazitäten für verschiedenste Forschungsrichtungen wie z.B. der Systembiologie, der Klimaforschung, der Forschung mit Synchrotronstrahlung, den Geisteswissenschaften und der Erdsystemforschung; die Entwicklung von IT-Diensten für Nutzer im Land Baden-Württemberg, z.B. zum einfachen Austausch von Daten a la Dropbox.

Darüber hinaus werden ausgewählte wissenschaftliche Highlights aus der gemeinsamen Forschung im Rahmen der *Helmholtz-Portfolioerweiterung LSDMA* anhand der modernen Lichtscheiben-Mikroskopie und der technologischen Analyse von Stromnetz-Messdaten dargestellt und auf Erfahrungen in der datenintensiven Forschung zusammen mit verschiedenen Wissenschaftscommunities eingegangen.

Ein weiterer Schwerpunkt des Vortrags ist das *Smart Data Innovation Lab (SDIL)*, das Anfang 2014 am KIT öffentlich vorgestellt wurde. Die Vision des SDIL ist der Aufbau eines bundesweiten Verbunds für *In-Memory Big Data Analyse* komplexer Datenbestände mit intelligenten Verfahren (aka Smart Data), der für die kooperative Forschung und Entwicklung durch Wirtschaft und Wissenschaft genutzt wird. Am KIT soll eine *leistungsstarke IT-Infrastruktur* von der Hardware- und Software-Industrie bereitgestellt werden, mit der die Wissenschaft gemeinsam mit der Wirtschaft hoch-performante Arbeit mit Big Data durchführen kann. Um eine realitätsnahe Forschung zu ermöglichen, liefern *Industriepartner Datenquellen aus der Praxis*, mit denen Forschung in strategisch wichtigen Feldern unterstützt wird. Diese Datenquellen werden ergänzt durch Daten der öffentlichen Hand sowie im Internet frei verfügbare Datenquellen.

Das SDIL richtet sich mit seinem Angebot zunächst an Forschungsarbeiten in den *strategisch wichtigen Feldern* Industrie 4.0, Energie, Smart Cities und Medizin. Andere Forschungsfelder werden später folgen. Um in diesen Bereichen mit einem umfassenden Angebot arbeiten zu können, wird SDIL mit *Data Innovation Communities* in jedem der strategischen Forschungsfelder arbeiten. Diese Communities werden von je einem Partner aus Wissenschaft und Wirtschaft als Moderator geleitet. Ziele der Data Innovation Communities sind die Diskussion und Definition von Forschungsschwerpunkten im jeweiligen Feld und die gemeinsame Akquise von weiteren europäischen Industrie- und Wissenschaftspartnern zur Stärkung der Innovation Community.

Industriepartner haben grundsätzlich die *volle Kontrolle* über die zur Verfügung gestellten *Daten*. Dies schließt sowohl die Erlaubnis von Zugriffen durch individuelle Forschungspartner ein als auch die Definition eines Lifecycle, der bspw. auch das Löschen der Daten beinhaltet. Weiter möchte das Smart Data Innovation Lab auch kleinen und mittleren Unternehmen zusätzliche Chancen bieten. Falls Software-Angebote von diesen Unternehmen das Portfolio des SDIL stärken können, kann das SDIL im Gegenzug dabei helfen, das Angebot der Unternehmen bekannter zu machen.

Gibt es sichere Software?

Dr. Thomas Liedtke, Prof. Dr. Bernhard Hohlfeld

thomas.liedtke@ics-ag.de,
bernhard.hohlfeld@ics-ag.de

Abstract: Steigender Komplexität in Industrie 4.0- und Big Data-Themen erfordern neue Lösungskonzepte. So werden Security-Themen normativ größtenteils noch stiefmütterlich behandelt. Der Vortrag beleuchtet Konzepte der IT-Security (Angriffssicherheit) in Abgrenzung zu Safety (Betriebssicherheit).

Der Vortrag ist in drei Blöcke strukturiert. Nach einer Motivation mit Beispielen wird auf den Unterschied zu Safety eingegangen. Am Ende soll der derzeitig Stand des Themas Security bei Mensch und Technik zusammengefasst werden, insb. das Umfeld und Ziele der z.Zt. entstehenden Normen..

1 Motivation

Spätestens seit dem SCADA-Maroochy-Abwasserunfall in Australien, der erfolgreichen Attacke auf Autos durch Malware-behaftete MP3-Dateien, den ersten von einem Kühlschrank versendeten Spam-E-Mails, ist in unserer Gesellschaft das Bewusstsein allgegenwärtiger IT-Security sprunghaft gestiegen. Der menschliche Faktor den Wert von Sicherheit (erst) dann zu erkennen, wenn sie nicht mehr gewährleistet ist, treibt auch hier den Stand der Technik.

Für Industrie 4.0-Anwendungen werden Anforderungen an technische Systeme stetig komplexer:

- Dezentralisierung von Steuerungen: Energiegewinnung/-verteilung, wachsender Einsatz "SMARTer"/unterschiedlichster Endgeräte, MES, Car2x, einfachste Funktionen werden über mehrere Steuergeräte verteilt, Steuergeräte kommunizieren über immer mehr Kommunikationskanäle/Busse miteinander, ...

- Neuartige unterschiedlichste Endgeräte und Bedienungskonzepte: Flexible Tablet-Steuerung einer begehbaren Maschine, …

- Erhöhte Anzahl von Sicherheitslücken. Jede neue Technologie, die Sicherheitslücken schließt, bringt neue mit sich. Anlagensteuerungen/-wartungen über Cloud-Rechenzentren

- Anlagenlebenszeiten werden länger, umgekehrt zu Software und Betriebssystemen

2 IT-Security und/ oder Safety

Je nach Domäne, Anwendungsfall und erfüllender (Sicherheits-) Funktion spricht man der Software einen bestimmten Safety Integrity Level zu. Das jetzt noch enthaltene Restrisiko wird (vom Mensch) akzeptiert. Bei Security spricht man von einem Evaluation Assurance Level, der die kriminelle Energie, die notwendig ist um ein System zu gefährden berücksichtigt. Zielwerte bei Safety und Security sind untereinander abhängig und tlw. sogar orthogonal (z.B. geschlossene Türen in Auto und Flugzeug).

Safety braucht Security. Umgekehrt darf Security Safety nicht stören. Was z.B. wenn bei einem Störfall eines weit entlegenen dezentralen Energiekraftwerks dem helfen wollenden Servicetechniker im Stress Passwort und Schlüssel für den entsprechenden Zugriff entfallen sind?

3 IT-Sicherheit

Spricht man von IT-**Sicherheit** meint man i.d.R. IT-**Security**. Im Gegensatz zu Safety gelten für Security tlw. ganz andere Randbedingungen:

- Schutzziele müssen zunächst definiert werden. Diese sind nicht automatisch gleich Mensch, Leib, Leben und Umwelt, sondern können z.B. auch Geld, zu schützendes Wissen und ein ausfallfrei laufender Online-Service sein (s. DoS).

- Eine Sicherheitslücke in der IT-Sicherheit wird – einmal bekannt – mit großer Wahrscheinlichkeit auch ausgenutzt werden. Bekannte Lücken müssen also schnell geschlossen werden und können nicht statistisch erfasst werden (i.G. zur Regelmäßigkeit eines Tsunamis der Stärke acht…)

- Eine Gefährdung wirkt nicht zufällig vom System auf seine Umwelt (z.B. der Airbag, der ohne Aufprall während einer Fahrt aufgeht), sondern die Umwelt (z.B. ein Mitarbeiter mit entsprechendem Knowhow und Motivation) attackiert ein System absichtlich um es zu „überlisten“.

- Fehlverhalten sind meist Ergebnisse von Angriffen über mehrere (Zwischen-) Stufen und nicht unabhängig. Sie können deshalb mit konventionellen Safety-Methoden nicht offenbart werden

- Ein attackiertes System soll unter allen Umständen weiterhin funktionieren. Ein sicherheits-gerichtetes System (im Sinne von Safety) muss einen sicheren Zustand erreichen (Flugzeug landet, Zug hält auf offener Strecke, ...).

Im Gegensatz zu Normen, die seit vielen Jahren die Sicherheit in Bezug auf Safety regeln (insb. IEC 61508, abgeleitete, sowie RTCA DO 178B/C et. al.) sind Normen

bezugnehmend auf Security erst noch im Entstehen. So gibt es derzeit vielfältige Ansätze aus unterschiedlichen Blickwinkeln:

- Common Criteria (ISO/ IEC 15408) versucht international Kriterien zur Sicherheits-Evaluierung von IT-Technik zu harmonisieren. Inzwischen ist ein umfangreicher Katalog zur Spezifikation von Sicherheitsanforderungen entstanden, der eine Vergleichbarkeit der Ergebnisse von Evaluierungen herstellen soll (Bsp. Bahntechnik)

- Entwicklungsprozessreife. Ähnlich wie SPICE oder CMMI gibt es Ansätze notwendige Prozesse für die IT-Sicherheit zu definieren (SAMM, BSIMM, SDL, …) und deren Erfüllung einer Reifegradstufe zuzuordnen

- Zertifizierungen: Beim BSI kann man Organisationen und Produkte basierend auf den Normen 27k ähnlich zur ISO-9000 zertifizieren lassen

- ISA 99/ IEC 62443 versucht einen Ansatz analog zu SIL-Einstufungen SAL (Security Assurance Level) vorzunehmen. Parameter sind hier Angriffe/ Hilfsmittel/ Wissen/ Ressource und Motivation.

Derzeit verfolgte Lösungskonzepte betreffen insbesondere

- konsequente(re)s "Security by Design" (durch Norm geregelt)

- Standardisierung von Protokollen wie z.B. OPC UA

- Schulung/ Know-How- und Bewusstseinserhöhung

Am Ende des Vortrags wird beispielhaft die Methodik CORAS (Cost Risk Analysis) vorgestellt, mit der sich ganzheitlich ein Prozess von der Definition von Schutzzielen bis zur Identifikation und Priorisierung von Risiko-Maßnahmen grafisch modellieren lässt. Sie visualisiert Risikobewertungen und hilft damit ein gemeinsames Verständnis zu bekommen.

"Eventually Consistent" - Neubetrachtung von Enterprise-Anwendungsarchitekturen unter BigData

Andreas Tönne

Director R&D
NovaTec GmbH
Dieselstraße 18/1
70771 Leinfelden-Echterdingen
andreas.toenne@novatec-gmbh.de

Ein zentrales Versprechen von BigData ist die Lösung von Problemen bezüglich scheinbar unmöglicher Datenvolumina und Änderungsfrequenzen. Die typischen Beispiele für erfolgreiche BigData Anwendungen decken dabei Probleme ab, die mit traditionellen Enterprise-Architekturen und -Technologien trotz Optimierung dieser Architekturen und Technologien kläglich scheitern – oder mit gutem Grund – überhaupt nicht erst versucht wurden.

Wir betrachten anhand einer Umstellung einer Enterprise-Middleware auf BigData Technologien die technischen und fachlichen Probleme und Konsequenzen einer Migration einer nach traditionellen Prinzipien entwickelten Enterprise-Anwendung nach BigData. Wir konzentrieren uns dabei auf das Spannungsfeld zwischen traditionell quasi selbstverständlichen nicht-funktionalen Eigenschaften der Anwendung, die sich aus dem ACID Prinzip ergeben, und dem sich aus dem CAP-Theorem und Lockfreiheit ergebenden Widerspruch von Konsistenz und Verfügbarkeit bei einer skalierbaren verteilten Anwendung. Dies ist kein theoretisches Problem sondern sehr real: es geht um den mit den Stakeholdern auszufechtenden Konflikt von (zeitnaher) Konsistenz der Daten und Skalierbarkeit der Plattform.

Unser Praxisbeispiel ist eine Middleware zur Datenintegration. Daten aus beliebigen heterogenen, strukturierten und unstrukturierten Datenquellen werden importiert und in ihrem Schema automatisiert vereinheitlicht integriert. Die Integration erfolgt dabei über die Ermittlung von inhaltlichen Verknüpfungen zwischen den Datensätzen anhand von semantischen Analyseverfahren. Anschaulich erhält man als Ergebnis einen Graphen von Objekten und zu einem Objekt über diesen Graphen eine Auswahl von inhaltlich "interessanten" verwandten Objekten.

Diese Middleware wurde vor ihrer Migration mit einer typischen Java EE Schichtenarchitektur unter Nutzung einer relationalen Datenbank implementiert. Die Lösung hat bei Proof-of-Concept-Projekten zu großem Kundenzuspruch in Anwendungsgebieten wie der Integration von Unternehmensdaten geführt. Der reale

Datenumfang hat aber gezeigt, dass die nicht-funktionalen Anforderungen die Leistungsfähigkeit der gewählten Architektur und Technologien um mehrere Größenordnungen überschreiten würde: Es geht um viele Millionen von Dokumenten mit vielen Milliarden von inhaltlichen Verknüpfungen.

Ein sehr charakteristisches Problem ist die Einhaltung von Konsistenzkriterien, die die horizontale Skalierbarkeit hart begrenzen. Man stelle sich vor, dass zwei inhaltlich zueinander bezogene Datensätze zeitgleich importiert werden, wodurch Konsistenzanforderungen des Graphen verletzt werden. In einer traditionellen Enterprise-Architektur wird dies zum Beispiel über Locking oder über Datenbankconstraints abgesichert. Derartige Konsistenzanforderungen sind regelmäßig in Enterprise-Datenmodellen anzutreffen. In einer hochgradig horizontal skalierenden, verteilten Architektur muss man dagegen sicherstellen, dass der Grad der Synchronisation zwischen den Anwendungsknoten minimal ist, damit die Knoten unabhängig voneinander rechnen können und nicht aufeinander warten müssen.

Wir haben uns in dem Projekt aufgrund des existierenden fachlichen Graphmodells für eine Graphdatenbank (Titan) mit NoSQL Storage (Cassandra) und verteiltem Index (ElasticSearch) entschieden. Den oben genannten Konflikt lösen wir durch den neuen Technologiestack aber nicht. Durch die bessere Skalierbarkeit der neuen Architektur wurde der Konflikt sogar noch verschärft und erfordert deshalb einen Paradigmenwechsel.

Zu erheblichem Diskussionsbedarf mit den Stakeholdern und erheblichem Forschungsaufwand hat geführt, dass der Übergang zu einer skalierbaren, verteilt rechnenden und speichernden Anwendungsarchitektur zwangsläufig zu Änderungen der funktionalen und nicht-funktionalen Eigenschaften einer Enterprise-Anwendung führen muss.

Für diese wird in der Regel stillschweigend das ACID Prinzip vorausgesetzt. Für uns ergibt sich daraus die Erwartung der Stakeholder, dass Daten sofort nach ihrer Erfassung konsistent verfügbar sind. In einer BigData Architektur mit hochgradig horizontal skalierenden, parallel ausgeführten Komponenten ist ACID aber nicht immer einzuhalten. Es kann für eine begrenzte Zeit nach einer Datenänderung nicht garantiert werden, dass die Konsistenz der Daten auf allen Knoten gilt, was mit "eventually consistent" bezeichnet wird.

Ganz praktisch ergab sich für uns die Notwendigkeit der Entscheidung, ob wir für das obige Beispiel der Erweiterung des Graphen mehr Wert auf Konsistenz (synchronisierte Änderungen) oder Antwortzeit (nicht synchronisierte Änderungen) legen wollten. Bei einer verteilten, nicht synchronisierten Datenänderung ergibt sich zwangsläufig ein Zeitfenster, in dem die Daten entweder potenziell inkonsistent aber dafür auf allen Knoten verfügbar sind oder in dem die geänderten Daten nicht auf allen Knoten gesehen werden. Wir haben uns für ein Zeitfenster der Inkonsistenz entschieden und damit eine sehr gute horizontale Skalierbarkeit und hohen Durchsatz beim Import von Daten erzielt. Dieses Zeitfenster wird durch periodische Korrekturprozesse auf Hadoop-Basis geschlossen. Wichtig für diese Entscheidung war eine Abschätzung der Auswirkungen

der Konsistenzfehler. Es gibt ebenso Einsatzszenarien der Middleware, in denen bestimmte Daten durch eine alternative Analytik konsistent aber zwangläufig mit deutlich schlechterem Durchsatz zu importieren sind.

Die Umstellung auf BigData hat nicht nur funktionale Änderungen und Einschränkungen bedeutet. Die neue Middleware hat in großem Maße von der Zuverlässigkeit der BigData-Komponenten Titan, Cassandra und ElasticSearch profitiert. Sowohl Cassandra als auch ElasticSearch lassen sich zur Lastverteilung und Hochverfügbarkeit sehr leistungsfähig clustern. Sie bieten eine zuverlässige Replikation an, die gerade für BigData Datenvolumen für eine schnelle Erholung von Ausfällen von Knoten sehr wertvoll ist.

Unser Fazit: Die Migration einer bestehenden Enterprise-Anwendung auf einen BigData-Technologiestack setzt voraus, dass eine Abstimmung mit allen relevanten Stakeholdern erfolgt. Es muss eine Abwägung und Entscheidung der möglichen Änderungen der bisherigen fachlichen und nicht-fachlichen Eigenschaften erfolgen. Ebenso muss beachtet werden, dass die hoch-parallele Verarbeitung, mit zum Beispiel Hadoop, zusätzliche völlig andersartige Algorithmen und Geschäftsregeln benötigt. Ohne eine Entscheidung zu Änderungen an den Anforderungen, und hier besonders an der Abwägung von Skalierbarkeit und Konsistenz, läuft die Migration Gefahr, die Erwartungen an die Skalierbarkeit und damit an den Durchsatz nicht erfüllen zu können.

Aufbrechen der Datensilos – Big Data

Forschungsfragen aus dem Bereich Industrial Analytics

Benjamin Klöpper, Jan Christoph Schlake

Industrial Software Systems
ABB AG Forschungszentrum Deutschand
Wallstadter Str. 59
68526 Ladenburng
benjamin.kloepper@de.abb.com
jan-christoph.schlake@de.abb.com

Abstract: Aufgrund isolierter Datenhaltung und –verarbeitung in Datensilos kann das volle Potential der Analyse historischer Daten im Umfeld der Prozessindustire oftmals nicht ausgenutzt werden. Big Data Technologien bieten Chancen, diese Situation zu verbessern. Gleichzeit stellen sich aber neue Herausforderungen, die den Einsatz von Big Data in diesem Umfeld behindern.

1 Kurzfassung des Vortrags

In einer typischen Prozessanlage findet sich in der Regel eine Vielzahl von *Datensilos*, gesammelte Daten werden in isolierten Informationssystemen gespeichert und verarbeitet. Ähnlich wie im Fall von funktionalen Silos [Ame88] wird durch die isolierte Speicherung und Verarbeitung der Daten das Lernen aus historischen Daten durch Vergleichen und Erkennen von Korrelation erschwert. In einem solchen Zustand ist Datenanalyse eine sehr mühselige Aufgabe und erfordert sowohl tiefgehendes Verständnis für die Domäne oder sogar die einzelne Anlage wie auch erhebliche Personen- und Zeitaufwände. Big Data Technologien versprechen die Möglichkeit, Lösungen für Industrial Analytics leichter bereitzustellen und flächendeckend auszurollen. Gleichzeitig stellt dieses Anwendungsszenario einige Herausforderungen an die Big Data Research Community.

Big Data Technologien werden entwickelt, um Daten in großen Mengen (high volume), mit hoher Geschwindigkeit (high velocity) oder sehr unterschiedlichen Formaten (variety) zu behandeln [MB12]. Eine vollständige Integration aller Daten einer Prozessanlage haben sicherlich diese Eigenschaft: historische Prozessdaten in unkomprimierter Form erreichen Terabyte an Datenvolumen, die Daten haben unterschiedlichste Formate wie Zeitreihendaten oder unstrukturierten Text (z.B.

Schichtbücher). Sollen die Daten direkt während der Ausführung analysiert werden, müssen Datenströme mit einer hohen Samplingrate unter Echzeitanforderungen bearbeitet werden. Eine Vielzahl von Software existiert bereits, um diesen Anforderungen zu genügen.

Die zentrale nicht-funktionale Anforderung von *Industrial Analytics* sind aber niedrige Investment- und Betriebskosten sowie Nutzbarkeit durch Nicht-Informatiker. Diese Anforderungen lassen sich mit den heutigen Big Data Technologien nur eingeschränkt erreichen. Insgesamt bestehen vor allem folgende Herausforderungen:

Geringste Kosten für die Datenspeicherung: Der Einsatz von Standardhardware zur Datenspeicherung stellt einen ersten Schritt zur kostengünstigen Speicherung von Daten dar. Allerdings werden nicht immer alle Daten in unkomprimierter Form für Berechnungen benötigt. Die Verwendung bspw. eines Clusters mit dem Hadoop File Systems [SK+10] verursacht hier wesentlich größere Kosten als die Speicherung der Daten auf Network Attached Storages (NAS). Das Konzept des Data Lakes ist für *Industrial Analytics* nicht wirtschaftlich. Es werden Konzepte benötigt, Daten auf unterschiedliche Speichertechnologien zu verteilen und diese je nach analytischem Bedarf transparent zwischen den einzelnen Speicherformen zu verschieben.

Unterstützung von Machine Learning Methoden: Frameworks wie Mahout [OA+11] oder mlib für Spark setzen zwar Machine Learning Algorithmen auf Big Data Technologien um. Jedoch ist der Satz an Methoden nicht so vollständig wie in vergleichbaren „Small Data“ Bibliotheken. Zusätzlich unterscheiden sich die Umsetzungen und Aufrufe teilweise erheblich, je nachdem ob eine Batch-, In-Memory oder graph-basierte Technologie eingesetzt wird. Dies bedeutet einen hohen Portierung und Re-Implementierungsaufwand für *Industrial Analytics* Projekte.

Einfache Datenvorverarbeitung und –bereinigung: Datenvorverarbeitung und -bereinigung sind oftmals die zeitlich größten Aufwände in einem Industrial Analytics Projekt. Hier fehlen sowohl wiederverwendbare Vorgehensmodelle als auch generische Tools, die den Domänenexperten ohne Data Mining-Expertise effektiv unterstützen

Datensicherheit: Prozess- und Produktionsdaten können geschäftskritische Informationen wie bspw. Produktionsmenge oder Rezepturen offenlegen. Daraus ergeben sich starke Anforderungen an die Datensicherheit. Cloud-deployments von Big Data Technologien, die günstige Unterhaltskosten bieten, sind hier kritisch zu betrachten. Um im industriellen Umfeld akzeptiert zu werden, sind überzeugende Sicherheitskonzepte erforderlich.

Literaturverzeichnis

[Am88] AME Study Group on Functional Organization: Organizational Renewal – Tearing Down the Functional Silos, AME Target, 4 – 16, 1988

[ME12] McAfee, A.; Brynjolfsson, Erik: Big Data: The Management Revolution, Harvard Business Review, October 2013

[SK+10] Shvachko, K.; Kuang, H.; Radia, S.; Chansler, R.: The hadoop distributed file system. In IEEE 26th Symposium on Mass Storage Systems and Technologies (pp. 1-10), 2010
[OA+11] Anil, R.; Dunning, T.; Friedman, E.: Mahout in action. Manning, 2011

Professionelles Testmanagement in Datenreinigungsprozessen

Dr. Holger Petersen
Senior IT Consultant
XQT - Test & Quality Management
msg systems ag
Robert-Bürkle-Straße 1
85737 Ismaning/München
holger.petersen@msg-systems.com

Abstract: In diesem Beitrag wird das Testmanagement bei Maßnahmen zur Qualitätsverbesserung in Datenbeständen behandelt. Hierbei fließen vorwiegend Erfahrungen aus Projekten bei Erst- und Rückversicherungen, Rentenversicherungsträgern und in der öffentlichen Verwaltung ein. Dies sind Bereiche, in denen die Speicherung und Verfügbarkeit der Daten aus vielen Jahrzehnten notwendig ist. Naturgemäß werden diese Daten mehrfach migriert und mit verschiedenen Applikationen in immer neuen Versionen bearbeitet, was sowohl Volumen wie auch Komplexität der Daten erklärt und das Testmanagement bei Qualitätsverbesserungen (Datenreinigungen) zu einer sehr anspruchsvollen Aufgabe macht.

Erhalt und Verbesserung der Datenqualität ist bei lange im Einsatz befindlichen Applikationen mit großen Datenvolumina eine zentrale Frage [Nau]. Typische Qualitätskriterien sind Korrektheit, Vollständigkeit, Aktualität und Konsistenz [SMB05]. Fehler im Datenbestand können vielfältige Ursachen haben, wie bewusste Falschangaben (z.B. bei Volkszählungen), Irrtümer (z.B. falsche oder lückenhafte Angabe der besuchten Schulen), Eingabefehler (z.B. Hörfehler bei telefonischem Kontakt oder Tippfehler), Fehler der Applikation (z.B. doppeltes Schreiben von Datensätzen) oder fehlerhafte Migrationen (z.B. Abschneiden von Kundennummern aus einem Altsystem). Einige Fehler können durch fehlererkennende oder -korrigierende Kodierung vermieden werden (IBAN, Kreditkartennummern) oder durch Abgleich mit einer Datenbasis (Adressen). In vielen Fällen ist eine solche Verhinderung von Fehlern aber nicht möglich.

Mängel der Datenqualität können weitreichende negative Auswirkungen haben. Dies ist bei Verärgerung der Kunden durch doppelte Belastung von Rechnungsbeträgen oder falsch berechnete Verzugszinsen sofort einsichtig. Aber auch Fehler, die keine sofortige Wirkung zeigen, können über den Weg falscher statistischer Auswertungen und Prognosen durch darauf beruhende Fehlentscheidungen dramatische Auswirkungen bis zur Existenzbedrohung von Unternehmen haben. Auch auf die Qualität von Testergebnissen können fehlerhafte Daten negative Wirkungen haben. Die Beseitigung von Datenqualitätsmängeln (Datenreinigung) und Verhinderung von neuen Fehlern ist somit eine wichtige und anspruchsvolle Aufgabe. Die Qualitätssicherung von Datenreinigungen bildet aus unserer Sicht einen Bestandteil professionellen Testens, das den gesamten Anwendungslebenszyklus begleiten sollte [Jac09].

Die Beratung zu Fragen der Datenreinigung zählt seit vielen Jahren zu den zentralen Themen des Crossbereichs XQT - Test & Quality Management der msg. Stetig zunehmende zu verarbeitende Datenvolumina in immer komplexeren Anwendungslandschaften bei gleichzeitig wachsendem Anspruch an die Softwarequalität haben dazu geführt, dass sich Spezialdisziplinen herausgebildet haben, die im Beitrag ausführlich dargestellt werden. Ein besonderer Vorteil des Crossbereiches XQT ist es, dass er in Kooperation mit den Branchenbereichen der msg systems ag an vielen Einführungsprojekten beteiligt war und ist, wodurch fundierte Kenntnisse der Applikationen erworben werden konnten, die in die Wartungsphase dieser Produkte eingebracht werden können.

In diesem Beitrag soll der Schwerpunkt auf Datenfehlern liegen, die durch technische Ursachen entstanden sind. Diese Fehler sind in vielen Fällen innerhalb des betrachteten Systems erkennbar und bei ausreichender Redundanz ist sogar eine Korrektur ohne zusätzliche Informationen möglich. Mindestens können Auswertungen zur Verfügung gestellt werden, die erforderliche zusätzliche Daten und Korrekturen charakterisieren.

Die Themen Testplanung, Teststufen, Testfallentwurf, Testdatenanlage und Testdurchführung werden basierend auf Projekterfahrungen im Hinblick auf die Besonderheiten von Datenreinigungstests besprochen.

Literatur

[BMB+] Rex Black, Judy McKay, Graham Bath, Debra Friedenberg, Bernard Homès, Kenji Onishi, Mike Smith, Geoff Thompson und Tsuyoshi Yumoto. *Certified Tester - Advanced Level Syllabus Testmanager VERSION 2012*. Deutschsprachige Ausgabe. Herausgegeben durch Austrian Testing Board, German Testing Board e.V. & Swiss Testing Board.

[Jac09] Recardo Jackson. Testmanagement: Professionelles Testen. *Informatik Spektrum*, 32(1):37–41, 2009.

[Nau] Felix Naumann. *Informatiklexikon: Datenqualität*. https://www.gi.de/service/ informatiklexikon/detailansicht/article/datenqualitaet.html (abgerufen am 23.02.2014).

[SMB05] Monica Scannapieco, Paolo Missier und Carlo Batini. Data Quality at a Glance. *Datenbankspektrum*, 14:6–14, 2005.

[Spi08] Andreas Spillner. *Systematisches Testen von Software — Ein Einstieg*. dpunkt.verlag, 2008.

[SRWL11] Andreas Spillner, Thoma Roßner, Mario Winter und Tilo Linz. *Praxiswissen Softwaretest — Testmanagement*. dpunkt.verlag, 2011. 3., überarbeitete und erweiterte Auflage.

Simulation Technology: Systems for Data Intensive Simulations

(SimTech@GI)

Simulation Technology: Systems for Data Intensive Simulations (SimTech@GI)

Dimka Karastoyanova[1], Dirk Pflüger[2],

[1]Universität Stuttgart, Institut für Architektur von Anwendungssystemen
Universitätsstr. 38, 70569 Stuttgart
karastoyanova@iaas.uni-stuttgart.de
[1]Universität Stuttgart, Institut für Parallele und Verteilte Systeme
Universitätsstr. 38, 70569 Stuttgart
Dirk.Pflueger@ipvs.uni-stuttgart.de

Isolated and monolithic software is not capable of meeting the requirements of nowadays complex simulations any more. Novel approaches are needed to cover all phases of the simulation pipeline – starting from the discretization of models, through software system development, and to the last stage of visualization of the simulation results. Especially challenging are the cases of coupled simulations, where simulation parameters and data have enormous impact, which imposes a strong need for a well-designed simulation workflow.

This workshop focuses on computer science and computational aspects that enable complex simulations. The contributions cover, for example, different aspects of the coupling of simulations, be it on the level of workflow choreographies or in the direct interplay of multi-physics simulations, and the quality and management of simulation results is addressed as well as the efficient integration of tools for their analysis.

We thank all those who contributed to this workshop, especially all reviewers and all authors.

Towards Quality-aware Simulations on Mobile Devices

Christoph Dibak, Boris Koldehofe

Institute for Parallel and Distributed Systems
University of Stuttgart
Universitätsstraße 38
70569 Stuttgart, Germany
{firstname.lastname}@ipvs.uni-stuttgart.de

Abstract: Numerical simulations are import for analyzing Big Data and realizing applications in the Internet of Things. Running numerical simulations on mobile devices makes analyzing and reasoning about Big Data ubiquitous. However, mobile devices are limited in energy and compute resources, and connectivity to a dedicated infrastructure like a cloud cannot always be assured. Therefore, we propose to run the simulation on a distributed environment consisting of a mobile device and the cloud. This environment has a number of constraints for compute and network resources that need to be considered for providing simulation results in time and with high quality. In this paper we propose an architecture for mobile simulations and list challenges for realizing them.

1 Introduction

Numerical simulations have many applications ranging from stress tests in computer aided designs (CAD) over traffic analysis to weather forecast. In engineering and natural sciences, dynamic complex systems and processes are often described by means of partial differential equations. Numerical simulations allow understanding the dynamic evolution of such systems that follow from such a description. Numerical simulations are also important for analyzing huge data sets in the context of Big Data and the Internet of Things. Given real-world measurements and data, simulations can predict important future critical states or situations within a complex system. Embedded in a control-loop, a controller can react to predicted states in order to optimize or stabilize the future evolution of the system.

Whenever the result of a simulation needs to be directly available on a mobile device, the term *mobile simulation* can be applied. For instance, a controller reacting to critical states and situations in traffic, navigation, or logistics is embedded in a car, ship or drone. Furthermore, in a wide range of applications like medicine, sports, or agriculture, decision makers like the doctor, trainer, or farmer require personalized simulations that depend on the current context, e.g. the status of a patient, the training activity, or plants.

A key challenge is how to realize mobile simulations so they yield i) sufficient quality of result, ii) offer the actors sufficient responsiveness and iii) are energy efficient for the mobile device. Naive methods in deploying mobile applications are likely to yield unsatisfactory

results in the context of mobile simulations. First, mobile simulation may be executed similar to traditional simulations on a high performance computing infrastructure, offered by a cloud provider. Although a simulation may benefit from strong compute resources, temporal loss of connectivity is common in mobile services and can severely impact the time until actors can react to simulated results, e.g. when no network is available. Second, loss of communication may be tackled by running the entire simulation on the mobile device themselves. Even though mobile devices are becoming more powerful and already replace their stationary counterparts in many fields like image and video processing, it should be observed that simulations are highly demanding regarding processing resources. Simulations like in fluid dynamics require solving linear equations with thousands of unknowns. Performing all computations of a simulation on the mobile device will therefore lead to high response time and energy usage or allow only low quality results by simplifying the underlying simulation model.

In this paper, we present our vision towards supporting mobile simulations that establishes a trade-off between the two deployment possibilities dynamically under run-time constraints like the connectivity of the mobile and constraints of the decision maker on time and quality to respond to results of a simulation. In the remainder of the paper we give an overview of what numerical simulations are and how mobile simulations can be realized using existing work. In Section 3 we propose an architecture to realize mobile simulations overcoming the drawbacks from a naive deployment. Section 4 describes challenges for implementing this architecture, possible methods and further aspects to be regarded during implementation of the proposed architecture.

2 Background

In numerical simulations, the behavior of the system is described by a mathematical model such as partial differential equations (PDEs). These equations are defined on continuous space and time also called *domain* of the simulation. As computers are not capable of dealing with continuity, time and space need to be mapped to a finite set of discrete points. The discrete points form a mesh on the continuous spaces. Using this mesh, the partial differential equations can be numerically solved using a system of linear equations.

Figure 1 gives an overview of the process. On the left, the numerical treatment of PDEs is depicted. The model is kept in partial differential equations. These PDEs are discretized using techniques such as finite elements, finite differences or finite volumes. Those discretization techniques allow transforming the simulation to a finite mathematical problem like linear equations. The mathematical problem is then solved using a task-specific Solver, e.g. to solve a system of linear equations iteratively. In this paper, we deal with the question how to run this process within a distributed environment consisting of the mobile device and a server hosted in the cloud (cf. right side of Figure 1). In order to run the numerical simulation on this environment, quality has to be adapted to run-time constraints of the user and the available network and compute resources.

There are a number of methods of how mobile simulations could be realized using existing

Figure 1: Numerical simulations have to be adopted to run-time constraints in order to run on a mobile distributed environment.

work which can be categorized into two main classes. On the one hand, there are methods reducing quality of simulation so that the entire simulation can be efficiently executed on the mobile device [ASG01, HDO11]. On the other hand, there are methods to execute programs on a distributed infrastructure [CBC+10, CIM+11, RSM+11, GRA12, Fuj99, BPOS12].

In *model reduction*, the numerical simulation is transferred into a lower dimensional space [ASG01, HDO11]. This reduces quality, while it also reduces computational cost. Using model reduction, it is possible to run the entire simulation on mobile devices. A drawback of the approach is that no additional compute resources accessible over a wireless link can be used. Utilizing many compute resources in the course of a simulation is often addressed in high performance computing by a parallelization of the simulation's execution [Fuj99, BPOS12]. However, the execution of a parallel simulation requires very frequent exchanges of states and synchronization between the parallel simulation instances. Therefore, parallelization yields good response time only when—unlike in a mobile environment—communication is subject to low latencies and high availability.

A promising alternative are *code offloading* techniques. Code offloading is a paradigm where some parts of a mobile computation is offloaded to the cloud in order to either speed up the application [RSM+11, GRA12] or to save energy on the mobile device [CBC+10, CIM+11]. While both aspects are important for mobile simulations, code offloading is oblivious to the quality of the results of an application. In particular, code offloading per se will not be able to benefit from different discretization parameters of the simulation model.

3 An Architecture for Mobile Simulations

The proposed architecture of the mobile simulation middleware supports the execution of a mobile simulation to be executed on two interconnected *nodes*, the mobile device and a server in the cloud. Both nodes are assumed to be connected through a wireless communication link such as WiFi (IEEE 802.11) or a cellular network. As the mobile device is carried by users, the wireless connection is subject to varying bandwidth, e.g. a mobile device might leave the coverage area of its company's WiFi and therefore has to

rely on cellular networks, providing significant lower bandwidth.

In order to initiate a simulation a request needs to be posed to the middleware in form of a tuple (t_S, t_D). The request defines that state of the simulated system at time t_S should be available on the mobile device before the deadline t_D expires. Our middleware should provide an adequate quality with respect to a trade-off between available resources and risk to exceed the deadline t_D.

Note, there is no perfect simulation. As it would cost too much compute resources to simulate all effects, some aspects of the real-world system have to be omitted during modeling and implementation [BZBP09]. In order to make run-time decisions our middleware requires to change some well-defined parameters during run-time that effect the trade-off between computation cost and quality. This may include time and space discretization grids (cf. Section 2). There are further parameters, e.g. the maximum error of an iterative solver after the discretization process or even different models for simulating the system. However, we will focus on discretization grids as they have a significant impact on computation time and state transfer volume.

The method of how quality can be adopted in the discretization process depends on the applied discretization method. For finite differences, only the mesh size can be changed. For more advanced discretization techniques like finite elements, test functions can also be changed [BZBP09]. However, as the method used for discretization is chosen by the programmer, she also has to specify how quality can be changed in her implementation.

Figure 2 depicts our architecture consisting of four components, named discretization controller, quality predictor, resource statistics and run-time protocol. These components will be explained in detail in the remainder of this section in the context of the following example: Consider a sailor on a boat near the cost and the sailor wants to find the fastest route to a nearby harbor. The fastest route depends on the wind flow near the coast. Wind depends on weather conditions as well as the structure of the coast like hills. To keep our simulation simple, we will neglect the flow of the sea and just focus on the wind flow. The behavior of wind flow will be modeled using Navier-Stokes equations [KPL00]. In our example, these equations are discretized using the finite volume method. For simplicity, we assume a quadratic 2D spatial domain. The quadratic domain is cut into squares of width Δx. The time domain is divided into equidistant time steps of length Δt. The simulation will provide the sailor with a map of the wind flow. This map is generated using data of nearby weather stations. In a more advanced example, sensor data of nearby boats might also be integrated into the simulation. To run and see the results of the mobile simulation, the sailor uses a mobile device like a tablet or smartphone. Dependent on the distance to the coast the communication link between the mobile device and the cloud can strongly vary in its characteristics. If the sailor is near the coast, the device might be connected through a 3/4G cellular network. When the sailor is too far off the coast, the mobile device may require satellite communication, or there may be temporarily no connection at all.

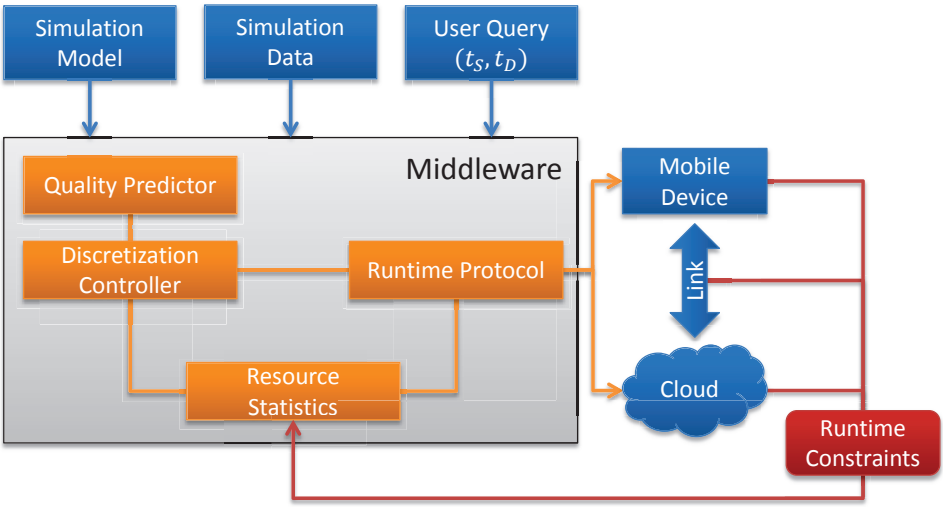


Figure 2: Components of the Architecture for Mobile Simulations

3.1 Quality Predictor

For multiple discretization parameters it is hard to find the optimal combination in order to have good quality with low computational cost. The question whether further compute resources should be invested for having finer spatial (in our example Δx) or finer temporal (in our example Δt) discretization to achieve maximum quality is not trivial [LDBNR13].

The *quality predictor* component is used to predict the quality of a fixed combination of discretization parameters. Given a combination of discretization parameters, it returns a number expressing the predicted quality of the result of the simulation. Quality can be defined by performing a comparison of the simulated to the real-world system. To obtain the real-world system state, either sensor data can be used, or the system has to be simulated using a reference combination of discretization parameters. Computing the outcome of the simulation using the reference discretization costs a lot of time. Therefor the spatial and temporal domain of the simulation needs to be constrained.

In order to allow for time efficient computations, the reference simulation should be run on a highly constrained domain. Neglecting errors of inaccurate sensors, the model of the simulation and other effects, the outcome of the reference simulation and observed sensor values can be treated as the 'real' behavior of the system. Finally, the behavior of the 'real' system and the simulation with given discretization parameters is compared by a user-defined metric which is suited best for expressing quality in the application.

In our example, the quality predictor component will receive requests $(\Delta x_0, \Delta t_0)$. It will use a very fine reference resolution $(\Delta x_{ref}, \Delta t_{ref})$, expressing the behavior of the real system. For predicting the quality, the spatial domain will be restricted to a smaller square. In order to obtain results that are valid for the entire domain, we can choose a region where

the velocity of the flow was very fast, or changed its direction frequently in the past. When the reference simulation with $(\Delta x_{ref}, \Delta t_{ref})$ and the requested simulation with $(\Delta x_0, \Delta t_0)$ are finished, their results will be compared. As metric, the maximum different velocity at any point in the simulated domain can be used. This value is returned by the quality predictor component.

3.2 Resource Statistics

The *resource statistics* component will collect data on load, bandwidth and latency on the nodes and the link connecting them. The data is interpreted statistically to predict available resources in future computation steps of a simulation.

There are two components relying on resource statistics. The discretization controller needs statistical information on how load and bandwidth will develop in the future to set the discretization parameters. During run-time, the run-time protocol needs to decide when to transfer state from one node to another to estimate the risk in exceeding the bandwidth. The available bandwidth depends on the location of the mobile device. Therefor the resource statistics component may benefit from known mobility patterns of the device to improve the prediction of available bandwidth. In particular, the mobility of the device is restricted in many applications. For example, if the simulation is executed on a car, positions will be limited to the street network. Additionally the route can be known in advance as the destination might be known to the car's navigation system. All this information can be used to predict available resources for future computations.

In our example, the resource statistics component might collect data on where 3/4G cellular network is available. As the smartphone or tablet knows its current position and its heading, it might be able to predict its future position. This position can be used to predict available resources for the future.

3.3 Discretization Controller

In numerical simulations, there are several ways of how quality can be adopted to the availability of resources. One way is to change the discretization of time and space. If the simulation consists of fewer discretization points, computation time and quality is reduced. Discretization parameters affect different resources, e.g. while temporal discretization just affects compute resources, spatial resolution also affects the volume of data having to be transferred over the air.

Goal of the *discretization controller* is to decide which discretization parameters to choose. Discretization parameters should be chosen such that the simulation i) has decent quality and ii) does not exceed the deadline t_D.

The discretization controller has to rely on the resource statistics to estimate the available computing and network resources. On the basis of these statistics it can find a set of

possible combination of discretization parameters. Every possible combination is fed into the quality estimator. The best quality combination is set as the discretization parameter for the simulation.

For example, the flow simulation running on a device on a boat can use the resource statistics component to obtain a set of feasible discretization parameters $P_f = \{(\Delta x_i, \Delta t_i)\}$. A simple implementation would annotate each of these parameter combinations $(\Delta x_i, \Delta t_i)$ with the expected quality q_i using the quality predictor component. After the annotation process it might select the parameter combination with maximum q_i for execution.

3.4 Run-time Protocol

During execution of the simulation, the *run-time protocol* controls which of the computations are performed on the mobile device and which computations are performed in the cloud. The run-time protocol is a distributed online algorithm executed on both the mobile device *and* the cloud which intends to minimize the resource usage on the mobile device so that the result of a simulation is available within the specified deadline t_D. To this end, the online algorithm needs to decide for each time-discretization step whether it should be executed on a node or not. It ensures that at least every time discretization step will be executed either on the mobile device or the server in the cloud. In order to perform a time-discretization step the results of the immediate preceding discretization step must be available on the device. Therefore, the run-time protocol also decides when to send updates between the cloud server and the mobile device. The online-decisions depend on the information about resource statistics which are also used to predict when meeting a deadline becomes infeasible. In this case, the run-time protocol needs to re-invoke the discretization controller to choose new discretization parameters at lower quality.

A simple run-time protocol for our sailing boat example would be for the device on the boat to wait for a computation result of the cloud. If the network suddenly becomes and stays unavailable, the simulation has to be processed on the mobile device. To save time and energy in such a case, the server in the cloud can transfer intermediate results to the mobile device from time to time.

4 Challenges

As the mobile device is subject to movement by the user, its bandwidth on the link to the server in the cloud varies. Through this variation there is the risk of exceeding the deadline as a state transfer from the cloud to the mobile device might take longer than expected. Figure 3 depicts such a situation. The discretization controller decided to split the computation into 5 discrete time steps. The figure depicts two executions, (1) and (2). On (1), the last step of the computation is run on the mobile device, where in (2), the simulation is completely run on server in the cloud. During the computation of time step 5 on the cloud, the bandwidth decreases. Thus it is no longer possible for execution (2) to deliver

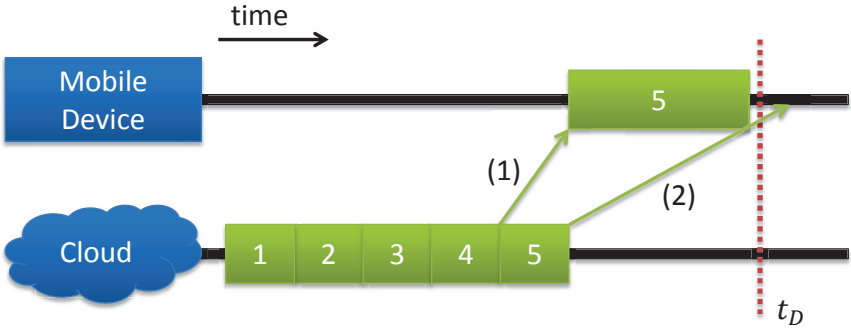


Figure 3: Varying bandwidth can cause the system to miss the deadline t_D.

the end state of the simulation to the mobile device before the deadline is reached. We will describe the following challenges using this example.

4.1 Dealing with Bandwidth Variations

In the previous example, one execution exceeds the deadline since the final state transfer takes longer than expected. A simple solution to this problem would be to run both executions, (1) and (2) at the same time. The run-time protocol is not restricted to atomic executions of time steps. It might compute time step 5 on both, the mobile device and the cloud. As the computation on the mobile device typically takes longer, the computation can be stopped when the state of time step 5 is successfully transferred to the mobile device. However, running the same computation on multiple nodes also consumes more resources. It will lead to increased bandwidth usage as multiple states need to be transferred over the wireless medium. It will also cost more energy for communication and computing.

In cases where the available bandwidth is low and updates to the states can be only transferred at low rate, the server may have finished with the computation of several time-discretization steps before successfully sending the update of a single state. In such a situation the run-time protocol has to decide whether the transfer of the old time step should be continued or if it should be stopped to transfer the new time step.

Another challenge is to find a good initial combination of discretization parameters. It would be easily possible to have a 6th time step if the bandwidth would be constantly as good as in execution (1). However, this would also increase the risk of not meeting the deadline. This risk is also increased by choosing a finer spatial discretization as the volume needed for state transfers increases. The discretization controller has to carefully optimize quality. While optimizing quality it has to take into account a constraint on the risk to exceed the deadline the user is willing to take.

To have more choices during run-time, data can be compressed while sending it over the link. Especially when there is only small bandwidth available, compression might be the only way of how the final state can be delivered to the mobile device in time. There are

two classes of methods of how data can be compressed. Either the compression does not affect quality (lossless compression) or it does (lossy compression). Lossless compression is typically asymmetric. It takes more computational power to compress data than to decompress it. Thus it would be most interesting to compress data on the way from the cloud to the mobile device. There is a trade-off between decompressing and receiving data over a wireless medium in terms of energy cost. Either the energy is consumed by the processor for decompression or by the communication system to receive the data. Using lossy compression, the run-time protocol might send low quality states in order to reduce the risk of exceeding the deadline. Computation can be continued with reduced quality on the mobile device to have a backup result if the link is down and the final result cannot be transferred before the deadline is reached. One easy way to realize lossy compression would be to simply omit some spatial points of the simulated area. However, there are more advanced techniques like *proper orthogonal decomposition*, which also has a background in Model Reduction [ASG01].

Dealing with varying bandwidth is a critical task for the design of mobile simulations. In order to be able to describe the risk, statistical models need to be developed. The run-time protocol needs to decide if it should compress states in order to transfer it over the wireless medium even when bandwidth is scarce. However, this needs careful modeling as the user not only has to set a risk of exceeding the deadline but also a risk to get non-optimal quality.

4.2 Limited Energy Resources

Another challenge for mobile simulations are limited energy resources. Energy plays an important role in mobile computing. As typical code offloading exploits the trade-off between computation and transfer cost, mobile simulations might need to run several computations on both computing nodes in parallel.

Energy consumption on mobile devices is already understood very well [BBV09, TAN+12]. There are several approaches of how energy consumption can be reduced, e.g. by shifting the transmission time to deliver messages in a bulk. Especially with multiple processors such as a graphical and a main processor, there might be different energy and computational characteristics for simulations. The usage of graphical processors for computation on smartphones is already provided by the latest Android version [And14]. Although there are no hardware vendors implementing these interfaces yet, they are expected to implement it in the near future.

Since the quality of simulations depend on the validation of simulation parameters against real-world measurements, e.g. in form of global sensor grid [BKR14], the energy-efficient execution of a mobile simulation may strongly depend on how the integration of sensor measurements is achieved. In particular, *participatory sensing* utilizes sensors on the mobile devices themselves to gain increased access to sensors to sense various environmental properties such as pollution or temperature [LML+10]. Sensor measurements of a participatory sensing system cost on the one side energy. On the other side, quality can be increased, as more information about the real-world system is available. If the discretization controller

decides to run the simulation using a certain combination of parameters, sensor data can be requested to meet required quality while minimizing energy cost.

While optimizing quality, mobile simulations should also be constraint in their energy consumption. Components need to be quality-aware. Methods to find out how much energy resources will be needed for an execution have to be developed. On top of that, methods to interact with sensor networks to provide just the right quality for the combination of discretization parameters are needed.

4.3 Monetary constraints

Computations in the cloud typically cost money and therefore impose additional constraints for the execution of the mobile simulation. There are numerous pricing mechanisms discussed in the literature to determine dynamically the resources in cloud dependent on the user behavior and the availability of resources [AWS12]. For instance, with *spot-instances* [AWS12] the price is set by the users. Every user decides how much she is willing to pay to run her spot-instance. If the price given by the provider is lower, the instance will be started. If the price rises above the user defined threshold, the instance will be stopped.

Especially for predicting the quality of a simulation the understanding of pricing mechanisms is highly important. To avoid dependency on a single spot instance, we may need to refine our architecture to benefit from multiple servers in the cloud to increase the availability of resources.

5 Conclusions

In this paper we have proposed an architecture for mobile simulations and described challenges we need to address as part our future work for their realization. In particular, we highlighted in the context of simple distributed execution model consisting of a mobile device and a server in the cloud, mobile simulations can benefit from a distributed execution on both nodes in order optimize quality under time and resource constraints. The validation of the requirements and evaluation of the conceptual design will be performed in the course of simulation models explored by researchers in the context of the Cluster of Excellence in Simulation Technology.

Acknowledgments The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXEC 310/1) at the University of Stuttgart.

References

[And14] Android Developers. Render Script, Juni 2014. http://developer.android.com/guide/topics/renderscript/compute.html.

[ASG01] Athanasios C. Antoulas, Danny C. Sorensen, and Serkan Gugercin. A survey of model reduction methods for large-scale systems. *Contemporary mathematics*, 280:193–220, 2001.

[AWS12] Amazon Web Services. How AWS Pricing Works, March 2012. http://aws.amazon.com/whitepapers/.

[BBV09] Niranjan Balasubramanian, Aruna Balasubramanian, and Arun Venkataramani. Energy consumption in mobile phones: a measurement study and implications for network applications. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 280–293. ACM, 2009.

[BKR14] Andreas Benzing, Boris Koldehofe, and Kurt Rothermel. Bandwidth-Minimized Distribution of Measurements in Global Sensor Networks. In *To appear in Proceedings of the 14th IFIP International Conference on Distributed Applications and Interoperable Systems (DAIS 2014)*, pages 1–15. Springer, Juni 2014.

[BPOS12] Jose J Blanco-Pillado, Ken D Olum, and Benjamin Shlaer. A new parallel simulation technique. *Journal of Computational Physics*, 231(1):98–108, 2012.

[BZBP09] Hans-Joachim Bungartz, Stefan Zimmer, Martin Buchholz, and Dirk Pflüger. *Modellbildung und Simulation*. Springer, Berlin, Heidelberg, 2009.

[CBC+10] Eduardo Cuervo, Aruna Balasubramanian, Dae-ki Cho, Alec Wolman, Stefan Saroiu, Ranveer Chandra, and Paramvir Bahl. MAUI: making smartphones last longer with code offload. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*, pages 49–62. ACM, 2010.

[CIM+11] Byung-Gon Chun, Sunghwan Ihm, Petros Maniatis, Mayur Naik, and Ashwin Patti. CloneCloud: elastic execution between mobile device and cloud. In *Proceedings of the sixth conference on Computer systems*, pages 301–314. ACM, 2011.

[Fuj99] Richard M Fujimoto. Parallel and distributed simulation. In *Proceedings of the 31st conference on Winter simulation: Simulation—a bridge to the future-Volume 1*, pages 122–131. ACM, 1999.

[GRA12] Ioana Giurgiu, Oriana Riva, and Gustavo Alonso. Dynamic software deployment from clouds to mobile devices. In *Middleware 2012*, pages 394–414. Springer, 2012.

[HDO11] Bernard Haasdonk, Markus Dihlmann, and Mario Ohlberger. A training set and multiple bases generation approach for parameterized model reduction based on adaptive grids in parameter space. *Mathematical and Computer Modelling of Dynamical Systems*, 17(4):423–442, 2011.

[KPL00] Hyun Goo Kim, VC Patel, and Choung Mook Lee. Numerical simulation of wind flow over hilly terrain. *Journal of wind engineering and industrial aerodynamics*, 87(1):45–60, 2000.

[LDBNR13] Philipp C Leube, Felipe PJ De Barros, Wolfgang Nowak, and Ram Rajagopal. Towards optimal allocation of computer resources: Trade-offs between uncertainty quantification, discretization and model reduction. *Environmental Modelling & Software*, 50:97–107, 2013.

[LML+10] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

[RSM+11] Moo-Ryong Ra, Anmol Sheth, Lily Mummert, Padmanabhan Pillai, David Wetherall, and Ramesh Govindan. Odessa: enabling interactive perception applications on mobile devices. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, pages 43–56. ACM, 2011.

[TAN+12] Narendran Thiagarajan, Gaurav Aggarwal, Angela Nicoara, Dan Boneh, and Jatinder Pal Singh. Who killed my battery?: analyzing mobile browser energy consumption. In *Proceedings of the 21st international conference on World Wide Web*, pages 41–50. ACM, 2012.

Coupling of Existing Simulations using Bottom-up Modeling of Choreographies

Andreas Weiß, Dimka Karastoyanova
Institute of Architecture of
Application Systems (IAAS)
University of Stuttgart
{andreas.weiss, dimka.karastoyanova}
@iaas.uni-stuttgart.de

David Molnar, Siegfried Schmauder
Institute for Materials Testing,
Materials Science and
Strength of Materials (IMWF)
University of Stuttgart
{david.molnar, siegfried.schmauder}
@imwf.uni-stuttgart.de

Abstract: As a contribution for eScience, we discuss the bottom-up derivation of scientific choreography models from existing simulation workflows interconnected as a multi-scale and multi-field simulation. Starting from a motivating scenario of only implicitly coupled simulation workflows for the studying of thermal aging of iron-copper alloys, we present a choreography life cycle supporting the bottom-up derivation of choreography models and the propagation of changes to the underlying simulation workflows in a round-trip manner. Furthermore, we discuss several distinct starting points for the derivation, namely explicitly and implicitly connected simulation workflow models and already running simulation workflow instances.

1 Introduction

In eScience, the main objective is to provide generic approaches and tools to support the whole eScience life cycle [HTT09] and different fields of natural and social sciences for the purpose of faster scientific exploration and discovery. In previous work in the scope of the Cluster of Excellence Simulation Techology (SimTech[1]), the Model-as-you-go approach was introduced [SK10], [SK13]. This approach is based on the workflow paradigm, i.e., simulation steps are modeled as workflow activities and implemented by existing software systems exposing their functionality as services [SHK+11]. Incomplete workflow models can be extended after they have already started in order to conduct the iterative, trial-and-error based modeling of eScience experiments [BG07]. Furthermore, workflow logic can be re-executed, i.e., already executed steps can be compensated and executed again with a different set of parameters as well as re-iterated for convergence of results. However, the existing Model-as-you-go approach and the corresponding implementing scientific workflow management system (sWfMS) only support single scale and single field simulations. Multi-scale and multi-field (so-called multi-*) simulations can only be supported if the orchestrated simulation software is already coupling the different scales/fields on the level of the mathematical formalization. Typically, descriptions of one or more scales/fields to another scale/field are used. Multi-scale simulations cover different scales within the same

[1] SimTech: http://www.iaas.uni-stuttgart.de/forschung/projects/simtech/

computer experiment, where the scales can either refer to time scales, e.g., nanoseconds to days, or to length scales, e.g., nanometers to meters. Multi-field simulations use different scientific fields in the same experiment, e.g., physics, biology, or chemistry. An example is the nucleation, growth and coarsening of second phase particles [MMC+12], also referred to as precipitation. With this approach, multiple time scales of thermal aging as well as length scales in terms of sample volumes become accessible by coupling two simulation methods, the kinetic Monte Carlo (KMC) and the Phase-field Method (PFM), each describing the phenomena of precipitation from a different point of view.

We have observed two basic scenarios that need to be supported: (i) a bottom-up approach where existing software implements different mathematical models and/or scales that need to be orchestrated, very often across organizations. (ii) a top-down approach where one organization needs to realize a particular multi-scale/field simulation and starts modeling and implementation from scratch. In both scenarios the major open issue with respect to modeling is how the interactions among the participating simulations and the data exchange can be represented. In this paper, we focus on the bottom-up aspect of modeling and executing already existing simulation workflows and software. Parts of a multi-* simulation are modeled as workflows, but intermediate results may have to be copied between the simulation software components and subsequent simulation methods may have to be triggered manually. Modeling support of multi-* simulations and automation as a consequence of explicitly modeled control and data flow between simulation software components would decrease simulation duration and increase user friendliness. Furthermore, flexibility such as Model-as-you-go support for multi-* simulations is also needed to cope with the iterative, trial-and-error modeling approach of scientists.

Towards this goal, we contribute a bottom-up view on a life cycle realizing modeling and execution of simulations (Sec. 3) and discuss different cases that occur when deriving an overall model of existing and interacting simulations (Sec. 4). We focus on using proven methods and technologies of the business domain [SK10]. In this work, we make use of *choreographies* to describe the global model of a multi-scale and multi-field simulation. Choreographies are a concept known from the business domain that enables independent organizations to collaborate and reach a common business goal. Choreographies provide a global view on the interconnection of independent organizations communicating without a central coordinator [DKB08]. While choreographies show the public interfaces of the collaboration, these interfaces are implemented by orchestrations, i.e., the so-called enacting workflows, realizing the private business logic of a single organization. The distinct organizations (and their workflows) are called *choreography participants*. To complete this paper, Sec. 5 shows the current state of our implementation supporting the bottom-up life cycle and choreography derivation. We compare our approach with related ones in Sec. 6 and conclude the article with an outline of future research topics in Sec. 7.

2 Motivation

In this section, we introduce a motivating scenario from the domain of material science. Molnar et al. have studied the thermal aging of iron-copper alloys [MMC+12] and emerging effects of existing precipitates on the mechanical behavior [MBHS12], [MBM+14]

of the single crystalline structures by coupling kinetic Monte Carlo (KMC), Molecular Dynamics (MD) and Phase-field Method simulations (PFM) sequentially. Each of these methods is working on a different time scale while, in addition, KMC and PFM are typically applied on different length scales. In this paper, we focus on the sequential coupling of KMC and MD simulations for the case of nano tensile tests of iron-copper alloys at different states of thermal aging [MBHS12], [MBM+14].

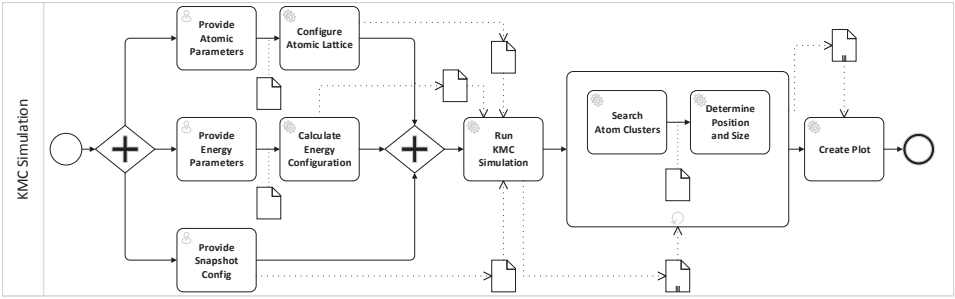


Figure 1: Simplified kinetic Monte Carlo (KMC) simulation workflow. Adapted from [SK13]

Fig. 1 shows a simplified workflow model of a KMC simulation using the custom-made simulation software **O**stwald ripening of **P**recipitates on an **A**tomic **L**attice (OPAL) [Bin06]. The modeling of OPAL as scientific workflow has been documented in [SHK+11]. OPAL simulates the formation of copper precipitates, i.e., atom clusters, due to thermal aging. The workflow receives a set of parameters such as atom concentration, energy values, and the number of intermediate snapshots to be taken from the scientist, configures the atomic lattice and calculates the energy configuration as input for the KMC simulation. According to the desired number of snapshots, the OPAL software saves the current state of the atom lattice at a particular point in time in a snapshot file. Every snapshot is then searched for atom clusters and their position and size. The result is visualized using an external visualization software.

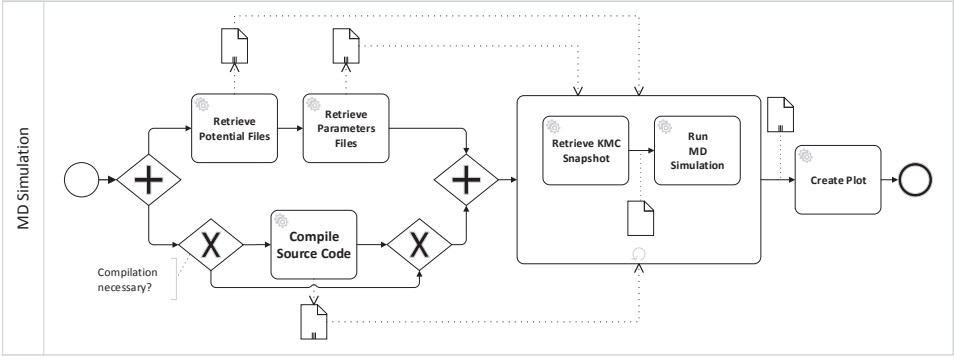


Figure 2: Simplified Molecular Dynamics (MD) simulation workflow

Fig. 2 depicts a simplified workflow of a Molecular Dynamics simulation. The MD simulation, which is implemented by the **I**TAP **M**olecular **D**ynamics (IMD) software pack-

age [SMT97], is used to study the tensile deformation of the snapshots generated in the kinetic Monte Carlo simulation. This simulation is computationally very costly and cannot be done in the KMC simulation due to the rigid lattice. The simulation services encapsulating the IMD software package and used as implementations for the activities in Fig. 2 have been developed in the scope of [Nem14].

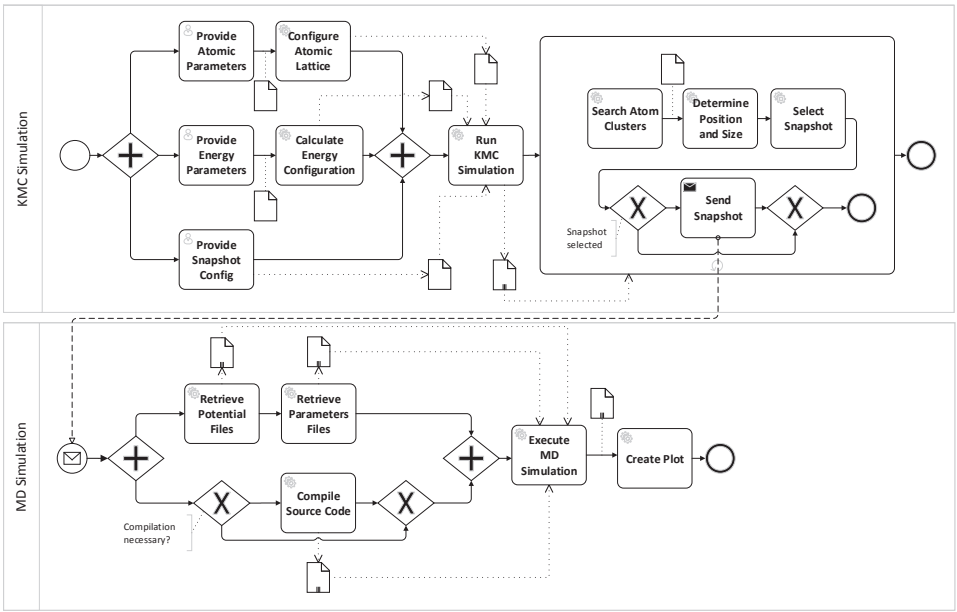


Figure 3: Bottom-up derived choreography forming a multi-scale simulation

In order to provide an adjusted simulation tool, the workflow compiles the source code for a particular computation platform if this has not already been done in a previous run. IMD runs on single cores as well as on computing clusters using the MPI interface[2] standard between computing instances. In parallel, the necessary inter-atomic potential and parameter files containing the simulation's boundary conditions are retrieved for the simulation run. Subsequently, for each selected KMC snapshot, an MD simulation instance is created and the computer-based tensile test is executed. The result is also visualized using an external visualization software.

Fig. 3 shows the target state after coupling both simulations. The coupled workflows form a choreography. Note that the visualization step after the KMC simulation has been removed in order to visualize after the combined multi-scale simulation. Moreover, the activity *Select Snapshot* uses some specific criteria to evaluate if a snapshot should be send to the MD simulation. The *Send Snapshot* activity sends every selected snapshot file or a reference to it to the molecular dynamics simulation. Without the workflow-based coupling the data transfer, the selection of the appropriate snapshot, and the subsequent triggering of the MD simulation for every selected KMC snapshot has to be conducted manually. This

[2] http://www.mcs.anl.gov/research/projects/mpi/

is especially cumbersome and error-prone if the number of generated snapshots is high. An automation of the coupling of both workflows would decrease the overall simulation time and manual errors due to file copying. However, automating data transfer using a implicit choreography is not sufficient to enable a flexible multi-* simulation modeling. We want to derive an explicit choreography model from the interacting workflows which can itself be adapted, e.g., by adding new participants, i.e., simulation methods to the overall multi-* simulation. These changes can then be propagated to the existing workflows to update them. In order to realize Model-as-you-go for choreographies, this has to be possible both with workflow models and already running instances without losing the instance state. Therefore, in Sec. 3 we provide a life-cycle supporting a round-trip modeling beginning bottom-up and in Sec. 4 we discuss different bottom-up derivation cases from existing workflows and to what degree this can be conducted automatically. Note that we do not limit our approach to sequential interaction of simulation methods/participants.

3 Bottom-up Life Cycle

In this section, we show a bottom-up view on a life cycle for modeling and executing multi-* experiments that are realized by choreographies introduced in [WK14]. A scientific Workflow Management system (sWfMS) implementing the life cycle enables the typical trial-and-error style of scientists when modeling scientific experiments [BG07], [SK10]. Since scientists want to be able to react to intermediate results during execution without modeling the experiment completely beforehand, the enactment of choreographies may be started even before the choreography model is completely specified. We call this *Model-as-you-go for choreographies*. The bottom-up life cycle is an extension of the scientific workflow life cycle introduced in [SK10], which itself has been implemented by an extended Business Process Management (BPM) life cycle. While in the traditional BPM life cycle there are several distinct roles, such as business analyst and IT specialist, in the domain of scientific experiments this is typically done by one role, the scientist. In order to ease the handling of the sWfMS, the technical complexities and the difference between workflow models and instances must be hidden, so that the actual phases for modeling on choreography and workflow level, the deployment, the execution, and monitoring are perceived as one experimentation phase by scientists.

Fig. 4 shows the bottom-up modeling approach view on the life cycle as the scientist experiences it. The life cycle starts with the *Workflow Modeling* phase and the modeling of executable workflows (1), which are interconnected. However, the interconnection is not explicitly captured by a choreography model. A graphical workflow editor is used for workflow modeling. The scientist is able to run the workflows and proceed to the *Execution and Monitoring* phase, the deployment is hidden behind the *Run/Resume* action. In this phase, scientists can suspend one or all running workflows using the *Suspend Workflow* action and thus returning to the Workflow Modeling phase in order to adapt them. Furthermore, the *Derive/Update* action can be used to derive a meaningful multi-* choreography model (2) from the interconnected workflows. If the choreography model has already been derived once, it is updated. Note that the derivation can be triggered either in the Execution and Monitoring phase under consideration of the already running workflow

instances and their state or directly from the Workflow Modeling phase only considering the existing workflow models. The derived choreography reflects the error handling, monitoring, and adaptation capabilities of the underlying executable workflows and services. It can be examined and adapted in a graphical choreography editor.

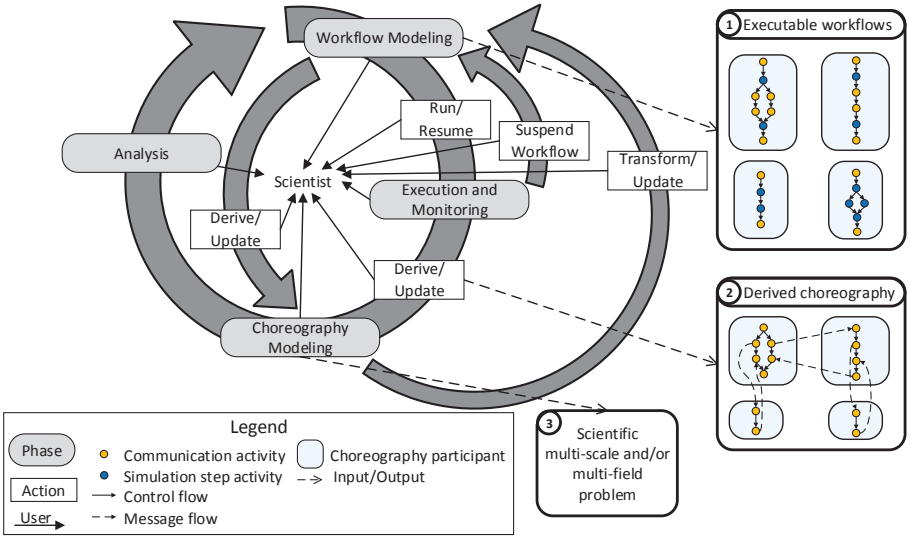


Figure 4: Bottom-up life cycle as perceived by the scientist

Finally, the derived choreography represents a specific scientific multi-scale and/or multi-field problem (3). The adaptation on the choreography level can be used to update the existing executable workflows in a *round-trip* fashion. Via the *Transform/Update* action the underlying workflow models can be modified and at the same time correctness of the changes is enforced. Thus, the scientist returns to the Workflow Modeling phase. The life cycle is concluded by the *Analysis* phase where the experiment results are evaluated.

A scientific Workflow Management system implementing this bottom-up view on the life cycle must have two additional technical phases (Fig. 5). The previously hidden *Deployment* phase realizes the Run/Resume action. In this phase, the executable scientific workflows are deployed onto execution engines and exposed as services that typically require the involvement of a service middleware, too. Similarly, the services that realize the experiment steps are also deployed on their execution environments. The Derive/Update action is technically realized by the *Derivation* phase. It comprises the automatic steps necessary to derive a meaningful choreography from the executable workflows or update the choreography model, respectively. Four adaptation cycles can be observed: In the *Execution and Monitoring* phase, the system is adapted along the functions and the logic dimension. Functions dimensions adaptation comprise the replacement of service implementations whereas the logic dimension can be adapted by refining abstract placeholders in the enacting workflows during run time. The other three adaptation cycles depict the capabilities of adapting the enacting workflows and the choreography along the logic dimension, i.e., the adaptation of the workflow and choreography constructs.

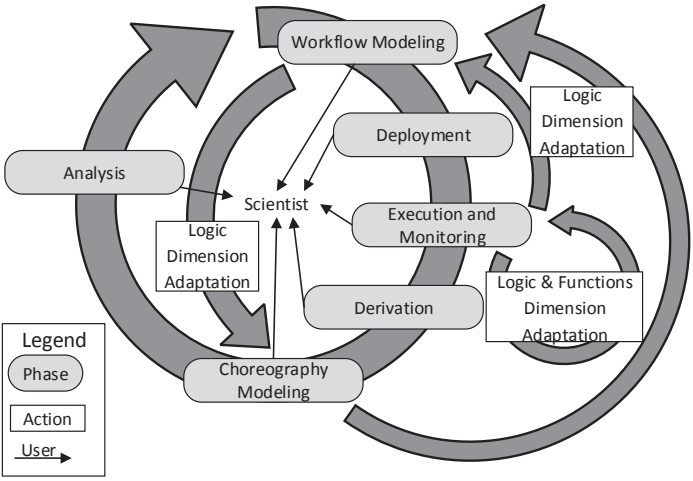


Figure 5: Bottom-up life cycle from the sWfMS perspective

4 Discussion of the Bottom-up Derivation

In this section we discuss the different starting points for a bottom-up derivation of a choreography model from existing simulation workflows. We assume that simulations using a single scale and single field have been modeled as workflows. When these workflows form a multi-* simulation, we distinguish between explicitly and implicitly connected workflows. *Explicitly connected* workflows contain explicitly defined links described in a workflow language. *Implicitly connected* workflows do not contain any explicit specification of a communication link between them. Nevertheless, the workflow models are implicitly connected by some manual or script-based action such as copying of files or invocation of a subsequent workflow. Note that these manual steps could be modeled as manual tasks in the used modeling language, however, we aim for automation when coupling workflows. The benefit of having a derived choreography model from an existing multi-* simulation is that it can be used as a starting point for iterative optimization procedures. Changes to the choreography model can be propagated to the underlying workflows. This helps to explicitly model the connection between implicitly connected workflows or to iteratively incorporate new simulation methods while the multi-* simulation is already running. A key question is whether the choreography derivation can be conducted automatically or needs manual steps and how we can support these manual steps.

Choreography derivation of explicitly connected workflows models: Fig. 6a shows a simplified illustration of explicitly connected workflow models picking up the motivating example from Sec. 2. A selected snapshot (or a reference to it) produced in the KMC simulation is directly sent to an MD simulation instance A possible derivation approach can use explicitly specified communication links between the simulation workflow models to discover the overall choreography. After the derivation the choreography model can be altered using a choreography editor, for example to incorporate a Phase-field Method simulation workflow into the overall simulation.

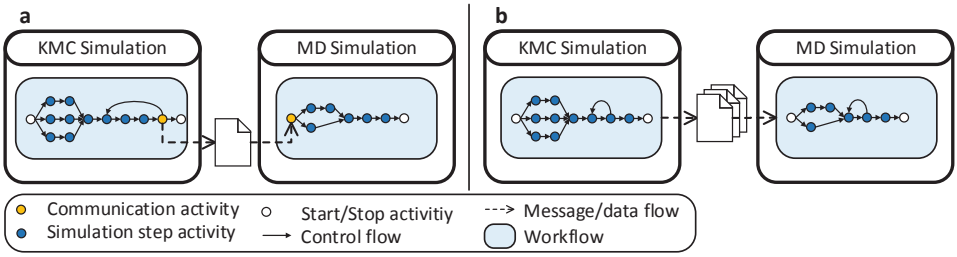


Figure 6: (a) Explicitly connected workflows, (b) implicitly connected workflows

Choreography derivation of implicitly connected workflow models: A more interesting use case for the bottom-up derivation are implicitly connected workflow models. In Fig. 6b, the two simulation workflows are only coupled implicitly. The modeled KMC simulation finishes the calculation for every given time step writing a snapshot to a file. Only when all time steps have been executed, the MD simulation can be started. The transfer of the selected snapshot files and the triggering of the subsequent simulation has to be conducted manually. An automatic derivation of a choreography model is not possible with a derivation approach relying on communication links specified in the workflow modeling language because there are no communication links available in this case. Mining the audit trails of the workflow execution engines is also not an option because the implicit connection will not occur in any execution log. However, modeling support for the definition of explicit interconnections between simulation workflows is still possible. A derivation algorithm has to use the involved workflow models and generate a new choreography participant for each workflow model. A scientist is then able to connect the generated choreography participants by manually drawing a message link in a choreography editor. The changed choreography model can be used to update the existing simulation workflows according to the scientific choreography life cycle presented in Sec. 3.

Choreography derivation of explicitly connected and running simulation workflows: The previous two cases only consider non-running multi-* simulations and the corresponding simulation workflow models. If the simulation workflows have already been started and are long-running because a scientist wants to simulate with a high level of detail, it might not be desirable to abort the simulation workflows in order to derive a choreography model. Rather, the derivation of a choreography model from explicitly connected and running simulation workflows must consider the execution state of the workflow instances. Before the derivation starts, all involved simulation workflows have to be suspended, i.e., the execution has to be paused. Subsequently, a derivation algorithm that recognizes explicitly described communication links between the workflow models can be used to derive a choreography model. Every workflow model is represented by a choreography participant in the choreography model. In contrast to the non-running workflow models before, the state of the paused workflow instance has to be attached to the corresponding choreography participant. If the scientist wants to re-execute, iterate or change the suspended workflow instances and, thus, the derived choreography, the attached instance state information has to be considered (see next paragraph).

Choreography derivation of implicitly connected and running simulations workflows: In the case of only implicitly connected, but already running simulation workflows collaborating as a multi-* simulation, the following approach can be used. The simulation workflows belonging to the same multi-* simulation are suspended. A choreography model is derived from the collaborating workflow models without relying on explicitly modeled communication links in the used workflow language. For every workflow model a choreography participant is generated. The execution state of the corresponding workflow instances is attached to the choreography participant – if several instances of one particular workflow model take part in the multi-* simulation, this also has to be reflected in the derived choreography model. In order to explicitly connect the choreography participants, communication activities have to be inserted on the sending and the receiving participant and a communication link has to be drawn using a graphical choreography editor.

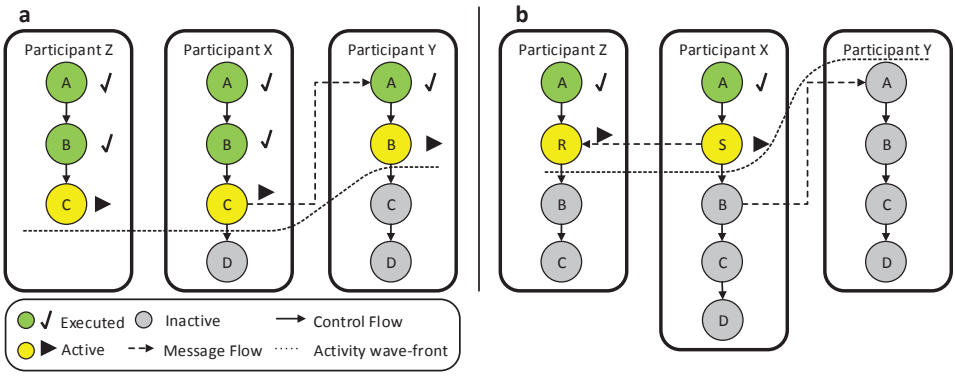


Figure 7: (a) Before the explicit connection of Participant A with Participant C , (b) after the explicit connection of Participant A with Participant C

This is straightforward if the communication link must be inserted ahead of the so-called activity wave-front [SK13], i.e., ahead of the set of already executed or currently active activities. If the insertion of the communication link has to be conducted behind the activity wave-front, i.e., between already executed or currently active activities, it is more complex. The activities after the new communication activity to be inserted may have to be compensated, i.e., semantically undone and re-executed to also consider the behavior of the newly inserted communication activity. The compensation is also necessary for further participants that have been already been invoked. This becomes clear when considering the example in Fig. 7. Fig. 7a shows a simplified scientific choreography with three Participants (X, Y, Z). While Participant X and Participant Y are explicitly connected via a message link from activity C in Participant X to activity A in Participant Y, Participant Z contributes to the overall multi-* simulation but is not explicitly connected to any other choreography participant. If the decision is made to explicitly incorporate Participant Z into the scientific choreography without aborting the already started multi-* simulation, this leads to the situation depicted in Fig. 7b. All activities following the new sending activity S and receiving activity R have to be compensated and old data snapshots [SK13] have to be loaded. Then, a message link can be inserted between the activities S and R and the changes can be propagated from the choreography model to the enacting workflow

instances. A detailed specification of the different compensation cases and the propagation to the workflow instances is out of scope in this paper but will be investigated in future.

5 Realization

In the following, we will discuss the realization options of a system supporting the life cycle and the derivation cases. As workflow language for simulation workflows, we use BPEL [OAS07], which has also been used in previous work in the scope of single scale and single field simulation workflows [SK13]. BPEL4Chor [DKLW07] is the choreography language, we have chosen to represent our bottom-up derived choreography models. BPEL4Chor is a non-executable choreography language forming an additional layer on top of the BPEL standard. In BPEL4Chor, the communication behavior is specified by the so-called Participant Behavior Descriptions (PBD), which are abstract, but more restricted BPEL processes. In the Participant Topology artifact, the structural aspects of a choreography such as the participants and the message links between them are specified. Simulation workflows can be modeled using the *Mayflower BPEL Designer* [SHK12], for choreography models we use our graphical choreography editor, the *ChorDesigner* [Sch14]. The derivation algorithms will be implemented in a *Transformer* component connecting both editors. The Transformer contains functions for both deriving choreography models and updating the underlying workflow models to support the modeling round-trip. The approach of Steinmetz [Ste07] can be applied in order to derive a BPEL4Chor choreography from explicitly connected simulation workflow models. The approach generates a participant for every workflow model and follows the declared BPEL Partner Links to derive the communication links between the participants. However, the approach is dependent on explicit communication links between workflows in order to derive a choreography model, thus, it is not directly suitable in the case of implicitly connected workflow models. In this case, we propose to generate choreography participants from the given BPEL workflows without relying on BPEL Partner Links. The BPEL workflows can be easily transformed into BPEL4Chor PBDs by turning them into abstract BPEL processes and removing Partner Links, Port Types and Operations as demanded by the Abstract Process Profile for Participant Behavior Descriptions [DKLW07]. For the derivation of a BPEL4Chor choreography model from running implicitly and explicitly connected simulation workflows, the derivation approach of Steinmetz has to be extended to attach the workflow instance state to the derived choreography participant. The Model-as-you-go operations for choreographies such as re-execution, iteration and iterative modeling of already started multi-* simulation workflow instances will be implemented in future.

6 Related Work

In [TKOAM08], a bottom-up approach for the interconnection of existing workflows to form choreographies spanning short-term virtual enterprises is presented. In order to preserve privacy for the participants, the approach only exposes the parts of existing workflows that are necessary for communicating with other participants using formal methods. Furthermore, the abstracted workflows are advertised and matched with potential partner workflows. In contrast, in our approach we do not conduct a semantic matching between

simulation workflows to find an complementary participant, since the simulation workflow may not be explicitly connectible when deriving a choreography model. Additionally, only exposing the external visible communication behavior is too strict for the scenario of simulation workflows when scientists want to perform iterative modeling in the choreography modeling phase. In [ea08], the authors introduce MUSE, a software for multi-scale and multi-physics simulations in the scope of astrophysics. Software modules implementing distinct physical methods on distinct scales are orchestrated using a Python layer. While this is a significant improvement compared to completely monolithic approaches, our approach is usable for different scientific fields and is based on simulation workflows, which also specify distinct steps inside the scope of one scale or field. Furthermore, we support the derivation of a global model in order to support the iterative addition of further simulation methods with respect to scales and/or scientific fields. In [DKB08], a generic choreography life cycle is presented. While the proposed life cycle supports choreography modeling beginning with a domain specific problem, i.e., top-down modeling, the life cycle presented in our work starts bottom-up with executable workflows from which a choreography model is derived.

7 Conclusions and Future Work

In this paper, we discussed the bottom-up derivation of scientific choreography models from existing simulation workflows interconnected as a multi-scale and multi-field simulation. Starting from a motivating scenario of only implicitly coupled simulation workflows for the studying of thermal aging of iron-copper alloys, we presented a choreography life cycle supporting the bottom-up derivation of choreography models and the propagation of changes to the underlying simulation workflows in a round-trip manner. Furthermore, we discussed several distinct starting points for the derivation, namely explicitly and implicitly connected simulation workflow models and already running simulation workflow instances. While we already have an choreography editor for adapting choreography models, in our future research we plan to implement the discussed derivation cases based on the choreography language BPEL4Chor as well as to develop concepts for the Model-as-you-go operations for choreographies.

Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

References

[BG07] R. Barga and D. Gannon. Scientific versus Business Workflows. In *Workflows for e-Science*, pages 9–16. Springer, 2007.

[Bin06] P. Binkele. *Atomistische Modellierung und Computersimulation der Ostwald-Reifung von Ausscheidungen beim Einsatz von kupferhaltigen Stählen*. PhD thesis, University

of Stuttgart, Stuttgart, 2006.

[DKB08] G. Decker, O. Kopp, and A. Barros. An Introduction to Service Choreographies. *Information Technology*, 50(2):122–127, 2008.

[DKLW07] G. Decker, O. Kopp, F. Leymann, and M. Weske. BPEL4Chor: Extending BPEL for Modeling Choreographies. In *ICWS'07*. IEEE, 2007.

[ea08] S. Portegies Zwart et al. A Multiphysics and Multiscale Software Environment for Modeling Astrophysical Systems. In *ICCS'08*, pages 207–216. Springer, 2008.

[HTT09] T. Hey, S. Tansley, and K. Tolle, editors. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, 2009.

[MBHS12] D. Molnar, P. Binkele, S. Hocker, and S. Schmauder. Atomistic multiscale simulations on the anisotropic tensile behaviour of copper-alloyed α-iron at different states of thermal ageing. *Philosophical Magazine*, 92(5):586–607, 2012.

[MBM+14] D. Molnar, P. Binkele, A Mora, R. Mukherjee, B. Nestler, and S. Schmauder. Molecular Dynamics virtual testing of thermally aged Fe-Cu microstructures obtained from multiscale simulations. *Computational Materials Science*, 81(0):466 – 470, 2014.

[MMC+12] D. Molnar, R. Mukherjee, A. Choudhury, A. Mora, P. Binkele, M. Selzer, B. Nestler, and S. Schmauder. Multiscale simulations on the coarsening of Cu-rich precipitates in a-Fe using kinetic Monte Carlo, molecular dynamics and phase-field simulations. *Acta Materialia*, 60(20):6961–6971, 2012.

[Nem14] Markus Nemet. Kapselung von bestehenden Simulationsanwendungen mit Hilfe von Web Services. Bachelor thesis, 2014.

[OAS07] OASIS. Web Services Business Process Execution Language Version 2.0, April 2007.

[Sch14] J. Schilling. Analyse und Erweiterung eines bestehenden Choreographiewerkzeugs. Student Thesis No. 2433, University of Stuttgart, 2014.

[SHK+11] M. Sonntag, S. Hotta, D. Karastoyanova, D. Molnar, and S. Schmauder. Using Services and Service Compositions to Enable the Distributed Execution of Legacy Simulation Applications. In *ServiceWave'11*, pages 1–12. Springer, 2011.

[SHK12] M. Sonntag, M. Hahn, and D. Karastoyanova. Mayflower - Explorative Modeling of Scientific Workflows with BPEL. In *CEUR Workshop'12*, pages 1–5. Springer, 2012.

[SK10] M. Sonntag and D. Karastoyanova. Next Generation Interactive Scientific Experimenting Based On The Workflow Technology. In *MS'10*. ACTA Press, 2010.

[SK13] M. Sonntag and D. Karastoyanova. Model-as-you-go: An Approach for an Advanced Infrastructure for Scientific Workflows. *Grid Computing*, 11(3):553–583, 2013.

[SMT97] J. Stadler, R. Mikulla, and H. R. Trebin. IMD: A software package for molecular dynamics studies on parallel computers. *International Journal of Modern Physics C*, 8(5):1131–1140, 1997.

[Ste07] T. Steinmetz. Generierung einer BPEL4Chor-Beschreibung aus BPEL-Prozessen. Student Thesis No. 2101, University of Stuttgart, 2007.

[TKOAM08] S. Tata, K. Klai, and N. Ould Ahmed M'Bareck. CoopFlow: A Bottom-Up Approach to Workflow Cooperation for Short-Term Virtual Enterprises. *Services Computing*, 1(4):214–228, 2008.

[WK14] A. Weiß and D. Karastoyanova. A Life Cycle for Coupled Multi-Scale, Multi-Field Experiments Realized through Choreographies. In *EDOC'14*, pages 1–8. IEEE, 2014.

Coupling Algorithms for Partitioned Multi-Physics Simulations

Benjamin Uekermann*, Bernhard Gatzhammer*, Miriam Mehl†

* Department of Computer Science, Technische Universität München
Boltzmannstraße 3
85748 Garching, Germany
uekerman@in.tum.de

† Institute for Parallel and Distributed Systems, Universität Stuttgart
Universitätsstraße 38
70569 Stuttgart, Germany
miriam.mehl@ipvs.uni-stuttgart.de

Abstract: Partitioned coupling approaches are an important tool in order to achieve a decent time-to-solution for multi-physics problems with more than two physical fields or changing combinations of fields. We study different approaches to deduce coupling schemes for partitioned multi-physics scenarios, by means of a simple, but yet challenging fluid-structure-fluid model problem. To our knowledge, this is the first time that a fully implicit black-box coupling scheme for partitioned multi-physics scenarios is described. This allows the simulation of a new range of applications in a partitioned way.

1 Introduction

The simulation of the interaction between three or more physical fields, so-called multi-physics scenarios, is a crucial part of many engineering or biomedical applications. This includes, for example, fluid-structure-acoustic interaction (e.g. [LKBD09]), where the acoustic field can be seperated in a complex near field and a simplified far field, or simulations of the heart, where electro-magnetism, fluid flow, and structure equations are coupled ([LAVH12]). Also scenarios with multiple flow fields such as partially filled tanks on container ships ([GM08]) belong to this class. If the coupling is restricted to a lower-dimensional manifold in the computational-domain, which is the case in the mentioned examples, we speak of surface-coupled problems. Such applications benefit clearly from a partitioned approach where the physical domain is split into single-physics fields, all fields are simulated with their own solver, and these solvers are coupled via common interfaces. Since the coupling is restricted to a surface (instead of the whole volume of the domain), the amount of data to be transferred between the solvers is comparable to the amount of data transferred in a distributed memory parallelization of the single fields. The splitting into single fields gives us the possibility to reuse existing software which is

crucial to maintain a decent overall time-to-solution. Furthermore, multi-physics scenarios often adhere different time and spatial scales, making different spatial and temporal discretizations of each physical field necessary. However, the splitting into single fields introduces possible stability issues that need to be tackled by sophisticated coupling algorithms. The counterpart of partitioned methods, the monolithic approach tries to set up one large system of equations for all fields, which is solved within one program. This approach implies an unfeasible complexity, if multiple physical fields are considered that all have to be reimplemented for each new combination of fields.

Depending on the type of interaction between the single physical fields, the task of finding a stable and accurate coupling method can be easy or rather difficult. We speak of a uni-directional interaction if the mutual influence between two physical fields P_i and P_j is only one-sided, e.g., P_i influences P_j but not vice versa. Such problems can be handled easily and efficiently even with a file-based one-time data transfer between the single field solvers and is, therefore, not considered in this paper. We focus on bi-directional interactions where both fields have an impact on each other. Depending on the strength of the bi-directional interaction, more or less sophisticated methods are required for the numerical coupling of the fields.

Whereas many coupling algorithms for systems involving two surface-coupled physical fields have been developed in the last decade, there are hardly any general and robust methods available for the coupling of multiple fields. For the two-field coupling which we refer to as bi-coupling schemes in the following, a variety of methods from explicit schemes, executing only a fixed number of single field solves per multi-field time step [FL00], to stable implicit schemes, iterating between the single field solvers until the time step equation converges to the monolithic solution, are known from literature. Explicit coupling schemes are known to yield stable time-stepping only in cases with a relatively weak bi-directional coupling. Implicit schemes are particularly necessary for fluid-structure interaction (FSI) with incompressible fluids, a typical case for a strong bi-directional interaction due to the so-called added mass effect (e.g. [CGN05, Bru09]). Established methods are Aitken underrelaxation ([KW08]), or quasi-Newton methods ([DBV09, VLDV07, Mic04, MY10]). Usually, these methods execute the single field solvers in a staggered way, i.e., one after the other which limits the scalability of such a simulation on massively parallel systems. As massively parallel computations are necessary for multi-physics scenarios, since only a high resolution of all fields allows to take advantage of the more complex modeling of multi-physics scenarios compared to single-physics scenarios, we developed implicit coupling algorithms executing the involved single field solvers in parallel to each other ([UBGM13, MUB$^+$13]).

On the other hand, not many multi-coupling algorithms have been developed. In [SBE$^+$14], a multi-coupling algorithm that needs full Jacobian information (non-black-box) is described and tested for lower dimensional problems. The necessity of full Jacobian information makes it very cumbersome to integrate this algorithm in 2D or 3D scenarios already in the case of non-black-box solvers.

In this work, we develop and discuss multi-coupling algorithms that are derived from either a simple composition of bi-coupling schemes or a generalization of the underlying ideas to a true multi-coupling. To our knowledge, this is the first time, that a fully-implicit black-

box multi-coupling algorithm is described that allows to simulate scenarios in a partitioned way. The results in Section 5 show that the easiest approach, i.e., the composition of bi-coupling schemes, is not sufficient for already moderately difficult problems.

All coupling methods used in Section 5 are implemented in `preCICE`, a coupling library ([Gat14]) developed in our group, which offers, besides different coupling algorithms, also methods for mapping data between non-matching meshes and communication routines. The application interface of `preCICE` is formulated on a very high level, allowing for a minimally invasive inclusion in any single-physics code. Once a solver is adapted to `preCICE`, different other solvers can be coupled in a nearly plug-and-play manner. Different solvers including commercial tools such as `Fluent` or `COMSOL`, open source tools such as `OpenFOAM`, and in-house codes (e.g. [UCG+14]), have been coupled successfully to `preCICE`. `preCICE` itself is open-source software[1]. For sake of brevity, implementation details of `preCICE` are left out in this work. The interested reader is referred to [Gat14].

The remainder of this work is organized as follows: Section 2 introduces a model problem containing three different physical fields. We use this model problem throughout this work to study multi-coupling algorithms. Section 3 discusses multi-coupling algorithms derived from a composition of bi-coupling schemes, whereas Section 4 studies multi-coupling methods based on a generalization of the underlying ideas of selected bi-coupling methods. Section 5 presents numerical results for our model problem and the derived algorithms. Section 6 concludes this work with some preliminary guidelines how to choose the right multi-coupling scheme for a particular scenario.

2 A Fluid-Structure-Fluid Model Problem

To include also commercial tools with a closed application interface, i.e., black-box solvers, we describe a solver P_i simply by an input-output relation

$$P_i : x_{\text{in}} \mapsto x_{\text{out}} .$$

Here, $x_{\text{in}} \in \mathbb{R}^N$ denotes a vector of nodal input values at the discretized coupling surface between P_i and other physical fields. P_i maps this input to output values $x_{\text{out}} \in \mathbb{R}^N$ by calculating one timestep or one iteration. A fluid solver in a so-called Dirichlet-Neumann coupling, for example, gets displacement values from (possible multiple) structure solvers at a common surface, calculates new pressure and velocity values, and finally gives back force values to the structure solvers at the same coupling surface. Note that black-box solvers, in particular, do not offer information concerning the Jacobian of P_i.

In general, we can model surfaced-coupled multi-physics scenarios as directed graphs where each vertex corresponds to a physical solver P_i and each edge from solver P_i to P_j means that an output variable x_k of P_i is an input variable of P_j. For the sake of clarity, we only consider a model problem comprising three solvers, two fluid solvers F_1 and F_2 and one structural solver S solving a scenario with two (possibly different) fluids separated by

[1] `http://www5.in.tum.de/wiki/index.php/PreCICE_Webpage`

an elastic structure. As the two fluid solvers do not have a direct interaction with each other, this model scenario corresponds to the graph shown in Figure 1.

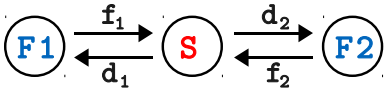


Figure 1: Multi-physics model problem, represented in a dependency graph. Two fluid solvers F_1 and F_2 compute force values f_1 and f_2, acting as input values for the structural solver S, who itself gives back displacement values d_1 and d_2 to the fluid solvers.

This model problem contains already many basic issues of general multi-physics scenarios. Therefore, we use it to study various multi-coupling approaches without the formalism needed to describe general multi-physics problems. We want to stress that all techniques developed in this work can be generalized in a straight-forward way to problems comprising more than three physical solvers. Also the implementation in `preCICE` allows for more complicated scenarios.

The densities ρ_{F1}, ρ_{F2}, and ρ_S in all three physical domains and the geometrical shape of the structural domain determine the strength of the three-field interaction, and, thus, have a crucial impact on the suitability and performance of coupling schemes. We use this fact to study different setups with strong or weak interaction between all three fields as well as strong interaction between two fields and weak interaction with the third field in Section 5. We introduce a variety of coupling approaches and discuss their theoretical applicability in the following two sections.

3 Composition of Bi-Coupling Schemes

This section discusses methods for multi-coupling that can be easily derived (and implemented) as a combination of existing bi-coupling methods. There are basically two ways to combine bi-couplings: concatenation (Section 3.1) and inclusion (Section 3.2).

3.1 Concatenation of Bi-Coupling Schemes

The straightforward idea to couple multiple physical solvers P_i is to use a bi-coupling scheme for each edge in the multi-physics graph as displayed in Figure 2 for our model problem from Section 2. In this setting, each bi-coupling scheme can be adjusted to meet particular needs. In particular, we can choose whether an explicit or an implicit scheme is required for each bi-coupling separately. Figure 3 shows that different choices lead to different execution orders of the physical solvers. All combinations sketched in Figure 3 are supported in `preCICE` (cf. [Gat14]). If two implicit schemes are used, they do, in general, not convergence at the same time. In this case, the early converged coupling scheme waits until convergence of all schemes.

Sophisticated implicit coupling schemes such as Aitken underrelaxation or quasi-Newton methods reuse information from previous iterations to estimate the response of the cou-

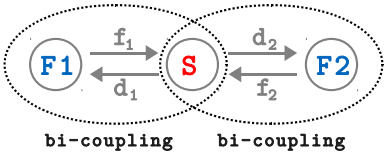


Figure 2: Schematic view of the three-field coupling for the model problem from Section 2 using a concatenation of two bi-coupling methods, one for each edge in the interaction graph.

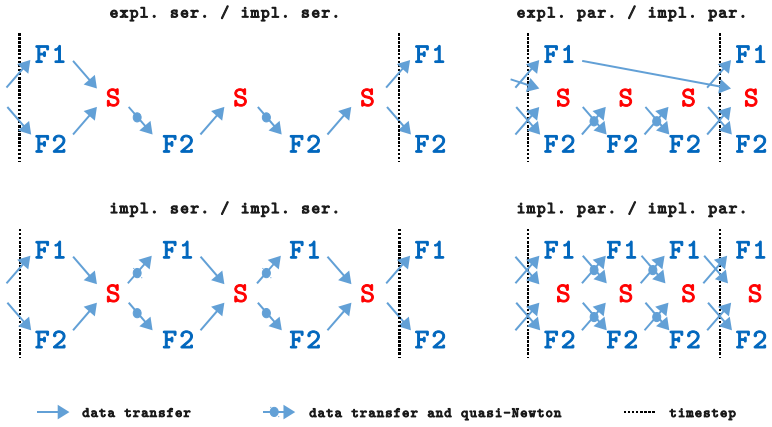


Figure 3: Execution orders for concatenation of different bi-coupling scheme. Whether the coupling scheme is explicit or implicit, serial or parallel, the execution order differs. The selection of sketches is restricted to reasonable and non-trivial combinations. For sake of simplicity, implicit coupling schemes are assumed to converge in 3 iterations.

pling partner, to choose the best coupling parameters. If, in our model problem, both coupling schemes are implicit, they hamper each other, since their adaptive choice of coupling parameters changes the response mechanism of the two coupled solvers seen from the third solver. In mathematical terms, this means that the Jacobian of both two-field subsystems in the three-field system changes substantially throughout the iterations which makes it almost impossible for the third solver to get any useful information on the response of such a subsystem. This is especially a problem if both fluid solvers F_1 and F_2 are indirectly strongly coupled which is the case for a very thin, elastic or lightweight structure. The impact of the described problem, depends, thus, on the geometrical shape and the stiffness of S. Numerical experiments in Section 5 confirm this conclusion. Obviously, this problem reappears in general multi-physics problems, if multiple implicit bi-coupling schemes influence each other. A simple concatenation of bi-coupling schemes does, in general, not result in a stable overall coupling. In Sections 3.2 and 4, we present possible remedies.

3.2 Inclusion of Bi-Coupling Schemes

The alternative to a simple concatenation of bi-coupling schemes is the inclusion of bi-coupling schemes. Here, two physical solvers coupled by a bi-coupling scheme are re-

garded as an entity from the outside and are coupled with the third solver using a bi-coupling approach again. Figure 4 shows a variant of this approach for our model problem: F_1 and S are bi-coupled and the converged entity $(F_1 S)$ is bi-coupled with F_2. Different from the methods in Section 3.1, this implies, that we have induced a nesting of iterations: In the inner iteration, we iterate between F_1 and S until convergence to the two-field monolithic solution. The result is then transferred to F_2 which returns its result to S. This outer coupling is then repeated until convergence to the three-field system. If we assume that M iterations are required for both the inner and the outer iteration, a total of M^2 solves is required for F_1 and S, whereas M solves are needed for F_2. Only if we apply an explicit scheme to the outside coupling, the overall scheme does not differ from a concatenation scheme.

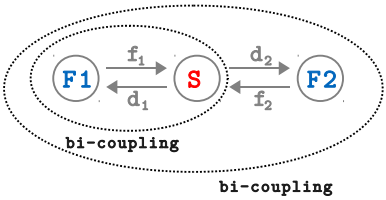


Figure 4: Schematic view of the three-field coupling for the model problem from Section 2 using an inclusion of two bi-coupling methods. Two coupled physical solvers are regarded as an entity from the outside and are coupled with the third solver. If two implicit schemes are applied, this results in two nested iteration loops.

We suppose that the inclusion of two implicit schemes always results in an overall stable coupling if we choose appropriate methods for each of the bi-couplings. The disadvantage of a multiple inclusion are its computational costs: the number of solver executions grows exponentially with the number of physical solvers if we assume that each coupling schemes needs a constant amount of iterations. This is the reason, why we did not implement such schemes in `preCICE`. In the next section, we develop a far more efficient and stable multi-coupling scheme.

4 Multi-Coupling Schemes

In this section, we generalize bi-coupling schemes to proper multi-coupling schemes by directly looking at the coupling of multiple solvers as depicted in Figure 5.

The basis of our coupling approach is the following observation: Each multi-physics problem can be reformulated as a fixed-point equation (FPE). This is widely used in classical partitioned FSI problems, where many coupling algorithms are derived from the FPE

$$(S \circ F)(d) = d\,. \tag{1}$$

S and F denote the structure and the flow solver, respectively, d is the displacement (or the velocity) of the structure surface. Below, f is the force exerted on the structure surface

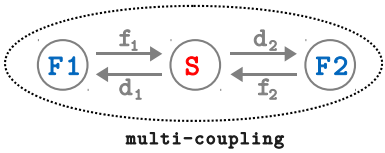


Figure 5: Schematic view of the three-field coupling for the model problem from Section 2 using an multi-coupling approach, i.e. considering all interactions in an overall coupling system.

by the fluid. We showed in [UBGM13, MUB+13], that the same solution is achieved from the alternative FPE in matrix-like notation

$$\begin{pmatrix} 0 & F \\ S & 0 \end{pmatrix} \begin{pmatrix} f \\ d \end{pmatrix} = \begin{pmatrix} f \\ d \end{pmatrix} , \tag{2}$$

which we refer to as the vectorial system. This has the advantage that the fluid and structure solver can be executed in parallel to each other leading to a better parallel efficiency. To generalize this idea, we use the more compact notation

$$(S, F)\,(f, d) = (f, d) \ .$$

This notation stresses the fact, that we consider all input variables, and also all output variables as a closed entity. However, the notation does not specify any longer, which physical solver uses which variables as input and output. We can easily generalize this idea to more complex multi-physics problems. Basically, we calculate one iteration of the complete multi-physics dependency graph, taking all coupling variables as input values, computing all physical solvers P_i once, and considering all new calculated coupling variables as output values. For our model problem introduced in Section 2, the associated FPE reads

$$(F1, S, F2)\,(f_1, d_1, f_2, d_2) = (f_1, d_1, f_2, d_2) \ .$$

This FPE can now be solved with any FPE solver. We showed in [MUB+13] that simple FPE solvers such as a fixed-point iteration or an Aitken underrelexation for classical FSI problems lead to a two times slower convergence when using (2) compared to (1). Sophisticated quasi-Newton FPE solvers, based on the solution of a least-squares system in every iteration, however, show only a slight degradation of the convergence rate. Such a quasi-Newton least-squares (QNLS) solver (cf. [HDHV09]) was applied for (1) in [DBV09] based on similar ideas in [Mic04, VLDV07] and for (2) in [UBGM13, MUB+13]. Algorithm 1 describes the QNLS technique for a general fixed-point equation $H(x) = x$. In a transient setting, the reuse of iteration values from previous time steps can lead to a far better efficiency (cf. [DBV09, MUB+13]). We write QNLS(n) for a quasi-Newton solver reusing information from n time steps. If this algorithm is applied for the complete multi-physics problem, we refer to the coupling scheme as the "multi interface quasi-Newton" scheme (MIQN(n)).

The fixed-point x that Algorithm 1 seeks for consists, in our model problem, of four different subvectors: $x = (f_1, d_1, f_2, d_2)$. The solution of the least-squares system in QNLS

Algorithm 1 Quasi-Newton least squares method in pseudocode (cf. [DBV09, HDHV09])

initial value x^0
$\tilde{x}^0 = H(x^0)$ and $R^0 = \tilde{x}^0 - x^0$
$x^1 = x^0 + 0.1 \cdot R^0$
for $k = 1 \ldots$ **do**
  $\tilde{x}^k = H(x^k)$ and $R^k = \tilde{x}^k - x^k$
  if $\|R^k\|/\|x^k\| < \epsilon$: **break**
  $V^k = [\Delta R_0^k, \ldots, \Delta R_{k-1}^k]$ with $\Delta R_i^k = R^i - R^k$
  $W_k = [\Delta \tilde{x}_0^k, \ldots, \Delta \tilde{x}_{k-1}^k]$ with $\Delta \tilde{x}_i^k = \tilde{x}^i - \tilde{x}^k$
  decompose $V^k = Q^k U^k$
  solve the first k lines of $U^k \alpha = -{Q^k}^T R^k$
  $\Delta \tilde{x} = W \alpha$
  $x^{k+1} = \tilde{x}^k + \Delta \tilde{x}^k$
end for

(Algorithm 1 uses a Q-R decomposition), depends highly on the mutual scaling of those subvectors. In classical FSI applications, force vectors often have substantially larger values than the displacements, in particular if displacement increments are used instead of absolute displacements relative to the initial state. In order to balance the influence of all subvectors we, thus, have to apply a normalization. In our studies, we found that a static scaling of the force vectors results in a sufficiently good balance. The potential of further improvements if using a dynamical scaling is subject of further research.

5 Numerical Results

5.1 Experiment settings

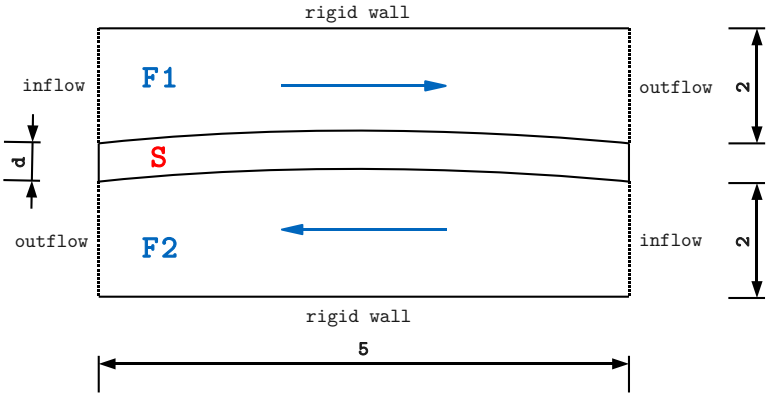


Figure 6: Scenario sketch: 2 fluid domains F_1 and F_2 are separated by an elastic beam S, which is fixed on both ends. Two variants of this scenario are studied, which differ in the vertical extent of the beam, $d_1 = 0.2$ or $d_2 = 1$ m.

Figure 6 depicts the geometry of the scenario that we study numerically. Two fluid domains F_1 and F_2 are separated by an elastic beam S, which is fixed on both ends. We distinguish two variants, which differ in the vertical extent of the beam: one thin variant S_1, $d_1 = 0.2$, and one thick variant S_2, $d_2 = 1.0$. The thin variant induces a stronger indirect interaction between the two fluid fields than the thick variant. The structure is modeled by a St. Venant-Kirchoff cantilever with Young's modulus $E = 5.6 \cdot 10^6$ kg/ms^2 and Poisson-ratio 0.4 . The density of the structure varies from 10^2 to 10^3 kg/m^3. Both fluids are modeled by the incompressible Navier-Stokes equations, with dynamic viscosity $\mu = 10^2$ kg/ms. The inflows show a parabolic velocity profile with mean velocity $U = 1.0$ m/s. The densities of both fluids, ρ_{F1} and ρ_{F2}, vary separately in the range from 10^0 to 10^3 kg/m^3, leading to either weak or strong interactions between both fluids and the structure.

We use `Alya`[2] as the physical solver for all three domains (cf. [UCG+14, HVAC09]). There is, however, no particular need for this choice. Solvers can be exchanged in a nearly plug-and-play manner to meet particular needs. We showed, for example, in [MUB+13] the bi-coupling of `COMSOL`, `Fluent`, or `OpenFOAM`. Both fluid domains are discretized with approximately 4000 elements each, whereas the structure consists of approximately 2250 (S_1), respectively 600 (S_2) elements. We use a matching grid at both interfaces. `Alya` uses finite elements for both, structure und fluids. The moving structure is resolved in the fluid by an arbitrary Lagrangian-Eulerian perspective, involving the movement of the underlying grid. For further numerical details of `Alya`, we want to refer to [UCG+14]. The timestep size is set to $dt = 10^{-3}$ s. To decide on the overall convergence, we use relative convergence criteria of 10^{-3} for all coupling variables, displacement increments and forces on both interfaces. Figure 7 shows the results for the thin structure at $t = 10$ s

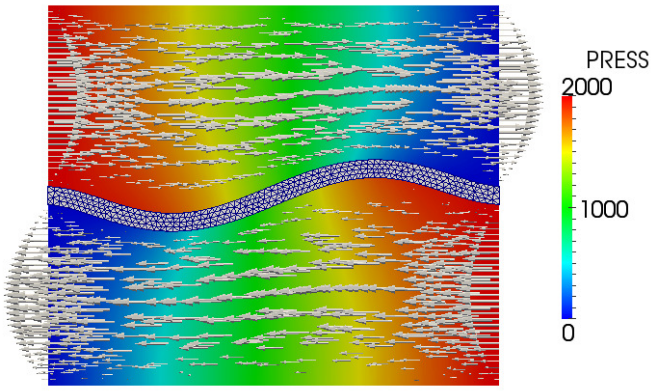


Figure 7: Pressure values (N/m^2), velocity vectors and grid deformation for the fluid-structure-fluid model problem at $t = 1.0$. This run uses the thin structure S_1, the densities $\rho_{F1} = \rho_{F2} = 10^3$ kg/m^3, $\rho_S = 10^1$ kg/m^3, and the MIQN(8) coupling scheme. For sake of clarity, the structural displacements are scaled by a factor of 5.

[2] `Alya` is developed at the Barcelona Supercomputing Center,
cf. `http://www.bsc.es/computer-applications/alya-system`.

5.2 Discussion

This subsection presents and discusses the results for the test case, described above. We vary the density of all three physical fields and the thickness of the structure. From literature (e.g. [CGN05, Bru09]), we know that the interaction between fluid and structure gets stronger, and thus the coupling gets more elaborate, if the density ratio ρ_F/ρ_S raises. Furthermore, by lowering the thickness of the structure, we can raise the indirect interaction between both fluid fields. Table 1 shows the average iteration numbers per timestep, listed for various density ratios, for both structures, and for the different coupling schemes, we introduced in Sections 3 and 4. By "IQN(5) / IQN(5)", we denote a concatenation of two interface quasi-Newton schemes (developed in [DBV09]), with re-used information from 5 timesteps, whereas CSS refers to the classical serial staggered scheme (the trivial explicit scheme, discussed in [FL00]). MIQN is the multi quasi-Newton coupling scheme, introduced in Section 4. Our experiences in [MUB+13] determines the choice of reused timesteps. For the MIQN scheme, we scale all force vectors by a factor of 10^{-7} (compare the discussion in Section 4).

ρ_{F1} - ρ_S - ρ_{F2}	Coupling	S_1	S_2
10^0 - 10^3 - 10^3	CSS / IQN(5)	4.62	4.62
10^0 - 10^3 - 10^3	CSS / IQN(0)	9.04	9.24
10^1 - 10^3 - 10^3	CSS / IQN(5)	crash	crash
10^1 - 10^3 - 10^3	IQN(5) / IQN(5)	crash	3.08 / 4.56
10^2 - 10^3 - 10^3	IQN(5) / IQN(5)	crash	crash
10^3 - 10^3 - 10^3	IQN(5) / IQN(5)	crash	5.10 / 5.18
10^3 - 10^2 - 10^3	IQN(5) / IQN(5)	crash	crash
10^3 - 10^1 - 10^3	IQN(5) / IQN(5)	crash	crash
10^1 - 10^3 - 10^3	MIQN(8)	5.64	5.62
10^2 - 10^3 - 10^3	MIQN(8)	5.88	5.76
10^3 - 10^3 - 10^3	MIQN(8)	7.40	7.00
10^3 - 10^2 - 10^3	MIQN(8)	11.78	11.72
10^3 - 10^1 - 10^3	MIQN(8)	20.90	17.20

Table 1: Average iteration numbers needed for convergence for various coupling configurations and different density ratios (all values in $\mathrm{kg/m^3}$). S_1 and S_2 refer to the thin respectively thick structure. The values are averaged over the first 50 timesteps. The upper part shows results for the concatenation of an explicit and an implicit scheme, the middle part for the concatenation of two implicit schemes (cf. Section 3.1), and the lower part for the newly developed multi-coupling scheme (cf. Section 4).

In Table 1, we can observe several aspects. The upper part of the table shows results for a weak $F_1 - S$ interaction and a strong $F_2 - S$ interaction. For this case, we concatenate an explicit with an implicit scheme. We can argue from the results that, in this case, the strong interaction dominates the overall system. This means that the implicit coupling is not hampered by the explicit scheme. We can observe this in the better convergence of CSS / IQN(5) compared to CSS / IQN(0). If the density of F_1 is raised to 10^1 $\mathrm{kg/m^3}$, the interaction between F_1 and S also becomes a strong one, leading to divergence if only an

explicit scheme is used. In general, the behavior of the overall system, in this case, is very similar to the bi-coupling of F_2 and S itself. The thickness of the structure, in particular, has no important influence here.

The middle block of Table 1 shows results for two strong interactions, demanding an implicit scheme on both interfaces. The concatenation of two implicit schemes, however, is not successful, since both interface systems hamper each other (compare the explanation in Section 3.1). It is striking that this problem, especially influences the test case with the thin structure S_1, since, here, both fluid fields possess a strong indirect interaction. In this case, all density settings lead to divergent runs. If the structure gets very thick, both interface systems get quasi independent from each other, and a concatenation of two implicit schemes can be successful. If we then lower the density of S, both interactions become stronger, and both interfaces influence each other again.

The lower block of Table 1 shows results for the same density settings, but now for the multi-coupling scheme MIQN. We observe a stable coupling for all settings. The number of iterations gets bigger for a stronger interaction between both fluid fields and the structure fields, but also for a stronger indirect interaction between both fluid fields.

6 Conclusions

In this work, we developed multi-coupling schemes based on either a composition of bi-coupling schemes or a true generalization of the underlying ideas of bi-coupling schemes. We tested those schemes by means of a simple, but yet challenging fluid-structure-fluid model problem. If one fluid-structure interface posseses a weak and the other one a strong interaction, the strong interaction dominates the overall stability and a concatenation of an explicit with an implicit scheme (cf. Section 3.1) leads to a stable multi-coupling scheme. If both fluid-structure interfaces show a strong interaction, a concatenation of two implicit coupling schemes does not result in an overall stable scheme, since both interface system hamper each other. For this case, we developed a true multi-coupling scheme, named "multi interface quasi-Newton" (MIQN) scheme by generalizing the ideas that we used in earlier work to deduce parallel FSI coupling schemes ([UBGM13, MUB$^+$13]. To our knowledge, this is the first time, that a fully implicit black-box multi-coupling scheme is described. This allows to simulate a new range of applications in a partitioned way.

We can easily generalize those ideas to multi-physics problems consisting of more than three physical fields. If one group of interface interactions is weak, and the other one strong, it should always be possible to couple all strong interactions with the MIQN scheme and then concatenate MIQN with explicit schemes for each weak interaction.

Acknowledgements: The financial support of the Institute for Advanced Study (IAS) of the Technische Universität München, and of SPPEXA, the German Science Foundation Priority Programme 1648 – Software for Exascale Computing are thankfully acknowledged. `Alya` is developed at the Barcelona Supercomuting Center, by Guillaume Houzeaux, Mariano Vázquez, et al. We want to thank, in particular, Juan Carlos Cajas, who helped to implement the `preCICE` adapter in `Alya`.

References

[Bru09] E. H. Van Brummelen. Added Mass Effects of Compressible and Incompressible Flows in Fluid-Structure Interaction. *Journal of Applied Mechanics*, 76, 2009.

[CGN05] P. Causin, J. F. Gerbeau, and F. Nobile. Added-mass effect in the design of partitioned algorithms for fluid–structure problems. Rapport de recherche, INRIA, 2005.

[DBV09] J. Degroote, K.-J. Bathe, and J. Vierendeels. Performance of a new partitioned procedure versus a monolithic procedure in fluid-structure interaction. *Computers & Structures*, 87, 2009.

[FL00] C. Farhat and M. Lesoinne. Two efficient staggered algorithms for the serial and parallel solution of three-dimensional nonlinear transient aeroelastic problems. *Computer Methods in Applied Mechanics and Engineering*, 182, 2000.

[Gat14] B. Gatzhammer. *Efficient and Flexible Partitioned Simulation of Fluid-Structure Interactions*. PhD thesis, Technische Universität München, Institut für Informatik, 2014.

[GM08] M. Graczyk and T. Moan. A probabilistic assessment of design sloshing pressure time histories in LNG tanks. *Ocean Engineering*, 35, 2008.

[HDHV09] R. Haelterman, J. Degroote, D. Van Heule, and J. Vierendeels. The quasi-Newton least squares method: A new and fast secant method analyzed for linear systems. *SIAM Journal on Numerical Analysis*, 47, 2009.

[HVAC09] G. Houzeaux, M. Vázquez, R. Aubry, and J. Cela. A massively parallel fractional step solver for incompressible flows. *J. of Comput. Phys.*, 228, 2009.

[KW08] U. Küttler and W. A. Wall. Fixed-point fluid–structure interaction solvers with dynamic relaxation. *Computational Mechanics*, 43, 2008.

[LAVH12] P. Lafortune, R. Aris, M. Vázquez, and G. Houzeaux. Coupled electromechanical model of the heart: Parallel finite element formulation. *Int. J. Num. Meth. Bio. Engng.*, 28, 2012.

[LKBD09] G. Link, M. Kaltenbacher, M. Breuer, and M. Döllinger. A 2D finite-element scheme for fluid-solid-acoustic interactions and its application to human phonation. *Computer Methods in Applied Mechanics and Engineering*, 198, 2009.

[Mic04] C. Michler. An interface Newton-Krylov solver for fluid-structure interaction. *International Journal for Numerical Methods in Fluids*, 47, 2004.

[MUB+13] M. Mehl, B. Uekermann, H. Bijl, D. Blom, B. Gatzhammer, and A. Van Zuijlen. Parallel Coupling Numercis for Partitioned Fluid-Structure Interaction Simulations. *submitted to SIAM SISC*, 2013.

[MY10] S. Minami and S. Yoshimura. Performance evaluation of nonlinear algorithms with line-search for partitioned coupling techniques for fluid-structure interactions. *Int. J. Num. Meth. Fluids*, 64, 2010.

[SBE+14] S. Sicklinger, V. Belsky, B. Engelmann, H. Elmqvist, H. Olsson, R. Wüchner, and K.-U. Bletzinger. Interface Jacobian-based Co-Simulation. *Int. J. Num. Meth. Eng.*, 98, 2014.

[UBGM13] B. Uekermann, H.-J. Bungartz, B. Gatzhammer, and M. Mehl. A parallel, black-box coupling algorithm for fluid-structure interaction. In *Proceedings of ECCOMAS Coupled Problems 2013*, Ibiza, Spain, 2013.

[UCG+14] B. Uekermann, J. C. Cajas, B. Gatzhammer, G. Houzeaux, M. Mehl, and M. Vázquez. Towards Partitioned Fluid-Structure Interaction on Massively Parallel Systems. In *Proceedings of WCCM XI/ ECCM V/ ECFD VI*, Barcelona, Spain, 2014.

[VLDV07] J. Vierendeels, L. Lanoye, J. Degroote, and P. Verdonck. Implicit coupling of partitioned fluid-structure interaction problems with reduced order models. *Computers & Structures*, 85, 2007.

Configurable and Collaborative Scientific Workflows

Michael Hahn and Dimka Karastoyanova

Institute of Architecture of Application Systems (IAAS)
University of Stuttgart
Universitätsstraße 38, 70569 Stuttgart
{hahn, karastoyanova}@iaas.uni-stuttgart.de

Abstract: The use of workflows to support and realize computer simulations, experiments and calculations is well-accepted in the e-Science domain. The different tasks and the parameters of the simulation are therefore specified in workflow models. Scientists typically work in a trial-and-error manner which means they do not know how the final workflow of a simulation has to look like. Therefore, they use a maybe insufficient workflow model as a basis and try to improve this model over multiple iterations to get a better approximation to the problem to solve. So in each iteration multiple trials are based on different variants of the same workflow model. Towards the goal of building variants of workflow models and enabling the reuse of existing scientific workflows in a controlled and well-defined manner, in this paper, we identify how configurable workflow models will support scientists to customize existing workflow models by their configuration. Therefore, we introduce possible configuration options for scientific workflows and how scientists can specify them. Furthermore, we show how configurable workflow models are a first step towards enabling the collaboration among scientists in creating scientific workflows.

1 Introduction

Workflows are well-accepted in the e-Science domain to support and realize computer simulations, experiments or calculations. Workflow models enable scientists to specify the tasks and parameters of a simulation. Due to the fact, that scientists typically work in a trial-and-error manner they use a maybe insufficient workflow model as a basis and try to improve this model over multiple iterations to get a better approximation to the problem to solve. Towards the goal of building variants of workflow models and enabling the reuse of existing scientific workflows in a controlled and well-defined manner, in this paper, we identify how configurable workflow models will support scientists to customize existing workflow models by their configuration.

Figure 1 shows an example for the configurability of a workflow model based on the executable workflow modeling language BPEL [BPE07]. On the right of Figure 1 is a set of variables which are used by the workflow model. The *maxIterations* variable specifies the maximum number of iterations of the *Create Parallel Instances* BPEL While loop activity. The model itself contains a set of service invocations (BPEL Invoke activities) that trigger the execution of existing scientific services, like the *CreateGrid* or *WriteIntermediateRe-*

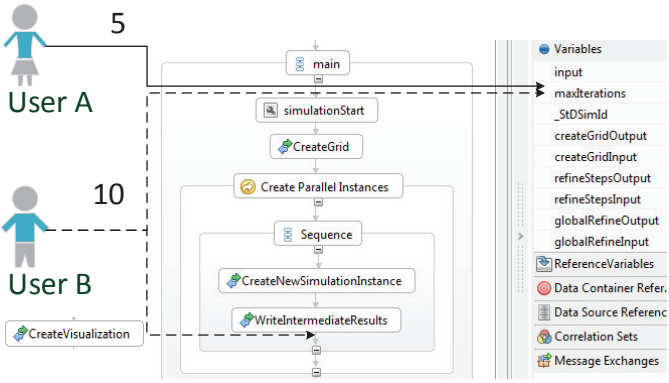


Figure 1: Example of a scientific workflow that is tailored to the needs of two different users through configuration

sults activities. On the left of Figure 1, two users (User A and User B) are shown. Both User A and B register specific values for the *maxIterations* variable. In addition to that, User B enriches the model with an additional activity (*CreateVisualization*) at the end of the while loop. If User A instantiates the workflow model the while loop is executed five times as specified over the registered configuration data. Each instance of the workflow model started by User B, executes ten loop iterations and at the end of each loop cycle the *CreateVisualization* activity is executed. We have already a prototypical realization of a multi-tenant aware workflow engine SCE^{MT}[1] that provides a subset of the configuration capabilities introduced in this paper. Furthermore, interested readers can find a demonstration[2] which shows the configuration of a workflow model with runtime data (user-specific variable values) using SCE^{MT} as described in the example.

The remainder of this paper is structured as follows: Section 2 provides an introduction to multi-tenancy, its capabilities and how the multi-tenancy support of a workflow engine can be utilized to enable the configurability of scientific workflows. Section 3 discusses the configurability of scientific workflows and introduces a set of possible configuration options. Since the support of collaborative work is an important topic when modeling and executing simulations – especially if scientists from different domains work together – the paper provides a discussion how such collaboration can be supported over configuration on the level of scientific workflows in Section 4. In Section 5, we compare our research against existing works. Finally, Section 6 summarizes our findings and presents a short outlook to future work.

2 Relationship between Multi-tenancy, Isolation and Configurability

Even though multi-tenancy is an application property typically considered in the context of Cloud computing, in fact multi-tenancy aware applications or services provide a variety

[1] SCE^{MT}: `http://www.iaas.uni-stuttgart.de/scemt/`
[2] SCE^{MT} Demo: `http://www.iaas.uni-stuttgart.de/scemt/#videos`

of capabilities which are beneficial regardless of how the application or service is provided to the users. In this section we therefore shortly introduce multi-tenancy as a property of applications and its capabilities, which also serves as a basis in the following sections.

In this paper we follow the definition of *multi-tenancy* provided in [SALM12], where the authors define multi-tenancy as "the sharing of the whole technological stack (hardware, operating system, middleware and application instances) at the same time by different tenants and their corresponding users". In the same work the authors also support the claim that resource sharing comes along with/or is a prerequisite for *isolation* (e.g. data, communication) of tenant's resources, *configurability* and *scalability*.

When talking about multi-tenancy it is important to distinguish two types of users: *tenants* and *tenant users*. A tenant is used to separate users of a multi-tenant aware application or service into organizational units like companies or departments which share the same view on the application or service [KMK12], [SALM12]. Since these organizational units are not necessarily completely disjoint and a user may belong to multiple tenants at the same time, tenant users enable the distinction of users belonging to more than one tenant. Multi-tenancy implies that applications or services are able to identify the resources belonging to a tenant (e.g. an institute) or tenant user (e.g. a scientist). For example such resources are data like simulation parameters or event data, or software artifacts like workflow models or scientific services.

The isolation of tenants and their users in nearly all parts of a system is the main requirement to provide multi-tenancy support [GSH+07, KMK12]. From a users point of view, isolation is an important capability of a system from which users can benefit. The following list describes the most important isolation types and some examples how the different types change the behavior of a system.

Data Isolation ensures that the data belonging to one tenant user is not accessible by any other user of the system. Chong et al. introduce three different approaches to realize data isolation on a database level [CCW06]. As a result, a multi-tenant database enables the sharing of tables between multiple users and simultaneously protects the data from any unintended access by other users.

Communication Isolation is a special kind of Data Isolation which isolates the communication (message exchanges) between two or more applications or services on a tenant user basis [SALM12]. This provides the ability to enforce the authentication of requests sent to (scientific) workflows or services by checking if an incoming request message belongs to the same tenant user as the workflow or service to invoke. Therefore, scientists can ensure that only allowed tenant users are able to invoke their scientific workflows or services.

Administration Isolation is provided if no tenant user is able to manage or administer resources which belong to another tenant user. For example, a SWfMS provides functionality for the administration of the workflow engine, workflow models and workflow instances. Administration Isolation enforces, that all resources of a tenant user can only be managed by himself.

Performance Isolation ensures that each tenant user gets the performance he paid for and that the potentially detrimental behavior of one tenant user does not adversely affect the system performance for other tenant users [WMTJ12, GSH+07]. In the domain of scien-

tific workflows, performance isolation can secure that all simulations running on a shared system are executed with predefined performance (not violating the performance of concurrently executed simulations) and the system automatically scales if the guaranteed performance can not be provided any longer.

Configurability enables the adaptation of the behavior or appearance of an application or service without the necessity to change the underlying implementation to the needs of tenant users [KMK12], [GSH+07]. Such customization is only possible if the application or service provide a set of configuration possibilities, which in turn enable users to perform ad-hoc customization of reusable artifacts to their needs. Since configurability is a main requirement for multi-tenancy to comply with the needs of all served users, multi-tenant aware applications provide natively a set of configuration options. For example, workflow models which provide commonly used functionality required in many scientific workflows can be adapted through configuration and therefore reused without the necessity to change the underlying model.

3 Configurable Scientific Workflows

Configurability enables users to customize predefined workflow models based on their needs by providing in addition a set of configuration data without the necessity to create new or adapt existing workflow models. These configuration data are used in all instances of a workflow model belonging to a single user. This allows users to customize an existing model by overriding predefined values (e.g. constants specified in a workflow model or transition conditions between activities) or registering customized control flow logic (e.g. additional activities or workflow fragments). Furthermore and because of this, workflow modelers are able to define more generic and reusable models (workflow templates) which can be adapted by the users themselves through configuration. For example, any user specific values of a scientific workflow (e.g. parameters or boundary conditions for computations) can be dynamically set upon workflow instantiation by using registered model configurations related to the user who instantiates the workflow model. Since the configuration data is stored independently of the workflow models, multiple tenant users can use the same workflow model as a basis for configuration without the necessity to maintain their own copy. If the shared workflow model has to be improved over time, only the single model has to be adapted and then the changes are dynamically reflected in all new created workflow instances of all tenant users.

We distinguish between three different types of configuration options for workflow models, the configuration of *workflow model runtime data*, *workflow model logic* and *workflow model appearance*. In the following each of these three categories and a selection of possible configuration options will be described.

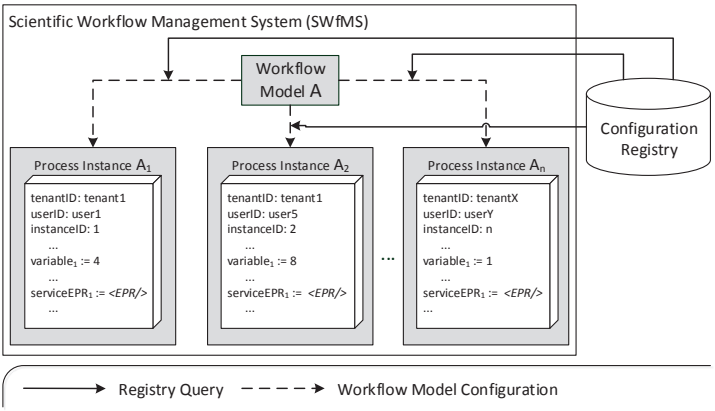


Figure 2: Example of a set of workflow instances using different runtime data which are assigned through the configuration of the workflow model

3.1 Configuration of Workflow Model Runtime Data

Figure 2 shows an example scenario for the configuration of *runtime data* of a workflow model. Runtime data can be registered for all elements of a workflow model which hold some data during the runtime (execution) of an instance of the workflow model. For example, the (initial) values of *variables* or *service endpoint references* can be specified over runtime data configurations. This means that a user can register a value for an element (e.g. a variable) which is used by the SWfMS to initialize the element whenever a new instance of the workflow model is created by the user. The registration and administration of the configuration data of the users can be realized over a *Configuration Registry*. To isolate the configuration data of all served tenant users, the registry has to provide multi-tenancy support. The configuration of variable values can be used for example to register user-specific constants or initial values. The configuration of service endpoint references makes it possible to change the target endpoint of a modeled service invocation. For example, if a tenant user hosts another (potentially adapted) version of a service invoked by the model, he can overwrite the specified service endpoint reference through configuration. Upon workflow instantiation the registered values are queried from the Configuration Registry by the SWfMS and used to initialize the corresponding workflow model elements for which they are registered. Based on the associated tenant information of the initial request which triggers the instantiation of the workflow model, the correct tenant specific runtime data is determined automatically. Since a workflow model may specify initial values for some of its elements, the configuration data has to be dynamically set during runtime to avoid that it is overwritten by such predefined initial values. As a result, the configuration of runtime data is also applicable to set/change the initial values or constants defined in a workflow model without the need to change the model. This makes it possible that multiple, parallel executed instances of the same workflow model triggered by different users, automatically use the correct configuration (runtime) data.

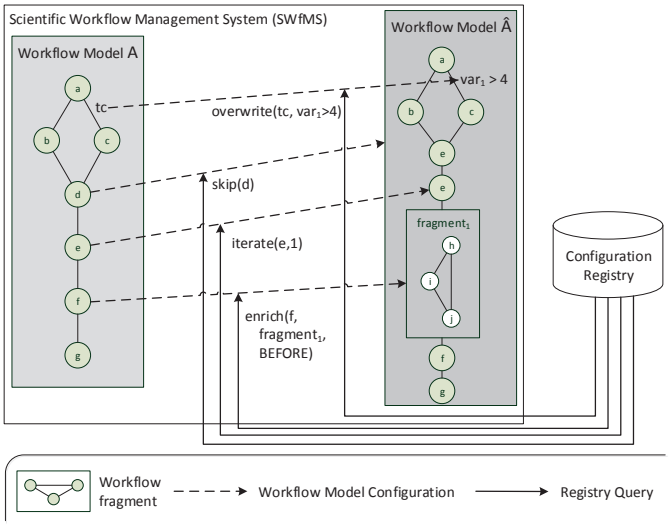


Figure 3: Example of creating a variant of a workflow model through the configuration of its control flow logic

3.2 Configuration of Workflow Model Logic

Another much more flexible and powerful option is the configuration of *workflow model logic* because users are then able to customize the control flow and therefore the behavior of the workflow. Based on the underlying workflow modeling language different configuration possibilities exist, in terms of different sets of modeling constructs which can be used for the configuration of a model. In general there are three different ways how the control flow can be configured.

First of all, the control flow can be ***enriched*** by the registration of additional workflow logic in form of workflow fragments (control flow snippets). A workflow fragment is a self-contained, well-formed, partially defined workflow model, like a sequence of activities [ELS+10]. These fragments are then dynamically weaved into the workflow model at the beginning of its execution (instantiation) triggered by a tenant user. As a result, the executed instance is based on an enriched, temporary workflow model which is the aggregate of the original (template) model and all registered workflow fragments. Figure 3 shows an example for such a scenario, where *Workflow Model A* is configured by enriching it with *fragment$_1$* to a temporary *Workflow Model Â*. The exact position where a fragment should be inserted into the workflow model, is specified by referencing an existing activity and a constraint which defines if the fragment should be placed before or after this activity (*enrich(f, fragment$_1$, BEFORE)*, Fig. 3).

Second, the control flow can be ***restricted*** by adding new transition conditions between two activities or by skipping activities as shown in Figure 3. There activity *d* is skipped and therefore not represented any longer in the configured temporary model Â (*skip(d)*).

Third, the existing control flow logic can be ***adapted*** by overwriting already defined tran-

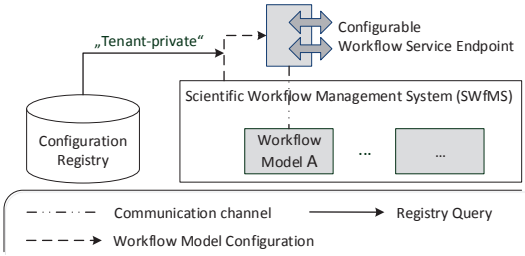


Figure 4: Example of configuring the accessibility of a workflow model

sition conditions or any other expressions (e.g. loop conditions in loop activities), or by configuring the repeated execution of modeled activities – also known as *iteration* [SK11]. Figure 3 shows an example for both adaptation possibilities. The defined transition condition *tc* of model A is configured by overwriting it with an another expression (*overwrite(tc, $var_1 > 4$)*) so that the link between activity *a* and *c* is only fired if $\text{var}_1 > 4$. Furthermore, activity *e* is configured to be iterated once (*iterate(e,1)*) and therefore is executed twice in each instance of the temporary workflow model Â. Sonntag et al. provide additional information on such an iteration operation for the repeated execution of activities [SK11].

All three configuration possibilities adapt the control flow of the workflow model on an instance level based on the tenant user. Available adaptation approaches from service compositions and workflows can be used to support such configuration.

3.3 Configuration of the Workflow Model Appearance

Figure 4 shows one possibility to configure the *workflow model appearance*, which means how the workflow is visible and accessible from the outside through its service interface. In the following we want to introduce and describe the configuration of the visibility and the level of Communication Isolation – more precisely the accessibility of the service interface – of a workflow model based on three constraints. These access constraints enable the owner of a workflow model to specify a scope with specific visibility and access restrictions. As a result, only well-defined groups of tenant users have access to the corresponding sets of workflow models: *public* – all tenant users of all tenants, *tenant-private* – all tenant users of one tenant and *user-private* – only one tenant user of one tenant.

Public: Workflow models that are configured as *public* do not have any access restrictions and therefore are visible to/accessible by all users of all tenants. Any incoming request sent to the service interface of a public workflow model is forwarded without any authentication. These workflow models can be used by any user of any tenant without any restrictions. The fact that a workflow model is public does not mean that its instances are also accessible by any user. If a public workflow model is instantiated by a tenant user, the instance is directly associated to this user. This secures all tenant-specific instances and their data from any unauthorized entities (Data Isolation). If a public workflow model is instantiated without any provided tenant user information, the instance is also public. In

this case, everyone is able to manage these instances via the administration functionality of the workflow engine. This access constraint enables for example the provisioning and sharing of workflow models which provide some useful domain-specific functionality to all tenants and their users while still enforcing Data Isolation on the workflow instance level. Furthermore, this constraint enables backward compatibility because non multi-tenant aware applications or services are able to send requests to public workflow models.

Tenant-private: This constraint isolates a workflow model on a tenant basis. Therefore the workflow model is only visible/accessible by the users of one specific tenant. The workflow engine only processes requests which reference the same tenant as the one associated to the workflow model to which the request is sent. For example, if the workflow model is associated to tenant A only requests of users of tenant A are processed, any other requests are rejected by the workflow engine. Requests without any associated tenant information have to be rejected. Similar to the public constraint case, the created instances of a tenant-private model belong to a single entity identified by the tenant information of the initial request sent to the service interface of a workflow model.

User-private: Workflow models which are configured as *user-private* are only visible to/ accessible by one specific user. The workflow engine only processes requests referencing the same tenant and tenant user as the one associated to the workflow model to which the instantiating request was sent. Requests without any associated tenant information are rejected. The created instances of a user-private model belong to a single entity identified by the tenant information of the initial request sent to the service interface of a workflow model. This is the default access constraint which is used when no configuration data is specified for the accessibility of the service interface of a workflow model.

3.4 Associating Configuration Data to Workflow Model Elements

In general, users should be able to identify and as a result, reference any element of a workflow model for which they want to specify configuration data. For example, an id, the name of the element or an XPath expression for XML[3] based model languages can be used to identify an element of a workflow model. The configuration possibilities of a workflow model range from complete configurability of the entire model to complete restriction of configurability (only some elements are configurable through a predefined set of configuration data). On the one extreme, the users of the workflow model want to be able to configure all parts of the model, while on the other, the providers of a workflow model want to restrict the configurability to some parts of the workflow using preset values in terms of security and control.

The configurability of a workflow model can be defined through the workflow model itself by specifying corresponding meta-data which mark all configurable elements and configuration possibilities. If there are no configurable (marked) elements modeled, a user is not able to configure anything. The enrichment of a model with additional activities can also be restricted using placeholders. Therefore, the workflow modeler should be able to

[3] W3C, Extensible Markup Language (XML) 1.0:`http://www.w3.org/TR/xml/`

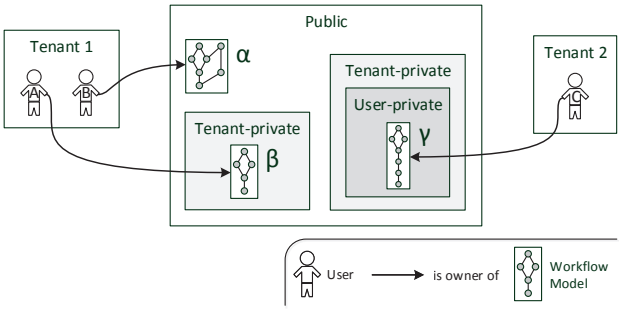


Figure 5: Specifying access scopes for workflow models based on predefined access constraints

insert such placeholders into a workflow model. The users can then register customized control flow logic only for a set of predefined positions in a workflow model which are represented by the modeled placeholders. The definition of different kinds of restrictions, how they can be specified in a workflow model and enforced during configuration time is out of the scope of this paper.

4 Discussion

Especially for scientific workflows, where a variety of scientists are working together to realize coupled simulations (e.g. multi-scale, multi-physics or multi-domain), collaboration is an important aspect. In our view, multi-tenancy can be used as a basis to enable the collaboration for scientific workflows. As described in Section 2, multi-tenancy provides already the required functionality because it makes it possible to establish the connection between users and their resources, isolate them in a defined manner and enables the user-based configuration of these resources. By default multi-tenancy realizes the strict isolation of resources among tenant users. If we support that this isolation can be configured in a well-defined and secure scope, then resources can be accessed and used in collaboration by multiple users. Therefore, tenant users should be able to specify access permissions for their resources (e.g. workflow models or instances) and any operations (e.g. read, change or delete) which are applicable to these resources to enable a secure collaboration between multiple tenant users. Furthermore, these access permissions should be easily assignable across tenants' boundaries, dynamically defined during all phases of a workflow model or workflow instance (e.g. modeling, deployment, runtime) and defined in a fine-grained manner on an operation level.

As a first step to enable the collaborative work on the level of workflow models, which means sharing one workflow model between multiple users, a mechanism to share them across tenant boundaries is required. The ability to configure the accessibility of a service interface described in Section 3 is such a mechanism which enables the sharing of a workflow model based on a set of predefined access constants. The SWfMS uses these constants to check if the request send to a service interface belongs to a tenant which is contained in the configured access scope of the workflow model.

Figure 5 shows how the access scopes can be used to enable the collaboration on a workflow model level. *User A* of *Tenant 1* is the owner of one tenant-private workflow model (*Workflow Model* β) which can be used by all users of the tenant (Users A and B). In contrast to that, *User B* is the owner of a public workflow model (*Workflow Model* α) which is visible/accessible by all users (A, B, C) of both tenants (1 and 2). *User C* of *Tenant 2* is the owner of a user-private workflow model (*Workflow Model* γ) which is isolated from all other users and therefore is only visible to User C. Based on Figure 5 consider the following scenario. User B has specified workflow model α which provides some useful and reusable functionality. User C requires exactly the same functionality as provided by workflow model α, but is not able to realize an equivalent model on his own. With the help of the introduced collaboration mechanism, User B is now able to simply enable User C to use his model without the necessity to transfer the workflow model to User B. So all users, no matter if they belong to the same organizational unit (tenant), can share their workflow models with others and as a result benefit from the joint use.

5 Related Work

Van der Aalst, Gottschalk et al. introduce configurable process models as a basis for reference modeling in [vdADG$^+$06] or [GvdAJV07]. They point out that a reference model which provides a general solution can be specialized for a concrete scenario over configuration. This enables the creation of new models by selecting specific parts of a configurable (reference) model that are relevant to the user [vdADG$^+$06, GvdAJV07]. Based on that, they describe "configuration as the inverse of inheritance" which means instead of enriching a model with new functionality, configuration restricts a general model to a concrete model. Therefore, they introduce two different techniques for the restriction of the control flow of a model, called *blocking* and *hiding*. In our work, we take into account additional configuration possibilities like the customization of runtime data, accessibility of the model or the enrichment of the control flow of a workflow model.

La Rosa et al. state that "a configurable process model captures the commonalities and the variability of a family of process models in a consolidated manner" [LRDTHM11]. As a result, users are able to specify values to individualize a configurable process model based on their requirements. Therefore the individualization of the configurable process models can be done automatically which removes a tedious and error-prone step in case the user requirements are reflected over the manual adaptation of the process model. They agree with the definition of configurable process models provided in [vdADG$^+$06] and [GvdAJV07]. In addition, they describe the shortcomings of existing configurable process modeling approaches and their lack of mechanisms for enable configurability not only for the control flow of process models. Therefore, they extend the Configurable Event-Driven Process Chain (C-EPC, [RvdA07]) notation to enable the definition of roles (e.g. a organizational unit) and objects (e.g. data). Based on that, they introduce a range of configurations for the way how those roles and objects and their association to tasks can be configured.

In [LL07] the authors introduce a formal model to define different customization options

for BPEL processes. Based on that and on the idea of reference models as defined in the works mentioned above, they provide the ability to specify reference process models containing possible customization options.

All the discussed related works use a reference model as a basis to enable configurability. The reference model therefore has to be specified using a modeling language which implicitly provides a set of configuration possibilities to individualize the model for a concrete scenario. In contrast to the business domain, scientific workflows are ever-changing models because scientists work in a trial-and-error manner to specify and improve the modeled simulations. In this case, it is not possible to predefine all variants of a scientific workflow in a reference model and as a result enabling configurability based on reference models is not an applicable approach for the scientific domain. Therefore, in this work we focus on enabling the configurability of workflow models without requiring a reference model specified with an enriched modeling language. This allows users to seamlessly build multiple variants of a workflow model or even of previously created variants just through the association of configuration data to the workflow model (see Sec. 3). As a result, any existing workflow model can be tailored on a per-user-basis through configuration using e.g. restriction, enrichment, replacement, overwriting.

6 Conclusion and Future Work

Our main goal in this work is to introduce the concept of configurable scientific workflow models. We described how multi-tenancy can be used as a basis to support the configuration of workflow models in a secure and well-defined manner. Furthermore, we introduced three different types of configuration options, which refer to the customization of different layers of a workflow model. For each of these three options, we presented concrete configuration possibilities. As a first step towards collaboration in scientific workflows, we presented our vision about how collaboration can be enabled through configuration.

Our current work focuses on the extension of our existing multi-tenant aware workflow engine SCE^{MT} to support all introduced configuration options and their validation using a case study.

In the future, we aim at supporting *configuration views*, in order to enable the customization of nearly all parts of a workflow model using the introduced configuration options. These views provide to the users a graphical representation of the resulting virtual models, which are the aggregate of the initial model and all specified configuration data.

Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310) at the University of Stuttgart.

References

[BPE07] Organization for the Advancement of Structured Information Standards (OASIS). Web Services Business Process Execution Language Version 2.0. OASIS Standard, 2007.

[CCW06] Frederick Chong, Gianpaolo Carraro, and Roger Wolter. Multi-Tenant Data Architecture, 2006.

[ELS+10] Hanna Eberle, Frank Leymann, Daniel Schleicher, David Schumm, and Tobias Unger. Process Fragment Composition Operations. In *Proceedings of APSCC 2010*, pages 1–7. IEEE Xplore, Dezember 2010.

[GSH+07] Chang Jie Guo, Wei Sun, Ying Huang, Zhi Hu Wang, and Bo Gao. A Framework for Native Multi-Tenancy Application Development and Management. In *E-Commerce Technology and CEC/EEE 2007*, pages 551–558, 2007.

[GvdAJV07] Florian Gottschalk, Wil MP van der Aalst, and Monique H. Jansen-Vullers. Configurable Process Models – A Foundational Approach. In Jrg Becker and Patrick Delfmann, editors, *Reference Modeling*, pages 59–77. Physica-Verlag HD, 2007.

[KMK12] Rouven Krebs, Christof Momm, and Samuel Kounev. Architectural Concerns in Multi-tenant SaaS Applications. In *Proceedings of CLOSER'12*, pages 426–431, 2012.

[LL07] Alexander Lazovik and Heiko Ludwig. Managing process customizability and customization: Model, language and process. In *Web Information Systems Engineering–WISE 2007*, pages 373–384. Springer, 2007.

[LRDTHM11] Marcello La Rosa, Marlon Dumas, Arthur HM Ter Hofstede, and Jan Mendling. Configurable Multi-perspective Business Process Models. *Information Systems*, 36(2):313–340, 2011.

[RvdA07] Michael Rosemann and Wil MP van der Aalst. A configurable reference modelling language. *Information Systems*, 32(1):1–23, 2007.

[SALM12] Steve Strauch, Vasilios Andrikopoulos, Frank Leymann, and Dominik Muhler. ESB^{MT}: Enabling Multi-Tenancy in Enterprise Service Buses. In *Proceedings of CloudCom'12*, pages 456–463, 2012.

[SK11] Mirko Sonntag and Dimka Karastoyanova. Enforcing the Repeated Execution of Logic in Workflows. In *Proceedings of BUSTECH'11*, pages 1–6. IARIA, September 2011.

[vdADG+06] Wil MP van der Aalst, Alexander Dreiling, Florian Gottschalk, Michael Rosemann, and Monique H. Jansen-Vullers. Configurable Process Models as a Basis for Reference Modeling. In *Business Process Management Workshops*. Springer, 2006.

[WMTJ12] Stefan Walraven, Tanguy Monheim, Eddy Truyen, and Wouter Joosen. Towards Performance Isolation in Multi-tenant SaaS Applications. In *Proceedings of MW4NG'12*, page 6. ACM, 2012.

R as an Integration Tool in High Performance Computing – Lessons Learned

Patrick Boba, Stefan Gries, Kay Hamacher

Computational Biology and Simulation
Technische Universität Darmstadt
Schnittspahnstraße 2
64283 Darmstadt
Germany
{boba,hamacher}@cbs.tu-darmstadt.de

Abstract: In this paper we discuss the usage of the statistical software R to integrate software components for large-scale simulations. To this end, we show how to interface different components written in various programming languages under the umbrella of R.

We base our study on well-known tools such as COPASI [HSG+06] and connect it to our own R package SOMNIBIEN (Simulation Of Metabolic Networks Influenced By Internal And External Noise). In particular, we show the need to use Fortran code for fast simulation via stochastic differential equations. This implies the need to interface with code in C, while offering the opportunity to leverage the superior analysis capabilities of R and its plotting capabilities.

Eventually, our system compiles models defined in an XML-dialect, translates them into C in the background and executes a Fortran solver for stochastic differential equations (SDEs) on them. Furthermore, we present the integration with download-capabilities from the XML-based BioModels Database [FH03a].

For the lessons learned, we describe pitfalls and performance bottlenecks and provide guidelines for future work on the integration of different code repositories into one R framework. Finally, we show performance evaluations on multi-core parallelization within our package.

1 Introduction

Nowadays network simulations play a major role in biological sciences and beyond. Especially in systems biology, simulations of genetic or metabolic networks have become a state-of-the art method to analyze properties and dynamics of those networks [KLW+09] and assign evolutionary meaning to the experimentally determined network topologies. While network science has identified several domains in which network simulations are an important research tool, here we will focus on biological, intracellular networks to illustrate our lessons. Here, computational biology poses an interesting challenge due to several problems plaguing the field: legacy code infrastructure, several often contradicting programming styles, poorly curated data sources with the need to manually interfere in

models and software, as well as hypothesis driven investigation approaches that demand both HPC resources and visual/interactive analytics.

In general, the network or graph topologies mentioned above can be represented in so-called flow charts (see Fig. 1). Here, vertices are molecular species quantified by their respective concentrations and the edges represent (enzyme driven) chemical reactions.

Changes in the molecular concentrations (A to E in Fig. 1) are characterized by rate constants controlling the fluxes (v_n) between the different chemical species. In the most simple case the stoichiometry of a reaction indicates that one compound is transformed into another compound with a specific reaction rate (see Eq. 1).

$$A \xrightarrow{v_1} B \tag{1}$$

$$\frac{\mathrm{d}A}{\mathrm{d}t} = -v_1 \quad \text{and} \quad \frac{\mathrm{d}B}{\mathrm{d}t} = v_1 \tag{2}$$

Each chemical reaction can be written as an ordinary differential equation (ODE) resulting from the stoichiometry of the involved components. For the example of Eq. 1 we derive Eq. 2 which states that metabolite A's concentration is reduced with speed v_1 and therefore at the same speed metabolite B is produced from A.

In general, stoichiometric coefficients c_{ij} exist for every metabolite M_i and every jth reaction with respective rate v_j. Then, we can write a coupled system of equations as follows:

$$\frac{\mathrm{d}M_i}{\mathrm{d}t} = \sum_{j=1}^{r} c_{ij} v_j \tag{3}$$

Typically, the right-hand side is more complex than in this simple, linear example. The general model is given as *coupled, non-linear* ODEs for which only numerical solutions can be obtained.

Now, this modeling approach is based on the so-called "perfect mixing" paradigm, which assumes that every metabolite and every other molecule is present in sufficient concentrations and more or less homogeneous throughout the cellular volume. But in reality, biology is driven by noise. Therefore, a more detailed model is given by a set of stochastic differential equations (SDEs). The noise then captures effects on the individual reactions like changes in the organisms environment (e.g., temperature, stress, etc.). We therefore need to simulate *and* analyze the output for an SDE system:

$$\mathrm{d}M_i(t) = f_i\left(M_1(t), \ldots, M_r(t), v_i^{\downarrow}, v_i^{\uparrow}\right)\mathrm{d}t + \beta \cdot M_i \cdot \mathrm{d}B_t^{(i)} \tag{4}$$

where $v_i^{\downarrow}, v_i^{\uparrow}$ are the reaction rates of the particular forward and backward reactions for the metabolite i. Furthermore, β scales the strength of the noise and $B_t^{(i)}$ is a Brownian process. Note that i.i.d. noise[1] with $B_t^{(i)} \neq B_t^{(j)}$ for $i \neq j$ as well as universal noise[2]

[1] e.g. to capture deviation from perfect mixing

[2] e.g. for temperature changes throughout the cell

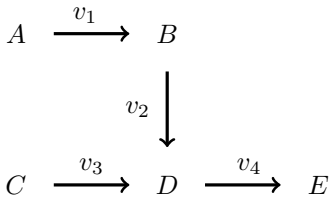


Figure 1: Reaction scheme of a metabolic network represented as a flow chart. The change in concentration of the metabolites indicated by uppercase letters (A to E) is controlled by the corresponding reaction kinetics (v_1 to v_4).

$B_t^{(i)} := B_t$ [GGJ+11] are included in this model of multiplicative noise. Such systems can then be engineered, e.g., optimized in, e.g., its accumulated output by heuristics [Ham13, Ham06] in biotechnology.

A typical work-flow in the life science, includes several manual, error-prone steps: downloading, local storage of model files, simulations by GUI-based tools without parallelization techniques, and final analysis with additional software. Such a work-flow makes it quite hard to achieve desirable goals: 1) reproduciblity is hindered due to the manual nature of file/directory-based documentation, 2) HPC-facilities cannot be leveraged for large-scale parameter scans and optimization efforts, and 3) efficiency is reduced due to a back-and-forth between analysis and simulation programs.

2 Our Contribution

Our primary goal was to develop an `R` package that provides users with a flexible and efficient code base to solve SDEs and analyze the results without migrating between different software systems and architectures. We contribute a) a software package [BGH14] and b) the lessons learned during the integration of the following components into a self-contained, both HPC-enabled and quasi-interactive simulation system:

- download capabilities for curated models – mostly to enable *reproduciblity* of simulations by others
- an `R`-integrated compilation facility of SBML[3]-models to C source code
- an *efficient* SDE solver including a global noise term based on Fortran90 [Ste08].
- interactive sessions via REPL[4]-based compilation and object-file linking mechanisms for the to-be-simulated systems, and

[3] SBML=Systems Biology Markup Language, an XML variant for models from established libraries [Huc03, sbm]

[4] REPL=Read-eval-print loop

- analysis routines in `R` typically employed in time series analysis, e.g., a parallel computation of the time-lagged cross-correlation functions between concentrations of different chemical species.

3 Approach

Our ultimate goal is the simulation, optimization, and investigation towards parameter sensitivity of bio-networks. A very versatile and powerful system to analyze time series and other (big) data is the `R` environment for statistical computing.

Furthermore, wet lab biologists and biotechnologists tend to investigate their systems interactively. Additionally, the involved parameter spaces are high-dimensional and thus demand at times the computational resources only offered by HPC infrastructures. Therefore, we decided to integrate all steps into one system that can be used either in batch mode for intensive simulations or interactively. A typical workflow is shown in Fig. 2.

Note that the major challenge we faced lies in the efficient and transparent integration of various external tools and data sources, while maintaining the possibility of user- and problem-specific analysis procedures. At the same time, the system is supposed to handle *any* SBML model and should not work only for a particular one.

Taken together, this lead us to use COPASI [HSG+06] as an SBML-to-C compiler (run in batch mode), SDERK [Ste08] as the numerical engine for the integration of the SDEs of Eq. 4; in particular to simulate SDEs rather than the less realistic models based on ODEs, C routines to interface components, OpenMP as a parallelization approach, and an R wrapper to steer compilation, linking, and calling of routines.

To give an impression of the achieved simplicity and abstraction of the integrated tools, we show 1 an example workflow in listing. The major steps are the download of the required network model from the database (`getModel`) followed by the creation of the shared object(`PrepareODEFile`). After extracting the network's initial metabolite concentrations and parameters, the simulation can be started at those default values with the `StartSim` command. A common analysis would, for example, continue with the cross correlation function (ccf) of the different metabolite time series.

3.1 Related Work

Besides COPASI [HSG+06], there exist several other (simple) simulation packages such as CellDesigner [Fun03], which is a GUI driven editor that allows editing of existing networks or building them from scratch based on SBML. Various other methods exist to give researchers without a simulation background the tools to simulate intracellular networks, e.g., PySB [LMBS13].

While all these systems address some of the challenges and fulfill some of the requirements from above, we lacked an integrated package – in particular one that interfaces with `R`.

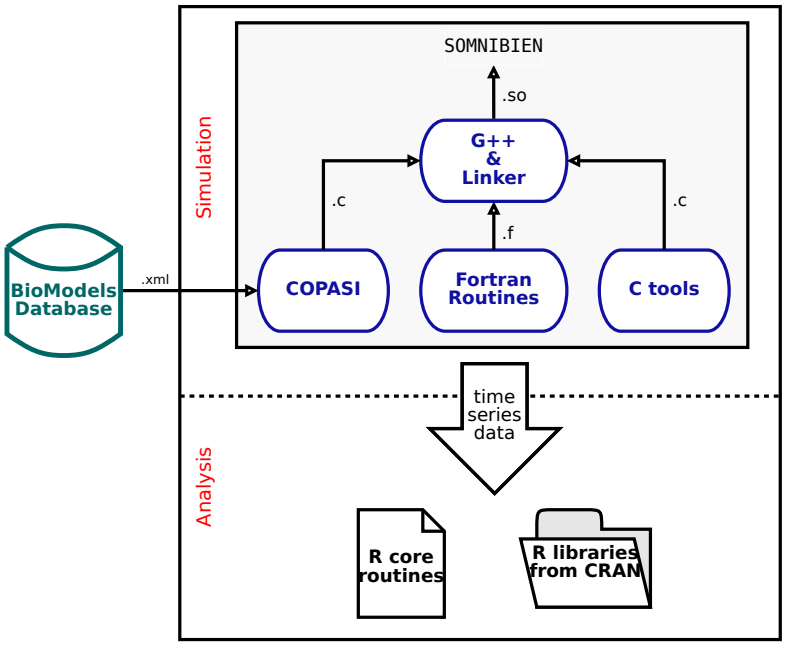


Figure 2: Basic workflow of our system. Models are downloaded as XML-files from the Internet or loaded from the file system. Various calls to external components (COPASI, g++ compiler, linker) compile SBML to C, combine this with numerical code in Fortran, and tool infrastructure in C into an object file, that can be called from within SOMNIBIEN and therefor R. This enables the user to analyze the simulated data *directly* within R using its superb analysis procedures as well as additional R-software from repositories such as CRAN [cra].

One could – in theory as it turned out – also combine existing R solvers such as deSolver [SPS10] with the R-based SBML-library SBMLR [Rad04] to achieve the same. Unfortunately, we were never able to a) integrate these components seamlessly, b) always get a working system that did not crash, and c) deal with the various SBML file format revisions and versions [Huc03, FH03b, sbm].

4 Lessons Learned

A first development approach intended to compile the Fortran and the C code containing the SDEs into shared objects during R execution. A pointer from the SDE function should then be passed to the also dynamically loaded solver in Fortran. Figure 3 illustrates the intended architecture of the package.

This leads to two problems, both caused by the fact that R treats function symbols from shared objects internally as objects. First, pointers to objects in R cannot be retrieved from

Listing 1: Typical workflow of the SOMNIBIEN package.

```
library("SOMNIBIEN")

## Downloading model from BioModels Database based on the ID (right now
    only curated Models are supported). This will  download an SBML file
    to your current working directory
getModel(42)

## Generating the shared object
PrepareODEFile("BIOMD0000000042.xml", filename="Bio042")

## Extracting the necessary information of the network from the C file.
    Note that PrepareODEFile generated a Copasi-style C file in background
InitConc <- GetCopasiInitialValues("BIOMD0000000042.c")
Parameters <- GetCopasiParameterValues("BIOMD0000000042.c")

## Starting the simulation
SimData <- StartSim(sfile="Bio042.so", parameters=Parameters,
    concentrations=InitConc, stepwidth=0.01)

## Evaluation with cross correlation function
EvalData <-  evalTimeseries(SimData, method="ccf")
```

the C/Fortran code, so passing function pointers to the SDERK-solver is not possible. Second, the Fortran interface to `R` does only supports primitives to be passed. Consequentially, `R` cannot pass functions or pointers to functions.

Due to this problem we had to develop additional C code which we name "controller". Figure 4 shows a schematic of the final architecture. The controller and the solver are compiled into a shared object during *installation* of the R package and thus only once. When the simulation is started via the `R` REPL or in `Rscript` batch mode, the wrapper compiles an SBML file containing the model into C; then the full SDE model file is compiled into a shared object that is in turn loaded *dynamically* from the controller, which passes the function pointer to the actual solver SDERK.

Since the Fortran solver requires the afore mentioned C files, we use the command line version of COPASI to convert the SBML markup into C code before running it with the `sderk90` solver. This whole process, however, is integrated into our `R` package, which will recognize if an SBML file is passed instead of a C file and automatically convert it to the required format in the background – letting the user experience a smooth integration without the need to understand the particulars "under the hood".

This involved flow of information, (function) pointers, etc. was necessary to achieve all of the stated above goals. In Listing 2 we illustrate in detail the code necessary for the pointer arithmetic and conversion to interface the components. The overall architecture fulfills all requirements from above and can be used by non-experts. However, it is also fragile against subtle changes in some of the components, in particular in the solver itself. The overall message at this point is that without loss of performance, one can integrate various tools and codes to offer an integrated, HPC-enabled simulation infrastructure.

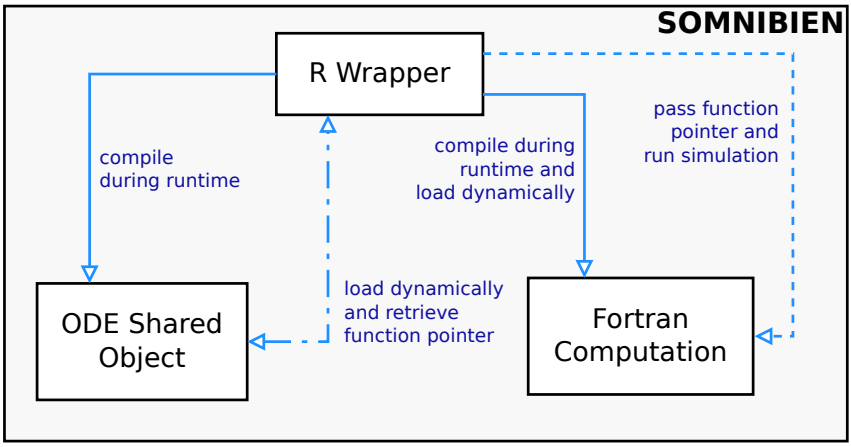


Figure 3: Initially intended package architecture. An interactive or batch-mode `R` process compiles the SDE function and the FORTRAN code during runtime and loads both. Then a pointer to the SDE function is retrieved and passed to the Fortran solver.

4.1 Parallelization

4.1.1 Parallelization of external noise

We further proceeded to improve the performance by various parallelization strategies. First, *external* noise in the form of parameter changes can easily be mapped onto SIMD systems as each simulation for a perturbed parameter set[5] is independent of any other simulation – thus, we face the rather undemanding challenge of parallelizing an embarrassingly parallel [Fos95] problem. This can be achieved within `R` itself, e.g., via the `snow` package [TRL09]. We have already discussed how the `snow` package performs very well in biophysical simulations [HWH10].

Therefore, at this point we can almost perfectly parallelize parameter scans and the analysis procedures following the myriads of simulations afterwards. This applied in particular to "grid scanning" of the parameter space as is frequently done, cmp. [XK07].

4.1.2 Parallelization of analysis steps

Analysis of trajectories and simulated data in general can sometimes be as time consuming as the simulation itself. Therefore, any efficiency improvement by, e.g., parallelization, is beneficial.

We have – as a proof of concept – implemented the computation of all cross-correlation functions (ccf) $a_{ij}(\tau) := \sum_{t=0}^{T-\tau} M_i(t) \cdot M_j(t+\tau)$ for the metabolite concentrations M_i

[5] such as the rate constants $v_i^{\downarrow}, v_i^{\uparrow}$ in Eq. 4

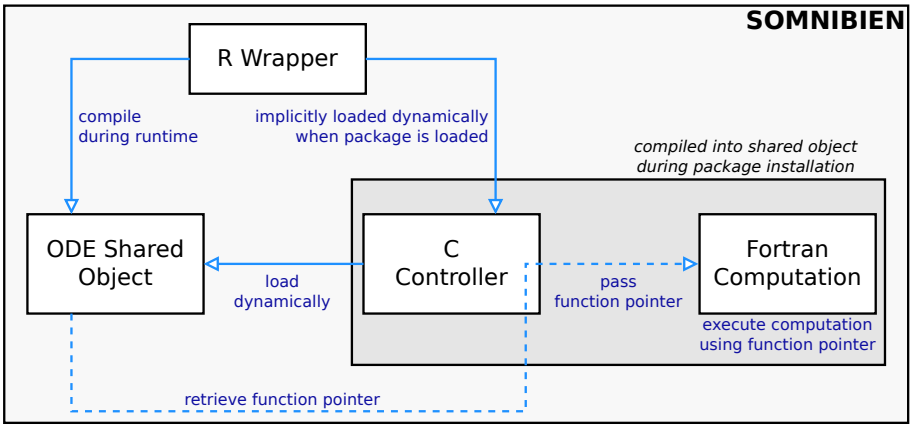


Figure 4: Overview of the implemented package architecture. The C and Fortran sources are compiled into one shared object during package installation. Each time the package is loaded in `R` using the `library` command, this shared object is loaded, too. The SDE shared object is compiled during `R` runtime. When starting a simulation the controller part of the SOMNIBIEN shared object loads the SDE shared object and retrieves a pointer to the SDE function, which is passed to and stored in the Fortran part during the simulation.

and M_j in parallel. This can be done very efficiently and easily with add-on packages in `R`, for example via the `doParallel` package. Since almost all analysis is done for all of the time series, this straight-forward approach is almost always applicable.

Note that both parallelization steps in Secs. 4.1.1 and 4.1.2 are very obviously parallel and thus – as long as there are more sub-jobs than processing elements perfectly parallelizable [MSM04].

4.1.3 Parallelization of single instances

Now, in optimization and network design one frequently deals with only one system and thus parameter set at a time – for example in Simulated Annealing in similar approaches [Ham13, Ham06], the optimization procedure would need to simulate just *one* replica of a system, propose a new one with varied parameters or graph topology, estimates its properties via simulation, and so on. Here, high-level parallelization over data-parallelism as discussed above would not be of any use. We therefore also tried to parallelize an *individual* simulation based on our architecture. To this end, we have employed the OpenMP parallelization paradigm to the evaluation of the right-hand side of the SDEs in Eq. 4 and modified the compilation routines to this effect.

A direct parallelization of the Runge-Kutta algorithm in `sderk90` is prohibitive, as the prediction and correction steps imply heavy thread creation and collapse, which caused a too large overhead during our tests to be of any use.

Our empirical results on the parallelization performance evaluation are shown for the net-

works of Tab. 1 in Figs. 5 and 6. Obviously there are only marginal, if at all efficiency improvements. Sometimes the OpenMP threading created so much overhead while the computational costs of the evaluation of the rhs was so small, that no break-even point exists. We conclude that the costs parallelization of a `single` instance is prohibitive.

Table 1: Overview of networks features: Number of metabolites, parameters and reactions

ID	Metabolites	Parameters	Reactions	ID	Metabolites	Parameters	Reactions
Mod042	15	25	25	Mod320	9	31	14
Mod160	19	47	41	Mod422	22	13	43
Mod209	6	49	12	Mod505	45	140	59

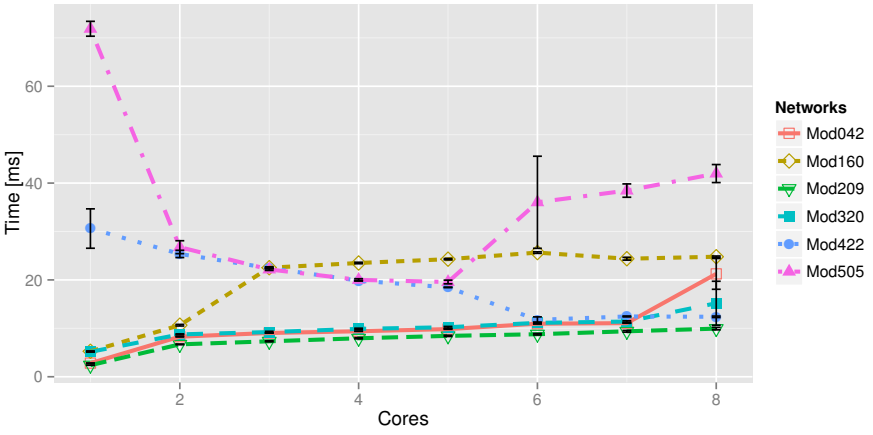


Figure 5: Execution times of an integration step averaged over the six different metabolic networks of Tab. 1 without internal noise ($\beta = 0$ in Eq. 4). Each network was simulated 100 times on the given number of cores. Error bars show the standard deviation.

5 Conclusion and Outlook

The package we presented, gives life scientists a tool to investigate the dynamics of (metabolic) networks without switching between different software systems, even allowing for large-scale computations on HPC structures. In the beginning we defined some important goals such as *interactive* and *HPC, stand-alone* usage; integration with `R` for analysis, as well as efficiency and "one-stop" usage. All these aspects are addressed by SOMNIBIEN.

Our contribution is a recipe how to deal with the particulars of `R` as a system omnipresent in the life sciences. Other approaches – such as the SimTech Workflow Management System [SKL10] – are more general and/or applicable to HPC settings but not frequently used in the life sciences where workstations and group clusters are standard.

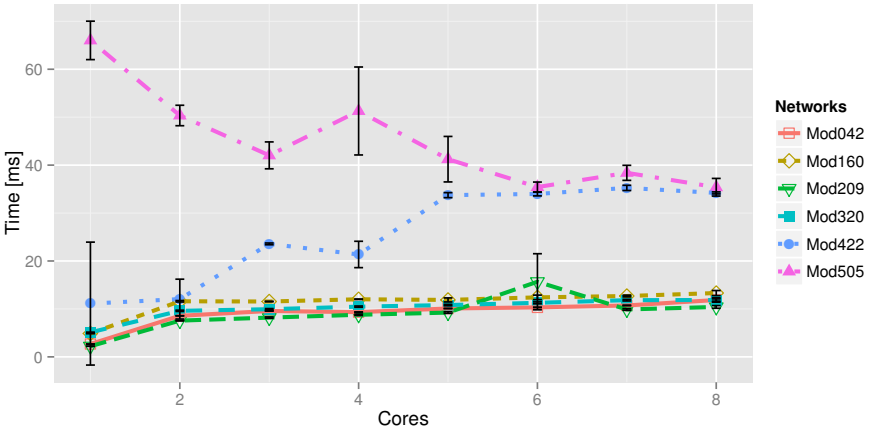


Figure 6: Same as in Fig. 5, but for a noise level of $\beta = 0.01$ in Eq. 4.

However, we had to overcome several obstacles which we encountered. Most of them are a direct results of R's layout and mechanism to handle objects and foreign-function-interfaces (FFI). We developed a mechanism via pointers in C and Fortran to integrate various tools. While the overall architecture is complex for the developer and maintainer, the user can handle the complex tasks at an abstract level. The overall method did not lead to any noticeable overhead in the computational efficiency.

Our parallelization strategies are two-fold: while the first one over independent replica parallellizes almost perfectly, our second approach to leverage OpenMP for a single instance showed problems due to high overhead costs. We hope that our short description helps other simulation and HPC researchers to integrate their systems with R, as it has become the predominant platform for end-users and analysts in recent years.

References

[BGH14] P. Boba, S. Gries, and K. Hamacher. SOMNIBIEN: Simulation Of Metabolic Networks Influenced By Internal And External Noise, 2014. *to be submitted*, available from http://www.cbs.tu-darmstadt.de/somnibien.

[cra] The Comprehensive R Archive Network. http://cran.r-project.org/.

[FH03a] A. Finney and M. Hucka. Systems biology markup language: Level 2 and beyond. Biochemical Society Transactions, 31(6):1472–1473, 2003.

[FH03b] A Finney and M Hucka. Systems biology markup language: Level 2 and beyond. Biochem Soc Trans, 31:1472–1473, 2003.

[Fos95] Ian Foster. Designing and building parallel programs : concepts and tools for parallel software engineering. Addison-Wesley, Reading, Mass, 1995.

[Fun03] A. etal Funahashi. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico, 1(5):159–162, 2003.

[GGJ+11] E. Gehrmann, C. Gläßer, Y. Jin, B. Sendhoff, B. Drossel, and K. Hamacher. Robustness of glycolysis in yeast to internal and external noise. Phys. Rev. E, 84:021913, 2011.

[Ham06] K. Hamacher. Adaptation in stochastic tunneling global optimization of complex potential energy landscapes. Europhys. Lett., 74(6):944–950, 2006.

[Ham13] K. Hamacher. A New Hybrid Metaheuristic - Combining Stochastic Tunneling and Energy Landscape Paving. In M. Blesa, C. Blum, P. Festa, A. Roli, and M. Sampels, editors, Hybrid Metaheuristics, volume 7919 of Lecture Notes in Computer Science, pages 107–117. Springer Berlin, 2013.

[HSG+06] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes, and Ursula Kummer. COPASI—a complex pathway simulator. Bioinformatics, 22(24):3067–3074, 2006.

[Huc03] M. et al Hucka. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics, 19(4):524–531, 2003.

[HWH10] F. Hoffgaard, P. Weil, and K. Hamacher. BioPhysConnectoR: Connecting Sequence Information and Biophysical Models. BMC Bioinformatics, 11(1):199, 2010.

[KLW+09] E. Klipp, W. Liebermeister, C. Wierling, A. Kowald, H. Lehrach, and R. Herwig. Systems biology. John Wiley & Sons, 2009.

[LMBS13] Carlos F Lopez, Jeremy L Muhlich, John A Bachman, and Peter K Sorger. Programming biological models in Python using PySB. Molecular Systems Biology, 9(1), 2013.

[MSM04] T. Mattson, B. Sanders, and B. Massingill. Patterns for Parallel Programming. Addison-Wesley, 2004.

[Rad04] Tomas Radivoyevitch. A two-way interface between limited Systems Biology Markup Language and R. BMC Bioinformatics, 5(1):190, 2004.

[sbm] Systems Biology Markup Language. `http://sbml.org`.

[SKL10] M. Sonntag, D. Karastoyanova, and F. Leymann. The Missing Features of Workflow Systems for Scientific Computations. In 3rd Grid Workflow Workshop (GWW), Lecture Notes in Informatics, pages P–160, 2010.

[SPS10] Karline Soetaert, Thomas Petzoldt, and R. Woodrow Setzer. Solving Differential Equations in R: Package deSolve. J. Stat. Softw., 33(9):1–25, 2010.

[Ste08] Daniel A. Steck. SDERK90, A Fortran 90 Integrator for Ito Stochastic Differential Equations, 2008. available at `http://steck.us/computer.html`.

[TRL09] Luke Tierney, A.J. Rossini, and Na Li. Snow: A Parallel Computing Framework for the R System. International Journal of Parallel Programming, 37(1):78–90, 2009.

[XK07] Z. Xie and D. Kulasiri. Modelling of circadian rhythms in Drosophila incorporating the interlocked PER/TIM and VRI/PDP1 feedback loops. J. Theo. Biol., 245(2):290 – 304, 2007.

Listing 2: Code snippet from the C controller, that mediates pointer between the Fortran solver `sderk90` and the shared object of the network's SDE system. `odefun` is the function to be integrated and `setFA` manipulates a global pointer variable. Furthermore, the code allows for a user-provided noise function in Eq. 4 via the β argument – if its name is given, then an object file is loaded and integrated into the final .so file.

```
...
// forward-define Fortran functions
void setFA(void (**)(double*, double**, double**, double*));
void setDrift(void (**)(double*, double**, double*));
void setBeta(double*);
...

// passing parameters as s-expressions
SEXP launchSim(SEXP infile, ..., SEXP trmethod)
{
...
    void * clib;
    PROTECT(infile = coerceVector(infile, STRSXP));
    // load given shared object and extract needed function symbol
    clib = dlopen(CHAR(STRING_ELT(infile, 0)), RTLD_NOW | TLD_LOCAL);
    // load dynamic library "infile"
    PROTECT(fnname = coerceVector(fnname, STRSXP));
    // allocation of 'ode_fun' using "fnname" function from "infile"
    *(void **)(&ode_fun) = dlsym(clib, CHAR(STRING_ELT(fnname, 0)));
        UNPROTECT(2);
    // pass function pointer to fortran
    setFA(&ode_fun);
    void * ex_drift;
    if(isNumeric(beta)) {
        PROTECT(beta = coerceVector(beta, REALSXP));
        // set beta value for internal noise
        setBeta(REAL(beta));
        UNPROTECT(1);
    } else if(isString(beta)) {
        // basically do the same with drift function as with ODE
        // function, if custom one is to be used.
        // function pointer declaration 'drift_fun':
        void (*drift_fn)(double*, double**, double*);
        PROTECT(beta = coerceVector(beta, STRSXP));
        PROTECT(betafname = coerceVector(betafname, STRSXP));
        // load dynamic library "beta"
        ex_drift = dlopen(CHAR(STRING_ELT(beta, 0)), RTLD_NOW | RTLD_LOCAL
            );
        // load "betafname" symbol from "beta" into 'drift_fun'
        *(void**)(&drift_fn) = dlsym(ex_drift, CHAR(STRING_ELT(betafname,
            0)));
        UNPROTECT(2);
        // set drift function pointer in Fortran
        setDrift(&drift_fn);
    } else {fprintf(stderr,"Error in beta argument"); exit(1); }
...
}
```

Towards Automated Analysis of Eye Tracking Studies using the Workflow Technology

T. Blascheck[1], K. Vukojevic-Haupt[2], D. Weber[2], D. Karastoyanova[2], T. Ertl[1]

[1]Institute for Visualization and Interactive Systems
[2]Institute of Architecture of Application Systems
University of Stuttgart
Universitätsstrasse 38, 70569 Stuttgart
tanja.blascheck@vis.uni-stuttgart.de
karolina.vukojevic@iaas.uni-stuttgart.de

Abstract: Eye tracking studies have become one means to evaluate user behavior. For example, eye tracking is used in marketing research, psychology, human-computer interaction, or visualization. Analyzing eye movement data can help to find where on a stimulus and at what areas of a stimulus a participant focused on. Eye tracking can be used besides classical benchmark metrics such as completion times and accuracy rates of correctly given answers. However, evaluating eye movement data is a time consuming task, as a large amount of data has to be analyzed. Typically, multiple software systems have to be used, each having a different format of the data. This leads to tedious work, as the analyst has to reformat the data and learn how to use different software systems. Therefore, we suggest to analyze eye movement data using workflow technology. This allows to automatically analyze the data, and have a reproducible result. The analyst is not concerned with the technicalities and can rather focus on the interpretation of the analyzed data. Allowing an automatic evaluation using the workflow technology, requires that the analysis functionality is available via web services. In this paper, we contribute a workflow system based on web service operations. The web service offers functionality of an eye tracking analysis framework for the automatic analysis of eye movement data. The workflow system can be used to model and execute different types of eye tracking evaluations.

1 Introduction

Analyzing eye movement data has become an important part in user studies. Nowadays, eye tracking is used in research areas such as psychology, marketing, human-computer interaction, or visualization [Duc02]. Classical benchmarks such as completion times and accuracy rates of correctly given answers can be extended by including eye movement data in the analysis. This helps to analyze where on a stimulus and at what objects of a stimulus a participant inspected. However, analyzing eye movement data is a time consuming and error prone task. For each step, the analyst needs a different software system and the data has to be converted into the right format. Using different software systems makes it hard to reproduce the analysis steps in a later study. To overcome those problems, in an earlier

work we have implemented an eye tracking analysis tool called eTaddy (Eye Tracking Analysis, conDuction, and Designtool for user studYs) [BRE13]. However, eTaddy has the limitations, that analysis steps have to be repeated manually and a complete analysis processes can not be applied to a different eye tracking study. To tackle those issues, we now contribute a workflow systems, which uses the functionality of eTaddy via web service operations. The web service operations access the existing functionality and database of eTaddy. Workflow technology is then used to orchestrate the web service functionality. The workflow system communicates with the web service operations via SOAP messages. Using workflow technology saves the analyst time to concentrate on the interpretation of the eye movement data and visualizations.

To the best of our knowledge no similar systems exist, which use workflow technology to analyze eye movement data. Classical evaluation of eye movement data is done manually using different software tools. Therefore, the implementation of eTaddy is a first step to allow an automatic analysis of eye movement data using workflow technology.

The rest of this paper is structured as follows. In Section 2 we give an overview of the analysis of eye tracking studies and the eTaddy tool. In Section 3 we introduce the concept of workflows for eScience. In Section 4 we show how we extended the existing eye tracking analysis framework with a web service interface and in Section 5 we present two simple eye tracking workflows as a proof of concept. We discuss our work in Section 6 and give an overview about relevant related work in Section 7. The paper closes with a conclusion and future work in Section 8.

2 Eye Tracking

During the last decade, eye tracking has become one means to analyze user behavior in user studies. Eye tracking is used in different research areas such as neuromarketing, human-computer interaction, psychology, or visualization [Duc02]. Analyzing eye movement data can give further insight into where on a stimulus and at what objects of a stimulus participants looked. Eye tracking can be used besides analyzing accuracy rates of correctly given answers, or completion times. During an eye tracking study, fixations and saccades are recorded. Fixations usually last about 200 to 300 ms. Common metrics for fixations are the number of fixations (fixation count), the fixation duration in milliseconds, and the fixation position. A saccade describes a rapid eye movement which typically lasts about 30 to 80 ms. During this timespan visual information is suppressed. A sequence of fixations and saccades is called scanpath. If specific regions of a stimulus are of interest for an analysis of eye movement data, areas of interest (AOIs) can be defined [HNA+11].

Analyzing eye movement data can either be done statistically or using visualizations. Using statistical significance tests like Student's t-test or an analysis of variance (ANOVA) the evaluation is quantitative. A qualitative evaluation of eye movement data can be achieved by using visualization techniques. This allows to explore the eye movement data and find hypothesis which can later be analyzed statistically. Typical visualization techniques for eye tracking are attention map or scanpath visualizations. An attention map as shown

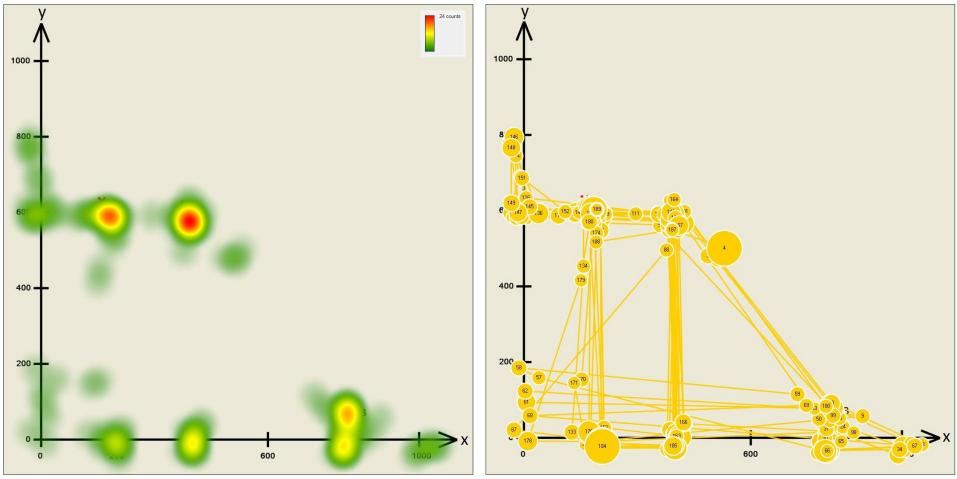


Figure 1: On the left, a classical attention map visualization is shown on top of a stimulus. The stimulus shows a synthetically generated coordinate system with arbitrary units where participants had to give the coordinates of different points. An attention map aggregates fixation data of all participants and represents them with a color coding. On the right, a scanpath visualization for one participant is shown. A scanpath represents fixations as circles, where the circle radius corresponds to the fixation duration. The saccades are represented by lines. Images created with Tobii Studio 2.2.8.[1]

in Figure 1 on the left is an aggregation of collected fixation data. The fixation position and duration of all participants is clustered and color coded. Green corresponds to few fixations and red to many fixations. An attention map allows a user to gain an overview about the eye movement data. On the right side of Figure 1 a scanpath visualization is shown for one participant. Each fixation is indicated by a circle where the radius of the circle corresponds to the fixation duration. Lines connecting fixations are saccades. This visualization technique allows to see the temporal order of eye movement data. However, it becomes cluttered easily if many fixations are shown or if scanpaths of multiple participants are shown.

The analysis process of eye movement data usually is time consuming. A complete analysis usually takes about a week to a month. Analyzing eye movement data of 30 participants, for three tasks with 30 stimuli would lead to 2,700 scanpath and over 10,000 fixations. Furthermore, the software for an eye tracking analysis is often distributed over multiple computers, i.e. the eye tracking computer and the computer of the analysts, and different software systems, i.e. the eye tracking software, the visualization tools, and the statistical tools. This means, that eye movement data has to be formatted to fit each software's requirements. If the domestic participant information or other data collected during the study is hand written, it first has to be transcribed into machine readable data. This has to be done manually by the analyst beforehand. Transcribing the data is error prone and the data has to be checked for those errors. The eye tracking data itself can be exported from the eye

[1] http://www.tobii.com/en/eye-tracking-research/global/products/software/tobii-studio-analysis-software/

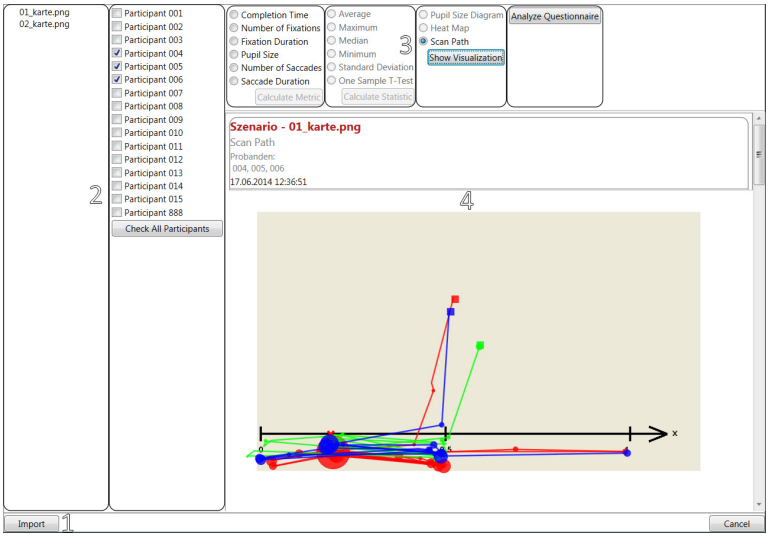


Figure 2: Eye Tracking Analysis, conDuction, and Designtool of user studYs (eTaddy) [BRE13] is a software to create eye tracking user studies, collect data, and analyze the eye movement data. For the analysis, the user has to import the eye movement data into a database (1). Afterwards, a stimulus, as well as participants can be chosen (2). Several eye tracking metrics, statistics, and visualization are available (3). The results of a calculation are shown in a timeline (4).

tracker automatically into a comma separated value file (csv). However, exporting one data set with 10,000 fixations takes about 15-30 minutes. Another issue is the reproducibility of an eye tracking analysis. All data generated during an analysis and all steps performed during the study have to be saved and logged carefully. This leads to even larger amounts of data which has to be stored and later analyzed.

To overcome some of these problems, in an earlier work we have implemented a software system called eTaddy (Eye Tracking Analysis, conDuction, and Designtool for user studYs) [BRE13]. eTaddy allows an analyst to design an eye tracking study and collect study data besides eye movement data. Furthermore, it supports an analyst by offering a single analysis tool for an evaluation of eye movement data. Figure 2 shows the analysis window of eTaddy. First, eye movement data has to be imported into a database (1). At the moment, eTaddy allows to import data from the eye tracking systems Tobii[2] and SMI[3]. Further parsers can be developed to include data from other eye tracking systems. The underlying data model is a general model for eye movement data and not dependent on a specific system. The domestic information is collected by eTaddy during the study process itself. After importing the eye movement data an analyst can choose a stimulus for investigation (2). Furthermore, multiple participants can be selected for an evaluation (2). eTaddy offers different eye tracking metrics, statistical calculations, and visualization techniques (3). An analyst can

[2]http://www.tobii.com/
[3]http://www.smivision.com/

calculate completion times, fixation counts, fixation durations, pupil sizes, saccade counts, and saccade durations. For a statistical evaluation, eTaddy offers to calculate the average, maximum, median, minimum, standard deviation, and Student's t-test. A pupil change diagram, an attention map, and a scanpath visualization can be displayed. This functionality is available in eTaddy using plug-ins allowing the user to easily add new functionality. Furthermore, a questionnaire with domestic information about participants is evaluated automatically. Every calculation or visualization technique is shown in the center part (4) of the window. Calculations are saved in the database and shown at the next start of eTaddy in chronological order. This allows an analyst to take up his evaluation at a later point in time.

The benefit of eTaddy is that it allows the user to evaluate eye movement data with one analysis tool. However, eTaddy still has some limitations. For example, each analysis step has to be manually initiated by the analyst. If a step has to be repeated several times this can become tedious and tiresome. Although, each calculation is saved in the database a reproducibility of the steps in a different eye tracking study is not possible. Therefore, a mechanism is needed that allows the user to define a general eye tracking analysis process which is then automatically initiated and executed.

3 Workflows for eScience

The usage of workflow technology for the automation of business processes is widely accepted in the domain of business applications [LR00]. Workflows are another core concept of service oriented architectures (SOA) [PG03] beside services, loosely coupled, and reusable components providing a unified interface. Workflows enable the orchestration of services to create new and potentially more complex functionality from existing functionalities in a user-friendly manner. This is achieved by modeling and the use of a workflow infrastructure that takes over deployment, control flow navigation, and interaction with the existing functions. Furthermore, workflow technology provides desirable non-functional properties like scalability, reusability, robustness, and sophisticated fault and exception handling.

In the last years, there is a notable movement towards workflow technology in the eScience domain [TDG06]. Here, we can distinguish two fundamental approaches. There are workflow systems especially developed for the eScience domain. As these systems typically have some deficits [TDG06], especially in the field of non-functional properties, a second approach is pursued. In this approach, conventional workflow technology, already used in business applications, is extended and adapted for the special needs of the eScience domain [GSK+11, SK10, SHK+11].

We have developed a scientific workflow management system (SimTech SWfMS) based on conventional workflow technology [SK10]. The workflow system was developed in the scope of SimTech - a cluster of Excellence and many related simulation technology projects.[4] We use the Business Process Execution Language (BPEL)[5] as a workflow

[4] http://www.simtech.uni-stuttgart.de/

[5] http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html

language. BPEL is a dominant industry standard for the orchestration of web services. To integrate existing legacy applications with BPEL, typically wrappers are developed, providing the functionality of these applications over a web service interface. The SimTech SWfMS has been successfully applied in different scenarios, i.e. for the automation of a Kinetic Monte-Carlo (KMC) simulation of solid bodies [SHK+11]. Through automation and particularly the ability of our workflow system to execute various instances of a workflow model in parallel, parameter studies can be carried out in a substantially greater scale and in much less time. In this way new research opportunities were opened for scientists.

Analysis of eye movement data is an iterative process, where results have to be interpreted and decisions have to be made based on these results. Usually, analysis steps are repeated during the evaluation. However, so far those analysis steps are not automated. To automate the analysis of eye movement data a workflow can be used modeling the ordered execution of the analysis steps. In the following, we show how we applied the workflow technology to the analysis of eye movement data.

4 Eye Tracking Web Service

To create an eye tracking workflow the individual analysis steps have to be offered as web service operations since web services enable reuse and seamless integration. The functionality for eye movement analysis in eTaddy is available using plug-ins. Therefore, a web service wrapper can be implemented that offers the functionality of each plug-in as one operation of the web service. During an eye tracking study a large amount of data is generated which has to be analyzed. This data is stored in a database that is used by eTaddy as well as the web service. Figure 3 shows the basic principle of the communication between eTaddy and the workflow system. The workflow system communicates via SOAP messages with the web service.[6] To reduce the size of the SOAP messages, the web service operations exchange references to data rather then the data itself whenever possible. Furthermore, the web service wrapper implementation accesses the plug-in functionality from eTaddy.

Figure 4 shows an example communication with a web service. At first a SOAP message is sent to the web service (1). This message contains information about the target operation as well as the input data for this operation. Each plug-in in eTaddy requires one stimulus and at least one participant. Therefore, the stimulus and a list of participants are sent in the message. The wrapper implementation then translates the SOAP message into a C# object (2). The incoming message does not contain the input data but a reference to this data. Thus, the relevant data has to be fetched from the database (4). Afterwards, the plug-in function can be called (5) by the wrapper implementation (3). The result from the calculation is then serialized into XML code again (6) and send as a SOAP response message back (7). In Section 5 we show how the web service described so far can be used by workflows for the automation of eye tracking analysis.

[6] http://www.w3.org/TR/soap12-part1/

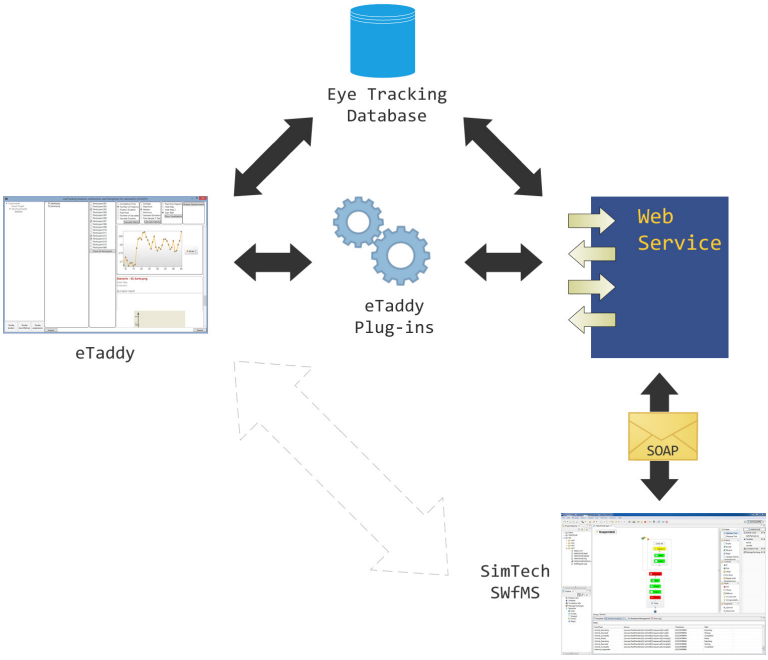


Figure 3: The functionality of eTaddy, namely the metrics, statistics, and visualizations, can be accessed through plug-ins. The eye movement data is saved in a database. A web service wrapper implementation exchanges references to the data in the database and communicates via SOAP messages with the workflow system where the eye movement analysis is modeled.

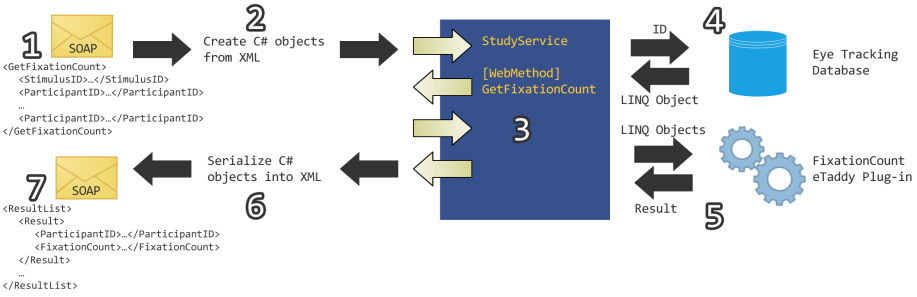


Figure 4: A workflow sends a request as a SOAP message containing the function and the necessary parameters. The parameters are cast into a C# object and sent to the web service. The web service retrieves the appropriate data from the database and calls the function. The result is then serialized into XML and sent back as a SOAP response message to the workflow.

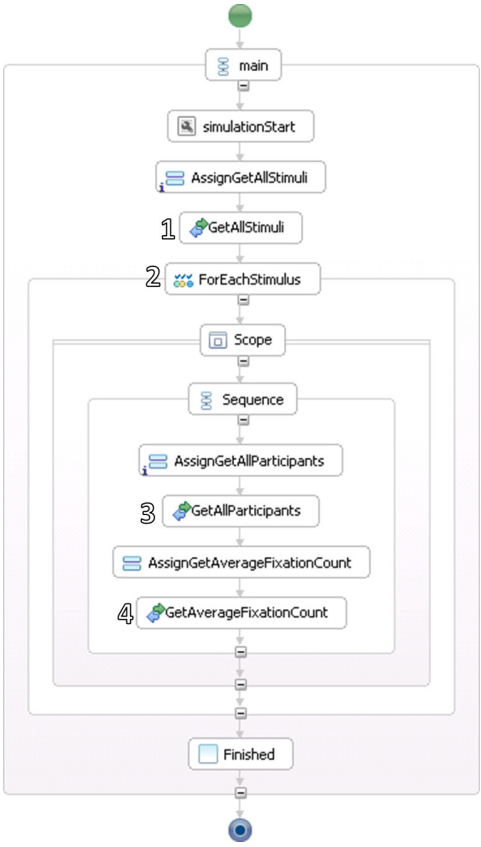


Figure 5: Example workflow calculating the average fixation duration of a given list of stimuli and a given list of participants. First, all required stimuli are fetched from the database (1), then for each stimulus (2) the list of participants is retrieved from the database (3), and then for each participant the average fixation count is calculated (4).

5 Eye Tracking Analysis Workflows

To demonstrate how our web service system can be used in practice we have implemented two simple workflows representing two examples for an eye tracking analysis and a proof of concept. The first workflow calculates the average fixation duration for each participant. The second workflow creates two attention map visualizations for two participant groups in parallel.

5.1 Average Fixation Duration

The first workflow was created to show how a functionality that is already available in eTaddy can be modeled as a workflow. The workflow shown in Figure 5 calculates the average fixation duration for all stimuli and each participant. First, the workflow calls the web service to retrieve all stimuli from the database (1). In the following loop (2), the workflow calls the web service to retrieve for each stimulus all participants from the database (3). Then, the workflow calls the web service operation which invokes the plug-in of eTaddy to calculate the average fixation count for all of these participants (4). Instead of performing several actions manually, the workflow allows to execute these actions in an automated manner. This process would be more difficult in eTaddy as each stimulus would have to be selected separately by the analyst. The implementation with a workflow allows to speed up the calculation process and to reduce errors.

5.2 Participant Group Attention Map

The second workflow shown in Figure 6 creates an attention map for two different participant groups. The participant groups are created based on the age of the participants. At first, the user has to select a stimulus. This is achieved by using a separate window, where the user can input the ID of a stimulus. Then, the workflow selects the web service to retrieve all participants from the database (1). Afterwards, the workflow uses a flow activity which describes the parallel execution of two sequences (2). In the left sequence, the workflow uses a data manipulation activity to select all participants younger than 22 years from the participant set (3a). Then the workflow calls the web service which invokes the eTaddy plug-in to calculate the attention map for this participant group (4a). Similarly, in the right sequence, the workflow selects all participants older than 22 years from the participant set (3b). Then, the workflow calls the web service which invokes the eTaddy plug-in to calculate the attention map for the second participant group (4b). This is a functionality which is not available in eTaddy. eTaddy does not offer the option to automatically group participants based on their domestic information. Furthermore, the calculation of the attention maps is done in parallel, leading to a shorter time span to receive results. This example also shows that it is easier to create new functionality for additional types of analysis faster, if workflows are used.

6 Discussion

Analyzing eye movement data using workflow technology allows the analyst to concentrate on the interpretation of data and visualizations. The analyst does not have to bother deciding which participants to group or selecting a calculation for each stimulus and selected participants. On the one hand, this allows the analyst to save time. On the other hand, the available workflow can later be executed again either with the same or different

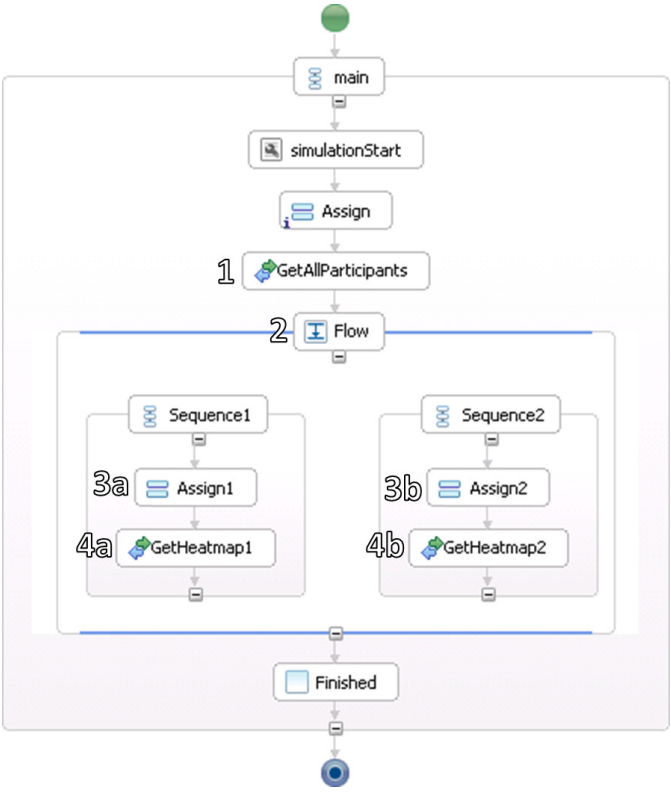


Figure 6: Example workflow calculating two attention maps for two groups of participants. First, all participants are retrieved from the database (1), then the attention maps for the two groups are created in parallel (2). On the left, the participants younger than 22 are used (3a) and then the attention map for this group is calculated (4a). Parallel to this, the participants older than 22 are used (3b) to create the second attention map (4b).

eye movement data. Furthermore, a workflow can be shared with other researchers and revised if necessary to the needs of other researchers in a short time. However, a complete automation of eye movement analysis is not possible without human interaction. Often, the analyst has to decide on a next step which does not depend on selected parameters, for example if an attention map or a scanpath is generated. Furthermore, the appropriate workflow for an eye movement analysis depends on the hypothesis to be verified. Therefore, the user needs to be able to influence the workflow behavior and decide on the next steps to be executed. In the current implementation this lack of integrating a human user is one limitation which will have to be tackled in future work.

7 Related Work

In Jäger et al. [JAZ+05] a similar approach like the one described in our paper has been followed to improve the analysis of geospatial data. The data, available in several heterogeneous data sources, has been made accessible over web service interfaces. Then, the Kepler scientific workflow system [ABJ+04] was used to model and execute geospatial data analysis processes. In contrast to our work, the main challenge was the heterogeneity of data sources. Whereas in our work we focus on the automation as a means to avoid errors and speed up the analysis process.

In Cummings et al. [CPP+08] a scientific workflow system is used to automate complex edge plasma simulation processes. The authors first provide arguments, why the manual or script based execution of such simulations is not feasible or possible. Then, they describe their solution to model the simulation process as a scientific workflow. Afterwards, they describe the automatic execution with a scientific workflow system. The analysis of eye tracking data is far less complex and can be executed by hand. However, the general benefits of the scientific workflow approach are very similar, although we apply it in a totally different domain. Automating simulation and analysis processes provides significant time savings, full control of the execution, and the ability to repeat exactly the same process at any time. Furthermore, to the best of our knowledge there currently exists no approaches for analyzing eye movement data automatically.

8 Conclusion and Future Work

Analyzing eye movement data using a workflow system is a time saver and allows the user to concentrate on data and visualization interpretation. In this paper, we have contributed an option to extend the functionality of the existing eye movement analysis framework eTaddy. We added a web service interface to eTaddy and then orchestrated the provided operations using workflow technology. The operations of the web service provide access to the existing functionality and database of eTaddy. Furthermore, we have demonstrated how possible workflows for eye movement analysis can look like. We used workflows to automate processes that had to be executed manually before. We have also shown that workflows can provide new functionality that was not available in the original eye tracking analysis system. Although, new functionality is available for the analysis using the workflow system so far the analyst has no option to interact with the workflow. This will be part of future work. We want to extend our workflows to include a human component. The analyst can then decide at specific parts of the workflow what next steps to execute. Furthermore, more complex analysis tasks will be modeled using the workflow technology.

References

[ABJ+04] I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. Kepler: An extensible system for design and execution of scientific workflows. In *Proceedings of the 16th International Conference on Scientific and Statistical Database Management*, pages 423–424, 2004.

[BRE13] T. Blascheck, M. Raschke, and T. Ertl. eTaddy - Ein integratives Framework für die Erstellung, Durchführung und Analyse von Eyetracking-Daten. In Gesellschaft für Informatik, editor, *Lecture Notes in Informatics (LNI) - Seminars*, volume S-12, pages 111–114. Köllen Druck+Verlag GmbH, 2013.

[CPP+08] J. Cummings, A. Pankin, N. Podhosrzki, G. Park, S. Ku, R. Barreto, S. Klasky, C. S. Chang, H. Strauss, L. Sugiyama, P. Snyder, D. Pearlstein, B. Ludäscher, G. Bateman, and A Kritz. Plasma edge kinetic-MHD modeling in tokamaks using Kepler workflow for code coupling, data management and visualization. *Communications in Computational Physics*, 4(3):675–702, 2008.

[Duc02] A.T. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Method, Instruments, & Computers*, 34(4):455–470, 2002.

[GSK+11] K. Görlach, M. Sonntag, D. Karastoyanova, F. Leymann, and M. Reiter. Conventional workflow technology for scientific simulation. In X. Yang, L. Wang, and W. Jie, editors, *Guide to e-Science*, Computer Communications and Networks, pages 323–352. Springer London, 2011.

[HNA+11] K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 1 edition, 2011.

[JAZ+05] E. Jaeger, I. Altintas, J. Zhang, B. Ludäscher, D. Pennington, and W. Michener. A scientific workflow approach to distributed geospatial data processing using web services. In *Proccedings of the 26th International Conference on Scientific and Statistical Database Management*, pages 87–90, 2005.

[LR00] F. Leymann and D. Roller. *Production workflow - concepts and techniques*. PTR Prentice Hall, 2000.

[PG03] M.P. Papazoglou and D. Georgakopoulos. Service-oriented computing: concepts, characteristics and directions. In *Proceedings of WISE*, 2003.

[SHK+11] M. Sonntag, S. Hotta, D. Karastoyanova, D. Molnar, and S. Schmauder. Using services and service compositions to enable the distributed execution of legacy simulation applications. In W. Abramowicz, I.M. Llorente, M. Surridge, A. Zisman, and J. Vayssière, editors, *Towards a Service-Based Internet*, volume 6994 of *Lecture Notes in Computer Science*, pages 242–253. Springer Berlin Heidelberg, 2011.

[SK10] M. Sonntag and D. Karastoyanova. Next generation interactive scientific experimenting based on the workflow technology. In R.S. Alhajj, V.C.M. Leung, M. Saif, and R. Thring, editors, *Proceedings of the 21st IASTED International Conference on Modelling and Simulation (MS 2010)*, 2010.

[TDG06] I. J. Taylor, E. Deelman, and D. B. Gannon. *Workflows for e-Science: Scientific workflows for grids*. Springer New York, 2006.

Data Analytics for Simulation Repositories in Industry

Rodrigo Iza Teran[1] and Jochen Garcke[1,2]

[1]Fraunhofer-Institut SCAI
Numerical Data-Driven Prediction
Schloss Birlinghoven
53754 Sankt Augustin
rodrigo.iza-teran@scai.fraunhofer.de

[2]Universität Bonn
Institute for Numerical Simulation
Wegelerstr. 6
53115 Bonn, Germany
garcke@ins.uni-bonn.de

Abstract: Simulations are used intensively in the developing process of new industrial products and have achieved a high degree of detail. In that workflow often thousand finite element model variants, representing different product configurations, are simulated within a few days or even overnight. Currently the decision process for finding the optimal product parameters involves the comparative evaluation of large finite element simulation bundles by post-processing each one of those results using 3D visualisation software. This time consuming process creates a severe bottleneck in the product design and evaluation workflow.

To handle these data we investigate an analysis approach based on nonlinear dimensionality reduction to find a low dimensional parameterisation of the dataset. In such a reduced representation, similar model variants are organised in clusters and the influence of the input variables can be analysed along such a parameterisation. We demonstrate the application of this approach to a realistic and relevant industrial example for robustness analysis of the bumper location in a frontal crash simulation. The approach has the potential to considerably speed up the virtual product development process by allowing an intuitive and interactive simultaneous evaluation of many product designs.

1 Introduction

Finite element simulations of physical behaviour are an essential tool in the development and improvement of industrial products for many years now [Mey07]. Here, an approximate representation of the object by a mesh is employed and, based on a mathematical model of the physical process, a numerical simulation is performed. For an accurate description such numerical approximations typically need to be very fine, in the order of million of unknowns in our case, with larger simulations needed for some other applications. The size of the resulting numerical data are huge, one simulation in time generates number of unknowns multiplied by the number of (saved) time steps numerical values. Furthermore, during the research and development process engineers easily generate several hundred variants of a specific finite element model that simulates different operating conditions of the product, multiplying the data once more. Post-processing software tools are readily available to display the 3D geometrical information of such a model and the results of one numerical simulation. Engineers analyse and compare the different simulations using their own engineering knowledge, although this is generally limited to the si-

multaneous analysis of only a few simulations at a time. The complex structure of the data and its sheer size, the required 3D visualisation of the geometry and the needed inspection of the associated design variables of each configuration prohibits a detailed comparative analysis of more than a few simulations by hand. There is need for a more efficient product development process which overcomes the current limitations.

The best known approach to tackle this data analysis challenge are Principal Component Analysis (PCA) to identify variation modes and its counterpart classical Multidimensional Scaling (CMDS) for finding a low dimensional embedding. For several thousand simulations, corresponding to a specific product development phase, the PCA can recover the principal variations in just a few components. The approach has already been successfully used for the analysis of numerical simulation data, see [AGHH08, MT08, TNNC10]. But in spite of their success so far, there are many situations where PCA or CMDS are not adequate. Being linear methodologies, they cannot cope with a presence of nonlinear correlations in the data, for example when dealing with time dependent highly nonlinear deformations. Approaches that attempt to overcome these limitations have been proposed in recent years, see [BGIT+13, Iza14]. In line of this research we use a nonlinear dimensionality reduction method called diffusion maps, which is a machine learning method known to be able to recover the so called intrinsic geometrical structure of the data in the presence of nonlinearities. Assuming the intrinsic geometry is low dimensional, recovering such nonlinear structures can be very effective for analysing a number of simulations simultaneously. The contribution of this work in comparison to [BGIT+13, Iza14] consists in constructing and using a realistic large scale industrial scenario to proof the effectiveness of nonlinear dimensionality reduction methods in the virtual product development process.

In section 2 we will introduce the applied data workflow, followed by section 3 about the essential basics of nonlinear dimensionality reduction. The application of the approach for the crash simulation bundle is presented in section 4. We finally give an overview of the potential application of the methodology and discuss also further efforts in this area.

2 A Workflow for Analysing Numerical Simulation Data

We describe a general workflow that allows for the analysis of high dimensional data from bundles of large finite element simulations. Four steps are involved in it. The first two steps in the analysis workflow involve costly data intensive computations. Preferably these take place in a parallel server infrastructure where the bulky data is stored, therefore avoiding transfer of the big data and exploiting the parallel HPC resources.

- Extraction - the raw data for the analysis are obtained directly from the simulation, these variables can be of different types such as scalars, vectors or tensors defined on nodes or elements of a finite element mesh.
- Preprocessing - this step is usually necessary to cope with the huge data size (millions of nodes and elements) and its complex nature. A usual preprocessing step

consists in the use of only subsets of the datasets in areas of interest (for example the supporting frontal beams of a car in a frontal crash), also sub-sampling and clustering can be employed [BGIT+13]. The data can also be transformed in different ways in order to obtain a compact representation; for example using principal component analysis (PCA).

- Dimensionality Reduction - in this step a low dimensional representation is obtained from the dataset that parameterises the information. We use here diffusion maps as a dimensionality reduction framework, but other dimensionality reduction methods can be used as well.

- Exploration - based on the low dimensional representation found after the dimensionality reduction. The simulation variables are organised in the low dimensional embedding space, i.e. each numerical simulation is represented as a point in the obtained low dimensional space. Due to the reduced dimensionality, the datasets can be efficiently explored.

3 Nonlinear Dimensionality Reduction with Diffusion Maps

Nonlinear dimensionality reduction is a very active area of research in machine learning in recent years, often known as manifold learning to emphasise the actual objective of such methods, namely the identification of low dimensional structures to represent high-dimensional information. Several methods have been introduced and we refer to [LV07, Wan11] for details about each of this methods. In this work we concentrate on so called kernel methods, which are based on the construction of a similarity matrix with coefficients calculated using a kernel function of the type $e^{-d(x,y)}$, where $x \in R^n$ and n can be very large (e.g. up to the total number of nodes of the mesh in our application). A specific distance function d can be used in the method like Euclidean or Geodesic (approximated by the graph distance on the mesh for our application). After building the similarity matrix a SVD decomposition is used to extract the eigenvectors corresponding to the largest eigenvalues. In this research work we use a specific variant of a kernel method called diffusion map, for details about the method see [LL06, CL06, Wan11]; for its use for the analysis of numerical simulation data see [Iza14].

Notice that in our analysis we use simulation data on a finite element mesh directly. A finite element mesh contains nodes and elements and we assume that the mesh connectivity is the same and use only the values defined at the nodes on the mesh as data set. If the mesh connectivity is different a reference mesh can be used to map the simulation values from the slightly different numerical simulation meshes to it.

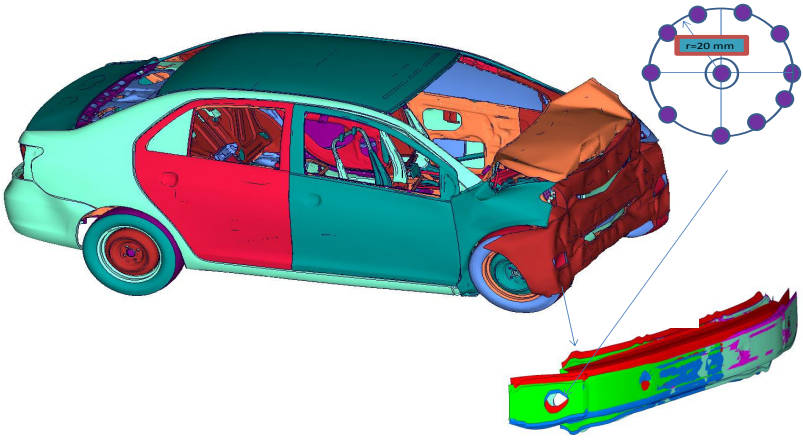


Figure 1: Variations of the bumper position along a circle.

4 Robustness Analysis in Frontal Crash

For this case study we investigate a current Toyota model with 998,000 nodes and corresponding elements, which is publically available from the NCAC[1]. As a realistic and difficult engineering problem, the position of the bumper was changed along a circle, see Figure 1, which results in observably different crash behaviour during a frontal impact with a barrier. A total of $M = 243$ numerical simulations where performed using LS-DYNA[2] and 26 intermediate time steps were saved for each run. The training dataset is saved in files in a binary format. So called post-processor software, can be used to read this data and extract (parts of) it. We use the software Animator[3] to extract specific components of a car or structure. The components we choose are the ones that are critical for the engineer during the analysis of the structural behaviour under crash (see Figure 2). Note that in [BGIT+13, Iza14] car crash data was analysed stemming from a simpler model from that repository, with more than one order of magnitude less grid points, to study the general analysis procedure. In this work we analyse simulation data in a size currently being used from a difficult engineering task, and further proceed to an interpretation of the results from an engineering perspective.

We apply the workflow explained in section 2 to all simulations, and extract in the preprocessing step as relevant structural parts the firewall and the structural beams, see Figure 2. We perform the data analysis at time step 14, when most of the crash affecting the frontal structure took place. For the entries of the feature vector we employ for each mesh point the distance to the corresponding mesh point of a chosen reference model, this gives a vector in $\mathbb{R}^n$, where n is the total number of nodes of the selected parts. The dimension-

[1]Finite Element Model Archive, `http://www.ncac.gwu.edu/vml/models.htm`
[2]Livermore Software Technology `www.lstc.com/products/ls-dyna`
[3]GNS mbH `http://gns-mbh.com/animator.html`

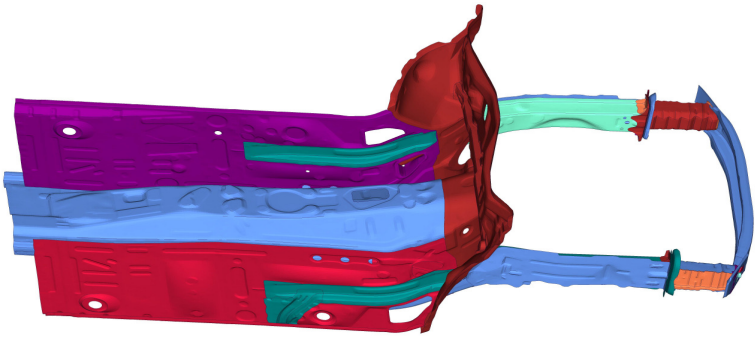


Figure 2: Selected structural parts relevant for the engineering analysis.

ality reduction is then performed with these feature vectors. For the exploration step of the workflow, each of the simulations is now represented as a point in a 3D plot which therefore gives a parametric representation of the 243 simulations in 3D.

To have a comparison with a standard linear approach we first performed a CMDS analysis for the same dataset using the first spectral coordinates as embedding for the exploration step, which is shown in Figure 4(a). Although simulations using a similar angle are arranged nearby, no distinct structures can be identified.

In Figure 3 the low dimensional embedding with diffusion maps is shown. It can clearly be seen that the simulations are organised according to a certain type of deformation mode. The colour of the deformed structure corresponds to the difference to the chosen reference model and the colour of the points in the 3D plot corresponds to a specific angle for the bumper location in degrees.

From the embedding plot for diffusion maps in Figure 3, it can be seen that at least three deformation modes can be associated to a range of positioning angles of the bumper. We extracted for clarity a beam from each mode in order to show the typical deformation in each mode and the angle dependence, see Figure 4.

The computational time for data loading and preprocessing in an off-line-phase takes less than an hour, which is mostly due to the data extraction, and needs to take place once for each data set. The actual data exploration afterwards can be done interactively. The selection of the corresponding simulations and the generation of the 3D pictures shown in Figure 4 took less than an hour. In comparison an engineer in charge of this project normally needs to analyse each of the deformation modes by hand and then classify them. A task that for that amount of simulations can require several days, if not weeks.

5 Overview and Perspectives

We studied a methodology for analyzing finite element simulations arising in the virtual product development process on a specific realistic industrial application. We could show

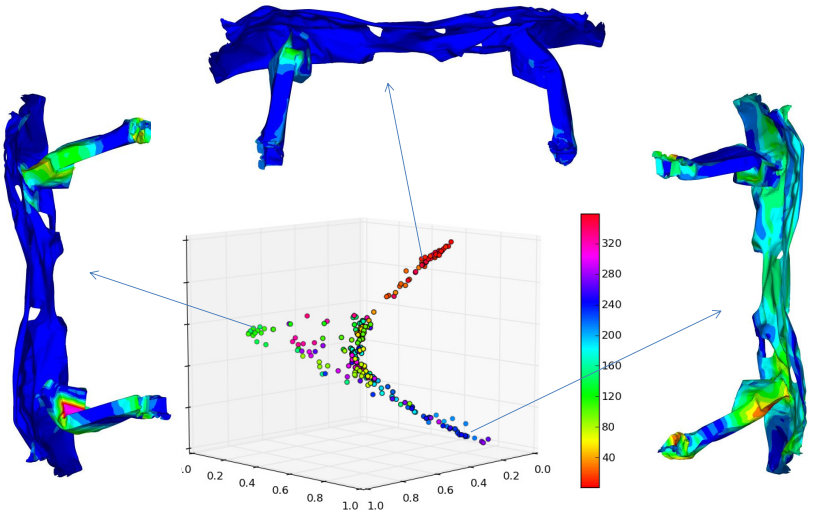
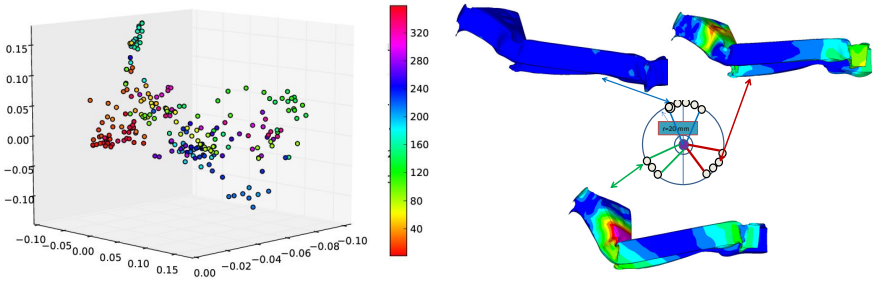


Figure 3: Embedding obtained by diffusion maps. Each simulation run is colour coded with the angle of the bumper variation.



(a) CMDS embedding (b) deformation modes

Figure 4: We show the CMDS embedding, where each simulation run is colour coded with the angle of the bumper variation, and the bumper deformation modes derived from the diffusion maps embedding from Figure 3.

that organising the data along low dimensional structures allows the simultaneous analysis of many simulations. A realistic example in crash simulation, whereby the position of a bumper is changed, shows that different deformation modes in the structural beams of the car can be easily identified in the low dimensional embedding. Using nonlinear techniques reduces the complexity and time for investigating such large bundles of huge numerical simulation data.

To deploy such an approach for the analysis of large scale simulation data in real-life environments, it will need to handle efficient data storage (including compression) which

will take place on data servers in the future instead of workstations as nowadays is the practice in industry. It will also need efficient transfer of the (relevant) data between server and client, as well as efficient data processing for analysis and visualisation procedures, like the one outlined in this work, on the server and the client.

Acknowledgements

The variations of the bumper positions were designed by Claudius Schoene of GNS mbH in the course of the KMU-innovativ project FEMMINER supported by the BMBF.

References

[AGHH08] S. Ackermann, L. Gaul, M. Hanss, and T. Hambrecht. PCA for Detection of Globally Important Input Parameters in Nonlinear Finite Element Analysis. In *Weimar Optimization and Stochastic Days 5.0*, 2008.

[BGIT+13] B. Bohn, J Garcke, R. Iza Teran, A. Paprotny, B. Peherstorfer, U. Schepsmeier, and C.-A. Thole. Analysis of car crash simulation data with nonlinear machine learning methods. In *Proceedings of the ICCS 2013*, volume 18 of *Procedia Computer Science*, pages 621–630. Elsevier, 2013.

[CL06] R. Coifman and S. Lafon. Diffusion Maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[Iza14] R. Iza Teran. Enabling the Analysis of Finite Element Simulation Bundles. *Internat. Jour. for Uncertainty Quantification*, 4(2):95–110, 2014.

[LL06] S. Lafon and A. B. Lee. Diffusion Maps and Coarse-Graining: A Unified Framework for Dimensionality Reduction, Graph Partitioning, and Data Set Parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1393–1403, 2006.

[LV07] J. Lee and M Verleysen. *Nonlinear Dimensionality Reduction*. Springer, 2007.

[Mey07] M. Meywerk. *CAE-Methoden in der Fahrzeugtechnik*. Springer, 2007.

[MT08] L. Mei and C.-A. Thole. Data analysis for parallel car-crash simulation results and model optimization. *Sim. Modelling Practice and Theory*, 16(3):329–337, 2008.

[TNNC10] C.-A. Thole, L. Nikitina, I. Nikitin, and T. Clees. Advanced mode analysis for crash simulation results. In *Proc. 9th LS-DYNA Forum*, 2010.

[Wan11] J. Wang. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer, 2011.

Big Data in Mobility and Logistics

(BDMobiLog 2014)

Big Data in Mobility and Logistics (BDMobiLog 2014)[1]

Abstract: The growing technical possibilities to gather data, e.g. sensors, mobile devices, social media, log files, cameras and microphones, have resulted in large and complex data sets, which today are known as Big Data. But which potential does Big Data hold for the domains mobility and logistics? Mobility and logistics are closely linked to each other. On one hand, traffic flows and traffic density have a strong influence on the efficiency of logistic processes (especially when it comes to transportation). On the other hand, the trend to e-mobility poses new challenges to the planning processes, e.g. the consideration of battery capacity during planning. Together with other factors, this leads to a rising complexity of logistic processes and networks, as well as a big amount of data to be handled. That is why there is already a growing amount of IT-based applications, which support and automate different steps of the logistic processes. Using the right analytic approaches Big Data can elevate these applications to a new level, improve the planning processes, and enable precise and well-founded decisions. This workshop brings together researchers and practitioners working on Big Data technologies in mobility and logistics. We accepted three submissions for presentation, which identify the potentials of Big Data in logistics and describe Big Data-based approaches in technical customer management as well as in public transportation.

Organizing Committee: Natalja Kleiner (FZI Forschungszentrum Informatik, Karlsruhe), Thomas Setzer (FZI Forschungszentrum Informatik, Karlsruhe), Hansjörg Fromm (Karlsruhe Institute of Technology), Kai Furmans (Karlsruhe Institute of Technology), Anne Meyer (FZI Forschungszentrum Informatik, Karlsruhe)

Program Committee: Andreas Abecker (disy GmbH, Karlsruhe), Boris Amberg (FU Berlin), Matthias Bender (FZI Forschungszentrum Informatik, Karlsruhe), Martin Birkmeier (FIR at the RWTH Aachen University), Felix Brandt (FZI Forschungszentrum Informatik, Karlsruhe), Simone Braun (CAS Software AG, Karlsruhe), Andreas Cardeneo (SAP Deutschland AG & Co.KG, Walldorf), Iris Heckmann (FZI Forschungszentrum Informatik, Karlsruhe), Mark Hefke (CAS Software AG, Karlsruhe), Niklas Hering (FIR at the RWTH Aachen University), Marcel Huschebeck (PTV Group, Karlsruhe), Peter Korevaar (IBM Deutschland GmbH), Marco Lewandowski (LogDynamics Lab, Bremen), Peter C. Lockemann (FZI Forschungszentrum Informatik, Karlsruhe), Jan Necil (FZI Forschungszentrum Informatik, Karlsruhe), Stefan Nickel (Karlsruhe Institute of Technology), Jens Nimis (HS Karlsruhe), Boris Otto (IML/TU Dortmund), Roland Schmidt (Robert Bosch GmbH, Reutlingen), Nenad Stojanovic (FZI Forschungszentrum Informatik, Karlsruhe), Jacqueline Wirnitzer (FZI Forschungszentrum Informatik, Karlsruhe), Katrin Zeiler (DHL Customer Solutions & Innovation, Troisdorf)

[1] http://www.fzi.de/en/aktuelles/bdmobilog-2014

Big Data in Logistics – Identifying Potentials through Literature, Case Study and Expert Interview Analyses

Volker Frehe, Thorsten Kleinschmidt, Frank Teuteberg

Osnabrück University
Accounting and Information Systems
Katharinenstr. 1
49074 Osnabrück
volker.frehe@uni-osnabrueck.de
tkleinschmid@uni-osnabrueck.de
frank.teuteberg@uni-osnabrueck.de

Abstract: In this contribution, we elaborate the current state of research and practice of Big Data in the field of logistics by means of a systematic literature review, a case study analysis and expert interviews. Although all interviewees are from Germany, viable perspectives and opinions of practitioners from worldwide operating companies were gained. Based on the analyzed information and the identified knowledge gaps, we developed implications for practice and for research. We call for an advanced interdisciplinary research, which integrate practitioners as early as possible. Practitioners should identify how Big Data can improve management decision or daily business. Management support was identified as essential in Big Data projects, besides department staff should be integrated and a holistic approach should be followed. Therefore, appropriate training for project members or hiring of new staff is needed. Thus, this paper offers fundamental new insights in the field of Big Data useful for practitioners and researchers.

1 Introduction and Motivation

The topic Big Data can be seen as a recent trend in the field of Information Systems (IS), equally significant for practitioners and academics [CCS12]. A survey of 100 senior executives of Fortune 1,000 companies revealed that more than 70% of the participants want to improve the "time to answer" through Big Data [Fer13]. In scientific literature, [CCS12] have identified a continuing trend in the past decade (2000-2011). In 2000, there were no publications on Big Data, in 2011, there were already 95. The amount of data to be stored is getting bigger, also due to data creation through the web, mobile devices and sensors [CCS12, Gob13]. Irrespective of the industry sector, the increasing data volume allows companies advantageous insights in their businesses [CCS12, Sch13]. The data acquisition is no problem anymore; the issue is rather to make the best possible use of the data for operational and strategic business [Gob13].

[CCS12] conducted a literature review of 3,602 articles as of 2000 in the area of Business Intelligence & Analytics (BI&A) and Big Data. One of the identified Top

Keywords was “supply chain”, the subject of logistics, however, did not emerge. A lot of data rises from sensors, telematics, GPS, etc., (techniques used in the logistics industry) which is why we consider the fields of freight transport and inventory management to have a high potential. This hypothesis is also supported by [SK13], who inter alia identify two Big Data topics with a high impact on logistics in a period of less than five years (The analysis of data in real-time to provide value-adding information to the costumer and the analysis of previously unused data sources).

This information leads us to the following research questions (RQ):

RQ1: What is the current scientific status quo of the topic Big Data in logistics?
RQ2: Which use cases of Big Data are already implemented in the area of logistics?
RQ3: Is there any potential for the use of Big Data in new logistics scenarios?

To answer RQ1, we performed a literature review, while RQ2 was investigated through a case study analysis. To examine RQ3 we carried out four expert interviews. The paper is structured as follows. In the following section, we present our research methods. Afterwards, we briefly describe the results of the literature review, the case study analysis and the expert interviews. Then, we combine and discuss the results. Finally, we make recommendations and implications for research and practice.

2 Research Methods and Review Scope

2.1 Systematic Literature Review and Analyzed Articles

Since it is recognized as key instrument for the efficient analysis of large numbers of different sources [WW02], we used the systematic literature review as research method, thereby following the steps established by [WW02] and [Dib04].

First, we defined the review scope. We used the EBSCO literature database (including EconLit) to cover peer-reviewed business journals as well as industry/trade magazines. To get a broader base of literature, we also performed a google scholar search. Furthermore, in order to consider high quality papers, we searched for all A-ranked journals (not covered by EBSCO) in the VHB jourqual logistics ranking.

The research topic has already partially been described in the Motivation section, which led to the use of the keywords “logistics”, “transport” and “warehouse” in combination with “Big Data”.

In EBSCO, we used the phrase <"Big Data" AND (Logistic* OR Transport* OR Warehouse*)> and searched in title, subject terms and abstract. This resulted in 44 articles, of which only five were published in academic journals. In google scholar, we used the phrase <allintitle: Logistics OR Transport OR Warehouse "Big Data">, which

yielded only six articles[1]. Two VHB A-ranked journals were not covered by EBSCO. By using the term <Big Data> we identified another three articles in these journals. For all identified academic papers we additionally performed a forward and backward search, which led to another 10 articles. We checked all papers for relevance by reading the abstract and thus received 28 relevant contributions.

None of the identified papers contains a systematic literature review of Big Data in the area of logistics. No related work could be identified. Thus, to the best of our knowledge, our contribution is the first to provide such insights.

According to the phases of [Dib04], the literature analysis and conclusion will be performed in the following sections.

2.2 Case Study Research

In addition, we aimed at capturing the status quo from practitioners. Therefore, we performed a case study research; a method used to examine phenomena in their natural context [My09].
We used Google for a broad, open search, applied the search term <"Big Data" "logistics" "case study"> and thus obtained more than 119,000 results. We successively went through the results and yielded 36 case studies on Big Data and logistics from the first 70 results. As the hit ratio then decreased significantly (we did not identify one relevant case study in the results 71 to 100) we terminated our search.
Overall, we identified 36 case studies and checked these for relevance. Case studies that only offered technical descriptions with no connection to business processes have not been considered further. We also excluded extremely superficial case studies that were rather promotion. The finally selected 14 case studies are listed in table 3.

2.3 Expert Interview

To get an overview of the potential use of Big Data in the area of logistics, we accomplished expert interviews, which are the most important technique for qualitative research [My09]. We applied the recommendations for construction, justification, ethical issues and tensions for interpretive research by [Wal06] and the principles of [My09] to apply a well-established research method.
We invited more than 150 German companies, which either come from the logistics industry or provide Big Data solutions, to participate in the interview. The return rate was below 16 % and most of the answers were refusals (due to a lack of time/resources or knowledge or because of corporate policies).
Lastly, five interviewees from three companies have been interviewed in four semi-structured expert interviews (cf. table 1).

[1] We limited the search to title, as we only want to identify paper with high relevance to the topic. As google scholar provides access to a huge database, which contains also grey literature, we have to focus on the high relevant one paper. An open google scholar search reveals over 11,000 papers.

#	Interviewee position	Company's activity	Employees	Area of operation
1	Director of eFullfilment	Stationary trade logistics	~ 5,000	> 40 countries
2	Project Manager for Information Management	International contract logistics	> 11,000	18 countries
3	**3.1:** Head of IS Governance & Compliance	Logistics services and freight transport	~ 11,000	> 50 countries
	3.2: Member team Information Management			
4	Global Lead of Business Productions Solutions			

Table 1: Overview of Interviewees

The interview guide consists of open questions and supplemental questions to discuss interesting topics in detail. First some organizational topics were explained and an introduction (especially the sentence „Big Data") was given, to achieve a common understanding) was given. Afterwards the interviewee was asked how Big Data is used in his firm (which data is analyzed, where does the data rise and which data is persisted, where is a Big Data analysis useful and how does Big Data develop in the next 3-56 years). Then a case study (use of telematics to reduce downtime in freight transport) was presented and the interviewee was asked to evaluate and discuss this case study. Before saying goodbye, there was open space to discuss unasked questions. The guide was sent to the experts at least one day before the interview to enable sufficient preparation. The interviews, which took place in the timeframe from 2013-08-01 to 2013-11-29, were recorded and transcribed.

3 Analysis of Results

3.1 Literature Review Results

To start with, we clustered the reviewed papers in different categories. As we did not limit our search to one special field of logistics, different topics are discussed (cf. table 2).

Topic	Paper handle topic
Supply Chain	[Big13], [WF13b], [WF13a], [Wil13], [Bie13], [Lu13], [Sav12], [Bru12], [Cec12], [MGI11], [Cha13], [Lap13]
Transportation (public and freight)	[Mil13], [Bie13], [Cob13], [Sco14], [Bur13], [MGI11], [McC13], [Ali12], [Dai12], [Cow13]
Sales	[Gro14], [Rit14], [Hol14], [Sav12], [Bru12]
Manufacturing/Production	[LK12], [Gro14], [Byr12], [MGI11]
Inventory Management / Warehousing	[Tre13], [MGI11]
not specified	[Ola13], [CCS12]
Procurement	[Gro14]

Table 2: Overview of Big Data logistics topics

The most frequent topic is Supply Chain (SC), which is treated in one third of the papers. The second major topic is Transportation (freight and public transport).
As the data volume increases with the number of parties involved, it is not surprising that Big Data plays a major role in SC. As already mentioned, also the transportation industry has, due to the huge amount of data that can be generated (e.g., through sensors, GPS, mobile devices), great concerns about Big Data. Surprisingly, the Inventory Management (or Warehousing) industry is mentioned in two papers only, despite technologies (like Barcode and RFID) that may generate large data volumes. However, Big Data research is currently under-represented in this industry.

The papers on SC differ substantially. Some papers only explain to some degree why Big Data is important for Supply Chain Management (SCM) [HTZ13] or give recommendations how to start Big Data [Big13]. Other papers explain the connection of Big Data and predictive analysis [WF13a] and clarify how Big Data can be used to ameliorate the quality of predictions [WF13b]. [Wil13] identifies an obstacle that ought to be overcome; he explains that currently most information is collected and analyzed in the individual departments, data aggregation and a more holistic analysis is often missing. Therefore, it is important to distinguish relevant from unimportant data [Lap13]. To make good predictions it is necessary to get customer as well as consumer data [Bru12]. Although there are new possibilities to gather such information, the related risks must not be neglected. For instance, in case of an internationally deployed SC possibly deviating or stricter laws of other countries must be known and adhered to [Wil13].
Among the articles on transportation, there are also some papers with basic information, which motivate companies to adopt Big Data [Sco14]. Others describe existing software solutions like SAS[2] and how these can be applied in the transportation sector [Bie13]. Furthermore, there are studies describing the topic's future development. A survey reveals that 30% of the shippers and 27% of the third party logistics (3PL) providers state that they were planning on adopting a Big Data initiative [Bur13]. However, 50% of the shippers and 3PL providers declare that for them, Big Data does not provide solutions for significantly improved, data-driven decision-making. There are several papers describing the process of information and data retrieval. [McC13] explains which Big Data processes may matter in the area of Transport Management Systems. These are, e.g., cleansing, harmonizing, combining and standardizing data, as well as data visualization. [Ali12] described the way of gathering information by crowdsourcing. The use of Geographical Information System is examined by [Mil13], who also indicates sensors, GPS and social media as new data sources for the transportation industry. The study by [MGI11] explains how information can be collected through new technologies like smartphones or GPS. The customer could use such data for smart routing, car monitoring and localized services. But this also holds benefits for the government (e.g., urban planning) and companies (e.g., localized advertising, optimized routing). [MGI11] further on provides useful information for companies and governments how to start with Big Data.
Although Inventory Management/Warehousing was no major topic, we give a brief overview of the two analyzed papers. The first paper describes how the analysis of Big

[2] http://www.sas.com/

Data can be used to optimize the flow of packages in a company, and how to use this for simulation. The findings could serve to plan and develop new distribution centers. The simulation leads to a more proactive approach [Tre13]. The second paper describes a scenario in which data gathered from barcodes are analyzed to reduce storage capacity and distribution [MGI11].
A detailed view on further analysis (e.g. the technologies and systems generating the data) can be accessed via http://uwi.uos.de/att/BDL-LiteratureConceptMatrix.pdf

3.2 Case Study Research Results

Overall, we identified 14 case studies, which we analyze in this section. Similar to the paper analysis, we grouped the case studies into three main thematic areas, namely transport (79% of the case studies), SC (14%) and manufacturing (7%). To allow for a deeper insight, we briefly summarize the case studies in table 3.

Name	**Description**	**Data Sources**	**Addressee**
Barcelona	The City of Barcelona analyzes Big Data collected from social networks, GPS signals and government systems. These data can be used to analyze the flow of people or visitors in case of festivals in order to identify the most suitable places for bike rental systems (bicing) or to arrange for adequate public transport. Furthermore, the data help to better respond in case of emergencies because the city is well-informed of the urban environment and agglomeration areas etc.	Social Networks, Mobile Apps, Websites, GPS, Government Systems	Departments (Employees), Citizens, Tourists
http://www.microsoft.com/casestudies/Microsoft-Excel-2010/City-of-Barcelona/City-Deploys-Big-Data-BI-Solution-to-Improve-Lives-and-Create-a-Smart-City-Template/710000003415			
Barnes&Noble	Barnes & Noble (B&N) use Big Data. They run a portal through which their suppliers are granted insights in their sales at B&N without having to buy software.	Business Data	Supplier
http://public.dhe.ibm.com/common/ssi/ecm/en/imm14097usen/IMM14097USEN.PDF			
Big Data in Logistics	The DHL Smart Truck is used to optimize route planning. Initial planning can dynamically be adjusted through data from current orders and traffic conditions. This results in CO2 and cost reduction by saving miles.	Business Data (Order, Shipment), Traffic Data	Employees (Truck Driver)
http://www.dhl.com/en/about_us/innovation/research_development/bigdata.html			
Daimler FleetBoard	The Daimler Fleet Board allows for the real-time analysis of data from vehicles, such as status, position, charge and order data. The data can be used for various purposes, e.g., to schedule maintenance by predictions or to identify the eco-efficiency of the drivers or vehicles. During use of the vehicle, the drivers are provided with information contributing to	GPS, Telematics, Business data	Employees (Truck Driver, Management)

	the improvement of the driving style. The software also grants comparisons with other companies (benchmarking).		
http://public.dhe.ibm.com/common/ssi/ecm/en/nic03008deen/NIC03008DEEN.PDF			
FleetRisk Advisors	Different data, like working hours, telematics data (distances, times) as well as data from the drivers are collected for risk analyses. The software provides risk assessments for each driver. Such information allows the company to detect drivers who bring about an increased risk of accidents.	Telematics, Business Data,	Employees
http://public.dhe.ibm.com/common/ssi/ecm/en/ytc03278usen/YTC03278USEN.PDF			
General Electric (GE)	GE installs sensors in its products (jet engines, turbines, trains) to make the items smarter. Thereby, the customers are supplied with real-time data that can help to improve services (i.e. optimized maintenance times).	Sensors	Employees, Costumers
http://sites.tcs.com/big-data-study/ge-big-data-case-study/			
Jet Airways India	Jet Airways India analyzes diverse data sources to provide accurate information on the carbon footprint for reporting to the EU. The analysis is broken down to the aircraft level and also takes into account the consumptions during maintenance activities.	unknown	Employees, Government
http://public.dhe.ibm.com/common/ssi/ecm/en/ytc03528usen/YTC03528USEN.PDF			
KTH	The KTH University in Stockholm (Sweden) collects various data from GPS, automotive radar sensors on roads, transport fees and weather in order to calculate the expenditure of time it takes to get from one place to another. Based on the data, different alternatives can be evaluated to optimize traffic within a metropolitan area.	GPS, Radar Sensors, congestion charging, weather, etc.	Researchers
http://public.dhe.ibm.com/common/ssi/ecm/en/imm14097usen/IMM14097USEN.PDF			
RabbitMQ	Lucid Logistics provides software to collect data from trucks (e.g., GPS). The data can be accessed by their customers via a SOAP API in real-time and can be used for own analysis.	GPS, telematics	Employees, Costumers
http://blog.gopivotal.com/case-studies-2/case-study-shift-from-delayed-polling-to-real-time-telematics-with-rabbitmq			
Schneider National	Schneider collects data from sensors in trucks for the measurement of position, road behavior, current fuel capacity and amount of charge. The data are used to optimize refueling: Based on the fuel level, the target, the gas stations (and prices) on the way, the software recommends	Sensors, GPS	Employees (Driver, leader)

	the best fuelling station. Further analyses like driving behavior are possible.		
http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf			
StatSlice	A provider of SCM and logistics services in the automotive industry was forced to constantly respond to customer requests. Through the implementation of a data warehouse with an interface for customers for real-time data analysis, this burden was eliminated.	Business data (various source systems)	Costumers
http://www.statslice.com/case-study-series-auto-industry-logistics-provider			
TimoCom	TimoCom analyzes data from the largest European freight and vehicle exchange portal. The results are delivered in real-time and provided to customers free of charge. These data are for instance used for pricing negotiations.	Business data from website	Employees (Management), Costumers, Press
http://www.pentaho.com/sites/default/files/uploads/resources/timocom.pdf			
UPS	By means of telematics sensors, UPS measures data, such as speed, direction, braking and drive train performance to optimize routes in real-time. In 2011, this resulted in the saving of 85 million miles (or 8.4 million gallons of fuel). In future, the data will also be used to optimize the over 2000 daily flights.	telematics sensors, online map data, Business data	Employees (Trucker)
http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf			
US Xpress	US Xpress analyzes various data (GPS, telematics data of trucks and data collected through drivers' tablets (e.g., via camera)). One major goal is to minimize waiting times for drivers/trucks, which can be achieved by providing real-time information to the user. They also perform a social network analysis to evaluate the drivers' opinions on the software.	GPS, telematics, video camera, GPS, business data	Employees (Trucker, Management)
http://www.computerweekly.com/news/2240146943/Case-Study-US-Xpress-deploys-hybrid-big-data-with-Informatica			

Table 3: Overview of Big Data logistics case studies

Sensors, telematics and GPS seem to be the most important data sources. The gathered information is mainly used by the companies themselves, not for the costumers.

3.3 Expert Interview Results

The five experts have been interviewed in four conversations. In this section, we want to reflect the main statements.

The first interview reveals that the company has not yet explicitly addressed the topic Big Data. However, it offers a web platform on which their customers can analyze data in near real-time. The data are gathered from traditional business systems like transport management systems and SAP. All data persist in relational databases in a company-owned data center. The bulk of data is stored only (mostly requested by costumers) and

has not yet been evaluated.
The interviewee thinks that appropriate knowledge/skills and staff are essential success factors for the implementation of Big Data: “If you want to be successful in online trading, you must be in a position to collect large amounts of data and to create reasonable evaluation algorithms, so as to ensure that the customer receives the requested data within the shortest period of time.”.

The company of the second interviewee is planning pilot projects on Big Data. But already now, a lot of data analyses are executed in near real-time. For this purpose a data warehouse management system (DWMS) is used in which data from different sources are loaded. Data are stored in different relational databases (DB2, Informix, Oracle) while QlikView20 is used as BI solution.
All analyses are focused on process optimization. “[…] it helps to be able to transparently illustrate the decision makers what is currently happening. Thus, the speed of decision-making can be raised.” For faster analyses a NoSQL[3] buffer layer is planned, which should enable near real-time access (15-30 minutes). A faster availability is not cost-effective at the moment. The company identified several success factors and restraints, such as lack of skills and poor adaptability of existing software solutions. The full support of the management as well as the collaboration of IT and the different departments is seen as essential. Potential fields of operations are seen in the warehouse control and workforce management, as these are essential cost-driving factors. "It is about getting the whole SC transparent per customer or per location, detecting bottlenecks or identifying problems early, but also to recognize reciprocal effects of various processes. This is in my opinion the key to success." Benefits may be improved staff scheduling and increased customer satisfaction by keeping them constantly informed.

The third company has been working on the analysis of structured and unstructured data for 10-15 years. Currently, the two interview experts are working on the implementation of a central intranet platform. Customer data, orders, GPS data are additionally stored to data from business software. The data are used by systems like a DWMS and web portal, as well as warehouse and air, sea and road management. As management support is mandatory, its awareness for restructuring had to be increased. The use of Big Data in the data-rich areas (transport and warehousing) may lead to competitive advantages.
The third expert of the company reported on another project. To optimize warehouse logistics, the company has been working in the field of Complex Event Processing (CEP) as part of a pilot project. First forecasts are positive but measurable results have not yet been achieved. Another pilot project with explicit reference to Big Data will be launched in 2014.
Again, unexamined data are stored for future analysis (e.g. further development of CEP or Big Data pilot project). Recent analysis via DWMS is available within an hour, while CEP nearly achieves real-time.

[3] NoSQL = Not only SQL; SQL = Structured Query Language

The storage and transport logistics are considered as the most promising areas for real-time analysis. Production costs could be reduced and production quality could be raised by early detection of faults.

The full interviews can be downloaded at: http://uwi.uos.de/att/BDL-InterviewTranscripts.pdf

4 Discussion of Results

Regarding to RQ1, this research reveals that the topic of Big Data in logistics is still in its infancy. Our open search resulted in only a small number of scientific papers. Two main areas (transport and SC) could be identified, others remain rather unattended.

Especially in the area of warehouse management, we expected a high data volume (through barcodes, RFID, robots and sensors) and thus a large base of literature. To uncover potential opportunities, we recommend starting research in unexplored areas like warehousing. Due to the wide scatter of topics in the two main areas no focus or trend could be identified. Most of the papers are rather technical or describe possible use cases. Acceptance and behavioral research are completely missing. This research should be carried out in areas where Big Data projects have already been implemented.

The case study research offers a similar picture. Again, SC and Transportation are the largest areas, nearly 80 % of the case studies deal with transportation. The result of the literature review is assisted by the case studies. Especially in the field of transportation, there are many cases where data are collected from various sources (sensors, GPS, even social networks), and used for analysis. The analysis is often applied for optimized routing or recognizing driving behavior. Thus, the information achieved will be directly consumed in the company and used for process optimization. Concerning RQ2, Big Data, especially the analysis of various data sources like sensors, GPS and tags, is demonstrated in some cases in the realm of transportation. In other logistics areas it is widely under-represented.

The impressions of the literature review and case study analysis are affirmed by the results of the expert interviews. Again, a lot of data are collected from different sources, but there is a lack of analytical possibilities, often due to missing (costly) infrastructure. The use of near real-time information is seen as sufficient in most companies. The benefits do not justify the disproportionate costs for real real-time. There are no real business cases among the interviewed companies. Referring to RQ3, potential is seen in the field of transport (which coincides with the literature analysis and case study analysis), as well as in the area of warehousing. The gathered information should mainly be used by decision makers. Management support and trained staff are seen as essential success factor in Big Data projects.

5 Conclusion

This contribution captures the status quo of Big Data in logistics by means of a literature review, a case study analysis and expert interviews. However, there are some limitations in our research. Less than one third of the identified articles are academic. And although we identified several case studies of good quality, we may have missed others due to the ranking of the Google results. The expert interviews were conducted in Germany and thus only reflect the domestic image.
Although Big Data is still in its infancy in research and practice, there are some initial experiences in transportation and SC. From the analysis, the following implications for research and practice are derived.

Main Contributions and Implications for Science

1. Initiate research projects
   There are more practical examples than scientific articles. Thus, researchers are to be encouraged to conduct research on the relevant topic Big Data in logistics. Especially the use of technologies like tags (RFID, barcode, etc.) and ubiquitous computing (mobile phones, smartphones, etc.) and the use of data generated in the web (social networks, blogs, etc.) is not investigated in detail, as most paper concentrate on the use of business data (cf. concept matrix provided in section 3.1).
2. Interdisciplinary research
   Often, the technical side is in the foreground of Big Data research. We therefore call for an interdisciplinary research, which is composed of the areas of Information Systems, Computer Science, Statistics, Business Management as well as Social Science.
3. Relevant research
   To avoid irrelevant research, practitioners should be integrated into the research process as early as possible.
4. Explore new areas
   In addition to the already partially established areas transportation and SC, the research should also explore new areas. E.g. a warehouse management system can be enhanced by the data of various sensors installed in the warehouse or the products itself. In this controlled environment the use of technologies like tagging (barcode, RFID, etc.) can be used in a valuable manner.
5. Measure user acceptance and behavior
   Since there are already implemented Big Data projects, advanced research in these areas can take place, e.g. investigate how the behavior of the user has changed and whether the technology is accepted.

Main Contributions and Implications for Practice

1. Build a Big Data strategy
   As technical burdens are overcome, first Big Data projects can begin. The target should be to identify how Big Data can improve management decision or daily business. Case studies can be used to identify practices and technologies used by other companies; however a best practice had not been established yet. In case of transport and supply chain management there are several use cases which can be used. In case of warehouse management or production only innovators should introduce Big Data projects as there is no reference.
2. Management Support
   Since Big Data may change processes, an adequate change management and management support in Big Data projects is essential.
3. Include departments
   It is important to understand what the user wants. Therefore, department staff should be integrated into the projects as early as possible and a holistic approach should be followed.
4. Staff & Skills
   In addition to the technical side (knowledge in information systems used for Big Data Analysis), employees who deal with Big Data (e.g. data analysts) also need to understand the processes (domain knowledge) and possess analytical skills. Therefore, appropriate training or hiring of new staff is needed.

Acknowledgments

The authors would like to thank the anonymous reviewers as well as the other project members, specifically Ms. Marita Imhorst, who provided valuable insights, help and substantive feedback during the research process.

References

[Ali12] Ali, Kashif; Al-Yaseen, Dina; Ejaz, Ali; Javed, Tayyab; Hassanein, Hossam S.: CrowdITS: Crowdsourcing in intelligent transportation systems. IEEE Wireless Communications and Networking Conference, 2012, p. 3307–3311.

[Bie13] Biederman, David: BIG DATA. In: Journal of Commerce 14 (2013) 5, p. 82–87.

[Big13] Big Data: "Big Data" Means Unstructured Data for Supply Chain Managers, Material Handling & Logistics, vol. 68, no. 3, p. 12, 2013.

[Bru12] Bruner, Jon: The Case for Big Data: "You Can't Do Anything Cool Until You Know How Much You've Sold," Forbes.com, 2012. URL: http://www.forbes.com/sites/jonbruner/2012/02/29/the-case-for-big-data-you-cant-do-anything-cool-until-you-know-how-much-youve-sold/ (18. February 2014).

[Bur13] Burnson, Patrick: New Study Released at CSCMP Addresses "Big Data." In: Supply Chain Management Review 17 (2013) 7, p. 76–78.

[Byr12] Byrne, Robert F.: Driving Profitable Growth with Big Data and Better Forecasts. In: Supply Chain Europe 21 (2012) 1, p. 40–41.
[CCS12] Chen, Hsinchun; Chiang, Roger H. L.; Storey, Veda C.: Business Intelligence and Analytics: From Big Data to Big Impact. In: MIS Quarterly 36 (2012) 4, p. 1165–1188.
[Cec12] Cecere, Lora: Big Data – Go Big or Go Home? Philadelphia, 2012 p. 1–18.
[Cha13] Chao, Chian-Hsueng: The Framework of Information Processing Network for Supply Chain Innovation in Big Data Era. In (Uden, L.; Wang, L. S. L.; Hong, T.-P.; Yang, H.-C.; Ting, I.-H. Hrsg.): The 3rd International Workshop on Intelligent Data Analysis and Management, 2013, p. 77–85.
[Cob13] Coby, Paul: Technology brings the industry together. In: Airline Business 29 (2013) 8, p. 16–177.
[Cow13]Cowen, Martin: THE BIG BANG. In: Airline Business 29 (2013) 4, p. 49-51.
[Dai12] Dai, Jinquan; Huang, Jie; Huang, Shengsheng; Liu, Yan; Sun, Yuanhao: The Hadoop Stack: New Paradigm for Big Data Storage and Processing. In: 2Intel Technology Journal 16 (2012) 4, p. 92–110.
[Dib04] Dibbern, Jens; Goles, Tim; Hirschheim, Rudy; Jayatilaka, Bandula: Information Systems Outsourcing : A Survey and Analysis of the Literature. In: The DATA BASE for Advances in Information Systems 35 (2004) 4, p. 6–102.
[Fer13] Ferguson, Renee Boucher: How Big Data is Influencing Big Companies, MIT Sloan Management Review, 2013. URL: http://sloanreview.mit.edu/article/how-big-data-is-influencing-big-companies/.
[Gob13] Gobble, MaryAnne M.: Big Data: The Next Big Thing in Innovation. In: Research-Technology Management 56 (Jan. 2013) 1, p. 64–67.
[Gro14] Groves, William; Collins, John; Gini, Maria; Ketter, Wolfgang: Agent-assisted supply chain management: Analysis and lessons learned. In: Decision Support Systems 57 (Jan. 2014), p. 274–284.
[Hol14] Holmes, Lawrie (AUTHOR): Sir Terry Leahy, Financial Management, vol. 42, no. 3, p. 20–24, Apr-2014.
[HTZ13] Huang, George Q.; Tsui, Kwok Leung; Zhong, Ray Y.: Big Data for Service and Manufacturing Supply Chain Management. 2013.
[Lap13] Lapide, Larry: The Promise and Pitfalls of Big Data. In: Supply Chain Management Review 17 (2013) 4, p. 4–5.
[LK12] Li, Ting; Kauffman, Robert J.: Adaptive learning in service operations. In: Decision Support Systems 53 (May. 2012) 2, p. 306–319.
[Lu13] Lu, Tianbo; Guo, Xiaobo; Xu, Bing; Zhao, Lingling; Peng, Yong; Yang, Hongyu: Next Big Thing in Big Data: The Security of the ICT Supply Chain. International Conference on Social Computing, 2013, p. 1066–1073.
[McC13] McCrea, Bridget: TMS: Trends driving adoption. In: Logistics Management 52 (2013) 11, p. 32–34.
[MGI11] MGI: Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.
[Mil13] Miller, Harvey J.: Beyond sharing: cultivating cooperative transportation systems through geographic information science. In: Journal of Transport Geography 31 (Jul. 2013), p. 296–308.
[My09] Myers, Michael D.: Qualitative Research in Business & Management. Sage Publications Ltd, London, 2009.
[Ola13] Olavsrud, Thor: Where to Find Big-Data Gold, CIO, vol. 26, no. 12, p. 11, 2013.
[Rit14] Ritson, Mark: Amazon has seen the future of predictability, Marketing Week, no. 01419285, p. 10, Jan-2014.
[Sav12] Savitz, Eric: The Big Value In Big Data: Seeing Customer Buying Patterns, Forbes.com, 2012. URL: http://www.forbes.com/sites/ciocentral/2012/09/25/the-big-value-in-big-data-seeing-customer-buying-patterns/ (18. February 2014).

[Sch13] Schultz, Beth: Big Data in Big companies. In: Baylor Business Review 32 (2013) 1, p. 20–21.
[Sco14] Scott, Alex: How Supply Chain Managers Dig Big Data. In: Material Handling & Logistics 69 (2014) 1, p. 25–26.
[SK13] Söntgerath, Markus; Kückelhaus, Veronika: Logistics Trend Radar - Delivering insight today. Creating value tomorrow. Troisdorf, 2013.
[Tre13] Trebilcock, Bob: The big picture on BIG DATA. In: Logistics Management 52 (2013) 10, p. 53–57.
[Wal06] Walsham, Geoff: Doing interpretive research. In: European Journal of Information Systems 15 (Jun. 2006) 3, p. 320–330.
[WF13a] Waller, Matthew A.; Fawcett, Stanley E.: Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management. In: Journal of Business Logistics 34 (Jun. 2013) 2, p. 77–84.
[WF13b] Waller, Matthew A.; Fawcett, Stanley E.: Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. In: Journal of Business Logistics 34 (Dec. 2013) 4, p. 249–252.
[Wil13] Williams, Cindy: More Big Things. In: World Trade: WT100 26 (2013) 12, p. 20–21.
[WW02] Webster, J.; Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: MIS Quarterly 26 (2002) 2, p. xiii–xxiii.

Towards a Big Data-based Technical Customer Service Management

Deniz Özcan, Michael Fellmann, Oliver Thomas

Information Management and Information Systems
Osnabrück University
Katharinenstr. 3
49074 Osnabrück
{deniz.oezcan, michael.fellmann, oliver.thomas}@uni-osnabrueck.de

Abstract: The increasing use of Information Systems (IS) in diverse business sectors and the growing degree of mobility has changed the traditional use of Information Technologies (IT) in the economy. While Big Data is a current issue in the telecommunication and trade sector, high data intensity is also in the service sector given. Especially in the Technical Customer Service (TCS) area, a large variety of information and data exist that are needed for a service process or that result from them. In this paper, a conceptual framework for Big Data-based Service Management of the TCS is presented revealing opportunities and improvements for both product and service development. These opportunities lead to generating value from the use of large datasets and applying advanced analytical methods to increase efficiency and service quality. Finally, with the help of the framework an optimized Product-Service-Lifecylce Management is envisioned.

1 Introduction

The machinery and plant engineering encompasses more than 6.000 companies and approximately 978.000 employees which makes it to the largest industrial employers in the German industry. The increase of the mechanical engineering production by 24 percent in the past 3 years clarifies the importance of this sector [Vdma13]. In addition to industrially manufactured products, industrial services are also of great importance. A majority of these services are inputs for the manufacturing sector, with the manufacturing sector as one of the largest buyers of business services [Edle13]. This results from the fact that in recent years suppliers of machines and plants have emerged from a mere producer to service providing producer, because customers are increasingly demanding integrated solutions to meet their individual needs [ThWL08a]. In this way, conventional products turn into customer-focused solutions that provide plant and machinery installers the opportunity for competitive differentiation [BhVF93]. The interface of the industrial service to the customer is provided by the Technical Customer Service (TCS). The TCS takes an active part in the utilization phase of a physical product and serves the purpose to maintain the performance of the property during the life cycle continuously. In addition, the TCS ensures that the investment of the users will

pay off in the long term [Harm99]. The traditional functions of the service technician include repair, inspection as well as maintenance activities that are performed on a service object at customers' site [Walt10a]. Due to the product-related operations with technical consumer goods, the TCS has become an essential part of the so-called Product-Service Systems (PSS) and requires intelligent support in its knowledge-intensive activity [ThWL08b].

A large number of systems exist aiming at supporting the service technician, e.g. diagnosis system, condition-monitoring, maintenance system and various teleservice system solutions. Besides, mobile assistance systems are increasingly used for processing service orders [MFÖK13]. The integration of mobile solutions into business processes owns a lot of potentials to revolutionize the service delivery at the "point of service" admittedly, but to date there is still no structured development of integrated systems that aim at real-time analysis of relevant service data and their communication and evaluation in an adequate way before, during and after a service process. In addition, appropriately processed data and thus important requirements from the customer's perspective do not or insufficient flow back to the research and development and into the product planning phases of the manufacturer. This is somewhat surprising regarding the already identified growing importance of service and support as a Big Data field of application [Bitk12]. In fact, it is challenging for traditional data analysis to cope with emerging requirements, imposed by the growing data volumes, the increasing number of data users, the growing number of data sources and the demand for real-time analysis [ZiLa13]. The challenge is even amplified by current surveys showing that in future the number of intelligent networked products and services will continue to multiply exponentially in all industries [Deut14]. To meet these challenges, an effective data preparation and use is required, which provides the integration, processing and deployment of service relevant data along the value creation process of the Product-Service System and including the value added partners with the aim to optimize the data usage so that these information can be integrated into new products and services. For this reason, an appropriate Big Data-based infrastructure is necessary to meet the demands of the growing data of the TCS.

In this paper, we investigate Big Data issues in the field of the Technical Customer Service and provide a conceptual Big Data-based Service Management framework for the TCS. Section 2 introduces to Service Management in the area of the TCS by providing fundamental knowledge about TCS processes and pointing out information needs as well as data occurrences within these service processes. In section 3, we categorize the Big Data topics and transfer them to the TCS domain. Further, Big Data analytic tools are presented. In section 4 a conceptual framework of Big Data leverage is described including potencials for the TCS. Section 5 reveals challenges of Big Data in the service sector and section 6 concludes our approach and gives an outlook to further research.

2 Service Management and the Technical Customer Service

2.1 Technical Customer Service Processes

The service process involves activities undertaken to realize and deliver the service [BoYa05]. A service process of the TCS consists of various functions and sub-processes, which may occure in different manifestations depending on the company. The master and transaction data serve as a base that enrich the functions and processes (Figure 1).

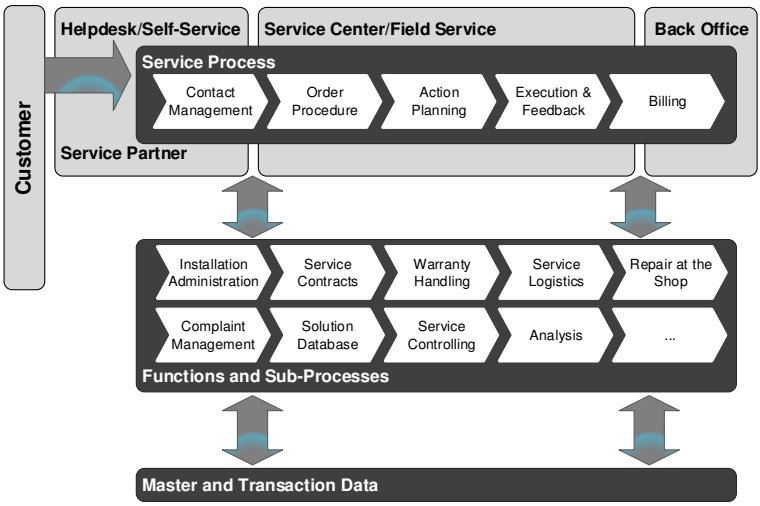


Figure 1: Functions, data and actors of a service process (referring to [Vdma08])

The customer initiates the service process by contacting the TCS so that the contact management takes place. After the contact management occurred the service process passes the stages order procedure, action planning, execution and feedback ending with the billing of the service order. The service process can pass through different departments that operate on the service process. Each department may require a slightly different view of the service processes where different requirements for the Service Management solution emerge. Stakeholders of those departments can be, for example, employees of the helpdesk, customer service employees, service field workers, accountants and logistics experts. Within the service process functions and sub-processes arise, e.g. installation administration, complaint management, which can vary depending on the service order. The execution of these functions and sub-processes needs access on master and transaction data and transmitts in turn data back after processing [Vdma08]. A wide variety of IT resources and capabilities are relevant to provide the transaction and master data required for the execution of the customer service process [RaMB05]. The integration of mobile IT improves the service process and supports the service technician at multiple levels by providing required information in real-time [ÖFMD13][BrTW08]. The service technician needs the relevant information about the state of the system and the documentation and order data either for remote operation or when he is at the customer's site in a timely manner. [BNST08][Walt09]. The

integration of mobile devices introduces higher levels of flexibility and personalization of the service technician and presents organizations opportunities to integrate new or enhanced business processes that would result in greater productivity, efficiency and effectiveness [CoHH08]. Information needs of the TCS causes that there is the necessity to integrate appropriate IS to support the operation and the staff efficiently. For ensuring the competitiveness and quality of the service, it is important to structure and process the information needs and existing data of the service phases and sub-processes with analytical methods that meet the demands of the growing (big) data.

2.2 Information Needs and Data Occurrence in Service Processes

Information can be declared as a fundamental component for the performance of a service process. For this reason, the usage of IS is a critical success factor to provide the TCS employees the needed information for a qualitative service delivery [RaMB05]. The activities of employees require in-depth knowledge, skills and sufficient trained IT support to cope with the wide range of tasks. Thus not only information about the service itself, but also information about the considered products are required [Stil03] which are likewise important for the entire value chain [Walt10b][ÖsSe06]. The information base must also include customer data due to the customer proximity of the service technician [DeWB09]. In all phases of the service process including the functions and sub-processes, the documentation of information occurs [Stil03]. Documents that are stored as data with IS, provide information that can be recycled [WaPK95]. The documented facts are also fundamental part in the subsequent assessment phase of service provision, since this information is used for after-sales purposes and conveys retroactively information by explicit analysis of the service provision [Harm99] [PoMi85]. While traditional Knowledge Management encompasses the activities of maintaining, structuring, representing and sharing of knowledge within an organization [ScLe02][Sofi14], requirements for Knowledge Management changed through the growing amount of large data sets caused by multiple systems and actors in the value creation process [KaWa14]. These generated or received data during a service process can be of structured, semi-structured or unstructured nature (Figure 2). Examples for structured data are customer master data that reside in fixed fields, like columns or rows [MCBB11]. Placed in a relational database the fields and their relations to one another are clearly defined [Cóte14]. Semi-structured data are also structured to a certain grade but they also own an unstructured part (e.g. e-mails in which the head of the mail follows a clear structure but the body can be of unstructured nature comprising contents and attachments of any type or XML and RSS-feeds). Unstructured data comprise data that do not reside in fixed fields and include free-form text or human language, e.g. articles, talks, untagged audio [Russ11]. Within a service process, the involved actors generate unstructured data by people to people communication in virtual communities, social networks or web logs. Also a communication between people and applications or machines exist like data resulting from e-commerce or data generated by using specific mobiles. Machine to machine communication includes data from sensors, GPS data or images and videos from cameras [KlTH13]. Embedded systems are now integrated into a variety of objects and thus endow them with intelligence. Equipped with sensors,

intelligent objects capture data from their environment and react upon them [Deut14] which occasions increasing data traffic without human interaction.

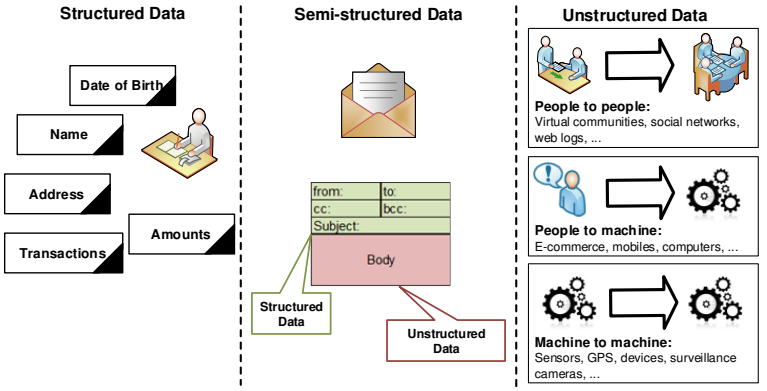


Figure 2: Data structures of a service process (referring to [KlTH13])

These different types of data are generated in the multimedia-based and knowledge-intense environment [Kant11] and need data-mining methods to handle them and make them analyzable.

3 Big Data

3.1 Big Data Topics and their Pertinence to the TCS

In recent years, Big Data has become a new ubiquitous term [BRMH13] but however no consistent definition exists [KlTH13]. Chen et al. (2012) see Big Data Analytics as a related field of Business Intelligence (BI). The term comprises large and complex data sets that require advanced IT for data storage, analysis and visualization [ChCS12]. Initial situation that triggers Big Data, includes data collection from enterprises of various department sections and even data beyond their own company, like data of customers and business partners [BrCM11]. Elementary components of the Big Data topic are four basic properties, namely volume, velocity, variety and veracity [BRMH13]. While the property volume relates to the amount of data, variety refers to its heterogenity. In the range of velocity measures and techniques need to be integrated to optimize the speed of data processing [Lyce13]. Veracity concerns data regarding their quality. Transferring the Big Data topics to the TCS we ascertained the following findings.

Variety. The traditional functions of the service technician comprise the tasks of maintenance, which include the maintenance, inspection and improvement [Din12]. They are mainly carried out at the customers' site on a service object. A challenge in the performance of maintenance tasks consists of the large number of manufacturers whose service objects underly various maintenance requests, so that the service technician

needs to be available of a wide range of subject-specific knowledge at the right ”point of service“. Effective data preparation and data use allow the automation of routine service and analysis of service cases, so that the service engineer can focus on the difficult cases [Bitk12]. Despite these efforts, the engineers are increasingly overtaxed without a comprehensive preparation and provision of diverse data on mobile assistance systems as a consequence of the enormous product variety and complexity of the complicated maintenance tasks [MFÖK13].

Velocity. The increasing interconnection of physical objects with information-processing components (also known as cyber-physical systems [GeBr12]), e.g. sensors, generates data which require a fast transfer. The continuous monitoring of sensor data enables predictive handling of technical problems and avoidance of far-reaching damages. For example, a trouble-free maintenance of an air conditioning, which is used for cooling server rooms, has to be guaranteed. If a fault is transmitted via sensors, the responsible object manager needs to be informed. Such errors, error messages and other information about the state of the plant have to be used to prevent impairments of the machinery. However, sensor data need to be transferred in real-time to the appropriate positions in order to avoid such downtime.

Volume. To facilitate analysis and conclusions in terms of product development, the collection, analysis and correct interpretation of all relevant data from the environment of the service object is required. Also data in form of huge amounts of unstructured text from service reports or product forums of the social web need to be used to draw conclusions on problems or opportunities for product and service improvement [MCBB11].

Veracity. Information generated by sensors but also additional information, such as customer feedback data may include noise, errors and distortions but information need to be accurate and reliable for decision making. Incompleteness and errors must be managed during data analysis but doing this correctly is a challenge [ABBD11].

As the previous sections have revealed, the problems that are at the heart of Big Data and that are characterized by the four “V”s also increasingly gain importance in the context of the TCS. Hence, Big Data analytic tools can be leveraged to tackle these challenges in the TCS associated to the four “V”s.

3.2 Big Data Analytic Tools

The analysis and mining of Big Data is challenging. In order to make use of Big Data deep analysis, instead of just generating simple report forms, is needed [CCDL13]. Data warehouse is a relational database system used for storing, analyzing and reporting functions of structured data employing Standard Query Language (SQL). Unstructured data are of a more complex nature what makes tradtitional methods of storing and managing impractical [Baks12]. “Not only SQL” (NoSQL) database management systems are designed for large-scale data storage of a huge quantity of data especially when the data’s nature does not require a relational model [Hoss13]. The MapReduce algorithm was developed to process large data collections and builds the foundation for

Apache Hadoop which is a distributed file system modeled on Google File Systems (GFS) and a distributed processing framework, using both MapReduce concepts and a distributed database called HBase [BoCL12]. Hadoop also includes a cluster file system that scales to store massive amounts of data, also called Hadoop Distributed File System (HDFS). The MapReduce algorithm is dived into two phases. In the map-phase input data is subdivided into smaller workloads and are assigned to a map-task that processes each block of data. The result is a sorted list that is spread in the reduce-phase. Out of the preliminary result from the map-phase, the final output is produced and written to the HDFS [Baks12][KlTH13]. Since Big Data become present, different storage systems have been developed [RSJG12]. However, in consideration of the developments, new IT infrastructure opportunities build the starting point towards the intelligent management of data and their usage, especially in the field of the TCS.

4 A Conceptual Big Data-based Service Management Framework

Much of the value that Big Data can create in the TCS sector requires the access and varied use of data from multiple sources across the extended enterprise [MCBB11]. Figure 3 shows the crucial partners participating in the general extended value-adding chain.

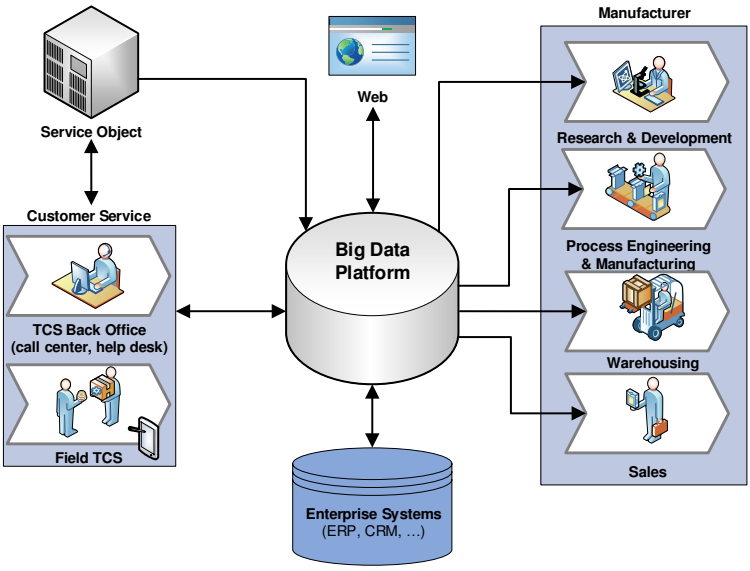


Figure 3: Big Data-platform for the TCS

The partners contribute to the product-supporting service provision and are differentiated into the work areas research and development, process engineering and manufacturing, warehousing and sales [FHBT11]. The manufacturer in total creates necessary data that are relevant for the service provision of the TCS (e.g. product data, manual or service handbook). However, the heterogeneous amount of data the TCS needs to be available of

and which is generated by the plurality of machines and equipment is difficult to control because the service objects underly different maintenance requests, e.g. operating instructions. In addition to the data and information provided by the work areas of the manufacturer the TCS itself generates also information about the service object, service order or customer within the back office or in the field. These can be structured or unstructured data comprising service reports, protocols, lack messages, bills or simple comments about the service order. Due to the increasing interconnectedness of physical objects with information-processing components, data of the service object itself are transmitted to the TCS and their devices. Sensor data must be integrated into the service process and transmit in real-time to relevant authorities. At the same time a further channel, the web, is claimed by the increased cross-linking and presence of producers and service providers on the internet. Here, information is shared among customers and also partly employees of the company. The bidirectional arrow directions indicate the data and information exchange between the actors and objects.

Hitherto, regarding the information flow between the partners of the value chain, research has focused on the integration of information across the value chain with the aim to provide improved (mobile) access to information that are stored in an integrated information system [TWLN07][BlNü10]. Taking these underlying conditions into account, an appropriate Big Data-based processing is needed to allow an exploratory, retrospective and prospective analysis of data in order to support the Service Management of the TCS. The proposed Big Data-platform should contain a generic collection of methods, algorithms and use case descriptions intended to facilitate and guide implementation of Big Data systems in the TCS domain. The data reflow from the platform constitutes the supply of information. The leverage of Big Data in the domain of the TCS is the use of customer-related data to provide insights for the benefit of all partners in the value chain. Data- and text-mining algorithms used for unstructured data allow their editing and reflow to relevant actors of the value-chain. Especially for research and development departments great opportunities are offered to accelerate product development, product improvements and service-efficient product development that minimize production costs and harness consumer insights [MCBB11]. Analyzing service-relevant data in real-time at the mobile point of service enables a preventive maintenance of the TCS by direct information evaluation and provision, e.g. through fault hypotheses and user-driven data analysis using “one click mining”-concepts [BMKT13]. Preventive actions improve customer satisfaction and provide for a more efficient warehousing [KuWD09]. Thus, the user is able to benefit from the analysis algorithms in the context of their daily work without requiring help of experts. This helps to improve the autonomy of the TCS and allows the automation of other service instances. The real-time processing of both structured and unstructured data plays especially for ensuring the operability of the systems an important role since it affects performance-critical product-related services. Also errors can be spotted in early development phases, since the cost for errors increase considerably in each stage of the product lifecycle [ABBD11]. A holistic optimization of the service lifecycle can be realized because manufacturers have tremendous potential to generate value from the use of large datasets, integrating data across the extended value chain and applying advanced analytical techniques to raise their productivity both by increasing efficiency and improving the quality of their products [MCBB11].

5 Challenges of Applying Big Data in the TCS Domain

Gathering and analyzing large amounts of data for both gaining real-time insight into the operation of the TCS and to inform other departments such as warehousing and sales has not been proposed to date. However, the core idea of controlling the growing mass of information and data through innovative solutions is associated with challenges. Data analysis, organization, retrieval and modelling of different structures and created by different departments of the value chain are foundational challenges. Furthermore, the interconnection of machines to other computer systems and the growing importance of the web afford additional data that may also possible sources of error: computer systems can have bugs or results can be based on erroneous data [ABBD11]. Moreover semantic challenges exist which makes it difficult to extract only those information that are relevant for the value-added content of a company out of an unstructured mass [ZaPG13]. In human collaboration, which is given in the TCS domain, the extra challenge is the uncertainty of the data collection devices. Smart devices used for Big Data analytics and even the so-called "Bring Your Own Device"-trend (BYOD) are critical regarding data privacy and can be seen as a further driver for data growth. Traditional instruments of the data protection law are barely to deploy although the Federal Data Protection Act commands that data which is gathered for different purposes need to be processed isolated (separation rule). Gathered data can only be processed and used for purposes that are compatible with the purpose of the collection (purpose limitation) but Big Data requires the decoupling of data from their purpose. Even with BYOD there is a risk, that data may be alienated [Conr13]. Further, attempts in mobile web attacks have increased, along with the growth of mobile device users. Especially mobile web services are affected by this because service methods of many businesses have changed to web-based services, and their dependence on web system has also increased with the rapid increase in mobile web applications [HaLe14]. As we can see a wide range of challenges exist affecting different scopes. Big Data users need to be aware of these challenges to work efficiently with Big Data solutions and to avoid adverse consequences.

6 Conclusion and Outlook

While the relevance of Big Data is widely discussed in the fields of social media, e-commerce [BRMH13] and healthcare [Cóte14], it is not prevalent in the machinery and plant engineering and it's accompanying TCS yet. Especially social media is seen as a showpiece for Big Data due to the variety of user data and the need to calculate and analyze them to control offers in real time and provide an optimized referral marketing. The service sector is equally affected by Big Data problems through the necessity of information and the increasing use of mobile IS which also cause a wide variety of data in structured, semi-structured or unstructured form. Expecially product-related services of the TCS require analytical tools that enable the effective and efficient processing of data before, during and afterwards the service process. For ensuring the competitiveness and quality of the service delivery it is important to structure and process the existing data. Big Data analytic tools support the return of processed data back from service

delivery into the planning phase of product development. This enables a holistic optimized product-service lifecycle management by submitting current issues, problems and suggestions to each actor of the extended value chain. This paper provides a conceptual framework for a Big Data-based Service Management of the TCS. Opportunities are revealed and challenges described that occure as side-effects of Big Data technologies. In a next step, it is planned to gather existing applications and required interfaces in the field of the TCS to specifiy appropriate Big Data-based analytic algorithms to enable the demand-actuated recall of integrated and rehashed service data.

References

[ABBD11] Agrawal, D.; Bernstein, P.; Bertino, E.; Davidson, S.; Dayal, U.: Challenges and Opportunities with Big Data 2011-1. Cyber Center Technical Reports, Purdue University, 2011.

[Baks12] Bakshi, K.: Considerations for Big Data: Architecture and Approach. In: 2012 IEEE Aerospace Conference (Mar. 2012), pp. 1–7.

[BhVF93] Bharadwaj, S.G.; Varadarajan, P.R.; Fahy, J.: Competitive Advantage in Service Industries: A Conceptual Model. In: Journal of Marketing 57 (1993) 4, pp. 83–99.

[Bitk12] BITKOM: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. Berlin 2012.

[BlNü10] Blinn, N.; Nüttgens, M.: Empowering Technical Customer Services with Intelligent Mobile Applications: A Contingency Theory-Based Approach. In: Proceedings of the IFIP, Working Papers on Information Systems, 10 (107), 2010.

[BMKT13] Boley, M.; Mampaey, M.; Kang, B.; Tokmakov, P.; Wrobel, S.: One Click Mining – Interactive Local Pattern Discovery through Implicit Preference and Performance Learning. In: IDEA'13, 2013, pp. 27–35.

[BNST08] Blinn, N.; Nüttgens, M.; Schlicker, M.; Thomas, O.; Walter, P.: Lebenszyklusmodelle hybrider Wertschöpfung: Modellimplikationen und Fallstudie an einem Beispiel des Maschinen- und Anlagenbaus. In: Multikonferenz Wirtschaftsinformatik 2008, pp. 711–722.

[BoCL12] Borkar, V.; Carey, M.J.; Li, C.: Inside "Big Data Management": Ogres, Onions, or Parfaits? In: EDBT/ICDT 2012, 2012, pp. 3–14.

[BoYa05] Boughnim, N.; Yannou, B.: Using Blueprinting Method for Developing Product-Service Systems. In: International Conference on Engineering Design, 2005, pp. 1–16.

[BrCM11] Brown, B.; Chui, M.; Manyika, J.: Are you ready for the era of "big data"? In: McKinsey Quarterly (2011), pp. 1–12.

[BRMH13] Buhl, H.U.; Röglinger, M.; Moser, F.; Heidemann, J.: Big Data – A Fashionable Topic with(out) Sustainable Relevance for Research and Practice? In: Business & Information Systems Engineering 5 (Feb. 2013) 2, pp. 65–69.

[BrTW08] Brocke, J.; Thurnher, B.; Winkler, D.: Improving IT-Service Business Processes by Integrating Mobility: The Role of User Participation – Results from Five Case Studies. In: Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE'08), 2008.

[CCDL13] Chen, J.; Chen, Y.; Du, X.; Li, C.; Lu, J.; Zhao, S.; Zhou, X.: Big data challenge: a data management perspective. In: Frontiers of Computer Science 7 (Apr. 2013) 2, pp. 157–164.

[ChCS12] Chen, H.; Chiang, R.H.L.; Storey, V.C.: Business Intelligence and Analytics: From Big Data to Big Impact. In: MIS Quarterly 36 (2012) 4, pp. 1165–1188.

[CoHH08] Coursaris, C.; Hassanein, K.; Head, M.: Mobile technology and the value chain: Participants, activities and value creation. In: International Journal of Business Science and Applied Management 3 (2008) 3, pp. 14–30.

[Conr13] Conrad, I.: Big Data und ByoD – Datenschatz und Datenschutz. In: Informatik-Spektrum 36 (Jun. 2013) 4, pp. 407–408.

[Cóte14] Cóte, M: Data Motility – The Materiality of Big Social Data. In: Cultural Studies Review 20 (2014) 1, pp. 121–149.

[Deut14] Deutsche Akademie der Technikwissenschaft (acatech): Smart Service Welt – Umsetzungsempfehlungen für das Zukunftsprojekt Internetbasierte Dienste für die Wirtschaft. 2014.

[DeWB09] Deuse, J.; Wischniewski, S.; Birkmann, S.: Knowledgebase für die kontinuierliche Innovationsarbeit im Technischen Kundendienst. In (Herrmann, T.A.; Kleinbeck, U.; Ritterskamp C. eds.): Innovationen an der Schnittstelle zwischen technischer Dienstleistung und Kunde. Physica, Heidelberg 2009, pp. 155–176.

[Din12] DIN: Grundlagen der Instandhaltung 31051. Beuth Verlag, Berlin 2012.

[Edle13] Edler, D.: Die Industrie – ein wichtiger Treiber der Nachfrage nach Dienstleistungen. In: DIW Wochenbericht (2013) 34, pp. 16–24.

[FHBT11] Fellmann, M.; Hucke, S.; Breitschwerdt, R.; Thomas, O.; Blinn, N.: Informationssystemarchitekturen zur Unterstützung technischer Kundendienstleistungen. In: Wirtschaftsinformatik Proceedings 2011, pp. 252–261.

[GeBr12] Geisberger, E.; Broy, M.: Integrierte Forschungsagenda Cyber-Physical Systems – acatech STUDIE. München 2012.

[HaLe14] Ham, Y.J.; Lee, H.W.: Big Data Preprocessing Mechanism for Analytics of Mobile Web Log. In: International Journal of Advance Softwared Computing 6 (2014) 1, pp. 1–18.

[Harm99] Harms, V.: Kundendienstmanagement. Dienstleistung, Kundendienst, Servicestrukturen und Serviceprodukte; Aufgabenbereiche und Organisation des technischen Kundendienstes. Neue Wirtschaftsbriefe, Herne 1999.

[Hoss13] Hossain, S.A.: NoSQL Database: New Era of Databases for Big data Analytics – Classification, Characteristics and Comparison. In: International Journal of Database Theory and Application 6 (2013) 4, pp. 1–14.

[Kant11] Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. Wiley & Sons, 2011.

[KaWa14] Kandalkar, N.A.; Wadhe, A.: Extracting Large Data using Big Data Mining. In: International Journal of Engineering Trends and Technology (IJETT) 9 (2014) 11, pp. 576–582.

[KlTH13] Klein, D.; Tran-Gia, P.; Hartmann, M.: Big Data. In: Informatik-Spektrum 36 (Apr. 2013) 3, pp. 319–323.

[KuWD09] Kuo, Y.F.; Wu, C.M.; Deng, W.J: The relationship among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. In: Computers in Human Behaviour 25 (2009) 4, pp. 887–896.

[Lyce13] Lycett, M.: "Datafication": making sense of (big) data in a complex world. In: European Journal of Information Systems 22 (2013), pp. 381–386.

[MCBB11] Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A. H.: Big data: The next frontier for innovation, competition, and productivity, 2011.

[MFÖK13] Matijacic, M.; Fellmann, M.; Özcan, D.; Kammler, F.; Nüttgens, M.; Thomas, O.: Elicitation and Consolidation of Requirements for Mobile Technical Customer Services Support Systems – A Multi-Method Approach. In: Proceedings of the International Conference on Information Systems (ICIS) 2013, 2013.

[ÖFMD13] Özcan, D.; Fellmann, M.; Matijacic, M.; Däuble, G.; Nüttgens, M.; Thomas, O.: Use Case based description of IT-supported customer service processes. In: Proceedings of XXII. International RESER Conference, 2013.

[ÖsSe06] Österle, H.; Senger, E.: Innovative Geschäftskonzepte im After Sales Service. Erfolgreich mit After Sales Services. Springer, Berlin 2006, pp. 37–54.

[PoMi85] Porter, M.E.; Millar, V.E.: How information gives you competitive advantage. In: Harvard Business Review July-Augus (1985).

[RaMB05] Ray, G.; Muhanna, W.A.; Barney, J.B.: Information Technology and the Performance of the Customer Service Process: A Resource-Based Analysis. In: MIS Quarterly 29 (2005) 4, pp. 625–652.

[RSJG12] Rabl, T.; Sadoghi, M.; Jacobsen, H.A.; Gómez-Villamor, S.; Muntés-Mulero, V.; Mankovskii, S.: Solving Big Data Challenges for Enterprise Application Performance Management. In: Proceedings of the VLDB Endowment, Vol. 5, No. 2, 2012, pp. 1724–1735.

[Russ11] Russom, P.: Big Data Analytics, TDWI Best Practices Report. Renton 2011.

[ScLe02] Schultze, U.; Leidner, D.E.: Studying Knowledge Management in Information Systems Research: Discourses and Theoretical Assumptions. In: MIS Quarterly 26 (2002) 3, pp. 213–242.

[Sofi14] Sofianti, T.D.: Framework for Measuring Knowledge Management for Service Development in Industrial Distributor. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), 2014, vol. II.

[Stil03] Stille, F.: Produktbegleitende Dienstleistungen gewinnen weiter an Bedeutung. In: Wochenbericht des DIW 21 (2003), pp. 336–344.

[ThWL08a] Thomas, O.; Walter, P.; Loos, P.: Product-Service Systems: Konstruktion und Anwendung einer Entwicklungsmethodik. In: Wirtschaftsinformatik 50 (Jul. 2008) 3, pp. 208–219.

[ThWL08b] Thomas, O.; Walter, P.; Loos, P.: Design and usage of an engineering methodology for product-service systems. In: Journal of Design Research 7 (2008) 2, pp. 177–195.

[TWLN07] Thomas, O.; Walter, P.; Loos, P.; Nüttgens, M.; Schlicker, M.: Mobile Technologies for Efficient Service Processes: A Case Study in the German Machine and Plant Construction Industry. In: Proceedings of the Americas Conference on Information Systems (AMCIS 2007), August 9–12, Keystone, USA, 2007, p. 49.

[Vdma08] VDMA: Entscheidungshilfe zur Einführung von Servicemanagementsystemen. VDMA Verlag, Frankfurt am Main 2008.

[Vdma13] VDMA: Maschinenbau in Zahl und Bild 2013. VDMA Verlag, Frankfurt am Main 2013.

[Walt09] Walter, P.: Modellierung technischer Kundendienstprozesse des Maschinen- und Anlagenbaus als Bestandteil hybrider Produkte. In (Thomas O.; Nüttgens M. eds.): Dienstleistungsmodellierung: Methoden, Werkzeuge und Branchenlösungen. Physica, Heidelberg 2009, pp. 129–145.

[Walt10a] Walter, P.: Technische Kundendienstleistungen: Einordnung, Charakterisierung und Klassifikation. In (Thomas, O.; Loos, P.; Nüttgens M. eds.): Hybride Wertschöpfung. Springer, Berlin 2010, pp. 24–41.

[Walt10b] Walter, P.: Hybride Wertschöpfung durch IT-basierte Integration: Am Beispiel des Technischen Kundendienstes im Bereich komplexer technischer Gebrauchsgüter. Logos, Berlin 2010.

[WaPK95] Watson, R.T.; Pitt, Leyland F.; Kavan, B.C.: Service Quality: A Measure of Information Systems Effectiveness. In: MIS Quarterly 19 (1995) 2, pp. 173–187.

[ZaPG10] Zaslavsky, A.; Perera, C.; Georgakopoulos, D.: Sensing as a Service and Big Data. In: Proceedings of the International Conference on Advances in Cloud Computing (ACC 2012), Bangalore, India.

[ZiLa13] Zilch, A.; Landrock, H.: Big Data – kein neuer Hype, aber eine neue Dimension, Experton-Group, Ismaning, 2013. URL: http://www.experton-group.de/research/ict-news-dach/news/article/big-data-kein-neuer-hype-aber-eine-neue-dimension.html (25. April 2013).

Insights, Analytics & Modelling – Wie Reisende den ICE nutzen und nutzen werden

Dr. Jürgen Antes, Dr. Robert T. Firla, Dr. Martin Thust

Business Intelligence Marketing
DB Fernverkehr AG
Stephensonstraße 1
60326 Frankfurt am Main
juergen.antes@deutschebahn.com
robert.firla@deutschebahn.com
martin.thust@deutschebahn.com

Abstract: Geschäftsentscheidungen für die DB Fernverkehr AG basieren auf umfassendem Wissen über aktuell getätigte und prognostizierte zukünftige Reisen. Um entsprechende Entscheidungsunterstützung zu leisten integrieren unsere BI-Systeme dazu zunehmend vielfältigere Datenquellen. Wir beschreiben Hürden bei Aufbau und Nutzung der Systeme aus Bahn-systemischer-, Daten- und Nutzerperspektive sowie unsere Ansätze zu Integration, Strukturierung und Visualisierung zum Umgang damit.

1 Einleitung

Die DB Fernverkehr AG bietet Fernreisemöglichkeiten mit ICE- und IC/EC-Zügen im Linienverkehr für ihre Reisenden an, der mit rund 400 Zügen und etwa 1000 kommerziellen Zugfahrten pro Tag erbracht und von ca. 130 Mio. Reisenden im Jahr genutzt wird. Wir erreichten 2013 ca. 4 Mrd. EUR Umsatz und rund 8% EBIT-Marge [DB2014].

Unsere Geschäftsentscheidungen benötigen umfassendes Wissen darüber, wie unsere aktuellen Angebote von den Reisenden real genutzt werden und Prognosen, wie Reisende dies zukünftig bei veränderten Angeboten tun werden. Beispiele sind Fragen zur Flottenerneuerung (Wie viele Züge? Wie schnell? Wie viel 1./2. Klasse Plätze?), zur Fahrplangestaltung (z.B. neue Strecken) oder zur Kontingentsteuerung der Sparpreise.

Stetig wachsende Herausforderungen (Wirtschaftlichkeit, Wettbewerb) führen auch bei uns zu einem stetig steigenden Bedarf an verfügbaren Informationen bzgl. Genauigkeit, Detaillierung und Vielfalt, damit automatisch zu steigender Komplexität der BI-Systeme und entsprechenden **Nutzungshürden**. Konkret möchten wir drei hervorheben:

- **Systemische**: Im Gegensatz zu Airlines gibt es bei uns **keine vollständigen** flug-, ticket- und personengenauen **Datenerfassung**[1] für alle Einzelreisen, z.B.

[1] Ein Extrembeispiele: die Bahncard 100 erlaubt beliebiges Reisen im Netz der DB ohne weiteres Ticket. Es erfolgt keinerlei Erfassung der damit getätigten Reisen, auch nicht bei der Fahrscheinkontrolle.

beinhalten Normalpreistickets das Recht von einem Start- zu einem Zielbahnhof zu reisen, legen aber weder Zugverbindung noch das genaue Datum fest.

- **Datenseitige**: Wir benötigen daher eine **Vielzahl sehr heterogener Datenquellen**[2] um die notwendigen Entscheidungsgrundlagen zu schaffen.

- **Nutzerseitige**: Hohe **Nutzerakzeptanz** ist nur bei **Vollständigkeit** der Informationen, Anbieten **fachspezifischer Sichten** (z.B. von der Reise her über Quelle-Ziel-Relationen oder vom Fahrplan her über Züge), intuitive **Nachvollziehbarkeit** und **Einfachheit** der Nutzung zu gewinnen.

Dem begegnen wir mit drei Ansätzen: **systematische Zusammenführung** aller relevanten **Daten**, **gleichartige Strukturierung** und damit einfache Vergleichbarkeit für **Ist- und Prognosedaten** sowie **sehr guten Visualisierungsmöglichkeiten**. Im Weiteren erläutern wir die dazu entwickelten „Nachbar"-Systeme TReK für Monitoring und Analyse der Ist-Reisen (Abs. 2), PRIMA zur Prognose und Bewertung von zukünftigen Angebotsszenarien (Abs. 3) und enden mit einer Skizze aktueller Fragestellungen (Abs. 4).

2 TReK – Wie Reisende den ICE/IC heute nutzen

Die Anforderung das laufende Geschäft sehr gut zu verstehen kummuliert in der Frage *„**Wo, wann und wie nutzen Reisende**[3] **den Fernverkehr der DB?**"*. *Wo* fragt nach Start- und Zielort der Reise sowie den dabei genutzten Zügen oder weiter gefasst, der Zugverbindung bzw. der Reisekette. *Wann* fragt nach Reisedatum sowie den einzelnen Zeitpunkten innerhalb der Reisekette (Abfahrt, Umstiege, Ankunft). *Wie* fragt u.a. nach Art des Tickets, Nutzung einer Bahncard, Preis, Klasse, Gruppengröße u.v.a. Die Bereitstellung von Daten zur Beantwortung dieser Fragen ist Aufgabe von TReK.

2.1 Datengrundlagen und Informationsschichten

Zentrale Datengrundlage von TReK sind die Buchungs- und Verbindungsdaten, die beim fahrplanbasierten Verkauf durch die IT-Systeme erfasst werden. **Stufe 1** umfasst die **Buchungsdaten für Sparpreise**, also zuggebundene Tickets inkl. der fest gebuchten Zugverbindung, **Stufe 2** die **Normalpreistickets** inkl. der angefragten Zugverbindungen. Aus Marktforschungen wissen wir, dass das Gros der Reisenden – trotz voller Flexibilität des Normalpreises bzgl. der gewählten Zugverbindung – die beim Ticketkauf angefragte Verbindung auch real nutzt.

Die Daten dieser beiden Stufen liefern also explizite Verbindungsinformationen und somit Transparenz über die Reisekette, umfassen jedoch nur 70% des Reiseaufkommens. Um Nutzerakzeptanz insb. im Punkt Vollständigkeit zu erreichen, sind die fehlenden 30% der Reisen z.B. solche mit Streckenzeitkarten oder Bahncard 100 zu ergänzen.

[2] z.B. Daten aus Verkaufssystemen, Reisendenzählungen, Marktforschungen, Daten zu Schienen und Bahnhöfen, Fahrplandaten, sozio-ökonomische Daten, Umfeldinformationen über Markt und Wettbewerber

[3] TReK kennt keine personenbezogenen Daten sondern nur anonyme Reisen

Dabei handelt es sich um Angebote deren Nutzung z.T. strukturelle Besonderheiten beinhalten (Reisezeitpunkt, Preis, Relation), womit eine einfache Hochrechnung unzweckmäßig ist. Der für **Stufe 3** beschrittene Weg basiert daher auf der **modellbasierten Erzeugung von Reiseketten** für die **fehlenden Angebote** unter Beachtung jeweils verfügbaren Informationen und unter Einsatz des Verkehrsprognosemodells PRIMA (siehe Abs. 3). Da die vorfügbaren Informationen zwischen den Angeboten stark variieren, wurden jeweils separate Modelle entwickelt, die die jeweiligen Reiseketten erzeugen, in identischer Struktur und Detaillierung wie die in Stufe 1 und 2 erfassten Reiseketten.

Der so geschaffene Datenbestand erlaubt es die 130 Mill. Reisen p.a. konsistent, durchgängig, einheitlich, in variierenden Sichten – zeitlich, räumlich, nach Relationen, nach Zügen, nach Tickets – und auch deren zeitliche Entwicklung auszuwerten.

2.2 Von den Daten zur Entscheidungsunterstützung

TReK ist als klassisches fünfschichtiges Data Warehouse-System realisiert:

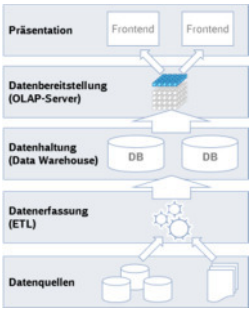


Abb. 1: Architektur und schematischer Aufbau des BI-Systems TReK

Wie bei (fast) allen BI-Systemen sind Beschaffung, Zusammenführung und Qualitätssicherung heterogener Datenquellen zentrale Kernaufgaben und Herausforderungen. Ergänzend zu den Reiseketten sind dies u.a. Infrastruktur-, Fahrplan- und Stammdaten, Prämissen und Geschäftsregeln. Die Datenerfassung und Aufbereitung umfasst auch Einzelmodule die Geschäftsregeln (u.a. Erlösaufteilung) implementieren. Das zentrale Data Warehouse ist mittels mehrschichtiger Datenbanken realisiert, die Datenbereitstellung mittels OLAP-Server. Da es uns letztlich um Erkenntnisgewinn und resultierende Wertschöpfung im Geschäft geht, kommt der Präsentation und dem Zugang zu den Systeminhalten eine besondere Bedeutung zu. Wir setzen auf eine Mischung aus klassischen BI-Komponenten (Standard- und *ad hoc*-Reports mit wachsendem *self-service*-Anteil) und spezifischen, kontextbezogenen Visualisierungen, die wir nun kurz vorstellen.

2.3 Visualisierung von Reisen und Reiseströmen

Nach unseren Erfahrungen ist es für die Nutzerakzeptanz essentiell, dass BI-Systeme den verschiedenen Anwendergruppen nicht „nur“ ein Meer an Daten anbieten, sondern dass dies jeweils fachspezifisch geeignet und zweckmäßig verdichtet geschieht. Visualisie-

rung kommt dabei besondere Bedeutung zu, da diese einen **Weg zur intuitiven Erfassbarkeit von Mustern in größeren Datenmengen** darstellt. Für die Visualisierung von Reisen und Reiseströmen wollen wir dies an zwei Beispielen erläutern.

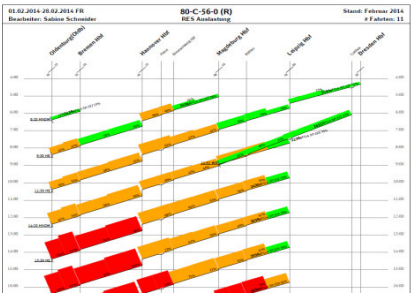


Abb. 2: Visualisierung Reiseströme im Weg-Zeit-Diagramm (Bildfahrplan)

Visualisierung entlang der Zugfahrten: Sogenannte Weg-Zeit-Diagramme (Bildfahrplan) sind eine klassische Darstellung von Fahrplänen. Sie zeigen entlang der X-Achse den Streckenverlauf mit Haltebahnhöfen und entlang der Y-Achse die Zeit. Einzelne Diagonalstriche zeigen Zugfahrten entlang von Ort und Zeit. Wir ergänzen dies durch die Visualisierung frei wählbarer Attribute entlang der einzelnen Zugfahrten: etwa die Balkenbreite als Besetzung und die Farbe als Auslastungsklasse. Vergleiche zwischen oder Identifikation von Strukturen oder Besonderheiten einzelner Fahrten sind auf diesem Wege substantiell einfacher als eine tabellarische Analyse der gleichen Daten.

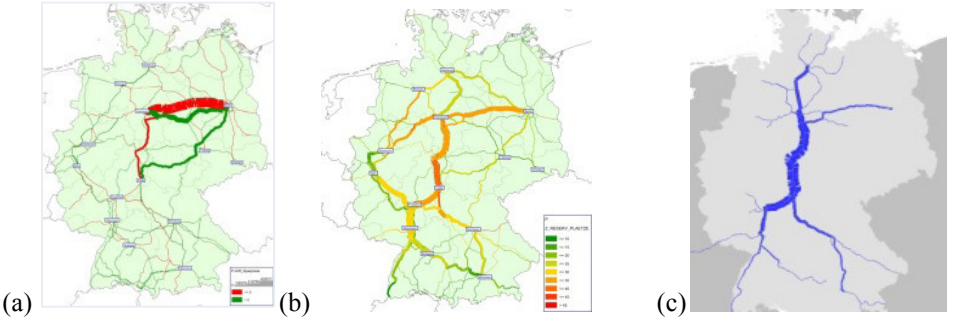


Abb. 3: Räumliche Visualisierung:
(a) Differenznetze, (b) Gesamtnetz inkl. spez. Kennzahlen (c) Spinnendarstellung

Visualisierung räumliche/geographische Dimensionen: Reiseströme in räumlicher Ausprägung darzustellen ist eine weitere Möglichkeit. Während dies für ortsbezogene Daten, z.B. Einsteiger je Bahnhof, ein etabliertes Feature vieler BI-Frontends ist, ist die Darstellung der Reiseströme (aus Streckensicht) deutlich komplexer, da die Einzelinformationen der Reisekette des Ticktes (von Bhf. A zu Bhf. B mit Zug C) den richtigen

Strecken zugeordnet werden müssen (*routing*). Dazu nutzen wir die Verkehrsplanungssoftware Visum[4]. Abb. 3 zeigt Beispiele für verschiedene mögliche Visualisierungen.

Obwohl Visualisierungen (fast) keine neuen Informationen beinhalten, sind wir sicher, dass die Erzeugung „bunter Bilder" nicht nur den Erkenntnisgewinn erleichtert. Vielmehr wird damit ein intuitiverer Zugang zu den Daten, eine hohes Maß an Nachvollziehbarkeit und damit ein deutlicher Beitrag zur Nutzerakzeptanz erreicht.

3 PRIMA – Wie Reisende den ICE/IC zukünftig nutzen werden

Für Investitionsentscheidungen in Infrastruktur oder Fahrzeuge sowie die Gestaltung des Fahrplans genügt es nicht, die Nutzung der Fernverkehrszüge im Status quo zu kennen, sondern es ist auch die zukünftige Nutzung unter veränderten Bedingungen vorherzusagen. Diese Prognose der Wirtschaftlichkeit (Umsatz, Kosten) von Fahrplanszenarien übernimmt das System PRIMA, dessen komplexester Teil die Nachfrageprognose ist.

3.1 Das Nachfrageprognosemodell

Die Prima-Nachfrageprognose ist zeit- und verbindungsfein, erzeugt dabei identische Strukturen (Reiseketten) wie TReK[5] und ermöglicht dieselben detaillierten Auswertungen. Damit ist eine wesentliche Erfolgsvoraussetzung gegeben: direkte und strukturgleiche Vergleichbarkeit der Prognose- zu Ist-Daten. Der Ausgangspunkt des Modells ist ein klassisches 4-stufiges Verkehrsmodell, (z.B. [SL97], [SKV05] Kap. 5B):

1. Erzeugung: *Wie viele* Reisen werden je Verkehrszelle generiert?
2. Verkehrsverteilung: *Wohin* (zu welcher Zelle) gehen die Reisen?
3. Verkehrsmittelwahl: *Womit* (z.B. Pkw, Flug, etc.) erfolgt die Reise?
4. Verbindungswahl: *Wann* und *wo lang*? (hier: Mit welcher Zugverbindung?)

Schwerpunkt der Prognose in PRIMA ist die **Verkehrsmittel- und Verbindungswahl** in **Abhängigkeit** von den **fünf Angebotskenngrößen** einer Bahnreise: Reise**zeit,** Fahr**preis,** Anzahl und Dauer **Umstiege** (zu knapp oder zu lang ist schlecht), **Zeitlage** (Wie oft/ wie gut passt das Angebot zur Nachfrage?), **Auslastung** (Wie viel Platz ist im Zug?).

PRIMA differenziert dazu aktuell 1.340 Verkehrszellen europaweit (460 in Deutschland), 4 Verkehrsmittel, 5 Reiseanlässe und 14 Tagestypen. Datengrundlagen klassischer Modelle sind räumlich aggregierte sozioökonomische Daten und Marktforschungen in Form stichprobenhafter Befragungen. Dieses Vorgehen ist state-of-the-art für neue Verkehrsmittel wie z.B. den Fernlinienbus in Deutschland (Abs. 3.2), für die es noch keine Nutzungsdaten gibt. Andererseits ist dieser Ansatz nicht fein genug auf dem Detailniveau von 13 Mio. Fernzugverbindungen je Tagestyp, die PRIMA generiert. Aktuelle

[4] Firma ptv AG, Karlsruhe
[5] Ausnahme: PRIMA kennt keine Ticketarten und Klassen

Ansätze, wie die genaue, detaillierte Kenntnis des Status quo durch TReK zur Verbesserung der Prognosen verwendet werden kann, werden in Abschnitt 3.3 vorgestellt.

PRIMA wird zur Bewertung von ca. 1.000 Fahrplanvarianten jährlich eingesetzt. Ein Szenario wird innerhalb weniger Stunden, i.d.R. über Nacht gerechnet. Fahrplanszenarien und Bewertungsergebnisse werden in einem eigenen Data Warehouse verwaltet und das aufgesetzte Berichtssystem ermöglicht – analog dem von TReK – standardisierte Auswertungen in zahlreichen Dimensionen insbesondere auch den paarweisen Vergleich von Fahrplanszenarien bzw. von Szenarien mit Ist-Daten.

Der hohe Automatisierungsgrad von PRIMA legt die Nutzung als *black box* nahe. Es ist uns jedoch wichtig, die Bewertungslogik transparent zu halten, damit Fahrplaner ihre eigene Einschätzung dagegen reflektieren und erfolgsversprechende Ideen generieren können. Auch hier helfen die Visualisierungsverfahren wie in 2.3 dargestellt erheblich.

3.2 Gemeisterte Herausforderungen: Fernlinienbus und Auslastung

Seit 2013 ist der Fernbusmarkt in Deutschland liberalisiert und Fernbuslinien dürfen nun in Konkurrenz zur Schiene angeboten werden. Der Markt entwickelt sich extrem dynamisch: Mittlerweile verkehren täglich mehr Fernbusse als Fernzüge im Linienbetrieb. Da es vor 2013 i.W. kein Fernlinienbusangebot gab, also Reisende im realen Verhalten nicht beobachtet werden konnten, war es eine große Herausforderung, die **Auswirkung des Fernbusses** als neues Verkehrsangebot **auf die Bahnnachfrage** vorherzusagen.

Der empirische Input für das PRIMA-Verkehrsmittelwahlmodell ist eine Marktforschung in Form eines individualisierten *choice based conjoint* [GSG09] mit 5.700 Befragten im Fernverkehr. Jedem Befragten wurde zu seiner berichteten Reise neben den drei realen Alternativen Pkw, Bahn und Flugzeug auch der Fernbus als hypothetische Alternative angeboten. Die Alternativen wurden dabei durch Angebotskenngrößen beschrieben und die Wahlentscheidungen bei Variation selbiger festgehalten. Das auf Basis der Marktforschung mit dem *Maximum Likelihood* Ansatz ([BAL85], Kap. 2) geschätzte Verkehrsmittelwahlmodell prognostiziert, dass sich ein Fernbusangebot parallel zum Kernnetz der DB Fernverkehr AG rentieren kann, dass jeder 4. Busnutzer von der Bahn wechselt und sich die Busnachfrage auch bei steigenden Buspreisen in den kommenden Jahren etwa verdoppeln könnte.

Eine weitere große Herausforderung war die **Integration der Angebotskenngröße „Auslastung“** in das Nachfragemodell PRIMA. Zuvor konnten Züge im Modell ohne Rückwirkung auf die Nachfrage beliebig und unrealistisch überlastet werden, Verdrängung wurde nicht berücksichtigt. Aus Marktforschungen ist bekannt, dass ein einfaches Abschneiden der Besetzung bei 100% Auslastung kein geeignetes Modell ist: Die ersten Kunden springen bereits bei 50% Auslastung ab, die letzten sind sogar bereit zu stehen.

Die Integration der Kenngröße „Auslastung“ in das Modell war vor allem algorithmisch schwierig. Anders als die übrigen vier Kenngrößen hängt die Auslastung direkt von der prognostizierten Nachfrage ab, wodurch sich ein negativ rückgekoppelter Wirkungszusammenhang ergibt: Hohe Auslastung senkt die Bahnnachfrage, diese wiederum senkt

die Auslastung. Um den Gleichgewichtszustand ([SKV05], Abschnitt 5.10) zwischen Angebot und Nachfrage zu finden, wurde ein intelligent gedämpftes, iteratives und sehr performantes Verfahren entwickelt. Die notwendige zweistellige Anzahl von Iterationen erfordert nur die doppelte Rechenzeit ggü. dem Modell mit nur vier Kenngrößen.

3.3 Aktuelle Herausforderungen und Ansätze zur weiteren Modellverbesserung

Planungen werden immer detaillierter (Einzelzüge, Verkehrstage) und stellen höhere Ansprüche an PRIMA: Hohe Prognosegüte auf hohem Detailniveau. TReK ist dabei für PRIMA Herausforderung und Chance zugleich. Es ermöglicht einerseits den Vergleich von Prognose vs. Ist auf maximalem Detailniveau und deckt damit die Schwächen eines Prognosemodells auf: Je höher das Detail, desto ungenauer sind statistische Modelle. Andererseits kann TReK eine Datengrundlage für ein besseres Modell sein. Wir sehen zwei Ansätze, Informationen aus TReK für PRIMA zu nutzen:

1. **Schätzung der Verhaltensparameter der (Zug)Verbindungswahl**
   Als Alternative zu hypothetischen, stichprobenhaften Aussagen einer Marktforschung liefert TReK für jede Relation explizit, unter welchen Angebotsbedingungen welche Bahnverbindungen in der Realität gewählt werden.
2. **Inkrementelle Nachfrageprognose auf Basis IST-Nachfrage**
   Aus PRIMA wird nur die Änderung der Nachfrage übernommen und auf die tatsächliche Nachfrage aus TReK angewendet. Gegenüber herkömmlichen Ansätzen [DFP12] möchten wir auf der Ist-Nachfrage nicht nur in räumlicher, sondern auch in zeitlicher Schneidung aufsetzen.

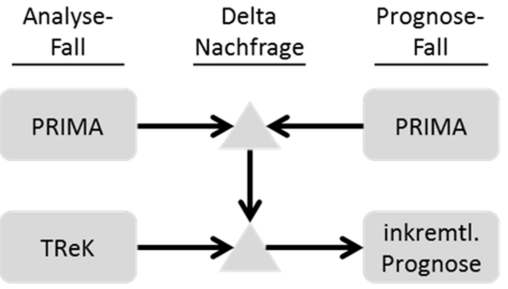


Abb. 4: Inkrementelle Nachfrageprognose auf Basis TReK

Um zu prognostizieren, wie viel Nachfrage die Bahn wann und wo von anderen Verkehrsmitteln gewinnen kann, ist es wichtig, deren Quelle-Ziel-Nachfrage genau zu kennen. Während das für den Flugverkehr aus Angebot und Buchungen der Fall ist, ist der MIV (motorisierte Individualverkehr) mit 80% Verkehrsleistungsanteil im Personenfernverkehr die große Unbekannte. Hier versprechen wir uns deutliche Fortschritte durch Nutzung anonymisierter Bewegungsdaten von Navigations- oder Mobilfunkgeräten.

Beide Datenquellen haben aber noch ihre Unzulänglichkeiten: Die Abdeckung von verfügbaren anonymisierten Pkw-Navigationsdaten beträgt derzeit nur etwa 1% des MIV-

Marktes und ist im Detail so disproportional, dass eine brauchbare Hochrechnung auf den MIV-Markt nicht ohne weiteres möglich erscheint. Die Verwendbarkeit von Mobilfunkdaten wird durch Anonymisierungstechniken sehr stark eingeschränkt. Der regelmäßige Wechsel der Telefon-ID verhindert die Nachvollziehbarkeit von längeren Fernreisen und die Nicht-Weitergabe von Einzelereignissen detaillierte Auswertungen.

4 Aktuelle Fragestellungen und Ausblick

Die durchgängige Digitalisierung der Gesellschaft und der Dinge (*internet of things*) prägen die Erwartungshaltungen unserer internen Kunden und die der Reisenden. Erstere erwarten einfachsten Zugang zu allen, stets top-aktuellen Daten auf allen Endgeräten. Letztere lernen, dass individuell relevante Ansprachen und Informationen von Unternehmen anderer Branchen beherrscht werden und erwarten dies zunehmend auch von uns. Daher halten wir umfassende, gut zugängliche Datengrundlagen und Fähigkeit diese als Marketingorganisation zweckorientiert nutzen zu können, für wesentliche Schlüsselkompetenzen. Unsere aktuellen Herausforderungen liegen nun insbesondere darin das Nützliche und Zweckmäßige vom technisch Möglichen zu trennen. Einige herausgehobene Themen und *trade-off* dazu sind:

- Wie kann viel *self-service* erreicht werden und wie viel/welche *Governance* soll dabei sein (u.a. Agilität Fachbereiche vs. Risiko Fehlinterpretationen)?
- Ist BI auf mobilen Endgeräten nützlich oder bloß schick?
- Wie aktuell müssen Daten sein ggü. wie vollständig/strukturiert?
- Welche Möglichkeiten zur persönlich relevanten Ansprache der Reisenden können wir u.a. mit neueren *Big-Data, predictive analytics* Ansätzen realisieren (z.B. *churn rate reduction*, *next best offer*, *liquid feedback*, …)?

Die DB Fernverkehr AG strebt danach in den Augen der Reisenden und der Mitarbeiter erste Wahl zu sein. Wir freuen uns mit unseren Themenstellungen in Zeiten gesellschaftlicher Veränderungen und einem hochdynamischen technischen Umfeld dazu einen Beitrag zu leisten.

Literaturverzeichnis

[BAL85] Ben Akiva, M.; Lerman, S.: Discrete Choice Analysis: Theory and Application to Travel Demand, MIT Press 1985

[DB14] Geschäftsbericht DB AG zum Jahr 2013

[DFP12] Daly, A.; Fox, J.; Patruni, B; Milthorpe, F.: Pivoting in Travel Demand Models. BTS Conference Paper, Australasian Transport Research Forum, 26-28 September 2012

[GSG09] Günthel, D.; Sturm, L.; Gärtner, C.: Anwendung der choice-based-conjoint-Analyse zur Prognose von Kaufentscheidungen im ÖPNV. TU Dresden. Diskussionsbeiträge aus dem Institut für Wirtschaft und Verkehr Nr.1/2009

[SKV05] Steierwald, S.; Künne, H-D.; Vogt, W.: Stadtverkehrsplanung: Grundlagen, Methoden, Ziele, Springer Berlin Heidelberg New York 2005

[SL97] Schnabel, Lohse: Grundlagen der Strassenverkehrstechnik und der Verkehrsplanung. Verlag für Bauwesen, Berlin, 1997.

Bereitstellung und Verwaltung von portablen und sicheren Cloud-Diensten

(CloudCycle14)

Bereitstellung und Verwaltung von portablen und sicheren Cloud-Diensten (CloudCycle14)[1]

Abstract: Der Workshop „Bereitstellung und Verwaltung von portablen und sicheren Cloud-Diensten“, organisiert durch die Universität Stuttgart und das TrustedCloud-Projekt „CloudCycle“, fand am 22. September 2014 im Rahmen der 44. Jahrestagung der Gesellschaft für Informatik in Stuttgart statt.

Ausgehend vom Leitthema der Jahrestagung, Big Data – Komplexität meistern, adressierte der Workshop die Frage, wie Cloud-Dienste sicher und portabel bereitgestellt und verwaltet werden können. Beides, Sicherheit und Portabilität, ist nötig um die Cloud gewinnbringend als eine Plattform für Big Data Anwendungen zu nutzen.

Neben den beiden eingeladenen Vorträgen aus Industrie und Forschung, „Next Generation Cloud Architecture” (Gerd Breiter, IBM) und „Building Applications for the Cloud: The Wrong, the Right, the Solid Way” (Frank Leymann, Universität Stuttgart), wurden 9 begutachtete Beiträge vorgestellt. Diese wurden von Autoren der Universitäten Bamberg, Pisa, Stuttgart und Tübigen, den Fraunhofer Instituten IAO, IDG, IESE und SIT, sowie innoQ, der ENX Association und der medisite Systemhaus GmbH eingereicht. Die ersten beiden Sessions befassten sich mit den beiden Kernthemen des Workshops, „Sicherheit“ und „automatisierte Bereitstellung und Verwaltung von portablen Diensten“. Abgerundet wurde das Programm durch eine Session mit begutachteten Beiträgen aus den Projekten SkIDentity, TRESOR und CloudCycle des durch das BMWi-geförderten Technologieprogramms „TrustedCloud“.

Prof. Bernhard Mitschang,
Prof. Frank Leymann,
Dr. Matthias Wieland,
Tobias Binz

[1] http://www.cloudcycle.org/workshop/

Enhancing Cloud Security with Context-aware Usage Control Policies

Christian Jung, Andreas Eitel, Reinhard Schwarz

Fraunhofer Institute for Experimental
Software Engineering IESE
Kaiserslautern, Germany
{christian.jung, andreas.eitel, reinhard.schwarz}@iese.fraunhofer.de

Abstract: Cloud environments strongly rely on virtualization infrastructure that provides virtual resources by abstracting from the physical hardware. Thus, cloud providers can cost-efficiently share physical hardware among multiple tenants, and a single virtual resource may span multiple physical resources at different geo-locations. From a tenant's perspective, the uncertainty about location and context of virtual resources is a potential security threat. For instance, tenants may want to enforce geo-fencing to prevent their applications and data from migrating to undesirable jurisdictions, untrusted co-tenants, or dubious locations. They may also want to ensure that certain virtual resources share (or expressly do not share) a common physical resource, for example, to improve fault tolerance or performance. To tackle these problems, we suggest a flexible policy decision and enforcement framework for enabling usage control in cloud environments. In support of this framework, we collect additional information from the cloud environment to enforce context-aware and therefore more fine-grained usage control policies. Our solution offers flexible controls for secure and resilient cloud management. The paper presents our policy enforcement framework IND^{2}UCE and its extension to enable context-ware policy enforcement on an exemplary cloud infrastructure using VMware products.

1 Introduction

The Cloud offers organizations the chance to move their IT services to foreign operated cloud environments. Due to the economy of scale, IT outsourcing is often more cost-efficient and more flexible than in-house service provisioning, and professional cloud operation can even provide more security and resilience.

However, for being competitive, the cloud providers will run their systems close to the tolerable load limit. To reach their goal, they may share each of their physical resources among multiple customers, even among direct competitors or tenants of varying trustworthiness. In general, sensitive data are mixed with nonsensitive data, and critical services are running together with noncritical services. This unpredictability of runtime context raises security concerns among potential cloud users.

To explore the cloud users' security needs and reservations, we invited about 60 different

companies to participate in a survey on cloud security and critical infrastructure IT [RS14]. Participants stated cost reduction, scalability, elasticity and especially availability, reliability, and resilience as the main advantages of cloud infrastructures for their organizations. However, the majority of the respondents claimed security and privacy as well as loss of control over key IT systems and the infrastructure itself as their main concerns regarding service deployment in cloud infrastructures. Our survey endorses results from previous surveys, for instance, performed by ENISA [CH09] or the TClouds project [BCH13].

Seemingly, the multi-tenancy problem, but also not knowing where data is really stored, prevents security and privacy aware enterprises from shifting their business software into cloud environments. To lower the barriers for cloud adoption, we need to put cloud users back into control by providing them security policies and corresponding mechanisms to express and enforce their protection needs for sensitive applications and data.

Contribution. We present an approach for security policy enforcement across multiple cloud abstraction layers. To this end, we adapted our policy enforcement framework IND^{2}UCE to the VMware virtualization environment, and we provided several enforcement components for VMware, such as policy enforcement and execution points as well as a policy information point. These components interact with the virtualization environment by retrieving events and by triggering detective, preventive, or corrective actions within the cloud environment when predefined conditions occur. Based on this framework, we can formulate usage control policies that are monitored and enforced in a VMware cloud.

2 Related Work

In [HKK12], Hamlen et al. propose a framework that contains different modules for policy decision and enforcement, such as a reasoning module. They propose to have control over data by replacing specific instruction (e.g., read method invocations) by surrounding the code with their own check (i.e., intercepting the method invocation and asking the reasoning module for a decision) before further processing.

The approach by Hamlen et al. is very similar to our IND^{2}UCE framework. However, we are focusing on controlling the management capabilities of the cloud environment with focus on the infrastructure level rather than the applications running in the cloud. Our future research will address security policies linking service level and infrastructure level. Furthermore, we provide a mechanism for context-aware enforcement of security policies.

Goyal et al. [GM09] present a “policy-based event-driven services-oriented architecture”. They state that policies are a set of rules that control behavior and usage of services. The services are managed by hierarchically distributed controllers named “mediators”.

Their architecture is focusing on the service layer, more specifically, on the enterprise service bus (ESB). The relation to the cloud is given, as such services may run within the cloud, but is not cloud-specific. Hence, our approach could be complementary to the approach presented by Goyal et al.

For KAoS [IHM13] was originally designed as an agent platform, but eventually became a set of platform-independent services for defining policies for the security, predictability, and controlability of both agents and traditional distributed systems. For policy enforcement, the KAoS Policy Services Architecture is provided. However, KAoS does not specifically address cloud environments.

VMware and other vendors developing virtualization software for cloud environments offer powerful management capabilities to their users (including security policies). However, they are focusing on their technologies and are often limited to specific system events. In our approach, policies can be defined for all available events in the system. Moreover, contextual information can be specified flexibly and can go beyond the information provided by the cloud infrastructure (e.g., geolocation or co-location of services).

3 Usage Control

Usage control extends access control by adding restrictions to future data usage once data access has been granted [PS04, PHB06, PS02]. Typical technologies for implementing enforcement mechanisms for usage control include ad-hoc solutions, runtime verification and complex event processing mechanisms.

3.1 The IND^{2}UCE Framework

The IND^{2}UCE framework ("Integrated Distributed Data Usage Control Enforcement"), illustrated in Figure 1, enforces usage control policies at different abstraction layers of a system. The framework is related to XACML [OAS10], a standard describing declarative access control policies and a processing model for it.

PDP	Policy Decision Point
PEP	Policy Enforcement Point
PXP	Policy Execution Point
PIP	Policy Information Point
PRP	Policy Retrieval Point
PMP	Policy Management Point
PAP	Policy Administration Point

Table 1: IND^{2}UCE Framework Component Abbreviations

The central component of the IND^{2}UCE framework is the PDP (cf. Table 1 for a list of abbreviations), a generic reasoning component. Each policy loaded into the PDP is represented by a state machine controlling policy evaluation. If external information or entities are referenced in the policy, the PDP will contact other framework components, such as PIPs for context information retrieval or PXPs for executing policy directives. Depending on the events that affect or are affected by the usage control policy, the PDP

subscribes to one or more PEPs.

The role of the PEPs is to intercept the policy-relevant events in the system and to enforce the decisions of the PDP on them. A PEPs submits intercepted events to the PDP, where they are analyzed according to the specified policies. In contrast to the PDP, the PEP is a technology-dependent component, which is integrated into the software or system to interact with events.

Basically a distinction is made between preventive and detective enforcement. The latter does not prevent policy violations, but merely detects them to ensure accountability, whereas the former alternatives actually enforce policy compliance. For preventive enforcement, the PEPs are intercepting and blocking events; corrective actions or notifications for detective enforcement are performed by PXPs. In contrast to the PEP, the PXP is directly triggered from the PDP and not by a system event. Hence, a PXP cannot intercept or interact with events, it only executes additional actions, such as sending notifications, on behalf of the PDP. Enforcement is explained in more detail in Section 3.2.

In the XACML OASIS Standard [OAS13], the PIP is described as a "system entity that acts as a source of attribute values". Missing attribute values are added by a so-called context handler component, which can happen with or without request by the PDP.

Accordingly, our PIP component offers additional information for the decision making in the PDP and has to be integrated into the system being observed. PIP attributes may, for example, refer to contextual or information flow tracking data, such as geo-location or network connectivity. Contrary to the XACML framework, our PIP is directly connected to the PDP without having a context handler in between, so that the PDP can configure the PIP according to policy information needs. After configuration, the PIP collects and maintains all required information autonomously.

We differentiate between two components that handle the specification and management of usage control policies — the PAP and the PMP. The PAP provides a human-computer interface to specify usage control needs as policies. It transforms those demands into a machine-readable format.

The PMP is responsible for managing all IND^{2}UCE components. It is the first component to be available in the system. All other components have to register at the PMP and have to name their available connection interfaces. If a PDP wants to subscribe to specific events, the first step is to ask the PMP for an appropriate PEP for these events. Moreover, the PMP is also responsible for deployment and revocation of usage control policies.

The PRP provides a secure policy storage, protecting them against malicious modification. The PRP is only accessed by the PMP for policy management and by the PDP for policy retrieval.

3.2 IND^{2}UCE Enforcement Mechanisms

Policy enforcement is based on an event-condition-action (ECA) paradigm, which is reflected by the XML structure of the policy specifications. An event is an observable system

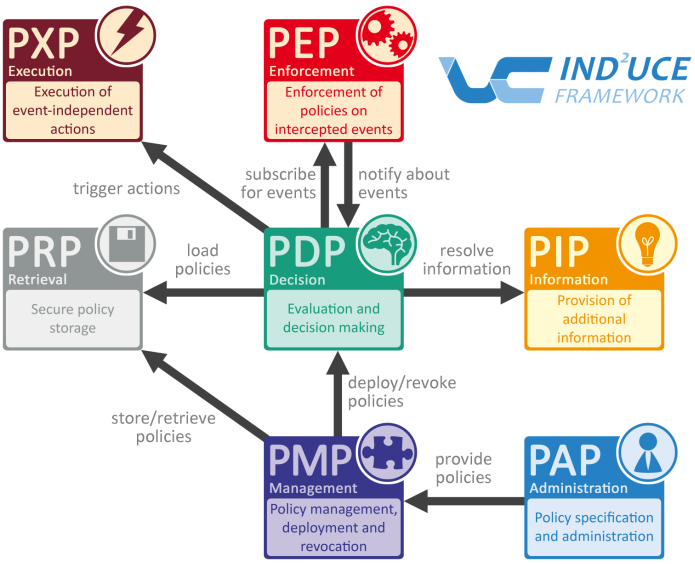


Figure 1: IND^{2}UCE Framework

activity happening at a given instant of time. Whenever an event is intercepted by a PEP, it is forwarded to the PDP. If the event matches the specified event description in a policy mechanism, the condition part of the mechanism is evaluated. If the condition evaluates to *true*, the instruction specified in the action part is executed.

In preventive policy mechanisms, the intercepted system event is caught before it is executed in the observed system. IND^{2}UCE offers four enforcement types for preventive mechanisms. To enforce, for example, the policy "no personal data is allowed to leave the system" on event "send message", there are the following options:

- *inhibition:* The event (i.e., sending the message) is rejected or simply dropped.
- *modification:* The event content is modified (e.g., person-related message fields are erased) before the event is permitted.
- *delay:* The event is delayed until a compensatory action has been executed (e.g., manual permission is requested before sending the message).
- *execution:* The event is permitted, but an additional action is performed (e.g., sending a notification to a privacy monitor).

Obviously, there is a fifth implicit enforcement type *allow*, which is the standard case. In the example, an *allow* decision from the PDP instructs the PEP to permit the message exchange.

The desired behavior is specified in the policy's authorization part. The behavior can also depend on the successful execution of an additional actions, so-called execution actions,

such as writing a log entry or sending a notification email. These additional actions are executed by an appropriate PXP component.

Contrary to preventive policy mechanisms, detective policy mechanisms contain only execution actions but cannot prevent undesired system events. Such a behavior may be useful for compensating undesired but uncritical system actions or for a billing system just sending notifications that a particular action has been executed.

Trigger Events For preventive policy mechanisms, the specification of a trigger event is mandatory, whereas for detective ones it is optional. If the trigger event is not specified, the condition is evaluated at every timestep with a granularity specified by the policy author (e.g., once every 30 seconds).

Conditions The condition part of a policy mechanism can simply be *true*, or it can be expressed in past temporal logic. We are using the Obligation Specification Language (OSL) [HPB+07] for expressing restrictions in the condition. OSL provides propositional, cardinal, and temporal operators. The condition part can also refer to additional attributes from PIPs, such as contextual or information flow tracking information. Regarding cloud environments, for instance, we can retrieve information on resource load of the virtual machines from a PIP that can be included in the decision making process.

Authorization and Execution Actions The action section of a policy mechanism may consist of several blocks: the authorization action and optionally several execution action blocks. For preventive mechanisms, an authorization block is mandatory in the policy, as the PEP needs a decision how to proceed with the intercepted event. If the condition evaluates to *true*, the specified authorization is returned to the PEP that intercepted the event to allow, modify, delay, or completely inhibit the execution of the desired action.

3.3 Context-specific Extensions to the IND^{2}UCE Framework

As mentioned before, the condition part can contain OSL operators as well as references to additional information sources (i.e., PIPs). Hence, the framework may be enriched with components for additional information retrieval. Such a component could be a *Context-PIP* that collects, aggregates, and evaluates arbitrary information.

To provide separation of concerns and to decouple the Context-PIP's tasks from one another, Context-PIPs are designed as layered components (see Figure 2). A fully developed Context-PIP implementation usually contains three layers: The *sensor layer* provides several sensor implementations that collect raw context data from the system under observation. The *data layer* persistently stores all collected sensor data in a suitable format, and it schedules the appropriate sensors as required by the configuration. The *evaluation layer* evaluates the aggregated sensor data. To this end, it contains multiple *evaluators*, which we will be described in Section 4. The evaluation layer consumes a context description as

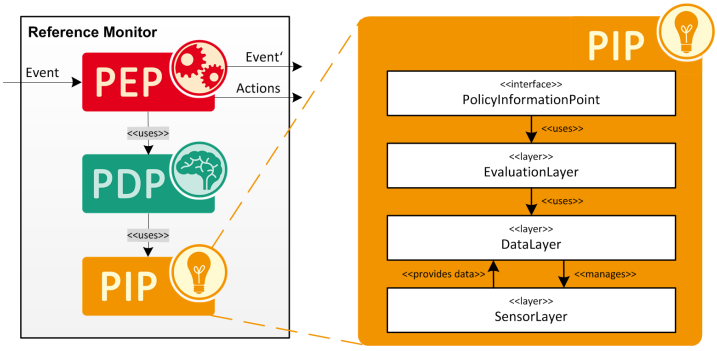


Figure 2: Context-PIP Layered Architecture

configuration, derives a context expression, and provides a unique identifier for referencing the context information. This reference is used by the PDP component for resolving context-dependent policy attributes.

Not all Context-PIP implementations have to provide all these three layers. In the simplest case, they have to provide only the evaluation layer, which is the entry point for configuring contexts and retrieving their evaluation result.

4 Policy Enforcement for VMware

We have chosen the VMware vSphere Hypervisor for implementing our policy enforcement. VMware is one of the major hypervisor vendors. However, VMware software is closed source, which limits our modification possibilities for hooking into the hypervisor. Our prototypical VMware cloud includes two physical hosts running VMware vSphere Hypervisor ESXi. A VMware vCenter Server manages the physical hosts as well as the virtual resources (e.g., network, CPU, memory). Virtualization operators can use either the client software VMware vSphere Client or a web interface for system management. The VMware vSphere Client uses a SOAP interface for interacting with the server. The web interface runs directly on the server using proprietary interfaces.

4.1 Hooking into VMware

We identified three viable solutions for connecting the IND^{2}UCE components with the VMware environment, as explained in more detail in the following.

4.1.1 Man in the Middle Approach

One approach for hooking into the VMware management communication is to insert a proxy component between vSphere Client and vCenter Server. This approach would require to implement our own SOAP interface that replaces the existing SOAP interface provided by VMware and to integrate our enforcement components directly in the code of the proxy implementation in order to control the event flow between client and server.

A clear advantage is that the proxy can intercept and then inhibit, modify, or delay any event before it will be performed on the VMware vCenter Server. This would enable us to implement both preventive and detective policy mechanisms, which offers comprehensive control over the virtualization infrastructure.

However, there are also some disadvantages. First, we would have to reimplement all SOAP interfaces, which is a tedious task. Second, the man in the middle approach affects the security of SSL/TLS in the system. The connection between client and server is protected using SSL/TLS, so we need to establish our own SSL/TLS connection between proxy and client and a second connection between proxy and server. Hence, we would need to provide our own X.509 certificates for the proxy, but deploying such digital certificates to our enforcement framework might be an undesirable modification for a cloud operator. Third, there might be events that are rejected by the system due to insufficient access privileges. Thus, we would have to check whether a specific user has the specific privileges by trying the event on the system, or we would have to implement the permission checks by our own (e.g., by querying the vCenter Server and also the Single-Sign-On server of VMware). To our knowledge, trying whether an event would succeed is not possible, and rebuilding the permission checking would again cause significant implementation overhead. Therefore, both solutions seem to be not very attractive. Moreover, a user can easily bypass the proxy by connecting directly to the VMware environment web interface.

4.1.2 vSphere Client Plugin

Another approach is to implement a vSphere Client plugin. This plugin could replace critical functions in the user interface with our IND^{2}UCE-enabled implementations. In this approach, the enforcement components run within the plugin.

Similar to a proxy, the plugin can intercept all events before they are performed on the server. Hence, a plugin approach enables both preventive and detective enforcement mechanisms.

However, this solution has similar drawbacks as the proxy approach. First, to intercept all possible user interactions, we need to rebuild the entire control center user interface. Second, user roles and privileges need to be handled within our enforcement framework. Third, from a security perspective, a client-based solution is undesirable as we lack control over the client configuration, so users can disable or circumvent it easily or simply connect to the original, unmodified web interface.

4.1.3 SOAP Interface vCenter Server

To avoid these drawbacks, a third option is to build a client software that connects to the SOAP interface of the vCenter Server and that is able to read system events and perform actions on the virtualization cluster. Such a software needs to retrieve events regularly from the management interface.

Such an approach is unaffected by VMware changes (except changes regarding the SOAP interface itself), and it does not depend on any other components or systems interacting with the vCenter Server. Hence, it will not disturb any other systems interacting with the cluster. Furthermore, instead of reimplementing all interfaces or rebuilding the control center user interface, we only have to implement suitable functions for intercepting and manipulating VMware events and for performing actions on the system. In particular, we do not have to take care of the user roles and privileges, as we are only observing events that already happened. The solution does not depend on the origin of the user or system interaction that it controls (thus, it cannot be circumvented), as long as events are produced.

However, the main disadvantage of this approach is the missing preventive enforcement. We can only monitor events that already happened and can perform some compensatory or corrective actions, but we cannot prevent the events from happening.

4.2 Implementation

Our current implementation is based on the latter approach. Although we loose preventive enforcement, it is not possible to circumvent or bypass our solution by simply using another interface or by accessing the ESX hosts directly, as all events will eventually be collected in the vCenter Server. Furthermore, IND^{2}UCE instrumentation can be transparently enabled and disabled without negatively affecting other components or systems interacting with the virtualization environment.

Our solution is based on Java and we implemented a PEP and a PXP for intercepting events and executing actions as well as a PIP component for retrieving additional information from the vCenter Server. All components run on a separate machine and connect to the vCenter Server via the SOAP interface. The components are using the VI Java API [Jin14].

The PEP implementation contains two main components: the EventCollector and the EventConverter. The *EventCollector* connects to the vCenter Server to retrieve events from the event history. This is done by polling for new events once a second. New events are forwarded to the EventConverter. We implemented several standard converters for various data types that could be part of a VMware event (e.g., BooleanConverter, IntegerConverter, CalendarConverter, EntityConverter). Thus, we are able to process arbitrary events as long as they introduce no new data fields in the VMware event.

Overall there are about 700 event types in VMware, and we tested 230 specific event types. We assume we can support nearly all 700 event types with the chosen approach. Events may affect various cloud aspects. For example, they can refer to virtual machines

(e.g., migration, life-cycle, or power-cycle events) as well as the entire cluster (e.g., life-cycle, resources, or high availability services). There are six event categories: cluster, hosts, virtual machines, data stores, networking, and roles/permissions. All these event types and their specific parameters can be used to define the trigger event in our policy specification.

The PXP component can perform specific actions on the VMware cluster. For our prototype, we implemented several actions in the categories virtual machines, cluster, and roles/permissions. Hence, we are able to control the power-cycle (e.g., PowerOn/Off, Reset, Suspend, Standby, Shutdown, Reboot) and the life-cycle (e.g., Reconfig, Relocate, Migrate, Clone, CreateSnapshot) of virtual machines running in the virtualization cluster. These actions and their parameters can be specified in the action part of the mechanisms.

Our PIP component retrieves additional contextual information from the virtualization cluster during the decision making of the policy mechanism. *Evaluators* retrieve and aggregate information from the vCenter Server. Due to their generic implementation, they can be used to evaluate any kind of data from the vCenter Server. For example, they can evaluate configuration details of a virtual machine or a physical host as well as run-time information, such as resource load or storage usage. Evaluation results are associated with *expressions*. We differentiate between logic, arithmetic, and comparison expressions. Expressions and evaluators may be nested and combined to model "contexts of contexts", and we can perform calculations and comparisons.

5 Policy-Enhanced Cloud Security

The following scenario may serve as an example for policy-enabled security enhancement in the cloud. Assume a tenant runs two critical infrastructure services on different virtual machines (VMs) of a cloud datacenter. Due to resilience requirements, the tenant requires that these critical services must not share the same physical resources. If the tenant or the cloud infrastructure operator starts migrating VMs to the same physical host, both critical services would run on the same hardware, which would violate the tenant's security constraint. Using IND^{2}UCE, we can prevent co-location by specifying a corresponding anti-affinity policy. The policy specifies that migrating critical VMs to the same physical host is automatically resolved, for example, by migrating one critical service away.

Although VMware offers affinity and anti-affinity policies, their virtualization management allows policy violation. In contrast, the IND^{2}UCE framework enforces policy compliance. Moreover, our context-aware extension allows smarter corrective mechanisms taking into account the overall load situation and preventing life-locks (e.g., ping-pong migrations). In addition, IND^{2}UCE not only provides VMware-specific actions, such as creating recovery points (VM snapshots), but can also interact with external components, such as a hardware firewall.

6 Conclusion and Future Work

In response to the emerging need for improving security in cloud environments, we adapted our policy enforcement framework IND^{2}UCE to offer usage control capabilities to VMware. We discussed different options how a policy enforcement and execution point can interact with VMware. With the chosen strategy of our research prototype, we cannot prevent any system event, but we can only react after the event has already happened. Therefore, we will explore other possibilities to interact with the cloud environment to enable preventive policy mechanisms, at least for some critical events. On the other hand, our enforcement components are only interacting with the management interfaces of the cloud environment, which minimizes interference and reduces maintenance effort of the IND^{2}UCE instrumentation.

We have begun to implement policy enforcement framework and hooks to interact with a cloud storage environment, based on HBase and Hadoop. Further research will investigate how policy enforcement components, more specifically PEPs and PXPs, can also be implemented for higher layers of cloud environments (e.g., for the service layer) and how policy interaction between infrastructure and service layer can further improve security. For instance, an application running in the service layer could produce events that trigger the creation of snapshots of the virtual machines or create new instances in the virtualization layer. We have to investigate how a secure communication can happen between applications running in the service layer and our policy decision and enforcement framework placed in the management network of the cloud environment.

To explore context-aware policy enforcement, we developed a component for retrieving contextual information from the cloud environment. Based on context information such as geo-location or physical co-location, we were able to enforce affinity, anti-affinity, and geo-fencing policies that can be used to improve resilience and data privacy, but also performance of cloud applications. Further work will research what kind of information is most useful for improving cloud security and how such information can be modeled in a user-friendly manner.

Currently, our machine-readable security policies written in XML are usually created with a simple text editor. This process is error-prone and can only be used by experts that know the available events and system actions that can be used within the policy description. Hence, we started to research approaches on how to build user-friendly specification interfaces that allow even unskilled users to specify their security demands. The use of different usability pattern and how different user groups with varying skill levels and expertise perceive their support will be studied further.

Finally, we will investigate policy deployment and redeployment to the cloud in a secure manner to realize, for instance, distributed policy decision and enforcement. We also need to distinguish between different policy scopes: server-attached policies (e.g., exclusive resource usage), service-attached policies (e.g., service-logic and behavior-dependent policies), and request-attached policies (e.g., dynamic adaptation depending on policies attached to data in the request).

Acknowledgment The research presented in this paper has been funded by the European Commission in the context of the Research Framework Program Seven (FP7) project SECCRIT (Grant Agreement No. 312758).

References

[BCH13] Roland Burger, Christian Cachin, and Elmar Husmann. Cloud, Trust, Privacy – Trustworthy cloud computing whitepaper. Whitepaper, TClouds Project, 2013.

[CH09] Daniele Catteddu and Giles Hogben. An SME perspective on Cloud Computing. Survey, ENISA, Nov 2009.

[GM09] P. Goyal and Rao Mikkilineni. Policy-Based Event-Driven Services-Oriented Architecture for Cloud Services Operation & Management. In Cloud Computing, 2009. CLOUD '09. IEEE International Conference on, pages 135–138, Sept 2009.

[HKK12] Kevin W. Hamlen, Lalana Kagal, and Murat Kantarcioglu. Policy Enforcement Framework for Cloud Data Management. IEEE Data Eng. Bull., 35(4):39–45, 2012.

[HPB+07] M. Hilty, A. Pretschner, D. Basin, C. Schaefer, and T. Walter. A Policy Language for Distributed Usage Control. In Joachim Biskup and Javier Lpez, editors, Computer Security – ESORICS 2007, volume 4734 of Lecture Notes in Computer Science, pages 531–546. Springer Berlin Heidelberg, 2007.

[IHM13] KAoS Policy Services Framework: User Guide. Technical report, Institute for Human & Machine Cognition (ihmc), Jan 2013.

[Jin14] Steve Jin. VMware VI (vSphere) Java API. `http://vijava.sourceforge.net/`, April 2014.

[OAS10] OASIS. eXtensible Access Control Markup Language (XACML) Version 3.0 - Committee Specification 01. `http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-cs-01-en.pdf`, August 2010.

[OAS13] OASIS. eXtensible Access Control Markup Language (XACML) Version 3.0 - OASIS Standard. `http://http://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.pdf`, January 2013.

[PHB06] A. Pretschner, M. Hilty, and D. Basin. Distributed usage control. Commun. ACM, 49(9):39–44, 2006.

[PS02] Jaehong Park and Ravi Sandhu. Towards Usage Control Models: Beyond Traditional Access Control. In Proceedings of the Seventh ACM Symposium on Access Control Models and Technologies, SACMAT '02, pages 57–64, New York, NY, USA, 2002. ACM.

[PS04] J. Park and R. Sandhu. The UCON ABC usage control model. ACM Trans. Inf. Syst. Secur., 7(1):128–174, 2004.

[RS14] Manuel Rudolph and Reinhard Schwarz. SECCRIT Online Survey on Cloud Security for Critical Infrastructure IT. Survey, Fraunhofer IESE, March 2014.

Sichere und benutzerfreundliche Schlüsselverteilung auf Basis von QR-Codes

Kevin Körner, Thomas Walter

Fachbereich Informatik, Arbeitsbereich Informationsdienste
Eberhard Karls Universität Tübingen
Wächterstraße 76
72074 Tübingen
kevin.koerner@uni-tuebingen.de
thomas.walter@uni-tuebingen.de

Abstract: Datenschutzalgorithmen im Webumfeld und darüber hinaus basieren auf der Nutzung geheimer Schlüssel. Diese sind einerseits durch Kombination heterogener Zeichen höchstmöglich zu sichern, andererseits müssen die Schlüssel zur Nutzung auf allen Endgeräten des Nutzers verfügbar sein. Praktische Ansätze wie die Verteilung von Schlüsseln über E-Mail oder der Einsatz gut merkbarer Wörterbuchpasswörter sind zwar benutzerfreundlich, jedoch bezüglich Vertraulichkeit und Sicherheit fragwürdig. Die Problematik der benutzerfreundlichen Schlüsselverteilung wird durch den wirtschaftlichen Erfolg von mobilen Endgeräten zusätzlich interessant. Zum einen steigt die Anzahl an Personen mit mehr als einem (mobilen) Endgerät, zum anderen bieten mobile Endgeräte neue Technologien für die Nutzerinteraktion. In jedem Fall ist eine Auseinandersetzung mit benutzerfreundlichen Methoden zum Schlüsselaustausch zwischen beliebigen (mobilen) Endgeräten sinnvoll. Das vorliegende Dokument diskutiert einen Ansatz zur Verteilung von Schlüsseln zwischen unterschiedlichen (mobilen) Endgeräten. Hierbei führen wir das Konzept der Schlüsselverteilung über zweidimensionale Barcodes am Beispiel des QR-Codes ein, diskutieren potentielle Schlüsselspeicher-Architekturen und erörtern praktische Anwendungsszenarien.

1 Einleitung

Aus der Privatisierung des WorldWideWeb Anfang der 1990er Jahre ist eine Vielzahl an Diensten zur Kommunikation und zum automatisierten Datenaustausch über unsichere Netzwerke entstanden; beispielsweise Chat, Online-Banking, E-Mail und Cloud-Computing. Um Abhörsicherheit, Nachrichtenintegrität und Authentifizierung der Kommunikationspartner auf unsicheren Kanälen zu gewährleisten, existieren diverse Sicherheitsprotokolle (siehe auch [Lam04]). Diese basieren dabei auf der Geheimhaltung von Schlüsseln: Entweder eines allen Kommunikationspartnern bekannten symmetrischen Schlüssels oder der privaten Schlüssel bei asymmetrischen Verfahren. Bei letzteren besteht zudem noch das Problem, dass die Identität der Kommunikationspartner vor dem initialen Nachrichtenaustausch sichergestellt werden muss.
Dementsprechend müssen die Dienstnutzer in jedem Fall über einen sicheren Kanal In-

formationen austauschen, bevor sie gesichert über einen unsicheren Kanal kommunizieren können. Die Autoren von [LV99] erörtern hierbei, weshalb es notwendig ist, die verwendeten Schlüssel immer aufwendiger und größer zu gestalten. Andernfalls steigt die Wahrscheinlichkeit eines praktikabel durchführbaren Brute-Force-Angriffs mit steigender Rechnerleistung. Lampson erörtert in [Lam04] zudem die Herausforderungen an sichere computerbasierte Kommunikation und arbeitet heraus, weshalb Computer in der, wie er sie bezeichnet, „echten Welt“ nicht als sicher angesehen werden dürfen.
Mit steigender Schlüsselkomplexität steigt jedoch auch der Bedarf an benutzerfreundlichen und sicheren Ansätzen zum Austausch der initial benötigten Informationen. Die Entwicklung solcher Ansätze hängt der Entwicklung von Datensicherheitsmethoden unseres Erachtens hinterher. Wir stimmen den Autoren von [WT98] dahingehend zu, dass es ein Zusammenspiel zwischen Benutzerfreundlichkeit bei Datenschutztechnologien sowie deren Akzeptanz seitens der Nutzer gibt: Je aufwändiger eine Technologie für deren Nutzer ist, desto geringer ist ihre Akzeptanz.
Ende des ersten Jahrzehnts des 21. Jahrhunderts etablierten sich zudem die Technologien für mobile Endgeräte im privaten Umfeld; insbesondere Smartphones und Tablets. Aktuelle Studien (z.B. [AB13]) sagen voraus, dass die Anzahl der verwendeten Smartphones sich in den kommenden Jahren deutlich steigern wird. Für die Nutzer stellt sich daraus resultierend die Frage, wie benötigte Schlüssel auf sicherem Weg zwischen Endgeräten verteilt werden können. Insbesondere lohnt eine Betrachtung der neuen Interaktionsmöglichkeiten mit den mobilen Endgeräten, die von konventionellen Techniken nicht unterstützt werden.

Im vorliegenden Dokument präsentieren wir ein Konzept zur benutzerfreundlichen Verteilung von Schlüsseln zwischen (mobilen) Endgeräten unter Verwendung von QR-Codes. In Kapitel 2 wird der von uns aufgefasste aktuelle Forschungsstand erörtert. Die grundlegenden Architekturkomponenten diskutieren wir in Kapitel 3, gefolgt von potentiellen Schlüsselspeicher-Architekturen in Kapitel 4. Kapitel 5 diskutiert die präsentierte Idee und beschäftigt sich mit praktischen Anwendungsszenarien. In Kapitel 6 folgt abschließend eine Zusammenfassung des Dokuments sowie ein Ausblick auf künftige Forschungen in diesem Bereich.

2 Themenbezogene Arbeiten

2.1 Diffie-Hellman-Schlüsselaustausch

Eines der bekanntesten Schlüsselaustauschprotokolle ist 1976 von Diffie und Hellman in [DH76] eingeführt worden. Dieses erlaubt es Kommunikationspartnern, symmetrische Schlüssel sicher über einen unsicheren Kanal auszutauschen. Unberechtigte Dritte können die für den Schlüsselaustausch verwendeten Nachrichten mithören, ohne aus diesen auf den ausgetauschten Schlüssel schließen zu können. Ein essentieller Nachteil des Verfahrens ist die Anfälligkeit gegen einen Man-in-the-middle-Angriff aufgrund der fehlenden Authentifizierung der Kommunikationspartner (vgl. [Lam04]). Ansätze wie digitale Signaturen oder Message Authentication Codes (MACs) können mit dem Diffie-Hellman-

Verfahren kombiniert werden, um die Sicherheitslücke zu beheben. Für die Kommunikationspartner bedeutet dies jedoch zusätzlichen Organisationsaufwand, wie den Einsatz vertrauenswürdiger Public-Key-Infrastrukturen (vgl. [M+99]) oder den Kommunikationspartnern bekannte Geheimnisse. Unser Konzept kann hierbei mit geringerem Aufwand zum vertrauenswürdigen ad hoc Schlüsselaustausch verwendet werden.

2.2 Off-the-Record-Messaging (OTR)

Off-the-Record-Messaging (OTR) ist ein Protokoll zum sicheren initialen Schlüsselaustausch zwischen Kommunikationspartnern und wurde 2004 in [BGB04] eingeführt. Das Protokoll verwendet dabei das Diffie-Hellman-Schlüsselaustauschverfahren zum Schlüsselaustausch und stellt die Identität der Kommunikationspartner über MACs sicher. Eine weitere in OTR umgesetzte Anforderung besteht darin, dass von aufgedeckten Schlüsseln nicht auf andere für Kommunikationen verwendete Schlüssel geschlossen werden kann; beispielsweise von ausgetauschten Langzeitschlüsseln auf kurzlebige Sitzungsschlüssel. Unser Konzept erlaubt es Kommunikationspartnern ad hoc Schlüssel und eindeutige Identifizierungsmerkmale für die Authentifizierung in OTR auszutauschen.

2.3 QR-Codes

Nach eigenen Angaben (siehe [DW14]) wurde die QR-Code-Technologie erstmalig 1994 vom japanischen Unternehmen DENSO WAVE als Erweiterung des eindimensionalen Barcodes um eine zweite Dimension veröffentlicht. Die zweite Dimension ermöglicht es auf derselben Fläche deutlich mehr Zeichen in einem Code zu kodieren als in einem eindimensionalen Barcode. Nach [DW14] hängt die Anzahl an in einem QR-Code kodierbaren Zeichen dabei von der ihn erzeugenden Anwendung ab. Die Anzahl reduziert sich dabei durch verschieden ausgeprägte Fehlerkorrektur-Stufen, die den Ausgleich fehlerhafter Übertragungen erlauben. Der Standard [ISO00] ermöglicht beispielsweise 2953 8-Bit-Zeichen bei niedrigster Fehlerkorrektur-Stufe in einem QR-Code zu kodieren. Zudem erlauben es Positionsinformationen innerhalb des QR-Codes aus beliebigen Aufnahmepositionen die korrekte QR-Code-Darstellung zu errechnen. Insbesondere mobilen Endgeräten erleichtert dies die Aufnahme und Weiterverarbeitung dieser Codes.
Anwendungen wie [AK08] oder [TH14] nutzen die genannten Vorteile für eine möglichst benutzerfreundliche Überführung von Informationen auf mobile Endgeräte. Diverse Bibliotheken sind auf diese Weise aus Projekten entstanden, welche es Anwendungsentwicklern erleichtern die QR-Code-Funktionalitäten in ihre Applikationen zu integrieren; beispielsweise [DT14]. Nachteile von QR-Codes werden in [K+10] ausführlich erörtert; zum Beispiel die erschwerte visuelle Verifizierbarkeit von QR-Code-Inhalten. Dies vereinfacht bekannte Angriffe wie SQL-Injection sowie Social Engineering und Phishing.
Trotz des intuitiven Ansatzes Schlüssel über QR-Codes zwischen (mobilen) Endgeräten zu verteilen, haben wir bisher bei unserer Nachforschung keine wissenschaftlichen Projekte

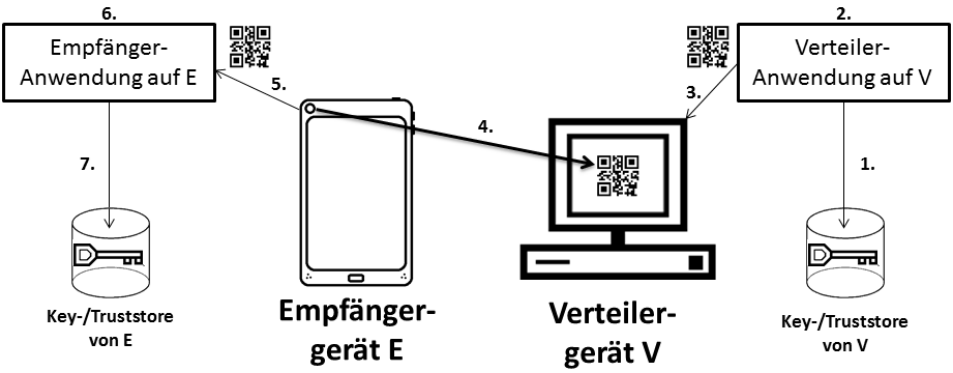


Abbildung 1: Konzept zum Schlüsselaustausch über QR-Codes

oder Veröffentlichungen zu der Thematik gefunden. Die proprietäre Anwendung *Threema* ([TH14]) bietet jedoch eine Möglichkeit öffentliche Schlüssel über QR-Codes zu verifizieren sowie private Schlüssel über solche zu verteilen.

3 Systemarchitektur

Kommunikationspartner müssen zusätzlichen Organisationsaufwand investieren, sofern sie etablierte Methoden zum sicheren Schlüsselaustausch verwenden möchten. Insbesondere gehen solche Protokolle davon aus, dass die Partner vor dem initialen Schlüsselaustausch ein gemeinsames Geheimnis ausgetauscht haben. Für den Austausch ist es daher unvermeidlich, dass alle Kommunikationspartner das identische Geheimnis zur Verfügung haben. Dies ist jedoch kritisch zu sehen, insbesondere wenn zwischen der Geheimnisvereinbarung und dem Schlüsselaustausch ein längerer Zeitraum liegt. Zudem müssen die Kommunikationspartner sicherstellen, dass die gewählten Geheimnisse nicht von Dritten mitgehört, erraten oder gefälscht werden können. Die im Folgenden präsentierte Architektur reduziert diese Probleme, indem der Schlüsselaustausch ohne zeitliche Verzögerung erfolgen kann und kein zusätzlich vereinbartes Geheimnis zwischen den Kommunikationspartnern erforderlich ist.

3.1 Systemkomponenten

Das von uns vorgeschlagene Konzept zur Schlüsselüberführung ist in Abbildung 1 dargestellt. Nachfolgend werden die Komponenten sowie deren Funktionen und Voraussetzungen an die (mobilen) Endgeräte erörtert. Dabei unterscheiden wir die verwendeten (mobilen) Endgeräte in ein Schlüssel verteilendes Gerät V und ein Schlüssel empfangendes Gerät E.

Keystore: Der Keystore ist ein gesicherter Speicherort auf dem Endgerät, der für die Aufbewahrung und Organisation von geheimen beziehungsweise privaten Schlüsseln verwendet wird; beispielsweise in pkcs12-Dateien oder in einer Datenbank. Der Keystore muss eine abgesicherte Schnittstelle besitzen, über die autorisierte dritte Anwendungen auf die hinterlegten privaten Schlüssel zugreifen können. Weiterhin ist sicherzustellen, dass in den Keystore ausschließlich vom Nutzer verifizierte Schlüssel gespeichert werden.

Truststore: Der Truststore ist ein gesicherter Speicherort auf dem Endgerät, der für die Aufbewahrung und Organisation von vertrauenswürdigen öffentlichen Schlüsseln verwendet wird; beispielsweise in crt-Dateien oder in einer Datenbank. Der Truststore muss eine abgesicherte Schnittstelle besitzen, über die autorisierte dritte Anwendungen auf die hinterlegten Schlüssel zugreifen können. Weiterhin ist sicherzustellen, dass in den Truststore ausschließlich vom Nutzer verifizierte öffentliche Schlüssel gespeichert werden.

Verteiler-Anwendung: Die Verteiler-Anwendung muss auf V verfügbar sein. Ihre Aufgabe ist es, aus der Textdarstellung eines Schlüssels den zugehörigen QR-Code zu erzeugen und diesen grafisch sichtbar darzustellen. Hierfür muss V eine grafische Ausgabe besitzen; beispielsweise einen Bildschirm. Zusätzlich kann die Verteiler-Komponente eine Passworteingabe bereitstellen, die es ermöglicht den Schlüssel vor der Überführung in die QR-Code-Darstellung zu verschlüsseln. Ohne Einschränkung gehen wir nachfolgend davon aus, dass zu verteilende Schlüssel bereits auf V gespeichert sind; zum Beispiel in V's Truststore oder Keystore.

Empfänger-Anwendung: Die Empfänger-Anwendung muss auf E verfügbar sein. Ihre Aufgabe ist es, QR-Code-Darstellungen aufzunehmen und aus diesen die kodierten Schlüssel zu berechnen. Hierfür muss E eine Bildaufnahme-Funktion besitzen; beispielsweise eine Kamera. Zusätzlich kann die Empfänger-Anwendung eine Passworteingabe bereitstellen, die es ermöglicht, einen übertragenen verschlüsselten Schlüssel zu entschlüsseln. In Abhängigkeit des übertragenen Schlüssels muss der Nutzer die Möglichkeit haben, den empfangenen Schlüssel entweder in E's Truststore oder in E's Keystore zu überführen.

Eine Aufteilung der Funktionen auf mehrere Anwendungen, die auf den (mobilen) Endgeräten installiert sind, ist empfehlenswert, jedoch nicht zwingend notwendig. Ebenso ist eine Anwendung realisierbar, in welche alle beschriebenen Softwarekomponenten integriert sind. In beiden Fällen kann ein (mobiles) Endgerät sowohl als verteilendes Gerät V als auch als empfangendes Gerät E Nutzung finden.

3.2 Schlüsselverteilung

Das Konzept der Schlüsselübertragung über QR-Codes ist in Abbildung 1 dargestellt. Es ist sowohl für öffentliche, private und symmetrische Schlüssel anwendbar. Möchte ein Nutzer einen auf Gerät V hinterlegten öffentlichen oder privaten Schlüssel an Gerät E

weitergeben, so greift er mit der Verteiler-Anwendung auf Vs Trust- beziehungsweise Keystore zu und wählt den zu verteilenden Schlüssel (1). Zusätzlich kann die Verteiler-Anwendung über eine Passworteingabe verfügen, um den ausgewählten Schlüssel für die Übertragung zu verschlüsseln. Aus dem (verschlüsselten) Schlüssel erstellt die Verteiler-Anwendung die QR-Code-Darstellung (2) und stellt diese auf Vs grafischer Ausgabe dar (3).
E fotografiert mit seiner Bildaufnahme-Funktion den QR-Code von Vs grafischer Ausgabe ab (4) und leitet diesen an die Empfänger-Anwendung weiter (5). Diese ermittelt den im QR-Code kodierten (verschlüsselten) Schlüssel (6). Sofern der Schlüssel verschlüsselt ist, muss Es Nutzer zusätzlich das zuvor verwendete Passwort für die Entschlüsselung des übertragenen Schlüssels eingeben. Dies kann beispielsweise direkt in der Empfänger-Anwendung geschehen. Anschließend speichert diese den übertragenen Schlüssel über die gesicherten Schnittstellen in Es Key- beziehungsweise Truststore (7). Nun kann der Schlüssel auch von Es Anwendungen verwendet werden.

4 Sicherung der gespeicherten Schlüssel

Die vorgestellte Architektur muss sicherstellen, dass ausschließlich berechtigte (dritte) Anwendungen auf die hinterlegten Schlüssel zugreifen können. Folgend werden die hierfür unseres Erachtens relevanten Ansätze präsentiert und diskutiert. Der Begriff *Schlüsselspeicher* steht in diesem Abschnitt stellvertretend sowohl für Trust- als auch Keystore.

4.1 Eigenverantwortliche Schlüsselverwaltung

Jede Anwendung kann neben ihren eigentlichen Funktionen einen eigenen Schlüsselspeicher betreiben. Dementsprechend ist jede Anwendung für die Verwaltung der von ihr benötigten Schlüssel eigenverantwortlich. Dies reduziert die Anzahl der im Schlüsselspeicher hinterlegten Schlüssel im Vergleich zu einem zentralisierten. Hieraus ergibt sich der Vorteil, dass im Falle eines korrumpierten Schlüsselspeichers die Schlüssel anderer Anwendungen nicht betroffen sind. Unserer Meinung nach ist es aus organisatorischer sowie aus datenschutzrechtlicher Sicht empfehlenswert, Schlüssel nur an den Stellen zu hinterlegen, an denen sie benötigt werden. Schlüsselsuche und fehlerhafte Schlüsselwahl kann somit zumindest verringert werden, was die Benutzerfreundlichkeit erhöht.
Aus der dezentralen Schlüsselhaltung ergeben sich jedoch auch diverse Nachteile. Anwendungsentwickler sind hierbei für die Implementierung eigener Verteiler- und Empfänger-Funktionalität verantwortlich oder müssen hinreichend geschützte Schnittstellen mit diesen realisieren. Zudem muss die Anwendung sich selbst um die Schlüsselorganisation kümmern, wie das Einfügen, Abrufen und Entfernen sowie der Schutz von Schlüsseln. Dieser Zusatzaufwand kann für die Entwicklung neuer Dienste und Anwendungen hinderlich sein. Zwar können Anwendungsentwickler hierfür Bibliotheken verwenden, deren Änderungen und Sicherheitsanpassungen jedoch von den Entwicklern kontinuierlich

in die Anwendung integriert werden müssen. Aus Nutzersicht bedeutet dies zusätzlichen Aufwand: Man muss alle Anwendungen auf allen (mobilen) Endgeräten aktualisieren, die diese Bibliotheken verwenden.
Ein aus Nutzersicht weitaus größerer Nachteil am dezentralen Ansatz ist die Verteilung von Schlüsseln, die in mehreren Anwendungen auf demselben (mobilen) Endgerät benötigt werden. In diesem Fall muss der Schlüssel durch mehrere Empfänger-Funktionen auf dasselbe (mobile) Endgerät überführt werden. Mit steigender Anzahl an Anwendungen steigt dabei für den Nutzer der organisatorische Aufwand. Besonders problematisch ist diese Herangehensweise, wenn ein Schlüssel ausgetauscht werden muss; beispielsweise wenn ein privater Schlüssel gestohlen wird. In diesen Fällen ist der Schlüssel zwangsläufig in allen ihn verwendenden Anwendungen auszutauschen. Folglich steigt mit wachsender Anzahl an Anwendungen zum einen der Aufwand für den Nutzer und zum anderen das Risiko, dass der Austausch an einer Stelle vergessen wird.
Das bereits erwähnte Threema ([TH14]) beinhaltet einen eigenen Speicher für die öffentlichen Schlüssel vertrauenswürdiger Personen und speichert den eigenen privaten Schlüssel.

4.2 Zentrale Schlüsselverwaltung

Der entgegengesetzte Ansatz zur dezentralen Schlüsselverwaltung ist, unabhängig von den Schlüssel nutzenden Anwendungen, das Betreiben einzelner zentraler Schlüsselspeicher. Die Extremform hierbei ist ein einzelner zentraler Schlüsselspeicher, der von allen Anwendungen genutzt werden kann. Ein Hauptvorteil aus Nutzersicht liegt im reduzierten Aufwand bei der Überführung von Schlüsseln, die in mehreren Anwendungen verwendet werden. Ebenso ist die Anpassung im Falle eines Schlüsseltausches einfacher: Der Nutzer muss den Schlüssel ausschließlich im zentralen Schlüsselspeicher aktualisieren. Die Folge ist eine Reduzierung des Risikos, dass Nutzer den Schlüsselaustausch in Anwendungen vergessen und in diesen mit dem veralteten oder unsicheren Schlüssel unbemerkt weiterarbeiten.
Aus technischer Sicht ist es vorteilhaft, dass sowohl Verteiler- als auch Empfänger-Funktionen über auf dem (mobilen) Endgerät zentrale Anwendungen bereitgestellt werden können. Diese können dadurch direkt als Teil des Schlüsselspeichers entwickelt und somit die Interaktion zwischen Verteiler-, Empfänger- und Schlüsselspeicher-Funktion gesichert werden. In jedem Fall ist die Aktualisierung der zentralen Verteiler-, Empfänger- oder Schlüsselspeicher-Funktion für den Nutzer weniger aufwendig, wenn sich verwendete Bibliotheken ändern, da die mit den Schlüsseln arbeitenden Anwendungen von diesen unabhängig sind.
Falls unberechtigte Dritte Zugriff auf einen zentralen Schlüsselspeicher erhalten, so sind alle dort gespeicherten Schlüssel als gestohlen anzusehen. Dies ist im Vergleich zur Aufteilung der Schlüssel auf mehrere unabhängige Schlüsselspeicher nachteilig. Hierbei handelt es sich um einen problematischen Gesichtspunkt, da alle Daten der mit dem Schlüsselspeicher interagierenden Anwendungen als gestohlen, manipuliert oder unsicher betrachtet werden müssen. Dadurch ist ein erheblicher Aufwand für den Nutzer unvermeidbar, um die Integrität der Daten zu prüfen, neue Schlüssel für die Anwendungen zu erstellen sowie

zu verteilen und die Daten unverzüglich neu zu verschlüsseln. Insbesondere wenn Dritte Zugriff auf Schlüssel aus kollaborativen Projekten bekommen haben, sind nicht nur die Daten des angegriffenen (mobilen) Endgeräts verloren, sondern auch die der Projektpartner. Dies muss an alle Projektpartner kommuniziert werden und es ist eine in der Gruppe durchführbare Lösung für die zuvor genannten Aufgaben zu finden.
Im zentralisierten Ansatz muss berücksichtigt werden, dass dritte Anwendungen mit dem Schlüsselspeicher interagieren können müssen. Erfahrungsgemäß ist es sinnvoll, standardisierte Schnittstellen für die Anwendungskommunikation zu verwenden. Der vom Schlüsselspeicher verwendete Standard ist dabei möglichst nicht zu verändern, da Änderungen an der Schnittstelle Änderungen an den zugreifenden Anwendungen nach sich ziehen. Dies verursacht wiederum Aktualisierungsaufwand für den Nutzer. Da die Schlüssel als kritische Elemente angesehen werden, muss der Schlüsselspeicher seine Schnittstellen vor unautorisiertem Zugriff schützen; zum Beispiel über eine Passwortabfrage.

4.2.1 Betriebssystemintegration

Eine mögliche Implementierungsart der zentralen Schlüsselverwaltung ist die Integration in das Betriebssystem des (mobilen) Endgeräts. Dies ermöglicht die Erstellung einer optimierten Gesamtanwendung aus Verteiler, Empfänger und Schlüsselspeicher und sie kann direkt in das Betriebssystem integriert werden. Nutzer müssen somit keiner dritten Anwendung für die Aufbewahrung ihrer Schlüssel vertrauen. Das Betriebssystem des (mobilen) Endgeräts ist zwangsläufig als vertrauenswürdig einzustufen, andernfalls dürften die Daten nie unverschlüsselt auf diesem vorliegen. Zudem ermöglicht die Integration, den Zugriff auf die Schlüssel vom Betriebssystem schützen zu lassen. Beispielsweise ist es denkbar, den Zugriff auf den Schlüsselspeicher auf Anwendungen zu beschränken, die vom Betriebssystemhersteller zertifiziert sind.
Nachteilig an dieser Art der Integration ist die Notwendigkeit, dass alle Anwendungen die mit den Schlüsseln arbeiten, mit den Betriebssystemschnittstellen interagieren können müssen. Daraus resultiert, dass die Anwendungen nicht betriebssystemübergreifend entwickelt werden können.

4.2.2 Externe Schlüsselspeicher

Zentrale Schlüsselspeicher können auch von Drittanbietern entwickelt und auf das Betriebssystem des (mobilen) Endgeräts installiert werden. Dies erlaubt es, mehrere zentrale Schlüsselspeicher auf einem Endgerät bereitzustellen, die für das zu verwaltende Schlüsselmaterial optimiert sind. Die Aufteilung der Schlüssel auf mehrere zentrale Schlüsselspeicher reduziert zumindest die Problematik eines erfolgreich angegriffenen Speicherorts.
Schlüsselspeicher können von externen Entwicklern auch in betriebssystemübergreifenden Technologien entwickelt werden. Dies erlaubt es Nutzern ihnen bekannte Anwendungen auf mehreren (mobilen) Endgeräten zu verwenden, auch wenn diese unterschiedliche Be-

triebssysteme haben.
Der Ansatz ist hinsichtlich des Zugriffs auf den Schlüsselspeicher nachteilig, da jeder Schlüsselspeicherentwickler seinen eigenen Schnittstellenstandard definieren kann. Daraus entsteht eine starke Kopplung zwischen den Schlüssel nutzenden Anwendungen und den Schlüsselspeichern. Weiterhin muss der Nutzer dem Entwickler des Schlüsselspeichers dahingehend vertrauen, dass er die aufbewahrten Schlüssel nicht weitergibt oder über „Hintertüren“ ausliest. Auch die eigentlich vorteilhafte Verwendung von betriebssystemübergreifenden Technologien ist kritisch zu betrachten, da nicht sichergestellt ist, ob ein Endgerät benötigte Grundlagen bereitstellt.

4.3 Aktualisierende Schlüsselverwaltung

Die Trennung der Verteiler- und Empfänger-Funktionalität vom Schlüsselspeicher ermöglicht eine aktualisierende Schlüsselverwaltung. Wie bei der dezentralen Schlüsselverwaltung kann jede Anwendung ihren eigenen Schlüsselspeicher betreiben, der über Schnittstellen der Anwendung zugreifbar ist. Dies erlaubt es, einer zentralen Empfänger-Anwendung in einem Schritt mehrere dezentrale Schlüsselspeicher zu aktualisieren; beispielsweise die Überführung eines benötigten Schlüssels in mehrere Anwendungen. Ebenso kann eine zentrale Verteiler-Anwendung die zu verteilenden Schlüssel aus mehreren dezentralen Schlüsselspeichern beziehen. Durch die zentrale Implementierung der Empfänger- und Verteiler-Anwendung auf dem (mobilen) Endgerät müssen neue Versionen ausschließlich an einer Stelle eingepflegt werden. Gleichzeitig können die Anwendungshersteller eigene Schlüsselspeicher implementieren und für ihre Anwendungsfälle optimieren.
Bei aktualisierender Schlüsselverwaltung sind standardisierte Schnittstellen zwischen Verteiler- beziehungsweise Empfänger-Funktionalität und den Schlüsselspeicher-Anwendungen erforderlich. Zudem müssen die Schnittstellen vor unautorisiertem Zugriff geschützt werden, was für den Nutzer Aufwand bedeutet, beispielsweise indem er für jede Schlüsselspeicher-Anwendung ein Zugriffspasswort vergibt.

5 Diskussion des Konzepts

5.1 Aufwand

Aufgrund der direkten Interaktion zwischen den (mobilen) Endgeräten benötigen die austauschenden Partner kein zusätzliches Geheimnis und keinen unsicheren Kanal wie E-Mail-Verkehr für den Schlüsselaustausch. Für den vertraulichen Austausch von öffentlichen Schlüsseln werden zudem keine digitalen Zertifikate und Public-Key-Infrastrukturen benötigt. Letztere können jedoch zusätzlich für die Gültigkeitsprüfung der öffentlichen Schlüssel genutzt werden.
Für den Verteilenden beschränkt sich der Aufwand darauf, auf seinem (mobilen) Endgerät

die Verteiler-Anwendung zu starten und den Schlüssel auszuwählen. Der Empfänger muss ausschließlich mit der Empfänger-Anwendung auf seinem (mobilen) Gerät das Display des Verteilenden-Geräts fotografieren. Hierbei sind die Positionsinformationen im QR-Code hilfreich, da die Position des Empfänger-Geräts somit weitestgehend vernachlässigbar ist. Zur eindeutigen Zuordnung zwischen Schlüssel und Verwendungszweck beziehungsweise Schlüssel und Person ist es empfehlenswert, dass der Empfänger Metadaten über den Schlüssel in sein Gerät einträgt.
Die direkte Interaktion ist auch der Hauptnachteil des Verfahrens. Um Schlüssel austauschen zu können, müssen die Kommunikationspartner eine Möglichkeit haben die Display- und Fotofunktionen ihrer (mobilen) Endgeräte einzusetzen; beispielsweise indem einander persönlich treffen. Unserer Meinung nach ist dies für das Verfahren jedoch nicht nachteilig, da die Möglichkeit im persönlichen Gespräch ad hoc Schlüssel austauschen zu können deutlich vorteilhafter ist, als zeitverzögert aufwendigere/weniger sichere Verfahren zu nutzen. Es ist zudem aus technischer Sicht zu hinterfragen, ob QR-Codes wie in [ISO00] spezifiziert über genügend Kapazität für das zu übertragende Schlüsselmaterial verfügen. Insbesondere asymmetrische Verschlüsselungsverfahren benötigen möglichst große Schlüssel, um ausreichend viele und hinreichend sichere Schlüssel bereitzustellen (siehe hierzu auch [LV99]). Alternative Darstellungsformen, die über ausreichend Kapazität verfügen oder eine Aufteilung des Schlüssels auf mehrere QR-Codes sind hierbei von Nutzen. Alternativen zum QR-Code sind hierfür beispielsweise verknüpfbare DataMatrix-Codes wie in [ISO06] spezifiziert.

5.2 Sicherheit

Durch die direkte Interaktion zwischen den (mobilen) Endgeräten ist der Schlüsselaustausch als sicher anzusehen. Fehler während der Übertragung können von der Empfänger-Anwendung anhand der QR-Code-Prüfsummen entdeckt werden. Damit erhält der Empfänger direkt Rückmeldung über den Erfolg und der Austausch kann gegebenenfalls wiederholt werden. Für geheime und private Schlüssel besteht die Möglichkeit diese vor ihrer Darstellung als QR-Code zusätzlich über Passwörter zu schützen. Kann die QR-Code-Darstellung nicht hinreichend geschützt werden, so können unberechtigte Dritte den geheimen Schlüssel nicht aus dem QR-Code ermitteln.
Die Implementierungen von Trust- und Keystores definiert die Sicherheit der gespeicherten Schlüssel auf den (mobilen) Endgeräten. Die Aufbewahrung von Schlüsseln ist jedoch eine Grundsatzfrage für Datenschutzsysteme. In Kapitel 4 haben wir diesbezüglich Vor- und Nachteile von Implementierungen diskutiert.

5.3 Anwendungsgebiete

Unser Ansatz erlaubt es, einer Person ohne zusätzlichen Aufwand ihre geheimen und privaten Schlüssel zwischen ihren (mobilen) Endgeräten zu verteilen. Diese Verteilung erfolgt dabei gesichert und weitestgehend automatisiert. Die Person muss nicht zeichenwei-

se Passwörter von einem Gerät in ein anderes übertragen, sondern es bedarf ausschließlich einer Fotografie des QR-Codes. Zudem ist dieser Übertragungsweg sicherer als der Versand von Passwörtern beispielsweise in E-Mails.
Ebenfalls vereinfacht der Einsatz den Austausch von Schlüsseln zwischen unterschiedlichen Kommunikationspartnern. Über zentral bereitgestellte QR-Code-Darstellungen können Gruppen einfach ihre Schlüssel untereinander verteilen. Beispielsweise kann ein Vortragender seinen öffentlichen Schlüssel auf sicherem Weg an seine Zuhörer verteilen, indem er sie in seine Präsentation einfügt. Dies erlaubt die vertrauenswürdige Übertragung von öffentlichen Schlüsseln ohne Public-Key-Infrastruktur.
Ähnlich zum Web-of-Trust-Modell kann ein (öffentlicher) Schlüssel über die transitive Vertrauensrelation ausgetauscht werden. Vertraut Person B einer Person A, so überführt sie den Schlüssel von A's Endgerät auf das von B. Benötigt Person C A's Schlüssel und Vertraut C Person B, kann der Schlüssel ebenfalls einfach von B's Endgerät auf C's übertragen werden. Diese Herangehensweise ist insbesondere dann von Interesse, wenn A und C sich nicht persönlich treffen können.
Ein weiteres Beispiel ist die Verteilung von Schlüsseln innerhalb einer Gruppe über Printmedien. Im Vorfeld eines Gruppentreffens kann ein administratives Gruppenmitglied die von der Gruppe zu verwendenden Schlüssel erstellen und in ihre QR-Code-Darstellung überführen. Diese können mit Zusatzinformationen ausgedruckt und beim Gruppentreffen verteilt werden. Während des Treffens haben alle Gruppenmitglieder die Möglichkeit, die geheimen Schlüssel auf ihre Endgeräte zu überführen und nach Abschluss des persönlichen Treffens werden die Ausdrucke vernichtet. Die Erstellung des Printmediums erfordert anfänglich zwar Zusatzaufwand, die sichere Überführung auf die (mobilen) Endgeräte der Teilnehmer ist jedoch benutzerfreundlich und sicher.

6 Zusammenfassung und Ausblick

Integrität, Vertraulichkeit und Identifikation werden im Webumfeld auf Basis von geheimen Schlüsseln sichergestellt. Dabei müssen gewählte Schlüssel einerseits möglichst sicher gewählt sein, andererseits werden sie auf allen sie verwendenden (mobilen) Endgeräten des Nutzers benötigt. Mit steigender Schlüsselgröße und -komplexität sowie wachsender Anzahl an Endgeräten, die mit den Schlüsseln arbeiten, steigt der Aufwand für die Nutzer ihre Schlüssel zu verteilen. Das im vorliegenden Dokument vorgestellte Konzept erlaubt es, Schlüssel zwischen (mobilen) Endgeräten unter Verwendung von QR-Codes auszutauschen. Wir haben dabei in Kapitel 3 die benötigten funktionellen Einheiten und in Kapitel 4 potentielle Schlüsselspeicher-Architekturen diskutiert. Abschließend haben wir das vorgestellte Konzept hinsichtlich Aufwand für die Anwender und Sicherheit für das Schlüsselmaterial betrachtet und ausgewählte Anwendungsfälle vorgestellt. Die Zielstellung des vorgestellten Konzepts ist hierbei in erster Linie der einfache, ad hoc mögliche und vertrauliche Schlüsselaustausch bei persönlichen Treffen von Personen und die Möglichkeit für einen Nutzer seine geheimen Schlüssel möglichst unkompliziert zwischen seinen (mobilen) Endgeräten zu verteilen.
Basierend auf dem vorgestellten Konzept entwickelt unsere Arbeitsgruppe derzeit proto-

typisch eine Anwendung. Sobald die Implementierung abgeschlossen ist, werden wir eine Evaluation über eine große Bandbreite an Arbeitsgruppen durchführen, um damit die von uns angenommene Nutzerfreundlichkeit zu überprüfen. Hierfür sind noch Anwendungen zu erstellen, die auf Trust- und Keystore unseres Protoypen zugreifen; beispielsweise mobile E-Mail-Clients oder Anwendungen, die mit Daten auf Cloud-Strukturen interagieren.

Literatur

[AB13] Ericsson AB. Ericsson Mobil Report: On The Pulse Of The Networked Society, November 2013.

[AK08] Hend S. Al-Khalifa. Utilizing QR Code and Mobile Phones for Blinds and Visually Impaired People. In *Computers Helping People with Special Needs*, Seiten 1065–1069. Springer Berlin Heidelberg, 2008.

[BGB04] Nikita Borisov, Ian Goldberg und Eric Brewer. Off-the-record Communication, or, Why Not to Use PGP. In *Proceedings of the 2004 ACM Workshop on Privacy in the Electronic Society*, Seiten 77–84. ACM, 2004.

[DH76] W. Diffie und M. Hellman. New Directions in Cryptography. *IEEE Transactions on Information Theory*, 22(6):644–654, 1976.

[DT14] The ZXing Developer-Team. Zxing, open source library to read 1D/2D barcodes. Webseite: "https://github.com/zxing/zxing/"[Zuletzt Aufgerufen: Mai 2014].

[ISO00] International Organization for Standardization. Information Technology - Automatic Identification and Data Capture Techniques - Bar code symbology - QR Code. ISO/IEC 18004:2000, 2000.

[ISO06] International Organization for Standardization. Information technology - Automatic identification and data capture techniques - Data Matrix bar code symbology specification. ISO/IEC 16022:2006(E), 2006.

[TH14] Threema GmbH. Threema - Seriously secure mobile messaging. Webseite: "https://threema.ch/de/"[Zuletzt Aufgerufen: Mai 2014].

[DW14] DENSO WAVE INCORPORATED. History of QR Code. Webseite: "http://www.qrcode.com/en/history/"[Zuletzt Aufgerufen: Mai 2014].

[K+10] Peter Kieseberg et al. QR Code Security. In *Proceedings of the 8th International Conference on Advances in Mobile Computing and Multimedia*, MoMM '10, Seiten 430–435. ACM, 2010.

[Lam04] Butler W. Lampson. Computer Security in the Real World. *Computer*, 37(6):37–46, 2004.

[LV99] Arjen K. Lenstra und Eric R. Verheul. Selecting Cryptographic Key Sizes. *Journal of Cryptology*, 14:255–293, 1999.

[M+99] M. Myers et al. RFC2560 X.509 Internet Public Key Infrastructure, Online Certificate Status Protocol, 1999.

[WT98] Alma Whitten und J. D. Tygar. Usability of Security: A Case Study. In *CMU-CS-98-155*, December 1998.

Reusing cloud-based services with TOSCA*

Antonio Brogi and Jacopo Soldani
Department of Computer Science, University of Pisa, Italy
{*brogi,soldani*}*@di.unipi.it*

Abstract: The OASIS TOSCA specification [OAS13b] aims at enhancing the portability and interoperability of cloud-based applications by providing a language to describe and manage them across heterogeneous clouds. A cloud-based application is modelled by a service template, an orchestration of typed nodes which can be in turn instantiated by matching [BS13] other service templates. In this paper we present a methodology to non-intrusively adapt a service template S into a new service template $newS$ which exactly matches a node type N, and hence to easily reuse any actual service modelled by S to deploy cloud-based applications that rely on N.

1 Introduction

Current cloud technologies suffer from a lack of standardization, with different providers offering similar resources in a different manner. Hence, migrating (parts of) an application from one cloud to another is a costly and error-prone process. As a result, cloud users tend to end up locked into the cloud platform they are using since it is practically unfeasible for them to migrate (parts of) their application across different clouds platforms [PMPC13].

In this scenario, OASIS recently released the *Topology and Orchestration Specification for Cloud Applications* (TOSCA), which aims at enhancing the portability of cloud-based applications by defining a language to describe and manage complex applications across heterogeneous clouds. More precisely, TOSCA [OAS13b] defines a XML-based language that permits to specify —in a vendor-agnostic way— topology and behaviour of complex composite applications as service templates that orchestrate typed nodes.

As stated in the TOSCA primer [OAS13a], a node type can be instantiated by substituting it with a "matching" service template. In our previous work [BS13], we provided a formal definition of when a service template exactly or plug-in matches a node type. Informally speaking, a service template S *exactly* matches a node type N if S has the same capabilities, requirements, properties, policies and interfaces of N. S instead *plug-in* matches N if S has "less" requirements and "more" capabilities, properties, and interfaces of N. It is worth noting that, while [BS13] formally defines the conditions for type-checking whether a node type N may be instantiated by a service template S, no methodology is provided in [BS13] to actually perform such an instantiation when S does not exactly match N.

In this paper we present a methodology which permits to non-intrusively adapt a service template S into a new service template $newS$ which exactly matches a node type N, thus allowing (the adaptation of) S to be (re)used in place of N [OAS13a]. We first consider the case of a service template S that plug-in matches a node type N. As we will see, $newS$ is obtained by extending —in a non-intrusive way— S by (artificially) adding requirements

*Work partly supported by project EU-FP7-ICT-610531 SeaClouds.

of N that are not requirements for S, and by hiding capabilities, properties, policies and interfaces of S that N does not have.

We then show that also a service template S which does *not* plug-in match a node type N may be non-intrusively adapted into a new service template $newS$ which exactly matches N. More precisely, we will show how the above mentioned adaptation can be properly extended so as to be successfully performed (i) when S exposes all the capabilities and properties as N, but in a *syntactically* different way, and/or (ii) when N declares all the requirements which are exposed by S, but in a *syntactically* different way, and/or (iii) when N features one or more interface operations which are not matched by any operation featured by S, while they can be matched by some compositions of S's operations.

It is important to observe that the adaptation of a TOSCA service template S to match a node type N does suffice to reuse any actual service modelled by S to deploy cloud-based applications that rely on N. This is thanks to the powerful way in which TOSCA supports the deployment of cloud-based applications. TOSCA permits to pack in a CSAR (*Cloud Service ARchive*) file an application specification together with the actual executable files to be deployed on a cloud platform. When a CSAR file is given in input to a TOSCA container, the latter takes care of deploying and executing the application specification contained in the CSAR file [OAS13a]. Therefore, in order to adapt an actual service modelled by a service template S to deploy an application that relies on a node type N, it suffices to adapt S into a new service template $newS$ that matches N — without having to generate an implementation of the adaptation specified by S.

It is also worth observing that, since the aforementioned adaptation methodology is non-intrusive, it can be performed both by the cloud provider offering the service S and by the application developer who needs the node N and discovers the availability of S.

Note that the adaptation works also in the case in which the CSAR of S should not be available, for instance when S is a proprietary service offered by a third party. In such cases it suffices to develop a simple proxy of the remote service modelled by S, and to pack it in a new CSAR file together with the application specification containing $newS$ (and the executable files associated with such a specification).

Finally, it is worth highlighting that, thanks to the features of TOSCA, the simple adaptation methodology that will be described in this paper considerably reduces the work needed to reuse cloud-based services if compared to the alternative of explicitly devising adapters as in traditional software adaptation approaches (e.g., [BCP06, GMMC13, KMNB$^+$09]).

The rest of paper is organised as follows. TOSCA and the matching between service templates and node types are introduced in Section 2. The methodology for adapting service templates is presented in Section 3. Finally, related work and some concluding remarks are discussed in Sections 4 and 5, respectively.

2 Background

In this section we briefly recall the main notions of the *Topology and Orchestration Specification for Cloud Applications* (TOSCA)[1] and the notions of matching between TOSCA services [BS13].

[1] Interested readers may refer to [OAS13b] and [OAS13a] for a comprehensive presentation of TOSCA, or to [BSW14] for a quick introduction to TOSCA.

2.1 TOSCA

The main aim of TOSCA is to enhance the portability of multi-cloud applications by enabling an interoperable description of application and infrastructure cloud services, of the relationships between service parts, and of the operational behaviour of services, independently of the supplier creating the service and of any particular cloud provider or hosting technology.

Syntactically speaking, TOSCA is an XML-based language for describing service templates. All definitions are contained in the XML *Definitions* element (the root of a TOSCA XML document). The *ServiceTemplate* element (Figure 1) defines all the topology and

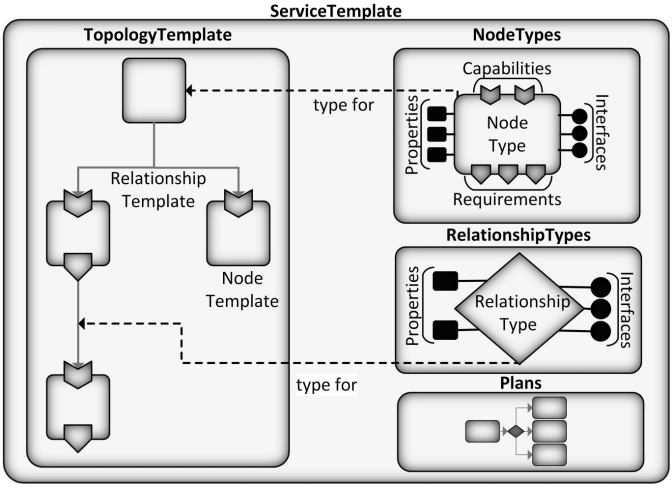


Figure 1: TOSCA *ServiceTemplate*.

management aspects of a service by means of *TopologyTemplate* and *Plans* elements. A *TopologyTemplate* specifies the topological structure of a service as a directed graph, whose nodes are *NodeTemplate*s and whose arcs are *RelationshipTemplate*s. *Plans* contains *Plan* elements that specify how to manage the associated service template during its whole lifetime. *BoundaryDefinitions* can be used to define which features are externally exposed by a *ServiceTemplate*[2].

*NodeTemplate*s are typed by means of *NodeType*s, which define the structure of the features whose values are specified in a *NodeTemplate*. Such features include properties, interfaces, requirements, capabilities, and policies. While properties and interfaces can be defined internally, requirements and capabilities must be typed by referring to external *RequirementType*s and *CapabilityType*s. Finally, *NodeTemplate*s and *ServiceTemplate*s can declare QoS information by exposing *Policy* elements, which must be typed by referring to *PolicyType*s. A *PolicyType* defines the structure of policy (and the *NodeType*s it is applicable to), while a *Policy* assigns actual policy values.

[2]For instance, a property p_{ext} can be defined on the boundaries of a service template as a reference to an internal property p_{int} (which can have a different name but must have a compatible type).

2.2 Matching service templates with node types

As stated in the TOSCA primer ([OAS13a], page 35): *"node types can be made concrete by substituting them by a service template"*. While the matching between *ServiceTemplate*s and *NodeType*s is mentioned with reference to an example, no definition of *matching* is given either in [OAS13b] or in [OAS13a]. In our previous work [BS13] we formally defined when a service template *exactly* or *plug-in* matches a node type.

A service template S *exactly* matches a node type N ($S \equiv N$) if the capabilities, requirements, properties, policies and interfaces exposed by S exactly match those of N, namely: (i) the requirements, capabilities and properties of S and N have the same name and type, and they are in a one-to-one correspondence, (ii) the policies exposed by S are applicable to N, and (iii) the interfaces of S and N have the same name, contain the same operations, and are in a one-to-one correspondence.

On the other hand, a service template S *plug-in* matches a node type N ($S \simeq N$) if, intuitively speaking, the former "requires less" and "offers more" than the latter. Namely, (i) for each requirement r of S there exists a requirement of N which has the same name as r and whose type is a sub-type of r's type, (ii) for each capability c of N there exists a capability of S which has the same name as c and whose type is a super-type of c's type, (iii) for each property p of N there exists a property of S which has the same name as p and whose (XML) type is a sub-type of p's type, (iv) the policies exposed by S are applicable to N, and (v) for each interface operation o of N there exists an interface operation of S which exactly matches o.

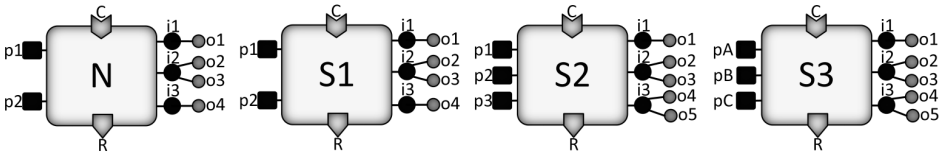


Figure 2: Example of node type and service templates to be matched.

Consider for instance the node type N and the service templates $S1$, $S2$ and $S3$ in Figure 2, where C is a capability of type $CType$, R is a requirement of type $RType$, $p1$, $p2$ and $p3$ are string properties, $i1$, $i2$ and $i3$ are interfaces, and $o1$, $o2$, $o3$, $o4$, and $o5$ are operations[3]. It is easy to see that $S1$ exactly matches N ($S1 \equiv N$) since $S1$ has the same capabilities, requirements, properties, and interfaces of N. The same does not hold for $S2$ and $S3$ ($S2 \not\equiv N$ and $S3 \not\equiv N$). $S2$ plug-in matches N ($S2 \simeq N$) because $S2$ and N expose the same requirements and capabilities, and $S2$ features "more" properties and interface operations than N. On the other hand, $S3$ does not plug-in match N ($S3 \not\simeq N$) since property names differ.

3 Adaptation Methodology

In this section we present a methodology to non-intrusively adapt a service template S into a new service template $newS$ which exactly matches a node type N, and hence to easily reuse an actual service modelled by S in the deployment of cloud-based applications that rely on N.

[3]For the sake of simplicity, we assume that S has no policies and that operations with the same name have the same input and output parameters.

3.1 Adaptation of plug-in matching services

Figure 3 describes the steps to perform when adapting a service template S, which does not exactly match ($S \not\equiv N$) but plug-in matches N ($S \simeq N$), into a new service template $newS$ that exactly matches N ($newS \equiv N$).

(1) Create a new service template $newS$ which initially contains S as the only node template in its topology.

(2) For each capability (property) of N
- (i) define a capability (property) with the same name and type on the boundaries of $newS$, and
- (ii) map the defined capability (property) onto the corresponding one of S.

(3) For each interface i of N
- (i) declare a new interface with the same name and operations on the boundaries of $newS$, and
- (ii) map each operation of the new interface onto the corresponding operation of i.

(4) Add a dummy node template *NoBe* (whose capabilities satisfy the requirements of S and whose requirements are the same of N) to the topology of $newS$.

(5) For each requirement of N
- (i) define a requirement with the same name and type on the boundaries of $newS$, and
- (ii) map the defined requirement onto the corresponding one of *NoBe*.

(where mapping f to f' simply means that f is a reference to f')

Figure 3: Adaptation of plug-in matching service templates.

We now employ the *GameScoreService* in Figure 4 to illustrate the adaptation described in Fig 3. *GameScoreService* is a service which manages score data of various games by associating to each *nick*name a list of $\langle$*game*,*score*$\rangle$ pairs. We assume that *GameScoreService* exposes a *DataConfidentialityPolicy*, that the *GameScoreApp* capability is of type *RecordingCapabilityType*, that the requirements *OSMySQL* and *OSApache* are of *OSRequirementType*, that the property *Version* is a string, and that the *nick* and *game* parameters (both representing names) are strings, while the *score* parameter is an integer.

Figure 4: Available *GameScoreService*.

Figure 5: Target *GScoreNodeType*.

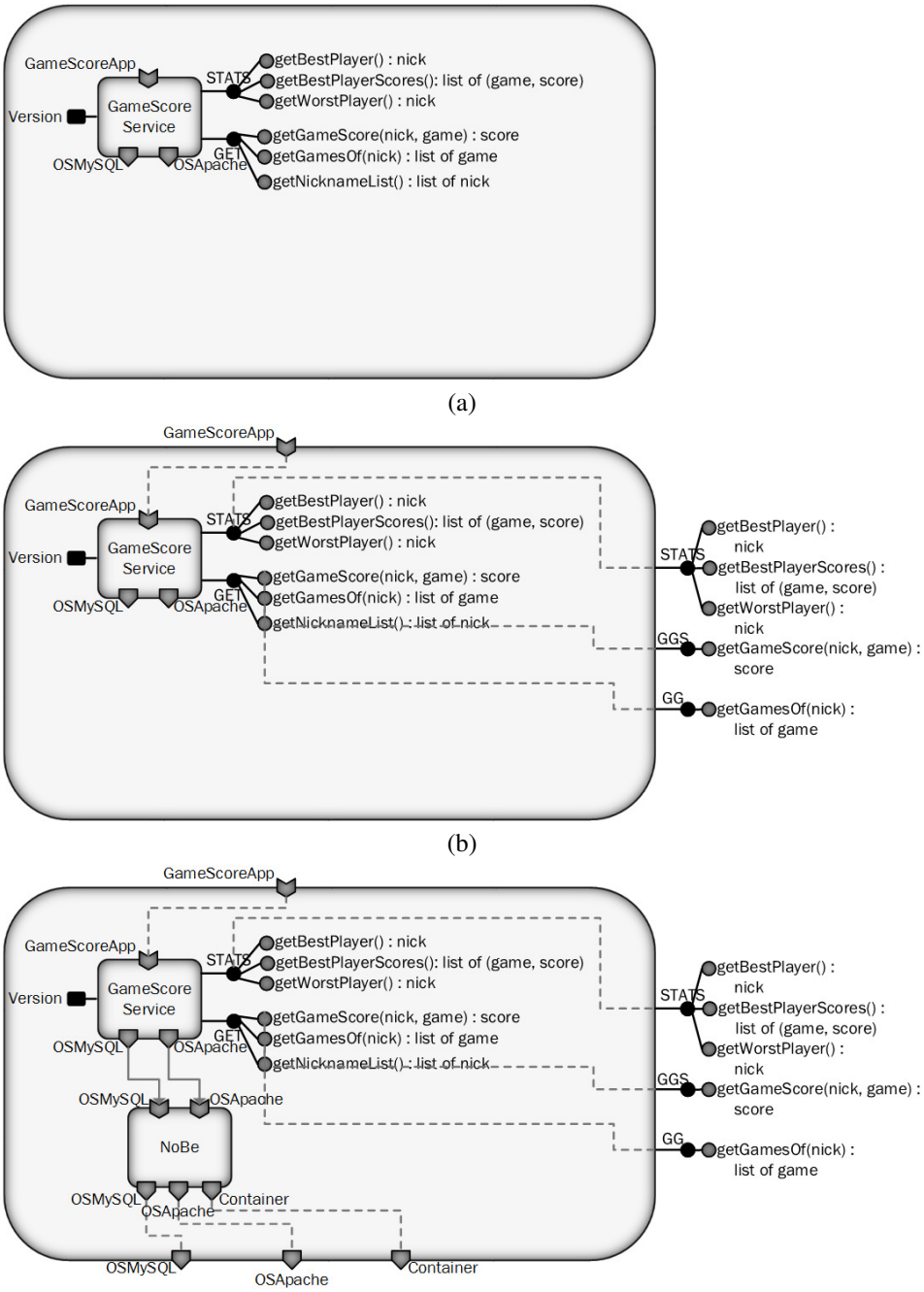


Figure 6: Example of adaptation of a plug-in matching service template: (a) after step 1, (b) after steps 2 and 3, (c) after steps 4 and 5.

Consider now the target *GScoreNodeType* (Figure 5). Suppose that the *DataConfidentialityPolicy* is applicable to *GScoreNodeType*, that the *GameScoreApp* capability is of type *GameScoreApplicationCapabilityType* (which is sub-type of *RecordingCapabilityType*), that *Container* is of type *ContainerRequirementType*, while *OSMySQL* and *OSApache* are both of type *LinuxRequirementType* (which is sub-type of *OSRequirementType*), and that the parameters *nick* and *game* are strings, while *score* is an integer. We can observe that, according to the (informal) definitions given in Sect. 2.2, *GameScoreService* does not exactly match, but it plug-in matches *GScoreNodeType*.

Figure 6 illustrates how the *GameScoreService* service template can be adapted into a new service template *GScoreService* which exactly matches *GScoreNodeType*, by applying the steps in Figure 3. Namely, (a) we create a new service template *GScoreService* which contains *GameScoreService* as the only node template, (b) we suitably define the *GameScoreApp* capability and the interface operations on the boundaries of *GScoreService*, and (c) we add the *NoBe* node to the topology of *GScoreService* so as to expose the desired requirements on its boundaries.

3.2 Adaptation of non plug-in matching services

We now extend the adaptation methodology introduced in the previous section so as to show also how a service template S which does *not* plug-in match a node type N may be non-intrusively adapted into a new service template $newS$ which exactly matches N.

(1) Create the adapted service template $newS$ which initially contains S as the only node template in its topology.

(2) For each capability c (property p) of N

(i) define a capability (property) with the same name and type on the boundaries of $newS$, and

(ii) map the defined capability (property) to that of S which is type-compatible with c (p) and whose name can be considered equivalent to the name of c (p).

(3) For each interface exposed by N, define an interface with the same name on the boundaries of $newS$. For each operation op exposed by N,

(i) define an operation with the same name and parameters in the corresponding interface exposed by $newS$, and

(ii) map the defined operation to
- an operation of S which you consider semantically equivalent to op (e.g., to an operation which exposes the same input and output parameters), or
- a plan which combines the operations of S so as to obtain an operation which you consider semantically equivalent to op.

(4) Add a dummy node template *NoBe* (whose capabilities satisfy the requirements of S and whose requirements are the same of N) to the topology of $newS$.

(5) For each requirement r of N

(i) define a requirement with the same name and type on the boundaries of $newS$, and

(ii) map the defined requirement to that of *NoBe* which is type-compatible with r and whose name can be considered equivalent to r's name.

(where mapping f to f' simply means that f is a reference to f')

Figure 7: Adaptation of *non* plug-in matching service templates.

Figure 7 describes how such an adaptation can be successfully performed (i) when S exposes all the capabilities and properties as N, but in a *syntactically* different way, and/or (ii) when N declares all the requirements which are exposed by S, but in a *syntactically* different way, and/or (iii) when N features one or more interface operation which is not matched by any operation featured by S, while it can be matched by some composition of S operations. The adaptation described in Figure 7 implements the relaxed matching conditions that were defined in [BS13] (in terms of ontology-based name equivalences). According to [BS13], the adaptation process cannot succeed if capability, property or requirement mismatches are not just syntactic or if operation mismatches cannot be solved by means of operation compositions. In other words, the adaptation in Figure 7 fails if (at least) one of the steps cannot be performed, while it succeeds if all steps are performed.

Consider for instance the target *GmsScoreNodeType* (Figure 8), where the *GmsScoreApp* capability is of type *GmsScoreApplicationCapabilityType* (which is sub-type of *RecordingCapabilityType*), *Container* is of type *ContainerRequirementType*, while *OS4MySQL* and *OS4Apache* are both of type *LinuxRequirementType* (which is sub-type of *OSRequirementType*), and the *nick* and *game* parameters are strings, while the *score* parameter is an unsigned integer. We observe that, according to the (informal) definition given in Sect. 2.2,

Figure 8: Target *GmsScoreNodeType*.

the service template *GameScoreService* (Figure 4) does not plug-in match the node type *GmsScoreNodeType*.

Figure 9 illustrates how the *GameScoreService* service template can be adapted so as to exactly match *GmsScoreNodeType*. First, (1) we create a new service template *GmsScoreService* which contains *GameScoreService* as the only node template. Since *GmsScoreApp* and *GameScoreApp* are type-compatible, (2) we adapt the capabilities by adding a *GmsScoreApp* to the boundaries of the adapted service and by mapping it to the *GameScoreApp* of the available service. Analogously, (3) we adapt the unmatched operations *getGmsScore* and *getGmsOf* by mapping them respectively on the *getGameScore* and *getGamesOf* operations. We also adapt the unmatched *getWorstPlayerScores* operation by generating the *WPScoresPlan* which suitably combines the available *getWorstPlayer*, *getGameScore* and *getGamesOf* operations. Finally, we adapt the unmatched requirements by (4) adding the *NoBe* node to the *GmsScoreService*'s topology and by (5) suitably defining the needed requirements on its boundaries.

Once all adaptation steps have been performed (Figure 9(c)) the obtained *GmsScoreService* exposes all the features exhibited by *GmsScoreNodeType*. This implies that they exactly match and hence that *GmsScoreService* can be employed to instantiate *GmsScoreNodeType*.

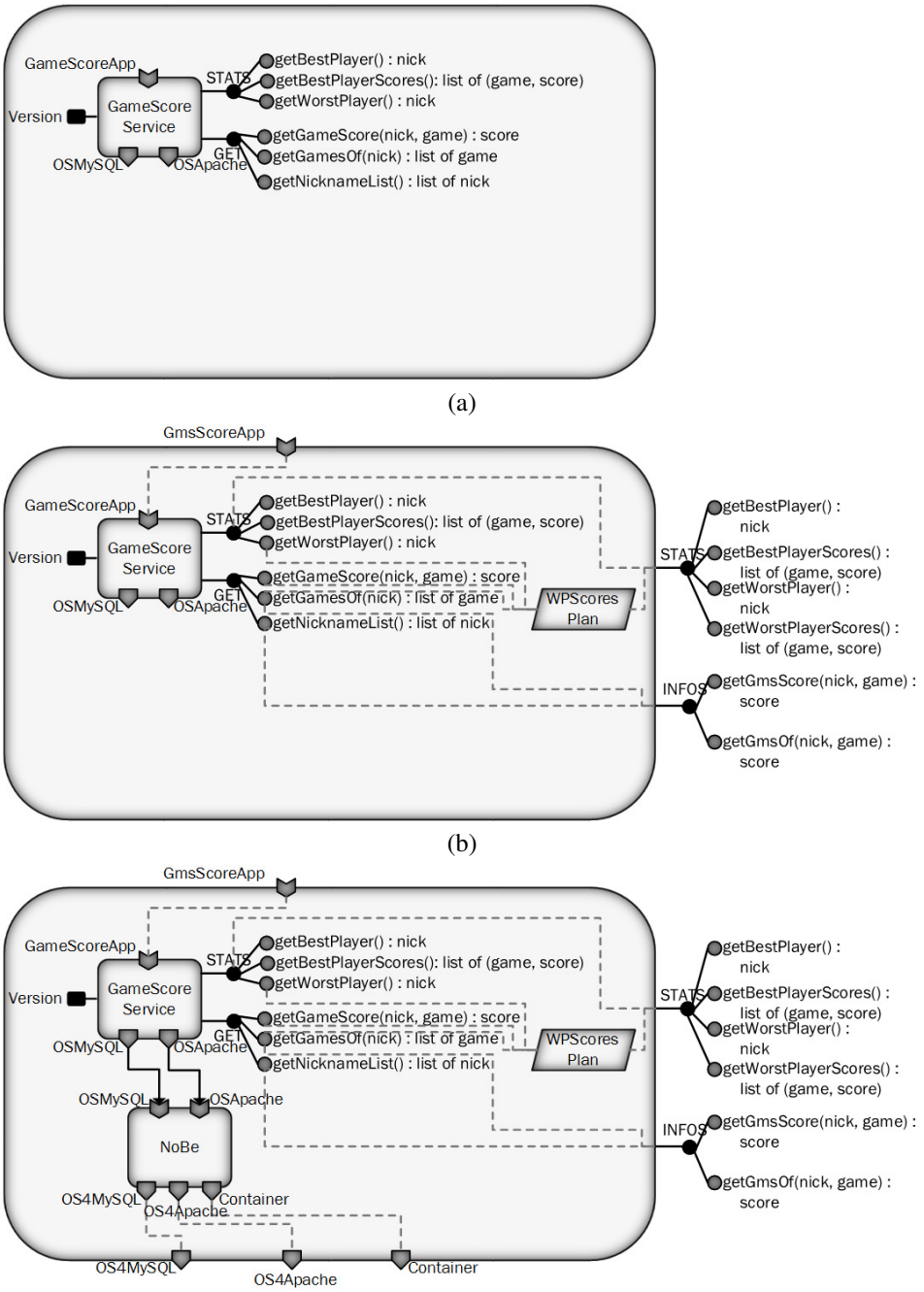


Figure 9: Example of adaptation of a non plug-in matching service template: (a) after step 1, (b) after steps 2 and 3, (c) after steps 4 and 5.

4 Related Work

The development of systematic approaches to reuse existing software is widely recognized as one of the crucial problems in system integration [XW05]. In spite of the increasing availability of cloud services, currently platform-specific code often needs to be manually modified to (re)use services in cloud-based applications. This is obviously an expensive and error-prone activity, as pointed out in [TKLF11], both for the learning curve and for the testing phases needed.

Various efforts have been recently oriented to to try devising systematic approaches to reuse cloud services. For instance, [MPC+11] and [HLT11] propose two approaches to transform platform-agnostic source code of applications developed with a model-driven methodology into platform-specific applications. In contrast, our approach does not restrict to applications developed with a specific methodology, nor it requires the availability of applications' source code, and it is hence applicable also to non open-source, third-party services.

[GMMC13] proposes a framework which allows developers to write the source code of cloud-based services as if they were "in-house" applications. Cloud deployment information must be provided in a separate file, and a middleware layer employs source and deployment information to generate the artifacts to be deployed on cloud platforms. It is worth noting that in [GMMC13] the reuse of a cloud service requires (always) invoking the middleware layer, while in our approach adaptation is performed only once. Moreover, [GMMC13] always requires to write source code, while our approach only requires to edit the application specification.

In general, most existing approaches to the reuse of cloud services support a from-scratch development of cloud-agnostic applications, and do not account for the possibility of adapting existing (third-party) cloud-based services. To the best of our knowledge, ours is the first approach which proposes a methodological approach for adapting existing cloud applications, by relying on TOSCA [OAS13b] as the standard for cloud interoperability, and to support an easy reuse of third-party services.

It is worth noting that the novelty of our approach does not reside in the type of adaptation techniques that we employ to adapt service templates. Indeed, our methodology exploits well-know adaptation patterns (e.g., [GHJV95, BBG+06]) to adapt TOSCA templates. The novelty of (applying) our approach is rather that, in contrast with traditional adaptation approaches (e.g., [BCP06, GMMC13, KMNB+09]), no additional code must be developed to reuse existing cloud-based services. This is thanks to the powerful way in which TOSCA supports the deployment of cloud-based applications. Indeed we exploit the possibility provided by TOSCA of mapping exposed features onto internal ones [OAS13b], and of entirely delegating the management of such mappings to TOSCA containers [OAS13a]. Last, but not least, it is also worth noting that our approach can be fully automated by employing ontologies to resolve syntactic differences, as suggested in [BS13].

5 Concluding remarks

In this paper we presented a methodology to non-intrusively adapt a service template S into a new service template $newS$ which exactly matches a node type N, and hence to

easily reuse any actual service modelled by S to deploy cloud-based applications that rely on N. Notably, as we already observed, the methodology can also be applied to adapt and reuse third-party services, whose source code is not available.

According to the TOSCA specification [OAS13b], we consider that a service template S can be used to substitute a node type N if S has "less" requirements, namely if for each requirement r of S there exists a requirement of N which has the same name as r and whose type is a sub-type of r's type. However, it is worth noting that there may be situations in which ignoring requirements of N that are not requirements of S may lead to loosing some relevant constraints on N (e.g., that a node template with type N should be linked to a private network). Interested readers may refer to [BSW14] for a discussion of this and other possible research directions for TOSCA.

We conclude the paper by mentioning some directions for our future work. One is obviously the implementation of a tool supporting the matching described in [BS13] as well as the adaptation methodology described in this paper. This is scope for our immediate future work, together with experimenting the methodology via such tool. Moreover, the matching and adaptation tool may also be fruitfully integrated in a plug-in for the TOSCA implementations that are currently under development (e.g., the Winery editor [KBBL13] and the OpenTOSCA engine [BBH+13]) as soon as a high-level API will be available

Another interesting direction for future work starts from the observation that cloud applications do share management infrastructure. The possibility of reusing fragments of available service templates is hence of clear practical interest. Extending the proposed methodology in this direction is also scope for our immediate work.

A further possible direction for future work is to investigate the applicability to the problem of adapting TOSCA service templates of more complex adaptation techniques, like those that input also an explicit specification of the desired adaptation (e.g., like [BCP06]).

References

[BBG+06] S. Becker, A. Brogi, I. Gorton, S. Overhage, A. Romanovsky, and M. Tivoli. Towards an Engineering Approach to Component Adaptation. In *Architecting Systems with Trustworthy Components*, volume 3938, pages 193–215. Springer, 2006.

[BBH+13] T. Binz, U. Breitenbücher, F. Haupt, O. Kopp, F. Leymann, A. Nowak, and S. Wagner. OpenTOSCA – A Runtime for TOSCA-based Cloud Applications. In *11th International Conference on Service-Oriented Computing*. Springer, 2013.

[BCP06] A. Brogi, C. Canal, and E. Pimentel. On the semantics of software adaptation. *Science of Computer Programming*, 61(2):136 – 151, 2006.

[BS13] A. Brogi and J. Soldani. Matching Cloud Services with TOSCA. In C. Canal and M. Villari, editors, *Advances in Service-Oriented and Cloud Computing*, volume 393 of *Communications in Computer and Information Science*, pages 218–232. Springer, 2013.

[BSW14] A. Brogi, J. Soldani, and P. Wang. TOSCA in a nutshell: Promises and perspectives. In *ESOCC 2014 - Proceedings of the 3rd European Conference on Service-Oriented and Cloud Computing*, Lecture Notes in Computer Science. Springer, 2014. *[In press]*.

[GHJV95] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. *Design Patterns: Elements of Reusable Object-oriented Software*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[GMMC13] J. Guilln, J. Miranda, J. M. Murillo, and C. Canal. A service-oriented framework for developing cross cloud migratable software. *Journal of Systems and Software*, 86(9):2294 – 2308, 2013.

[HLT11] M. Hamdaqa, T. Livogiannis, and L. Tahvildari. A reference model for developing cloud applications. In F. Leymann, I. Ivanov, M. van Sinderen, and B. Shishkov, editors, *CLOSER 2011 - Proceedings of the 1st International Conference on Cloud Computing and Services Science*. SciTePress, 2011.

[KBBL13] O. Kopp, T. Binz, U. Breitenbücher, and F. Leymann. Winery – Modeling Tool for TOSCA-based Cloud Applications. In *11th International Conference on Service-Oriented Computing*. Springer, 2013.

[KMNB+09] W. Kongdenfha, H. R. Motahari-Nezhad, B. Benatallah, F. Casati, and R. Saint-Paul. Mismatch Patterns and Adaptation Aspects: A Foundation for Rapid Development of Web Service Adapters. *IEEE Transactions on Services Computing*, 2(2):94–107, 2009.

[MPC+11] B. Martino, D. Petcu, R. Cossu, P. Goncalves, T. Mahr, and M. Loichate. Building a Mosaic of Clouds. In *Euro-Par 2010 Parallel Processing Workshops*, volume 6586 of *Lecture Notes in Computer Science*, pages 571–578. Springer, 2011.

[OAS13a] OASIS. Topology and Orchestration Specification for Cloud Applications (TOSCA) Primer, Version 1.0. "`http://docs.oasis-open.org/tosca/tosca-primer/v1.0/tosca-primer-v1.0.pdf`", 2013.

[OAS13b] OASIS. Topology and Orchestration Specification for Cloud Applications (TOSCA), Version 1.0. "`http://docs.oasis-open.org/tosca/TOSCA/v1.0/TOSCA-v1.0.pdf`", 2013.

[PMPC13] D. Petcu, G. Macariu, S. Panica, and C. Craciun. Portable Cloud applications - From theory to practice. *Future Generation Computer Systems*, 29(6):1417 – 1430, 2013.

[TKLF11] V. Tran, J. Keung, A. Liu, and A. Fekete. Application Migration to Cloud: A Taxonomy of Critical Factors. In *Proceedings of the 2nd International Workshop on Software Engineering for Cloud Computing*, SECLOUD '11, pages 22–28. ACM, 2011.

[XW05] X. Xiong and Z. Weishi. The Current State of Software Component Adaptation. In *Semantics, Knowledge and Grid, 2005. SKG '05. First International Conference on*, pages 103–103, 2005.

Automatic Topology Completion of TOSCA-based Cloud Applications

Pascal Hirmer[1], Uwe Breitenbücher[2], Tobias Binz[2], Frank Leymann[2]

Pascal.Hirmer@ipvs.uni-stuttgart.de
[1]Institute for Parallel and Distributed Systems
[2]Institute of Architecture of Application Systems
University of Stuttgart

Abstract: Automation of application provisioning is one of the key aspects of Cloud Computing. Standards such as the Topology and Orchestration Specification for Cloud Applications (TOSCA) provide a means to model application topologies which can be provisioned fully automatically. Many automated provisioning engines require that these topologies are complete in the sense of specifying all application, platform, and infrastructure components. However, modeling complete models is a complex, time-consuming, and error-prone task that typically requires a lot of technical expertise. In this paper, we present an approach that enables users to model incomplete TOSCA application topologies that are completed automatically to deployable, complete models. This enables users to focus on the business-relevant application components and simplifies the creation process tremendously by minimizing the required effort and know-how. We prove the technical feasibility of the presented approach by a prototypical implementation based on the open source modeling tool Winery. In addition, we evaluate the approach by standards-compliance and performance.

1 Introduction and Background

In recent years, Cloud Computing gained a lot of attention due to its economical benefits through payment on demand, flexibility, scalability and self-service [MG11]. One key to enable this is full automation of application provisioning and management. To make this possible, standards have been developed to automate the provisioning process. One example is the *Topology and Orchestration Specification for Cloud Applications (TOSCA)* [OA13a], a standard that enables creating portable application models including management functionalities [Bi14]. Such topology models are called *Topology Templates* and describe the application's components and their relationships as directed topology graph. Topology Templates are usually modeled by *Cloud Service Developers* with the aim to deploy their developed applications in a Cloud environment. However, Cloud Service Developers – as defined in [OA13b] – are no experts in Cloud infrastructure and platform components. For non-experts, modeling a Topology Template for applications consisting of multiple different components that run on heterogeneous Cloud offerings is a complex task: (i) the configuration of components must be defined, (ii) interoperability between systems has to be considered, and (iii) the artifacts for deployment must be at-

tached to enable a fully automated provisioning of the application. Thus, a lot of technical know-how regarding Cloud infrastructures, platforms, providers and software components is required. As a result, modeling complex applications leads to huge effort causing high costs.

The result of this discussion is that it is of vital importance to ease the modeling of complex Cloud applications to reduce the required know-how. As we tackle these issues, the driving research question of this paper is: *How can the (i) manual effort and (ii) technical expertise currently required for modeling TOSCA topology models be minimized?* We answer this question by presenting an approach that enables modelers to create *incomplete topology models* that are completed automatically by the system we present. Note that in the context of this paper, the term *incomplete* is used as a synonym for the term *undeployable*. An *undeployable* topology can't be used for automatic provisioning due to missing components. In these incomplete models, only the business-relevant components have to be specified by the modeler—the required underlying infrastructure and platform components including their relations are inserted fully automatically while guaranteeing a syntactically and semantically correct model. The resulting models can be used as input for automated TOSCA provisioning engines that require complete models. The presented approach tackles the mentioned issues in three different dimensions: first, the complexity of modeling is reduced as only business-relevant information has to be provided. Second, the required manual effort for modeling is reduced. Third, as the approach is based on TOSCA, all models are portable which helps to avoid vendor lock-in. All these properties help to decrease the costs of developing, provisioning, and maintaining Cloud-based applications, thus, leading to economical advantages when modeling applications and their management using TOSCA. We prove the technical feasibility of the approach by a prototypical implementation and evaluate the concept in terms of standards-compliance and performance.

The remainder of this paper is structured as follows: after presenting related work in Section 2, we introduce the basics of TOSCA that are required to understand the presented approach in Section 3. In the following Section 4, we provide a motivating scenario which is used throughout the paper to explain the developed concepts. In Section 5, the main contribution of this paper is presented: we show a concept to automatically complete incomplete TOSCA topologies. We evaluate in Section 6, validate the approach in Section 7, and give a conclusion as well as an outlook on future work in Section 8.

2 Related Work

This section discusses related works which have been published in this research field and serve as basis for the presented approach. The article "Managing the Configuration Complexity of Distributed Applications in Internet Data Centers" [Ei06] presents an approach for a model driven provisioning and management of applications for the TOSCA standard. Using transformations, topologies are transformed to several levels of abstraction. Similar to the approach presented in this paper, the resulting topology can be used for provisioning using a provisioning engine. Arnold et al. [Ar07] present a pattern based

approach for the deployment of service-oriented architectures. By defining patterns for certain deployment scenarios (e.g. high availability), the deployment can be managed dependent on the requirements. These pattern are very similar to the concept of Requirements and Capabilities in the TOSCA standard. Binz et al. [Bi12] present an approach to use topologies to describe complex enterprise IT systems. These topologies are called *Enterprise Topology Graphs (ETG)* and describe the system components and their connections similar to TOSCA. However, while TOSCA defines topologies for the provisioning of applications, ETGs describe the state of already provisioned systems. Breitenbücher et al. [Br12] present Vino4TOSCA, a visual notation of the TOSCA standard. In this paper we use the Vino4TOSCA notation to present our approach graphically. Vino4TOSCA is also used for the implementation of the modeling tool Winery [Ko13]. Eilam et al. [Ei11] and Kalantar et al. [Ei06] present a combination of a model and workflow driven approach for the deployment of applications. A deployment model describing the final state is used to create a workflow model containing operations to transform an initial topology to the final state. Using a workflow engine this transformation can be processed automatically. Képes [Ke13] presents an implementation of the Provisioning API for OpenTOSCA (cf. Section 5.3.1), which can be used for the automatic completion of topologies that shall be provisioned in the OpenTOSCA Runtime Environment. Finally, Brogi et al. [BS13] present an approach to match Service Templates and Node Types, also using the concepts of TOSCA Requirements and Capabilities.

3 TOSCA

In this section, we introduce the basic concepts of TOSCA, which are relevant to understand the approach presented in this paper. TOSCA is an OASIS standard introduced in November 2013 to describe Cloud applications in a portable way. TOSCA-based descriptions define (i) the structure as well as (ii) management functionalities of Cloud-based applications, e.g., management functionalities can be implemented using executable management plans. Although TOSCA is a relatively new standard, several tools exist that ease modeling, provisioning, and management of TOSCA-based applications. The open source ecosystem *OpenTOSCA*, for example, includes a graphical modeling tool called *Winery* [Ko13] and a plan-based provisioning and management Runtime Environment [Bi13] which can be used to provision and manage TOSCA applications fully automatically by executing management plans. The TOSCA Primer [OA13b] defines the roles *Cloud Service Consumer*, *Cloud Service Developer* and *Cloud Service Provider*. The approach presented in this paper mainly supports the *Cloud Service Developer*, who models TOSCA topologies to provision and manage an application in the Cloud. Further details on the TOSCA standard can be found in the official OASIS TOSCA specification [OA13a], TOSCA Primer [OA13b], or Binz et al. [Bi14].

The core of the application description in TOSCA is the *Topology Template*, a directed graph containing *Node Templates* (vertices) and *Relationship Templates* (edges). Node Templates may describe all components of an application, including all software and hardware components. The relations between those Node Templates are represented by Rela-

tionship Templates. Node and Relationship Templates are typed by *Node Types* and *Relationship Types*, respectively. Types define the semantics of the templates, as well as their properties, provided management operations, and so on. Types can be refined or extended by an inheritance mechanism. *TOSCA Requirements* are used to define requirements and restrictions for a Topology Template. Based on its Requirements, it can be determined if a Node Template must be connected and, as a consequence, if further Node Templates are missing in the Topology Template to satisfy all Requirements. Each Requirement is derived from a *Requirement Type* defining the structure and semantics of the Requirement, which makes it possible to process Requirements based on their QName. To *fulfill* the Requirements of a Topology Template, every Node Template containing a TOSCA Requirement has to be connected to a Node Template containing a suitable *TOSCA Capability* using a Relationship Template. Which type of Capability is demanded by the Requirement is described by the XML attribute *requiredCapabilityType* of its Requirement Type. Besides being attached to Node Templates, Requirements can also be attached to Node Types as Requirement Definitions which makes them Requirements for all Node Templates of this Node Type. TOSCA specifies an exchange format called *Cloud Service Archive* (CSAR) to package Topology Templates, types, associated artifacts, plans, and all required files into one self-contained package. This package is portable across different standards-compliant TOSCA Runtime Environments [Br14a].

4 Motivating Scenario

This section introduces the motivating scenario that is used throughout the paper to illustrate our approach. It describes the modeling and completion of a TOSCA Topology Template to provision a Web-application on Amazon EC2[1]. The application used in this scenario consists of a Java-based frontend and a MySQL database to store its data. This application shall be hosted in the Amazon Cloud to be accessible remotely. As described in Section 1, in some TOSCA runtimes (e.g. OpenTOSCA), a complete and technically detailed TOSCA Topology Template has to be modeled to enable automatic provisioning. To simplify the modeling of such – currently XML-based – topologies, the Cloud Service Developer may use the graphical TOSCA modeling tool Winery [Ko13]. However, even with graphical modeling, the application programmer requires extensive knowledge of the technical details for provisioning Cloud applications as the whole topology model has to be specified in detail. The approach presented in this paper enables also inexperienced users to provision applications in the Cloud using TOSCA. This is enabled by allowing users to model incomplete topologies that do not specify the whole topology in detail but only the business-relevant components. An incomplete model for this motivation scenario is shown in Figure 1 on the left: the shown Topology Template is incomplete as it only contains two Node Templates, one representing the Java-based Web application packed as Java Web Archive (WAR), and the MySQL database, without defining the underlying infrastructure etc. However, this topology cannot be provisioned by runtimes that require complete models (e.g., OpenTOSCA): (i) the Webserver for running the Java frontend and

[1] https://aws.amazon.com/de/ec2/

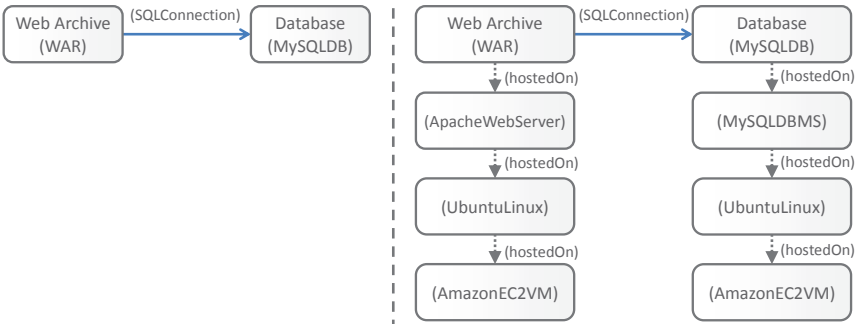


Figure 1: Incomplete (left) and complete (right) Topology Templates of the Motivation Scenario

database management system components are missing, (ii) the operating systems are not specified, and (iii) no information is provided about where the stacks have to be deployed. Thus, to make an automatic provisioning of this application possible, the missing technical information about the infrastructure and middleware components have to be defined. An example of a complete topology model that may result from applying our approach is shown in Figure 1 on the right: this model contains all middleware and infrastructure components and provides, therefore, a complete model that can be used to provision the application (we omitted properties, operations etc. to simplify the figure). In the next section, we present our approach for topology completion that supports the completion of incomplete Topology Templates.

5 Automated Completion of Incomplete TOSCA Topology Templates

In this section, we present the main contribution of this paper in the form of an approach for automatic TOSCA Topology Template completion. We describe the presented approach as method that is subdivided into five steps, which are shown in Figure 2 and explained in detail in the following subsections: (i) modeling of incomplete Topology Templates, (ii) target runtime selection, (iii) automated completion to deployable model, (iv) optional manual refinement of the completed solution, and (v) final deployment of the application.

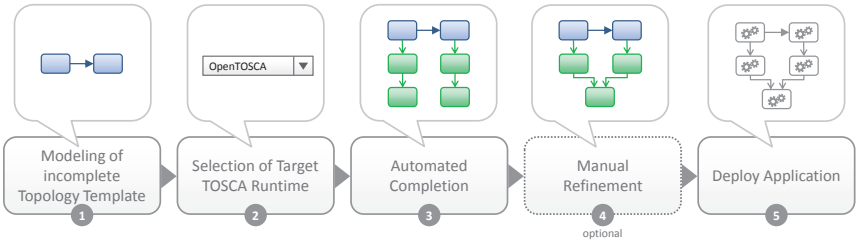


Figure 2: Automated TOSCA Topology Template Completion Method

5.1 Step 1: Modeling of Incomplete Topologies using TOSCA Requirements

In the first step, an incomplete Topology Template defining Requirements (cf. Section 3) as hints for the automatic completion is modeled. We define an incomplete topology as *a Topology Template that cannot be used for automated provisioning due to missing components*. Thus, considering the motivating scenario, users do not want to care about details such as the employed Cloud provider or the used operating system. However, there are two general requirements that have to be fulfilled: the application has to be deployed on a Web server and the MySQL database on a compatible database management system (DBMS). Consequently, these components must be added automatically by the topology completion based on the Requirements. To enable this, TOSCA Requirements have to be declared in the incomplete topology model. Therefore, the Requirements are either (i) defined directly by the Requirement Definitions of the corresponding Node Types (which is the common case) or (ii) added manually by the Cloud Service Developer (e.g., to additionally refine Requirements). In our example, two Requirements "WebserverContainerRequirement" and "MySQLDBMSRequirement", as shown in Figure 3, are defined. The result is an incomplete Topology Template that contains all application-specific components including a description of the corresponding Requirements that must be fulfilled to run the application.

Figure 3: Incomplete TOSCA Topology Template with Requirements

5.2 Step 2: Selection of Target TOSCA Runtime Environment

The TOSCA standard specifies a meta-model for describing Cloud application structures including all components and relationships among them. However, the standard currently neither predefines Node Types nor Relationship Types. Thus, there are *no* well-defined semantics of Node and Relationship Types. As a consequence, different TOSCA Runtime Environments (which are not standardized) typically support different features, containers, and Cloud providers. For example, the OpenTOSCA Runtime Environment [Bi13] currently supports automatically deploying applications on Amazon EC2 but not on Microsoft Azure. Thus, it supports the deployment based on the *AmazonEC2* Node Type but not on the *Azure* Node Type. Therefore, the completion of incomplete Topology Templates must consider the final TOSCA Runtime Environment and the supported features. To be precise, the completion must be aware which Node and Relationship Types can be provisioned by the employed runtime. Therefore, we have to select the target environment in this step to provide the required information for completing the model appropriately.

5.3 Step 3: Automatic Completion of Incomplete TOSCA Topologies

TOSCA topology completion can be done in two ways, the Black Box and the Glass Box approach: In the *Black Box approach*, used for example by inexperienced users, there is no user interaction during the completion of the topology, i.e., all decisions are made automatically. An incomplete TOSCA Topology Template (cf. Section 5.1) serves as input and one or more completed topologies are returned. If there is more than one resulting topology, the user can choose after the completion finished. In contrast, the *Glass Box approach* enables the user to control and observe the topology completion step by step. Therefore, if multiple possibilities are found to fulfill one Requirement or to connect two Node Templates, the user will be prompted to decide interactively. Using this approach is reasonable if the user has a concrete vision of the resulting completed topology.

5.3.1 Basics

Before presenting the topology completing algorithm in Section 5.3.2, this section introduces concepts on how to connect two Node Templates, how to check a topology for being provisionable in a TOSCA Runtime Environment, and introduces a Type Repository.

Connecting Node Templates: During the topology completion, Node Templates will be added to the topology. These have to be connected to the existent topology using Relationship Templates. To help determining the right Relationship Type to connect two Node Templates, the TOSCA standard allows to define valid source and target Node Types in each Relationship Type. These elements contain a reference to the Node Types that can be connected using the Relationship Type. Furthermore, these elements can also contain Requirement and Capability Types. To find a Relationship Type that is able to connect two Node Templates, all available Relationship Types are checked. If the type of these Node Templates is contained in the ValidSource and ValidTarget elements of a Relationship Type it can be used. In case the Node Templates contain Requirements or Capabilities, it is checked by their types if the Relationship Type is able to connect them. After the algorithm finished, the Relationship Type is instantiated to a Relationship Template – containing references to the connected Node Templates – and inserted into the topology.

Provisioning API: The completion algorithm requires information about the supported features of the target TOSCA Runtime Environment which shall provision the topology. This information is made available through the *Provisioning API*, which must be provided by each TOSCA Runtime Environment that shall be supported by our approach. The following three operations are implemented by the Provisioning API: (i) Check if a given Topology or Node Template can be automatically provisioned by the respective TOSCA Runtime Environment, (ii) propose Node and Relationship Templates to be inserted below a given Node Template that shall be provisioned, and (iii) check if the TOSCA Runtime Environment is able to handle a certain TOSCA Requirement directly without inserting additional nodes. In case a Requirement can be handled this way, it doesn't have to be fulfilled by the topology completion algorithm.

Type Repository: To complete a Topology Template, Node and Relationship Templates are inserted based on predefined Node and Relationship Types, which are made available to the completion algorithm through a type repository.

5.3.2 Topology Completion – Algorithm

In this section, we present the algorithms used to complete incomplete TOSCA Topology Templates to be provisioned on a given target TOSCA Runtime Environment. In addition, the flag *FulfillAllRequirements* indicates whether all Requirements have to be fulfilled or just the Requirements that cannot be handled by the TOSCA Runtime Environment. The completion of a topology is done in two steps: (i) fulfillment of Requirements and (ii) verification of completeness. To simplify the presentation, the Black Box approach is discussed here, i.e., proceeding without any user interaction. In the Glass Box approach, in case there are multiple possibilities to process, the user is prompted to choose one.

Step i: Fulfillment of Requirements: First, the Requirement analysis iterates over all the Node Templates of a Topology Template and their Types to check for existing Requirements or Requirement Definitions. In case a Requirement exists at a Node Template or Type, it is checked for fulfillment: If the Node Template containing a Requirement is connected to a Node Template containing a suitable Capability, the Requirement is already fulfilled. If required (flag *FulfillAllRequirements* is set to *false*), it is also checked whether the TOSCA Runtime Environment can already handle the Requirement using the Provisioning API introduced in Section 5.3.1. If that's the case, the Requirement can also be marked as fulfilled. For non-fulfilled Requirements, the required Capability Type is found by checking the *requiredCapabilityType* attribute of its Requirement Type. Thereupon, a Node Type is searched in the repository offering the suitable Capability Type. In case a suitable type can be found, a Node Template is inserted into the Topology Template by connecting it to the Node Template containing the Requirement (cf. Section 5.3.1). In this way all Requirements are processed and fulfilled. Finally, the analysis phase is restarted to check whether new Requirements have been added through the types of the inserted Node Templates. This leads to a recursive algorithm finishing when all Requirements have been fulfilled.

Step ii: Verification of Completeness: A topology without non-fulfilled Requirements does not guarantee that it can be provisioned in the selected TOSCA Runtime Environment. This is the case if the Runtime Environment requires further components in the topology for provisioning. Those components possibly were not added by the topology completion due to missing TOSCA Requirements or Node Types in the Type Repository. Therefore, using the Provisioning API introduced in Section 5.3.1 it is checked whether the completed topology can be automatically provisioned in the used TOSCA Runtime Environment or not. If the topology cannot be provisioned, the API returns a Topology Template containing the missing Templates which will be inserted into the topology. In case any Node Templates are added in Step ii, the first step is re-processed because new Requirements could have been added through the types of the inserted Node Templates.

Result: After the algorithm has finished, the topology is complete and can be used for provisioning in the selected TOSCA Runtime Environment.

5.4 Step 4: Manual Refinement

After the execution of the completion algorithm, the user can refine the resulting topology manually. For example, the user might want to switch the Cloud provider, which can be done by changing the type of the VM Node Templates. Note that it is necessary to check the provisionability of the refined topology again after changing it manually.

5.5 Step 5: Deploy Application

In the last step of the method, the complete application topology gets deployed on the selected TOSCA Runtime Environment. In Breitenbücher et al. [Br14a], we show how Topology Templates are interpreted and provisioned fully automatically using OpenTOSCA.

6 Evaluation

In this section, we conduct a qualitative and quantitative evaluation based on a toolchain, standards compliance, as well as the completion's performance.

Toolchain—End-to-End Prototype Support: The presented approach bridges the gap between the modeling of TOSCA topologies and the provisioning of applications. We developed the open source tool Winery that provides an user interface to model TOSCA topologies graphically and offers a management backend to manage types and artifacts. Types and topologies can be exported as CSAR files. Furthermore, a TOSCA Runtime Environment has been developed called OpenTOSCA [Bi13] which is able to run those CSARs. Using the plan generator presented in [Ke13] and [Br14a], provisioning plans can be generated and injected fully automatically into the CSAR running in the OpenTOSCA Runtime Environment. The last link of the toolchain is the "Vinothek" [Br14b] providing a graphical end-user interface to provision new application instances using OpenTOSCA. With our implementation of the presented approach, Topology Templates can be completed and provisioned fully automatically using the OpenTOSCA ecosystem. In conclusion, we provide tools to support an end-to-end TOSCA toolchain, as shown in Figure 4.

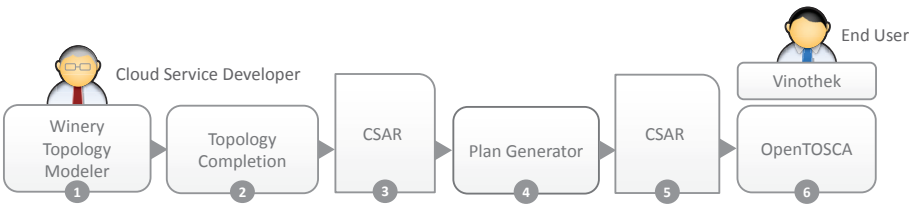


Figure 4: End-to-end Tool support

Standards Compliance: Standards are a means to achieve reusability, interoperability, portability, and maintainability resulting in higher productivity and lower cost. The approach presented in this paper addresses these issues as it exclusively employs the OASIS standard TOSCA as exchange format and is integrated into a chain of further TOSCA-compliant tools. Thus, the approach enables portability [Bi14] and automatic provisioning of applications based on standards.

Performance Measurement: Several performance measurements have been conducted, showing the runtime of the topology completion algorithm. The measurements are based on timestamps taken of the prototypical implementation. As displayed in Table 1, the completion algorithm has an exponentially growing run time depending on the amount of Requirements being completed. The measured runtime can be explained with the recursive algorithm being used.

Table 1: Topology Completion Duration

# Requirements	Completion Time ∅	Time / Requirement ∅
50	31 ms	0.62 ms
250	47 ms	0.18 ms
500	63 ms	0.12 ms
2 500	296 ms	0.11 ms
5 000	1 014 ms	0.20 ms
25 000	34 897 ms	1.39 ms
50000	168600 ms	3.372 ms

7 Validation & Prototype

The validation is based on the prototypical implementation we developed and integrated into the TOSCA modeling tool *Eclipse Winery* [Ko13]. Winery is an Eclipse open source project[2] and the Winery Wiki provides further information on the topology completion facilities we integrated[3]. Using Winery, the user graphically models incomplete topologies using the Vino4TOSCA visual notation [Br12]. The completion process is started from Winery by clicking the button "Complete Topology" and selecting the approach to use, as well as the targeted TOSCA Runtime Environment. In case the Glass Box approach is chosen, in every step, the completion gives feedback about the Node Templates that could be inserted into the topology and prompts the user to select the desired one (cf. Figure 5). When using the Black Box approach, the topology will be completed without further user interaction. The completed topology is then displayed in Winery.

[2] http://eclipse.org/winery/

[3] https://wiki.eclipse.org/Winery/Addons/Winery_Topology_Completion

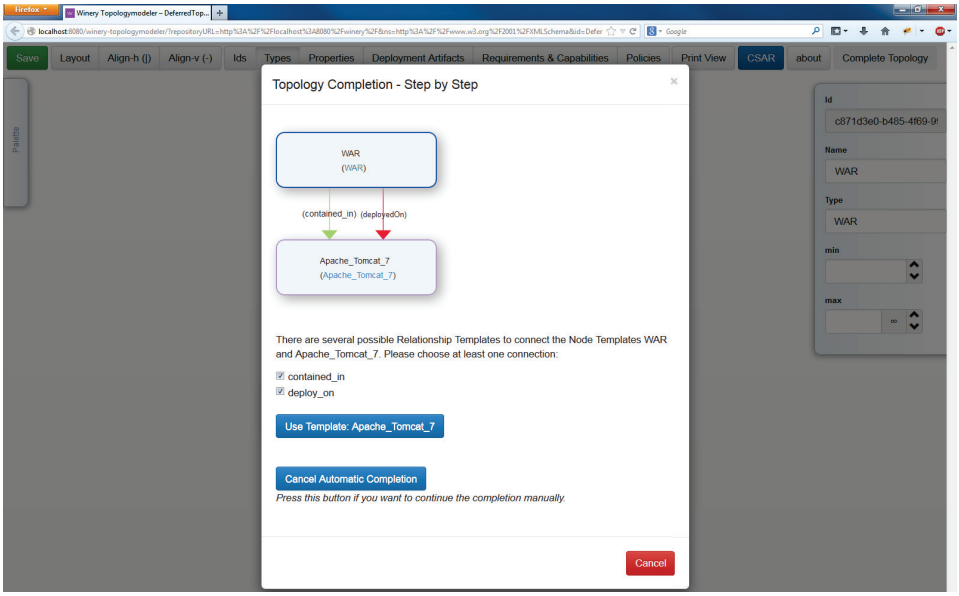


Figure 5: Topology Completion in Winery

8 Conclusion and Future Work

In this paper, we presented an approach to simplify the modeling of TOSCA topologies by automatically completing incomplete topologies. The completed topologies can then be automatically provisioned in the respective TOSCA Runtime Environment. We enable the user to create incomplete topologies that only define application components without giving information about platforms and infrastructure. These topologies are completed by adding middleware and infrastructure components until the topology is complete and can be processed by a provisioning engine. The user can control the completion using the presented Glass Box approach enabling him to determine which components will be added to the topology. The completion can also be processed without any user interaction by using the Black Box approach to reduce the modeling effort to a minimum. As mentioned before, the presented solution is usually used by Cloud Service Developers, exclusively. These developers want to benefit from the advantages of the TOSCA standard to enable automatic provisioning and management of their applications in the Cloud, without caring about infrastructure and platform details. Using the presented approach, it is sufficient to define information about the application to be deployed and ignore all other details. In practical use, this leads to an easy, low-effort way to deploy applications in the Cloud while also profiting from the advantages of the TOSCA standard. In conclusion, the presented solution is highly applicable for practical scenarios. In the future, the presented concepts can be improved by including the completion of TOSCA properties (e.g. IP addresses of servers, user / passwords of database etc.). In this approach the TOSCA properties are completed only with predetermined default values. Furthermore, the current approach can

be extended in the future by using TOSCA Policies to define non-functional requirements (e.g. security, costs, availability) to steer the topology completion.

Acknowledgment

This work was partially funded by the BMWi project CloudCycle (01MD11023).

References

[Ar07] Arnold et al. Pattern Based SOA Deployment. In *ICSOC*, pages 1–12. Springer, 2007.

[Bi12] Tobias Binz et al. Formalizing the Cloud through Enterprise Topology Graphs. In *CLOUD*, pages 742–749. IEEE, June 2012.

[Bi13] Tobias Binz et al. OpenTOSCA - A Runtime for TOSCA-based Cloud Applications. In *ICSOC*, pages 692–695. Springer, December 2013.

[Bi14] Tobias Binz et al. *TOSCA: Portable Automated Deployment and Management of Cloud Applications*, pages 527–549. Advanced Web Services. Springer, January 2014.

[Br12] Breitenbücher et al. Vino4TOSCA: A Visual Notation for Application Topologies based on TOSCA. In *CoopIS*, pages 416–424. Springer, September 2012.

[Br14a] Breitenbücher et al. Combining Declarative and Imperative Cloud Application Provisioning based on TOSCA. In *IC2E*, pages 87–96. IEEE, March 2014.

[Br14b] Breitenbücher et al. Vinothek - A Self-Service Portal for TOSCA. In *ZEUS*, pages 69–72. CEUR-WS.org, March 2014.

[BS13] Brogi and Soldani. Matching Cloud Services with TOSCA. In *Advances in Service-Oriented and Cloud Computing*, pages 218–232. Springer, 2013.

[Ei06] T. Eilam et al. Managing the configuration complexity of distributed applications in Internet data centers. *Communications Magazine, IEEE*, 44(3):166–177, March 2006.

[Ei11] T. Eilam et al. Pattern-based Composite Application Deployment. In *IM*, pages 217–224. IEEE, May 2011.

[Ke13] Kalman Kepes. Konzept und Implementierung eine Java-Komponente zur Generierung von WS-BPEL 2.0 BuildPlans für OpenTOSCA. Bachelor Thesis: University of Stuttgart, Institute of Architecture of Application Systems, July 2013.

[Ko13] Oliver Kopp et al. Winery – A Modeling Tool for TOSCA-based Cloud Applications. In *ICSOC*, pages 700–704. Springer, December 2013.

[MG11] Mell and Grance. The NIST Definition of Cloud Computing. Technical Report 800-145, National Institute of Standards and Technology (NIST), September 2011.

[OA13a] OASIS. Topology and Orchestration Specification for Cloud Applications. http://docs.oasis-open.org/tosca/TOSCA/v1.0/os/TOSCA-v1.0-os.html, Nov 2013.

[OA13b] OASIS. TOSCA Primer. http://docs.oasis-open.org/tosca/tosca-primer/v1.0/cnd01/tosca-primer-v1.0-cnd01.pdf, November 2013.

Towards Uniform BPEL Engine Management in the Cloud

Simon Harrer, Jörg Lenhard, Guido Wirtz

Distributed Systems Group
University of Bamberg
An der Weberei 5
96047 Bamberg
Germany
firstname.lastname@uni-bamberg.de

Tammo van Lessen

innoQ Deutschland GmbH
Krischerstr. 100
40789 Monheim am Rhein
Germany
tammo.van-lessen@innoq.com

Abstract: The *Web Services Business Process Execution language* (BPEL) is a standard for modeling and executing automated processes and is tailor-made for service orchestration. BPEL specifies a serialization format which every BPEL implementation has to understand, thus allowing for the portability of processes among runtime engines. Although the modeling and execution of BPEL processes is portable between engines to a large degree, the lifecycle management of BPEL processes is not standardized and varies a lot for different engines. This paper presents a first approach for a uniform and cloud-based lifecycle management of BPEL processes and engines. We infer a uniform interface for the lifecycle management from the capabilities of current engines and provide a prototypic implementation of a tool that manages processes and engines on a TOSCA-compliant infrastructure.

1 Introduction

The *Web Services Business Process Execution Language 2.0* (WS-BPEL, or BPEL for short) [OAS07] is an OASIS standard specifying a process language with which executable Web Services-based orchestrations can be modeled. If these orchestrations or processes are modeled in a standard conformant manner, they can be executed on any standard conformant BPEL engine, ensuring portability and avoiding vendor lock-in. Similar to Java's concept of *write once, run anywhere*, BPEL was developed with the mantra *model once, run anywhere* in mind [KKL06]. However, the standardization initiative did not take all management related tasks into account. In terms of the BPM lifecycle [vdAtHW03] depicted in Fig. 1, the phases of *system configuration*, *process enactment*, and *diagnosis* are not standardized and differ for every BPEL implementation. As a consequence, the seemingly simple task of deploying a BPEL process on a BPEL engine requires engine dependent meta data files, called deployment descriptors, to be present (phase system configuration) and engine specific deployment methods have to be called (phase process enactment). Also tasks such as monitoring and log file formats (phase diagnosis) are not standardized. This aggravates the automation of deploying BPEL processes in a cloud environment, as every user has to automate these tasks for every engine they use. This situation is typical for these types of

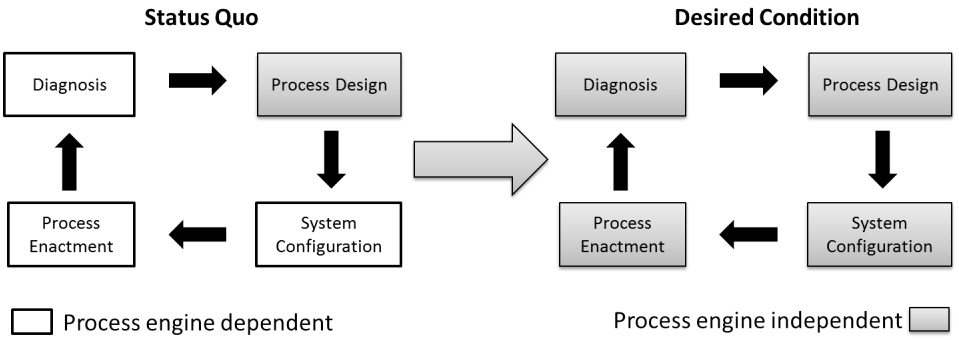


Figure 1: The status quo and the desired condition of the BPM lifecycle (adapted from [vdAtHW03])

systems and no different for the implementations of other process specifications, such as the *Business Process Model and Notation* (BPMN) [OMG11].

Since the finalization of BPEL in 2007, a plethora of open source and proprietary BPEL engines have emerged. These have a varying degree of standard conformance, as demonstrated in recent studies [HLW12, HLW13], which implies a varying degree of portability for BPEL processes in general. In other words, these studies revealed the strengths as well as weaknesses of the available BPEL engines and the need to select an engine according to the BPEL activities used in a BPEL process. For instance, some engines do not support concurrent execution of activities, which impedes the execution of processes that make use of concurrency. Because of this, a company may use multiple engines for different processes depending on their feature requirements, which in turn increases the maintenance effort, as administrators have to manage multiple engine installations at the same time. We aim to overcome these issues by providing a uniform BPEL management layer for these tasks to reduce maintenance costs and to enable the use of BPEL engines in the cloud. It can be expected that this uniform layer can be applied to engines for other process languages as well.

TOSCA [OAS13b] is an emerging OASIS standard for a holistic management of complex application topologies in cloud environments and therefore a good fit for the task at hand. It describes the application's structure using a typed graph: Node Templates represent the components of the application and relationship templates the relationships between the components. Each node template has a node type. A node type may offer interfaces to manage its lifecycle [BBLS12]. The presented BPEL management layer is a first step towards a BPEL engine node type allowing a uniform management of BPEL engines by a TOSCA runtime.

The paper is structured as follows. The next section discusses related work on the automated deployment and management of applications in cloud environments in general and of BPEL processes in particular, as well as on the assessment of BPEL portability. Section 3 presents the main contribution of the paper: A Web service-based approach to manage BPEL engines, the deployed processes and their instances using a uniform interface. Section 4 outlines the

implemented prototype. Finally, Section 5 concludes the paper and provides an outlook on future work.

2 Related Work

Related work separates in work on the automated deployment and management of BPEL processes, the assessment of their portability, the deployment of complex application topologies using TOSCA and the usage of uniform APIs in cloud computing.

The BPEL engine test system (*betsy*[1]) [HL12] consists of a test suite to evaluate the conformance of BPEL engines to the BPEL specification, as well as a tool to automatically determine the conformance degrees by running the test suite against multiple BPEL engines. Hence, betsy implements the logic to manage multiple BPEL engines as part of its test bed. This includes the (re)installation, startup, shutdown, the deployment of a process, and the retrieval of the log files of an engine. In [HRW14], Harrer et al. extended betsy with the creation of *vbetsy*, a tool that tests BPEL engines that are being provisioned and run in dedicated virtual machines in an effort to allow for a more efficient and fast testing process. As part of this extension, they extracted two interfaces: the `EngineLifecycle` interface to install, start, and stop the engine, and the `EngineActions` interface to deploy a process and retrieve the log files of the engine. The methods of the `EngineActions` interface are accessible via the network by exchanging Java objects over TCP, while the `EngineLifecycle` methods are only accessible locally. In our approach, we reuse these existing methods and integrate them into a holistic approach for uniform BPEL management. Moreover, instead of accessing these methods locally or via exchanging Java objects over TCP, we provide portable and platform-independent WSDL-based Web services for all engine capabilities. Similar to this, [vLLM+08] describes a management framework for BPEL engines. However, this framework is more of a suggestion for engine vendors how such a management framework should look like and is not aiming at the unification of proprietary management APIs. Here, we do not propose a framework to be implemented by engine vendors, but build a unified framework on top existing proprietary management APIs, without requiring these to be adapted.

In [HLW12, HLW13] betsy is used to evaluate the conformance of five open source and three proprietary engines by means of the test suite described in the previous paragraph. These results have been used in [LW13b] to compute the degree of portability of a BPEL process using software metrics and to evaluate the quality of four mappings from several modeling languages to BPEL [LW13a]. This computation is done automatically by the *bpp*[2] tool. This tool is a static analyzer that parses BPEL processes and detects portability issues in them. As part of this work, we extended and integrated the tool to support the automatic selection of a suitable BPEL engine for a specific BPEL process by rejecting the engines that do not support the BPEL features used in the process. A similar, policy-based selection approach is described in [MvLW+09]. However, it is focused on the automatic

[1] The tool is open source and publicly available at `https://github.com/uniba-dsg/betsy`.

[2] Bpp stands for *BPEL Portability Profile*. The tool is open source and publicly available at `https://github.com/uniba-dsg/bpp`.

routing of requests to the best suited service implementation and not on finding a suitable BPEL engine for a certain process.

Lego4TOSCA [HLNW14] presents a generic architecture for the implementation of TOSCA node types. Concrete node type implementations are called *implementation artifacts*, and are designed to work together to realize management functionality. For instance, the Apache Tomcat implementation artifact uses the Windows and the Ubuntu implementation artifact to execute operations on the respective virtual machine.

In cloud computing, the management interfaces of cloud vendors vary greatly despite the fact that they support a common set of management operations [KW14]. Hence, there is a trend towards uniform APIs for both the management of Infrastructure as a Service (IaaS) and Platform as a Service (PaaS) environments to provide a uniform access to these common operations. For instance, the *Open Cloud Computing Interface* (OCCI) [OCC11] provides a uniform API for IaaS, whereas *Cloud Application Management for Platforms* (CAMP) [OAS13a] proposes a uniform API for PaaS. Our approach solves a similar problem regarding BPEL engine vendors and their APIs, i.e., the provisioning, deployment and monitoring of infrastructure and services. In addition, a uniform API for BPEL engines in the form in which we propose it, can easily interface with other uniform APIs and be integrated in cloud provisioning and deployment plans, e.g., using TOSCA.

3 Approach

In this section, we present our idea of a uniform BPEL management layer, called *UBML*, for accessing all BPEL-related services. The big picture is depicted in Figure 2 which outlines the relationship of the management layer to multiple BPEL engines and a typical usage scenario in the form of a standard conformant BPEL process that should be executed on a suitable engine. The UBML comprises seven services which abstract the common callable functionality of the BPEL engines and provides an API to the user which supports multiple application scenarios. One of these scenarios, which can be seen as *typical* usage scenario is shown by the numbered edges from the standard conformant BPEL process to the different UBML services in Fig. 2. In the first step, we select the best available engine for this particular process using the *Engine Selection* service. Next, we install and start the selected engine via the *Engine Provisioning* and *Engine Lifecycle* components, respectively. The *Process Deployment* component can then be used to deploy the standard conformant process onto this running engine. This component takes care of every aspect of the deployment, including the generation of a valid and engine dependent deployment descriptor. When the process is now deployed, we can monitor its runtime execution via the log files (*Logfile Access*) of the engine, perform the recovery of activities (*Process Management*) and access the *Audit Trail*.

As the uniform BPEL management layer abstracts from all engine specific APIs, we are able to provide the services of this layer as Web services which can be integrated into Cloud-based provisioning processes. Moreover, we contribute to the effort of maintaining the portability advantages of the BPEL specification: *model once, run anywhere*.

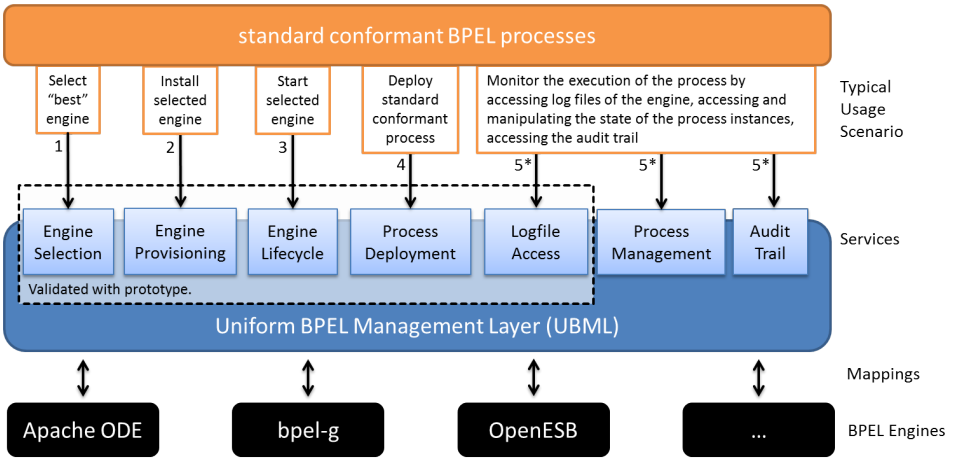


Figure 2: The Uniform BPEL Management Layer (UBML)

This section is organized according to the order in which the services are used in the typical usage scenario. At first, the *Uniform Engine Selection* service is detailed in Section 3.1 followed by the *Uniform Engine Provisioning* and the *Uniform Engine Lifecycle* in Section 3.2 and Section 3.3, respectively. Next, Section 3.4 describes the *Uniform Process Deployment* whereas the three remaining sections (3.5, 3.6 and 3.7) outline our approach on uniform monitoring, activity recovery, and troubleshooting.

3.1 Uniform Engine Selection

BPEL engines differ in terms of their standard conformance (e.g., the support for the various BPEL language elements, such as the `<validate>` activity), their functional capabilities (e.g., some engines are able to invoke REST services via proprietary extensions or provide special support for in-memory enactment to improve the performance of process execution) or their nonfunctional requirements (e.g., support for additional Web Services standards, such as WS-ReliableMessaging [OAS09] and WS-Security [OAS06]). To automatically select the engine which is providing the best service for the given BPEL process, we take two kinds of data into account: First, we analyze the language elements the process uses and match it against a standard conformance database, which contains information about which language element is supported by which engine. The language benchmarks of [HLW12, HLW13] are the primary source for this database and a static analyzer tool proposed and used in [LW13b, LW13a] performs the analysis of the process. This way, we can determine which engines are certain to be unable of supporting the process to be deployed and can restrain from using these engines in the following steps. Moreover, we can rank engines according to the amount of standard conformance they provide. Second, we use policy attachments to describe functional or nonfunctional requirements in an abstract manner. That way, for each engine, its requirements and capabilities are captured using

a policy description language, such as WS-Policy. Furthermore the process models to be deployed can be enriched with policies describing their requirements on a suitable engine. Such policies may define alternative sets of requirements of which at least one must be supported and may be embedded directly into the process via extensions or be attached to it as a separate file.

Based on the policy descriptions, the deployment component will find and select the best matching engine (applying a configurable ranking algorithm in case multiple engines match the requirement) and will compute an effective policy (i.e., the set of common requirements and capabilities). This effective policy can then be used to generate the vendor specific deployment descriptor to configure a certain behavior.

3.2 Uniform Engine Provisioning

After a suitable BPEL engine has been selected the engine needs to be provisioned on an appropriate hardware and operating system. Some engines have preferences on which hardware or operating system they run in a more optimized manner, which has to be taken into account for the engine provisioning process. Moreover, at installation time, configuration options must be possible to set, e.g., the port on which the engine is accessible. Furthermore, as the engines may require additional software, e.g., databases, container, etc., these have to be taken into account as well. In summary, the *Engine Provisioning* service takes care of selecting the appropriate server and operating system and is able to install or uninstall the engine and its dependent components. To achieve this, it needs to know which steps are needed for a successful (un)installation for each engine and will expose this functionality through a generic interface.

3.3 Uniform Engine Lifecycle

When BPEL engines are provisioned automatically, it is important to also manage their lifecycle in a uniform fashion. The *Uniform Engine Lifecycle* service exposes a generic interface, which allows for starting and stopping all known engines. That way, the UBML can automatically manage the resource consumptions of the managed engines and smartly decide whether an engine can be stopped (i.e., only if there are no processes deployed anymore).

3.4 Uniform Process Deployment

The deployment of a BPEL process to an engine is nontrivial, as it requires the creation of an engine specific deployment descriptor and the usage of an engine specific deployment method. As shown in [HL12, LHW13], there are major differences in the complexity of the deployment descriptors. There are engines which do not require additional descriptors,

but also others that require three per process. Regarding the deployment methods, there are multiple strategies as well, for instance deployment through a command line interface, through a Web service or manually through the browser. We aim to overcome these differences by a) generating deployment descriptors automatically based on the information already present in the BPEL and WSDL files, and b) by providing a uniform Web service for deploying processes on a total of seven engines in different configurations and versions.

3.5 Uniform Logfile Access

The location of the log files depends heavily on the engine and environment. Although most engines have a single log file, there typically also are multiple logs for the environment or container in which the engine is running. As all these files may be scattered into different places for each engine, there is a need to retrieve the files for troubleshooting and maintenance in a uniform way. We capture these files using an engine independent Web service that collects all files and returns them as a collection. That way, the IT support staff has a single point of access for all troubleshooting information.

3.6 Uniform Process Management

Most BPEL engines provide sophisticated management APIs. These interfaces enable applications to suspend/resume, debug or recover process instances. In addition, some of these engines allow for manipulating instance data (i.e., variables and partner links) or to rewind a process to a specific point in the past and to resume it from this position. Our management component aims at unifying these APIs in order to provide a single API that is capable to monitor all supported engines.

3.7 Uniform Audit Trail

Audit trails are an important asset when analyzing the execution of processes for a purpose such as process mining. Several tools such as ProM[3] or fluxicon's Disco[4] support that task but in some cases the only requirement is to display which parts of the process have been properly executed already. Most engines are able to provide a detailed audit trail but unfortunately there is no commonly implemented standard for their format. When abstracting from the concrete engine that is actually executing the given process, there is also the need to unify the audit trail. In order to achieve this, our monitoring component can translate engine specific audit data to a generic audit trail data format such as MXML [Don05].

[3] For more information, visit the project page at `http://www.promtools.org/prom6/`

[4] For more information, visit the project page at `http://www.fluxicon.com/disco/`

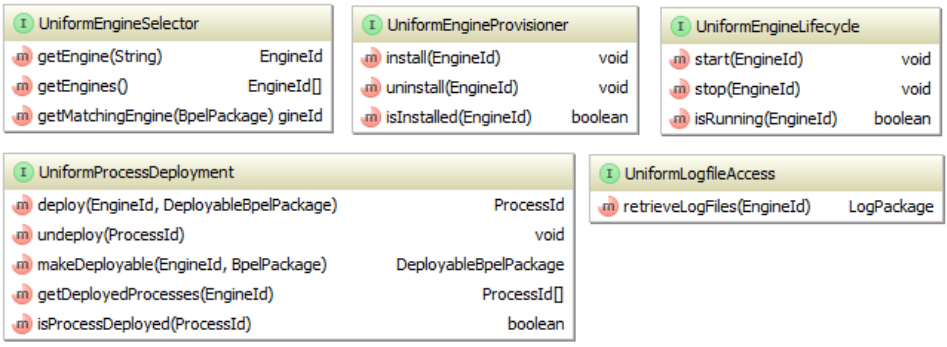


Figure 3: Class Diagram of the Uniform BPEL Management Layer implemented by the prototype.

4 Prototype

The prototype discussed in this section validates the feasibility of our proposed approach from Section 3. As shown in Figure 2, our tool implements the five UBML services: *Engine Selection*, *Engine Provisioning*, *Engine Lifecycle*, *Process Deployment* and *Logfile Access*. For each of these services, the prototype provides a callable Web service that maps the uniformly defined actions to engine specific calls. At the moment, UBML supports seven different open source BPEL engines in different configurations (e.g., in-memory or persistent execution) and versions. In particular, we support Apache ODE in versions 1.3.5 and 1.3.6 with the in-memory option, bpel-g in version 5.3 with the in-memory option, ActiveBPEL v5.0.2, Orchestra v4.9, OpenESB in versions 2.2, 2.3 and 2.3.1, Petals ESB in versions 4.0 and 4.1 and WSO2 Business Process Server versions 2.1.2, 3.0.0, and 3.1.0. The interfaces of the five services can be seen in the class diagram in Figure 3. The services are published as WSDL 1.1 Web services, and group low-level methods according to their task. Engines are identified with the string-based `EngineId` while deployed BPEL processes are referred to using the QName[5]-based `ProcessId` which also contains an `EngineId`. The standard conformant BPEL process and its related files (e.g. WSDL, XSD or other files) are sent to the interfaces as a `BpelPackage` without any engine specific information while the `DeployableBpelPackage` is an archive that includes the relevant files and structure to be deployable on a specific engine. Log files are transferred similarly in the form of packages using the `LogPackage` class.

The *Engine Selector* service is implemented as a registry or repository. It either provides a list of all available engines or the best engine for a specific BPEL process. For determining the latter, we use an extended version of the *bpp* tool to which we transfer the path to the BPEL file within the `BpelPackage`. Using the conformance data of the engines, the tool returns a list of engines that are able to support the process. Moreover, it also provides the number of supported features per engine, thus, providing the data to create a ranking of the engines that support the process by their overall support of the specification from which the topmost is selected. The other four services are implemented by reusing and extend-

[5]QName stands for qualified name, i.e., an identifier with a namespace [W3C09].

ing the engine specific logic of *betsy* [HL12]. The `UniformEngineProvisioning` component allows installing and uninstalling any supported engine as well as checking whether it is installed or not. To start, stop, or check the status of an engine, our tool provides corresponding methods in the `UniformEngineLifecycle` service. The largest service is the `UniformProcessDeployment` service, which supports the deployment and undeployment of a standard conformant BPEL process. The deployment, however, requires a preprocessing step as the `BpelPackage` has to be made deployable via the `makeDeployable` method which returns the desired `DeployableBpelPackage` that is required by the `deploy` method. Whether a process is already deployed can be checked with the method `isProcessDeployed` while a list of all deployed processes of a single engine can be retrieved via `getDeployedProcesses`. The smallest component is the `UniformLogfileAccess` which simply returns a package with all the log files for a specific engine.

Listing 1: Composition of UBML Services

```
1  class CompositeProcessProvisioningService  {
2      ProcessId makeProcessAvailable(BpelPackage bpelPackage) {
3          EngineId engineId = UniformEngineSelection.getMachingEngine(
               bpelPackage);
4          if(!UniformEngineProvisioning.isInstalled(engineId)) {
5              UniformEngineProvisioning.install(engineId);
6          }
7          if(!UniformEngineLifecycle.isRunning(engineId)) {
8              UniformEngineLifecycle.start(engineId);
9          }
10         DeployableBpelPackage deployableBpelPackage =
               UniformProcessDeployment.makeDeployable(engineId,
               bpelPackage)
11         return UniformProcessDeployment.deploy(engineId,
               deployableBpelPackage );
12     }
13     void makeProcessUnavailable(ProcessId processId) {
14         UniformProcessDeployment.undeploy(processId);
15         EngineId engineId = processId.getEngineId();
16         ProcessId[] processIds = UniformProcessDeployment.
               getDeployedProcesses(engineId);
17         if(processIds.length == 0) {
18             UniformEngineLifecycle.stop(engineId);
19             UniformEngineProvisioning.uninstall(engineId);
20         }
21     }
22 }
```

Based the lower-level operations depicted in Figure 3, we can build the functionality outlined in the typical usage scenario from Figure 2 and shown in Java-based pseudo code in Listing 1 as part of the `CompositeProcessProvisioningService` class. In the method `makeProcessAvailable`, we only need to pass in a `BpelPackage` and get the `ProcessId` of the, then running, process back. This operation orchestrates the *Engine Selector* component to select an engine that best supports the BPEL features of the provided process, the *Engine Provisioning* service to install the selected engine if it is not already installed, the *Engine Lifecycle* service to start the selected engine if it is not already started, and the *Process Deployment* component to create the `DeployableBpelPackage` and

Figure 4: The EngineControl GUI for controlling engine related tasks.

deploy it. In contrast, the method `makeProcessUnavailable` implements the reverse logic which allows to undeploy a process using the *Process Deployment* service and freeing resources. If the engine which previously undeployed the process has no other processes deployed, we can easily terminate it via the *Engine Lifecycle* service (i.e., freeing RAM and CPU power) and uninstall it via the *Engine Provisioning* service (i.e., freeing disk space). Furthermore, multiple other additional services and operations can be built, e.g., a dashboard showing the list of all currently installed and started engines and the deployed and running processes.

Our prototype is open source, publicly available[6] and written in Java 8 and Groovy 2.3. The Web services are exposed using the JAX-WS[7] 2.0 Java API and can be started using the `ubml.ws.WSMain` class, which sets up the Web services using the following pattern `http://localhost:1234/INTERFACE_NAME`. These addresses can be integrated into TOSCA plans to enable the uniform management of BPEL-related tasks. Using the Gradle[8] task `run`, the Web services start automatically. By executing `gradlew run`, Java 8 is the only runtime dependency, everything else is installed automatically on demand. This enables an easy deployment of this tool on any virtual machine to set up BPEL engines in the cloud.

Furthermore, the tool also contains a simple GUI application which can be started via the `gradlew enginecontrol` command on the command line. This GUI can control the operations of the three services *UniformEngineProvisioning*, *UniformEngineLifecycle*, and *UniformLogfileAccess* as shown in Figure 4.

Several limitations still remain for this prototype. First, only a single engine can currently be run on the same machine, because the engines run on their default ports, which can create conflicts. In the future, we aim to provide configuration options during provisioning to fix this. Second, the automatic creation of the deployable BPEL package including the deployment descriptors does only work for BPEL processes from the betsy test suite. We want to extend this feature to arbitrary BPEL processes in the future, however, this is

[6] See the project page for more information: `https://github.com/uniba-dsg/ubml`

[7] JAX-WS 2.0 is described in the Java Specification Request 224 which is available at `https://jcp.org/en/jsr/detail?id=224`.

[8] Gradle is a build management tool with which the UBML prototype is built. It is available at `http://www.gradle.org/`.

considered quite time consuming as this requires parsing much information from the BPEL and WSDL files. Third, both the `undeploy` and the `isDeployed` commands are only implemented for Apache ODE. We aim to implement the mapping for the other remaining six open source engines as well. Fourth, the prototype only runs on Windows 7 64bit. As many virtual machines use a Linux-based operating system, we plan to port the mappings to run on Linux as well.

5 Conclusion and Future Work

This paper presents a Web Services-based approach to manage BPEL engines and deployed processes and we provide a prototypic implementation of the approach. The current limitation is that the logic to manage the BPEL engines has to be installed manually on the virtual machine where the engine should be deployed on. Our plan is to leverage virtualization, i.e., port the prototype to the Lego4TOSCA universe. This allows us to provision arbitrary virtual machines without any user intervention. Moreover, as this approach only supports open source BPEL engines, we aim to support proprietary BPEL engines as well. Another subject of future work is the construction of a unified monitoring API across all vendors. Finally, the presented work focuses on BPEL engines and our next step is to apply the presented concepts on process engines for BPMN [OMG11].

Acknowledgment We would like to express our gratitude to Oliver Kopp for fruitful discussions on this topic.

References

[BBLS12] Tobias Binz, Gerd Breiter, Frank Leymann, and Thomas Spatzier. Portable Cloud Services Using TOSCA. *IEEE Internet Computing*, 16(03):80–85, May 2012.

[Don05] B. F. Van Dongen. A Meta Model for Process Mining Data. In *In Proceedings of the CAiSE WORKSHOPS*, pages 309–320, Porto, Portugal, 2005.

[HL12] Simon Harrer and Jörg Lenhard. Betsy–A BPEL Engine Test System. Technical Report 90, Otto-Friedrich Universität Bamberg, July 2012.

[HLNW14] Florian Haupt, Frank Leymann, Alexander Nowak, and Sebastian Wagner. Lego4TOSCA: Composable Building Blocks for Cloud Applications. In *Proceedings of the 7th IEEE International Conference on Cloud Computing (CLOUD 2014)*, Anchorage, Alaska, USA, 2014. IEEE.

[HLW12] Simon Harrer, Jörg Lenhard, and Guido Wirtz. BPEL Conformance in Open Source Engines. In *Proceedings of the 5th IEEE International Conference on Service-Oriented Computing and Applications (SOCA'12), Taipei, Taiwan*, pages 1–8. IEEE, December 17-19 2012.

[HLW13] Simon Harrer, Jörg Lenhard, and Guido Wirtz. Open Source versus Proprietary Software in Service-Orientation: The Case of BPEL Engines. In *ICSOC*, volume 8274

of *Lecture Notes in Computer Science*, pages 99–113, Berlin, Germany, 2013. Springer Berlin Heidelberg.

[HRW14] Simon Harrer, Cedric Röck, and Guido Wirtz. Automated and Isolated Tests for Complex Middleware Products: The Case of BPEL Engines. In *Software Testing, Verification and Validation Workshops (ICSTW), 2014 IEEE Seventh International Conference on*, Cleveland, Ohio, USA, April 2014. Testing Tools Track.

[KKL06] Rania Khalaf, Alexander Keller, and Frank Leymann. Business processes for Web Services: Principles and applications. *IBM Systems Journal*, 45(2):425–446, 2006.

[KW14] Stefan Kolb and Guido Wirtz. Towards Application Portability in Platform as a Service. In *Proceedings of the 8th IEEE International Symposium on Service-Oriented System Engineering (SOSE)*, Oxford, United Kingdom, April 7–10 2014. IEEE.

[LHW13] Jörg Lenhard, Simon Harrer, and Guido Wirtz. Measuring the Installability of Service Orchestrations Using the SQuaRE Method. In *Proceedings of the 6th IEEE International Conference on Service-Oriented Computing and Applications (SOCA'13)*, Kauai, Hawaii, USA, December 16-18 2013. IEEE.

[LW13a] Jörg Lenhard and Guido Wirtz. Detecting Portability Issues in Model-Driven BPEL Mappings. In *SEKE*, Boston, Massachusetts, USA, June 2013.

[LW13b] Jörg Lenhard and Guido Wirtz. Measuring the Portability of Executable Service-Oriented Processes. In *Proceedings of the 17th IEEE International EDOC Conference*, Vancouver, Canada, September 2013.

[MvLW+09] Ralph Mietzner, Tammo van Lessen, Alexander Wiese, Matthias Wieland, Dimka Karastoyanova, and Frank Leymann. Virtualizing Services and Resources with ProBus: The WS-Policy-Aware Service and Resource Bus. In *Proceedings of the 7th International Conference on Web Services (ICWS) 2009*, Los Angeles, CA, USA, July 2009. IEEE Computer Society.

[OAS06] OASIS. *Web Services Security*, February 2006. v1.1.

[OAS07] OASIS. *Web Services Business Process Execution Language*, April 2007. v2.0.

[OAS09] OASIS. *Web Services Reliable Messaging*, February 2009. v1.2.

[OAS13a] OASIS. *Cloud Application Management for Platforms*, July 2013. Version 1.1 – Draft 03.

[OAS13b] OASIS. *Topology and Orchestration Specification for Cloud Applications*, November 2013. v1.0.

[OCC11] OCCI. *Open Cloud Computing Interface - Core*. Open Grid Forum, 2011.

[OMG11] OMG. *Business Process Model and Notation*, January 2011. v2.0.

[vdAtHW03] Wil M. P. van der Aalst, Arthur H.M. ter Hofstede, and Mathias Weske. Business Process Management: A Survey. In *Proceedings of the International Conference on Business Process Management*, Eindhoven, The Netherlands, 2003. Springer Berlin Heidelberg.

[vLLM+08] Tammo van Lessen, Frank Leymann, Ralph Mietzner, Jörg Nitzsche, and Daniel Schleicher. A Management Framework for WS-BPEL. In *Proceedings of the 6th IEEE European Conference on Web Services 2008*, pages 187–196, Dublin, Ireland, November 2008. IEEE Computer Society.

[W3C09] W3C. *Namespaces in XML*, December 2009. v1.0.

Realisierung von vertrauenswürdigem Cloud-Computing mit SkIDentity: Der Anwendungsfall in Wertschöpfungsketten der Automobilindustrie

Guntram Flach, Fraunhofer IGD
Joachim-Jungius-Str. 11, 18059 Rostock
guntram.flach@igd-r.fraunhofer.de

Michael Kubach, Eray Oezmue, Fraunhofer IAO
Nobelstr. 12, 70569 Stuttgart
Eray.Oezmue@iao.fraunhofer.de

Immo Wehrenberg, ENX Association
Bockenheimer Landstraße 97-99, 60325 Frankfurt am Main
immo.wehrenberg@enx.com

Abstract: Ziel des SkIDentity-Projektes ist es, eine tragfähige Brücke zwischen den sicheren elektronischen Ausweisen (eID) und den heute existierenden bzw. sich entwickelnden Cloud-Computing-Infrastrukturen zu schlagen. Somit können vertrauenswürdige Identitäten für die Cloud bereitgestellt und komplette Prozess- und Wertschöpfungsketten sicher gestaltet werden. Hierfür werden existierende Komponenten, Dienste und Vertrauensinfrastrukturen zu einer umfassenden, rechtskonformen, wirtschaftlich sinnvollen und sicheren Identitätsinfrastruktur für die Cloud integriert und in breitenwirksamen Pilotprojekten erprobt. Eine solche Pilotapplikation ist ein Projektworkspace, der auf die Bedürfnisse der verteilten Wertschöpfungsketten der Automobilindustrie abgestimmt ist. Die SkIDentity-Technologie ermöglicht über einen eID-Broker den verschiedenen an einem Projekt beteiligten Entwicklern eine starke Authentisierung am Projektworkspace mit den unterschiedlichen bereits in ihrem Unternehmen vorhandenen Credentials im Sinne eines föderierten Identitätsmanagements. Diese Arbeit stellt die SkIDentity-Technologie vor und illustriert deren Vorteile anhand der Pilotapplikation.

1 Einleitung

Für ein vertrauenswürdiges Cloud Computing werden sichere und zuverlässige Mechanismen zur Authentisierung benötigt. Gleichzeitig ist eine möglichst nutzerfreundliche Gestaltung der Bedienoberfläche und des zugehörigen Prozesses notwendig, um eine hohe Nutzerakzeptanz zu erreichen (vgl. [Senk12]). Nur eine Lösung, die gute Benutzbarkeit mit hoher Sicherheit verbindet, wird die Authentisierung an Cloud Computing Systemen nachhaltig sicherer machen können.

Das Ziel des SkIDentity-Projektes[1] ist es daher, eine Architektur zu schaffen, die von vornherein darauf ausgelegt ist, flexibel mit Lösungen verschiedenster Anbieter zusammen zu arbeiten. Diese Anbieter können dabei einen oder mehrere Authentisierungsdienste anbieten oder auch Identitätsdaten und Authentifikation für die von ihnen betriebenen Anwendungen nutzen; beliebige Kombinationen sind möglich. Dabei werden sowohl die technischen und organisatorischen Aspekte, als auch die rechtlichen Voraussetzungen berücksichtigt.

Im Ergebnis sollen die Nutzer ihren bevorzugten „Ausweis“ (Credential) für die starke Authentisierung bei verschiedenen Anwendungen nutzen können. Dies verringert die Zahl der vom Benutzer zu merkenden oder zu verwaltenden Zugangsinformationen oder Credentials und erspart ihm das wiederkehrende Vertraut machen mit neuen Authentisierungsmechanismen. Diese Vereinfachung erhöht gleichzeitig die Sicherheit, da die ablaufenden Schritte und Abfolgen für den Benutzer nachvollziehbarer, verständlicher und wiedererkennbar werden.

2 Anwendungsszenario

Die Automobilindustrie ist unter anderem dadurch gekennzeichnet, dass die Automobilhersteller (OEM) wesentliche Teile der Wertschöpfung nicht selbst erledigen, sondern auf die verlängerte Werkbank ihrer Zulieferer auslagern. Diese Zulieferer entwickeln und stellen für die OEMs nicht lediglich einfache Teile wie Schrauben her, sondern tragen zum Endprodukt durch die Entwicklung und Produktion komplexer Module, wie etwa kompletter Sitze, entsprechend der Vorgaben des OEMs, bei. Diese Tier1-Zulieferer arbeiten ihrerseits wiederum mit Tier2-Zulieferern zusammen, von welchen sie Komponenten beziehen. Insofern wird von einer verteilten Wertschöpfungskette gesprochen.

Der hohe Wettbewerbsdruck in der Automobilindustrie macht eine Senkung der Time-to-Market unumgänglich. Dies wird unter anderem durch eine interaktive Zusammenarbeit der Ingenieure von OEMs und Zulieferern in Multibenutzerapplikationen erreicht. Aufgrund der komplexen netzwerkartigen Wettbewerbsstruktur der Automobilindustrie arbeiten jedoch sowohl OEMs als auch Tier1- und Tier2-Zulieferer für unterschiedliche Komponenten zur gleichen Zeit mit zahlreichen jeweils konkurrierenden Partnern zusammen. Aus diesem Grund ist eine wirksame Zugriffskontrolle für den Austausch und die gemeinsame Bearbeitung technischer Spezifikationen und Komponentendesigns unbedingt erforderlich (Siehe Abbildung 1).

---

[1] www.skidentity.de; das SkIDentity Projekt zählt zu den Gewinnern des „Trusted Cloud“ Technologiewettbewerbs (www.trusted-cloud.de) des Bundesministerium für Wirtschaft und Technologie (BMWi) und zielt auf die Bereitstellung vertrauenswürdiger elektronischer Identitäten für Cloud Computing Dienste ab. Es wird von einem interdisziplinären Expertenteam unter Leitung der ecsec GmbH und mit Beteiligung der ENX Association, der Fraunhofer Institute IAO und IGD, der OpenLimit SignCubes GmbH, der Ruhr Universität Bochum, der Universität Passau, der Urospace GmbH und der VDG Versicherungswirtschaftlicher Datendienst GmbH durchgeführt. Darüber hinaus wird das SkIDentity Projekt von maßgeblichen Verbänden, wie dem Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM), dem EuroCloud Deutschland_eco e.V., dem ProSTEP iViP e.V. und dem TeleTrusT – Bundesverband IT-Sicherheit e.V. sowie renommierten Unternehmen wie der DATEV eG, der easy Login GmbH, der media transfer AG, der SAP AG und der SiXFORM GmbH unterstützt.

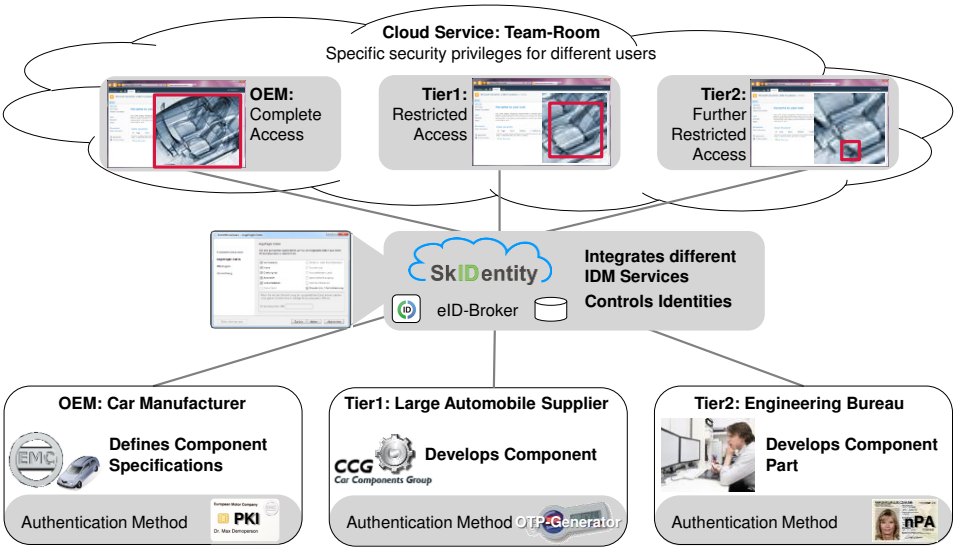


Abbildung 1: Rahmen für das Experimentalsystem

Diese Zugriffskontrolle stellt bereits innerhalb eines Unternehmens eine Herausforderung für das Identitätsmanagement (IdM) dar. Sind jedoch mehrere unabhängige Unternehmen (OEMs, Zulieferer und gegebenenfalls Service/Cloud-Provider) beteiligt, nimmt die Komplexität der Aufgabe einer vertrauenswürdigen Identifizierung aller beteiligten Ingenieure im Sinne eines föderierten Identitätsmanagements deutlich zu. Beispielsweise verwenden die unterschiedlichen Unternehmen jeweils eigene Authentifizierungsverfahren und Sicherheitspolicies, die vom Service/Cloud-Provider zu implementieren sind. Außerdem werden Ingenieure durch die Unternehmen ihrem Bedarf entsprechend dem Projekt zugeordnet oder abgezogen, weshalb die Zugriffsberechtigungen immer aktuell zu halten. Die ökonomische und technische Leistungsfähigkeit insbesondere kleinerer Akteure ohne größere IT-Abteilung stößt somit aufgrund der Vielzahl zu erfüllender Anforderungen schnell an Grenzen.

Der Pilotapplikation stellt in diesem Setting nun einen gemeinsamen Workspace als Cloud-Anwendung zur Verfügung, auf welchen unterschiedliche Akteure mit unterschiedlichen Berechtigungen zugreifen können. Die Authentifizierung erfolgt über den eID-Broker mittels der bereits in den Organisationen vorhandenen Credentials. Sie wurde in die Architektur von SkIDentity integriert und derart entwickelt, dass sie als Experimentalsystem für weitere Forschungs- und Entwicklungsarbeiten sowie für öffentliche Präsentationen genutzt werden kann.

Ausgehend vom geschilderten Anwendungsszenario und den genannten Anforderungen wird im nächsten Abschnitt die SkIDentity-Architektur kurz vorgestellt, bevor darauf folgend einige Realisierungsaspekte genauer präsentiert werden.

3 Systemarchitektur

Im Rahmen des Projektes wurde die in Abbildung 3 dargestellte Systemarchitektur entwickelt. Das SkIDentity-System baut auf dem Konzept des „Föderierten Identitätsmanagements“ (vgl. [HRZ10], [SAM05], [Open07] und Abbildung 2) auf und verfeinert die Komponenten „Client“, „Service Provider“ und „Identity Provider“, so dass unterschiedlichste Authentisierungsmechanismen, -dienste und -protokolle in einer einheitlichen und sicheren Weise in beliebigen Anwendungen genutzt werden können.

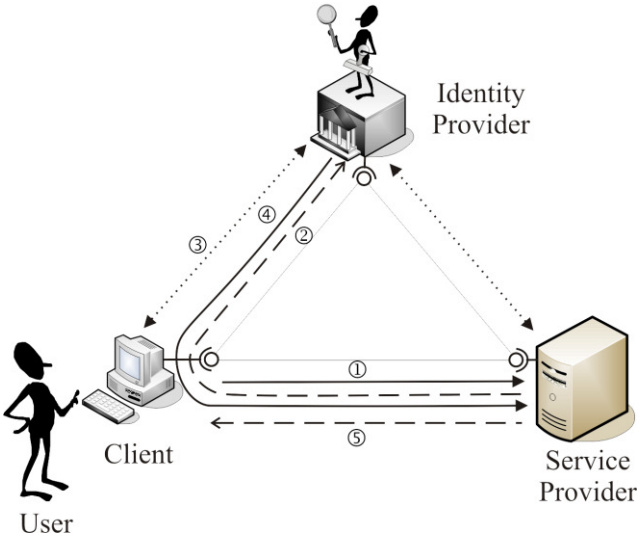


Abbildung 2: Föderiertes Identitätsmanagement

Hierdurch können insbesondere auch die zunehmend verbreiteten eID-Karten, die von staatlicher Seite ausgestellt werden, in der SkIDentity-Infrastruktur genutzt werden. Hierzu zählt insbesondere der neue Personalausweis (nPA) in Deutschland oder vergleichbare Dokumente anderer Staaten. Diese Karten werden wegen des behördlichen Vertrauensankers als sicher wahrgenommen und erreichen nach vollständiger Einführung eine hohe Verbreitung. Beim Entwurf der Systemarchitektur wurde deshalb insbesondere auf Konformität mit den Vorgaben des deutschen Personalausweisgesetzes geachtet [HoHH12].

Wie in Abb. 3 ersichtlich, umfasst die SkIDentity-Referenzarchitektur für die starke Authentisierung in der Cloud Systemkomponenten beim Benutzer, Systemkomponenten beim Dienstanbieter, sowie entsprechende Infrastrukturkomponenten.

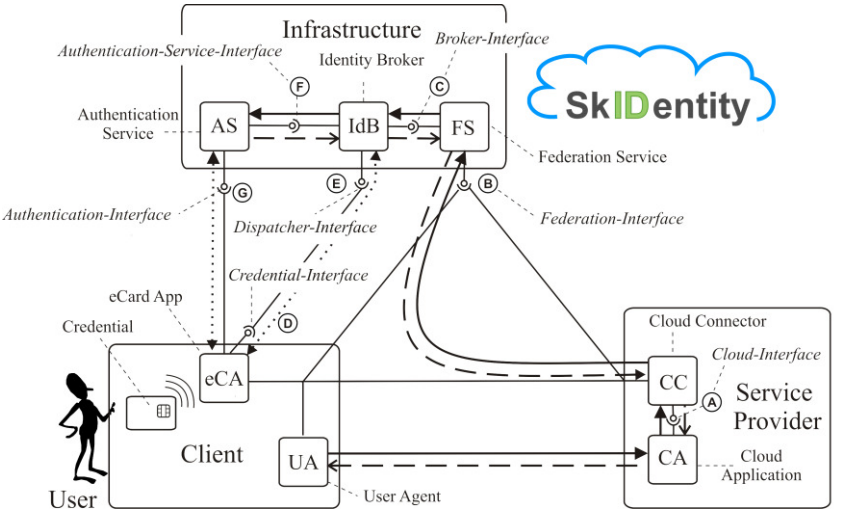


Abbildung 3: SkIDentity-Referenzarchitektur

4 Architekturkomponenten

Die für den Lösungsansatz notwendigen Überlegungen werden im folgenden Abschnitt durch eine Auswahl verschiedener Realisierungsaspekte kurz vorgestellt.

Identity Broker (IdB)

Der IdB ist die zentrale Komponente in der SkIDentity-Infrastruktur, die bei einer vom CC bzw. FS übermittelten Authentisierungsanfrage zunächst ermittelt, welche Credentials beim Benutzer vorhanden sind, so dass ein geeigneter Dienst für die Benutzerauthentisierung ausgewählt werden kann.

Federation Services (FS)

Hierbei handelt es sich um optionale Dienste, die entsprechende Föderationsprotokolle etwa gemäß [SAM05], [Open07] oder [Ham10] unterstützen und bei entsprechend ausgestalteter Authentisierungspolitik ein Single Sign-On ermöglichen. Hierdurch muss sich ein Benutzer nur einmal authentisieren, um (über einen gewissen Zeitraum) eine Vielzahl von Anwendungen nutzen zu können.

Cloud Connector (CC)

Über den CC wird die CA in die SkIDentity-Infrastruktur integriert (Siehe Abbildung 3). Hierbei wird mit dem Federation Service (FS) oder direkt mit dem IdB kommuniziert. Der CC wird bei der CA betrieben. Er erweitert die CA um die Fähigkeit, Identitäten und Authentifikation mit Hilfe eines vom IdB oder von einem FS unterstützten Protokolls zu nutzen. CAs die eine solche Funktionalität von Hause aus bieten, haben einen CC bereits integriert. Sollte eine solche Funktionalität nicht verfügbar sein, kann der CC entweder direkt in die CA integriert werden (etwa durch Nutzung einer Programmbibliothek) oder

als eigenständiger Prozess die Daten von IdB bzw. FS entgegennehmen und der Applikation in einer für sie geeigneten Form weiterleiten.

Authentication Services (AS)

Der AS wird vom IdB aufgerufen, um die tatsächliche Benutzerauthentisierung durchzuführen. Das hierfür einzusetzende Authentisierungsprotokoll hängt vom Credential ab. Beispielsweise wird für die Authentisierung mit dem nPA das Extended Access Control (EAC) Protokoll gemäß [BSI12] sowie eine ENX-spezifische Firmen-Karte genutzt.

5 Demonstrator

Das Automotive-Szenario wurde bereits in Kapitel 2 ausführlich erläutert. Ein beziehungsweise mehrere Unternehmen in einem Hersteller-Lieferantennetzwerk in der Automobilindustrie entscheiden sich für die Nutzung der SkIDentity-Infrastruktur zur sicheren Authentisierung in von ihnen gemeinsam genutzten Cloud Services. Beispielhaft hierfür steht die OwnCloud-Plattform im realisierten Demonstrator[2]. Die Anmeldung via SkIDentity kann über einen in den jeweiligen Service eingebundenen SkIDentity-Button erfolgen. Der Benutzer wird dann von SkIDentity an einen verfügbaren IdP verwiesen und meldet sich dort mit einem verfügbaren sowie erlaubten Credential an. Ist die Authentisierung erfolgreich, erfolgt die Anmeldung im Cloud Service und der Benutzer kann diesen wie vorgesehen nutzen.
Neben der Evaluierung des derzeitigen Demonstrator-Ansatzes wird in zukünftigen Arbeiten die Auswertung der qualitativen Marktanalyse für ein föderiertes Identitätsmanagement durchgeführt. Die Ergebnisse der Analyse fließen in die gesamte SkIDentity-Architektur sowie insbesondere in das Design der Pilotapplikation und der im nächsten Schritt zu entwerfenden Business Cases ein.

Literaturverzeichnis

[BSI12] BSI: Advanced Security Mechanism for Machine Readable Travel Documents – Extended Access Control (EAC), Password Authenticated Connection Establishment (PACE), and Restricted Identification (RI), Technical Directive (BSI-TR-03110, 2012

[Ham10] Hammer-Lahav, E.: The OAuth 1.0 Protocol, Request For Comments – RFC 5849, http://www.ietf.org/rfc/rfc5849.txt, 2010

[HoHH12] Hornung, G., Horsch, M., Hühnlein, D.: Mobile Authentisierung und Signatur mit dem neuen Personalausweis, DuD, 36(3):189-194, 2012

[HRZ10] Hühnlein D., Roßnagel, H., Zibuschka, J.: Diffusion of Federated Identity Management, in F. Freiling (Hrsg.), Tagungsband „Sicherheit 2010“, GI-Edition Lecture Notes in Informatics (LNI) 170, 2010, Seiten 25-37, http://www.ecsec.de/pub/Sicherheit2010.pdf

[Open07] OpenID Foundation: OpenID Authentication 2.0. Final, December 5, 2007. http://openid.net/specs/openid-authentication-2_0.html

[SAM05] S. Cantor, J. Kemp, R. Philpott, E. Maler: Assertions and Protocol for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard, 15.03.2005, http://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf , 2005

[Senk12] Senk, C.: Future of Cloud-based Services for Multi-factor Authentication: Results of a Delphi Study, Proceedings of 3rd International Conference on Cloud Computing, CLOUDCOMP, 2012

[2] https://www.skidentity.de/de/demo/

Policy-Framework – Eine Methode zur Umsetzung von Sicherheits-Policies im Cloud-Computing

Alexander Blehm, Volha Kalach, Alexander Kicherer, Gustav Murawski,
Tim Waizenegger, Matthias Wieland

Institut für Architektur von Anwendungssystemen (IAAS),
Institut für Parallele und Verteilte Systeme (IPVS)
Universitätsstraße 38
70569 Stuttgart
{blehm.ar, olga.kalach}@gmail.com
{Matthias.Wieland, Tim.Waizenegger}@ipvs.uni-stuttgart.de

Abstract: Cloud-Computing gewinnt immer mehr an Bedeutung bei der kosteneffizienten und skalierbaren Bereitstellung von IT-Diensten. Damit sich Cloud-Computing jedoch durchsetzen kann, muss die Sicherheit und Compliance der Dienste garantiert werden, d. h. die Einhaltung von Gesetzen, Richtlinien und Datenschutzvorgaben. Um diese Ziele zu erreichen, wird in diesem Beitrag ein Policy-Framework vorgestellt, welches die Umsetzung von Sicherheits-Policies im Cloud-Computing ermöglicht. Eine Policy beschreibt dabei nicht-funktionale Anforderungen an Cloud-Dienste. Des weiteren werden verschiedene prototypisch umgesetzte Policies beschieben und es wird ein Moodle-System als Anwendungsbeispiel für einen Cloud-Dienst mit den vorhandenen Policies annotiert. Dadurch erfolgt eine Evaluation des Policy-Frameworks.

1 Einleitung

Unter Cloud-Computing versteht man das Bereitstellen von IT-Infrastrukturen über ein Netzwerk, wie beispielsweise das Internet. In diesem Beitrag werden nur anwendungsspezifische Infrastrukturen betrachtet, die Cloud-Dienste genannt werden. Bekannte Beispiele sind Google Drive oder OwnCloud. Die Architektur eines Cloud-Dienstes kann durch eine Topologie modelliert werden. Darunter verstehen wir in diesem Beitrag einen hierarchischen Bauplan für die Anwendung und alle anderen Komponenten des Cloud-Dienstes. Um diesen Bauplan zu beschreiben, gibt es die standardisierte Sprache Topology and Orchestration Specification for Cloud Applications (TOSCA) [Kni13]. Cloud-Dienste, die in TOSCA beschrieben sind, haben den entscheidenden Vorteil, dass sie automatisiert bereitgestellt werden können. Cloud-Computing trifft aber häufig auf Kritik bezüglich der Sicherheit. Sämtliche Daten liegen nicht mehr im Einflussbereich des Endnutzers, sondern im Bereich des Betreibers der Cloud-Infrastruktur. Die Umsetzung von effektiven Sicherheitsstandards bringt hohe Kosten mit sich. Aufgrund von mangelnden und unausgereiften Alternativen wird auf die bestehenden Sicherheitslösungen der Technologie vertraut, ohne ihre Funktionalitäten und Limitationen ausreichend zu kennen [MSM08].

OpenTOSCA ist eine Open-Source-Implementierung einer TOSCA-Laufzeitumgebung [BBH+13]. Sie wird verwendet, um Cloud-Dienste automatisiert bereitzustellen. Dies wird auch „Deployment“ genannt. Um während des Deployments Sicherheitsanforderungen umzusetzen, wurde das Policy-Framework entwickelt. Dieses erlaubt es, Policies umzusetzen und beim Design des Cloud-Diensts zu nutzen. Policies definieren Richtlinien, die nicht-funktionale Anforderungen, die an Cloud-Dienste stellen. In diesem Beitrag werden vor allem Sicherheits-Policies betrachtet. Eine Policy kann an einen Cloud-Dienst annotiert werden, was ausdrückt, dass die darin definierten Anforderungen für diesen Dienst gelten sollen. Die Syntax und die Struktur einer Policy werden durch den TOSCA-Standard definiert [Kni13]. Dieser definiert dazu sogenannte Policy-Types und Policy-Templates. Die so definierten Policies können in der OpenTOSCA Umgebung bereitgestellt, und von allen Cloud-Diensten derselben Umgebung genutzt werden. Ein Vorreiter der hier beschriebenen Policies wurde schon vom Bundesamt für Informationstechnik unter dem Namen „IT-Sicherheitsleitlinie“ definiert [fSidI]. Mit den Instituten IAAS und IPVS der Universität Stuttgart wurden in einem Studienprojekt mehrere beispielhafte Policies für OpenTOSCA implementiert. Diese Policies, deren Funktionsweise sowie eine generische Anleitung zur Erstellung neuer Policies werden in diesem Beitrag beschrieben.

Dieser Beitrag ist wie folgt aufgebaut: In Kapitel 2 wird ein Beispielszenario für einen Cloud-Dienst mit Sicherheitsanforderungen vorgestellt. Danach werden in Kapitel 3 spezifische Policies und deren Implementierungen im Policy-Framework gezeigt. Daraufhin beschreibt Kapitel 4 eine Methode, welche die Integration dieser Policies in Cloud-Dienste ermöglicht. Kapitel 5 evaluiert das Policy-Framework anhand von Laufzeitmessungen. Abschließend fasst Kapitel 6 unsere Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Arbeiten.

2 Beispielszenario für einen abgesicherten Cloud-Dienst

Im Folgenden wird ein Beispielszenario für einen Cloud-Dienst mit annotierter Policy vorgestellt. Abbildung 1 stellt die Topologie dieser Anwendung in der Vino4Tosca Notation [BBK+12] dar. Ein Dienst kann manuell oder mit Werkzeugen wie Winery [KBBL13] modelliert und anschließend in einem standardisierten Format gepackt werden. Ein Standard-Format für OpenTOSCA ist beispielsweise das „Cloud Service ARchive“ Format, kurz „CSAR“[BBH+13]. Dieses CSAR-Archiv wird anschließend in OpenTOSCA geladen und ist dann zum Deployment bereit. Die in Abbildung 1 gezeigte Moodle-Topologie besteht aus einem Web-Server-Zweig für die PHP-Anwendung und aus einem Datenbank-Zweig, der auf einer separaten virtuellen Maschine läuft. Für diesen Cloud-Dienst schreibt eine Passwort-Policy Richtlinien für sichere Passwörter vor. Man sagt auch: Es ist eine Passwort-Policy am Moodle-Cloud-Dienst annotiert. Wichtige Passwörter sind in diesem Beispiel: ein Passwort für den Admin-Account der Anwendung, ein Passwort für den Root-Account der MySQL Datenbank und ein Passwort für den darin erstellten Datenbankbenutzer der Moodle-Anwendung. In der OpenTOSCA Laufzeitumgebung ist ein Business Process Execution Language Plan, kurz BPEL-Plan, dafür verantwortlich, dass das Deploy-

ment der Moodle-Topologie in der richtigen Reihenfolge durchgeführt wird. Beim Ausführen des Plans greift die Passwort-Policy und prüft, ob die Passwörter sicher genug sind, oder ob sie durch sichere, neu generierte Passwörter ersetzt werden müssen.

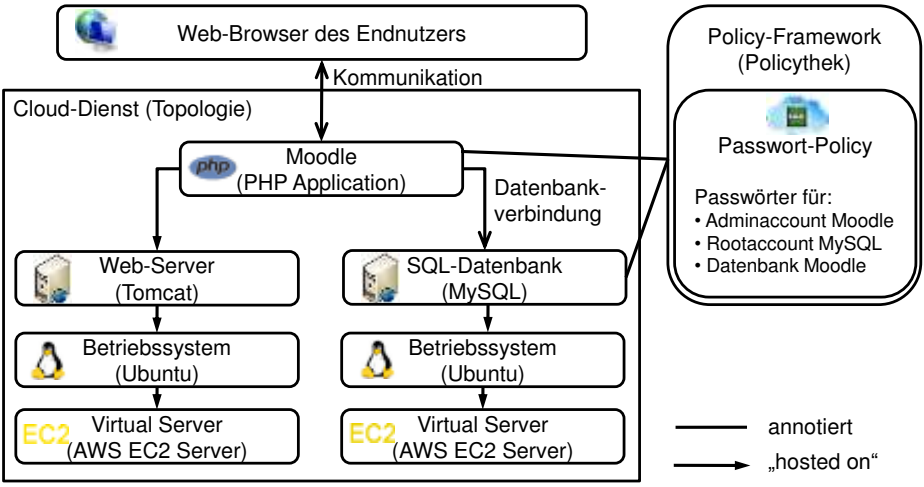


Abb. 1. Die Topologie einer Moodle-Anwendung mit Passwort-Policy

3 Prototypisch umgesetzte Policies im Policy-Framework

In diesem Kapitel werden die im Rahmen des Studienprojekts Policy-Framework prototypisch umgesetzten Policies vorgestellt. Auf Basis dieser ersten Policies wurde eine allgemeine Methode zur Umsetzung und Bereitstellung von weiteren Policies für die Realisierung anderer nicht-funktionaler Anforderungen erstellt. Diese wird in Kapitel 4 beschrieben. Die Definition der Syntax und die Struktur der Policies werden durch den TOSCA-Standard definiert [Kni13], in den folgenden Abschnitten geht es um die Implementierung der Policies, damit diese bei dem automatischen Bereitstellen des Cloud-Dienstes eingehalten und umgesetzt werden.

3.1 Einschränkung der Datenschutzregion durch die Region-Policy

Das Benutzen von virtuellen Maschinen (kurz VMs) zum Betrieb von Cloud-Diensten kann beim Mieten von VMs im Ausland zu Verstößen gegen das Datenschutzgesetz führen, was die Datenhaltung möglicherweise illegal macht. Ebenso möchte man unter Umständen für seine Cloud-Dienste nur bestimmte Rechenzentren zulassen, deren Betreibern man vertraut. Die Region-Policy ist eine Policy, welche mit einem Webservice und einer dazugehörigen Datenbank Regionen verwaltet, wie in Abbildung 2 dargestellt. Bei Cloud-Diensten, an denen eine Region-Policy annotiert ist und bestimmte zugelassene Regionen definiert sind, wird von dem Webservice vor dem Erstellen einer virtuellen Maschine (auch Instanziieren genannt) überprüft, ob diese in

einer gültigen Region instanziiert wird. Ein Deployment des Systems in einer nicht zugelassenen Region wird durch einen Abbruch verhindert.

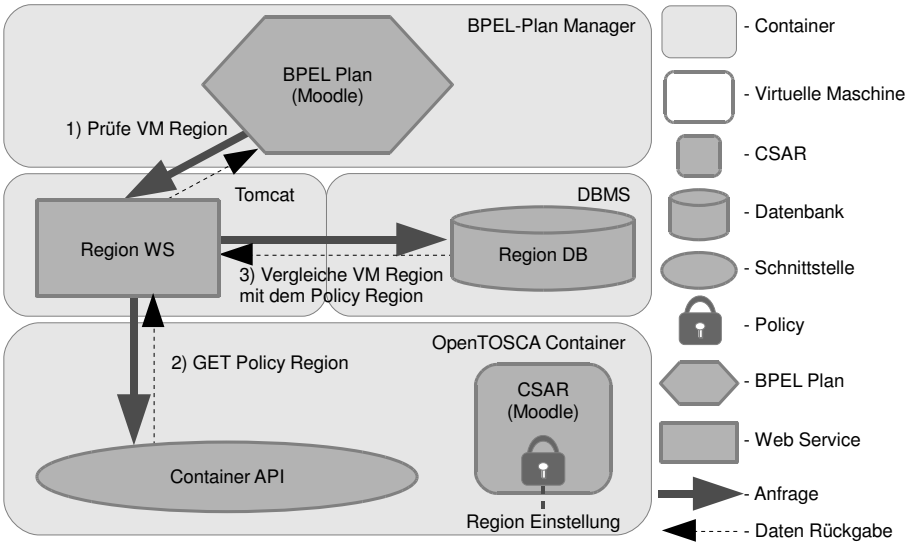


Abb. 2. Der Region-Policy Webservice prüft die Zulässigkeit einer Region für das Deployment der Anwendung.

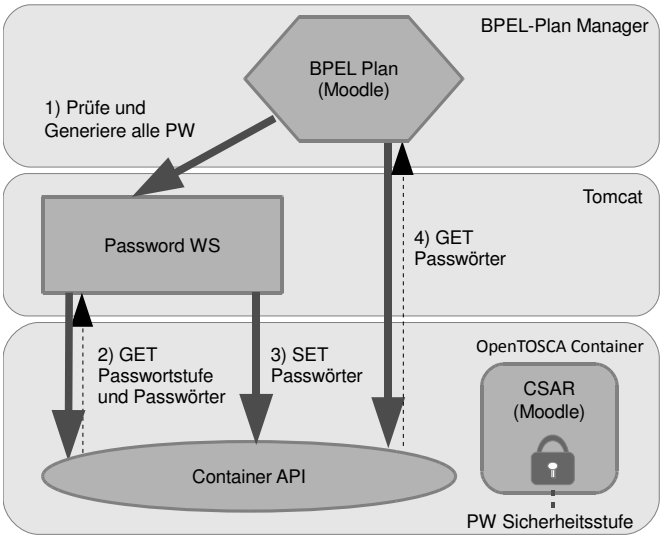


Abb. 3. Der Passwort-Policy Webservice erzeugt sichere Passwörter für die Anwendung.

3.2 Absicherung der Passwortstärke durch die Passwort-Policy

Passwörter sind im Bereich von Cloud-Diensten wichtig, weil initiale Passwörter oft trivial, oder Standardpasswörter wie „passwort“ sind. Solche Passwörter können heutzutage schnell und automatisiert erraten werden, was ein klares Sicherheitsrisiko darstellt [Eal03]. Die Passwort-Policy kann für bestimmte Teile eines Cloud-Dienstes Passwörter generieren, falls sie an dem Cloud-Dienst annotiert ist. Die Passwörter werden gemäß vordefinierter Sicherheitsstufen und Parameter erstellt. Auch hierfür wurde ein Webservice entwickelt, welcher diverse Methoden zur Generierung von Passwörtern anbietet. Wie dieser analog mit einer Anwendung funktioniert, sieht man in folgender Abbildung 3.

3.3 Energieeinsparung durch die Green-Policy

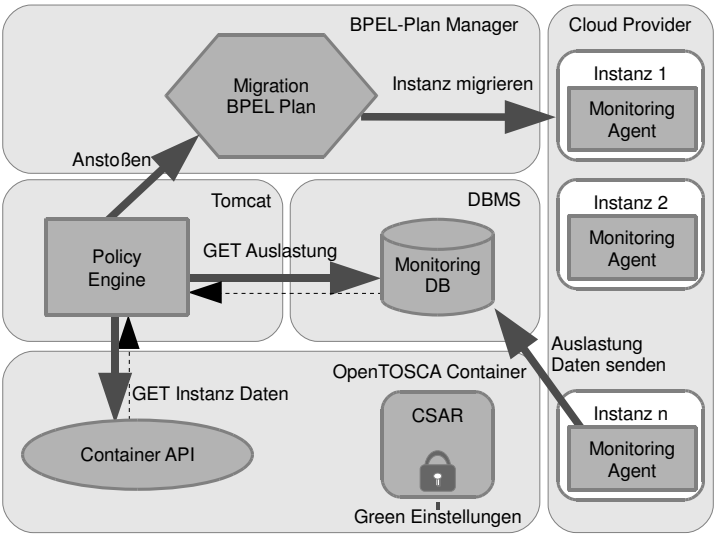


Abb. 4. Die Policy Engine stößt Migrationen nach Analyse gesammelter Instanz-Daten an.

Cloud-Dienste, deren Anwendungen nur sehr wenig Rechenkapazität von einer VM fordern, unterlasten die VM. Die Partei, die für die VM zahlt, muss höhere Kosten als nötig aufbringen, und die Partei, die die VM bereitstellt, hat höhere Stromkosten als nötig und ist so auch weniger umweltfreundlich. Umgekehrt kann es auch sein, dass ständig überlastete VMs der Forderung nach Rechenkapazität des Cloud-Dienstes nicht mehr gerecht werden und die Leistung des Cloud-Dienstes abnimmt. In beiden Fällen lässt sich das Problem lösen, wenn man für unterlastete VMs eine kleinere, leistungsschwächere VM wählt und für überlastete VMs eine größere und leistungsstärkere.

Die CPU-, Arbeitsspeicher- und Festplattengrößen sind vom Cloud-Provider abhängig. Bei der prototypischen Green-Policy Implementierung wurde Amazon als Cloud-Provider genutzt. Amazon gibt vordefinierte Größen für VMs an, die in allen der vorher genannten Werte mit jeder rechenstärkeren Größe auch entsprechend steigen. Wir

bezeichnen diese Kombination im folgenden als VM-Größe. Mit der Green-Policy hat der Cloud-Anbieter durch Angabe einer Liste von VM-Typen (sortiert nach ihrer Rechenleistung) die Möglichkeit automatisch auf nächstkleinere VMs bei Unterlastung und nächstgrößere VMs bei Überlastung zu wechseln. Dieser Vorgang wird hier auch Skalieren genannt, also das Anpassen der VM-Größe an die Anforderung des Cloud-Dienstes. Der Prozess des Wechselns und Kopierens aller Daten auf eine andere VM wird Migration genannt und ist in Abbildung 4 im Zusammenspiel der Green-Policy-Komponenten dargestellt. Die Green-Policy besteht aus einem Web-Service, der den Namen Policy Engine trägt. Er speichert Lastdaten seiner Instanzen in einer Datenbank und stößt mit Durchschnittswerten zu bestimmten Zeitperioden Migrationen an, wenn diese nötig sind. Folglich hat jede Instanz eines Cloud-Dienstes, an dem eine Green-Policy annotiert ist, ein Überwachungsprogramm installiert, das Monitoring Agent genannt wird und Lastdaten an die zentrale Datenbank der Policy Engine schickt.

3.4 Absicherung der gespeicherten Daten durch die Verschlüsselungs-Policy

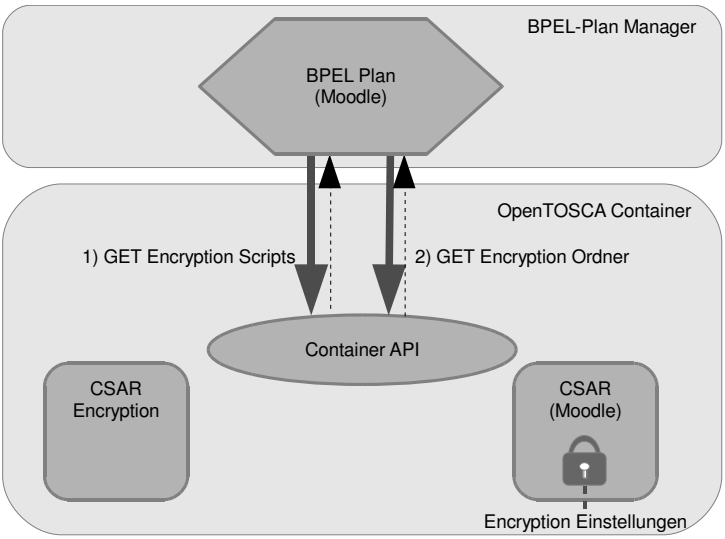


Abb. 5. Die Encryption CSAR beinhaltet alle Skripte, die den Encrpytion Ordner einer zu installierenden Anwendung verschlüsseln.

Wenn eine Datenbank oder ein Ordner mit sensiblen Daten eines Cloud-Dienstes unverschlüsselt auf einer VM liegt, so sind diese Daten vor Datendiebstahl nicht geschützt. Ein Beispiel für so einen Diebstahl ist das Kopieren der Daten von der VM auf ein Massenspeichergerät. Da die Daten unverschlüsselt sind, lassen sie sich im Klartext auf jedem anderen Computer lesen. Das Ziel der Verschlüsselungs-Policy ist es, sensible Daten einer Anwendung zu verschlüsseln. Der entstandene Schaden bei oben genanntem Diebstahl-Szenario ist dann weniger kritisch, da sensible Daten dann nicht mehr gelesen werden können. Die Verschlüsselungs-Policy nutzt das Programm

Ecrypt-FS [ecr], um Ordner zu verschlüsseln. Dies geschieht beim Deployment automatisch, wenn am Cloud-Dienst eine Verschlüsselungs-Policy annotiert ist. Der zu verschlüsselnde Ordner wird dabei als Parameter der Policy angegeben. Abbildung 5 beschreibt den Aufbau.

4 Methode zur Integration von Policies in Cloud-Dienste

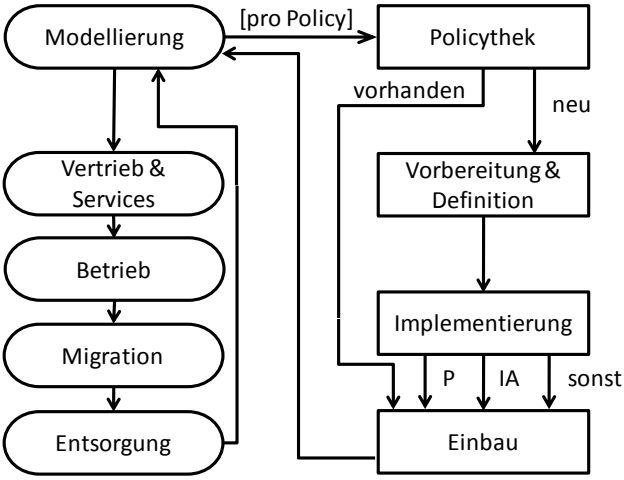


Abb. 6. Der Cloud-Cycle (links) und der Policy-Cycle (rechts) mit der Policythek

In diesem Kapitel wird eine allgemeine Methode zur Erstellung neuer Policies für die Realisierung nicht-funktionaler Anforderungen und zum Einsatz bestehender Policies in Cloud-Diensten beschrieben. Die Methode besteht aus vier Schritten, welche sich am Lebenszyklus einer Policy orientieren. Der Lebenszyklus von Policies (Policy-Cycle) baut wiederrum auf dem Lebenszyklus des zugehörigen Cloud-Dienstes (Cloud-Cycle) auf. Dies ist in Abbildung 6 dargestellt. Links sieht man den Cloud-Cycle, rechts den Policy-Cycle für eine annotierte Policy.

Der Policy-Cycle beginnt mit der Überprüfung, ob eine passende Policy nicht schon in der Policythek existiert (Schritt 1 siehe Kapitel 4.1), die man nutzen und sofort einbauen könnte. Wenn noch keine Policy existiert, muss eine entsprechende Policy zuerst vorbereitet und definiert werden (Schritt 2 siehe Kapitel 4.2). Anschließend wird in Kapitel 4.3 (Schritt 3) gezeigt, was mindestens nötig ist, um eine Policy zu implementieren und welche verschiedenen Möglichkeiten zur Implementierung existieren. In Kapitel 4.4 wird dann der Einbau einer Policy in einen Cloud-Dienst genauer beschrieben (Schritt 4).

4.1 Schritt 1: Policythek – Auswahl einer Policy

Im Rahmen des Policy-Frameworks wurde die Policythek erstellt. Sie zeigt die Verfügbarkeit der wichtigsten Bestandteile des Policy-Frameworks: die Policies. Hier kann eingesehen werden, ob Policies in einer OpenTOSCA Laufzeitumgebung installiert sind.

Im Policy-Framework wurden Policy-Pakete zur Installation von Policies definiert, die eine ähnliche Struktur wie CSAR-Pakete haben. Diese Pakete werden wie jedes CSAR-Paket in OpenTOSCA hochgeladen und automatisch als Policy-Paket erkannt. Danach erscheint die Policy in der Policythek mit ihrem Logo. Durch einen Klick auf das Logo öffnet sich eine Kurzbeschreibung der Policy. Der entscheidende Vorteil des Policy-Frameworks ist es, dass Policies mit ihren Webservices und anderen Daten nun an einer zentralen Stelle in der OpenTOSCA Umgebung liegen, von der aus jeder Cloud-Dienst dieser Umgebung Zugriff darauf hat. Wenn hier eine Policy existiert, die von einem Cloud-Dienst benötigt wird, so kann man diese einbauen, wie in Schritt 4 erklärt wird. Wenn keine passende Policy existiert, muss eine neue erstellt werden, wie in Schritt 2 beschrieben wird.

4.2 Schritt 2: Vorbereitungen und Definitionen einer neuen Policy

Will man eine Policy mithilfe des Policy-Frameworks umsetzen, so muss man zuerst folgende Überlegung anstellen: Ist die umzusetzende Policy generisch, sprich, kann die Policy ohne große Änderungen an der Policy Implementierung bei verschiedenen Cloud-Diensten annotiert und eingesetzt werden? Trifft dies nicht zu, so ist eine Implementierung über das Policy-Framework nicht anzuraten, sondern es sollte eine individuelle Implementierung für den jeweiligen Cloud-Dienst erfolgen. Bei generisch einsetzbaren Policies geht man unter Berücksichtigung folgender Kriterien vor:

- Spezifikation der Policy: Welche genaue Funktion soll die Policy umsetzen? Hierbei sollten textuelle Beschreibungen für die Entwickler und für die Endanwender, sowie eine Spezifikation erstellt werden. Zudem wird ein Logo für die Policy benötigt, damit diese in der Policythek angezeigt werden kann.

- Entwurf der Policy: Welche Funktionalitäten müssen von dem zu erstellenden Webservice und möglichen weiteren Programmen und Scripten geboten werden, um die Funktion der Policy umzusetzen?

- Architektur der Policy: Welche Dateien werden für die Umsetzung dieser Policy benötigt? Dazu gehören TOSCA Policy-Definitionen, Webservices, Scripte, Programme, Datenbanken und Ähnliches.

4.3 Schritt 3: Implementierung der benötigten Systemteile

Sind alle Vorüberlegungen geklärt, so können alle benötigten Dateien und Dokumente erstellt werden. Sind diese Dateien und Dokumente erstellt worden, so können diese

in ein Policy-Paket gepackt werden, welches eine spezielle Form eines CSAR-Paketes darstellt. Ein solches Policy-Paket enthält folgende Dateien, die für die Umsetzung der Policy nötig sind:

- den Ordner Definitions mit allen nötigen .xml Dateien
- den Ordner IAs mit eventuellen Webservices, wobei IA hier für Implementation Artifact steht
- den Ordner Imports. Er enthält die zu den Webservices gehörenden .wsdl-Dateien
- den Ordner TOSCA-Metadata mit der Datei TOSCA.meta
- den Ordner POLICY-Metadata, in welchem die Dateien für die Policy-spezifische Webseite der Policythek liegen. Darin ist mindestens eine Datei namens data.xml einzufügen mit einer Beschreibung der Policy und einer Referenz auf das Icon. Der Aufbau einer beispielhaften data.xml kann hier abgerufen werden: http://pastebin.com/1d8tsZvp
- den Ordner Scripts mit optionalen Scripten.

Alle diese Dateien werden in einem Ordner gespeichert und als .csar Datei gepackt. Dieses Paket kann nun als Policy-Paket in OpenTOSCA hochgeladen werden. Sobald es in der Policythek erscheint, ist es erfolgreich installiert worden und die Policy kann von nun an von allen Cloud-Diensten dieser OpenTOSCA-Laufzeitumgebung genutzt werden.

4.4 Schritt 4: Einbau der Policy in ein CSAR-Paket mittels Annotation

An dieser Stelle wurde ein Policy-Paket erfolgreich als Policy in einer OpenTOSCA Laufzeitumgebung installiert oder existiert dort bereits. Nun muss man Änderungen am Gegenstück vornehmen: dem Cloud-Dienst selbst, genauer gesagt am CSAR-Paket, das den Cloud-Dienst beschreibt. Hier muss die Policy am entsprechenden Service-Template oder Node-Type annotiert werden. Dies geschieht durch die Nutzung des TOSCA-Policy-Tags. Dies ist nötig, damit der Cloud-Dienst beispielsweise den Webservice ansprechen kann, den eine Policy bietet. Hier unterscheidet man je nach Implementierung des Webservices nach den zwei in [WWB+13] vorgestellten Ansätzen:

Der erste Ansatz ist der IA-Approach. Dieser ermöglicht es dem Webservice, mit der OpenTOSCA Container-API zu kommunizieren und automatisch die Struktur eines Cloud-Dienstes nach Stichwörtern zu durchsuchen. Wir stellen dies am Beispiel einer Passwort-Policy mit entsprechendem Webservice vor. Der Webservice prüft automatisch bei allen nötigen Elementen, ob ein neues Passwort zu generieren ist und generiert dieses. Dies hat den Vorteil, dass der CSAR-Ersteller, der auch Lösungspaketersteller genannt wird, nicht für jedes Element, das ein Passwort benötigt eine

Abfrage in den BPEL-Plan setzt, sondern er muss nur einen generischen Webservice Aufruf in den Workflow einfügen (Beispiel siehe http://pastebin.com/pWcAcZ5F). Allerdings muss der Lösungspakethersteller darauf achten, dass er eine Namenskonvention einhält für alle Elemente, die ein Passwort enthalten, damit der Passwort-Webservice diese auch als solche erkennt - beispielsweise durch die Konvention jedem Knoten, der ein sicheres Passwort haben soll, das Wort passwort anzuhängen.

Der zweite Ansatz ist der P-Approach, oder Plan-Approach, bei dem keine Kommunikation des Web-Services mit OpenTOSCA stattfindet. Das hat zur Folge, dass der Lösungspakethersteller genau für jedes seiner Elemente, die ein sicheres Passwort brauchen, im BPEL-Plan eine separate Anfrage an den Webservice der Policy schickt, sich eine zweite Anfrage zur Passwort-Generierung baut und ein Passwort generieren lässt. Das muss er für jedes seiner Elemente machen, das so ein sicheres Passwort braucht. Ein Beispiel für die erste Anfrage wird in http://pastebin.com/kj6U4CM3 dargestellt.

5 Evaluation

Zur Evaluation wurden die Laufzeiten zur Instanziierung zweier Anwendungen gemessen. Die Moodle Anwendung wurde mit der Passwort Policy, der Region Policy und einmal ohne Policy gemessen. Die SimpleBlog Anwendung wurde einmal mit der Encryption Policy und einmal ohne Policy gemessen. Alle Messungen wurden auf der gleichen OpenStack Infrastruktur (Ein Management und ein Hypervisor Server, je 12 Intel Xeon E5-2630 CPUs bei 2,3GHz, 256GB Speicher) durchgeführt und für jede Anwendung wurden die gleichen Instanzgrößen gewählt (8 vCPUs, 16GB Hauptspeicher, 20GB Festplatte). Die Verschlüsselung-Policy verändert die Instanz des Cloud-Dienstes, wohingegen die Passwort- und Region-Policy ihn lediglich umkonfigurieren. Es hat sich gezeigt, dass die Passwort- und Region-Policy die Instanziierungszeit des Cloud-Dienstes nicht verlängern (siehe Abbildung 8). Dies folgt daraus, dass die Policies bereits in der Policythek vorinstalliert sind und den Großteil der Laufzeit des Deployments die Instanzerstellung der VMs auf der verwendeten OpenStack-Plattform ausmacht. Diese Laufzeit hängt darum vor allem von der aktuellen Auslastung der OpenStack-Infrastruktur ab. Sie wird nur vernachlässigbar von den Policies beeinflusst.

Einen Unterschied stellt die Verschlüsselungs-Policy dar (siehe Abbildung 9). Diese muss bei jeder Instanziierung für jeden zu verschlüsselnden Ordner ausgeführt werden, wodurch sich die Laufzeit entsprechend verlängert. Auch hier sind die Laufzeitverlängerungen durch Annotation der Policy vernachlässigbar, da die Instanziierung des Cloud-Dienstes selbst den Großteil der Laufzeit ausmacht.

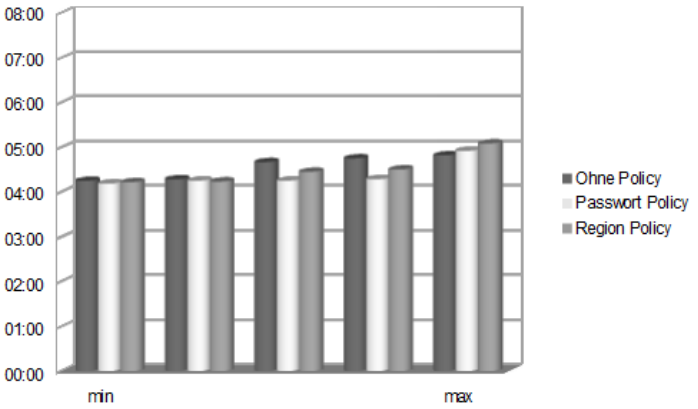


Abb. 7. Laufzeiten einer Instanziierung einer Moodle-Anwendung ohne Policy (dunkelgrau) mit Passwort-Policy (hellgrau) und mit Region-Policy (grau)

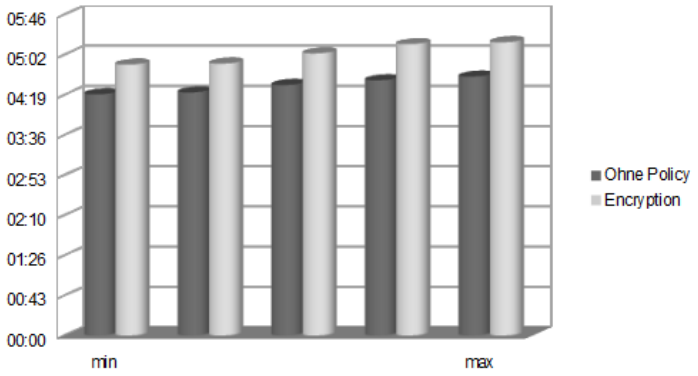


Fig. 8. Laufzeiten einer Instanziierung einer SimpleBlog-Anwendung. Ohne Policy (dunkelgrau) und mit Verschlüsselungs-Policy (hellgrau).

6 Zusammenfassung und Ausblick

Das Policy-Framework wurde entwickelt, um zu untersuchen, wie Policies generell in Cloud-Dienste integriert werden können. Es wurde erreicht die Policies zentral zu verwalten, ihren Aufbau möglichst generisch zu beschreiben und sie von der Entwicklung der Cloud-Dienste abzugrenzen. Dies wird durch die vorgestellten Prototypen aufgezeigt. Des Weiteren ist zu beachten, dass die Ausführung der Policies aufseiten des Cloud-Anbieters geschieht, sodass zwischen Kunde und Cloud-Anbieter ein Vertrauensverhältnis erforderlich ist, um sicherzustellen, dass die Policies eingehalten werden. Um dies zu unterstützen, sollte ein umfassendes Logging aufseiten des

Cloud-Anbieters durchgeführt werden, welches eine Nachvollziehbarkeit der ausgeführten Prozesse erlaubt.

In zukünftigen Arbeiten könnten, analog zur Standardisierung von TOSCA-Node-Types, standardisierte TOSCA-Policies definiert werden, sodass Entwickler neuer Node-Types und Dienste deren Unterstützung bereits vorsehen können. Dies hätte den Vorteil, dass die standardisierten Policies in jeder TOSCA-Laufzeitumgebung verfügbar wären und somit nicht installiert werden müssten und dadurch bei der vorgestellten Integrations-Methode der erste Schritt entfallen könnte.

Danksagung: Die Autoren danken dem Bundesministerium für Wirtschaft (BMWi) für die Förderung im Rahmen des Migrate! (01ME11055) und CloudCycle (01MD11023) Projekts. Die Autoren danken auch allen Teilnehmern, Betreuern und Kunden des Studienprojekts „Policy-Framework“ für die Zusammenarbeit.

Literaturverzeichnis

[BBH+13] T. Binz et. al. OpenTOSCA - A Runtime for TOSCA-based Cloud Applications. In Proceedings of 11th International Conference on Service-Oriented Computing *(ICSOC'13)*, volume 8274 of LNCS, S. 692–695. Springer Berlin Heidelberg, Dezember 2013.

[BBK+12] U. Breitenbücher et.al. Vino4TOSCA: A Visual Notation for Application Topologies based on TOSCA. In Proceedings of the 20th International Conference on Cooperative Information Systems (CoopIS 2012), Lecture Notes in Computer Science. Springer-Verlag, September 2012.

[Eal03] K. Ealy. A New Evolution in Hack Attacks: A General Overview of Types, Methods, Tools, and Prevention. S. 8-9, 2003.

[ecr] Ecrypt FS Homepage . http://wiki.ubuntuusers.de/ecryptfs.

[fSidI] Bundesamt für Sicherheit in der Informationstechnik. Konzeption von Sicherheitsgateways., https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Internet sicherheit/Konz_SiGw_pdf.pdf?__blob=publicationFile. Version 1.0S. 16.

[KBBL13] O. Kopp et. al. Winery - A Modeling Tool for TOSCA-based Cloud Applications. In Proceedings of 11th International Conference on Service-Oriented Computing *(ICSOC'13)*, volume 8274 of LNCS, S. 700–704. Springer Berlin Heidelberg, Dezember 2013.

[Kni13] P. Knight. Topology and Orchestration Specification for Cloud Applications Version 1.0., November 2013.

[MSM08] M. Messerschmidt, P. Schülein, und M. Murnleitner. Der Wertbeitrag der IT zum Unternehmenserfolg. S. 57, 2008.

[WWB+13] T. Waizenegger et. al. Policy4tosca: A policy-aware cloud service provisioning approach to enable secure cloud computing. On the Move to Meaningful Internet Systems: OTM 2013 Conferences, volume 8185 of LNCS, pages 360–376. Springer Berlin Heidelberg, 2013.

On the Measurement of Data Protection Compliance of Cloud Services

Thomas Kunz, Annika Selzer, Ulrich Waldmann

Fraunhofer Institute for Secure Information Technology SIT
Rheinstr. 75, 64295 Darmstadt
{thomas.kunz|annika.selzer|ulrich.waldmann}@sit.fraunhofer.de

Abstract: Companies want to benefit from the numerous advantages of cloud services such as flexibility and cost efficiency. However, cloud services vary considerably with respect to the security and privacy mechanisms provided. Moreover, security-aware companies complain the lack of transparency concerning the security measures and processes the cloud provider has installed. As a solution for the latter one, auditors may evaluate cloud providers and issue certificates attesting whether the cloud provider meets the agreed requirements. However, due to the characteristics of cloud computing, on-site inspections in the data centers of a cloud provider do not seem to be realistic. In this paper we show how metrics can be derived from data protection requirements and how these metrics can be expressed in the form of formal policies, in order to be used for an automated evaluation of cloud services[1].

1 Introduction

Cloud services promise numerous advantages for companies with respect to flexibility and cost efficiency. They are hosted in data centers that are professionally managed, physically protected, and equipped with failsafe hardware. However, in business context, security and data protection aspects become very important [CPA12]. This is one of the reasons why many companies still have concerns regarding the use of cloud services: on the one hand, companies are still responsible for the protection of their data, even if they shift the storage and the processing of the data into the cloud. But on the other hand, they do not trust the cloud providers in terms of confidentiality and data protection even though they expect the cloud providers to act according to their contractual agreements.

In [KSW14] and [KNW13] we proposed an automatically generated data processing certificate for cloud services, in order to compensate this problem. This solution is based on the standardized audit suggested by the working group "Legislation" of the "Trusted Cloud"-technology program [Sel13][B+12][2]. We use secure log files as a trustworthy data base for the evaluation of a cloud service. For this reason, we introduced a secure logging

[1]This contribution has been created within the project CloudCycle (`http://www.cloudcycle.org`). CloudCycle is funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) according to a decision of the German Federal Parliament within the funding priority "Trusted Cloud".

[2]`http://www.trusted-cloud.de`

method which is resistant against internal manipulation attempts through tamper-proof and confidential log data. In order to guarantee an automated control of the cloud service, we assumed that the underlying security and data protection requirements are described in a formal policy. Unfortunately, typical data protection requirements based on legal regulations like the Federal Data Protection Act (BDSG) are too generic for an automated evaluation. They do not define what has to be measured to control the requirements and which measured values are sufficient to fulfill the requirements.

In this paper we show how metrics may help to derive formal data protection policies from rather abstract data protection requirements. The metric definitions are essential for specific measurements that in turn are compared with parameters specified in the policies. This way an automated evaluation of cloud services seems possible.

The structure of this paper is as follows: Related work is discussed in Section 2. Section 3 shows how generic data protection requirements can be concretized in the form of metrics. Section 4 illustrates the translation of the metrics into formal data protection policies. The paper ends with an outlook in Section 5.

2 Related Work

To be able to measure the quality of data protection, quantitatively measurable indicators are needed – both for the technical components and for the organizational aspects. A promising project is the definition of security metrics to support management decisions in information security which is led by the Center for Internet Security (CIS). The current version from 2010 contains 28 definitions of metrics in the fields of incident management, vulnerability and patch management, application security, configuration management, change management, and financial indicators [CIS10]. However, the metrics focus rather on organizational and financial security indicators than on data protection indicators as required in BDSG.

[Jul09] also deals with typical security metrics like "number of security incidents". The paper introduces three properties that all security metrics have to fulfill (validity, accuracy, and precision), and gives a formal definition of the term "security metric". Furthermore, it classifies security metrics by means of the input data analyzed by the metrics. In [Böh10], Böhme covers security metrics from an economic perspective. In [Bez10], Bezzi provides an information theoretic approach for privacy metrics (e. g. k-anonymity) in order to express and compare disclosure risk metrics.

3 Derivation of Metrics from Data Protection Requirements

Data protection requirements derive primarily from legal regulations such as the BDSG. In Germany, the contract data processing according to Sec. 11 BDSG is crucial within the cloud computing context. One main requirement is taking technical and organizational

measures to protect the personal data processed by the processor. The annex of Sec. 9 BDSG includes eight requirements for technical and organizational measures like access control and disclosure control. In the remainder of this section, we will consider examples of measures and metrics for the following two requirements of the BDSG annex:

- Protection of data against accidental destruction or loss (availability control) and
- Separate processing of data collected for different purposes.

These generic requirements are hardly precise enough for an automated evaluation of cloud services. Looking at the examples above, it is unclear how to decide whether a cloud provider sufficiently cares for the availability of the customers' data and whether the customers' data have been separated adequately or not. A concise specification of technical and organizational measures is required, in order to ensure the conformance with these requirements. So, the next step is to search for appropriate measures.

3.1 Selection of Measures

Since cloud customers often rely on the IT-Grundschutz methodology [BSI08] that includes threat analysis, requirements and measures regarding different IT security aspects, it may be a good source of measures for data protection. The methodology suggests that a list of possible requirements is divided into three protection categories: "normal", "high", and "very high". Then it may be easier to identifier the damages that could occur under loss of protection and to find appropriate countermeasures.

The availability control obviously comprises many different aspects as it regards all parts of a process, i. e. data, IT systems, facilities and personnel. The creation of backups and redundancy are standard measures, others are the usage of virus scanners or the defense against DoS attacks. The protection category may influence the strength of selected measures, as regards measures of backups e. g., the backup frequency and storage period. Measures to contribute to the separate data processing may be implemented, e. g. on application level (use of separate databases or software with multi-client capability), on virtual level (providing each customer with an own virtual machine) or, for even higher protection, on hardware level.

It should be noted that protection requirements are often interdependent and even opposing each other like availability and confidentiality to control disclosure: whereas perfect availability could be realized by publishing in the Internet, the availability of data should probably be restricted to authorized parties. On the other hand, one measure may contribute to more than one requirement, e. g. cryptographic encryption may help to prevent unauthorized access as well as disclosure. In general, a whole bundle of carefully balanced measures is needed to fulfill generic requirements and adhere to the data protection principles of data economy, data avoidance and transparency. A survey of measures in relation to several data protection requirements is given in [Pro12].

Additionally, cloud customers may define further individual protection requirements, such

as backup locations restricted to a certain Federal State in Germany. Both requirement types determine the measures a cloud provider has to implement. In the following, we derive two sample metrics from a backup measure and a measure for separate data processing.

3.2 Definition of Metrics

After deploying selected measures, it has to be determined how to decide if a measure has been adequately implemented. This can be done by a system of metrics that effectively relates to measurements of desired compliance features. Each metric should represent a part of the real compliance state and have the following characteristics. First, a good metric is quantifiable as it is based on quantitative measurements. Second, it is repeatable yielding the same result in a repeated assessment, and third, it is comparable as it is based on a linearly ordered set of normalized values. Each metric should provide useful information for a continuous business improvement towards a defined target value. Most metric definitions include measurement units combined with a time unit (e. g., per week or mean-time between two activities) or a proportion to a total (e. g., percentage) [ENI12], and a compliance target (setpoint) that will be considered in relation to regular measurements (actual values).

The Center of Internet Security (CIS) has developed a set of agreed security metrics including the underlying formula and units, recommended measurement frequencies, target values and possible measurement sources [CIS10]. The target describes the good and bad value ranges - including ideal targets, e. g. zero security incidents, or defines a precise threshold to decide whether the measured value can be considered compliant with the generic security requirement. The following two tables show data protection metrics defined with similar attributes. Table 1 shows the definition of a sample metric for the "backup" measure. It may contribute to the indicator for the degree of availability control.

Name	**Mean Time Interval between Backups**
Description	The average time between the last two backups of all storage media with the data of a specific customer.
Formula	The Mean Time Interval between Backups (MTIB) is calculated by determining the number of hours between the date of the last backup and the date of the current backup for each storage medium. These results are averaged across the number of the relevant storage media: $$MTIB = \frac{\sum(CurrentBackup_Date - LastBackup_Date)}{Count(Data_Storage_Media)} \quad (1)$$
Units	Hours per backup
Frequency	Daily, weekly, monthly, quarterly, annually
Targets	MTIB values should trend lower over time. Generally, the target time for MTIB depends on the desired availability level, e. g. normal: 24 hours; high: 6 hours; very high: 1 hour.

Table 1: Availability metric "Mean Time Interval between Backups"

In general, for a single measure more than one metric is required. For example, in the case of backups this may include further metrics such as "The mean distance between the storage location of the original data and their backup", and "The average storage period of backup data". As a result, we would end up with a whole catalog of metrics. Table 2 shows the definition of a sample metric for a "Customer-specific separate data processing". It may contribute to the indicator for the degree of separate processing of data collected for different purposes.

Name	**Separate Processing Coverage**
Description	This metric attempts to answer the question "Are the data of a specific customer processed separately from other customers' data?". The metric's unit is the percentage of total storage and processing media that are exclusively allocated for a specific customer or have data-protecting multi-client capability. Examples of media may include applications, virtual machines, storage systems like databases, server components, hardware components.
Formula	Separate Processing Coverage (SPC) is calculated by determining the number of storage and processing media with data-protecting multi-client capability and then averaging this across the total number of storage and processing media: $$SPC = \frac{\sum(MultiClient_StorageProcessingMedia)}{Count(Media)} * 100 \quad (2)$$
Units	Percentage of storage and processing media
Frequency	Daily, weekly, monthly, quarterly, annually
Targets	The expected trend for this metric over time is to remain stable or increase towards 100%. The SPC scope depends on the desired separation level, e. g. normal: on application level by means of access control rules; high: on VM basis or separate database with access control rules; very high: on hardware level

Table 2: Metric "Separate Processing Coverage"

How can an effective system of metrics contribute to a continuous control of compliance features? The individual metrics should be aggregated, in order to calculate the significant compliance indicators, e. g. one indicator for each data protection requirement (each based on several metrics). The cloud provider may provide currently determined indicators that enable an evaluation of data protection-relevant qualities at every stage of a cloud service (service offer, instantiation, configuration, usage, termination). A broker service may use the indicators for advertising and matching the service qualities with individual customer requests. The cloud provider may empower the cloud customers to verify the compliance during operation by means of a reporting dashboard that visualize the relevant indicators.

However, there are obstacles to implement a customer-centric system of metrics. Most existing cloud services do not care about sector-specific requirements or even customers' individual requests. This is because "One size fits all" cloud solutions seem more effective and cost-efficient from the perspective of cloud providers. In the next section we show how machine-readable policies can be derived from metrics, in order to enable an automatic evaluation of cloud services. This may help to overcome the economic obstacles to customized services and metrics.

4 Data Protection Policies

Machine-readable policies are written in a formal language and may convey the properties of metrics into a technical control system for cloud services. The metrics precisely define what has to be measured to control the requirements and which measured values are sufficient to fulfill the requirements. The metrics and the corresponding thresholds serve as input for policies which in turn are the base to automatically evaluate the requirements.

Further use cases for such policies exist: applications may use them to automatically match cloud services to the customer's requirements. They may be used by cloud management environments to enable an automated enforcement of the security requirements. For these use cases, the policies should be standardized, in order to enable a widespread support through many cloud management environments.

A recently initiated standardization effort from OASIS, the "Topology and Orchestration Specification for Cloud Applications" (TOSCA) [LMPS13], provides an interesting approach. Using TOSCA, cloud service creators are able to describe the topology, the deployment, and the management of portable cloud services. TOSCA allows an automatic deployment and management of cloud services through the cloud platform. In addition, TOSCA allows the definition of policies that describe requirements of any kind. Current research deals with the definition of non-functional requirements in TOSCA as policies and how these policies can automatically be processed by a cloud environment [WWB+13].

In the following, we present two example policies that implement the data protection metrics in Section 3. The chosen syntax is close to the syntax of the TOSCA policies.

```
<PolicyTemplate type="BackupInterval">
  <Properties>
    <cc:Threshold>
      <cc:SecurityLevel value="normal">
        <cc:Frequency interval="24" timeUnit="hours"/>
      </cc:SecurityLevel>
    </cc:Threshold>
  </Properties>
</PolicyTemplate>
```

Listing 1: Policy for Backup Frequency

```
<PolicyTemplate type="DataSeparation">
  <Properties>
    <cc:Threshold>
      <cc:SecurityLevel value="high">
        <cc:Separation value="VM_Level"/>
      </cc:SecurityLevel>
    </cc:Threshold>
  </Properties>
</PolicyTemplate>
```

Listing 2: Policy for Data Separation

In order to enable an automated control of the cloud service, the data protection requirements are described in metrics and then transformed into formal policies. As part of a formal description of a cloud service the policies may contribute to customized services that are - due to the automation potential - not necessarily more complex or costly than the conventional pre-built solutions.

A software component capable of interpreting these policies is required for an automated evaluation of cloud services. Such a component has to implement the formulas presented in Section 3 to measure the thresholds defined in the policies. Furthermore, for performing the analysis, the component needs access to a trustworthy data basis such as secure log data [KSW14].

5 Outlook

In this paper we have shown how metrics can be derived from abstract data protection requirements and how in turn these metrics can be expressed as formal policies. Such policies may contribute to an automatical control of the technical and organizational measures implemented by the cloud provider. These controls could be performed at certain time intervals, e. g. every year, and even at any time during regular operation as well. The results of the controls may serve as input for data protection certificates as proposed by the working group "Legislation" of the technology program "Trusted Cloud" [B+12].

The metrics and corresponding indicators can be used for an independent certification of individual cloud services. Most existing cloud certificates do not take account of sector-specific requirements. In general, customers are not able to evaluate the quality of certificates or to compare the offers of different certification bodies. Furthermore, most available cloud certificates refer to general qualities of the cloud provider rather than to specific cloud services [SLS13].

Future research should focus on developing new tools that enable companies and authorities to describe and measure automatically standardized data protection requirements. In order to make the privacy characteristics of different cloud providers comparable, a catalog of data protection requirements and measures including measurement characteristics has to be defined. Such a catalog could serve as a guidance for cloud providers to implement sector-specific requirements (such as for health care or financial services) as well as for auditors to define corresponding certification practice statements and evaluate the level of technical and organizational measures taken by the cloud provider.

Further efforts should address the technical, economic and legal challenges of implementing a system of data protection metrics, its effectiveness (attainment of objectives) and efficiency (minimized investment). For example, a cloud provider should not be able to selectively optimize the specified measurements without really improving the data protection level (strategy resilience). These challenges are within the scope of the BMBF-sponsored follow-up project VeriMetrix[3].

[3] http://www.verimetrix.de

References

[B+12] G. Borges et al. Datenschutzrechtliche Lösungen für Cloud Computing. Kompetenzzentrum Trusted Cloud, Oct 2012.

[Bez10] M. Bezzi. An information theoretic approach for privacy metrics. *Transactions on Data Privacy*, 3(3):199–215, 2010.

[Böh10] R. Böhme. Security Metrics and Security Investment Models. In I. Echizen, N. Kunihiro, and R. Sasaki, editors, *Advances in Information and Computer Security*, volume 6434 of *Lecture Notes in Computer Science*, pages 10–24. Springer Berlin / Heidelberg, 2010.

[BSI08] BSI Standard 100-2: IT-Grundschutz-Vorgehensweise. Technical report, Bundesamt für Sicherheit in der Informationstechnik, 2008.

[CIS10] The CIS Security Metrics v1.1.0. Technical report, The Center for Internet Security, 2010.

[CPA12] R. Contu, L. Pingree, and E. Ahlm. Predicts 2013: Security Solutions. Technical report, Gartner, Nov 2012.

[ENI12] Critical Cloud Computing - A CIIP perspective on cloud computing services, Version 1.0. Technical report, European Network and Information Security Agency (ENISA), Dec 2012.

[Jul09] K. Julisch. A Unifying Theory of Security Metrics with Applications. Technical report, IBM Research – Zurich, 2009.

[KNW13] T. Kunz, P. Niehues, and U. Waldmann. Technische Unterstützung von Audits bei Cloud-Betreibern. *Datenschutz und Datensicherheit – DuD*, 37(8):521–525, 2013.

[KSW14] T. Kunz, A. Selzer, and U. Waldmann. Automatic Data Protection Certicates for Cloud-Services based on Secure Logging. In *WTC 2013 : Wissenschaftliche Ergebnisse des Trusted Cloud Technologieprogramms*, 2014.

[LMPS13] P. Lipton, S. Moser, D. Palma, and T. Spatzier. Topology and Orchestration Specification for Cloud Applications (TOSCA),Version 1.0, Mar 2013.

[Pro12] T. Probst. Generische Schutzmanahmen für Datenschutz-Schutzziele. *Datenschutz und Datensicherheit - DuD*, 36(6):439–444, 2012.

[Sel13] A. Selzer. Die Kontrollpflicht nach §11 Abs. 2 Satz 4 BDSG im Zeitalter des Cloud Computing – Alternativen zur Vor-Ort-Kontrolle des Auftragnehmers durch den Auftraggeber. *Datenschutz und Datensicherheit – DuD*, 4:215–219, 2013.

[SLS13] S. Schneider, J. Lansing, and A. Sunyaev. Empfehlungen zur Gestaltung von Cloud-Service-Zertifizierungen. *Industrie Management – Zeitschrift für industrielle Geschäftsprozesse*, 4:13–17, 2013.

[WWB+13] T. Waizenegger, M. Wieland, T. Binz, U. Breitenbücher, and F. Leymann. Towards a Policy-Framework for the Deployment and Management of Cloud Services. In *SECURWARE 2013 : The Seventh International Conference on Emerging Security Information, Systems and Technologies*, 2013.

Sicheres Cloud Computing im Gesundheitswesen

Alexander Grzesik [a,1], Torsten Frank [a,1]
[a] medisite Systemhaus GmbH

Zusammenfassung. Die Nutzung der Cloud zum Austausch von Daten und dem Einsatz von Programmen hat in den vergangenen Jahren zunehmend an Bedeutung gewonnen. Dies erstreckt sich sowohl über den privaten Bereich, wird in zunehmendem Maße aber auch im wirtschaftlichen Umfeld beobachtet und teilweise bereits produktiv eingesetzt. Dem Einsatz der Cloud im Gesundheitswesen sind dabei im besonderen Maße Einschränkungen und Vorgaben gesetzt, da hier rechtliche Aspekte wie Datenschutz und -sicherheit gesondert beachtet werden müssen. Im Rahmen der Trusted Cloud Initiative [1] des Bundesministeriums für Wirtschaft und Energie (BMWi) soll diesem Umstand mit der gezielten Förderung von drei Cloud-Forschungsprojekten im Gesundheitswesen Rechnung getragen werden. Das Projekt TRESOR – Trusted Ecosystem for Standardized and Open cloud-based Resources – verfolgt dabei als eines der geförderten Projekte den Anspruch, eine offene Cloud-Plattform für das Gesundheitswesen zu schaffen, welche von Beginn an zusammen mit klinischen Anwendern konzipiert wurde und die besonderen Gegebenheiten im Gesundheitswesen berücksichtigt [2].

Keywords. Gesundheitswesen, Cloud, PaaS, Datenschutz, Datensicherheit

Einleitung

Das Internet ist als Bestandteil des Alltags nicht mehr wegzudenken. Wurde es im vergangenen Jahrzehnt primär als Kommunikationsmedium verwendet, gewinnt seit kurzer Zeit zunehmend die Nutzung der durch das Internet nutzbaren Ressourcen an Bedeutung. Insbesondere die Verfügbarkeit von Rechenkapazität und Speicherplatz im Internet hat, verbunden mit zunehmenden Verbindungsgeschwindigkeiten, eine völlig neue Möglichkeit der Nutzung desselben geschaffen. In diesem Zusammenhang wird zunehmend der Begriff der „Cloud“ geprägt, welche eine Kombination der genannten Ressourcen als Infrastruktur im Netz zur Verfügung stellt.

Die Nutzung von Cloud-Infrastrukturen hat sich in den vergangenen Jahren stark verbreitet. Angefangen im privaten Bereich mit Programmen wie Dropbox™ oder den Cloud-Diensten der Google Docs™, bis hin zur Nutzung von Cloud-Infrastrukturen bei rechenintensiven Aufgaben bspw. im universitären Bereich, gewinnt die Cloud zunehmend an Bedeutung und bietet Privatpersonen und vor allem aber auch Firmen und Institutionen eine Alternative gegenüber dem eigenen Vorhalten derartiger Strukturen. Der traditionelle „Server im Keller“ wird zunehmend durch netzbasierte Infrastruktur abgelöst, wobei Faktoren wie Skalierbarkeit und Datenbackup wichtige Kriterien und Argumente sind, eine derartige Struktur zu nutzen.

1 Corresponding Author.

Das Gesundheitswesen steht dezentraler Datenhaltung oder allein dem Transport von Daten über das Internet traditionell eher distanziert gegenüber. In Anbetracht des hohen und schützenswerten Guts der Patientendaten ist dies durchaus auch verständlich, zumal das Patientengeheimnis von hoher Bedeutung im Vertrauensverhältnis zwischen Arzt und Patient ist. Nichtsdestotrotz weisen aber das Internet und vor allem die Möglichkeiten, welche die Nutzung von Cloud-Infrastrukturen bietet, auch wiederum Vorteile auf, deren Nutzung auch im Gesundheitswesen zunehmend erwogen wird.

Das Bundesministerium für Wirtschaft und Energie (BMWi) verfolgt mit der Förderung von Cloud-Forschungsprojekten im Rahmen des Technologieprogramms Trusted Cloud das Ziel der Entwicklung und Erprobung innovativer, sicherer und rechtskonformer Cloud Computing-Lösungen. Von diesen neuen, cloud-basierten Diensten sollen insbesondere mittelständische Unternehmen profitieren. Es werden grundlegende Technologien sowie Cloud-Anwendungen für unterschiedliche Bereiche entwickelt, insbesondere für Industrie, Handwerk, Gesundheit und den öffentlichen Sektor. Die Forschungs- und Entwicklungsaktivitäten haben im September 2011 begonnen und werden bis Anfang 2015 laufen.

Die TRESOR Architektur

Das Forschungsprojekt TRESOR hat zum Ziel, ein sicheres und datenschutzkonformes Cloud-Ecosystem aufzubauen. Mit dem speziellen Fokus auf das Gesundheitssystem adressiert TRESOR einen Anwendungsbereich, in dem Datenschutz und Datensicherheit von hoher Bedeutung sind und wo vor allem auf gesetzlicher Ebene hohe Hürden und Vorgaben für den Einsatz von Technologie und Telekommunikation gegeben sind [3].

TRESOR konzentriert sich exemplarisch auf Gesundheitseinrichtungen wie etwa Krankenhäuser oder Ärzte aber auch mittelständische und Industrieunternehmen aus dem Gesundheitswesen, da diese Zielgruppen besonders viele Charakteristiken für eine „Trusted Cloud“ aufweisen. So fordert dieser Bereich neben den sehr hohen Standards bezüglich Datenschutz und Datensicherheit aufgrund der Vielzahl der beteiligten Akteure zudem hohe Anforderungen an Interoperabilität, Skalierbarkeit und Verfügbarkeit. Um diesen Anforderungen Rechnung zu tragen, kombiniert die TRESOR Architektur modernste Technologien in einer Infrastruktur, die hohe Skalierbarkeit und Verfügbarkeit mit geprüften Sicherheitsverfahren und Software-Architekturen vereint.

Überblick über die TRESOR Architektur

Die TRESOR Architektur setzt sich aus vier Hauptkomponenten zusammen:

- Den Schnittstellen für die TRESOR Dienstnutzer
- Dem TRESOR Ecosystem, welches die TRESOR Dienste vermittelt
- Den TRESOR Dienstanbietern, welche ihre Dienste über die TRESOR Plattform den Nutzern zur Verfügung stellen
- Dem IaaS-Anbieter, welcher die Infrastruktur für die TRESOR Plattform zur Verfügung stellt

Aufgrund der modularen Struktur des TRESOR Projektes ist gewährleistet, dass Hauptkomponenten dynamisch gewechselt werden können. Dies kann beispielsweise bedeuten, dass entsprechend Länder-spezifischer Vorgaben verschiedene IaaS-Anbieter die Infrastruktur für Dienstnutzer zur Verfügung stellen. Dies kann ggf. dann notwendig sein, wenn Patientendaten nur innerhalb eines bestimmten Landes gehostet werden dürfen und dafür spezielle Anbieter genutzt werden müssen. Im folgenden Abschnitt werden die einzelnen Komponenten näher beschrieben.

Die Nutzer der TRESOR Dienste im Krankenhaus bemerken gar nicht, dass sie Cloud-Dienste im täglichen Betrieb einsetzen. Dies wird dadurch erreicht, dass die Nutzer diese Dienste nicht direkt aufrufen, sondern diese in die Infrastruktur des Krankenhauses, integriert sind. So sind diese Dienste beispielsweise über Schnittstellen direkt in das Patientendaten-Managementsystem (PDMS) des Krankenhauses integriert. Das PDMS ruft dann im Falle der Nutzung eines Cloud-Dienstes über einen TRESOR-Proxy, diese Dienste auf, wobei Authentifizierung und Autorisierung zur Nutzung des Dienstes automatisch vermittelt werden. So können beispielsweise im Krankenhaus schon vorhandene Benutzerverwaltungssysteme wie ein Active Directory unmittelbar eingebunden werden.

Über den TRESOR Proxy werden nach erfolgter Authentifizierung Cloud-Dienste des TRESOR Ecosystems aufgerufen und die entsprechenden Daten zur Verfügung gestellt, die für die Nutzung der Dienste notwendig sind.

Das TRESOR Ecosystem stellt als zentrale Verwaltungsinfrastruktur alle Komponenten zur Verfügung, die für eine optimale Nutzung der Cloud-Dienste notwendig sind. Dies umfasst zum einen Komponenten für die Verwaltung der Dienste und die Balancierung von deren Auslastung über einen Broker. Zum anderen umfasst das Ecosystem aber auch Komponenten zur Abrechnung der Dienstnutzung sowie Service-Komponenten, die beispielsweise Dienste-Entwicklern Infrastrukturen zum Testen oder Deployment von Anwendungen zur Verfügung stellen.

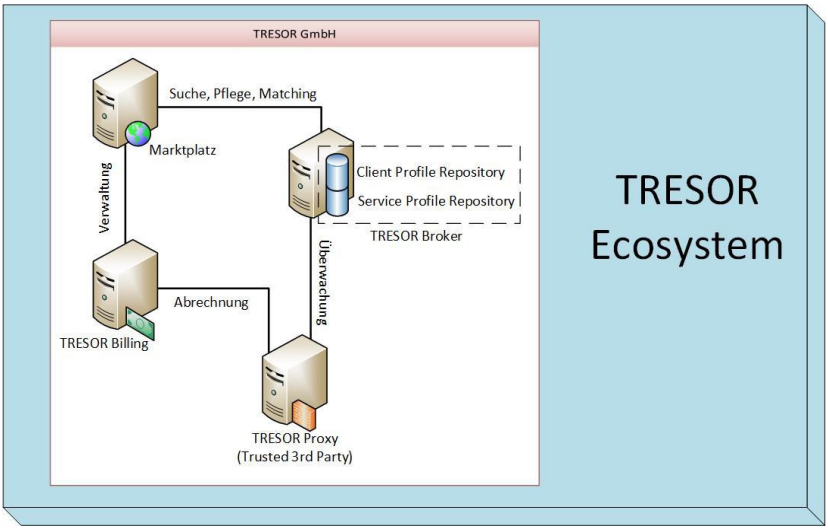


Abbildung 1. Das TRESOR Ecosystem

Ein weiterer Kernbestandteil des TRESOR Ecosystems ist eine neue offene PaaS-Plattform (PaaS – Plattform as a Service) für die Bereitstellung und Nutzung standardisierter cloud-basierter System- und Anwendungsdienste, welche an den Cloud Broker als vertrauenswürdigen Mediator zwischen den Klienten und den Cloud-Anbietern des Ecosystems gekoppelt ist.

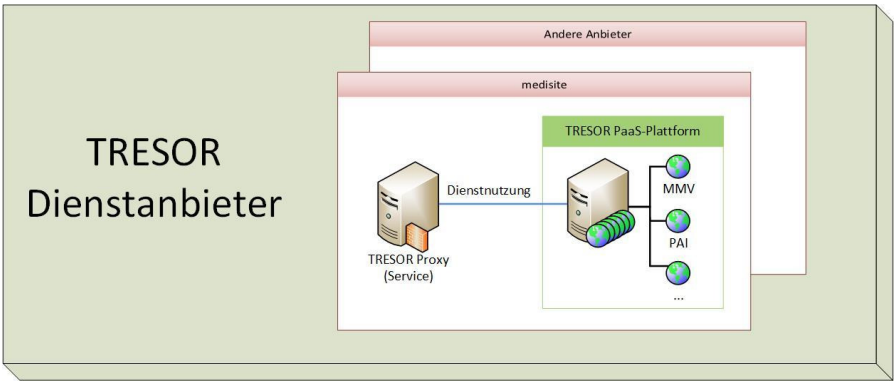


Abbildung 2. TRESOR Dienstanbieter

Die PaaS Plattform stellt die Application Server Infrastruktur bereit, in welche die Cloud-Dienste eingestellt werden. Aufgrund der Nutzung offener Standards für die Dienstentwicklung und der Bereitstellung von Entwicklungswerkzeugen wie

beispielsweise APIs aber auch schon umfassender Dienstekomponenten zum Beispiel für Authentifizierung oder Verschlüsselung, ermöglich die PaaS Plattform ein schnelles Bereitstellen neuer Dienste und senkt die Investitionskosten für Dienstanbieter auf ein Minimum, da diese keine eigene technische Infrastruktur zur Entwicklung der Dienste vorhalten müssen. Durch Bereitstellung von dedizierten Testumgebungen wird Dienstanbietern außerdem die Möglichkeit gegeben, mittels frühzeitiger und ggf. auch automatischer Testprozeduren aktuelle Software-Entwicklungsstandards wie testgetriebene Entwicklung im Rahmen des Software Life Cycles zu nutzen, um Dienste sowohl sicher als auch performant zu gestalten.

Die TRESOR Plattform wird von einem IaaS-Anbieter gehostet. Dieser Anbieter stellt die notwendige Infrastruktur zur Verfügung, welche für die Nutzung der Cloud-Dienste notwendig ist.

Der IaaS-Anbieter stellt neben Rechenleistung und Speicherkapazität auch die notwendige Netzwerkstruktur bereit, um dynamisch skalieren zu können und die Erreichbarkeit, jederzeit gewährleisten zu können. Erfordert ein Cloud-Dienst höhere Rechenleistung, kann auf diese Art jederzeit die Rechenleistung durch dynamisches Hinzufügen von Rechenknoten erhöht werden, um eine zügige Bearbeitung der Aufgabe des Dienstes zu ermöglichen.

Vorteile der TRESOR Plattform

Neben den bereits dargelegten allgemeinen Vorteilen einer Cloud-Infrastruktur wie der dynamischen Skalierbarkeit von Rechenleistung, Speicherkapazität und Netzwerkinfrastruktur weist insbesondere die TRESOR Plattform noch eine Reihe weiterer Vorteile auf, die in diesem Kapitel näher betrachtet werden sollen.

Sicherheit

Der Schutz von Daten während der Speicherung oder dem Transport über ein Netzwerk ist von hoher Wichtigkeit, und das nicht nur, aber vor allem auch im Gesundheitswesen. Neben den rein rechtlichen Aspekten des Schutzes von Patientendaten sind einem persönlich aber auch die Bedenken vertraut, was eigentlich mit den durch den Arzt bei einer Untersuchung oder der Analyse im Labor erhobenen Daten geschieht und wer Zugriff auf diese hat. Allein die Diskussion um die Einführung der Gesundheitskarte in Deutschland hat sehr deutlich den schmalen Grad in der Balance zwischen Datenschutz und einer effektiven Nutzung und Verfügbarkeit der Patientendaten gezeigt.
Von dieser Diskussion ausgehend stellt sich durchaus die Frage, wie man Daten bei der Nutzung von Cloud-Diensten optimal schützen kann, sodass kein Dritter darauf Zugriff erhält. Diesbezüglich werden Maßnahmen sowohl auf technischer, als auch auf organisatorischer Ebene getroffen.

Auf technischer Seite ist vor allem die Sicherheit der Server und Datenbanken zu gewährleisten. Dies umfasst zum einen die Absicherung der Server und Datenbanken durch aktuell gehaltene (‚gepatchte') Systeme, in welchen Updates zeitnah eingespielt

werden und wo Firewalls und Virenscanner den Zugriff auf Systeme und Daten verhindern oder zumindest erschweren. Ferner wird auf organisatorische Ebene auch der physische Zugriff auf die Infrastruktur kontrolliert. Dies umfasst ein klares Berechtigungskonzept, wer beispielsweise den Serverraum betreten darf, aber auch, wer die Berechtigung zu administrativen Tätigkeiten auf den Servern besitzt.

Auf Seiten der Cloud-Dienste sollten bei der Entwicklung derselben grundlegende Konzepte wie das Prinzip der Datensparsamkeit und die Nutzung von Sicherungsmöglichkeiten der verwendeten Programmiersprache berücksichtigt werden. Dem Prinzip der Datensparsamkeit kommt dabei eine besondere Sorgfaltspflicht zu, welche die Betreiber der TRESOR Plattform in besonderem Maße wahrnehmen. Durch diese wird bei Diensten, welche auf der Plattform zur Verfügung gestellt werden, regelmäßig überprüft, welche Daten zu welchem Zweck genutzt werden.

Einer der wichtigsten Aspekte in Bezug auf Datensicherheit ist allerdings deren Verschlüsselung. Diese soll im Falle, dass Daten doch in falsche Hände geraten, sicherstellen, dass niemand etwas mit den Daten anfangen oder aus diesen Rückschlüsse zum Beispiel auf bestimmte Patienten und deren Krankheiten oder Behandlungen ermöglichen. Um dies zu gewährleisten, werden in TRESOR die Daten vom Client bis zum Server durchgehend verschlüsselt übertragen. Dies umfasst zum einen die medizinischen Daten aber auch Daten, die beispielsweise für die Authentifizierung der Benutzer verwendet werden. Für die Verschlüsselung werden dabei mehrere Verfahren zur Verfügung stehen. Dies ermöglicht es, aufgrund der Vorgaben eines Dienstenutzers (z.B. entsprechend den gültigen Policies in einem Krankenhaus) bestimmte Verschlüsselungsverfahren zum Sichern der Daten zu verwenden.

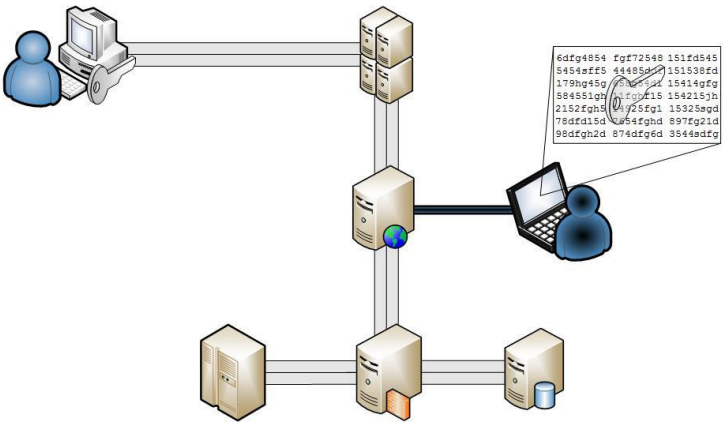


Abbildung 3. Durchgehende Verschlüsselung

Redundanz

Durch redundante Serverinfrastrukturen in verteilten Rechenzentren ist bei TRESOR jederzeit gewährleistet, dass sowohl gespeicherte Daten als auch Dienste zur Verfügung stehen und dies, auch wenn eines oder mehrere Rechenzentren ausfallen.

Gerade im medizinischen Bereich ist es wichtig, jederzeit auf notwendige Daten eines Patienten oder einer Behandlung zugreifen zu können.

Ein weiterer Vorteil einer redundanten Infrastruktur ist die Möglichkeit, Aktualisierungen von Diensten so zu gestalten, dass der Dienstnutzer davon in der Regel nicht beeinflusst ist. In einer redundanten Cloud-Infrastruktur kann die Aktualisierung knotenweise erfolgen, so dass es für den Benutzer keine Down-Zeiten gibt, in welchen er Dienste oder Daten nicht erreichen kann.

Lokalisierung von Daten und Diensten

Bei der Cloud Nutzung stellt sich oft die Frage, wo Daten eigentlich gerade gespeichert werden oder über welche Länder und Standorte diese versandt werden. Insbesondere im Zusammenhang mit medizinischen Daten gelten gesonderte Anforderungen. Entsprechend den Empfehlungen der Europäischen Gemeinschaft [4] und den gesetzlichen Vorgaben über den Datentransfer in Drittstaaten [5] kommt den Standorten von Servern und Datenbanken eine besondere Bedeutung im Rahmen von Cloud-Projekten zu.

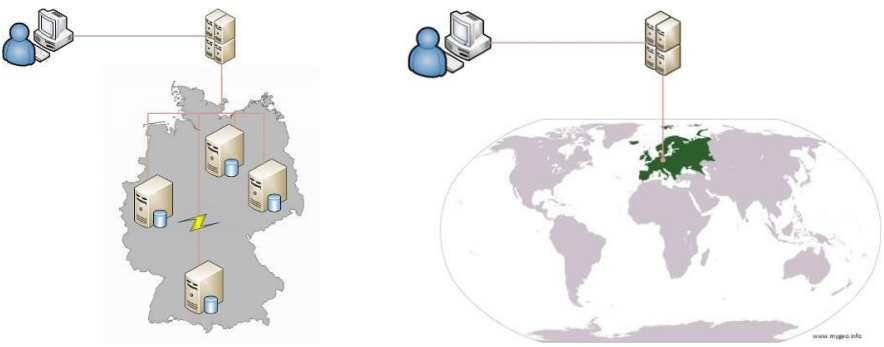


Abbildung 4. Redundante Serverstandorte entsprechend den Anforderungen

Das TRESOR Projekt setzt diese Anforderungen von Beginn an um. Der modulare Aufbau der Projekt-Infrastruktur ermöglicht flexible Lösungen, wenn beispielsweise in bestimmten Ländern restriktivere Anforderungen bezüglich der Datenhaltung bestehen. Zum Beispiel im Falle der Schweiz, wo Patientendaten das Land nur verlassen dürfen, wenn die Datenschutzgesetze des Ziellandes mit denen der Schweiz vergleichbar sind, was eine Übertragung von Patientendaten beispielsweise in die USA nur sehr eingeschränkt erlaubt [6,7].

Anwendungsfälle

Die Umsetzbarkeit der entwickelten Lösungen des Cloud-Ecosystems wird anhand von Anwendungsszenarien aus dem Bereich der Patientenversorgung demonstriert.

Die medienbruchfreie medizinische Verlaufsdokumentation (MMV) ist von hoher Bedeutung insbesondere bei der Verlegung von Patienten zwischen Krankenhäusern. An dieser Stelle bietet ein Cloud-Service zahlreiche Möglichkeiten zur Einsichtnahme von Daten und zum Anpassen (Matching) der Daten von unterschiedlichen Mandanten. Die Nutzung der Cloud bietet hier insbesondere neue Möglichkeiten des Zugriffs auf Daten. So kann im Empfängerkrankenhaus ein Arzt über den Cloud-Dienst Zugriff auf den medizinischen Verlauf des Patienten im anderen Krankenhaus nehmen, um für eine optimale Weiterführung der Behandlung des Patienten zu sorgen. Dies geschieht natürlich nur, wenn die Daten freigegeben wurden und der Arzt bzw. die Pflegekraft die entsprechende Berechtigung zum Zugriff auf diese besitzt.

Beim pharmakologischen Arzneimittel-Interaktionscheck (PAI) wird mittels eines Cloud-Diensts Zugriff auf eine Arzneimittel-Interaktionsdatenbank genommen, um zu überprüfen, ob Wechselwirkungen zwischen angeordneten Medikamenten vorliegen. Ein Vorteil eines Cloud-Dienstes ist hier die Möglichkeit des Zugriffs auf tagesaktuelle Daten, die von den Datenbankbetreibern zur Verfügung gestellt wurden und bisher manuell in den Datenbestand des Krankenhauses eingepflegt werden müssen.

Zusammenfassung

Mit den beteiligten Projektpartner - der medisite Systemhaus GmbH, der T-Systems International GmbH, der bitplaces GmbH, der TU-Berlin, dem Deutschen Herzzentrum Berlin und dem Paulinenkrankenhaus Berlin - stellt TRESOR ein interdisziplinäres Kooperationsprojekt dar. Es wird erstmals eine "Trusted Cloud" - Infrastruktur bereitstellt, die in der Lage ist, alle relevanten gesetzlichen Vorschriften, Sicherheits- und Datenschutzrichtlinien sowie individuelle Richtlinien mittelständischer Unternehmen und des öffentlichen Sektors zu berücksichtigen. Die **exemplarische** Umsetzung für den Anwendungsbereich der Patientenversorgung wird zukunftsweisend für die weitere Entwicklung von Cloud Services im Gesundheitswesen sein.

Referenzen

[1] Trusted Cloud Projekt des BMWi, http://www.trusted-cloud.de/ (07.05.2013)
[2] TRESOR Homepage des BMWi, http://www.trusted-cloud.de/de/1690.php (07.05.2013)
[3] Patientenrechtegesetz §630 BGB
[4] Richtlinie zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten und zum freien Datenverkehr 95/46/EG
[5] §§ 4b, 4c BDSG
[6] Art. 6 DSG CH
[7] Leitfaden für die Bearbeitung von Personendaten im medizinischen Bereich, Eidgenössischer Datenschutz- und Öffentlichkeitsbeauftragter (EDÖB)

Gamification und Virtuelle Welten

(GameWorlds)

Workshop „Gamification und Virtuelle Welten“

Leitung:

Prof. Dr. Stefan Stieglitz (Universität Münster)
Dr. Danny Pannicke (Technische Universität Berlin)
Prof. Dr. Susanne Robra-Bissantz (Technische Universität Braunschweig)
Prof. Dr. Kai-Ingo Voigt (Universität Erlangen-Nürnberg)
Prof. Dr. Rüdiger Zarnekow (Technische Universität Berlin)

Inhalt:

Der hedonistische Aspekt der Computernutzung gewinnt zunehmend an Bedeutung. Mit der Verbreitung des Internets, unterstützt durch mobile Technologien, sind Computer im letzten Jahrzehnt zum Multimedium geworden und durchdringen bzw. transformieren nahezu alle Bereiche des täglichen Lebens. Unterhaltung, kreative Selbstverwirklichung und freudvolle Interaktion bilden zentrale Aspekte der Nutzung. Das wesentliche Merkmal hedonistischer Informationssysteme besteht darin, dass die Interaktionen mit diesen Systemen sich – zumindest zum Teil – aus sich selbst heraus rechtfertigen. Neben der privaten Nutzung kommen dabei zunehmend auch Anwendungen im Business-Bereich in den Blick. In diesem Zusammenhang wird aktuell insbesondere diskutiert, wie spieltypische Elemente und Prozesse in spielfremden Kontexten nutzbar sind („Gamification“) und wie Spiele zur Klärung ernsthafter Fragestellungen beitragen können („Serious Games“). Viele Organisationen sind dabei, die Potenziale des Spielerischen zu entdecken und versprechen sich davon motivationale und kreativitätsfördernde Effekte.

Der Workshop auf der Informatik 2014 soll der Konsolidierung der bisherigen Forschung im deutschsprachigen Raum dienen und einen Ausgangspunkt für die weitere systematische Bearbeitung dieses Themenfeldes markieren. Die einführende Keynote (nicht im Tagungsband enthalten) von Herrn Simon Hassannia (Deutsche Telekom AG) ist dem Thema „Virtuelle Plattformen und Gamification im Unternehmen“ gewidmet. Die sieben angenommenen Beiträge adressieren verschiedene wichtige Aspekte des eingegrenzten Forschungsfeldes wie den Einsatz von Serious Games im Kontext organisationaler Fragestellungen, Methoden zur Auswertung der in Spielen entstehenden Daten und Spiele im Kontext von Ideenwettbewerben.

Ein zentrales Anliegen des Workshops ist es, interessierte Wissenschaftler und Praktiker zusammenzubringen und ein Forum für den weiteren Austausch zu schaffen.

Online Ideation Games as Means to Change

Christian W. Scheiner

Lehrstuhl für Industrielles Management
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20
90403 Nürnberg
christian.scheiner@fau.de

Abstract: Companies have to adapt to change if they want to stay competitive. Inducing and managing change processes are, however, highly complex and difficult tasks, especially when individual beliefs, knowledge, and routines are addressed. Despite the importance of this aspect, research on concrete measures is still underrepresented. Within the present study, an online ideation game is examined as a means to change. The aim of this study is therefore to explore whether participation in an online ideation game induce change by fostering learning and unlearning. A longitudinal, quantitative approach is chosen to detect changes over time. An adjusted opinion leadership scale of Childers [Ch86] is applied as a measure of change on an individual level. The findings of this study demonstrate that an online ideation game can induce change processes.

1 Introduction

Most of the time, companies operate in a competitive environment, which is stable and clearly understood. As a consequence, organizational competences and resources have been aligned over time to this environment. This ensures competitiveness and a company's survival. Even if Dewald and Bowen [DB10] are noticing a shortening of stability in this environment, periods of stability still last relatively long and are characterized by incremental changes and innovations. This in turn further determines and shapes organizational routines and processes [MF80, RT94]. Revolutionary changes, however, punctuate these periods of stability and create a new competitive framework for companies [RT94]. Factors contributing to fundamental changes can be found within and outside of organizational boundaries. Examples of internal factors are structural changes in the organization, or changes in the composition of the top management level [RT94]. Technology and market turbulence illustrate examples of external factors urging companies to reconsider and change their competitive position, their competencies, their technologies, and their routines. Technological turbulence describes, therein, the rate of change in which technologies are improved or are replaced by new technologies [JK93]. Market turbulence reflects changes in the preferences of buyers, wide-ranging wants and

needs, changing buyer structure, and a need to offer constantly new products [HH04, JK93]. In industries where market turbulence is high, companies are subsequently challenged to be highly innovative and to introduce new products under uncertain conditions. A company can respond adequately to changes through its capability to learn and to further develop upon existing knowledge. A substantial body of research has, therefore, been dedicated to learning in organizations and organizational learning in a variety of aspects [e.g. Ma91, LM88a, KA07, ALB03, LL05]. Kane and Alavi [KA07] examined, for instance, the influence of information technology in the context of organizational learning, and found that IT-enabled learning mechanisms influence organizational learning positively. Adjusting to change, however, is not solely accomplished by continuing and improving the status quo. Companies have to reconsider taken-for-granted assumptions of how to conduct business, and to abandon identified impediments to change. Dewald and Bowen [DB10] state in this context that whether "small or large, it is difficult for any organization invested in 'old ways' to abandon those known ways" (p. 198). Other authors have used terms such as core rigidities [Le92], incumbent inertia [LM88b, TG10], or competency traps [LM88a] to describe this phenomenon. Bettis and Prahalad [BP96] explain this difficulty with the robustness of existing dominant logics and routines. Routines comprise, according to Levitt and March [LM88a], procedures, forms, rules, conventions, technologies, strategies, beliefs, frameworks, paradigms, cultures, knowledge, and codes, and are thus the backbone and the essence of an organization. Tsang and Zahra [TZ08] argue, therefore, in favor of removing these routines and ways of thinking in order to enable change. The ability to unlearn plays a pivotal role in this context. Pratt and Barnett [PB97] define unlearning as "a process of discarding obsolete and misleading knowledge that is stimulated by opposing emotions which enable new responses and mental maps to develop" (p. 82). Starbuck [St96] draws attention to the individual in this process and sees unlearning as "a process that shows people they should no longer rely on their current beliefs and methods" (p. 727). These definitions have one characteristic in common: they characterize unlearning as an intended, actively initiated and conducted process, which distinguishes it from its meaning in colloquial use as a form of forgetting [TZ08].

Insights of how to initiate and achieve learning and unlearning within an organization is, despite its importance, still scarce. Exceptions are the works of Starbuck [St96], Becker et al. [BHA06], and Becker [Be10], who give general recommendations of how to achieve this objective. Starbuck [St96], for example, describes eight viewpoints that can support unlearning, while Becker et al. [BHA06] and Becker [Be10] suggest using training sessions, documented policies and procedures, structures, and information sessions to induce unlearning. Due to its characteristics, online ideation games could enrich the existing toolkit. Although they are mainly used to access the creative potential of people within and outside of an organization, online ideation games have indicated their inherent potential for this purpose (see e.g. [ABM12]). The use of an online ideation game as means to change is, however, still a new field of application. The aim of this study is therefore to explore whether participation in an online ideation game can foster change in individuals. In order to examine change, a longitudinal, quantitative approach has been chosen for this study.

This study is structured as follows: In chapter 2 a definition of online ideation games is given, before the research questions are derived from literature in chapter 3. Chapter 4 describes the sample subject, the data collection and the operationalization of change. The empirical findings are presented in chapter 5. Finally, chapter 6 comprises the discussion, implications, and possibilities for future research.

2 Online Ideation Games

Online ideation games are idea competitions which have been modified in a way that gives them a game-like appeal. Idea competition can be defined, in general, with the help of common characteristics [ABM12, BM10]. In idea competitions an invitation is sent to a private audience, or to the general public, to submit ideas for predefined topics. Submissions can occur in the form of rough ideas, to fully developed solutions [BM10]. Submissions have to be made within a given period of time, which can range from a few hours, to forms without a set end-date [BM10], and are evaluated by experts, the community, or a combination of both groups [PW06, ABM12]. After closing the competition and evaluating contributions, winners are announced.

The modification of idea competitions to an online ideation game is accomplished by the use of game mechanics [SW13] and can be subsumed under the term gamification [De11]. Game mechanics are the constitutive building blocks of games that occur in a structured and predefined way, as they enable the implementation of game characteristics [SW13]. In concordance with definitions of Schell [Sc08] and Fullerton et al. [FSH04], game mechanics are the means to induce actions and behaviors of players towards the goals of a game. Witt et al. [Wi12] argue, therefore, that developers who apply and implement game mechanics to an idea competition, or to another serious context, are not concerned in principle with the fact whether participants actually perceive a serious system as a game, but how the system has to be constructed and designed in order to ensure that participants behave and act in an intended way.

3 Derivation of Research Question

Research on idea competitions and games gives reason to believe that they can play an essential role as means to change. Although idea competitions are, in principle, applied to gain access to knowledge and ideas of employees and people outside of organizational boundaries to solve existing challenges, or to be provided with new impulses for innovations, Adamczyk et al. [ABM12] point to existing research that shows further relevant benefits of the usage of idea competition in the context of this study. Pack et al. [Pa04], for instance, found that 95 percent of participants of an idea competition reported a positive attitude toward the competition subject because of participation. Moreover, those participants had improved in theoretical and practical related areas. Oppliger [Op01] describes another example where participants were able to learn and improve their skills by taking part in an idea competition. Romanello [Ro05] ascribes similar benefits and outcomes to the usage of idea competitions as an integral part in the set of

teaching tools and techniques. What all those examples have in common is that participants are confronted with new information and challenges, and have to adjust to these new ramifications in order to solve given tasks. In order to do so, participants have to be willing to change, and have to accept new information and knowledge. Pardo del Val and Fuentes [PF03] provide a summary of research that shows how change can be impeded by refusal to accept information that is not desired or expected [BSH92, SGH78], by adhering to old assumptions and thoughts [BSH92], or by distorting reality due to unchallenged implicit assumptions [SGH78]. All these findings show that a major obstacle to change lies not within the organizational structures and given organizational procedures, but within employees, and that overcoming these obstacles is highly difficult and complex, as beliefs and opinions have to be addressed. By taking part in an online ideation game, participants work on tasks and deal with predefined topics; therewith, they can gain deeper insights. This can result in a more differentiated view on, and a more profound knowledge towards, a topic.

As an online ideation game is examined in this study, research on games will also need to be considered in order to assess its value as a means to change. Therein, it is necessary to reflect the role of game mechanics due to their initial importance in transforming an idea competition into an online ideation game. Scheiner and Witt [SW13], in their theoretical paper on game mechanics as triggers of motivation and creativity in idea competitions, argue that game mechanics can exert an influence on participants with respect to changing attitudes and behavior in several ways. Game points, which are assigned automatically for a predefined shown behavior, and social points, which are granted by other players, provide immediate feedback to participants. Hence, only such behavior, which is seen as appropriate by the designers and other players, is rewarded with points. Participants can learn, as a consequence, in the form of trial-and-error experiences, what is valued, and can adapt to it [LRV01, Cs90]. The game mechanics level and leaderboards depict in an accumulated way the accomplished and offer by that the opportunity to compare own performance with those of others. Because of the possibility of comparison, participants may adjust to improve their own position in the online ideation game, Moreover, the game mechanic story can foster and induce change within individuals. Story can be used in creating an imaginary frame, enabling people to consider things from a different perspective. This is seen as a crucial element to overcome confined mental corridors [CCD08]. Online ideation games incorporate, in addition, often the game mechanic exchange. Exchange gives participants the opportunity to communicate, support, and collaborate with each other. Hence, participants can learn from each other, and learn how to behave in this specific social group. This provides grounds to become an active part of a community within the online ideation game, and to satisfy the need of social belonging [Su97].

Bearing the previous arguments in mind, participation in an online ideation game could be a valuable means to change. The following research question asks therefore:

Research question: Does participation in an online ideation game lead to change within individuals?

4 Sample Subject, Data Collection, and Operationalization of Change

4.1. Sample Subject

The online ideation game EVOKE was selected as the sample subject. EVOKE was funded by the World Bank Institute, developed by Natron Baxter Applied Gaming, and designed by Jane McGonigal and Kiyash Monsef. It was chosen as a winner of the 2nd Annual Games for Change Award for its excellence in "games for change". Each week, a new challenge was presented. The presented challenges comprised the topics "social innovation", "food security", "power shift", "water crisis", "future of money", "empowering women", "urban resilience", "indigenous knowledge", "crisis networking", and "what happens next". Challenge 1 dealt with social innovation in general, and challenge 10 focused on how to improve the system. Challenge 1 and 10 had subsequently no direct connection to specific social challenges and were therefore excluded from the study. Participants had one week to submit their ideas, before the next challenge was assigned.

Participation in EVOKE required, in general, no previous knowledge towards the given challenges. The inclusion of several sources of information in EVOKE ensured that participants had the necessary knowledge to enable the development of ideas and potential solutions.

4.2 Data Collection

To answer the derived research questions, two independent longitudinal panel studies were conducted. An online questionnaire was used to collect the responses. Each participant logged into the online survey with a unique, randomly created code. This ensured, on the one hand, anonymity for participants. On the other hand, data sets in t1 and t2 of respondents could be identified and matched afterwards. This enabled the longitudinal panel analysis.

Only those participants from the two independent longitudinal panel studies who had completed EVOKE, and had additionally answered all questions in the questionnaires regarding the opinion leadership scale, were included in the statistical analysis . As a consequence, the data of 27 participants from the first and 61 participants from the second study could be used for the empirical analysis. Both studies were consolidated, as the ramifications for participation in the two studies were structured and organized identically. The sample comprised 54 male and 33 female participants. One participant did not state their gender.

4.3 Operationalization of Change

In order to measure, on a longitudinal base, changes within participants of an online ideation game, a form of measurement had to be chosen that allowed the tracking of changes in relevant attributes. The opinion leadership scale illustrated a valuable

approach. Although, opinion leadership has been part of scientific research for many decades, no common understanding exists regarding which measure should be generally used. In this study, change was measured with the opinion leadership scale of Childers [Ch86] (see Table 1).

A necessary prerequisite for the measurement of opinion leadership illustrates that anonymity is ensured, which was given in this study. Opinion leadership was measured for each social challenge before the social challenge was addressed, and after the completion of EVOKE. Opinion leadership belongs to the category of self-report measures. Self-report measures have been criticized in the past, as the accuracy of own-performance-evaluation has been doubted [LLE88]. Kratzer and Lettl [KL09] show, however, that opinion leadership can hardly be examined objectively, and that a broad stream of literature gives evidence to the fact that self-report measures achieve high accuracy levels. They point furthermore to previous findings, where high correlations between self-reports with "'more 'objective' measures" [KL09, p. 652] were found, where self-ratings achieved superior findings than objective ratings, and where errors caused by self-reported techniques were negligible. Additionally, it was not the purpose of this study to identify opinion leaders, but to examine whether the self-assessment to addressed topics had changed. The opinion leadership scale fulfilled this objective by comprising elements focusing on a person's perceived knowledge of a topic, and the perceived status of this person in her/his social environment.

Cronbach's alpha was calculated for all opinion leadership constructs with respect to each social challenge and to each time interval (t1, t2). In all cases, Cronbach's alpha was above the value of .700. Each construct fulfilled, as a consequence, the prerequisite of reliability.

Items
Please rate yourself on the following scales relating to your interactions with friends and neighbors regarding …[1]
1. In general, do you talk to your friends and neighbors about …[1]: Very often (1) (2) (3) (4) never (5)
2. When you talk to your friends and neighbors about …[1] do you: Give a great deal of information (1) (2) (3) (4) give very little information (5)
3. During the past six months, how many people have you told about …[1]? Told a number of people (1) (2) (3) (4) told no one (5)
4. Compared with your circle of friends, how likely are you to be asked about …[1]? Very likely to be asked (1) (2) (3) (4) not at all likely to be asked (5)
5. In discussions of …[1], which of the following happens most often? You tell your friends about …[1] (1) (2) (3) (4) your friends tell you about …[1] (5)
6. Overall, in your discussions with friends and neighbors are you: Often used as a source of advice (1) (2) (3) (4) not used as a source of advice (5)

[1]food security, power shift, water crisis, future of money, empowering women, urban resilience, indigenous knowledge, crisis networking (independently addressed)

Table 1. Adjusted opinion leadership measure

In order to ensure the appropriateness of the chosen statistical examination, data was tested on normal distribution with a Kolmogorov-Smirnov test. Power shift, empowering women, future of money, indigenous knowledge, crisis networking, and water crisis were normally distributed in t1 and t2, while food security and urban resilience deviated in t2 from normal distribution. All categories, which were normally distributed in t1 and t2, were subsequently analyzed with a paired t-test. Food security and urban resilience were analyzed with the nonparametric Wilcoxon test.

5 Findings

Figure 1 shows the general self-report on opinion leadership regarding the specific given social challenges in EVOKE. The self-report among respondents can be divided into two subgroups. In the first subgroup are those challenges in which participants believed themselves to possess a more consolidated knowledge and took, as a consequence, a more active role in their interpersonal relationships. Power shift, and the future of money belong to this group. The second subgroup comprises the topics, empowering women, urban resilience, food security, water crisis, crisis networking, and indigenous knowledge. Participants rated, with respect to these topics, their interpersonal role as tending to be rather passive, and their knowledge subsequently as somewhat unconsolidated and limited.

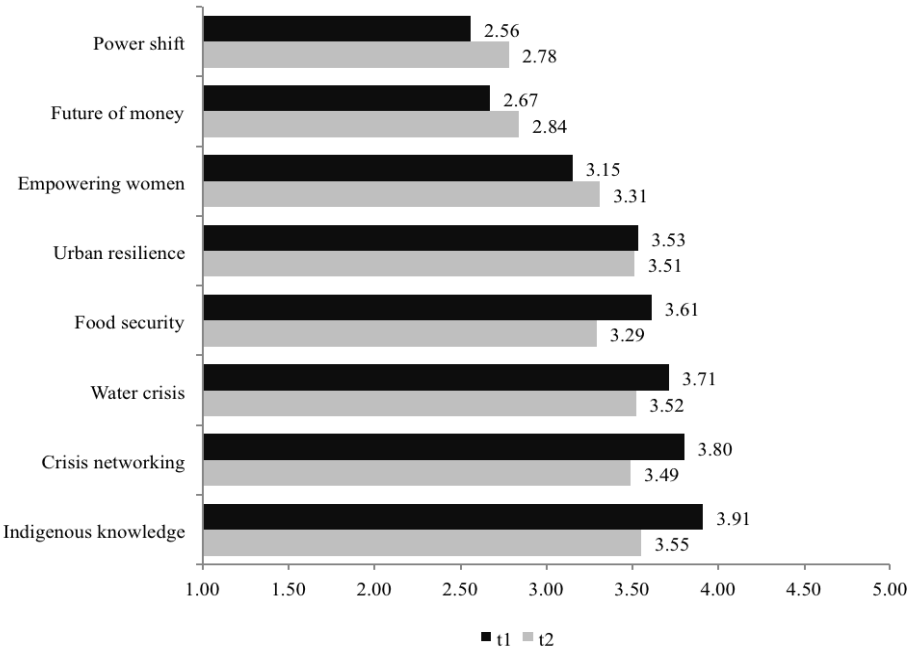


Figure 1: Opinion leadership (N=88) (1= very high level of opinion leadership to 5= very low level of opinion leadership)

Figure 1 also shows that self-report on social challenges were not stable over time, but differed between t1 and t2. To test whether these changes were statistically significant, a paired t-test was conducted for those challenges with a normal distribution, and a Wilcoxon test for those where the underlying data deviated by at least one interval from normal distribution. The findings are shown in Table 2. Apart from urban resilience, all changes were significant to $p<0.05$. Respondents underwent, as a consequence, a major change in their own perception during their participating in EVOKE. The self-assessment towards their own interpersonal role and knowledge with respect to power shift, future of money, and empowering women decreased from t1 to t2, while it increased for all other social challenges, with the exception of urban resilience.

Social challenge	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Power shift	-.22348	1.01202	.10788	-.43791	-.00906	-2.072	87	.041*
Empowering women	-.16477	.69641	.07424	-.31233	-.01722	-2.220	87	.029*
Future of money	-.16667	.78092	.08325	-.33213	-.00121	-2.002	87	.048*
Indigenous knowledge	.36553	.95888	.10222	.16236	.56870	3.576	87	.001*
Crisis networking	.30682	.93324	.09948	.10908	.50455	3.084	87	.003*
Water crisis	.19886	.69106	.07367	.05244	.34528	2.699	87	.008*

		Negative Ranks	Positive Ranks	Ties	Total	Z	Asymp. sig. (2-tailed)
Food security		53	26	9	88	-3.900	.000*
	Mean rank	44.81	30.19				
	Sum of ranks	2375.00	785.00				
Urban resilience		43	38	7	88	-.465	.642
	Mean rank	40.91	41.11				
	Sum of ranks	1759.00	1562.00				

*sign. $p<0.05$

Table 2. Paired t-test and Wilcoxon test

6 Discussion, Implications, and Possibilities for Future Research

6.1 Discussion and Implications

The findings of this study show that changes in the self-perception of opinion leadership occur in almost all topics covered. An exception is urban resilience, where no significant change occurs. The self-assessment towards power shift, future of money, and empowering women decreases and increases towards in food security, water crisis, crisis

networking, and indigenous knowledge. Concerning food security, water crisis, crisis networking, and indigenous knowledge, participants see themselves, subsequently, as more proficient and state to take a more active role in conversations. Regarding power shift, future of money, and empowering women, the opposite effect is given. Here participants perceive themselves as less knowledgeable and less proficient than in the beginning.

It is important to emphasize that the direction of change is not one-sided, but is given in both directions. This means that participation in this online ideation game seems to have induced change by fostering learning and unlearning processes simultaneously. Participants seem to have questioned their existing knowledge base in those topics where they felt they possessed a reasonable knowledge before participating in the online ideation game. They probably realized that they were not as competent as they thought. At the same time, participants see themselves, after the completion of the online ideation game, as more competent in topics where they assessed their knowledge to be rather limited at the beginning.

The findings of this study emphasize the value of an online ideation game as a means to change, as it indicates that deeply rooted, cognitive obstacles and beliefs can be challenged and addressed. Without such a possibility of influence, people tend to stick to existing and developed beliefs, and to ignore or interpret information in a way that is aligned to existing knowledge and beliefs [BSH92, SGH78, PF03]. This is especially given for unlearning, which can be achieved solely when people firmly believe that existing knowledge is inappropriate [BSH92, PF03]. The findings of this study offer, hence, an important managerial implication: Online ideation games seem to be a promising and valuable instrument that can extend the organizational toolkit in the context of continuous and fundamental change.

6.2 Possibilities for Future Research

Despite the value of the findings, this study possesses limitations which offer possibilities for future research. The limitations of this study arise from the chosen sample, the chosen methodology, and the content of addressed challenges.

Concerning the sample of this study, only students have been tested. Future research has to analyze whether the identified effect can also be found in an organizational setting, where the age structure differs to this study. Concerning the methodology, the study did not examine the reasons that determined change within the group of participants, but analyzed instead whether changes occurred over time. Future research could, subsequently, open up this black box by choosing a cross-sectional approach in contrast to the longitudinal panel study. Regarding the content of given challenges, only social topics were addressed within the chosen online ideation game. Within profit-oriented organizations, the structure and composition of existing challenges are different. Future research could investigate, therefore, whether similar results can be achieved in a setting with different challenges.

References

[ABM12] Adamczyk, S.; Bullinger, A. C.; Moeslein, K. M.: Innovation contests: a review, classification and outlook. Creativity and Innovation Management, 21 (4), 2012, 335-360.

[ALB03] Akgün, A. E.; Lynn, G. S.; Byrne, J. C.: Organizational learning: a socio-cognitive framework. Human Relations, 56 (7), 2003; pp. 839-868.

[Be10] Becker, K.: Facilitating unlearning during implementation of new technology. Journal of Organizational Change Management, 23 (3), 2010; pp. 251-268.

[BHA06] Becker, K.; Hyland, P.; Acutt, B.: Considering unlearning in HRD practices: an Australian study. Journal of European Industrial Training, 30 (8), 2006; pp. 608-621.

[BM10] Bullinger, A. C.; Moeslein, K.: Innovation contests – where are we? In Proceedings of the Sixteenth Americas Conference on Information Systems, 2010; pp. 1-8. Peru. Retrieved from http://aisel.aisnet.org/amcis2010/28

[BP96] Bettis, R. A.; Prahalad, C. K.: The dominant logic: retrospective and extension. Strategic Management Journal, 16 (1), 1996; pp. 5-14.

[BSH92] Barr, P. S.; Stimpert, J. L.; Huff, A. S.: Cognitive change, strategic action, and organizational renewal. Strategic Management Journal, 13 (3), 1992; pp. 15-36.

[CCD08] Coyne, K. P.; Clifford, P. G.; Dye, R.: Thinking from inside the box. Harvard Business Review, 85 (12), 2008; pp. 71-78.

[Ch86] Childers, T.: Assessment of the psychometric properties of an opinion leadership scale. Journal of Marketing Research, 23(6), 1986; pp. 184-188.

[Cs90] Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper and Row, New York, 1990.

[DB10] Dewald, J.; Bowen, F.: Storm clouds and silver linings: responding to disruptive innovations through cognitive resilience. Entrepreneurship Theory and Practice, 34 (1), 2010; pp. 197-218.

[De11] Deterding, S. et al.: Gamification: toward a definition. In CHI 2011 Gamification Workshop Proceedings, Vancouver, BC, Canada; pp. 12-15, 2011.

[FSH04] Fullerton, T.; Swain, C.; Hoffman, S.: Game Design Workshop. Designing, Prototyping and Playtesting Games. CMP Books, San Francisco, 2004.

[HH04] Hult, G.; Hurley, R.: Innovativeness: Its antecedents and impact on business performance. Industrial Marketing Management, 33 (5), 2004; pp. 429-438.

[JK93] Jaworski, B. J.; Kohli, A. K.: Market orientation: antecedents and consequences. The Journal of Marketing, 57 (3), 1993; pp. 53-70.

[KA07] Kane, G. C.; Alavi, M.: Information technology and organizational learning: an investigation of exploration and exploitation processes. Organization Science, 18 (5), 2007; pp. 796-812.

[KL09] Kratzer, J.; Lettl, C.: Distinctive roles of lead users and opinion leaders in the social networks of schoolchildren. Journal of Consumer Research, 36 (4), 2009; pp. 646-659.

[Le92] Leonard-Barton, D.: Core capabilities and core rigidities: A paradox in managing new product development. Strategic Management Journal, 13 (S1), 1992; pp. 111-125.

[LL05] Lumpkin, G. T.; Lichtenstein, B. B.: The role of organizational learning in the opportunity- recognition process. Entrepreneurship: Theory & Practice, 29 (4), 2005; pp. 451-472.

[LM88a] Levitt, B.; March, J. G.: Organizational learning. Annual Review of Sociology, 14, 1988; pp. 319-340.

[LM88b] Lieberman, M. B.; Montgomery, D. B.: First-mover advantages. Strategic Management Journal, 9 (S1), 1988; pp. 41-58.

[LLE88] Locke, E. A.; Latham, G. P.; Erez, M.: The determinants of goal commitment. Academy of Management Review, 13 (1), 1988; pp. 23-39.

[LRV01] Linder, M. -O.; Roos, J.; Victor, B.: Play in organizations. Working Paper 2, Imagination Lab Foundation, Lausanne, 2001.

[Ma91] March, J. G.: Exploration and exploitation in organizational learning. Organization Science, 2 (1), 1991; pp. 71-87.

[MF80] Miller, D.; Friesen, P. H.: Momentum and revolution in organizational adaptation. Academy of Management Journal, 23 (4), 1980; pp. 591-614.

[Op01] Oppliger, D. E.: University – Pre-college interaction through FIRST Robotics competition. Paper presented at the International Conference on Engineering Education, Oslo, 2001.
[Pa04] Pack, D. J. et al..: Fire-fighting mobile robotics and interdisciplinary design-comparative perspectives. IEEE Transactions on Education, 47 (3), 2004; pp. 369-376.
[PB97] Pratt, M. G.; Barnett, C. K.: Emotions and unlearning in Amway recruiting techniques: Promoting change through 'safe' ambivalence. Management Learning 28 (1), 1997; pp. 65-88.
[PF03] Pardo del Val, M. ; Fuentes, C. M.: Resistance to change: a literature review and empirical study. Management Decision, 41 (2), 2003; pp. 148-155.
[PW06] Piller, F. T.; Walcher, D.: Toolkits for Idea Competitions: A Novel Method to Integrate Users in New Product Development. R&D Management, 36 (3), 2006; pp. 307-318.
[Ro05] Romanello, T.: Collaborative competition? A great way to teach and motivate. The Physics Teacher, 43, 2005; pp. 76-78.
[RT94] Romanelli, E.; Tushman, M. L.: Organizational transformation as punctuated equilibrium: an empirical test. The Academy of Management Journal, 37 (5), 1994; pp. 1141-1166.
[Sc08] Schell, J.: The art of game design. Burlington: Morgan Kaufmann Publishers, 2008.
[SGH78] Starbuck, W.; Greve, A.; Hedberg, B.: Responding to crisis. Journal of Business Administration, 9 (2), 1978; pp. 111-137.
[St96] Starbuck, W.H.: Unlearning ineffective or obsolete technologies. International Journal of Technology Management, 11 (7–8), 1996; pp. 725-737.
[Su97] Sutton-Smith, B.: The ambiguity of play. Harvard University Press, Cambridge, 1997.
[SW13] Scheiner, C. W.; Witt, M.: The Backbone of gamification – A theoretical consideration of play and game mechanics. In Horbach, M. (Ed.): Informatik 2013 – Informatik angepasst an Mensch, Organisation und Umwelt, Lecture Notes in Informatics (LNI) – Proceedings, Series of the Gesellschaft fuer Informatik, Vol. 220, 2013; pp. 2372-2386.
[TG10] Tripsas, M.; Gavetti, G.: Capabilities, cognition, and inertia: evidence from digital imaging. Strategic Management Journal, 21 (10), 2010; pp. 1147-1161.
[TZ08] Tsang, E. W. K.; Zhara, S. A.: Organizational unlearning. Human Relations, 61 (10), 2008; pp. 1435-1462.
[Wi12] Witt M. et al.: Creative process engagement in a multiplayer online-ideation game. In Hans-Ulrich Heiß; Peter Pepper; Holger Schlingloff/Jörg Schneider (Eds.): Informatik schafft Communities; Lecture Notes in Informatics (LNI) - Proceedings, Series of the Gesellschaft fuer Informatik, Vol. 208, 2012; pp. 978-991.

Serious Fun - Effects of Gamification on Knowledge Exchange in Enterprises

Thomas Wiegand, Stefan Stieglitz

Department of Information Systems
University of Muenster
Leonardo-Campus 11, 48149 Muenster, Germany
t.wiegand@i2solutions.de
stefan.stieglitz@uni-muenster.de

Abstract: In recent years, gamification has become one of the major trends in information systems. However, until now the kinds of gamification approaches which exist and how they could be applied in the context of knowledge exchange in enterprises has been under-researched. By briefly summarising the current status of research in the fields of social capital, knowledge management and motivational theory we present a potential 'missing link': Gamification. Based on this, we suggest a classification of gamification approaches and instruments.

1 Introduction

Nowadays, entrepreneurs need to manage efficient collaboration between people, processes, and technology in order to meet the demand of increasing competitive environments, innovation pressures, and growing complexity. In order to generate competitive advantages it becomes increasingly important to successfully combine and share specialised knowledge [Ro12]. However, knowledge is created and held by individuals, and not by organizations [Gr96]. Looking at the experts, with different cognitive, relational, and structural backgrounds, points to the differences in individual characteristics and social requirements. Thus, successful collaboration needs to overcome different barriers and grow "social capital" [Mo13]. Research during the past few decades has focussed on different incentive instruments (e.g. monetary bonuses) in order to align employee behaviour with organisational goals and better understand the source of human motivation [Pi10]. Gamification is one of the latest hypes in motivational research [ZC11].

During the last four years, the question of how gamification can support business-to-consumer (B2C) scenarios, e.g. to increase customer involvement and retention, has been discussed [ZL10]. The potential of gamification however might go far beyond customer 'high scores' and 'badges'. Being a powerful driver for goal-oriented behavioural change on the one hand, and a driver of human interaction on the other, gamification has the potential to revolutionise the way people work, collaborate, and develop [Yu14]. It targets the very basic needs of people: Social belonging, competition, reward, self-expression, fun, and recognition, thereby making it a promising instrument for enterprises for the coordination of social capital and knowledge exchange.

However, in order to utilise the benefits of gamification, researchers and organisations need to have a better understanding about different approaches to actually create and apply gamification. This is also important from a design-science perspective. In our paper we want to make a first step in this direction by gathering existing approaches in the context of gamification and interweaving these with their sources (motivational theories) and specific targets (mitigating knowledge-exchange barriers). In the next section, we therefore provide a brief overview on knowledge exchange and motivational research. Following this, we introduce gamification and we deal with the question of how gamification is 'created'. Therefore, we derive gamification instruments from literature and discuss them. The paper ends with a discussion and conclusion of our findings.

2 Knowledge exchange, motivation theory, and incentive systems

The knowledge-based theory of the firm states that knowledge management is a crucial factor in team collaboration since knowledge is the most important resource of a firm [Gr96]. Moreover, it is held by individuals (not firms), and is thus socially complex. In many contexts, social capital is required for effective knowledge sharing. This requires a holistic approach, taking into account technical-, organisational-, and human-oriented factors [BH99] aligned by an appropriate goal (incentive) system.

Analysing the relationship between social capital and knowledge exchange shows that social capital serves as an enabler for knowledge management and that both have a positive reinforcing relation [Mo13]. However, building up social capital requires trust [NG98]. Without trust, communication between team members is restricted. Therefore, companies should support participative management structures with an open feedback culture [Fu06]. At the same time, barriers to knowledge exchange should be removed or lowered. Knowledge-exchange barriers are understood as factors which prevent the flow of knowledge throughout an organisation. Academic literature has already identified several knowledge barriers [Ri05, CL05, WA05] on different levels, such as: (1) Human-oriented (low awareness of values, lack of trust, lack of communication, different values and beliefs, i.e. culture, lack of management support, lack of encouragement, and corporate culture); (2) organizational (lack of knowledge management processes and goals, lack of time and resources); and (3) technical (lack of appropriate infrastructure, lack of integration of information systems).

Following [AHK10], one of the most relevant reasons for the existence of knowledge barriers is the absence of incentives for team members to engage in knowledge management activities. Thus, adapting appropriate incentive systems in order to build engagement for knowledge management activities seems a complex but promising approach. In order to create engagement, incentives systems need to address motivation. Lu and Wu summarise motivational theories as follows: "[…] needs-based motivations are the primary impetus for people to engage in various behaviors, and such motivations can be broadly categorized into two major groups: extrinsic and intrinsic" [LW13]. Intrinsic motivation drives people's behaviour without external stimulus while extrinsic motivation always involves external inducement (e.g. the 'carrot and stick' metaphor). In most

working environments extrinsic rewards (e.g. monetary bonuses when achieving predefined goals) are common nowadays.

However, during the past number of years psychologists have begun questioning the effects of extrinsic rewards when it comes to more complex conceptual and creative tasks. Pink states that extrinsically incentivised people narrow their focus, hindering high performance when working in conceptual jobs; only on straightforward tasks did these mechanisms function [Pi10]. According to Pink, autonomy, mastery, and purpose are key for 21st century intrinsic incentive systems [Pi10]. Literature about intrinsic motivation in the past has focussed primarily on aspects of joy. However, intrinsic motivation might stem from many other factors [Lo13] (see Table 1):

Table 1. Human work related needs

Pos	Human Needs/Wants (Work Related)	Mentioned by
A)	Social Exchange / Belonging	R M H C S
B)	Image / Recognition	R M H C
C)	Influence / Power	R C
D)	Competition / Status	R H C
E)	Achievement / Reward	M H C
F)	Mastery	H C P S
G)	Self-Expression	M C
H)	Relevance / Purpose / Idealism	R M P S
I)	Autonomy / Independence	R C P S

(**R** = Reiss 2004 (Reiss Profile) [Re04]; **M** = Maslow / Frager 1943 (Hierarchy of Needs) [MF87]; **H** = Herzberg 2008 (Two Factor Theory) [He08]; **C** = McClelland 1978 (Need Theory) [Mc78] **S** = Deci / Ryan 2000 (SDT) [DR00]; **P** = Pink 2010 (Drive) [Pi10])

Considering the broad variety of incentives, it is difficult to understand why organisations in recent decades did not at least combine their incentive systems of sole extrinsic, utilitarian nature with more intrinsic, hedonistic aspects. Lowry et al. state that "[t]angible rewards, deadlines, directives, and threats (common to extrinsic motivation) are examples of factors that undermine perceptions of self-efficacy and control (and of related joy and satisfaction)" [Lo13]. Figure 1 provides an overview of how different incentive systems are split between intrinsic and extrinsic systems focussing on pleasure or productivity. Hamari found that gamification as instrument can help turning solely utilitarian systems (productivity-oriented) into more hedonistically oriented (i.e. pleasurable) ones [Ha13]. Information technology can be used for both productivity and pleasure at the same time, representing a form of dual-purposed motivation system [Ce06].

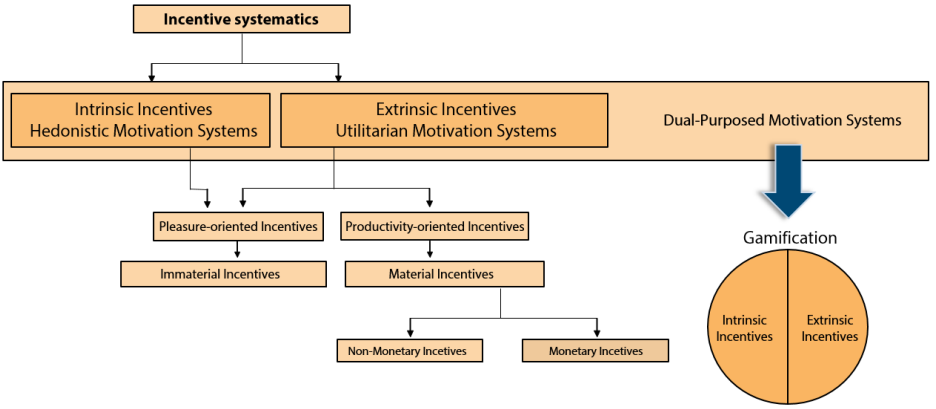


Figure 1. Incentive systems and gamification extended from [Za05]

Based on these observations, it seems promising to take a closer look at how to apply gamification as a motivational system which meets work-related needs on the one hand while also lowering the identified barriers of knowledge exchange.

3 Gamification

Deterding refers to gamification as "the use (rather than the extension) of design (rather than game-based technology or other game related practices) elements (rather than full-fledged games) characteristic for games (rather than play or playfulness) in non-game contexts (regardless of specific usage intentions, contexts, or media of implementation)" [DDKN11]. Following him, gamification should help to align organisational and employee objectives which according to the principal-agent theory differ depending on the relevant objectives. While organisations strive for revenue and resource maximisation, employees try to maximise their utility. The abstract concept of utility can be translated into work-related needs and wants.

Organisations should understand gamification as a strategy to 'design' behaviour, foster innovation, and develop human resources [Ga12]. When combining the findings of Reeves and Read [RR09], and Zichermann and Cunningham [ZC11], which both belong to the pioneers of gamification, the following effects on employees in a business-to-business (B2B) setting can be observed: (1) Shortened feedback cycles (from year to day); (2) clear goals and transparent rules (objectivity); (3) creation of participation, attention, and interaction (team building); (4) cutting of big projects into small achievable pieces; (5) substitution of extrinsic rewards by integration and relationships; (6) fostering of innovation (by supporting creativity as well as trial and error); and (7) creation of social capital in dispersed settings. However, when analysing the current status of real-life B2B gamification applications, these promising effects often do not meet expectations. One of the main reasons for the failure of gamification identified is insufficient game design. Business tasks or incentive systems do not become successful in the long run simply because points or badges are added to a software tool. Bogost criticizes the

buzz word ‘gamification’; while trying to create positive emotions by using the term ‘game’ there often is no further relation to the positive effects a game has in real-life [Bo11].

Following the same path, Nicholson created a user-centric theoretical framework for meaningful gamification. He notes: “If users have a positive and meaningful game based experience that is well connected to the underlying non-game setting, then the organization will benefit in the long term” [Ni12]. The critical success factor of a gamification system is the integration of user-centred game design elements instead of organisation-centred ones.

4 How to create gamification?

Hamari et al. conducted an extensive literature review about motivational affordances related to gamification. They found ten approaches that are used to realise gamification. Overall, they analysed 24 publications. In terms of frequency, they most commonly found ‘points’ (13), ‘leaderboards’ (10), and ‘achievements / badges’ (9) [HKS14]. However, they also found other instruments such as ‘challenges’ (7), ‘levels’ (6), ‘story / theme’ (6), ‘feedback’ (6), ‘clear goals’ (4), ‘rewards’ (4), and ‘progress’ (4) [HKS14].

Moreover, they note that outside pressures (e.g. extrinsic rewards) hinder intrinsic motivation and thereby also reduce gamification effects as it is suggested by self-determination theory. When clustering the identified gamification instruments into those of an intrinsic or extrinsic nature, contradictions are raised depending on the viewpoint taken. Gamification belongs to the dual-purposed motivation systems (combining intrinsic / hedonistic and extrinsic / utilitarian systems) [DRK99]. It can be argued that gamification as an instrument for goal-oriented behavioural change is extrinsic by definition, since there are external influences (e.g. goals) which do not stem from the individual in an intrinsic sense. On the other hand one could argue that varying human needs drives the roots of intrinsic motivation. If those needs can be fulfilled by ‘extrinsic’ gamification instruments a paradox arises, which no longer allows the classification of a single instrument as purely intrinsic or extrinsic. Many bonuses (e.g. extra money, social incentives, days off) are driven by an incentive mechanism which is extrinsic in nature. However, depending on the inner-need level of the individual, achievement / reward, social exchange / belonging, autonomy / independence might cause different levels of intrinsic motivation depending on current individual preferences. Thus extrinsic incentives can drive intrinsic motivation. Deci et al. analysed 128 studies on the effects of extrinsic rewards on intrinsic motivation. They concluded that engagement-contingent, performance-contingent, and completion-contingent rewards undermine free-choice intrinsic motivation [DRK99].

4.1 Gamification instruments

A major issue when trying to empirically measure singular gamification effects is that they often exhibit high complexity and interdependencies. For example, not every indi-

vidual likes games, and those who do, do not like the same game mechanisms. Furthermore, when moving from sole individual instruments (like private score/point mechanisms) to more group integrated mechanisms (like leaderboards or feedback instruments) many emotional and psychological issues interact causing high complexity and prohibiting a clear cause and effect relationship study. As a consequence, gamification instruments can be sorted (ascendingly) in terms of expected interdependency and social complexity. Based on our literature analysis we classified the different instruments, as shown in Figure 2.

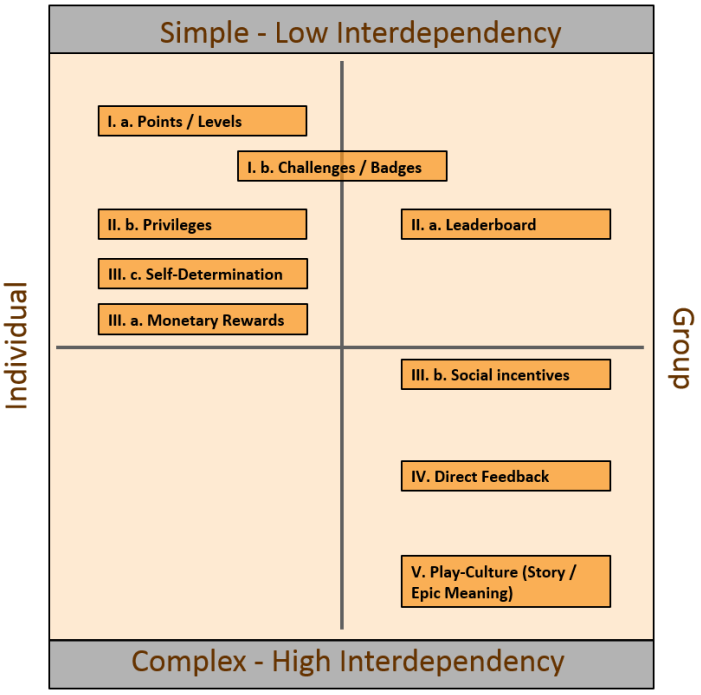


Figure 2. Selected gamification instruments (coordination variables)

Group I (progress, recognition, status)

I. a. Points and Levels

Ong notes that "[p]oints are a running numerical value given for any single action or combination of actions". They are given if desired behaviours are practised in order to provide frequent and clear feedback to users. The cumulative nature of points keeps users attracted. Points and levels, in their basic function, allow tracking of progress and provide a feeling of individual progression and reward. People like to take 'baby steps' which become visible to users as soon as they receive even a small amount of points. Progressing to higher levels provides positive feedback to the user (achievement / satisfaction). It seems questionable whether private points and levels alone will be able to

create reward feelings in the long run. If transparently communicated between users, points and levels can represent status.

I. b. Challenges / Badges
Challenges are tasks or goals which if successfully completed by users result in a badge reward. A badge consists of a small picture (graphic: e.g. a trophy) and a title (e.g. "Master Diplomat"). If badges are not visible to other users, they are comparable to levels (with a specific focus). They reward user actions and address collector instincts. However, they become powerful, but also socially complex, as soon as they are visible to others. For instance, the military has used the principle of badges and status for centuries. When deploying badges it is important to understand that badges must be meaningful, i.e. they should become progressively more challenging to earn and they should never all be earned [St12].

Group II (competition, status rewards)

II. a. Leaderboard
A leaderboard shows the top performers ordered by rank to all users (e.g. based on points). Leaderboards are considered as being one of the most effective tools to create competition and status. However, they also bear many pitfalls if not deployed wisely. The long-term perspective has to be taken into account, since if the scoring is not bound to a certain time-frame it might soon become frustrating for 'low performers'. Nicholson found in a gamification classroom experiment that after the novelty of leaderboards had worn off, engagement faded for most students while a few 'high performers were fighting for the top: "[…] students later reflected that the leaderboards were a demotivating factor; once the gap grew between the leaders and the rest of the class, there was little reason to pursue more of these points" [Ni13].

II. b. Privileges
Privileges are classified as non-monetary rewards which are earned in combination with other instruments (e.g. by reaching a certain score, level, or badge). The power of privileges can be seen in the real world when looking at different frequent-flyer bonus programmes or credit card types. It's a combination of social status and 'something money can't always buy' (e.g. the limousine shuttle or access to a special lounge). However, it also bears some risk of social complexity (e.g. jealousy / inequality) as soon as the privileges become transparent to other users.

Group III (rewards)

III. a. Monetary Rewards
Monetary bonus payments are the 'old-fashioned' reward structure of today's business world. Arguably they don't belong to gamification instruments since they undermine intrinsic motivations and might lower performance. However, depending on the intensity of innate desires [Ge13] and user characteristics, it might well be that there is room for monetary rewards in gamification.

III. b. Social Incentives
Social incentives should be understood as social 'events' which are conducted as rewards for achievements. Even though they do cost monetary and time resources they are

interesting, since they offer a lot of design flexibility and aim at creating social capital and bonds via social exchange. Design alternatives could be, for example, dinner vouchers (with the management or other winners), but might as well be a weekend for the whole team (regardless of performance) combined with some privileges (e.g. choice of destination for winners). Social incentives foster social exchange, belongingness, and trust, thereby creating valuable social capital.

III. c. Self-Determination
Self-determination should be understood as an alternative individual reward by giving a person a surplus of autonomy and independence. It follows the intrinsic motivation path by giving people freedom of choice in their work setting. This might be as simple as a random day off, but might also be freedom of choice to work on selected topics in order to address other needs like relevance, self-expression, or mastery. Even though used as a reward, it addresses needs other than social incentives, monetary rewards, or privileges.

Group IV (recognition)

IV. Direct Feedback
In 2009, Brim and Asplund found in a survey conducted in the United States that "[e]mployees receiving predominantly negative feedback from their manager are over 20 times more likely to be engaged than those receiving little or no feedback" [BA09]. Continuous feedback (not only from management but from a 360° viewpoint) seems to have a huge impact on employee engagement. According to Csikszentmihalyi's flow theory, immediate feedback is a central requirement in order to dive into the flow (deeply engage in an activity) [Cs97]. Moreover, direct feedback as gamification instrument may foster social capital. However, it must be classified as socially complex with high interdependencies.

Group V (self-expression, relevance, social belonging)

V. Play-Culture (Story / Epic Meaning)
Play-culture focusses on integrating interesting gameplay and story at the core of the gamification system. Badgeville Inc. state that "epic meaning" in a gamification system is necessary to let the users feel they are working on something big and meaningful [Ba14]. That in turn will increase engagement, loyalty, influence, and fun [Ba14]. Play-culture, if it is able to deeply engage users in the game, has the potential to overcome cultural barriers (by building an own in-game culture). This can help to overcome different beliefs and intensify social exchange, thereby also fostering trust and social capital. Gameplay can offer transparent guidelines on how users should develop as individuals and in teams. Furthermore it offers room for self-expression (e.g. via avatars). According to Kuo, the most recent gamification systems (termed gamification 1.0) are missing this central element of meaning, story, and play [Ku13]. Kuo describes the next stage, gamification 2.0, as follows: "[…] we need to combine our learning from Gamification 1.0 and look at real world business pain points and map core management activities and operations to gamified actions. [...] Points, leaderboards and badges and other such pure game elements are the fundamental building blocks, but we need to look beyond […]" [Ku13]. Being undoubtedly a key instrument for meaningful gamification play-culture bears as many risks as chances. Due to the high specificity of each story and the differ-

ences in perceptions of users with regard to "Epic Meaning" as well as the interdependencies with all previously mentioned instruments, play-culture is the most complex gamification instrument.

4.2 Expected effects of gamification instruments

Gamification was introduced as a concept between intrinsic motivation and extrinsic incentives. The previously derived gamification instruments can help to overcome knowledge-exchange barriers in an enterprise context. Based on our literature review we suggest that certain gamification instruments address certain work-related needs and certain knowledge-exchange barriers (Figure 3).

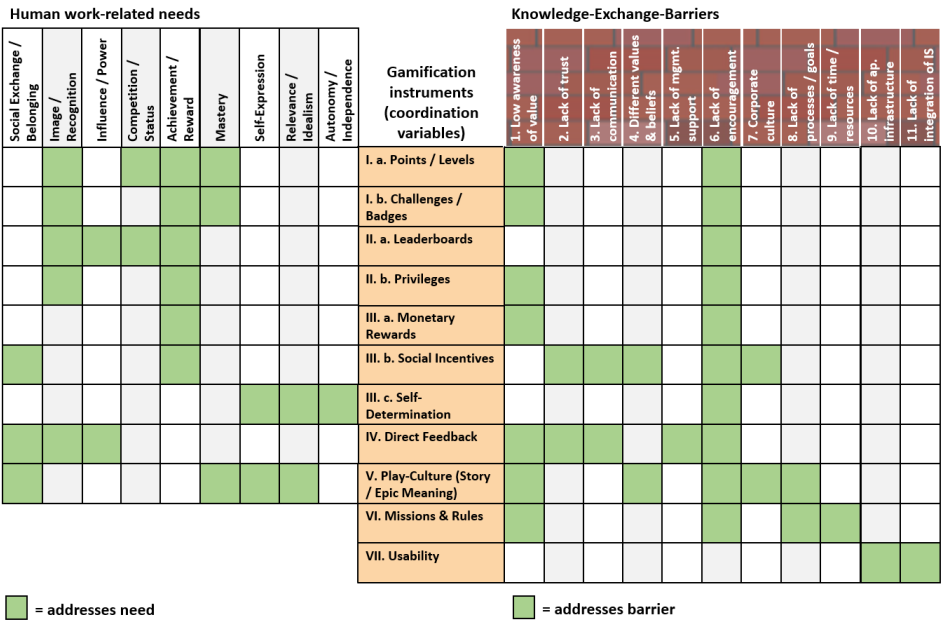


Figure 3. Overview: Gamification instruments and their expected effects

Figure 3 presents nine work-related needs (left side) which serve as triggers for different gamification instruments (e.g. the need for achievement triggers, points and levels – social belonging). Green spots in the matrix visualise the expected relationships. The authors suggest that gamification is triggered most often by the need of achievements and rewards (six matches).

On the right side of the matrix, knowledge-exchange barriers are listed. The green spots show the expected mitigating influences of the gamification instruments on the specific barriers (e.g. social incentives are expected to lower the barriers "Lack of Trust", "Different Values / Beliefs" and "Lack of Encouragement"). Even though the relationships still need to be verified empirically (which will be conducted in a subsequent paper) some preliminary conclusions can be drawn. All instruments are expected to create en-

couragement, thereby underlining the power of gamification as a dual-purposed motivation (incentive) system. Seven instruments are expected to create awareness for knowledge management, which is not surprisingly given the nature of an incentive system. However, Social Incentives, Direct Feedback, and Play-Culture (each mitigating an expected five barriers) seem powerful but, as noted previously, are also socially complex instruments.

5 Conclusion

Being a powerful driver for goal-oriented behavioural change on the one hand, and a driver of human interaction on the other hand, gamification as incentive system has the potential to revolutionise the way people work, collaborate, and develop. Gamification should help to align organisational and employee objectives which, according to the principal-agent theory, do differ depending on the relevant objectives. It can be argued that gamification is extrinsic by definition, since its instruments (e.g. points and levels) are external influences which do not stem from the individual in an intrinsic sense. Hamari found that gamification as instrument can help to turn solely utilitarian systems (productivity-oriented) into more hedonistically (pleasure-oriented) ones [Ha13]. Information technology can be used for both productivity and pleasure at the same time representing a form of dual-purposed motivation system combining the utilitarian and hedonistic dimensions.

Gamification addresses the previously identified intrinsic motivations via different instruments. Eleven gamification instruments are derived which have the expected potential to serve as stimulators between the identified human work-related needs (fostering intrinsic motivation) and the knowledge-exchange barriers (lowering the barriers). Among the most promising but also most challenging (social complexity and interdependency) are Social Incentives, Direct Feedback, and Play-Culture. Organisations should understand gamification as a strategy to 'design' behaviour, foster innovation, and develop human resources. However, today's gamification systems often lack a sufficient game design. In this context, Nicholson created a user-centric theoretical framework for meaningful gamification [Ni12]. This paper is based on expected effects between needs, gamifications instruments, and knowledge-exchange barriers. Those expectations rest on deduction and on current litertaure reviews however have not been empirically researched so far. The next step in this research is to refine the research model in a way which allows empirical testing of the expected effects between needs, gamification instruments, and knowledge-exchange barriers.

References

[AHK10] Ajmal, M.; Helo, P.; Kekäle, T.: Critical factors for knowledge management in project business. In Journal of Knowledge Management, 2010, 14; pp. 156–168.

[Ba14] Badgeville Inc.: Game mechanics – epic meaning. Gamification wiki definitions. http://badgeville.com/wiki/Game_Mechanics/Epic_Meaning, accessed 27 Apr 2014.

[BA09] Brim, B.; Asplund, J.: Driving engagement by focusing on strengths. http://businessjournal.gallup.com/content/124214/driving-engagement-focusing-strengths.aspx, accessed 27 Apr 2014.
[BH99] Brandt, D.; Hartmann, E.: Editorial: Research topics and strategies in sociotechnical systems. In Human Factors and Ergonomics in Manufacturing, 1999, 9; pp. 241–243.
[Bo11] Bogost, I.: Gamification is bullshit. My position statement at the Wharton Gamification Symposium. http://www.bogost.com/blog/gamification_is_bullshit.shtml, accessed 4 Jan 2014.
[Ca82] Caillois, R.: Die Spiele und die Menschen. Maske und Rausch. Ullstein, Frankfurt/Main, Germany, 1982.
[CL05] Chua, A.; Lam, W.: Why KM projects fail: A multi-case analysis. In Journal of Knowledge Management, 2005, 9; pp. 6–17.
[Cs97] Csikszentmihalyi, M.: Finding flow. The psychology of engagement with everyday life. Basic Books, New York, NY, 1997.
[DDKN11] Deterding, S.; Dixon, D.; Khaled, R.; Nacke, L.: From game design elements to gamefulness: Defining "gamification". In Proceedings of the 15th International Academic MindTrek Conference, 2011.
[DR00] Deci, E. L.; Ryan, R. M.: The 'what' and 'why' of goal pursuits: Human needs and the self-determination of behavior. In Psychological Inquiry, 2000, 11; pp. 227–268.
[DRK99] Deci, E. L.; Ryan, R. M.; Koestner, R.: A meta-analytic review of experiments examining the effects of extrinsic rewards in intrinsic motivation. In Psychological Bulletin, 1999, 125; pp. 627–668.
[Fu06] Fuchs, M.: Sozialkapital, Vertrauen und Wissenstransfer in Unternehmen. Dt. Univ.-Verl., Wiesbaden, Germany, 2006.
[Ga12] Gartner: Gamification 2020: What is the future of gamification?, 2012.
[Ge13] Gears, D.: Gamification in business: Designing motivating solutions to problem situations. In Proceedings of the CHI'13 Conference on Human Factors in Computing Systems, Paris, France, April 27–May 2, 2013.
[Gr96] Grant, R. M.: Toward a knowledge-based theory of the firm. In Strategic Management Journal, 1996, 17; pp. 109–122.
[Ha13] Hamari, J.: Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. In Electronic Commerce Research and Applications, 2013, 12; pp. 236–245.
[He08] Herzberg, F.: One more time: How do you motivate employees? Harvard Business Press, Boston, MA, 2008.
[HKS14] Hamari, J.; Koivisto, J.; Sarsa, H.: Does gamification work? A literature review of empirical studies on gamification. In Proceedings of the 47th Hawaii International Conference, Hawaii, USA, January 6–9, 2014.
[NG98] Nahapiet, J.; Ghoshal, S.: Social capital, intellectual capital, and the organizational advantage. In The Academy of Management Review, 1998, 23; pp. 242–266.
[Ku13] Kuo, I.: Moving beyond points and badges: Gamification 2.0. Gamification Co. http://www.gamification.co/2013/04/03/moving-beyond-points-and-badges-gamification-2-0/, accessed 23 Feb 2014.
[Lo13] Lowry, P.: Proposing the hedonic-motivation system adoption model (HMSAM) to increase understanding of adoption of hedonically motivated systems. In Journal of the Association for Information Systems, 2013, 14; pp. 617–671.
[LW13] Lu, X.; Wu, J.: Effects of extrinsic and intrinsic motivators on using utilitarian, hedonic, and dual-purposed information systems: A meta-analysis. In Journal of the Association for Information Systems, 2013, Vol. 14: Iss. 3, Article 1.
[Mc78] McClelland, D. C.: Managing motivation to expand human freedom. In American Psychologist, 1978, 33; pp. 201–210.

[MF87] Maslow, A. H.; Frager, R.: Motivation and personality. Harper and Row, New York, NY, 1987.
[Mo13] Monavvarian, A.; Nasser, A.; Akhavan, P.; Ashena, M.: Developing social capital for facilitating knowledge management practices. In International Journal of Social Economics, 2013, 40; pp. 826–844.
[Ni12] Nicholson, S.: A user-centered theoretical framework for meaningful gamification. Paper presented at Games+Learning+Society 8.0, Madison, WI, 2012.
[Ni13] Nicholson, S.: Exploring gamification techniques for classroom management, 2013.
[Pi10] Pink, D. H.: Drive. The surprising truth about what motivates us. Canongate, 2010.
[Re04] Reiss, S.: Multifaceted nature of intrinsic motivation: The theory of 16 basic desires. In Review of General Psychology, 2004, 8; pp. 179–193.
[Ri05] Riege, A.: Three-dozen knowledge-sharing barriers managers must consider. In Journal of Knowledge Management, 2005, 9; pp. 18–35.
[Ro12] Rockett, K.: Perspectives on the knowledge-based society. An introduction to the special issue. In Economics: The Open-Access, Open-Assessment E-Journal, 2012.
[RR09] Reeves, B.; Read, J. L.: Total engagement. Harvard Business School Press, Boston, MA, 2009.
[St12] Stampfl, N. S.: Die verspielte Gesellschaft. Gamification oder Leben im Zeitalter des Computerspiels. Heise, Hannover, Germany, 2012.
[Ce06] Chesney, T.: An acceptance model for useful and fun information systems. In Human Technology, 2006, 2; pp. 225–235.
[WA05] Wong, K. Y.; Aspinwall, E.: An empirical study of the important factors for knowledge-management adoption in the SME sector. In Journal of Knowledge Management, 2005, 9; pp. 64–82.
[Yu14] Yu-kai, C.: Octalysis: complete gamification framework. http://www.yukaichou.com/gamification-examples/octalysis-complete-gamification-framework, accessed 22 Feb 2014.
[Za05] Zaunmüller, H.: Anreizsysteme für das Wissensmanagement in KMU. Dt. Univ.-Verl., Wiesbaden, Germany, 2005.
[ZC11] Zichermann, G.; Cunningham, C.: Gamification by design. Implementing game mechanics in web and mobile apps. O'Reilly Media, Sebastopol, CA, 2011.
[ZL10] Zichermann, G.; Linder, J.: Game-based marketing. Inspire customer loyalty through rewards, challenges, and contests. Wiley, Hoboken, NJ, 2010.

Towards Gamification Analytics - Requirements for Monitoring and Adapting Gamification Designs

Benjamin Heilbrunn[1], Philipp Herzig[1], Alexander Schill[2],

[1]SAP AG, Chemnitzer Straße 48, 01187 Dresden, Germany
{benjamin.heilbrunn, philipp.herzig}@sap.com
[2]TU Dresden, Nöthnitzer Straße 46, 01062 Dresden, Germany
alexander.schill@tu-dresden.de

Abstract: Application and gamification data contain valuable information about users and their behavior. This data can be used to measure the success of gamification projects, to analyze user behavior, and to continuously improve gamification designs. However, to support the process of gamification analyses, appropriate tool support is missing. Moreover, the specific requirements for methods and tools have not been studied yet and remain unclear. Therefore, this paper presents a theoretical model of 22 requirements that can be used to assess existing solutions or construct novel methods and tools for gamification analytics. The presented model is validated based on comprehensive expert interviews.

1 Introduction

Gamification is a novel topic which receives increasing attention from researchers as well as practitioners. Inspired by the user experience of computer games, gamification describes the idea of using game design elements in non-gaming contexts [DSN+11]. Recent studies show that gamification has the potential to considerably improve user experience and user engagement in a variety of software application domains [HKS, HSA12].

In literature [HAWS14], experts structure the methodology of gamification projects into four high-level phases: (1) *Business Modeling and Requirements*, where the application context is analyzed and business goals are documented. (2) *Design*, where the gamification design is developed and playtested. (3) *Implementation* and related activities, where the design is implemented as software artifacts and functionally tested. And finally (4) *Monitoring and Adaptation*, where the business goal achievement is measured and subsequent design adaptations are conducted. To support the rapid implementation and integration of gamification designs in phase three, researchers and industry have come up with generic gamification platforms, e.g., [HAS12, Bun14, Bad14]. However, as we will show in this paper, there is still an unsatisfied demand for tools that support phase four, i.e., the monitoring and adaptation of gamification designs. Experts request such gamification analytics tools to measure the success of gamification design changes, to better understand the user behavior, and to learn when a gamification design requires adjustment [KH13].

In this paper we present 22 user requirements for gamification analytics which were validated by semi-structured interviews with 10 gamification experts. The rest of the paper is structured as follows: Section 2 introduces the used research methodology. Section 3 presents our hypothetical user requirements model. Section 4 describes the expert sample and the model assessment results. Finally, Section 5 summarizes our work and provides an outlook to future research.

2 Methodology

In the following, we introduce our research methodology which is used for developing and validating the user requirements model.

Initially, we created a hypothetical user requirements model based on existing gamification literature, game literature, and our own experience in gamifying various business applications. Besides a textual description of these requirements, we also developed conceptual mock-ups to showcase the functionality in accordance with [May99].

To validate and extend our hypothetical user requirements model, we invited gamification experts from a variety of functions and domains to semi-structured interviews. The interviews were conducted with web conferencing tools and recorded for later analysis. The interview structure comprised four parts: (1) *Introduction* – The interviewer briefly introduced himself, the direction of his work, the interview procedure, and the intended exploitation of results. (2) *Professional background* – To get to know the experts, they were asked to introduce themselves and their professional career. (3) *Experience and expertise in gamification* – Experts were asked to elaborate on their activities in the field of gamification. Questions addressed the state of the art in gamification projects, as well as the identification of problems with today's project methodologies and technologies. (4) *Discussion of mock-ups* – The last step of the interview comprised the presentation and discussion of the conceptual mock-ups. The experts were first introduced to the demo scenario, then guided through the mock-ups. For each presented feature they were asked to provide their opinion and to share feedback.

3 Hypotheses Model

This section presents our initial model of user requirements hypotheses which comprises five categories of requirements: *Application KPI Monitoring* helps gamification experts to observe the development of application-related KPIs. *Gamification Element Analytics* supports them to understand the development of the game state and how users interact with game elements. *Gamification Design Adaptation* enables gamification experts to test the impact of changes to the gamification design. *User Groups of Interest* allows experts to focus analyses on subsets of the users. Finally, *Simulation* empowers experts to simulate arbitrary gamification designs with existing behavior data.

3.1 Application KPI Monitoring

Definition of Application KPIs Gamification literature emphasizes the importance of defining clear business goals and measuring the success of gamification designs towards their achievement [Kap13, KH13, WH12, HAWS14, Rim14]. Key Performance Indicators (KPIs) based on user behavior can be used to operationalize business goals. Accordingly, we argue that the collection of relevant data is one of the essential prerequisites for realizing gamification analytics. A gamification analytics tool should be able to calculate application-related KPIs such as *New Blog Posts Per User and Month* in a community. In the context of KPI definition, we define the following three concrete requirements:

R1) Custom KPIs – Gamified applications typically have domain-specific KPIs. Gamification experts should be able to define these KPIs based on the available application log data which is typically available in form of event streams, databases, or log files. The definition of KPIs should be possible at any point of time, allowing experts to adjust and refine KPIs according to their informational needs and available event data.

R2) Pattern Based KPIs – Experts should be able to formulate KPIs that count the number of particular pattern occurrences in the behavior data of users. This supports experts in measuring the success of game elements which aim at influencing behavior patterns. Experts might, for instance, be interested in a KPI that determines the proportion of community users who actively read the community rules before posting their first question.

R3) KPI Goal Values – The experts should be able to define and adjust KPI goal values whose fulfillment will be monitored automatically by the gamification analytics system (see R6). For instance, that *New Blog Posts Per User and Month* should be at least *0.7*.

Presentation of Application KPIs In the following, we describe our user requirement hypotheses regarding the presentation of application KPIs. Figure 1 shows the corresponding conceptual mock-up of the application KPI monitoring screen which was used for discussion during the interviews.

R4) Dashboard – Gamification experts should be able to get a comprehensible overview of the state and over time development of application KPIs. This can be achieved by a visual dashboard that combines charts with descriptive statistics.

R5) Change Markers – Experts should be able to understand the impact of historical changes in the gamification design on the development of application KPIs. This can be achieved by annotating KPI curves with markers that indicate past design changes.

R6) Goal Markers – Experts should be aware of how individual KPIs perform in relation to their goal value. The defined KPI goal value should be shown together with the actual KPI value and deviations should be indicated. This might help experts to immediately notice undesired changes and gives them the chance to take appropriate action such as exploring the data for better insights or adapting the gamification design to increase engagement.

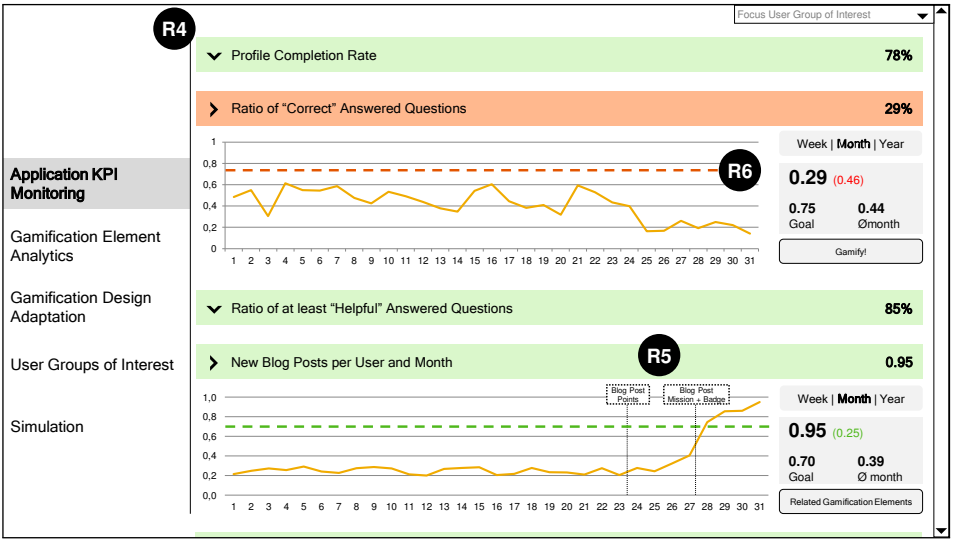


Figure 1: Conceptual Mock-up of Application KPI Monitoring Dashboard

3.2 Game Element Analytics

Game State Overview Gamification experts should have an overview of the game state and its development over time. Exploring the relation between game states and the corresponding users might help experts to detect design flaws or other needs for design adjustment. A design flaw could be, for instance, that users spend significantly more time in a level than expected. A need for adjustment might arise from the fact that most users already reached the final level. In particular, we define the following requirements:

R7) Gamification Feedback Rate – Feedback is an important element of games [SZ04, ZC11, WH12]. We define *Gamification Feedback* as any state change in the game that is perceived by the user as success, e.g., gaining points, or receiving a badge. Correspondingly, the *Feedback Rate* describes the amount of feedback per time. Experts should be able to inspect the feedback rate over time, corresponding descriptive statistics, and annotations representing past design changes. This insight might help them to qualify all other observations, and can be a starting point for investigating unexpected user behavior. For example, a game with an average of 0.1 feedbacks per user hour and a maximum of 20 feedbacks per user hour might have significant flaws in the design of its mechanics.

R8) Point Distributions – Experts should have insight into the distribution of points over users. This might help them to detect flaws in the balance of point amounts for gamified actions, e.g., when 1% of the users own 90% of the points.

R9) Achievable Game Elements – Gamification experts should have insight into the overall statistics of badges, levels, missions, and other achievable game elements. Moreover, they should see the users' progress in the game. Drill-down functionality should be available to allow a deeper investigation of particular game elements (see R10-R12). This can be

achieved with a dashboard that shows completion rates and temporal statistics of game elements. As an example, a gamification design might require adaptation, when already 60% of the users reached the highest level.

Detailed Statistics of Game Elements From the game state overview, gamification experts should be able to drill down to more detailed information on the relation between users and achievable game elements such as badges, levels, or missions. In particular, we define the following requirements:

R10) User Distribution on Game Element State – Gamification experts should be able to gain insight about the distribution of users on the states of particular game elements. For missions typical states would, for instance, be *Mission Completed*, *Mission Active* and *Not Assigned to Mission* [Dor12]. This should help them to understand how the users progress in context of the game element. Experts could, for instance, notice that only a few users completed a particular mission, while most others are stuck in one particular sub-goal of that mission. This might be an indicator that the design of the mission needs adjustment.

R11) Temporal Statistics – Experts should be able to see how long users need for the completion of particular game elements. Therefore, they should be able to browse the following distributions: *Time to Completion* – Time period between the start of user existence and game element completion. *Time to Assignment* – Time period between the start of user existence and the assignment to the game element. *Time Active* – Time period between assignment and completion of the game element. For example, noticing that users typically complete a mission faster than expected, might be an indicator for necessary adjustments.

R12) User Characteristics – Gamification experts should be able to explore which statistically significant properties users have in common, who share the same state on a game element of interest. We distinguish game properties and user properties. Game properties originate from the user's state in the game, e.g., *owns badge A*, while user properties originate from the information the application has about the user, e.g., *from geographical region Europe*. By revealing significant factors of user engagement in context of a particular game element, experts could optimize the game design for their individual audience. When experts notice that a mission is significantly more often completed by European users, they could, for example, start investigating the reasons and adapt it to raise its attractiveness in all relevant geographical regions.

3.3 Gamification Design Adaptation

A/B Testing Tests with experimental and control groups (A/B tests) are a widely used method for evaluating the effects of changes in a particular context. They have also been proposed for validating gamification design ideas [Kap13, KH13]. With A/B testing, the effects of game design changes can be verified before activating them for the whole user base. Thus, we formulate the hypothesis that an integrated way of conducting experiments and analyzing their results could be one of the key features of gamification analytics. A/B testing comprises the following two requirements:

R13) Experiment Creation – Experts should be able to create an experiment by defining its name, description, the size of the experimental group, target KPIs, desired KPI impact (increase or decrease), and the actual design changes which are subject of the experiment. After specifying the mentioned parameters and starting the experiment, a user group with the selected experiment size should start interacting with the new design. From this point on the analytics tool should analyze the difference between their behavior and the behavior of the rest of the users. This allows experts to analyze the generated impact (see R14).

R14) Experiment Result Analysis – As an intermediate and final result of A/B tests, a gamification analytics tool should show the experts a summary of observed effects in user behavior. Moreover, it should indicate, whether the effects are statistically significant in comparison to the control group. This supports objective decision making in the design adaptation process. As a result of keeping a new design idea, a new annotation should be created in all relevant graphical charts, indicating that a design change was conducted (see R5). Experiment results should be archived for durable access to the result data which led to a design decision. Figure 2 shows the conceptual mock-up of the A/B test result view.

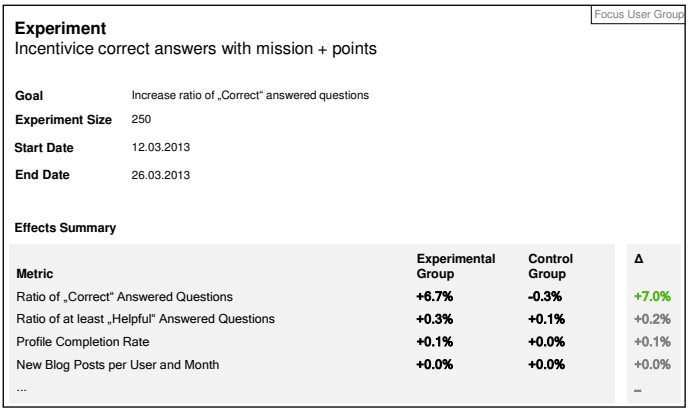


Figure 2: Conceptual Mock-Up of A/B-Test Result View

Direct Design Adaptation R15) In addition to design decisions based on A/B-tests, gamification experts should also be able to conduct direct changes to the gamification design resulting in the creation of change markers in the KPI visualizations (see R5). Direct changes might be necessary in cases when A/B tests are not suitable, e.g., with small user groups or when time constraints apply.

3.4 User Groups of Interest

Experts should be able to focus their analyses on user groups which are of special interest. In the following, we describe our hypothesis regarding their definition and use.

Definition of User Groups of Interest We assume the following three techniques to be relevant for the definition of user groups:

R16) Criteria Based – The experts should be able to define groups based on criteria which are evaluated against the users' properties. This approach is applicable when the exact criteria are well known before creating the user group. Such a group could, for instance, contain all users who are located in the geographical region *Europe* and who at the same time reached game level *9*.

R17) Cluster Analysis – Cluster Analysis aims at finding similar groups in a set of objects [ELLS11]. The experts should be able to conduct a cluster analysis on relevant properties of users to discover groups which are of interest for them. This approach is applicable when the exact criteria of the user group are not known a priori.

R18) Manual Selection – Experts should be able to manually compose a user group. This can be useful in the analysis of user groups, whose members' behavior is of special interest and whose members are known a priori. Gamification experts might, for instance, want to compose a user group out of community members with a high reputation.

Filtering of Overviews by User Groups of Interest R19) The experts should be able to filter overviews by selecting a user group of interest. This should be possible at all places, where statistical overviews are shown. Affected are the application KPIs, as well as the game element statistics and the result presentation of A/B tests.

3.5 Simulation

R20) Simulation is a common tool in game design [Dor12]. In gamification design it is also considered as useful to support early design decisions [Rim14]. Gamification experts should be able to simulate their design ideas with existing user and behavior data. Given that an appropriate dataset of historical user behavior exists, a simulation can help to identify major flaws in the mechanics of a new gamification design. The simulation results should be explorable in the same way as real data by viewing application KPIs, game element analytics, and the opportunity of defining user groups of interest.

4 Evaluation

In the following, we present the results of our semi-structured expert interviews with regards to the proposed requirements. First, we document the characteristics of our expert sample. Second, we describe the results of the model assessment. Finally, we list additional requirements which were raised during the conduction of the interviews.

4.1 Sample

In this section we characterize the interviewed experts by presenting their professional background and gamification experience. The interviews were conducted with 10 experts, took place in February and March 2014, and had a typical duration of 2-3 hours.

Professional Background To the date of the interview, the experts were involved in 36 ($\mu = 3.6, \sigma = 3.9$) currently running or already finished gamification projects. The average experience in the field of gamification projects was $\mu = 2.0$ years ($n = 9, \sigma = 1.2$). A breakdown of the experts' functions and domains of projects which were closer discussed, is given in Table 1. In addition, Figure 3 shows the audience size of projects that are already operative.

Expert Function	N
Development	3
Gamification Consultancy	2
Community Management	2
Project Leadership	1
Research	1
Software Architect	1

Project Domain	N
Training and Education	4
Social Media	3
Software Development Tool	2
Customer Relationship Management	1
Marketing	1
IT-Support	1
Serious Games	1

Table 1: Expert Functions and Project Domains

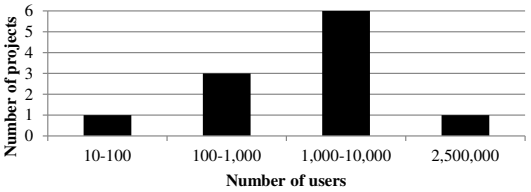


Figure 3: Audience size of experts' finished projects

Experience and Expertise In consistence with gamification literature [Kap13, KH13, WH12, HAWS14, Rim14], we based our requirements model on the assumption that gamification projects have a clear problem definition and measurable business goals. Eight experts reported that their gamification projects started with a clear problem definition. Furthermore, they agreed that they had clear quantitative goals or were still working on their definition. Only the two gamification consultants stated that they often get involved, when the target problems are not well-defined yet. Thus, analyzing the business problem and defining KPIs for success measurement is one of their first activities in a project. We conclude that clear business goals and measurable business goals typically exist.

The power of analytics depends on the richness of the available data-sources. For gamification analytics it is essential to have the necessary data from application side, i.e., user properties and user behavior, and from game side, i.e., the game state. All experts reported that the required application data for measuring the success of game elements is typically available. Seven experts reported, that they already used own tools in order to generate

focused reports on relevant KPIs, mainly for management reporting purposes. However, the fact that application and gamification data typically reside in separate systems, makes a joint analysis very expensive. The upfront investment to build a system for explorative data analysis is too high for typical gamification projects which have a reported size of typically 3-5 project members. Accordingly, all experts expressed a strong demand for appropriate tools and methods which help them to better understand how game elements affect the behavior of the users in an application.

4.2 Model Assessment

All experts agreed to the overall structure of our requirements model as it was directly reflected in the corresponding mock-ups and none of them was missing another high level group of requirements. Table 2 shows a detailed overview of the assessment results for each of the presented requirements. We categorized expert feedback into the following categories: [✓] *Agree* – The expert expressed agreement and stated that the requirement is relevant for gamification analytics, [–] *neutral* – The expert did not express a strong opinion towards the discussed requirement, and [✗] *disagree* – The expert expressed that the discussed requirement is not relevant for gamification analytics. In the following paragraphs we will discuss aspects with noteworthy input from the interviews.

Requirement			✔	–	✖
Application KPI Monitoring	R1	Definition of Custom KPIs	10	0	0
	R2	Definition of Pattern Based KPIs	10	0	0
	R3	Defintion of KPI Goal Values	10	0	0
	R4	Dashboard	10	0	0
	R5	Change Markers	10	0	0
	R6	Goal Markers	10	0	0
Game Element Analytics	R7	Feedback Rate	10	0	0
	R8	Point Distributions	9	1	0
	R9	Achievable Game Elements Statistical Overview	10	0	0
	R10	User Distribution on Game Element State	10	0	0
	R11	Temporal Statistics	10	0	0
	R12	User Characteristics	10	0	0
Gamification Design Adaptation	R13	Experiment Creation	10	0	0
	R14	Experiment Result Analysis	10	0	0
	R15	Direct Design Adaptation	10	0	0
User Groups of Interest	R16	Definition Based on Criteria	6	5	0
	R17	Definition Based on Cluster Analysis	4	6	0
	R18	Definition Based on Manual Selection	1	3	6
	R19	Filtering of Overviews by User Groups	10	0	0
Simulation	R20	Simulation and Result Analysis	7	3	0

Table 2: Model Assessment Results

Application KPI Monitoring All experts agreed to the key concept of monitoring relevant KPIs of the target application. The interviewees participated very actively in its discussion and provided many examples of relevant KPIs from their project contexts. In addition to the definition of KPIs (R1), three experts expressed the wish to be able to define KPIs even

with historical event data. This can be helpful to gain new insights in the process of data exploration and interpreting effects that took place in the past.

Game Element Analytics All experts confirmed that the feedback rate (R7) is useful to qualify other observations. Since, neither too much, nor too less feedback is desirable, they confirmed its value for balancing the amount of feedback of gamification designs. Six experts explicitly emphasized the importance of this high level metric. Moreover, two experts expressed that they would like to have access to more detailed statistical figures and the ability to drill down on the graphical chart in order to start investigating the users that are in the range of interest. All experts agreed on the concept of visualizing point distributions via visual charts combined with descriptive statistics (R8). However, one expert was not sure, whether such an information would really help to understand if the design has flaws, or not.

Gamification Design Adaptation Only two of the interviewed experts reported that they already conducted tests with experimental and control groups. However, all of them agreed that a tool supported workflow would be a strong benefit in the process of adapting gamification designs. Additionally, most of the interviewees explicitly emphasized their desire to conduct A/B tests. They perceived it as a reliable and objective information source that shows the effects of game design changes while also revealing negative side effects.

User Groups of Interest Most of the experts' examples for user groups in their project contexts were criteria based, e.g. filtering by organization unit. Six experts stated that they would like to define user groups based on criteria (R16). The remaining four experts were interested, but did not express a strong opinion towards this requirement. Four experts reported that they were interested in using cluster analysis to discover interesting groups of users with similar characteristics (R17). The remaining six experts were interested, but did not express a strong opinion towards this requirement. One expert reported that he would like to be able to compose user groups manually (R18). In contrast, three experts stated, that they probably would not use such a feature. The remaining six experts were interested but did not express a strong opinion towards this requirement.

Simulation Seven experts agreed on the idea of being able to execute existing sets of behavior event data against arbitrary gamification designs, while only two of them reported that they already conducted simulations in their practice. The interviewees reported that simulation could be helpful for balancing a game, e.g., point amounts, or determining the progression speed. We consider the absence of appropriate technological tools as the main reason why simulations are not conducted more often.

4.3 Additional Requirements

Our initial requirements model did not cover all relevant requirements. Using a threshold of at least two mentions during the interviews, two additional requirements were included into the model. Both of them fall into the category of *Game Element Analytics*.

User Interaction Tracking for Game Elements in the User Interface R21) Three of the interviewees explicitly requested the ability to see how users interact with game elements

in the user interface of the gamified application. They argued that this would help them to understand how attractive individual game elements are. Moreover, the analytics should determine which effects the interaction with a game element has on user behavior, e.g., how viewing a leader board may influence the engagement of users.

Alerting R22) Half of the interviewed experts raised the requirement that they would like to be alerted, when the statistics of a particular game element fulfill certain conditions. The mentioned conditions comprised: *Violation of a Threshold*, or *Violation of a Value Range* – Gamification experts sometimes have an a priori goal how certain key figures of their game elements should look like. In this case, the tool can help them to define and monitor the fulfillment of those intentions, e.g., that no more than 5% of the users should be in the highest level. Finally, *Anomaly Detection* – Automated notification about uncommonly strong changes in the statistics of a game element, e.g., that 1,000 users received a particular badge during a day, while the typical amount is 10–50.

Figure 4 shows the final user requirements model after incorporating the new requirements.

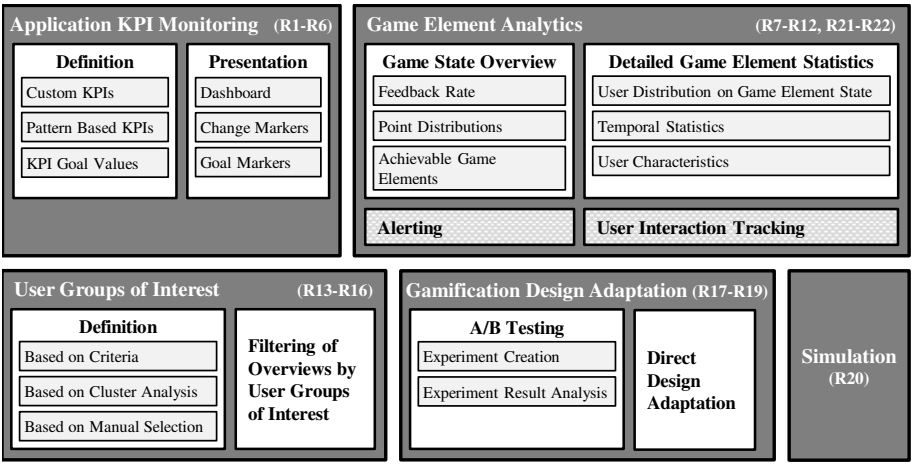


Figure 4: Final User Requirements Model for Gamification Analytics

5 Summary and Outlook

In this paper we presented a model of 22 user requirements for supporting the monitoring and adaptation of gamification designs. The model was validated and extended by 10 semi-structured interviews that were conducted with gamification experts. Our interviews mainly show three aspects: First, gamification projects typically have clear and measurable business goals. Second, the necessary data resources are typically available, however appropriate tools for data analysis and exploration do not exist. Third, the proposed model of an integrated gamification analytics tool was broadly accepted and considered as relevant.

Our future research efforts will focus on quantitative validation and prioritization of the user requirements model, and the design of an appropriate architecture that allows to integrate gamification analytics into today's common system landscapes.

References

[Bad14] Badgeville. Website, March 2014. http://www.badgeville.com.

[Bun14] Bunchball. Website, March 2014. http://www.bunchball.com.

[Dor12] Joris Dormans. *Engineering emergence: applied theory for game design*. Creative Commons, 2012.

[DSN+11] Sebastian Deterding, Miguel Sicart, Lennart Nacke, Kenton O'Hara, and Dan Dixon. Gamification. Using Game-design Elements in Non-gaming Contexts. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '11, pages 2425–2428, New York, NY, USA, 2011. ACM.

[ELLS11] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. *Cluster Analysis*. Wiley, 2011.

[HAS12] Philipp Herzig, Michael Ameling, and Alexander Schill. A Generic Platform for Enterprise Gamification. In *Proceedings of the 2012 Joint Working IEEE/IFIP Conference on Software Architecture and European Conference on Software Architecture*, WICSA-ECSA '12, pages 219–223, Washington, DC, USA, 2012. IEEE Computer Society.

[HAWS14] Philipp Herzig, Michael Ameling, Bernahrd Wolf, and Alexander Schill. Implementing Gamification: Requirements and Gamification Platforms. In *Gamification in Education and Business*. Springer, 2014.

[HKS] Juho Hamari, Jonna Koivisto, and Harri Sarsa. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*.

[HSA12] Philipp Herzig, Susanne Strahringer, and Michael Ameling. Gamification of ERP Systems-Exploring Gamification Effects on User Acceptance Constructs. *Multikonferenz Wirtschaftsinformatik, GITO*, pages 793–804, 2012.

[Kap13] Karl M. Kapp. *The Gamification of Learning and Instruction Fieldbook*. Pfeiffer, 2013.

[KH13] Janaki M. Kumar and Mario Herger. *Gamification at Work: Designing Engaging Business Software*. The Interaction Design Foundation, 2013.

[May99] Deborah. J. Mayhew. *The Usability Engineering Lifecycle: A Practitioner's Handbook for User Interface Design (Interactive Technologies)*. Morgan Kaufmann, 1999.

[Rim14] Gal Rimon. Website, 04 2014. http://www.gamification.co/2013/07/17/how-to-measure-the-roi-of-gamification.

[SZ04] Katie Salen and Eric Zimmerman. *Rules of Play*. MIT Press, 2004.

[WH12] Kevin Werbach and Dan Hunter. *For the Win: How Game Thinking Can Revolutionize Your Business*. Wharton Digital Press, 2012.

[ZC11] Gabe Zichermann and Christopher Cunningham. *Gamification by Design*. O'Reilly Media, 2011.

Smarter Games durch Big Data Analysen: Potentiale und Handlungsfelder

Nicolai Hanner, Thorsten Pröhl, Rüdiger Zarnekow

Fachgebiet für IuK-Management
Technische Universität Berlin
nicolai.hanner@tu-berlin.de, t.proehl@tu-berlin.de, ruediger.zarnekow@tu-berlin.de

Abstract: Big Data ist zu einem Trend der Wissenschaft und Praxis geworden, der verschiedene Anwendungsgebiete in unterschiedlichsten Industrien betrifft. Die Computer- und Videospieleindustrie stellt ein weiteres Anwendungsfeld dar, welches bisher noch kaum untersucht wurde. Vor allem im Rahmen von Online Spielen entstehen eine Vielzahl von Daten, wie Daten zum Spieler- oder Zahlungsverhalten sowie Informationen aus umliegenden Communities. In diesem Beitrag werden die Rahmenbedingungen für den Einsatz von Big Data Analysen in der Computer- und Videospieleindustrie vorgestellt. Darauf basierend werden mögliche Datenquellen identifiziert und passende Analysemethoden vorgestellt. Im Ergebnis soll auf Potentiale für den Einsatz von Big Data Technologien für Anwendungsfelder in der Computer- und Videospieleindustrie hingewiesen werden.

1 Einleitung und Motivation

Die heutige Computer- und Videospieleindustrie ist mit einem geschätzten weltweiten Umsatz von 70 Mrd. US Dollar (im Jahr 2013) zu einer globalen Industrie gewachsen [Pe13]. Gleichzeitig steigt der Wettbewerbsdruck auf die Unternehmen [HEZ14]. Heute sind Spieleentwickler und -publisher mehr denn je von erfolgreichen Spiele-Titeln abhängig [RG06], [Ts07]. Doch die kurzen Lebenszyklen von Videospielen und die zunehmende Konkurrenz erhöhen den wirtschaftlichen Druck auf die Wertschöpfungsketten. Daher gilt es die Erlöse eines Spiels zu optimieren und die Spieler möglichst lange an ein Spiel zu binden. Die Lösung sehen viele Publisher und Entwickler in serviceorientierten Geschäftsmodellen, zum Beispiel der Betrieb von Massively Multiplayer Online Games (MMOG) oder Browser Games. Damit erreichen Publisher und Entwickler eine engere und längere Bindung der Spieler an das Spiel und die eigene Infrastruktur. Für diese Branche eröffnen sich darin viele neue Möglichkeiten, von den generierten Daten der Spieler Informationen abzuleiten und diese wiederum für die Entwicklung als auch für die Verbesserung des Spielerlebnisses einzusetzen. Bisher werden Informationen und Feedback von Spielern bei der Produktentwicklung von Spielen berücksichtigt [AL07], [Fi13]. Durch den Einsatz von Big Data ergibt sich eine neue Anwendungsmöglichkeit, die nicht nur eine nachträgliche Anpassung von Spieleinhalten an eine Datenauswertung beinhaltet, sondern darüber hinaus die Echtzeitauswertung und Echtzeitreaktion der Spiele auf die Spieler ermöglicht. Im Rahmen dieses Artikels soll konzeptionell hergeleitet werden, wie Big Data Techniken

und Analysen in der Computer- und Videospieleindustrie eingesetzt werden können. Der Beitrag beantwortet dabei folgende Forschungsfragen:

1. Welche Art von Daten kann wo in der Computer- und Videospieleindustrie für Big Data Analysen gesammelt werden?
2. Welche Analysemethoden sind für das Szenario der Computer- und Videospieleindustrie besonders geeignet?
3. Welche Potentiale und Handlungsfelder ergeben sich durch den Einsatz von Big Data für die zukünftige Entwicklung und den Betrieb von Computer- und Videospielen?

Der Beitrag gliedert sich wie folgt: Zunächst werden die theoretische Hintergründe erläutert (Abschnitt 2). Dazu gehört die Darstellung der Computer- und Videospieleindustrie (Abschnitt 2.1) sowie ein Überblick über Big Data (Abschnitt 2.2). In Abschnitt 3 erfolgen die Untersuchung der Datenherkunft und die Auswahl geeigneter Analysemethoden aus dem Bereich Big Data. In Abschnitt 4 werden die gewonnen Erkenntnisse zusammengetragen und diskutiert, um mögliche Einsatzgebiete für die Entwicklung und den Betrieb aufzuzeigen. Der Beitrag endet mit einer Zusammenfassung.

2 Hintergrund

2.1 Computer- und Videospieleindustrie

Die klassische Computer- und Videospieleindustrie besteht primär aus den Plattformen des PCs und der Konsole. Die Produktionskette lässt sich in Entwicklung, Publishing und Vertrieb gliedern [CG06]. Publisher stellen in diesen Wertschöpfungsketten die stärksten Wirtschaftsakteure dar [MTP12], [RG06]. Heute ist diese Industrie geprägt durch einen hohen Innovationsgrad (in der Entwicklung von Software und Technologien), einer hohen Marktdynamik und einer sich schnell wandelnden Wertschöpfungsstrukturen [MH13], [RG06].

Durch die zunehmende Digitalisierung der Vertriebswege, Kommunikationskanäle und dem Wachstum mobiler Endgeräte wurde die Industrie um weitere Plattformen wie Online Games und Mobile Gaming erweitert [GRS11], [WCC13]. Online Games sind Spiele die ausschließlich über das Internet gespielt werden können, wohingegen Mobile Gaming die Ausführung von Spielen auf mobilen Endgeräten wie das Smartphone oder das Tablet beschreibt. Diese mobilen Endgeräte stellen mittlerweile nach dem PC die am häufigsten genutze Spiele-Plattform in Deutschland dar [Bi12a]. Der Wandel zum E-Business eröffnete neue Chancen für die Industrie. Beispielsweise zeigen Broekhuizen et al., dass durch den E-Commerce eine Möglichkeit entstanden ist die nachgelagerten Vertriebs- und Distributionsketten oder den Publisher überflüssig zu machen. Somit kann das Produkt direkt an den Kunden geliefert werden [BLR13]. Der Wandel zum E-Commerce führte zu einer Zunahme von möglichen Erlösmodellen [Me09]. Von Onlinespielern wurden häufig monatliche Gebühren für die Bereitstellung des Spieles

verlangt [WCC13]. Im Zusammmenhang damit stehen auch serviceorientierte Erlösmodelle wie zum Beispiel die Bereitstellung von kostenlosen Probespielzeiten [MTP12], [OT11]. Ein weiteres alternatives Erlösmodell, das sich bei den Online wie auch bei den Mobile Games durchgesetzt hat, ist der Verkauf von virtuellen Gütern [Me09], [WCC13].

Durch die Entwicklung von serviceorientierten Geschäftsmodellen und der Disintermediation in der Computer- und Videospieleindustrie wird der Einsatz von Big Data realisierbar. Durch die neuen Vertriebskonzepte, wie App-Stores und der Direktvertrieb digitaler Versionen, können die Aufgaben von Entwicklern, Publishern und Distributoren verschmelzen [BLR13]. Dies führt zu einem unmittelbaren Kundenkontakt zwischen Entwickler und Spieler, so dass die Entwickler einen direkten Zugriff auf die möglichen Informationen aus den Spielen und deren Umgebung haben. Das bisher gängige Prinzip der chronologisch geordneten Lebenszyklen der Spiele, wie Finanzierung, Entwicklung, Produktion, Vertrieb und Erlös, wird von serviceorientierten Geschäftsmodellen aufgebrochen [Jo06], [ZW10]. Insbesondere in den MMOG oder den digitalen Welten findet das eigentliche Spieleerlebnis erst durch Interaktionen mit anderen Spielern statt. Um dies zu gewährleisten, ist das Spielen nur über eine Internetverbindung möglich, ferner werden einzelne Teile oder das gesamte Spiel mit Hilfe der Infrastruktur des Publishers betrieben. Dieser sorgt für den reibungslosen Betrieb des Spiels und die kontinuierliche Fortentwicklung der Spielinhalte. Dies ermöglicht es dem Publisher oder den Entwicklern die Aktionen der Spieler stetig zu beobachten. Zudem erhält diese eine starke Kontrolle über die dem Spieler gezeigten Inhalte[1] und können diese möglicherweise in Echtzeit ändern sowie anpassen.

2.2 Big Data

Big Data steht als Synonym für den kontinuierlichen Fortschritt im Umgang mit Daten und charakterisiert den Wandel von klassischen zu modernen Business Intelligence (BI) Ansätzen wie auch Systemen. Bisher existiert in der Literatur kein einheitlicher Konsens über den Begriff Big Data, so dass zur Erklärung oft das Konzept der drei „V“ herangezogen wird [Sc12], [Si12]. Nach diesem Konzept werden Daten hinsichtlich der drei Dimensionen Volume (Datenmenge), Variety (Datenvielfalt) und Velocity (Datenverarbeitungsgeschwindigkeit) charakterisiert. Oft wird auch von den vier „V“ gesprochen, wobei das Vierte für den Begriff Veracity (Richtigkeit) steht. Dieses Konzept wird beispielsweise von dem Unternehmen IBM verwendet.

Das permanente Ansteigen von Datenvolumina kommt durch die zunehmende Digitalisierung der Datensätze, Dateien und Messdaten, die für Analysen zur Verfügung stehen, zusande [Bi12b]. Die Größenordnung ab der Datenmengen als Big Data zählen, hängt von der jeweiligen Branche ab.

[1] Bei klassischen Computer- und Konsolenspielen kann der Spieler unter anderem auf ein Update verzichten. Dies ist bei MMOG nicht möglich, da der Server den Zugang zum Spiel mit einer veralteten Version verhindern würde.

Die Datenvielfalt beschreibt die Anbindung von unstrukturierten und semistrukturierten Daten zu Analysezwecken. Im Gegensatz zu den klassischen Analysen, die sich vorrangig mit strukturierten Daten befassen, werden die neuartigen Analysen durch den Einbezug der nicht strukturierten Daten wesentlich komplexer. Der Fokus von den aktuellen Analysen liegt hierbei auf der Bestimmung der Polarisation von den Freitexten (Stimmung von den Nutzern) oder der Erschließung von den Beziehungsmustern in sozialen Netzwerken [Ag11].

Der Grund für die hohen Verarbeitungsgeschwindigkeiten ist eine permanente wachsende Anzahl an generierten Daten (Sensornetzwerken), welche möglichst in Echtzeit ausgewertet werden müssen, die als Entscheidungsgrundlage fungieren [Bi12b].

Moderne Datenanalysemethoden gestatten die Analyse großer Datenmengen aus unterschiedlichsten Datenquellen in nahezu Echtzeit. Viele dieser Datenanalysen basieren auf statistischen Verfahren, Mustererkennungen oder maschinellem Lernen und analysieren gleichzeitig die strukturierten Daten aus operativen Systemen [va12], [ICH11], [CDH10]. Die Konzepte MapReduce und In-Memory Technolgien bilden die technische Grundlage für solche Analysemethoden. Seit längerer Zeit bieten verschiedene Unternehmen Plattformen und Werkzeuge an, um Big Data Analysen durchzuführen. Besonders das maschinelle Lernen gewinnt durch eine Parallelisierung (MapReduce) der auszuwertenden Trainingsdatensätze bedeutend an Geschwindigkeit und ermöglicht zunehmend validere Analyseergebnisse [Ch06].

3 Datenquellen und Analysemethoden

Zur Identifikation der möglichen Datenquellen folgt dieser Artikel dem Ansatz von Marchant und Henning-Thurau [MH13]. Die Autoren beschreiben den zum Konsumenten gerichteten Teil der Wertschöpfungskette detailliert. Sie gehen von einer vertikalen Ebene aus, der der Spielinhalt, die Spiele-Plattform und der Spiele-Konsument (oder die Spiele-Konsumenten) zugeordnet sind. Die dazugehörige horizontale Ebene beinhaltet neben der Spiele-Plattform, die Kommunikations- (Social Media und traditionelle Medien) und die Vertriebskanäle (physische und digitale Kanäle). Die Autoren ergänzen ihr Framework durch Empfehlungssysteme und Konsumenten/Gesellschaft. Für die Betrachtung in diesem Artikel wird die, um ein Spiel aktive, Community den Spiele-Konsumenten untergeordnet. Da die Community um ein Spiel von großer Bedeutung für das Spieleerlebnis ist [AI03], [GRS11]. Darunter sind alle Aktivitäten der Spieler außerhalb des eigentlichen Spiels gemeint, etwa die aktive Teilnahme in Foren.

Bei der Betrachtung der Datenquellen lassen sich verschiedene Datenarten unterscheiden. Ein Beispiel für strukturierte Daten ist der Kopf (Header) einer E-Mail in dem dedizierte Felder für u. a. Sender und Empfänger existieren, wohingegen der Inhalt (Body) einer E-Mail als unstrukturiert aufgefasst wird. Weiterhin stellen Audio- und Videodaten sowie Bilder unstrukturierte Daten dar. Daten die kontinuierlich und automatisch erfasst werden, werden als Sensordaten bezeichnet. Anhand des Frameworks von Marchand und Henning-Thurau wurden verschiedene Datenquellen

identifiziert. Abbildung 1 zeigt die identifizierten Datenquellen. Die im Folgenden detailliert beschrieben werden.

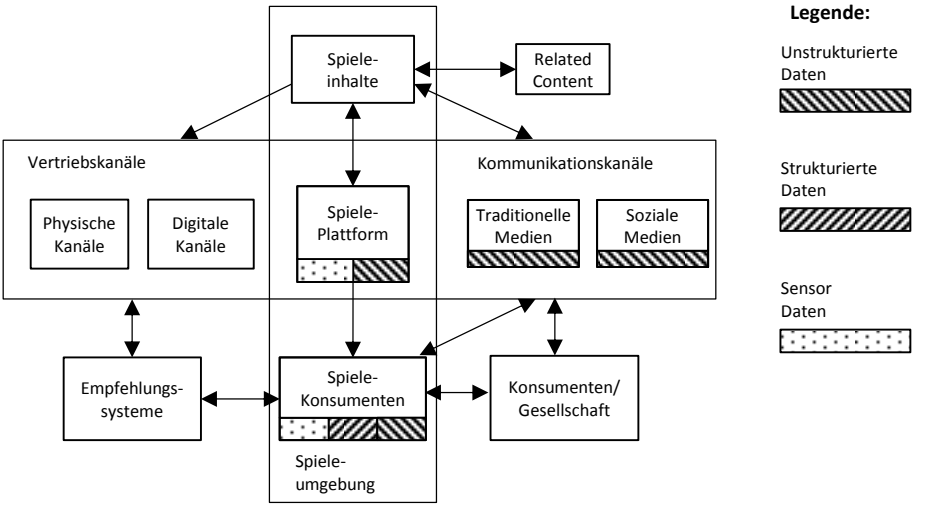


Abbildung 1: Konsumentensicht der Computer- und Videospieleindustrie nach [MH13]

Die gegenseitige Hilfe der Nutzer bei Problemen rund um das Spiel ist ein zentraler Bestandteil der Communities (Spiele-Konsumenten) eines Spiels. Innerhalb von Foren werden geeigneten Strategien oder die Wahl des besten Equipments diskutiert. Ferner gehören die von Spielern erstellte Anleitungen oder Videotutorials zum Bereich der Communities. Aus diesen Datenquellen können unstrukturierte Daten, wie Texte oder Videos, gewonnen werden.

Auf der Ebene der Kommunikationskanäle fallen insbesondere unstrukturierte Daten an. Eine Fülle an Texten oder Ton/Videoträgern können aus klassischen Medien oder dem Social Media gewonnen werden. Dazu gehören Vorberichte und Tests über das Spiel oder kritische Meinungen. Weiter sind diese Inhalte häufig auch in Form von Ton- und Bildmaterial vorhanden.

Betrachtet man den einzelnen Spieler, können zum einen unstrukturierte Daten aus der Kommunikation der Spieler, etwa durch die Kommunikation im Spiel oder durch die Kommunikation in anderen Medien oder Foren, gewonnen werden. Zum anderen liegen Sensordaten aus dem Spiel, wie die sämtlichen Interaktionen des Spielers mit der Spielumgebung oder durch die über das Spiel bestimmte Interaktion mit anderen Spielern (Handel mit Gütern oder der Kampf zwischen den Spielern) vor. Die Spieleplattform selbst liefert Sensordaten über die Performance und Funktion der Spiele. Neben den Aktionen der Spieler lassen sich auch strukturierte Daten aus deren Kaufverhalten gewinnen. Dies betrifft Free to Play Erlösmodelle, bei denen keine regelmäßigen Zahlungen, wie Monatsgebühren, zu erwarten sind, sondern unterschiedliche Transaktionen stattfinden können.

Tabelle 1 beinhaltet die Ergebnisse zur Datenherkunft sowie -art und zeigt, welche Techniken und Analysemethoden jeweils dafür eingesetzt werden können.

Tabelle 1: Geeignete Analyemethoden für Daten aus dem Spielebetrieb

Datenherkunft	Datenart	Kategorie	Analysemethode
Spiele-Konsumenten	strukturiert	Datenanalysen	Fraud Detection
Spiele-Plattform	strukturiert; Sensordaten	Datenanalysen	Predictive SLA
Spiele-Konsumenten	strukturiert; unstrukturiert; Sensordaten	Datenanalysen	Operational Intelligence
Community	unstrukturiert	Textanalysen	Text Opinion Mining
Community	unstrukturiert; schwach strukturiert	Textanalysen	Topic Models
Traditionelle/ soziale Medien	unstrukturiert; schwach strukturiert	Webanalysen	Web Opinion Mining
Community	Graphen; unstrukturiert bis strukturiert	Netzwerkanalysen	Expert Discovery

Die Methoden der **Fraud Detection** identifizieren kriminelles Verhalten interner wie externer Subjekte. Gegenstand dieser Forschung sind Spam- und Zugriffserkennung, Identifikation von Transaktionsbetrug und Entdeckung illegaler Aktivitäten [Ph10]. Mit Hilfe neuer Technologien lassen sich betrügerische Absichten mit höherer Wahrscheinlichkeit und annähernd in Echtzeit erkennen [QS08]. Präventionssysteme nutzen Data-Mining-Techniken, wie Mustererkennung, um auffälliges Verhalten zu identifizieren [MCD13]. Die Identifikation latent krimineller Aktionen erfolgt zumeist durch Auswertung zur Verfügung stehender Attribute, wie Höhe und Frequenz von Transaktionen. Quantifizierbare Attribute werden definierten und unterschiedlich gewichteten Kontrollen unterzogen. Der Gesamtscore einer Transaktion ist das Ergebnis und kann als Indikator für die Sicherheit genutzt werden.

Predictive SLAs stellen eine spezielle Form von Vorhersagen dar, welche durch den Einzug von Service Level Agreements (SLA) zwischen IT und Fachbereich entstanden sind. SLA sind Vereinbarungen zwischen einem IT Service Provider und seinem Kunden, die Service Level Ziele und Verantwortlichkeiten des IT Service Providers vertraglich regeln [Of11]. Im Fokus der Analyse steht die Überprüfung der SLA-Konformität bei der Ausführung der IT Services. Dazu werden historische Daten, wie Zeitpunkt, Dauer und Auslastung, konstant gemessen und zukünftige Entwicklungen durch Extrapolation vorausgesagt [ICH11]. Der Provider kann gegebenenfalls präventiv Kapazitäten derartiger Service Assets erhöhen [Le10]).

Die **Operational Intelligence** ist keine reine Analysemethode, sondern stellt vielmehr die Anwendung von BI-Technologien auf operationaler Ebene eines Unternehmens dar. Es werden alle verfügbaren Eventdaten aus Sensornetzwerken, Data Center Infrastructure Management (DCIM) und Business Process Mining unter Anreicherung externer Informationen in Echtzeit bewertet [CDH10], [Az06]. Ziel der Analyse ist die Vorhersage und Analyse von Events [CDH10]. Im Rahmen der Operational Intelligence werden permanent Abfragen von Bedingungen und Mustererkennungen auf Datenströmen durchgeführt. Mit Hilfe der Operational Intelligence sollen Informationen

generiert werden, die einen möglichst aktuellen Stand aller Daten wiederspiegeln, weshalb derartige Analyse oftmals als Real-time BI bezeichnet werden [Az06], [SR08].

Ziel des **Text Opinion Minings** ist die Identifizierung subjektiver Informationen aus Texten, die von Personen verfasst wurden [PL08]. Im Vordergrund steht hier die Extraktion von Meinungen und Emotionen anhand der Polarität der getroffenen Aussagen. Als Grundlage dienen Ontologien, die im Laufe der Zeit durch maschinelles Lernen kontinuierlich erweitert werden [Li11]. Techniken des Natural Language Processing (NLP) unterstützen die Mensch-Maschine-Interaktion, welche die menschliche Sprache für Maschinen interpretierbar machen [Ch03]. NLP ermöglicht Computern Informationen unter Berücksichtigung der Semantik und Syntax zu gewinnen. Die Techniken umfassen das Clustern von Texten, statistische Methoden, wie Hidden-Markov-Modelle zur Ermittlung der Wahrscheinlichkeit nacheinander auftretender Textbausteine [MS99].

Topic Models beschreiben Algorithmen zur Entdeckung von Themen und Inhalten in umfangreichen Dokumenten, die anschließend verlinkt und geordnet werden [Bl12]. Schlüsselwörter werden mit einer Wahrscheinlichkeit versehen, mit der sie zu einem bestimmten Thema gehören. Aus den identifizierten Schlüsselwörtern werden Ontologien erzeugt, die themenverwandte Begriffe zusammenfassen. Das System lernt anfangs meist mit manuell erstellten Trainingsdaten und entwickelt sich anschließend unabhängig weiter [HKM10]. Zu einer Suchanfrage relevante Inhalte eines Dokuments lassen sich mit Hilfe der Topic Models genauer identifizieren, da die Suchergebnisse themenbasiert auf Grundlage der Ontologien gebildet werden. Im Gegensatz zur Schlüsselwortsuche, bei der die Dokumente in erster Linie nach der Häufigkeit des auftretenden Schlüsselwortes geordnet werden, bewerten Topic Models die Nützlichkeit eines Dokuments anhand der themenrelevanten Inhalte [WC06].

Web Opinion Mining ist in seiner grundsätzlichen Funktionsweise dem Text Opinion Mining gleichzusetzen, wobei der Unterschied in den zugrundeliegenden Daten und den gewünschte Ergebnissen liegt. Web Opinion Mining analysiert das Sentiment von Texten im Web und nicht nur einzelne Dokumente. Hierbei stehen Analysen von Meinungen im Vordergrund, die in sozialen Netzwerken und auf Websites erstellt wurden [BL11]. Weiterhin wird die Auswertung von Meinungen im Bereich sozialer Netzwerke durch die Einbeziehung von Community Detection zusätzlich unterstützt. Das Web Opinion Mining realisiert insbesondere Analysen von Konsumentenmeinungen in als relevant identifizierten sozialen Gruppen, die direkte Auswirkungen auf das Marketing und Produktdesign eines Unternehmens haben [BL11]. Ein weiteres Anwendungsgebiet liegt im Bereich des Crowdsourcing, womit ein Ansatzpunkt zur Identifikation von Problemen auftaucht, die mit unternehmensinternen Ressourcen nicht gelöst werden können [MCD13]. Es wird eine Plattform geschaffen, auf der sich Nutzer oder Experten zumeist in Form von Freitexten zu Fragestellungen eines Unternehmens äußern können. Das Web Opinion Mining kann hierbei zur automatisierten Analyse der Aussagen bezüglich bestimmter Attribute eines Produktes oder Services eingesetzt werden und unterstützt somit aktiv die Recherche in den beim Crowdsourcing anfallenden großen Mengen von Texten, die andernfalls manuell untersucht werden müssten.

Die **Expert Discovery** beschreibt Methoden zur Identifikation von Wissensträgern in sozialen Netzwerken und deren Fähigkeiten in bestimmten Aufgabenbereichen [LLT11]. Hierzu werden vom Nutzer abgedeckte Themen in Texten, Fähigkeitsprofilen und Websites sowie soziale Konstellationen identifiziert [HK10]. Wenn ein Nutzer eine Suchanfrage zu einem Problem stellt, so werden die relevanten Wissensträger im Umfeld identifiziert, die zur Lösung des Problems beitragen können. Grundlage der Analyse ist die Identifikation der Fähigkeiten und die Evaluierung der Professionalität.

4 Diskussion

Die Ergebnisse aus Kapitel 3 zeigen, dass aus dem Spielbetrieb Daten für Big Data Analysen gewonnen werden können. Primär sind es Sensordaten aus dem Spiel und den darin stattfindenden Interaktionen sowie unstrukturierte Daten aus der Kommunikation im Spiel und um das Spiel. Zusammenfassend beinhaltet folgende Tabelle neben den bisher identifizierten Datenarten und Analysemethoden einige mögliche Einsatzzwecke im Spielbetrieb.

Tabelle 2: Anwendungsfälle für Analysetechniken

Datenart	Datenherkunft	Kategorie	Analysemethode	Anwendungsfall
strukturiert	Spielerverhalten	Datenanalysen	Fraud Detection	Anti-Cheating
strukturiert; Sensordaten	Spiele-Plattform	Datenanalysen	Predictive SLA	Ausfall der zugrunde liegenden Systeme
strukturiert; unstrukturiert; Sensordaten	Spiele-Konsumenten	Datenanalysen	Operational Intelligence	Steuerung des Spiels
unstrukturiert	Community	Textanalysen	Text Opinion Mining	Stimmung der Nutzer in Foren oder Sozialen Netzwerken
unstrukturiert; schwach strukturiert	Community	Textanalysen	Topic Models	Themen der Nutzer in Foren und im Spiel
unstrukturiert; schwach strukturiert	Traditionelle/ soziale Medien	Webanalysen	Web Opinion Mining	Wahrnehmung des Spiels in den Medien
Graphen; unstrukturiert bis strukturiert	Community	Netzwerkanalysen	Expert Discovery	Identifikation von Lead Usern

Anti-Cheating: In Spielen, die den Wettbewerbsfaktor unter Spielern unterstreichen oder solchen, in denen virtuelle Güter einen besonderen Wert darstellen, sind Spieler dazu geneigt, sich mithilfe illegaler Methoden, einen spielerischen Vorteil zu verschaffen. Häufig vergeht geraume Zeit, bis dieses Verhalten im Spiel entdeckt wird. Oftmals nur, wenn sich vom Betrug betroffene Spieler an den Spielbetreiber wenden. Hier kann durch die Nutzung von Fraud Detection die Entscheidungen, ob es sich um ein nicht regelkonformes Verhalten handelt, deutlich beschleunigt werden. Durch die Auswertung von Attributen kann regelwidriges Verhalten direkt vom System erkannt werden, z.B. Spieler, die zu hohe Schadenswerte erzeugen oder in kürzester Zeit einen

unüblich hohen Betrag virtueller Währung erlangen. Dies kann validiert oder direkt gemeldet werden. Weiterhin ist das automatische Einleiten weiterer Schritte denkbar.

Die frühzeitige Erkennung von **Ausfällen des zugrunde liegenden Systems**, kann aus der Thematik der Predictive SLA abgeleitet werden. Häufig kommt es beim kontinuierlichen Betrieb von Spielen zu ungewollten Ausfällen. Zwar bestehen keine SLA Vereinbarungen zwischen Spieler und Betreiber, doch müssen die Betreiber den kontinuierlichen Betrieb gewährleisten, um ihre Spieler nicht zu verlieren.

Steuerung des Spiels: Insbesondere die Spieleentwicklung und der Inhalt können durch Big Data Analysen durch Operational Intelligence beeinflusst werden. Die Auswertung des Spielerverhaltens auf Basis der zugrundeliegenden Daten ermöglicht eine direkte Anpassung an selbiges. Hier sind verschiedene Anwendungen denkbar.

Der Schwierigkeitsgrad kann individuell an die Fähigkeiten des Spielers angepasst werden, dabei könnte die künstliche Intelligenz des Spiels die Steuerung an das Verhalten des Spielers und ähnlicher Spieler anpassen. Durch die Anwendung von Mustererkennung auf das Spielerverhalten können typische Spielertypen identifiziert werden, denen dann eine bestimmte Schwierigkeit zugewiesen wird. In fortlaufender Beobachtung kann automatisch evaluiert werden, ob der Spieler mit dem ihm zugewiesenen Schwierigkeitsgrad über- oder unterfordert ist. Theoretisch ist dabei eine Echtzeitanpassung möglich, so dass sich beispielsweise schon der nächste Gegner passender verhalten kann. Dadurch erhält der Spieler einen für ihn optimalen Schwierigkeitsgrad, der die Balance zwischen Frust und Erfolg aufrechterhält.

Dieses ist ebenso bei dem direkten Aufeinandertreffen von Spielern denkbar. Insbesondere in MMOG ist dies ein gängiges Szenario. Überall dort, wo bisher einfache Algorithmen das Aufeinandertreffen bestimmten (z.B. in Abhängigkeit des Levels oder der gesammelten Erfahrungspunkte), können mit Hilfe von Big Data Analysen passende Gegner in Echtzeit gefunden werden. Dies kann deutlich dynamischer durch die Auswahl von Muster erfolgen, etwa durch die Auswertung von ähnlichen Spielstrategien.

Die aus dem Kaufverhalten gewonnen Daten können ebenfalls auf Muster hin untersucht werden. Hier kann insbesondere die Vorhersage von Kaufevents von Bedeutung sein (Zeitpunkt und Umfang der Transaktion). In Folge dessen lassen sich bestimmte Zahlungs-Events steuern oder nur der relevanten Zielgruppe angezeigt werden.

Themen und Stimmung der Nutzer in Foren und im Spiel: Die Analyse der Spielermeinungen und Stimmung in und um das Spiel ist insbesondere für den Entwicklungsprozess von großer Relevanz. Da bei serviceorientierten Geschäftsmodellen das Spiel häufig angepasst wird, beinhaltet dies die kontinuierliche Weiterentwicklung des Spiels. Hier kann neben den Informationen aus dem Spiel (z.B. welche Charakterklasse am häufigsten gewinnt), auch die Meinung der Spieler für die fortlaufende Entwicklungen verwendet werden. Die automatisierte Auswertung einer möglichst weitreichenden Datengrundlage kann wichtige Informationen über Trends und Forderungen der Spielercommunity liefern. Empfundene Mängel im Spieledesign und in der Mechanik, oder gewünschte Add-ons und Weiterentwicklung sind nur einige

mögliche Beispiele. Die Stimmung der Spieler in den verschickten Nachrichten kann Aufschluss über zu große Spielehürden oder frustrierende Spieleelemente geben.

Wahrnehmung des Spiels in den Medien: Die Nutzung und Auswertung von unstrukturierten Daten aus Medien und sozialen Netzwerken können für die Spieleentwicklung genutzt werden, wie Pretests einer Zeitung, die bestimmte Elemente im Spiel kritisieren oder diese in den sozialen Medien diskutiert werden. Aus diesen Informationen können zudem Marketing-Kampagnen geplant werden. Welches Spieleelement beworben werden sollte oder wie Medien auf veröffentlichte Details, wie Spiele-Trailer, reagiert haben, sind zentrale Fragen, die für die Planung von Marketing-Kampagnen von Relevanz sind.

Identifikation von Lead-Usern: Im Rahmen der Spielercommunity gibt es häufig besonders aktive Nutzer, die sich für das Spiel und dazugehörige Themen einsetzten. Unter der Bezeichnung „Lead-User“ nehmen diese eine Meinungsführerschaft ein. Die Identifikation und Reaktion auf diese Nutzer kann zu einer besseren Akzeptanz der anderen Spieler führen. Weiterhin können diese enger in den Entwicklungsprozess eingebunden werden oder Moderationsrollen in der Community übernehmen.

5 Zusammenfassung und Ausblick

Im Rahmen dieses Artikels wurden die Potenziale und Handlungsfelder von Big Data in der Computer- und Videospieleindustrie beleuchtet. Es konnten verschiedene Datenquellen für Analysen ermittelt werden. Die dort gewonnenen Daten wurden aufgrund ihrer Herkunft und ihrer Art einzelnen Analysemethoden zugeordnet. Durch die Betrachtungen der Analysemethoden und deren Übertragung auf die Entwicklung und den Betrieb von Videospielen konnten Potentiale und Handlungsfelder identifiziert werden.

Der vorliegende Artikel nähert sich dem Thema in Form einer konzeptionellen Ausarbeitung, weshalb keine quantitativen Aussagen gewonnen wurden. Für die zukünftige Forschung könnten vor allem zwei Perspektiven von Interesse sein. Einerseits könnte mithilfe qualitativer Studien (z.B. Fallstudien mit Anwendung von Reifegradmodellen) der Status quo und die Anwendbarkeit in der Industrie geprüft werden. Andererseits bedarf es einem wissenschaftlichen Fokus auf die Mechanismen die in Spielen wirken. Durch die Entwicklung geeigneter Modelle kann weitere Forschung einen wesentlichen Beitrag zur effizienten Anwendbarkeit von Big Data leisten.

Literaturverzeichnis

[Ag11] Aggarwal, C.: An Introduction to Social Network Data Analytics. In (Aggarwal, C. Hrsg.): Social Network Data Analytics. Springer Verlag, New York, 2011; S. 1–15.

[AI03] Aoyama, Y.; Izushi, H.: Hardware Gimmick or Cultural Innovation? Technological, Cultural, and Social Foundations of the Japanese Video Game Industry. In Research Policy, 2003, 32; S. 423–444.
[AL07] Arakji, R.; Lang, K.: Digital Consumer Networks and Producer-Consumer Collaboration: Innovation and Product Development in the Video Game Industry. In Journal of Management Information Systems, 2007, 24; S. 195–219.
[Az06] Azvine, B. et al.: Real Time Business Intelligence for the Adaptive Enterprise: The 8th IEEE International Conference on and Enterprise Computing, E-Commerce, and E-Services, 2006.
[Bi12a] Bitkom: Für die junge Generation gehören Computerspiele zum Alltag, 2012a.
[Bi12b] Bitkom: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte. In BITKOM, 2012b.
[BL11] Barbier, G.; Liu, H.: Data Mining in Social Media. In (Aggarwal, C. Hrsg.): Social Network Data Analytics. Springer Verlag, New York, 2011; S. S. 327-352.
[Bl12] Blei, D.: Probabilistic topic models. In Communications of the ACM, 2012, 55; S. 77–84.
[BLR13] Broekhuizen, T. L.; Lampel, J.; Rietveld, J.: New Horizons or a Strategic Mirage? Artist-Led-Distribution versus Alliance Strategy in the Video Game Industry. In Research Policy, 2013, 42; S. 954–964.
[CDH10] Castellanos, M.; Dayal, U.; Hsu, M.: Live Business Intelligence for the Real-Time Enterprise. In (Sachs, K.; Petrov, I.; Guerrero, P. Hrsg.): From Active Data Management to Event-Based Systems and More. Springer, Berlin, Heidelberg, 2010; S. 325–336.
[CG06] Cadin, L.; Guérin, F.: What Can We Learn from the Video Games Industry? In European Management Journal, 2006, 24; S. 248–255.
[Ch03] Chowdhury, G.: Natural language processing. In Annual Review of Information Science and Technology, 2003, 37; S. 51–89.
[Ch06] Chu, C. et al.: Map-Reduce for Machine Learning on Multicore. In (Schölkopf, B.; Platt J.C; Hoffman, T. Hrsg.): Neural Information Processing Systems. MIT Press, 2006; S. 281–288.
[Fi13] Fields, T. V.: Game Industry Metrics Terminology and Analytics Case Study. In (Seif El-Nasr, M.; Drachen, A.; Canossa, A. Hrsg.): Game Analytics. Springer London, London, 2013.
[GRS11] Gidhagen, M.; Ridell, O. P.; Sörhammar, D.: The Orchestrating Firm: Value Creation in the Video Game Industry. In Managing Service Quality, 2011, 21; S. 392–409.
[HEZ14] Hanner, N.; Erek, K.; Zarnekow, R.: Literaturübersicht: Geschäftsmodellkomponenten in der Computer- und Videospieleindustrie: Multikonferenz Wirtschaftsinformatik, 2014.
[HK10] Horowitz, D.; Kamvar, S.: The anatomy of a large-scale social search engine: Proceedings of the 19th International Conference on World Wide Web, 2010; S. 431–440.
[HKM10] Hoang, D.; Kaur, J.; Menczer, F.: Crowdsourcing scholarly data: Proceedings of the 19th International Conference on World Wide Web, 2010; S. 1–8.
[ICH11] Ivanovic, D.; Carro, M.; Hermenegildo, M.: Constraint-Based Runtime Prediction of SLA Violations in Service Orchestrations. In (Kappel, G.; Maamar, Z. N. H. Hrsg.): Service-Oriented Computing - 9th International Conference. Springer Verlag, Berlin/Heidelberg, 2011; S. 62–76.
[Jo06] Johns, J.: Video Games Production Networks: Value Capture, Power Relations and Embeddedness. In Journal of Economic Geography, 2006, 6; S. 151–180.
[Le10] Leitner, P. et al.: Runtime Prediction of Service Level Agreement Violations for Composite Services. In (Dan, A.; Gittler, F.; Toumani, F. Hrsg.): Service-Oriented

Computing. ICSOC/ServiceWave 2009 Workshops. Springer, Berlin, Heidelberg, New York, NY, 2010; S. 176–186.
[Li11] Liu, B.: Web Data Mining – Exploring Hyperlinks, Contents, and Usage Data. Springer Verlag, Berlin, 2011.
[LLT11] Lappas, T.; Liu, K.; Terzi, E.: A Survey of Algorithms and Systems for Expert Location in Social Networks. In (Aggarwal, C. Hrsg.): Social Network Data Analytics. Springer Verlag, New York, 2011; S. 215–241.
[MCD13] Minelli, M.; Chambers, M.; Dhiraj, A.: Big data, big analytics – Emerging business intelligence and analytic trends for today's businesses. John Wiley & Sons, Hoboken N.J., 2013.
[Me09] Messinger, P. R. et al.: Virtual Worlds — Past, Present, and Future: New Directions in Social Computing. In Decision Support Systems, 2009, 47; S. 204–228.
[MH13] Marchand, A.; Hennig-Thurau, T.: Value Creation in the Video Game Industry: Industry Economics, Consumer Benefits, and Research Opportunities. In Journal of Interactive Marketing, 2013, 27; S. 141–157.
[MS99] Manning, C.; Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge, 1999.
[MTP12] Moreno, C.; Tizon, N.; Preda, M.: Mobile Cloud Convergence in GaaS: A Business Model Proposition: 45th Hawaii International Conference on System Sciences, 2012; S. 1344–1352.
[Of11] Office of Government Commerce Service Strategy: ITIL Service Strategy, London, 2011.
[OT11] Ojala, A.; Tyrvainen, P.: Developing Cloud Business Models: A Case Study on Cloud Gaming. In IEEE Software, 2011, 28; S. 42–47.
[Pe13] Peter Warman: Global Monetization of Games. In Newzoo, 2013.
[Ph10] Phua, C. et al.: A Comprehensive Survey of Data Mining based Fraud Detection Research. In Computing Research Repository for Artificial Intelligence, 2010.
[PL08] Pang, B.; Lee, L.: Opinion Mining and Sentiment Analysis. In D.W. Oard, M. Sandernson: Foundations and Trends in Information Retrieval, 2008, 2; S. 1–135.
[QS08] Quah, J.; Sriganesh, M.: Real-time credit card fraud detection – using computational intelligence. In Expert Systems with Applications, 2008, 35; S. 1721–1732.
[RG06] Readman, J.; Grantham, A.: Shopping for Buyers of Product Development Expertise: How Video Games Developers Stay Ahead. In European Management Journal, 2006, 24; S. 256–269.
[Sc12] Schroeck, M. et al.: Analytics: Big Data in der Praxis – Wie innovative Unternehmen ihre Datenbestände effektiv nutzen. In IBM Institute for Business Value, 2012..
[Si12] Singh, S. S. N.: Big Data analytics: International Conference on Communication, Information & Computing Technology, 2012; S. 1–4.
[SR08] Sahay, B.; Ranjan, J.: Real time business intelligence in supply chain analytics. In Information Management and Computer Security, 2008, 16; S. 28–48.
[Ts07] Tschang, F. T.: Balancing the Tensions Between Rationalization and Creativity in the Video Games Industry. In Organization Science, 2007, 18; S. 989–1005.
[va12] van Aalst der, W.: Process Mining – Overview and Opportunities. In ACM Transactions on Management Information Systems, 2012, 3; S. 7:1-7:17.
[WC06] Wei, X.; Croft, W. B.: LDA-based document models for ad-hoc retrieva: SIGIR '06 Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieva, 2006; S. 178–185.
[WCC13] Wu, C.-C.; Chen, Y.-J.; Cho, Y.-J.: Nested Network Effects in Online Free Games with Accessory Selling. In Journal of Interactive Marketing, 2013, 27.
[ZW10] Zackariasson, P.; Wilson, T. L.: Paradigm Shifts in the Video Game Industry. In Competitiveness Review: An International Business Journal incorporating Journal of Global Competitiveness, 2010, 20; S. 139–151.

Steigerung der Effektivität durch Prozessmodellierungs-Tools – Änderbarkeit und Transparenz von Prozessen am Beispiel der Business Simulation Apollo 13

Maike Müller, Laura Lenz, Martin R. Wolf

Institut für IT Organisation & Management
FH Aachen
Eupener Str. 70
52066 Aachen
mueller_maike@gmx.de
l.lenz@itom.ac
m.wolf@itom.ac

Abstract: Die Nutzung von Prozessmodellierungsmethoden oder –werkzeugen kann erheblichen Einfluss auf die Effektivität von Prozessen haben. Das gilt insbesondere für Situationen in denen Personen unter Stress stehen oder ungeübt sind. In diesen Fällen geben Prozessmodelle konkrete Empfehlungen, nach denen sich die handelnden Personen richten können. In Experimenten mit der Business Simulation Apollo 13 haben wir den Effekt des Einsatzes von Prozess-modellierungsmethoden und –werkzeugen untersucht. Bereits bekannte Theorien z.B. über geeignete Verhältnisse von Kommunikationsinhalten konnten bestätigt werden. Darüber hinaus haben wir eine besondere Bedeutung der Übertragbarkeit von Prozessmodellen in konkrete Handlungen identifiziert.

1 Einleitung

Spielerische Simulationen, ob analog oder (teilweise) digital, sind ein effektiver Weg, um Lernenden die Dringlichkeit von vorhandenen Problemen aufzuzeigen. Genau bei dieser Dringlichkeit, dem sogenannten `sense of urgency' setzt die Simulation *Apollo 13 – an ITSM case experience* ™ der Firma *GamingWorks* an: Teilnehmer sollen im Rahmen dieser Simulation erleben, welche Probleme und Tücken die Arbeit in einem IT Service Desk mit sich bringt. Kontextfremd wird hier am Beispiel der wahren Geschichte des Raumschiffs Apollo 13 geübt. Die Crew des Shuttles hatte damals so massive Probleme, dass es am Ende nicht mehr um die Landung auf dem Mond ging, sondern nur noch um das Überleben der Besatzung. Um die Sicherheit der Crew zu gewährleisten muss das Team des Mission Control Centers auf der Erde die technischen Probleme der Astronauten, schnell und systematisch bearbeiten. Dafür ist das Zusammenspiel von einzelnen Kompetenzfeldern notwendig.

In der Regel herrscht insbesondere am Anfang der Apollo 13 Simulation allgemeine Ratlosigkeit. Die Regeln, die eigenen Kompetenzen, sowie das gemeinsame Vorgehen sind weder verstanden, noch konnten sie geplant werden. Im Laufe des Spiels entwickeln die Spieler im Idealfall ein System, also einen Ablaufprozess, der mehr oder weniger

strikt eingehalten wird. Verantwortlichkeiten werden geklärt, verteilt, und der Prozess wird Runde für Runde optimiert. Genannter Prozess wird klassischerweise auf einem Flipchart, einer Tafel o.ä. modelliert und ist die ganze Zeit für alle Teilnehmer sichtbar. Die sich hieraus ergebende Fragestellung ist, ob der Prozess schneller, transparenter, sowie veränderbarer konstruiert und eingehalten werden kann, wenn den Übenden alternative Instrumente bzw. Tools zur Prozessmodellierung an die Hand gegeben werden. Genau das ist die Frage, die mit dieser Arbeit beantwortet werden soll. Denn wäre dies der Fall, ließe das die Hypothese zu, dass firmeninterne Prozesse mithilfe von Modellierungstools einfacher und übersichtlicher gestaltet werden können, um einen effektiven Ablauf zu gewährleisten.

2 Business Simulationen

Generell ist eine Business Simulation, ob analog oder (teilweise) digital, ein interaktiver Workshop, bei dem die Teilnehmer mit Fragestellungen und Problemen konfrontiert werden, die direkt oder indirekt in ähnlicher Weise in ihrem Arbeitsumfeld auftreten können [MT11]. Auch komplexe betriebswirtschaftliche Zusammenhänge lassen sich im Rahmen einer Business Simulation in einem überschaubaren Szenario abbilden.

Ein Kernaspekt einer Business Simulation ist die Zusammenarbeit relevanter Mitarbeiter, also solcher, die auch im „Ernstfall“ in Reaktionsmechanismen eingebunden wären (ibid.). Jeder Teilnehmer bekommt eine Rolle mit bestimmten Aufgaben, Verantwortlichkeiten und Kompetenzen zugewiesen. Neben der Auseinandersetzung mit der eigenen Rolle und den damit verbundenen Aufgaben muss auch die Arbeit des Teams koordiniert und in den meisten Fällen durch einen hierfür ausgebildeten Spielleiter kontrolliert werden. Simulationen sind üblicherweise so konzipiert, dass Defizite in der Teamarbeit zu einem schlechteren Abschneiden des gesamten Teams führt. Aus diesem Grund profitieren nicht nur neu gebildete Teams von dieser Art Training, sondern auch eingespielte Teams, die routinierte Abläufe ungern durchbrechen. Business Simulationen helfen dabei, durch gemeinsame Planung und Umsetzung von entwickelten Arbeitsabläufen die Zusammenarbeit zu verbessern. Auch Ausprobieren ist im Rahmen von Simulationen ohne geschäftsrelevante Fehler möglich. Business Simulationen sollen Wissen vermitteln, dass die Teilnehmer im Arbeitsleben gebrauchen und auch anwenden können. Dabei werden die Lerninhalte so in die Simulation eingebunden, dass die Probanden ihr Wissen anwenden müssen, um die Simulation erfolgreich abzuschließen. Stress und Spaß gehen daher in der Regel Hand in Hand.

In einer Business Simulation können Organisationsprozesse und Strukturen getestet werden ohne den, im Unternehmensalltag vorherrschenden Erfolgsdruck. Es lassen sich auch Ursachen und Wirkungszusammenhänge zwischen Prozessen verdeutlichen [Apn.d.].

2.1 Apollo 13 Simulation

„Apollo 13 – an ITSM case experience ™ “ ist eine Business Simulation, bei der nicht nur Konzepte und Prozesse des IT Service Managements (z.B. ITIL) theoretisch vermittelt werden, sondern auch von den Teilnehmern in einem praxisnahen interaktiven Simulationsspiel direkt erlebt werden (ibid.). Die Simulation basiert auf den Vorkommnissen der NASA Apollo 13 Mission von 1970. Die Mondmission endete beinahe in einer Katastrophe. Nur durch das Zusammenspiel von Teamwork, effektiven Prozessen und Technologie gelang es der Mission Control, das Raumschiff und die Besatzung zu retten und lebend zur Erde zurück zu holen. Während der Simulation werden die Teams mit den gleichen Problemen und Vorkommnissen konfrontiert wie die Beteiligten damals. Teamwork und die Verwendung der IT Service Management Prozesse ist essentiell, um die Simulation erfolgreich zu bestehen und um die „Besatzung“ zu retten.

Für die Apollo 13 Simulation wurde gezielt kein IT Szenario gewählt. Die Teilnehmer werden unter Anleitung eines zertifizierten Teamleiters, der die Rolle der Crew einnimmt, in eine ihnen völlig fremde Situation versetzt, nämlich in das Mission Control Center der NASA. Nachdem die Teilnehmer in der ersten Runde die Rakete zusammengebaut und einen Kommunikationsprozess entwickelt haben, erleben sie den Start der Apollo 13 Mondmission. Zu besetzende Rollen sind hier beispielsweise das CapCom (d.h. Kommunikationsverbindung zur Raumkapsel), der Flight Director und verschiedene fachliche Spezialisten. Von den agierenden Parteien müssen im Verlauf der Simulation Prozesse entwickelt, überarbeitet und simuliert werden, um den Erfolg der Mission und damit das Überlegen der Crew sicherzustellen. Im Laufe der Mission wird das Team mit verschiedenen Störungsmeldungen und Anfragen der Crew konfrontiert. Die ITIL-Prozesse bilden dabei die Basis für Entscheidungen und Problemlösungen.

3 Prozessmodellierung

In jedem größeren Team kann – unabhängig von der Frage, wie die Teammitglieder zusammenarbeiten – Prozessmodellierung eingesetzt werden, um die Zusammenarbeit zu koordinieren. Durch Prozessmodellierung werden Verantwortlichkeiten im Idealfall lückenlos zugeteilt und denkbare Eventualitäten und die Reaktionen hierauf abgebildet. Ideen und Aufzeichnungen hierzu sind so alt wie die Interaktion der Menschen selbst, beginnend in der Steinzeit bis hin zum digitalen Zeitalter.

Definitionsgemäß beschreiben Geschäftsprozessmodelle die Eigenschaften und das Verhalten der Prozessbeteiligten sowie ihre Wechselwirkung mit der technischen und organisatorischen Umgebung [Wi07]. Um Modelle für Geschäftsprozesse zu erstellen muss zuerst analysiert werden, welche Fakten vorrangig betrachtet werden und welche lediglich ergänzend mit einbezogen werden sollen.

Je nach Lerninhalt, verwendet man verschiedene Ansätze zur Modellbildung. Es gibt funktionale, datenorientierte und objektorientierte Ansätze [Fl11]. Diese stellen bei der Modellierung die Funktionen, die Daten beziehungsweise die Objekte in den Vordergrund. Der Paradigmenwechsel von der Ablauforganisation zur Objektorien-

tierung, der in den letzten Jahrzehnten in der Informatik stattgefunden hat, lässt sich deshalb auch in der Modellbildung beobachten. Hier hat sich der Fokus vom Prädikat zum Objekt verschoben. Der objektorientierte Modellierungsansatz ist mit der Unified Modeling Language (UML) zurzeit der am weitesten verbreitete Modellierungsansatz [Ke09]. Ein vergleichsweise neuer Modellierungsansatz ist der subjektorientierte Ansatz [Fl11]. Dieser stellt das Subjekt, also den Handelnden in den Mittelpunkt. Bei den anderen Ansätzen wurde dieser bisher vernachlässigt oder nur am Rande erwähnt, obwohl es für einen funktionierenden Prozess wichtig ist, zu wissen, Wer für Was verantwortlich ist.

Welcher Ansatz zur Modellierung eines Prozesses genutzt wird, hängt auch vom jeweiligen Anwendungsgebiet ab. Werden Prozesse für Menschen dokumentiert, so sollten diese gut nachvollziehbar sein und eventuell auch mit zusätzlichen, erklärenden Anmerkungen versehen werden. Sollen Prozesse hingegen von einem Computersystem automatisch ausgeführt werden, müssen sie andere Anforderungen erfüllen [Al09]. Sollen Prozesse von einer Software ausgeführt werden, müssen sie strikt der vorgegebenen Notation folgen. Als Notation wird eine einheitliche Sprache mit bestimmten Symbolen bezeichnet, die jeweils für die einzelnen Elemente eines Prozesses stehen. Außerdem sind grammatikalische Regeln definiert, die beschreiben wie die Symbole miteinander kombiniert werden dürfen.

Für dieses Experiment wurde sowohl eine Prozessmodellierungsmethode, als auch ein Prozessmodellierungswerkzeug verwendet. Als Modellierungsmethode wurde die subjektorientierte Methode S-BPM ausgewählt. „S-BPM steht für ‚Subjektorientiertes Business Process Management' und bezeichnet eine formale Notation für die Beschreibung und Ausführung von Geschäftsprozessen. Dabei steht das Subjekt, also der einzelne Akteur im Vordergrund" [Men.d.]. Der Vorteil von S-BPM besteht darin, dass die modellierende Person im Vordergrund steht und damit die natürliche „Art und Weise, wie Menschen im allgemeinen Sachverhalte erfassen, denken und kommunizieren. Durch den Fokus auf das Subjekt können Akteure ihre Prozesse aus der Ich-Perspektive modellieren und auch sofort erleben" [Fl11].

Als Prozessmodellierungswerkzeug wurde das Qualitätsmanagementsystem von *ConSense* gewählt. Der Vorteil der ConSense-Suite besteht darin, dass die Prozesse auch ohne Vorkenntnisse des Tools sehr einfach modelliert und verändert werden können [KBK07].

4 Experimente zur Messung der Änderbarkeit und Transparenz von Prozessen

4.1 Hypothesen

Generell war es das Ziel des Experiments, die Veränderlichkeit, die Transparenz, sowie die Effektivität der von den Apollo 13-Teilnehmern erstellten und ausgeführten Prozesse zu messen. Dem Experiment lagen deshalb die folgenden Hypothesen zugrunde, die mit der Versuchsreihe entsprechend getestet werden sollten:

- Die Funktionalität eines Prozesses wird stärker von der verwendeten Modellierungsmethode beeinflusst als durch eine Prozessmodellierungssoftware. Die Untersuchung soll zeigen, dass zur Verbesserung von Geschäftsprozessen eine Software zur Dokumentation und Bereitstellung der Prozesse zwar sehr nützlich ist, aber die Qualität der Prozesse hauptsächlich von der Art der Modellierung beeinflusst wird.

- Mehr Kommunikation bedeutet nicht mehr Effektivität. Kommunikation sollte vielmehr gezielt erfolgen, damit die Nachricht den Empfänger vollständig erreicht und der Empfänger den Inhalt auch versteht. Gehen auf dem Übertragungsweg Informationen verloren oder versteht der Empfänger die Nachricht nicht, kann dies zu Fehlern führen.

4.2 Versuchsaufbau

Zunächst wurden die insgesamt 28 Teilnehmer des Experiments in drei Versuchsgruppen eingeteilt. Den Gruppen wurden verschiedene Modellierungstools zugeteilt, einmal ein Präsentationskoffer, einmal die subjektorientierte Prozessmodellierungsmethode und einmal das Prozessmodellierungstool von der Firma ConSense (letztere wurden den Teilnehmern im Vorhinein durch die Autoren erklärt). Die Freiwilligen, die zumeist aus dem greifbaren Umfeld der Hochschule kamen, waren zu etwa der Hälfte Studenten. Der Rest setzte sich aus verschiedenen Berufsgruppen zusammen.

Die Versuchsreihe bestand aus insgesamt drei Simulationen mit jeweils neun, beziehungsweise einmal zehn Teilnehmern. Die Teilnehmer wurden im Voraus zu ihren Vorkenntnissen in Bezug auf Prozessmodellierung und die Apollo 13-Mission, sowie zu ihrer Computeraffinität und ihrem Verhalten im Team befragt. Die Befragung ermöglichte eine möglichst gleichmäßige Verteilung der Probanden auf die einzelnen Versuchsgruppen (mit Blick auf die Erfahrung im Bereich Prozessmodellierung). Um identische Voraussetzungen für die Gruppen zu schaffen, wurde die erste Runde, in der die Rakete eigentlich zusammengebaut wird, nicht gespielt. Den Gruppen wurde die Rakete mit allen Komponenten vorgegeben, um zu vermeiden, dass der weitere Spielverlauf und das Ergebnis durch unterschiedlich hochwertige Raketen, sowie den sich hieraus ergebenden finanziellen Aufwand beeinflusst werden. Gespielt wurden die Runden zwei, drei und vier.

Für die Simulation wurde der Raum nach den Vorgaben des Apollo 13-Handbuchs hergerichtet. Die Tische und auch Sitzplätze der einzelnen Rollen sind darin vorgeschrieben (Abb. 1). Den Teilnehmern war es nicht erlaubt, die Position der Tische während der Simulation zu verändern. Ergänzt wurde die Raumausstattung noch durch zwei Kameras und Mikrophone, um die Simulation zu Auswertungszwecken aufzuzeichnen. Bei der dritten Versuchsgruppe, die mit einer Software arbeitete, wurde die Raumausstattung zusätzlich durch einen Computer und einen zusätzlichen Beamer erweitert.

Die Versuchsgruppen bekamen für ihre Simulation verschiedene Vorgaben. Gruppe 1 spielte die klassische Version der Simulation. Das heißt, dass diese Gruppe keine Vorgaben darüber bekam, wie sie den Prozess modellieren sollte. Ihr einziges Hilfsmittel

war ein Moderationskoffer. Die Gruppe 2 sollte ihren Prozess mithilfe der Subjekt-orientierten Modellierungsmethode (S-BPM) entwickeln und bekam ein Flip Chart, gesonderte Modellierungsblätter mit Beispiel und einen Moderationskoffer als Hilfsmittel. Die dritte Gruppe bekam zur Modellierung die Prozessmanagementsoftware der Firma ConSense. Diese wurde den Teilnehmern kurz erklärt um zu vermeiden, dass ein Bedienungsproblem das Ergebnis negativ beeinflusst.

Nach der letzten Runde der Simulation wurden die Teilnehmer zu ihren Erfahrungen während der Simulation befragt. In den Fragebögen wurden die Teilnehmer zu Aspekten der Teamarbeit, zu angewendeten Prozessen und Vorgehensweisen, sowie zur Kommunikation befragt. Zur Beantwortung der Fragen wurde eine Likert-Skala verwendet. Die Teilnehmer konnten aus fünf Antwortmöglichkeiten wählen (sehr klar, klar, neutral, unklar, sehr unklar). Um die durch den Fragebogen gesammelten Daten vergleichen und darstellen zu können, wurde aus den Werten der gewichtete Mittelwert berechnet. Dafür wurden den Antwortmöglichkeiten die Zahlenwerte von eins bis fünf zugeordnet.

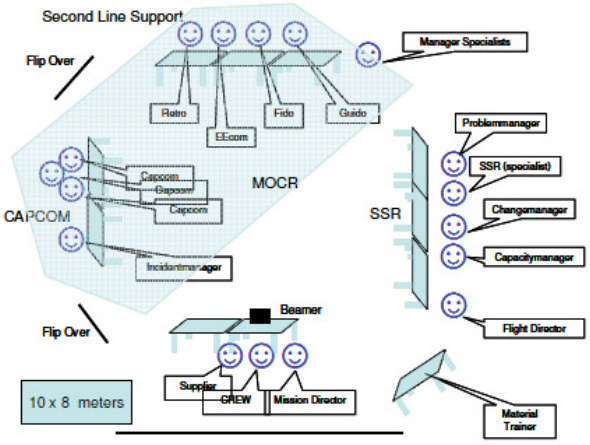


Abbildung 1: Aufbau Apollo 13 Simulation (vgl. Teachers Manual Apollo 13)

5 Ergebnisse und Beobachtungen

Bevor im Folgenden die Ergebnisse präsentiert werden, wird darauf hingewiesen, dass diese nicht den Anspruch der empirischen Validität erfüllen, da hierfür zahlreiche weitere Versuche mit weit mehr Versuchspersonen hätten durchgeführt werden müssen. Die nachstehenden Ergebnisse sind somit indikativer Natur, erlauben aber Schlussfolgerungen, die in weiteren Testreihen validiert werden müssten.

5.1 Kommunikation

Innerhalb eines Teams ist Kommunikation ein wichtiger Faktor, der aber oft erst näher betrachtet wird, wenn es zu Missverständnissen und Problemen kommt. Hierbei ist zu beachten, dass die Informationen vollständig sein und die Kommunikation gezielt erfolgen sollte [WWF13]. Werden Informationen nicht vollständig weitergegeben, kann dies zu Problemen führen, da Aufgaben zum Beispiel nicht korrekt oder nicht vollständig gelöst werden können. Erfolgt Kommunikation nicht gezielt, also nicht direkt an den Empfänger gerichtet, besteht das Risiko, dass die Nachricht verloren geht.

Die Qualität der Kommunikation innerhalb der Teams wurde von allen Gruppen als zufriedenstellend eingeschätzt. Bei der Frage, ob die übermittelten Informationen vollständig und korrekt waren, gehen die Ergebnisse allerdings auseinander

Während bei Gruppe 2 und 3 die Werte nach Runde 1 stark ansteigen und in einem guten Bereich liegen, steigt der Wert von Gruppe 1 erst nach der zweiten Runde stark an, erreicht aber nicht das Niveau der anderen zwei Gruppen (Abb. 2). Außerdem war auffällig, dass Gruppe 3 während der Simulation im Gegensatz zu den anderen Gruppen sehr ruhig interagierte. Es war sehr leise im Raum, die gesamte Interaktion wirkte gesteuerter und die Gruppe zeigte eine Art „Herdenverhalten“. Bei der Übermittlung wichtiger Informationen, wie z.B. bei der Modellierung der Prozesse, war oft die komplette Gruppe an einem Ort versammelt. Die Werte lassen darauf schließen, dass eine gezielte und ruhige Kommunikation zu besseren Ergebnissen führt.

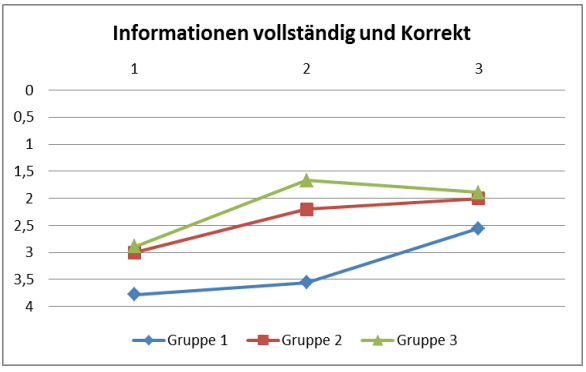


Abbildung 2: Informationsgehalt

Im Zusammenhang mit der Kommunikation ist auch das Verständnis für die Aufgaben und Zuständigkeiten der eigenen Rolle ein wichtiger Aspekt [WWF13]. Gruppe 3 ist hierfür ebenfalls ein gutes Beispiel. In Runde 1 wurde das Verständnis der eigenen Rolle als sehr niedrig bewertet. In Runde 2 steigt dieser Wert signifikant an. Parallel dazu ist zu beobachten, dass sich auch die Werte für die Kommunikation nach Runde 1 verbessern. Lag der gewichtete Mittelwert für die Frage „Hatten Sie Schwierigkeiten den

richtigen Ansprechpartner zu finden?“ in Runde 1 noch bei 3,44, stieg der Wert in Runde 2 auf 1,78 an.

5.2 Inhalt der Kommunikation

Zur Analyse der Kommunikation wurden die Teilnehmer in der Nachbefragung dazu aufgefordert, den Inhalt ihrer Kommunikation prozentual in drei Kategorien zu gliedern: Inhalt, Vorgehen und Beziehung.

Die Einteilung und Bewertung erfolgte nach den Vorgaben von R. Fisch [Fi94]. Dieser analysierte in dem Text „Eine Methode zur Analyse von Interaktionsprozessen beim Problemlösen in Gruppen“ das Kommunikationsverhalten von Diskussionsgruppen. Agiert eine Gruppe effektiv, sollte sich die Kommunikation zu 25% - 30% mit dem Vorgehen, zu 40% - 60% mit dem Inhalt und mit bis zu 20% mit der Beziehung beschäftigen. Auch Badke-Schaub (2001) und Lüdcke et al. (2000) argumentieren ähnlich [BS01 & LBWSBS00]. Die Kommunikation sollte zu zwei Dritteln aus Inhalt und zu einem Drittel aus dem Vorgehen bestehen. Der Beziehungsanteil sollte hingegen nur einen sehr kleinen Prozentwert erreichen.

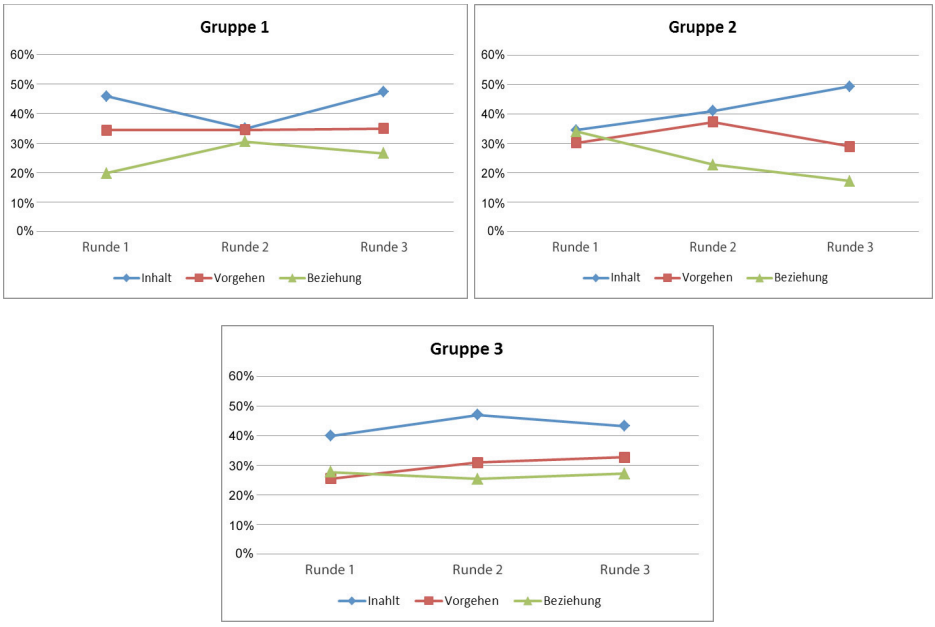


Abbildung 3-5: Rundenauswertung

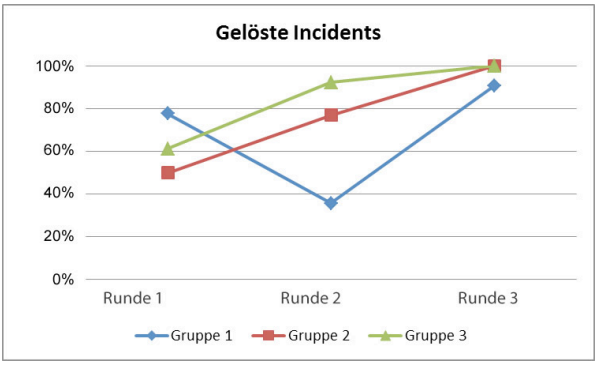


Abbildung 6: Gelöste Incidents

Für eine Stresssituation – wie durch die Apollo 13-Simulation erzeugt – war es ebenfalls zutreffend, dass sich nur ein geringer Teil der Kommunikation mit der Beziehung zu anderen befasst. Abbildung drei bis fünf zeigen die Werte für die Aufteilung der Kommunikation der Gruppen für die drei Runden. Vergleicht man die Diagramme ist auffällig, dass sowohl Gruppe 1, als auch Gruppe 2 jeweils eine Runde aufweisen, die im Gegensatz zu den anderen Runden stark negativ abweichende Werte zeigt. Bei Gruppe 1 ist es Runde 2 und bei Gruppe 2 ist es Runde 1. Die Kommunikationsanteile verteilen sich hier nahezu gleichmäßig auf die drei Kategorien. Aus den Ergebnissen lässt sich ablesen, dass Gruppe 1 und Gruppe 2 in den betreffenden Runden mit Abstand die geringste Incident Lösungsquote erzielt haben (Abb. 6).

5.3Transparenz und Änderbarkeit der Prozesse

Damit Prozesse von allen Beteiligten auch angewendet werden können, muss eine gewisse Transparenz geschaffen werden. Der modellierte Prozess muss demnach für jeden, der daran beteiligt ist, zugänglich gemacht werden. Jede Gruppe der Versuchsreihe hatte dafür ihre eigene Strategie. Da Gruppe 1 keinerlei Hilfsmittel außer einem Moderationskoffer zur Modellierung zur Verfügung gestellt bekam, nutzten sie die sich darin befindlichen Karten, um ihren Prozess damit auf dem Boden dar zustellen (Abb. 7). Durch dieses Vorgehen war der Prozess jederzeit für jeden sichtbar und zusätzlich blieb er veränderbar. Gruppe 2 verwendete zum einen die Modellierungs-blätter für den individuellen Prozess, die jeder an seinem Platz zur Verfügung gestellt bekommen hatte. Zum anderen hat die Gruppe den übergreifenden Anteil des Prozesses auf einer Pinnwand angezeichnet (Abb. 8). Auf diese Weise war der Prozessanteil, der für alle von Bedeutung war, zwar für alle sichtbar, allerdings nicht mehr so einfach abzuändern. Die dritte Gruppe hatte den Vorteil, dass sie als Vorgabe eine Modellierungssoftware gestellt bekam. Die Software ermöglichte eine gute Änderbarkeit und durch den angeschlossenen Beamer war der Prozess ebenfalls für jeden sichtbar.

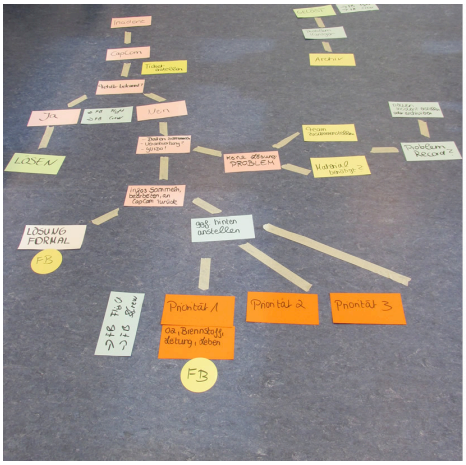


Abbildung 7: Prozess Gruppe 1

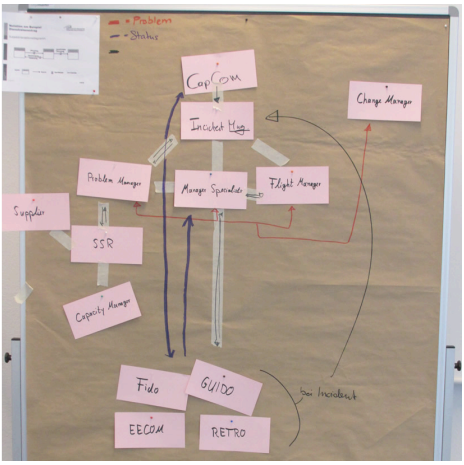


Abbildung 8: Prozess Gruppe 2

Betrachtet man die Arbeitsweisen der Gruppen lassen sich auch die Werte bei der Frage nach der Änderbarkeit des Prozesses erklären. Die Gruppen bewerteten die Änderbarkeit der modellierten Prozesse zwar von Anfang an als gut, die Werte von Gruppe 2 waren jedoch die schlechtesten.

5.4 Zusammenhang zwischen Methode und Ergebnis

Bei Gruppe 1 wurde das Vorgehen vor der eigentlichen Prozessmodellierung detailliert besprochen und erst dann wurde mit der Modellierung des Prozesses begonnen. Den Prozess modellierte die Gruppe mittels Karten auf dem Fußboden. So konnten sie die einzelnen Bestandteile bei Bedarf verschieben. Dadurch erreichte die Gruppe eine hohe Transparenz und gute Änderbarkeit bei ihrem Prozess. Die komplette Gruppe beteiligte sich an der Absprache wie vorgegangen werden soll und beteiligte sich auch aktiv an der Modellierung.

In Runde zwei, in der die Gruppe am schlechtesten abschnitt, war auffällig, dass die Probanden etwas hilflos beim Lösen der Probleme wirkten. Insgesamt wurden auch nur ca. 35% der Incidents gelöst. Bei der Übermittlung der Lösungen fehlten teilweise wichtige Informationen. Das kostete die Gruppe Zeit und minderte den Erfolg.

Gruppe 2, die ihren Prozess nach S-BPM Vorgaben modellieren sollte, bekam zusätzlich zu dem Moderationskoffer auch ein Flip-Chart. Bei Gruppe 2 war auffällig, dass die Modellierungsmethode vom Team nicht angenommen wurde. Schon bei der Erklärung der Methode traten Unklarheiten auf. Die Gruppe beschäftigte sich vorrangig mit der Modellierung des Hauptprozesses, aber kaum mit ihren internen Prozessen. Insgesamt verwendete die Gruppe die Methode nur sporadisch. Den Hauptprozess (Subjektinteraktionsdiagramm) erstellte die Gruppe nach ihren eigenen Vorstellungen. Der interne Prozess (Subjektverhaltensdiagramm) wurde nur von einem Teil der Probanden erstellt.

Die Kommunikation von Gruppe 2 wirkte in Runde 1 ziemlich unkoordiniert. Lange Zeit konnte gar kein Incident gelöst werden. Insgesamt konnte die Versuchsgruppe in Runde 1 nur die Hälfte der eingegangenen Incidents lösen.

Gruppe 3 sprach die vorhandenen Informationen (z.B. wie die einzelnen Rollen zusammenarbeiten sollen) zunächst gemeinsam durch und modellierte dann den Prozess gemeinsam am PC. Von Anfang an wirkte die Gruppe während der Spielphasen ruhiger als die anderen Gruppen. Viele auftretende Probleme wurden, auch während der Spielrunden, gemeinsam besprochen. Die ruhige Herangehensweise und die gute Teamarbeit führten zusammen mit der gut angenommenen Software bei Gruppe 3 zu positiven Ergebnissen. Die Ergebnisse steigerten sich von ca. 60% der gelösten Incidents in der ersten Runde bis zu 100% in der letzten Runde.

6 Interpretation und Fazit

Mit dem Apollo 13-Experiment konnten einige bekannte Theorien aus der Kommunikationswissenschaft nachvollzogen werden. Die Verteilung der Kommunikationsinhalte hatte, ebenso wie die Vollständigkeit der Informationen, einen Einfluss auf die Effektivität der einzelnen Gruppen. Auch zeigt sich, dass der Interaktionsaufwand durch gezielte Kommunikation verringert werden kann.

Interessant ist jedoch die Beobachtung, dass trotz vermeintlich gleich großer Transparenz der Prozesse unterschiedlich stark auf die Prozesse als Handlungsleitfaden zurückgegriffen wurde. Insbesondere bei Gruppe 1 und 2 konnte beobachtet werden, dass sich die Teilnehmer nicht an die vereinbarten Prozesse hielten, was sich dann auch in schlechteren Ergebnissen wiederspiegelte. Es reicht offensichtlich nicht aus, an der Erstellung des Prozesses beteiligt gewesen zu sein und den Prozess vor Augen zu haben.

Unsere Vermutung, die wir auch mit weiteren Experimenten empirisch belegen möchten, ist, dass von drei Faktoren abhängt, ob Personen einem modellierten Prozess (besser) folgen oder nicht:

1) kognitive Einbeziehung der involvierten Personen bei der Modellierung des Prozesses

2) Übertragbarkeit des Prozessmodells in die konkreten eigenen Handlungen

3) Flexibilität/Änderbarkeit des Prozessmodells zur Anpassung an Prozessänderungen

Insbesondere der zweite Punkt ist interessant, weil die einfache Übertragbarkeit von Prozessmodellen in konkrete Handlungen bei den meisten Modellierungsmethoden keine Rolle spielt. Die korrekte Abbildung der Handlungsstränge spielt meistens eine größere Rolle als die Frage, ob ein bestimmtes Symbol (z.B. ein Rechteck oder eine Raute) leicht in die entsprechende Handlung umgesetzt werden kann.

Prozessmodelle, die leicht in konkrete Handlungen zu übersetzen sind, müssen Personen hauptsächlich in Situationen unterstützen, in denen sie sich nicht sicher sind, welche Handlung sie als nächstes ausführen sollen. Nur in diesen Fällen greifen die Personen

auf die Prozessmodelle zurück. Die Prozessmodelle müssen dann erstens Orientierung geben (Wo im Ablauf bin ich gerade?), und zweitens eine Handlung vorschlagen (Was muss ich jetzt machen?), die dann auch möglichst einfach von einem abstrakten Symbol in eine konkrete Handlung übersetzt werden kann. Hier könnten z.B. Bilder oder Szenarien, aber auch kognitive Anker oder eingeübte Denkmuster zum tragen kommen.

Die subjektorientierte Modellierungsmethode S-BPM bietet hier einen vielversprechenden Ansatz. Zwar hat die Anwendung der Methode in den Experimenten nicht zu besseren Ergebnissen geführt, dies führen wir jedoch darauf zurück, dass die Methode zu Beginn nicht hinreichend erklärt und ausprobiert worden ist. So waren die positiven Effekte, die der Einsatz der Methode mit sich bringen könnte, den Versuchsteilnehmern nicht klar.

Literaturverzeichnis

[A09] Allweyer, T.: BPMN 2.0 - Business Process Model and Notation: Einführung in den Standard. Books on Demand GmbH, Norderstedt, 2009.

[Apn.d.] Apollo 13 (unbekannter Autor). (nicht datiert). Apollo 13 – an ITSM Case Experience. http//www.apollo13-simulation.de, aufgerufen am 30.4.2014.

[BS01] Badke-Schaub, P. [e-Journal]: Kritische Situationen als Analyseeinheit komplexer Handlungen. In Trimpop, R.; Zimolong, B.; Kalveram, A.: Psychologie der Arbeitssicherheit und Gesundheit. Neue Welten – Alte Welten. 11. Workshop 2001, Asanger, Heidelberg, 2001; S. 137-142.

[Fi94] Fisch, R. [e-Journal]: Eine Methode zur Analyse von Interaktionsprozessen beim Problemlösen in Gruppen. Gruppendynamik, 25, 1994; S. 149-168.

[Fl11] Fleischmann, A.: Subjektorientiertes Prozessmanagement: Mitarbeiter einbinden, Motivation und Prozessakzeptanz steigern. Hanser, München, 2011.

[FSSOB11] Fleischmann, A.; Schmidt, W.; Stary, C.; Obermeier, S.; Börger, E.: Subjektorientiertes Prozessmanagement Mitarbeiter einbinden, Motivation und Prozessakzeptanz steigern“ Hanser, München, 2011.

[Ke09] Kecher, C.: UML 2 – Das umfassende Handbuch. Galileo Computing, Bonn, 2009.

[KBK07] Killich, S.; Bruns, I.; Künzer, A.: Qualitäts- und Integrierte Managementsysteme in Unternehmensnetzwerken. In: Becker, T.; Dammer, I.; Howaldt, J.; Killich, S.; Loose, A.: Netzwerkmanagement. Mit Kooperation zum Unternehmenserfolg, 2. Auflage. Springer, Berlin. 2007.

[LBWSBS00] Lüdcke, R.; Birkhofer, H.; Wallmeier, S.; Stempfle, J.; Badke-Schaub, P [e-Journal]: Effizienzverbesserung durch gezielte Führung. Optimierung eines Produktentwicklungsprozesses. ZWF, 2000, 95, S. 398-402.

[Men.d.] Metasonic (unbekannter Autor). (nicht datiert). Subjektorientiertes Geschäftsprozessmanagement. http://www.metasonic.de/sbpm, aufgerufen am 30.4.2014.

[MT11] Metz, M.; Theis, F.: Digitale Lernwelt – Serious games. Bertelsmann, Bielefeld, 2011.

[WWF13] Watzlawick, P.; Weakland, J.; Fisch, R.: Lösungen – Zur Theorie und Praxis menschlichen Wandels, 8. Auflage. Hans Huber, Bern, 2013.

[Wi07] Wilhelm, R.: Prozessorganisation, 2. Auflage. Oldenbourg, München, 2007.

A game-based evaluation model for a successful cooperation in cloud computing

Stine Labes

Chair of Information and Communication Management
Technische Universität Berlin
Straße des 17. Juni 135
10623 Berlin
stine.labes@tu-berlin.de

Abstract: While the cloud concept has passed the hype of emerging technologies, community clouds became more popular, but are rarely addressed in publications so far. Within our research, we iteratively developed a business board game that contributes to give a more elaborate and realistic user experience for a complex cooperative situation and the strategy development in a community cloud. With action research, the game closely fits to a relevant defined problem and we validated its efficacy within a workshop with expert IT practitioners. Finally, we conclude with benefits for academic research and practical implementations.

1 Introduction

Over the last few years, cloud computing got established as a technology concept in the information technology (IT) industry. With this concept, customers can obtain freely scalable IT resources (e.g. servers, storage, applications, and network resources) in an on-demand manner via networks (intranet or internet) and pay usage-based fees [MG09], [We09]. These resources "as a Service" avoid long-term capital expenditures, can load for cost reductions and promote the flexibility of IT provisioning, particularly in the budget-limited public sector.

There exist four cloud provisioning models with an increasing legal separation of clients combined with a decreasing opening to publically accessible network resources: public (multi-tenancy, internet), private (single-tenancy, intranet), hybrid (combination of private and public), and the community cloud (restricted multi-tenancy, merger of intranets from several organizations). While the hype on public and private cloud services seems to be over, the community model rises as a relevant provisioning model in practice [Ga13], [Ha10], [Bu12]. Organizations with the same needs, legal requirements, goals or the same data can cooperate, merge their intranets, and use a common IT resource pool without dependences on third party vendors [BM09b]. This will overcome the lack of scalability and resource efficiency of private clouds and the loss of trust with public cloud services [BM09b], [Li12], [Re11].

As Henry Ford already mentioned, “Coming together is a beginning; keeping together is progress; working together is success” (Henry Ford, 1863-1947). For participating in a long-term cooperation, a high involvement is required for the community that takes into account the cooperative and competitive behaviour for each member. To cope with the sensitivity for this situation in advance, members will benefit from a simulation situation where they can test different strategies before they come to the decision to take part in the contract-related productive mode of the community cloud.

In this study, we are interested in examining if a business board game can contribute to a more elaborate and realistic user experience for the cooperative situation, the decision-making and the strategies to be pursued. For this approach, we will first give some theoretical background on cooperation in cloud computing and related work on the promoted game context. Afterwards we introduce our research approach and formulate the research questions. Then the background for the business board game is described as well as the game design and the evaluating experiments with IT service providers. The outcomes of our research will be relevant to comparable complex situations. We end with a discussion of the results and draw conclusions for academic research and practice.

2 Theoretical Foundation

Embodying cooperation in cloud computing, a community cloud describes a cloud provisioning model that “aspires to combine distributed resource provision from Grid Computing, distributed control from Digital Ecosystems and sustainability from Green Computing […], while making greater use of self-management advances from Autonomic Computing” [BM09a]. Following this conceptualization, we further propose the differentiation of two types of communities in the cloud: provider-based and customer-based community clouds (see Figure 1). In offering or using a distinct cloud service, either the providers or the customers can collaborate and cooperate.

A customer-based community cloud can be established if the same cloud service is used and shared by an association of customers. For example, a customer community cloud can be a governmental cloud service (e.g. software application) offered by one big cloud service provider for the cooperating institutions of a dedicated country or state with similar needs and interests. A provider-based community cloud is a service offered by a merger of cloud service providers with similar services and objectives. For instance, they collaborate to collectively provide a cloud service, to extend their range of services or to share their IT resource pools to gain resource efficiency. An exemplary scenario for a provider community is the assurance of high availability for services (e.g. disaster management). A mutual community cloud, i.e. a provider- and a customer-based community cloud at the same time, is for example a community of research and education facilities, where all community members have access to actual research data and can support each other by processing this research data. With the use of private clouds, high requirements for data protection are met as well. Therefore, also medical units or other fields with critical data can profit from this provisioning model. With our game-based research, we address the provider-based community cloud types.

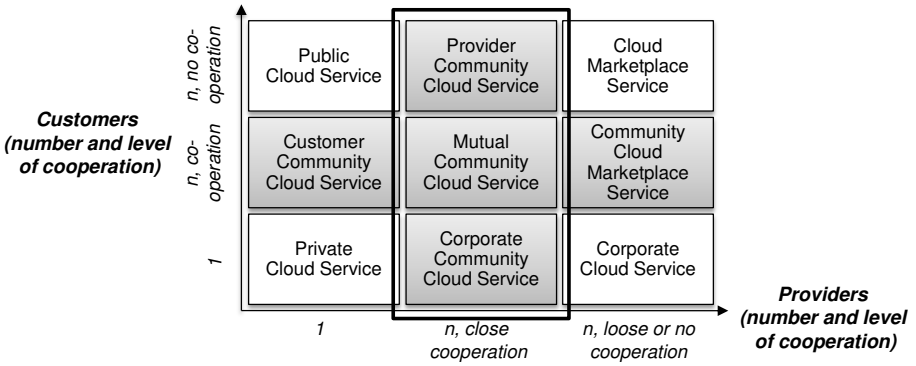


Figure 1: Types of a community cloud

3 Related Work

In academic research, community clouds are eminently underrepresented. Only a few authors have focused on this cloud provisioning model so far. Practical implementations of customer-based communities are seen in the public, medical or gaming sector, while provider-based communities are rare and still laboratory projects only.

Reasons for this gap can be high community entry barriers, high risks and investments, but uncertainty of success and a late reward. To address these issues, a possibility is required, where potential members can test their real behaviour in a simulated environment [LPN09] to decrease the uncertainty. Traditional methods for this goal do not provide the optimal opportunity to link abstract concepts with the background in practice [BC08] or have a higher financial effort [St08]. Therefore, serious games are proposed as evaluation tool [LPN09]. These games are systems with a specific intent to assist, persuade or deceive the player's behaviour [BGM87], [CP09], [Fo03], [OH09]. Other authors focus on the user activity of play and call it "meaningful play" [SZ04]. A similar approach are games with a human computer interface, supported by different authors. To support the practical context of such games, the design of the game must closely describe the context situation and decision processes [Be95], [Jo03], [Jo03]. Moreover, the game should differentiate which elements needs to be controlled or eliminated to finally get a relevant model [DG12] for community (cloud) cooperation in practice. A game manager plays an important role and provides immediate feedback to the players [Ki05]. Complementing the considerations, some authors provide summaries of related work for serious gaming in the business context [HV09], [LLS13], [Fa09].

Similar approaches are implemented in other research projects. One publication deals with a game-based approach for a similar complex situation, where astronauts can test personalized support situations in advance of a long duration mission [Sm10]. Other authors validate the game simulation method as a vehicle for testing an environment of decision-making for decision support systems [Be09] or the use of an enterprise resource planning system [FH11]. Complementing this, some authors propose a business game

taxonomy to develop an international and open database for business games [GBN13] or to classify game-playing research [LLS13]. Considering this classification, game research rather focuses individualistic but less cooperative or competitive game design so far [LLS13].

4 Research Approach

The goal of our research is to develop a game that emulates the situation of a provider-based community cloud to promote the imagination of the players. This game should help to analyse the players' behaviour to finally derive implications for the success of a community cloud. Following this intention, we distinguish two research questions:

1. How should a business board game look like to emulate a provider-based community cloud and create a realistic experience for participating players?

2. Does the business board game produce a more elaborate user experience for the cooperative situation, the decision-making and the strategies to be pursued?

Within the scope of design science research [He04], we use a multi-method approach to follow a proposed four-step game development approach [LPN09]. First, the (1) problem definition is answered with action research, where we accompanied a community cloud project, conducted several interviews and a workshop with the community cloud members. From the results, we derive preconditions, basics and design principles for the (2) game conception. The game design aims to fulfil the development principles of persuasive system design [OH09], [PT12] for serious games or "games with a purpose" [AD08], [De11], [HSA12]. These guidelines propose e.g. enjoyment factors, an accurate game result insurance, and evaluation measures [AD08], [De11], [HSA12]. After the rough conceptualization, we (3) play the game and iteratively revise the specifications. Finally, we (4) analyse the game and validate the efficacy of the game approach with a workshop experiment with four practitioners having expertise in IT service providing but not as a community cloud member.

5 Community Cloud Game

5.1 Problem Definition

The game bases on the practical intention to implement a community cloud as part of the project "Government Green Cloud Laboratory (GGC-Lab)". The GGC-Lab project is funded by the German Federal Ministry of Economics and Energy to analyse the possibilities of environmental sustainability and resource efficiency of the IT in the public sector. The project covers the use case of providing intelligent and cloud-based software application services based on a community of cooperating IT service providers. The project consortium consists of currently four public IT service providers, where each provider serves his own customers in the public sector. While those customers use

software applications, the IT service provider needs to handle the resulting computing load. In the GGC-Lab, this load can be decentrally processed on the distributed IT infrastructures of the community members. In case of capacity shortage or cost saving targets, each member has the choice to transfer the load to the community. Based on efficiency criteria, a resource controller balances the load to save costs and generate the best community profit. The efficiency is described by the energy costs for processing the load order. Other balancing strategies are considered as well, like the amount of energy, the emission of carbon dioxide or a sustainable energy mix. Summing up, the project intends to optimize the IT utilization, get a flexible resource approach, consolidate and standardize the service provisioning, increase the service quality and finally reduce the service costs.

This situation is emulated with the game, where the players benefit from a simulation framework for developing and testing strategies of potential community cloud members. Within the community game, the players have to process different load orders, can optimize their resource efficiency, and will be influenced by environmental events.

5.2 Game Conception

Along with the project GGC-Lab, the board game provides a fixed framework for long-term community cooperation. Static features display the community cloud construction and describe the game design. Procedural features model the community cloud processes and describe the playing process.

5.2.1 Static features of the game conception

The design of the board game considers a game board that represents the community cloud. Further, every player has a smaller personal board that depicts the own infrastructure resources (building and IT capacity) and is connected with the community game board. A deck of shuffled load order cards is positioned in the middle of the game board with the backside facing up. Controlled by a thrown dice, the players move on a game path in predefined rounds (see Table 1), influenced by random events:

- **Sit out**: “Your community contract is in revision, you have to miss a round.”
- **Heat**: “The temperature is rising and causes extra cooling effort for your IT. Your costs per processed unit increase about +5€.”
- **Cold**: “The temperature is decreasing and causes a less cooling effort of your IT. Your costs per processed unit decrease about -5€.”
- **Increasing energy costs**: “The energy price at the energy stock market increases. Your costs per processed unit increase about +5€.”
- **Decreasing energy costs**: “The energy price at the energy stock market drops. Your costs per processed unit decrease about -5€.”
- **Server infrastructure failure**: “Problems with your IT infrastructure force you to deliver your received load order to the community.”
- **Network failure**: “Problems with your network connection force you to process your received load order on your own.”

Feature	Characteristic	Background of the problem definition
Path event	The players move on a path of random event- and non-event-fields.	By operating a data centre, changing conditions and not anticipated events occur.
	The event fields influence the efficiency and the ability of the player to process or deliver load orders.	The data processing can be influenced by random environmental, technical, or contractual impacts.
Playthroughs	Fixed number of playthroughs per game, 2 to 4 are recommended.	This means the medium-term view of the cooperation.
Rounds	Fixed number of 4 rounds per playthrough.	It is the short-term equivalent with accounting periods.
	Changing starting player per round in clockwise direction.	This covers the varying order situation and balance the positions of the members.

Table 1: Static features of the game path

Moving on the game path, the players can optimize their own infrastructure resources (building and IT capacity) and need to process load orders. Load units are represented by cubed gaming pieces that differ in colour (shades of grey) to show the payment value of the load unit. Other cubed gaming pieces (white) signify capacity units in optimization mode. The concrete design of these features is justified with the practical background of the projects' problem definition (see Table 2).

Feature	Characteristic	Background of the problem definition
Capacity	Every player has 9 capacity units for load processing; a capacity unit can process a load unit.	Every IT service provider of the community has its own server infrastructure dedicated for community orders.
	The capacity units will be occupied for a round when they are optimized or process a load order.	While the server infrastructure is under construction for efficiency or busy, the capacity units will not be available.
Efficiency	The player starts with an efficiency of level "A" with 100€ costs per processed load unit.	At the beginning, the costs are relatively high and aligned with the revenue possibilities and an efficiency incentive.
	The optimization of capacity units costs 20€ per unit and upgrade ("A+++" is maximum) and causes -1€ costs for the efficiency in the next rounds.	The optimization costs for the capacity units are aligned with the amortization through energy savings of the IT investment after assumed 2 accounting periods.
	The optimization of the building and air-conditioning technology with 50€ per upgrade ("A+++" is maximum) has an effect of -4€ costs for the efficiency in the next rounds.	The optimization costs for the building and air-conditioning technology is aligned with the amortization through energy savings of the investment after assumed 12.5 accounting periods.
Orders	The load volume varies between 2 and 10 load units; the frequency distribution is equal for each load.	Variety of orders, from small to not grantable with own capacity. The frequency distribution is experimented and iteratively balanced with the capacities.
	The payment varies between 90€, 100€ and 110€ per processed load unit; the frequency distribution is equal for each payment.	Without optimization, the revenue covers the average costs of a members load processing. Underpaid orders are usual, e.g. to maintain an important customer.

	If a job can not be processed completely, a penalty (10% of the order value) occurs to the player who brings in the order.	This case corresponds with a violation of service level agreements. The handling in practice inspires the height of the penalty.
Finance	Players have a seed capital of 500€.	Members have limited financial background.
	Each round has fixed costs of 20€.	The height is iteratively balanced with the average costs and the revenue.

Table 2: Static features of the game content

5.2.2 Procedural features of the game conception

Starting the game process, the first player has to be drawn by a random method. He starts the first round in the first playthrough and can continue the game process (see activity chart in Figure 2, left). Each move of the active player contains sub-processes for optimization and load processing; both are pictured in another activity diagram (see Figure 3). After each player was in active mode, a balance of accounts follows (see Figure 2, right) and the next round starts. After pre-defined number of playthroughs, the game ends and a score needs to be calculated to determine the winner.

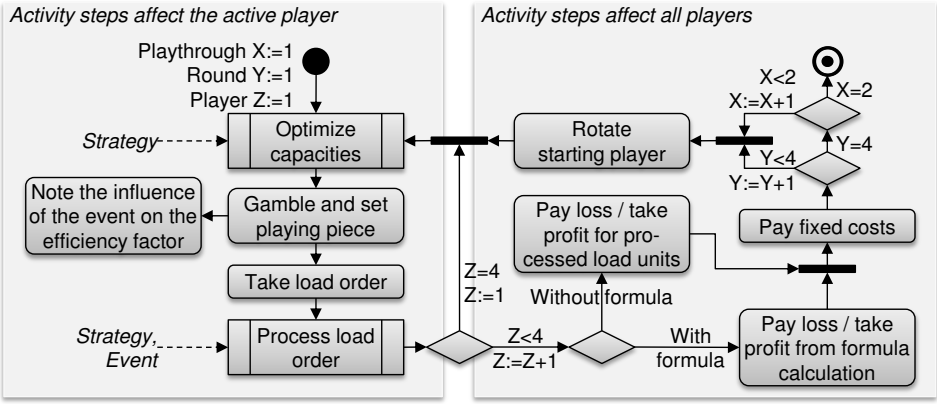


Figure 2: Activity diagram for the rule design in the game process

The first sub-process describes the optimization decision: a player can optionally optimize his resources for the next rounds but with the optimization mode, he occupies his resources for the actual round and has limited capacities to process load units and earn revenue. The more efficient a player is, the higher is his profit from the load order processing. The second sub-process describes the load processing decision: a player receives a load order and has to process the load within the limits of capacity or deliver load units to the community. By providing load units to the community, the next best efficient player has the first choice to process the load. This continues until the load order is fully processed or the delivering player needs to pay a penalty.

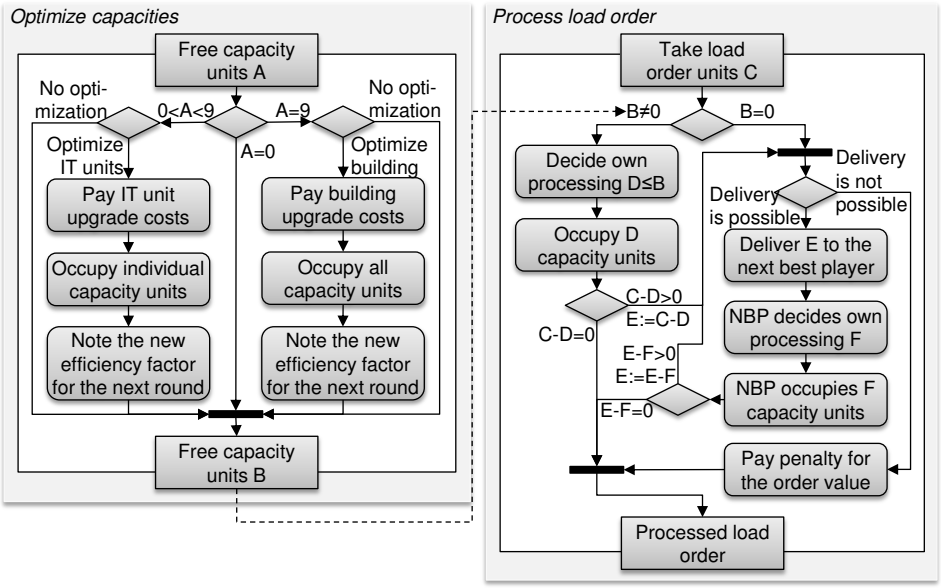


Figure 3: Activity diagram for the rule design in the sub-processes

5.3 Game Play and Game Strategy

Before playing the community cloud game, the players need to agree on a community strategy. There exist two opposite strategy positions: complete individual optimum (egoistic) and perfect community optimum (altruistic). While an egoistic strategy aims at maximizing the own infrastructure utilization and account balance, the altruistic strategy tends to optimize the overall savings. The final strategy can be positioned between both extremes and some policies for acting should be formulated, e.g. when is a player allowed to optimize his resources or when should a player deliver his load order to the community. A game manager can enforce the agreed strategy and plays an active role by reflecting the game decisions to the players.

The goal of the game is to be successful with the developed strategy. To measure this fact and to increase the players' enjoyment, we calculate a score for active and efficient behaviour. Each player wants to gain the highest score. Within an egoistic strategy, each player's goal is to optimize his account and have a higher balance than the other players do; therefore, the score includes the profit of the own load processing and the proportionally account balance. The other extreme is to renounce the freedom of choice for the load delivery to the community, so that only the efficient players will process the load orders. Following this strategy, the score includes the profit of the transferred overall profit and the proportionally efficiency factor.

To combine the advantages of both strategies, a hybrid strategy considers the resource and cost optimization of the community, but also optimizes and balances the individual

optimum. Along with the approach of persuasive systems [LOD12], [OH09], [PT12], such a strategy should be incentivized within the game, e.g. with a special revenue sharing model. This model can (a) provide a commission for delivered load orders to incentivize the transfer of load units to the community. Moreover, the formula uses a profit share to (b) facilitate a compensation of the effort of joining and participation in the community and another share (c) according the efficiency performance of the members to incentivize the active involvement. The score for the members includes the profit of the own and the delivered load processing as well as the proportional account balance and efficiency factor.

We tested the game elements several times, e.g. with different load order volumes and compensations or costs for optimizations, to finally get a playable and interesting game. To give the reader a visual impression of the game, the prototype of the designed game is shown in the following picture (see Figure 4).

Figure 4: Game prototype

5.4 Game Analysis

We validated the game-based approach with an experimental workshop to demonstrate the effectiveness and reliability of the game. Four expert IT providers played the game to give qualitative feedback and develop the scope of pursued strategies and policies. Within this workshop, we tested two scenarios and compared the results of the incentivizing revenue sharing formula with the normal revenue sharing:

1. Egoistic strategy with the goal of maximization the own account balancing
2. Cooperative strategy with the goal of optimization the overall account balancing
    a. Regular revenue sharing option
    b. Revenue sharing formula option

We played both game variants with two playthroughs. The feedback of the players was eminently positive. Initially, the game was perceived as 'complex' but with the increasing understanding, the players perceived the simulation situation as very realistic and felt the fun factor when they can process a well-paid load order. Within the egoistic game variant, the players have the major challenge to decide for optimizing the IT or keep open the possibility of load processing for profit making. Playing the cooperative variant, the players got deep into the situation and became aware of the important things to decide and the common policies to be agreed upon. Once, the players had identified the cooperative goal and their individual roles, they improved their teamwork with strong coordination skills. They developed and pursued strategies and give ideas for changing or extending the rule-set to make the game more realistic.

Comparing the results from the egoistic and the cooperative game variants, we can detect tendencies for a more profitable processing in the cooperative variant (see Figure 5). The total number and average revenue of the processed units was approximately identical in both played variants, which promotes a comparability of the results. Comprehensible, the proportion of delivered units to the community is higher in the cooperative than the egoistic strategy. The delivery causes a higher profit in the cooperative game, even though the optimization effort was slightly smaller. The overall balancing of the accounts is a little higher with the cooperative strategy. This difference should increase with the number of playthroughs.

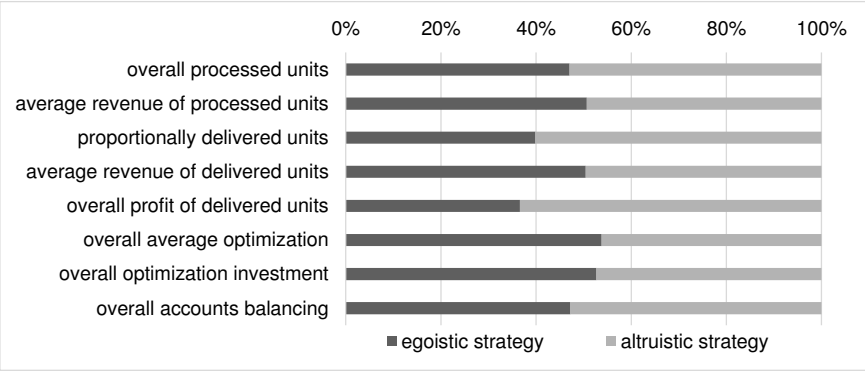


Figure 5: Comparison of the egoistic and the altruistic game strategy

The generated profit surplus with the cooperative strategy can be allocated to the players to incentivize their altruistic behaviour. This can be implemented with the proposed revenue sharing formula and aspire a more levelled balance of accounts (see Figure 6).

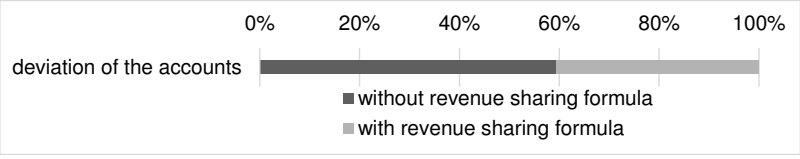


Figure 6: Deviation of the accounts within the altruistic game strategy

6 Conclusion and Outlook

With our research, we contribute with a novel evaluation method for testing realistic behaviour in a simulated situation. This situation is the participating and strategy developing in a community cloud, a relatively immature field in research and practice. Hence, our project is rather positioned as an invention, regarding the knowledge contribution framework of the design science research [GH13].

Answering the first research question, we developed a board game, closely derived from action research in a real project, and validated it with playing experiments in an expert workshop. Regarding the second research question, the workshop confirmed the expected intensification of the sensitivity for the cooperation scenario and the players appreciated the game for the simulation possibility.

Our benefits for academic research are manifold. For our research, we can determine the scope of strategies and goals for cooperating IT service providers in a community cloud. The game gives the possibility to analyse which strategies or rules are successful. Furthermore, we have a tool to simulate an implemented revenue sharing mechanism that helps to set incentives for a cooperative behaviour of the community members. Our theoretical implication is a transferable method to use a serious game for the analysis and understanding of the participants' behaviour.

In practical terms, the game can be used for workshops to inspire the participants with the possibilities of cooperative behaviour in cloud computing and help to shift traditional opinions. We have demonstrated that the concept is an enjoyable and inexpensive method to enable complex decision-making in a simulated environment.

Looking to future research options, the game should be analysed with more playthroughs to verify the results of the strategy comparison. Moreover, different experiment settings can be compared with control groups, the number of players could be raised, and a higher number of game processing gives more evaluated answers. Further, a structured survey can analyse how the game-based approach affects the players' opinion. Finally, to implement and automate the game process, an online version of the game could be valuable for more quantitative analyses.

7 References

[AD08] Ahn, L. v.; Dabbish, L.: Designing Games with a Purpose. In Communications of the ACM (CACM), 2008, 51.

[BC08] Ben-Zvi, T.; Carton, T. C.: Simulation Gaming in Technology Management. In Proceedings of the Americas Conference on Information Systems (AMCIS), 2008, 14.

[Be09] Ben-Zvi, T.: Simulating Decision Support Systems: A Laboratory Experiment. In Proceedings of the Americas Conference on Information Systems (AMCIS), 2009, 15.

[Be95] Bevan, N.: Human-Computer Interaction Standards. In Proceedings of the International Conference on Human Computer Interaction, 1995, 6.

[BGM87] Benbasat, I.; Goldstein, D. K.; Mead, M.: The Case Research Strategy in Studies of Information Systems. In Management Information Systems Quarterly (MISQ), 1987; pp. 369–386.

[BM09a] Briscoe, G.; Marinos, A.: Community Cloud Computing. In Proceedings of the IEEE International Conference on Cloud Computing (IEEE CLOUD), 2009, 1; pp. 472–484.
[BM09b] Briscoe, G.; Marinos, A.: Digital Ecosystems in the Clouds: Towards Community Cloud Computing. In Proceedings of the IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST), 2009, 3; pp. 103–105.
[Bu12] Butler, B.: Are community cloud services the next hot thing? In Computerworld, 2012, 2012-05-01.
[CP09] Chatterjee, S.; Price, A.: Healthy Living with Persuasive Technologies: Framework, Issues, and Challenges. In Journal of the American Medical Informatics Association, 2009, 16; pp. 171–178.
[De11] Deterding, S. et al.: Gamification: Using Game Design Elements in Non-Gaming Contexts. In Proceedings of the Conference on Human Factors in Computing Systems (HCI), 2011.
[DG12] Dunn, R. A.; Guadagno, R. E.: My avatar and me - Gender and personality predictors of avatar-self discrepancy. In Computers in Human Behavior, 2012, 28; pp. 97–106.
[Fa09] Faria, A. J. et al.: Developments in Business Gaming. A Review of the Past 40 Years. In Simulation & Gaming, 2009, 40; pp. 464–487.
[FH11] Foster, S.; Hopkins, J.: ERP Simulation Game: Establishng Engagement, Collaboration And Learning. In Proceedings of the Pacific Asia Conference on Information Systems (PACIS), 2011,15.
[Fo03] Fogg, B. J.: Persuasive Technology. Using Computers to Change What We Think and Do. In Morgan Kaufmann Series in Interactive Technologies, 2003.
[Ga13] Gartner: Hype Cycle for Cloud Computing. In Gartner, 2013.
[GBN13] Greco, M.; Baldissin, N.; Nonino, F.: An Exploratory Taxonomy of Business Games. In Simulation & Gaming, 2013, 44; pp. 645–682.
[GH13] Gregor, S.; Hevner, A. R.: Positioning and Presenting Design Science Research for Maximum Impact. In Management Information Systems Quarterly (MISQ), 2013, 37; pp. 337–355.
[Ha10] Haff, G.: The rise of the community cloud. In CNET, 2010, 2010-09-16.
[He04] Hevner, A. R. et al.: Design Science in Information Systems Research. In Management Information Systems Quarterly (MISQ), 2004, 28; pp. 75–105.
[HSA12] Herzig, P.; Strahringer, S.; Ameling, M.: Gamification of ERP Systems - Exploring Gamification Effects on User Acceptance Constructs. In Proceedings of the Multikonferenz Wirtschaftsinformatik (MKWI), 2012.
[HV09] Hasan, H.; Verenikina, I.: Serious Games: The importance of play in network centric organizations. In Proceedings of the Americas Conference on Information Systems (AMCIS), 2009, 15.
[Jo03] Jokela, T.: Assessments of Usability Engineering Processes: Experiences from Experiments. In Proceedings of the Hawaii International Conference on System Sciences (HICSS), 2003, 36.
[Ki05] Kiili, K.: Digital game-based learning: Towards an experiential gaming model. In Internet and Higher Education, 2005, 8; pp. 13–24.
[Li12] Liang, J.: Government Cloud: Enhancing Efficiency of E-Government and Providing Better Public Services. In Proceedings of the International Joint Conference on Service Sciences (IJCSS), 2012.
[LLS13] Liu, D.; Li, X.; Santhanam, R.: Digital Games and Beyond: What happes when Players compete? In Management Information Systems Quarterly (MISQ), 2013, 37; pp. 111–124.
[LOD12] Lehto, T.; Oinas-Kukkonen, H.; Drozd, F.: Factors affecting perceived persuasiveness of a behavior change support system. In Proceedings of the International Conference on Information Systems (ICIS), 2012, 33.
[LPN09] Lang, F.; Pueschel, T.; Neumann, D.: Serious Gaming for the Evaluation of Market Mechanisms. In Proceedings of the International Conference on Information Systems (ICIS), 2009, 30.
[MG09] Mell, P.; Grance, T.: A NIST Notional Definition of Cloud Computing. In National Institute of Standards and Technology (NIST), 2009.
[OH09] Oinas-Kukkonen, H.; Harjumaa, M.: Persuasive Systems Design: Key Issues, Process Model, and System Features. In Communications of the AIS (CAIS), 2009, 24.
[PT12] Pisan, Y.; Tan, C. T.: Argumentation Models in Serious Games. In Proceedings of the International Conference on Computer Science & Education (ICCSE), 2012, 7; pp. 1319–1322.
[Re11] Repschläger, J. et al.: Konzeption einer Community Cloud für eine ressourceneffiziente IT-Leistungserstellung. In Proceedings of the Informatik Konferenz, 2011.
[Sm10] Smets, N. J. J. M. et al.: Game-based evaluation of personalized support for astronauts in long duration missions. In Proceedings of the International Astronautical Congress (IAC), 2010, 59.
[St08] Streefkerk, J. W. et al.: Evaluating Context-Aware Mobile Interfaces for Professionals. Handbook of Research on User Interface Design and Evaluation for Mobile Technology, 2008.
[SZ04] Salen, K.; Zimmermann, E.: Rules of Play. The MIT Press, New York, 2004.
[We09] Weinhardt, C. et al.: Cloud-Computing - Eine Abgrenzung, Geschäftsmodelle und Forschungsgebiete. In Wirtschaftsinformatik (WI), 2009.

Ein Serious Game als neue qualitative Erhebungsmethode für die Customer Journey beim Automobilkauf

Silke Plennert, Susanne Robra-Bissantz

Lehrstuhl Informationsmanagement
Institut für Wirtschaftsinformatik
Technische Universität Braunschweig
Mühlenpfordtstr. 23
38106 Braunschweig
s.plennert@tu-bs.de
s.robra-bissantz@tu-bs.de

Abstract: Kaufentscheidungen verlagern sich immer mehr in die digitale Welt, der Einfluss digitaler Kontaktpunkte auf die Beziehung zwischen Unternehmen und Kunde wächst. Dadurch werden Kaufentscheidungsprozesse auch komplexer und schwerer nachvollziehbar. Um Erkenntnisse über die digitaler und komplexer werdende Customer Journey zu generieren, müssen sich auch die Forschungsmethoden verändern. Eine Lösungsmöglichkeit bietet der Einsatz von Spielen. Die vorliegende Arbeit beschreibt ein Forschungsdesign, anhand dessen am Beispiel des Automobilkaufs ein teilweise elektronisches Serious Game zu diesem Zweck entwickelt, verbessert und evaluiert wird. Mit dem Spiel als Erhebungsmethode sollen die Schwächen klassischer sowie prozessbegleitender und rückblickender Methoden eliminiert werden, indem Spielelemente gezielt eingesetzt werden, um besser an Insights zu gelangen – beispielsweise indem sich die Probanden besser in die Entscheidungssituation hineinversetzen können.

1 Einleitung

1.1 Motivation

Der Einfluss digitaler Kontaktpunkte auf die Beziehung zwischen Unternehmen und Kunde wächst [MW12, S. 155; Tn10]. Nicht nur, dass ein immer größer werdender Teil der Produkte online gekauft wird [He12, S. 1-7], auch Kaufentscheidungen verlagern sich zunehmend in die digitale Welt [He13a, S. 15]. Dadurch werden Kaufprozesse auch komplexer und schwerer nachvollziehbar.

Um sie besser verstehen zu können, ist es notwendig, neue Methoden zu entwickeln – hier setzt die vorliegende Arbeit an. Ziel des Forschungsvorhabens ist die Entwicklung einer neuen Marktforschungsmethode in Form eines teilweise elektronischen Serious Games zur Erfassung von Informationen über den Kaufentscheidungsprozess von End-

kunden. Durch den Einsatz des eigens entwickelten Spiels sollen tiefgehende Bedürfnisinformationen – sogenannte Insights – besser aufgedeckt werden als mit bisher angewandten Methoden. Die so gewonnenen Ergebnisse können beispielsweise bei Unternehmen in die Entwicklung von elektronischen Diensten einfließen, welche den Kaufprozess sinnvoll unterstützen können.

1.2 Wissenschaftliches Vorgehen

Die wissenschaftliche Herangehensweise orientiert sich an „Design Science Research". Hierbei handelt es sich um ein Forschungsparadigma, welches das Ziel unterstützt, neue Artefakte zu entwickeln, die bestehende Probleme auf innovative Weise lösen [He04, S. 77]. Ein Artefakt kann z.B. ein Konstrukt, ein Modell oder auch eine Methode sein [MS95]. Ein Forschungsbeitrag im Sinne des Design Science erfordert dabei nach der Identifikation und Beschreibung eines relevanten organisationalen Problems den Nachweis, dass hierfür nicht bereits eine adäquate Lösung existiert. Von Bedeutung ist zudem, dass nach der Entwicklung eines völlig neuen Artefakts, welches das identifizierte Problem adressiert, dieses evaluiert werden muss. [Ka09; MS08, S. 726; He07, S. 91]

Analog zu dem Design-Science-Vorgehen wird zunächst die Customer Journey und deren Relevanz vorgestellt, um zu zeigen, dass Unternehmen großen Bedarf danach haben, den Kaufentscheidungsprozess ihrer Kunden nachzuvollziehen (Kapitel 2). Danach wird aufgezeigt, dass die aktuell eingesetzten Methoden aufgrund diverser Schwächen nicht ausreichend sind, um zufriedenstellende Insights zu generieren (Kapitel 3). Deshalb wird dann beleuchtet, welche Potenziale Spiele bieten, um bessere Erkenntnisse zu liefern (Kapitel 4), bevor die Entwicklung sowie der aktuelle Stand der neuen Methode vorgestellt und ein Ausblick auf die Evaluation gegeben wird (Kapitel 5).

2 Relevanz der Customer Journey

Der klassische Kaufentscheidungsprozess sieht wie folgt aus: Ein Kunde hat das Bedürfnis des Erwerbs einer Dienstleistung oder eines Produkts, in seinem Bewusstsein befindet sich hierfür eine Reihe von Marken. Diese werden durch Marketingmaßnahmen der verschiedenen Anbieter reduziert, wodurch sich die Anzahl der vertrauten bzw. in Betracht kommenden Marken verringert, bevor schließlich die Entscheidung und der Kauf erfolgen. [Co09] Dieser sogenannte Trichterprozess ist in Abbildung 1 dargestellt.

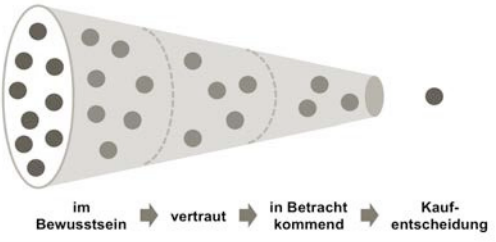


Abbildung 1: Der traditionelle Trichterprozess (in Anlehnung an [Co09])

Der traditionelle Trichter versagt zunehmend darin, die Vielzahl der Kontaktpunkte (engl. Touchpoints) und Kauffaktoren zu erfassen. Die riesige Auswahl an Produkten und digitalen Interaktionskanälen in Verbindung mit anspruchsvollen, gut informierten Verbrauchern erfordert neue Ansätze im Marketing. [MM11, S. 233]

Ein Ansatz ist die Customer Journey – bisweilen auch Consumer Journey genannt –, deren Ziel es ist, alle Situationen zu ermitteln, in denen die Verbraucher beeinflusst werden können [Co09, S. 1f]. Die Customer Journey schildert die „Reise" des Kunden vom ersten Kaufgedanken bis hin zum tatsächlichen Kauf und bezieht dabei alle Kontaktpunkte ein [He13b; Ne08].

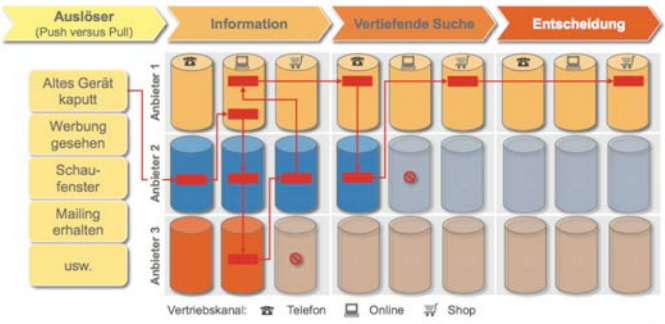


Abbildung 2: Beispielhafte Customer Journey [Do12, S. 53]

Abbildung 2 zeigt beispielhaft die Customer Journey eines Mobilfunkkunden [Do12, S. 53]. Hier ist stark vereinfacht dargestellt, dass es zwischen Kunden und Anbietern online und offline diverse Kontaktpunkte in den verschiedenen Phasen des Kaufprozesses gibt. Zusätzlich existieren jedoch noch viele weitere Möglichkeiten, wie der Kunde mit unterschiedlichen Marken oder Produkten in Verbindung kommen kann, ohne dass das Unternehmen direkt beteiligt ist. Bei solchen Touchpoints kann es sich beispielsweise um Gespräche mit Freunden, Preisvergleichswebseiten oder Empfehlungen über Soziale Netzwerke handeln.

Da die Customer Journey immer komplexer und nicht mehr steuerbar wird, sowie mehr und mehr online erfolgt, müssen auch die von Unternehmen durchgeführten Maßnahmen darauf ausgerichtet sein. Dies geschieht, indem sie elektronische Dienste entwickeln, welche an den Touchpoints ansetzen. Um diese Kontaktpunkte sowie die jeweiligen Bedürfnisse der Kunden zu identifizieren, müssen jedoch zuerst die relevanten Customer Insights erhoben werden.

Der Begriff Customer Insight (auch Consumer Insight) ist ein gerade in den letzten Jahren immer häufiger sowohl in wissenschaftlichen Arbeiten als auch in den Medien auftauchendes Schlagwort, dem es noch an einer einheitlichen Definition mangelt. Der englische Begriff Insight kann wörtlich übersetzt werden mit Einsicht, Erkenntnis oder Einblick[1]. Föll definiert Customer Insights in ihrer Forschungsarbeit, in welcher sie die bis dahin veröffentlichte Literatur zu diesem Thema systematisch aufarbeitet, als „spezi-

[1] Übersetzung laut http://dict.leo.org.

fische, erleuchtende, neuartige Kombination von Erkenntnissen über den Konsumenten“ [Fö07, S. 38]. Nach Riekhof handelt es sich dabei allgemeiner um „Erkenntnisse über Kundenverhalten, Kundenerwartungen und Kundenbedürfnisse" [Ri10, S. 9]. Unger et al. hingegen verwenden den Begriff enger für "die Gesamtheit der inneren Beweggründe, subjektiven Wahrnehmungen, Bedürfnisse, Präferenzen in einem spezifischen Produktfeld" [UFM13, S. 48].

Es handelt sich dabei somit um einen Einblick in die Konsumenten. Man will z.B. nicht nur wissen, was und wo die Kunden kaufen bzw. nicht kaufen, sondern auch die Motive und Bedürfnisse, die dahinter stecken. Bei Insights zur Customer Journey will man z.B. erfahren, welche Touchpoints den Entscheidungsprozess eines Kunden beeinflussen.

3 Schwächen aktueller Erhebungsmethoden

Für die Erhebung von Customer Insights kommen oft klassische qualitative Marktforschungsmethoden, wie z.B. Leitfadeninterviews oder Fokusgruppen, zum Einsatz. Daneben werden aber auch immer mehr Methoden speziell an die Customer Journey angepasst oder komplett neu dafür entwickelt. Es lassen sich hier prozessbegleitende und rückblickende Verfahren unterscheiden. Zur ersten Kategorie gehören z.B. Tagebuchstudien, bei denen die Probanden im Laufe ihres Entscheidungsprozesses immer wieder relevante Ereignisse aufschreiben sollen. Ein Beispiel für die zweite Kategorie sind Interviews, in denen die Befragten retrospektiv Auskunft über die erfolgten Ereignisse und deren Wirkung auf ihre Kaufentscheidung geben sollen. [Do11]

Ein Nachteil, der insbesondere auf Fokusgruppen und Interviews zutrifft, ist die große Bedeutung der Befragungssituation. Oft erhält man unwahre (z.B. sozial erwünschte) Antworten, weil sich die Probanden der Befragungssituation bewusst sind und durch den Moderator oder Interviewer sowie durch die anderen Testpersonen beeinflusst werden [Bo12, S. 55f]. Ferner ist der Entscheidungsprozess größtenteils nicht rational, sondern emotional. Kunden sind deshalb z.B. in Fokusgruppen selten in der Lage, rückblickend den wahren Prozess zu beschreiben. [We03, S. 33] Auch Belz et al. kommen zu dem Schluss, dass „Ergebnisse aus Befragungen von Kunden ... oft wirklichkeitsfremd [sind]“ [Be11, S. 37].

Prozessbegleitende Methoden haben dagegen das Problem, dass vor der Erhebung genau solche Probanden identifiziert und akquiriert werden müssen, die sich aktuell kurz vor Beginn eines Kaufentscheidungsprozesses befinden. Auch besteht die Gefahr, dass durch die Untersuchung selbst der Prozess beeinflusst wird. Zudem finden üblicherweise trotzdem zusätzlich am Ende des Prozesses sogenannte Follow-Up-Interviews statt [Do11, S. 31], was wiederum den Aufwand erhöht.

Weiterhin existieren diverse Möglichkeiten des Webtrackings, um die Customer Journey online zu verfolgen. Nicht unberücksichtigt darf bei dieser Art der Analyse der Wechsel der Medien bleiben. So können Online-Programme, die die Cookies der Verbraucher im Internetbrowser sammeln und auswerten, keine Daten aus anderen Medien und Informationskanälen denselben Kunden zuordnen. [Gü12] Zudem können die Trackingmethoden

nur die tatsächlichen Handlungen – also beispielsweise Klicks – der (potenziellen) Kunden erfassen, jedoch nicht die dahinterstehenden Gründe. Die Folge ist, dass wie bei den prozessbegleitenden Methoden zusätzlich Befragungen durchgeführt werden müssen [Do12, S. 54].

Aus den hier vorgestellten und weiteren Schwächen bestehender Methoden resultiert das Problem, dass Insights im Allgemeinen und dadurch auch in Bezug auf die Customer Journey noch nicht ausreichend erfasst werden können. Auch viele Forscher haben das Problem erkannt, z.B. stellen Belz et al. fest, dass die verwendeten Methoden in der Insight-Generierung überholt sind und größtenteils nur hypothetische Ergebnisse produzieren [Be11, S. 37]. Deshalb wird im folgenden Abschnitt kurz vorgestellt, welches – in diesem Zusammenhang noch ungenutzte – Potenzial Spiele und Gamification bieten können. Dies wird anschließend bei der Entwicklung einer neuen Erhebungsmethode herangezogen, welche versucht, dadurch die Schwächen bestehender Methoden zu eliminieren.

4 Serious Games und Gamification

Der in den letzten Jahren immer größer gewordene Trend des Einsatzes von Serious Games und Gamification ist in den unterschiedlichsten Bereichen zu beobachten. Beispielsweise existieren Umsetzungen mit wissenschaftlichen Untersuchungen in der Medizin [Ha10], im Bereich des E-Learning [CCF13] oder im Innnovationsmanagement [WSR11]. Die positiven Auswirkungen der Anwendung von Spielen und Spielmechaniken wurden schon hinreichend wissenschaftlich nachgewiesen: Hierbei handelt es sich beispielsweise um erhöhte Kreativität und Motivation des Spielers, oder auch um die Erreichung eines sogenannten Flow-Zustands (z.B. [WR12; Ch07; Cs90, S. 30f; Be69]).

Auch in der Marktforschung bzw. Insight-Generierung wird aktuell immer mehr über Gamification nachgedacht, und es existieren bereits einige Umsetzungen (z.B. [FH12; To12]). Vereinzelt werden aber auch schon komplette Spiele als Erhebungsmethode eingesetzt [Ja12]. Allerdings wurde deren Erfolg bisher kaum wissenschaftlich untersucht; zudem existieren noch keine Spiele zur Untersuchung von Kaufentscheidungsprozessen bzw. der Customer Journey. Ein solches Spiel wird in dieser Arbeit vorgestellt.

An dieser Stelle soll kurz erläutert werden, wie sich das entwickelte Spiel in den Gesamtkontext aus Spielen und Gamification einordnen lässt. Nach der allgemein anerkannten Definition von Deterding et al. für Gamification handelt es sich dabei um die Anwendung von Spielelementen in einem spielfremden Kontext [De11, S. 2]. Deterding et al. entwickeln auch eine grafische Abgrenzung aller spielbezogenen Begriffe (siehe Abbildung 3). Neben Gamification sind dort auch Game (Spiel), Toy (Spielzeug), Playful Interaction (spielerische Interaktion) und Serious Games eingeordnet. Letztere sind in der Grafik unter „komplette Spiele“ einsortiert – im Gegensatz zu bloßen Spielelementen oder dem reinen Spielen, ohne dass ein Spieldesign dahintersteckt.

Der Unterschied zwischen einem Serious Game und einem Spiel (Game) wird in dieser Grafik jedoch nicht deutlich. Er besteht nach Abt darin, dass bei Serious Games nicht der

Unterhaltungswert im Vordergrund steht, sondern ein anderes Ziel verfolgt wird, wie z.B. Problemlösungen zu erarbeiten oder etwas zu lernen [Ab87]. Sawyer und Smith stellen über 20 Jahre später fest, dass Serious Games mittlerweile eine viel größere Bandbreite an möglichen Zielen umfassen und teilen sie deshalb in sieben Kategorien ein [SS08].

Abbildung 3: Abgrenzung der spielbezogenen Begriffe [De11, S. 2]

Das Spiel der vorliegenden Arbeit ist ein Serious Game, und zwar aus der Kategorie der „Games for Science and Research" nach Sawyer und Smith. Es soll die Probanden zwar auch unterhalten, der Hauptzweck liegt jedoch eindeutig in der Insight-Generierung. Im Folgenden wird dennoch auch weiter der Begriff „Spiel" verwendet.

5 Ein Serious Game als neue qualitative Erhebungsmethode

Das hier vorgestellte Forschungsvorhaben hat als Ziel, ein Serious Game zu entwickeln und zu evaluieren, das den Kaufentscheidungsprozess abbildet und dabei Informationen über die Kundenbedürfnisse erhebt. Als Anwendungsbeispiel wird der Automobilkauf gewählt, da es sich hierbei um eine extensive Kaufentscheidung handelt. Das bedeutet, dass hier im Unterschied zu vielen anderen Produkten und Dienstleistungen der Kaufentscheidungsprozess typischerweise umfangreicher ist, keine kognitive Programmierung beim Kunden vorliegt, während ein sehr hohes Involvement sowie eine hohe Kaufmotivation verbunden mit einem hohen Kaufrisiko vorhanden sind [Di02, S. 366]. Somit ist hier auch der Bedarf an Customer Insights besonders groß.

Im Spiel sollen die Probanden preisgeben, welche Kontaktpunkte sie beeinflussen und welche konkreten Informationen sie als wichtig empfinden. Verschiedene gezielt eingesetzte Spielelemente fördern dabei einerseits die Erkenntnisgewinnung direkt. Andererseits unterstützen einige Elemente dies auch indirekt, indem sie z.B. dafür sorgen, dass die Probanden ein positives Spielerlebnis haben. Dies sorgt beispielsweise dafür, dass für die Probanden die Befragungssituation nicht mehr im Vordergrund steht, sondern die Spielerfahrung. Und auch dafür, dass sie wertvollere Insights preisgeben, da sie sich besser in die Entscheidungssituation hineinversetzen können und somit den wahren Prozess nachspielen und nicht nur davon erzählen.

5.1 Entwicklungs- und Evaluationsprozess des Spiels

Die Entwicklung und Evaluation der neuen Erhebungsmethode erfolgt in drei Phasen, welche in Abbildung 4 dargestellt sind.

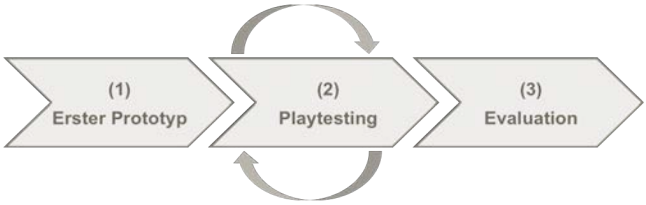


Abbildung 4: Vorgehen bei der Spielentwicklung und -evaluation

In Phase 1 wird die erste Version des Spiels entwickelt. Aus der Idee entsteht hier zunächst eine „Geschichte“ des Spiels, die dann mithilfe von Spielelementen umgesetzt wird. Der erste Prototyp wird relativ simpel mit Stift und Papier realisiert. Hier und in der folgenden Phase kommen verschiedene Ansätze des Game Design zum Einsatz (siehe hierzu bspw. [Sc08; Fu08; Ca06; SZ03]).

In Phase 2 wird ein kontinuierlicher Verbesserungsprozess durchlaufen, welcher aus dem wiederholten Test – d.h. Spielen – und daraus entstehenden Weiterentwicklungen des Spiels besteht. Diese Vorgehensweise wird auch als Playtesting bezeichnet [Sc08]. Die Verbesserungen betreffen sowohl die Insight-Generierung als auch das Spielerlebnis. Ob das Spiel bzgl. Letzterem ausgereift ist, wird immer wieder mit den von Schell vorgeschlagenen acht Filtern überprüft, welche unter anderem die künstlerische Sicht des Designers, die Engineering-Perspektive und die „Business-and-Marketing“-Sicht umfassen [Sc08, S. 76-78].

Sobald das Spiel dieser Überprüfung standhält, erfolgt der Vergleich mit anderen Methoden (Phase 3). Dies geschieht einerseits über eine Literaturrecherche. Andererseits durch ein selbst durchgeführtes Experiment, indem eine etablierte Methode sowie die neue Methode durchgeführt und die Ergebnisse einander gegenübergestellt werden.

5.2 Anwendung des Spiels

Das Spiel wird im Labor eingesetzt, d.h. bis zu vier Probanden werden zur gemeinsamen Teilnahme als Spieler eingeladen. Ein Moderator erklärt die Spielregeln und ist auch während des Spiels in der Nähe, um ggfs. mit Anmerkungen oder Fragen eingreifen zu können. Zudem werden die Probanden von ihm dazu aufgefordert, möglichst alle ihre Handlungen zu kommentieren. Dieses Vorgehen bezeichnet man als „Think aloud“, eine Methode, die eingesetzt wird „to gain insight in the knowledge and methods of human problem-solving“ [SBS94, S. 1f]. Zur Datensicherung des Spielablaufs sowie der Gespräche wird eine Videoaufzeichnung durchgeführt.

Das Spiel selbst besteht aus der Kombination eines Brettspiels und einer Android-basierten Smartphone-Applikation. Das Spielziel aus Spielersicht ist es, als Erster die vorgegebene Zahl an Informationsfeldern zu erreichen und anschließend auf dem Ziel-

feld anzukommen. Aus Anbietersicht ist das Ziel, dass die Probanden möglichst viele Informationsfelder erreichen, um dann wertvolle Insights einerseits aufzuschreiben und andererseits darüber zu sprechen.

5.3 Spielelemente und Spielerlebnis

Die Entwicklung und Verbesserung des Spiels erfolgt insbesondere nach den Richtlinien des Game Design von Schell [Sc08] sowie Salen und Zimmerman [SZ03]. Ein Spiel besteht aus unterschiedlichen Spielelementen, welche sich nach Schell in vier Basiskategorien einordnen lassen: Story (Geschichte), Mechanics (Spielmechaniken), Aesthetics (Ästhetik) und Technology (Technologie). Sie sind alle gleichermaßen wichtig und beeinflussen sich in vielen Aspekten gegenseitig. Das Zusammenwirken aller Spielelemente bestimmt das empfundene Spielerlebnis. [Sc08, S. 41-45]

Im Folgenden wird nun aufgezeigt, welche Spielelemente in der aktuellen Version des Spiels die Erhebung der Customer Insights einerseits direkt unterstützen und andererseits indirekt, indem ein positives Spielerlebnis herbeigeführt wird. Dazu werden zuerst die wichtigsten Elemente erläutert, bevor ein Überblick ihres Zusammenwirkens erfolgt. Es wird versucht, die Spielelemente nach den Kategorien von Schell aufzuteilen, auch wenn dies nicht immer trennscharf möglich ist.

Geschichte

Die Geschichte besteht aus der Abfolge aller Handlungen im Laufe des Spiels, welche linear oder verzweigt sein kann [Sc08, S. 41]. Eine gute Geschichte beinhaltet dabei mindestens zwei Aspekte: Ein Ziel für den Spieler und Hindernisse auf dem Weg dorthin. Durch das Überwinden Letzterer ergeben sich oft interessante Konflikte zwischen den Spielern. [Sc08, S. 270f]

Vor Spielbeginn wählt zunächst jeder Spieler ein Auto, mit dem er sich im Spiel fortbewegen möchte, aus einer vorgegebenen Liste. Diese Wahl wird mit den weiteren Informationen gespeichert und hat auch Auswirkungen auf den Spielverlauf. Denn je nach Fahrzeugtyp können die Spieler sich nun entweder auf dem Dorf, auf der Autobahn oder in der Stadt besser fortbewegen. Außerdem werden den Probanden vier Kategorien für Informationsquellen gezeigt (Internet, Werbung, Gespräche und persönliche Erfahrung), aus denen sie zwei bis vier wählen müssen, die für sie persönlich bei der Entscheidung für ein Auto relevant sind. Nachdem jeder Spieler ein Startfeld gewählt hat, wird reihum gewürfelt und gezogen. Die Autos können nur auf den Straßen bewegt werden, welche in einzelne Felder aufgeteilt sind, die jeweils einer Augenzahl auf dem Würfel entsprechen.

Es geht nun darum, schneller als die Mitspieler Felder mit Informationsquellen anzufahren, welche die möglichen Kontaktpunkte darstellen. Immer wenn so ein Feld erreicht wird, soll der Proband einige wertvolle Informationen preisgeben, die er laut aussprechen und in der App eintragen soll. Diese Informationen beinhalten die gewählte Kategorie, das Medium und auch die konkreten Informationen, die er bezüglich seiner Kaufentscheidung dort einholt (z.B. in der Kategorie Internet könnte ein Proband angeben,

dass er auf facebook seine Freunde fragt, mit welchen Fahrzeugtypen sie gute Erfahrung bzgl. des Fahrverhaltens bei hoher Geschwindigkeit gemacht haben). Als Feedback für das erfolgreiche Sammeln einer Information erhält der Spieler einen Marker, welcher anzeigt, dass er nun einen Schritt näher am Sieg ist. Wer sechs Marker erhalten hat, kann sich auf den Weg zum Ziel (Parkplatz) in der Mitte des Spielfelds machen. Sobald er es erreicht hat, hat er das Spiel gewonnen und damit auch beendet.

Zusätzlich zu dieser Hauptgeschichte gibt es auch einige Nebengeschichten. Beispielsweise müssen die Spieler regelmäßig tanken fahren. Und sie können durch den Einsatz eines „Blitzerwürfels" versuchen, schneller in einem Spielzug voranzukommen, während sie dabei jedoch das Risiko eingehen, geblitzt zu werden.

Spielmechaniken

Unter Spielmechaniken fasst Schell alle Prozeduren und Interaktionen des Spiels zusammen. Darunter fallen u.a. alle Objekte und Regeln sowie der Aspekt des Glücks. [Sc08, S. 41]

Zu den Objekten zählen hier alle Spielfiguren, Würfel, Spielmarken, Karten etc. – unabhängig davon, ob sie als reale Objektive existieren oder elektronisch umgesetzt sind. Die Probanden haben Spielfiguren (Autos), mit denen sie sich auf dem gemeinsamen Spielbrett fortbewegen können. Jeder erhält zudem ein Smartphone, auf dem die installierte Applikation einerseits Spielfunktionen erfüllt wie z.B. Ereigniskarten ziehen und speichern. Andererseits tragen die Spieler dort die vom Forscher gewünschten Informationen bezüglich ihrer Kaufentscheidung ein, welche daraufhin an eine zuvor eingestellte E-Mail-Adresse versendet werden.

Weitere wichtige Objekte sind die bereits erwähnten Ereigniskarten, von denen zwei Kategorien existieren. Ein Spieler zieht immer dann eine Karte, wenn er mit seiner Figur auf einem Ereignisfeld (orangefarbene Markierungen in Abbildung 5) landet oder eines überschreitet. Die erst Variante sind Handkarten: Sie können gegen Mitspieler eingesetzt (z.B. um einen Parkplatzzugang zu blockieren) und bis dahin „auf der Hand" behalten werden. Sie haben insbesondere das Ziel, das Spiel interaktiver zu gestalten und den Spielspaß zu erhöhen. Durch sie werden sogenannte private Informationen geschaffen, die – im Gegensatz zu den öffentlichen Informationen wie z.B. der Position der Spielfiguren auf dem Brett – immer nur einem Spieler bekannt sind. Die zweite Variante sind Plus-Minus-Karten, die eine direkte Auswirkung auf den aktuellen Spielzug haben. Hier entscheidet immer der Würfel, ob es ein positiver oder negativer Effekt ist. Diese Karten sollen die Probanden zum weiteren Nachdenken über ihre Bedürfnisse beim Autokauf anregen, indem sie verschiedene Situationen und Fahrzeugeigenschaften thematisieren. Beispielsweise kann ein negativer Effekt sein, dass der Spieler eine Runde aussetzt, da er ohne Freisprecheinrichtung zum Telefonieren anhalten muss.

Regeln sind laut Schell der wichtigste Bestandteil, da sie alle weiteren Mechaniken erst ermöglichen. Er unterscheidet insgesamt acht verschiedene Regelarten, von denen folgende drei hier relevant sind: Operationale, geschriebene und Verhaltensregeln. [Sc08, S. 144-146] Operationale Regeln beschreiben, was die Spielern tun müssen, um zu spielen. Beispielsweise dass sie erst würfeln müssen, um sich fortbewegen zu können. Mit

den geschriebenen Regeln ist das Regelwerk gemeint, dass einem Spiel beigelegt wird. Die Spieler sollten es lesen, um die operationalen Regeln zu verstehen. Wichtig ist hier, dass es nicht zu komplex ist. Mit den Verhaltensregeln sind implizite Regeln gemeint, die jedem Spieler bekannt sind und spielunabhängig gelten. Hiermit ist insbesondere gemeint, dass fair gespielt wird, z.B. indem kein Spieler in einer Runde übersprungen wird.

Glück ist ein essentieller Bestandteil eines spaßbringenden Spiels, da dadurch Unsicherheit und Zufall entstehen und somit Überraschungen für die Spieler [Sc08, S. 153]. Der Glücksfaktor ist hier beispielsweise durch den Würfel und das Ziehen einer zufälligen Ereigniskarte realisiert.

Ästhetik und Technologie

Die Ästhetik eines Spiels äußert sich insbesondere in seinem Aussehen und seiner Haptik, kann aber auch den Gehör-, Geruchs- oder Geschmackssinn ansprechen. Mit ihr kann das Spiel nicht nur ansprechend angefertigt werden, sondern auch die Spielwelt realer gestaltet, dem Spieler Mechaniken verdeutlicht, seine Vorstellungskraft beflügelt werden und einiges mehr. Die ästhetischen Aspekte haben einen großen Einfluss auf die anderen Spielelemente, indem sie z.B. die Geschichte zum Leben erwecken, sind aber insbesondere mit der Technologie eng verbunden. Die Technologie ist das Medium, mit dem die Ästhetik dargestellt wird. Damit sind alle Materialien und Interaktionen gemeint, welche das Spiel ermöglichen, von Papier und Bleistift bis zu einem Computerprogramm. [Sc08, S. 42-43]

Abbildung 5: Das aktuelle Spielfeld

Die ästhetischen und technologischen Elemente äußern sich hier insbesondere durch das Spielfeld und die Android-Applikation. Das Spielfeld wird mit Adobe Photoshop kreiert, ebenso wie die Symbole und Hintergründe für programmierte App.

Abbildung 5 zeigt die aktuelle Version des Spielfelds. In den Ecken befinden sich Dörfer mit den Startpunkten für bis zu vier Spieler. Dazwischen ist die Stadt zu sehen, die von der Autobahn eingefasst wird und das Zielfeld (den Parkplatz) umschließt. Die runden gelben Felder sind die bereits angesprochenen Informationsquellen. Das Spielfeld enthält noch weitere Felder, beispielsweise markieren die orangefarbenen Straßenfelder das Ziehen von Ereigniskarten.

Drei beispielhafte Screenshots aus der Android-Applikation sind in Abbildung 6 zu sehen. Der linke Screenshot zeigt den Startbildschirm, auf welchem die Auswahl der persönlich relevanten Quellenkategorien und des Autos erfolgt. Der Mittlere zeigt einen Bildschirm während des Spiels: Links oben können Ereigniskarten von einem virtuellen Stapel gezogen werden; rechts oben befindet sich eine Tankanzeige; in der Mitte wird gewürfelt; unten sind die sechs Informationsfelder zu sehen, die in diesem Fall bis auf eins alle leer sind, d.h. der Spieler hat erst ein solches Feld erreicht und dort Informationen zu einer Internetquelle angegeben. Rechts ist die Ansicht, die der Spieler sieht, wenn er ein leeres Informationsfeld anklickt. Das kann er immer dann tun, wenn er auf dem Spielfeld ein entsprechendes Feld anfährt.

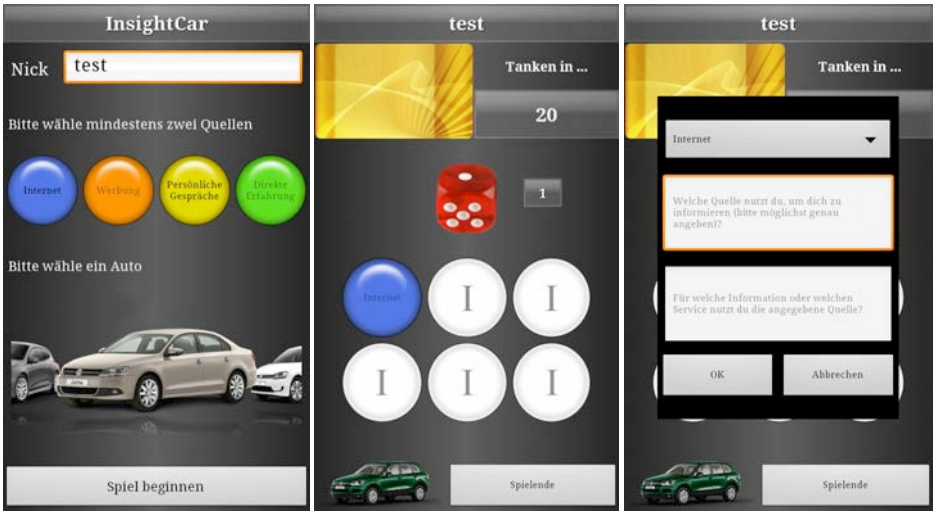


Abbildung 6: Drei beispielhafte Screenshots aus der App

Es ist theoretisch auch möglich, das Spiel komplett ohne App-Unterstützung zu spielen, die Verwendung bietet jedoch einige Vorteile. Beispielsweise können die erhobenen Informationen so direkt elektronisch erfasst und ausgewertet werden. Zudem können so sehr einfach die Ereigniskarten bearbeitet sowie weitere Anpassungen vorgenommen werden.

5.4 Generierung von Customer Insights

Das Zusammenwirken von Geschichte, Mechaniken, Ästhetik und Technologie erzeugt bei den Spielern ein bestimmtes Spielerlebnis. Je besser die Elemente nach den Richtlinien des Game Design gestaltet sind, desto positiver fällt das Erlebnis aus und desto mehr Spaß haben die Spieler. Hierdurch werden die bereits nachgewiesenen Effekte wie Motivations- und Kreativitätserhöhung gefördert. Dies führt indirekt dazu, dass die Insight-Generierung verbessert wird.

Durch die Balance aller Spielelemente sollen sich die Spieler zudem immer zwischen den Polen Über- und Unterforderung befinden, um so einen Flow-Zustand hervorzurufen [Cs90, S. 30f]. Dadurch soll erreicht werden, dass sie sich in die Spielwelt und somit besser in die Kaufentscheidungssituation hineinfühlen können als bei anderen Erhebungsmethoden. Auch hier wird somit ein indirekter Effekt auf bessere Insights erwartet.

Direkte Auswirkungen auf die Erhebung haben hingegen die bereits erwähnten Elemente wie die ästhetische Gestaltung der Spielwelt und die Ereigniskarten, insbesondere jedoch die Geschichte und das Spielziel. Die Customer Journey und die Kontaktpunkte werden im Spiel durch das Anfahren der Informationsquellen abgebildet. Dort geben die Probanden ihre grundsätzlich präferierten Medien, die jeweils gewählte Quelle sowie die konkreten dort abgerufenen Informationen bzw. durchgeführten Handlungen preis. Beispielsweise gab ein Testproband an, er würde sich über bestimmte Ausstattungsdetails persönlich bei einem lokalen Händler informieren. Dies regte unter den Spielern eine Diskussion an, warum er sich diese Information denn gerade dort (und nicht z.B. auf der Herstellerwebsite) einholen würde.

5.5 Erste Erkenntnisse

Durch das Playtesting konnten bereits sehr viele Verbesserungen des Spiels erfolgen. Diese wurden einerseits mittels Beobachtung des Spielens erfasst, andererseits wurden die verschiedenen Spieler immer wieder befragt. Einige Erkenntnisse werden im Folgenden exemplarisch genannt:

- Von den Testspielern empfundene Ungerechtigkeiten wurden eliminiert, indem z.B. das Spielfeld so angepasst wurde, dass der Weg zur Stadt aus jedem Dorf nun gleich lang ist, und indem die Informationsfelder anders verteilt wurden.
- Die Komplexität wurde reduziert, indem das zu Beginn sehr lange Regelwerk nach und nach stark gekürzt wurde. Beispielsweise wurden einige Regeln durch eindeutigere Kennzeichnungen auf dem Spielfeld überflüssig.
- Die wahrgenommene Interaktivität und der Spielspaß wurden durch die Einführung der Handkarten erhöht.
- Die Linearität des Spielablaufs konnte durch die Ergänzung von Nebenhandlungen wie auch durch die Handkarten aufgebrochen werden. Somit wurde z.B. erreicht, dass nicht ein Spieler mit viel Würfelglück sehr schnell das Spielziel erreichen kann, während die Mitstreiter weit zurückliegen und somit kaum Informationen preisgeben.

Darüber hinaus konnte die Befürchtung widerlegt werden, dass die Probanden keine sinnvollen Informationen eingeben oder nicht darüber sprechen.

5.6 Nächste Schritte

Nachdem das Spiel immer weiter entwickelt und verbessert wurde, erfolgt die Evaluation. Dazu wird die Methode in mehreren Durchgängen immer wieder mit verschiedenen Probanden angewendet. Die Ergebnisse werden in einem ersten Schritt mithilfe einer Literaturrecherche mit denen von bisherigen Untersuchungen verglichen, welche sich auch mit dem Automobilkauf beschäftigen.

Im zweiten Schritt erfolgt dann ein Experiment, in welchem die neue und eine bestehende Methode (z.B. qualitatives Interview) unter vergleichbaren Bedingungen durchgeführt werden, um die Ergebnisse direkt gegenüberstellen zu können. Zu den Bewertungskriterien, die für den Vergleich herangezogen werden, zählen u.a. die Qualität der Ergebnisse sowie der erforderliche Aufwand bei der Durchführung der Erhebung.

6 Fazit

Um Erkenntnisse über die digitaler und komplexer werdende Customer Journey zu generieren, müssen sich auch die Forschungsmethoden verändern. Die bisher angewandten Methoden haben diverse Schwächen, eine Lösungsmöglichkeit bietet der Einsatz von Spielen.

Die vorliegende Arbeit beschreibt ein Forschungsdesign, anhand dessen ein Serious Game zu diesem Zweck entwickelt, verbessert und in einem letzten Schritt mit anderen Forschungsmethoden verglichen wird. Mit dem Spiel als Erhebungsmethode sollen die Schwächen klassischer sowie prozessbegleitender und rückblickender Methoden eliminiert werden, indem Spielelemente gezielt eingesetzt werden, um besser an Insights zu gelangen – beispielsweise indem sich die Probanden besser in die Entscheidungssituation hineinversetzen können.

Eine Einschränkung des Projekts liegt darin, dass bisher nur Studenten und wissenschaftliche Mitarbeiter als Testpersonen eingesetzt wurden. Zukünftige Forschungsarbeiten sollten hier folglich die Repräsentativität für den Anwendungsfall des Automobilkaufs beachten, um die Generalisierbarkeit der Ergebnisse sicherzustellen.

Das Serious Game existiert aktuell für den konkreten Fall des Automobilkaufs. Allerdings besteht die Möglichkeit, das Spiel auch für andere Dienstleistungen oder Produkte sowie auch für gänzlich andere Prozesse einzusetzen. Geplant ist zudem eine komplette elektronische Umsetzung des Spiels, z.B. als Browserspiel oder als mobile Applikation.

Literaturverzeichnis

[Ab87] Abt, C.: Serious Games. University Press of America, Lanham 1987.

[Be11] Belz, C.; Huber, D.; Okonek, C.; Rutschmann, M.: Reales Kundenverhalten – reales Marketing. In (Belz, C.; Hrsg.): Innovationen im Kundendialog. Gabler, Wiesbaden, 2011; S. 36-66.

[Be69] Berlyne, D.E.: Laughter, humor, and play. In (Lindzey, G.; Aronson, E.; Hrsg.): The Handbook of Social Psychology. Addison-Wesley, Reading, 1969; S. 795-852.

[Bo12] Boateng, W.: Evaluating the Efficacy of Focus Group Discussion (FGD) in Qualitative Social Research. In: International Journal of Business and Social Science 3 (7). 2012; S. 54-57.

[Ca06] Caillois, R.: The Definition of Play, The Classification of Games. In (Salen, K; Zimmerman, E.; Hrsg.): The Game Design Reader: A Rules of Play Anthology. MIT Press, Cambridge, 2006; S. 122-155.

[CCF13] Cheong, C.; Cheong, F.; Filippou, J.: Quick Quiz: A Gamified Approach for Enhancing Learning. In: Proceedings of Pacific Asia Conference on Information Systems. Jeju Island, 2013.

[Ch07] Chen, J.: Flow in Games (and Everything Else). In: Communications of the ACM 50 (4). 2007; S. 31-34.

[Co09] Court, D.; Elzinga, D.; Mulder, S; Vetvik, O.J.: The consumer decision journey. In: Mc Kinsey Quarterly 3. 2009.

[Cs90] Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper and Row, New York, 1990.

[De11] Deterding, S.; Khaled, R.; Nacke, L.E.; Dixon, D.: Gamification: Toward a Definition, In: CHI 2011 Workshop Gamification. 2011. Online verfügbar unter: http://gamification-research.org/wp-content/uploads/2011/04/CHI_2011_Gamification_Workshop.pdf (letzter Zugriff am 16.04.2014).

[Di02] Diller, H.: Grundprinzipien des Marketing. GIM, Nürnberg, 2002.

[Do11] van Douwe, U.: Kundenzufriedenheit fortwährend oder retrospektiv analysieren (Teil 2). In: Research & Results 5. 2011; S. 30-32.

[Do12] van Douwe, U.: Customer Journey: Einblicke in die Kundenbeziehung. In (vsms Verband Schweizer Markt- und Sozialforschung; Hrsg.): Jahrbuch 2012 Markt- und Sozialforschung. Zürich, 2012; S. 52-55.

[FH12] Füller, J.; Hutter, K.: „Im Spiel liegt die Wahrheit“ – Games zur Insights-Generierung. In: Marketing Review St. Gallen 2. 2012; S. 26-32.

[Fö07] Föll, K.: Consumer Insight – Emotionspsychologische Fundierung und praktische Anleitung zur Kommunikationsentwicklung. DUV, Wiesbaden, 2007.

[Fu08] Fullerton, T.: Game Design Workshop: A Playcentric Approach to Creating Innovative Games. Morgan Kaufmann, Burlington, 2008.

[Gü12] Günther, V.: Customer Journey – Neue Fährtenleser für digitale Spuren. In: kressreport, 8. 2012; S. 20-21.

[Ha10] Halan, S.; Rossen, B.; Cendan, J.; Lok, B.: High Score!-Motivation Strategies for User Participation in Virtual Human Development. In: Proceedings of the international conference on Intelligent Virtual Agents. Springer, Berlin/Heidelberg, 2010; S. 482-488.

[He04] Hevner, A.R.; March, S.T.; Park, J.; Ram, S.: Design Science in the Information System Research. In: MIS Quarterly 28 (1). 2004; S. 75-105.

[He07] Hevner, A.R.: A Three Cycle View of Design Science Research. In: Scandinavian Journal of Information Systems 19 (2). 2007; S. 87-92.

[He12] Heinemann, G.: Der Neue Online-Handel: Erfolgsfaktoren und Best Practices. Gabler, Wiesbaden, 2012.

[He13a] Heinemann, G.: No-Line-Handel – Höchste Evolutionsstufe im Multi-Channeling. Springer, Wiesbaden, 2013.

[He13b] Heinemann, G.: Social Media als Spiegelbild des neuen Kaufverhaltens im Handel. In (Bruhn, M; Hadwich, K.; Hrsg.): Dienstleistungsmanagement und Social Media. Springer, Wiesbaden, 2013; S. 87-104.

[Ja12] Janke, K.: Die neue Lust am Mitmachen. In: Jahreskongress mafo 2012. Horizont/ planung & analyse Sonderpublikation. 2012; S. 8-10.

[Ka09] Kalb, H.: Design Science bei der Schwester der Wirtschaftsinformatik. 2009. http://lswiim.wordpress.com/2009/02/13/design-science-bei-der-schwester-der-wirtschaftsinformatik-teil-1 (letzter Zugriff am 16.04.2014).

[MM11] Mirchevska, T.; Markova, S.: Emergence of the Social Consumer – The impact of personalization, localization and social collaboration on consumerism. In: Annals of the University of Petroşani, Economics, 11(3). 2011; S. 229-236.

[MS08] March, S.T.; Storey, V.: Design Science in the Information System Discipline, An introduction to the special issue on Design Science Research. In: MIS Quarterly 32 (4). 2008; S. 725-730.

[MS95] March, S.; Smith, G.: Design and Natural Science Research on Information Technology. In: Decision Support Systems 15 (4). 1995; S. 251-266.

[MW12] Munzinger, U.; Wenhart, C.: Marken erleben im digitalen Zeitalter. Springer, Wiesbaden, 2012.

[Ne08] Nenonen, S.; Rasila, H.; Junnonen, J.-M.; Kaernae, S.: Customer Journey – a method to investigate user experience. In: (Alexander, K.; Hrsg.): W111 Research Report. Usability of Workplaces Phase 2. International Council for Research and Innovation in Building and Construction, Rotterdam, 2008; S. 54-63.

[Ri10] Riekhof, H.-C.: Customer Insights, Hypothesen und Markt-Wirkungsmodelle. In (Riekhof, H.-C.; Hrsg.): Customer Insights: Wissen wie der Kunde tickt. Gabler, Wiesbaden, 2010; S. 9-23.

[SBS94] Someren, M.W.; Barnard, Y.F.; Sandberg, J.A.C.: The Think Aloud Method. Academic Press, London, 1994.

[Sc08] Schell, J.: The Art of Game Design. Morgan Kaufmann, Burlington, 2008.

[SS08] Sawyer, B.; Smith, P.: Serious games taxonomy. Vorgestellt auf dem Serious Games Summit der Game Developers Conference 2008. Online verfügbar unter: http://www.dmill.com/presentations/serious-games-taxonomy-2008.pdf (letzter Zugriff am 02.05.2014).

[SZ03] Salen, K.; Zimmerman, E.: Rules of Play: Game Design Fundamentals. MIT Press, Cambridge, 2003.

[Tn10] TNS Infratest: Digital Life. 2010. Online verfügbar unter: http://2010.tnsdigitallife.com (letzter Zugriff am 16.04.2014).

[To12] Toubia, O.; de Jong, M.; Stieger, D.; Füller, J.: Measuring Consumer Preferences Using Conjoint Poker. In: Marketing Science 31 (1). 2012; S. 138-156.

[UFM13] Unger, F.; Fuchs, W.; Michel, B.: Mediaplanung – Methodische Grundlagen und praktische Anwendungen. Springer, Berlin, Heidelberg, 2013.

[We03] Wellner, A.: The new science of focus groups. In: American Demographics 25 (2). 2003; S. 29-33.

[WR12] Witt, M.; Robra-Bissantz, S.: Sparking Motivation and Creativity with "Online Ideation Games". In: Lecture Notes in Informatics (LNI), Tagungsband P208, Braunschweig, 2012; S. 1006-1023.

[WSR11] Witt, M.; Scheiner, C.; Robra-Bissantz, S.: Gamification of Online Idea Competitions: Insights from an Explorative Case. In: Lecture Notes in Informatics (LNI), Tagungsband P192, Berlin, 2011.

Enterprise Architekturen mit Big Data & Cloud

(EABDC)

Enterprise Architekturen für Big Data & Cloud

Alfred Zimmermann[1], Gunther Piller[2], Dimka Karastoyanova[3],
Oliver Bossert[4], Ulrike Steffens[5], Rainer Schmidt[6]

[1]Hochschule Reutlingen, Fakultät Informatik, Reutlingen
[2]Fachhochschule Mainz, Fachbereich Wirtschaft / Wirtschaftsinformatik, Mainz
[3]Universität Stuttgart, IAAS, Stuttgart
[4]McKinsey&Company, Frankfurt
[5]HAW Hamburg, Fakultät TI - Department Informatik, Hamburg
[6]Hochschule München, Fakultät Informatik und Mathematik, München

alfred.zimmermann@reutlingen-university.de
gunther.piller@fh-mainz.de
dimka.karastoyanova@iaas.uni-stuttgart.de
oliver_bossert@mckinsey.com
ulrike.steffens@haw-hamburg.de
rainer.schmidt@hm.edu

1 Positionierung

Software- und Unternehmensarchitekturen sind unerlässlich, um Informationssysteme auf die Erreichung von Unternehmenszielen auszurichten. Neue Modelle für flexible digitale Unternehmensarchitekturen mit Big Data und Cloud-Ökosystemen machen eine enge Kooperation verschiedener Partner aus Wissenschaft, Anwendungsunternehmen, öffentlichen Organisationen, Softwarehersteller und Hosting-Unternehmen notwendig.

Anforderungen bezüglich Flexibilität und Agilität von Informationssystemen im Kontext von Big Data und Cloud-Computing können besonders gut von serviceorientierten Architekturen bewältigt werden. Auf Basis der Analyse und Modellierung von Geschäftsprozessen werden Services so definiert, dass sie bei sich verändernden Anforderungen flexibel und schnell zusammengestellt werden können. Wesentliches Ziel ist, dass einzelne Services in agil veränderbaren Geschäftsprozessen wiederverwendet werden können und somit die Harmonisierung von Anwendungslandschaften durch zusammenarbeitende Partner unterstützt wird. Der Ordnungsgrad und die kosteneffiziente Gestaltung integraler Landschaften von Informationssystemen soll durch EAM – Enterprise Architecture Management – passend zu den neuen Möglichkeiten von Services & Cloud Computing, Big Data, sowie kollaborativen Geschäftsprozessen deutlich verbessert werden.

Die Workshop-Idee knüpft an die erfolgreichen Vorgängerworkshops auf den Informatik-Jahrestagungen 2011-2013 der GI - Gesellschaft für Informatik - an. Zudem baut er auf aktuelle Forschungsarbeiten der GI-Fachgruppe Software-Architektur und des

Arbeitskreises Unternehmensarchitektur der GI im Dialog zwischen der Wirtschaft und Wissenschaft auf.

Das Thema des Workshops „Enterprise Architekturen mit Big Data & Cloud“ ist durch die umfassende Sichtweise der Beiträge aus Wissenschaft und Praxis inhaltlich breit angelegt. Der Workshop richtet sich gleichermaßen an Vertreter aus der Wissenschaft, an IT-Manager und an Enterprise- und Systemarchitekten von Anwendern, Herstellern und Dienstleistern, die sich mit IT-Unternehmensarchitekturen und Systemen, serviceorientierten Architekturen, Cloud Computing Architekturen und dem Einsatz dieser Technologien in zunehmend kollaborativ organisierten Communities beschäftigen.

2 Ziele

Ziel des Workshops „Enterprise Architekturen mit Big Data & Cloud“ ist es, zentrale Herausforderungen an Unternehmensarchitekturen und service-orientierte Systeme durch Big Data und Cloud Computing transparent zu machen und Lösungswege aufzuzeigen.

Im Zentrum stehen hierbei drei wesentliche Aspekte und deren interdisziplinäre Verknüpfungen:

(i) Enterprise Architekturen für eine agil ausgerichtete Informatik in zukunftsweisenden Unternehmen,
(ii) Architekturen für Big Data-Technologien, Methoden und Anwendungssysteme,
(iii) Anwendungsarchitekturen und Architekturen für Services & Cloud Computing.

Als zugehörige Grundlagen und Forschungsansätze stehen zudem zur Diskussion: Digitale Ökosysteme und Architekturen, Referenzmodelle und Referenzarchitekturen, Architektur-Metamodelle, Capability-Maps, Frameworks, Semantik-Unterstützung, Visualisierung, Interaktion, Geschäftsprozessmanagement, große strukturierte und unstrukturierte Daten, Analyse-, Entscheidungs- und Diagnostikmodelle, Mobilitätssysteme, Architektur-Werkzeuge, Methoden, Prozesse, Architektur-Patterns, Architektur Governance und Management, Qualität, u.a.

3 Programm

Das Programm des ganztägigen Workshops umfasst im Kern die Präsentationen zu sechs für die Publikation ausgewählten und mehrfach begutachteten Aufsätzen. Darüber hinaus haben wir drei thematisch und strategisch ergänzende eingeladene Vorträge - passend zum Thema des Workshops - ausgewählt.

Die Key-Note von Oliver Bossert von McKinsey Frankfurt berichtet über Enterprise Architekturen als Voraussetzung für digitale Innovation und konzentriert sich auf neuartige Erkenntnisse über den Stellenwert digitaler Architekturen für die Innovation von Geschäftsmodellen und IT-Systemen. Die praktischen Erfahrungen des Autors als strategieorientierter Unternehmensberater und einer Vielzahl von praktischen Beispielen aus unterschiedlichen Branchen werden wirkungsvoll mit einer theoretischen Fundierung zusammengeführt.

Der zweite eingeladene Vortrag von Christian Schweda von der leanIT42 GmbH München über Erfahrungen und Perspektiven von Unternehmensarchitekturen in Theorie und Praxis fokussiert auf den agilen Einsatz von Modellen zur Enterprise Architecture & IT-Management mittels flexibel auswählbarer Bausteine. Jahrelange intensive Erfahrungen aufgrund seiner originellen Forschungsarbeit an der TU München werden mit praktischen Erkenntnissen beim industriellen Einsatz von schlanken EA-Modellen und Werkzeugen erfolgreich zusammengeführt.

Der Aufsatz von Rainer Schmidt, Michael Möhring, Alfred Zimmermann, Ralf-Christian Härting, Andreas Graule, Marius Schwarzer und Kristian Große berichtet über neue Ansätze zur Gestaltung datenzentrierter Unternehmensarchitekturen für den Bekleidungseinzelhandel. Sie basieren auf dem Einsatz von RFID zur Gewinnung von Kundenprofilen in den Niederlassungen und dem Einsatz von Big-Data basierten Auswertungs- und Analysemechanismen. Mit den vorgestellten Konzepten ist es Unternehmen des Bekleidungseinzelhandels möglich, ähnlich wie Versandunternehmen, individuelle Ansprachen des Kunden und Angebote zu entwickeln.

Der Aufsatz von Alfred Zimmermann, Kurt Sandkuhl, Rainer Schmidt, Dierk Jugel, Matthias Wißotzki und Michael Möhring berichtet über adaptive Digitale Enterprise Architekturen für Big Data und Cloud-Systeme. Big Data und Cloud Systeme werden zunehmend von mobilen, benutzerzentrierten und agil veränderbaren Informationssystemen im Kontext von digitalen sozialen Netzwerken genutzt. Metaphern aus der Biologie für lebendige und selbstheilende Systeme und Umgebungen liefern die Basis für intelligente adaptive Informationssysteme und für zugehörige serviceorientierte digitale Unternehmensarchitekturen.

Der Aufsatz von Kurt Sandkuhl und Matthias Wissotzki behandelt die Idee der Übertragung der Konzepte aus Wertschöpfungsketten von Materialflüssen hin zu Datenflüssen im Unternehmen. Der Aufsatz postuliert ein neuartiges, datenzentriertes EAM - Enterprise Architecture Management. Dabei werden drei Ansätze kombiniert: (i) Informationsbedarfe werden mittels Information Demand Patterns beschrieben, (ii) Wertschöpfungsketten aus der Informationslogistik werden Basiskonzepte für die Wertschöpfung von datenzentrierten Lieferketten betrachtet, (iii) Wissensservices als Basiselemente einer Architektur für EAM von datenzentrierten Lieferketten.

Die Autoren Santiago Gómez Sáez, Vasilios Andrikopoulos und Frank Leymann bearbeiten im Aufsatz Performance-aware Application Distribution in the Cloud einen neuartigen Forschungsansatz zur Analyse des Verteilungsgrades und der adaptierbaren Topologie von Applikationsschichten für massiv skalierbare Cloud–Umgebungen. Mittels einer empirischen Workload-Analyse wird die initiale Evaluation einer Persistenzschicht einer verteilten Applikation unter verschiedenen Cloud-Szenarien untersucht.

Immer mehr Unternehmen setzen auf IT-Services aus der Cloud. Diese Entwicklung verändert auch die Auswahlverfahren für IT-Produkte und Dienstleistungen. Oliver Bossert, Michael Falkenthal, Mark Mauerwerk und Gunther Piller beschreiben den Einfluss von Cloud Services auf Auswahlverfahren für IT-Lösungen. Im Rahmen einer explorativen Untersuchung werden beispielhaft Entscheidungsprozesse und Kriterien zur Bewertung von Angeboten für betriebliche Anwendungssysteme sowie für IT-Plattformen und Anbieter aufgezeigt.

Die Autoren Andreas Jede und Frank Teuteberg untersuchen Cloud-gestützte Supply-Chain-Architekturen – Entwicklung eines Referenzmodells auf Basis einer

systematischen Literaturanalyse. Der vorliegende Beitrag bietet eine interdisziplinäre Literaturanalyse. Dabei werden 102 Publikationen, die den CC-Einsatz in SCM-Bereichen thematisieren, im Hinblick auf Problemfelder, IT-Artefakte sowie Erfolgs- und Einflussfaktoren analysiert und zu einem Referenzmodells zusammengeführt.

Die abschließende Key-Note von Ulrike Steffens von der HAW Hamburg und Co-Sprecherin des Arbeitskreises EAM (Enterprise Architecture Management) der GI berichtet über das neu aufgebaute Architektur- und Integrationslabor der HAW.

Der Workshop „Enterprise Architekturen für Big Data & Cloud“ umfasst vier Sitzungen, die jeweils durch Pausen für informelle Diskussionen umrahmt sind. Um einen adäquaten Gesamtrahmen zu setzen, haben wir in der einführenden Session I die beiden einleitenden eingeladenen Vorträge positioniert. Die Session II umfasst die Themen EAM – Enterprise Architekturen für Big Data und Cloud, während die Session III die Aspekte Cloud Computing und Big Data detailliert. Den Abschluss bildet die Session IV mit dem eingeladenen Vortrag über das Integrationslabor für experimentelle Studien bei flexibel verteilten Systemen an der HAW Hamburg sowie abschließende Reflexionen und wissenschaftliche Diskussionen der Teilnehmer.

4 Programmkomitee

Oliver Bossert – McKinsey Frankfurt
Karlheinz Blank – T-Systems Stuttgart
Robert Hirschfeld – Hasso Plattner Institut Potsdam
Ulrich Kalex – alfabet Berlin
Michael Gebhart – KIT Karlsruhe
Dimka Karastoyanova – Universität Stuttgart
Stefan Kühne – Universität Leipzig
Martin Mähler – IBM Böblingen
Gunther Piller – Fachhochschule Mainz
Michael Pretz – Daimler Stuttgart
Karl Prott – Capgemini Hamburg
Manfred Reichert – Universität Ulm
Wilfried Reimann – Daimler Stuttgart
René Reiners – FIT Sankt Augustin
Kurt Sandkuhl – Universität Rostock
Rainer Schmidt – Hochschule München
Christian Schweda – München
Ulrike Steffens – HAW Hamburg
Gottfried Vossen – Universität Münster
Alfred Zimmermann – Hochschule Reutlingen

5 Organisationsteam

Alfred Zimmermann – Hochschule Reutlingen
Dimka Karastoyanova – Universität Stuttgart
Ulrike Steffens – HAW Hamburg
Gunther Piller – Fachhochschule Mainz
Oliver Bossert – McKinsey Frankfurt
Rainer Schmidt – Hochschule München

Datenzentrierte Unternehmensarchitekturen im Bekleidungseinzelhandel

Rainer Schmidt[1], Michael Möhring[2], Ralf-Christian Härting [2], Alfred Zimmermann[3]
Andreas Graule[1]; Marius Schwarzer[1]; Kristian Goßen[1]

[1]Fakultät für Informatik und Mathematik
Hochschule München
Lothstraße 34, 80335 München
rainer.schmidt@hm.edu

[2]Fakultät für Wirtschaftswissenschaften
Hochschule Aalen
Beethovenstraße 1, 73430 Aalen
michael.moehring@htw-aalen.de; ralf.haerting@htw-aalen.de

[3] Fakultät Informatik
Hochschule Reutlingen
Architecture Reference Lab
Alteburgstraße 150, 72762 Reutlingen
alfred.zimmermann@reutlingen-university.de

Abstract: Der lokale Bekleidungseinzelhandel steht unter immer stärkerem Konkurrenzdruck durch Versandunternehmen. Zusätzlich bestehen durch gewachsene Architekturen eine Reihe von Wachstumshemmnissen. Daher sollen hier eine Reihe von Ansätze zur Gestaltung datenzentrierter Unternehmensarchitekturen für den Bekleidungseinzelhandel vorgestellt werden. Sie basieren auf dem Einsatz von RFID zur Gewinnung von Kundenprofilen in den Niederlassungen und dem Einsatz von Big-Data basierten Auswertungs- und Analysemechanismen. Mit den vorgestellten Konzepten ist es Unternehmen des Bekleidungseinzelhandels möglich, ähnlich wie Versandunternehmen, individuelle Ansprachen des Kunden und Angebote zu entwickeln.

1 Aktuelle Situation und Herausforderungen

Der Bekleidungseinzelhandel in Deutschland ist mit über 59 Milliarden Euro jährlichem Konsumausgaben deutscher Verbraucher ein wichtiger Bestandteil der Einzelhandelsstruktur in Deutschland und wächst kontinuierlich [Hand13]**.** Besonders die Kunden, die Textilien im höheren Preissegment nachfragen, schätzen eine kompetente Beratung in einem Fachgeschäft und weichen daher meist nicht auf Online- Händler aus [Derh00] [KeHa06]. Weiterhin steigen die Konsumausgaben der Deutschen für Kleidung kontinuierlich an, welches die Bekleidungsbranche insgesamt positiv beeinflusst [Dest13]. Im Gegensatz zu etablierten Onlinehändlern wie Amazon.de, Zalando oder der Otto

Group, fällt es aber Einzelhändlern mit klassischem Filialen schwer Cross-Selling Umsätze zu steigern, da Konsumenten und Konsuminteressen oft nicht so transparent analysiert werden können, wie dies beispielsweise in Onlineshops (bspw. über Kauf- und Suchhistorie) der Anwendungsfall ist.
Es stellt sich daher die Herausforderung, wie durch eine geeignete Gestaltung der Unternehmensarchitektur die Unternehmen des Bekleidungseinzelhandles in die Lage versetzt werden können, echtzeitnahe Analysen des Kundenverhaltens zu erstellen und auf dieser Basis ihre Cross-Selling-Umsätze zu steigern.
Daher sollen hier datenzentrierte Unternehmensarchitekturen konzipiert werden, die Einzelhändler im Bekleidungshandel unterstützt, ihre Cross-Selling Umsätze zu steigern. Sie basiert auf einer grundlegenden Organisation der bestehenden IT-Services und stützt sich auf neue Technologien aus dem Big Data Umfeld sowie RFID Komponenten. Auf diese Weise kann die hohe Heterogenität und mangelnde Integration [BeWi04] [BeHP12] der bestehenden Systeme überwunden werden. Im nachfolgenden werden die Grundlagen zu den dazu benötigten Big Data Technologien, RFID und mögliche Geschäftsprozessszenarien aufgezeigt und in eine mögliche Enterprisearchitektur überführt.

2 Gestaltungsprinzipien für Unternehmensarchitekturen

Bei der Konstruktion einer geeigneten Architektur müssen verschiedene Architekturprinzipien zur Reduktion von Komplexität und Erhöhung der Flexibilität mittels Gestaltungsprinzipien umgesetzt werden [HeSt11]. Da Unternehmen in der Bekleidungsindustrie eine Vielzahl von IT-Systemen und darin implementierten Services betreiben [BeWi04] [BeHP12], sollte vor allem das Prinzip der losen Kopplung (geringe Abhängigkeit der Elemente einer Architektur voneinander) berücksichtigt werden [HeSt11]. Um dieses Gestaltungsprinzip zu befolgen, sollten implementierte sowie bereitgestellte Softwareservices vor allem eine hohe Kohäsion (hoher Zusammenhang der Teilfunktionen in einem Service) aufweisen [HeSt11]. Weitere wesentliche Gestaltungsprinzipien, welche zu berücksichtigen (teilweise überlappend) sind, sind in nachfolgender Tabelle 1 dargestellt [HeSt11] .

Gestaltungsprinzipien
- Prinzip des Entwurfs für Veränderungen - Separation-of-Concerns-Prinzip - Information-Hiding-Prinzip - Abstraktionsprinzip - Rückverfolgbarkeitsprinzip, etc.

Tabelle 1: Gestaltungsprinzipien nach [HeSt11, S. 52]

Service-orientierte Architekturen (SOA) [CLSF05] können die oberen Anforderungen erfüllen und so Geschäftsprozesse durch verschiedene IT-Services unterschiedlichster Informations- und Anwendungssysteme bilden. Service-Orientierte Enterprise Architekturen [NuSc09] wenden Services als Gestaltungsparadigma an und unterscheiden vier Ebenen von Services (siehe Abbildung 1). Business Services unterstützen Geschäftsprozesse direkt. Ihrerseits werden sie durch Komposition von Software-Services bereitgestellt. Software-Services basieren wiederum auf Platform-Services. Ihrerseits verwenden Sie wiederum Infrastruktur-Services. Vorteil bei Verwendung dieses Paradigmas ist u.a., dass Geschäftsprozesse stringent abgebildet werden und zugrundeliegende Services wiederverwendet und bei datenintensiven Diensten (wie in diesem Szenario BI-Anwendungen) skalierbar sind. Medien- und schwerwiegende Systembrüche aufgrund mangelnder IT-Integration, wie bisher vorliegend, können vermieden werden.

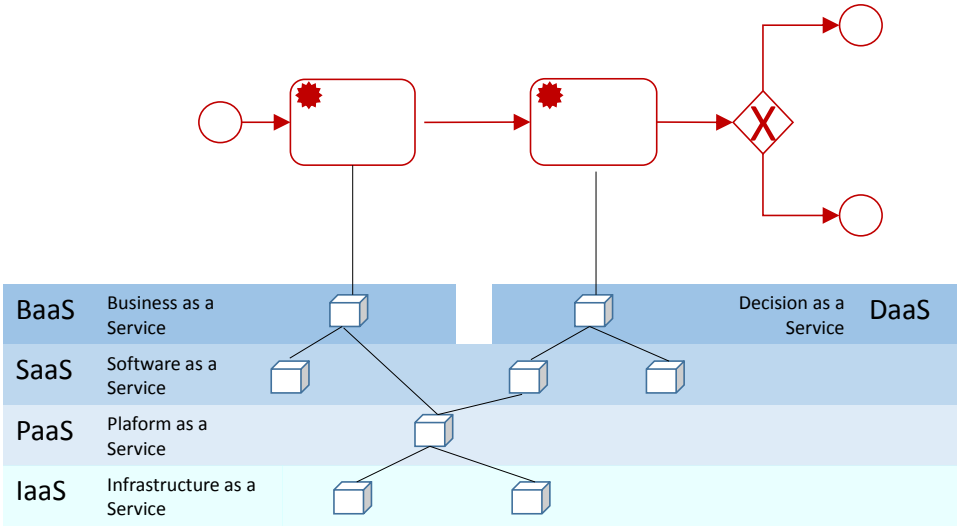


Abbildung 1: Service-Orientierte Unternehmensarchitektur

3 Entwicklung datenzentrierter Architekturen für den Bekleidungseinzelhandel

Bei einer europaweit agierenden Sportfachhandelskette des Mittelstands werden die Herausforderungen für die Entwicklung datenzentrierter Architekturen deutlich. Beim jeweiligen Händler vor Ort vorliegende IT-Systeme sind wenig integriert. Datenübernahmen (beispielsweise aus dem Kassen-IT-System) erfordern in der Regel System- oder Medienbrüche, welches die Geschäftsprozessqualität und damit die Kosten und Zeit zur Ausführung in einem hohen Maß negativ beeinflusst. Eine Vielzahl von heterogenen, nicht integrierten IT-Systemen wird betrieben und zusätzlich filialspezifisch um zusätzliche spezielle (ggf. auch Eigenprogrammierungen) IT-Systeme in der

Komplexität gesteigert. Eine Anbindung oder Bereitstellung an Business-Intelligence-Systeme existiert nicht. Kundenkarten existieren, werden jedoch nicht über RFID sondern über einen 2D-Barcode ausgelesen. Die Bereitstellung von Stamm- und Bewegungsdaten aus Warenwirtschaftssystemen wird nur stark zeitverzögert (aufgrund u. a. der Rechenintensität) gewährleistet, da skalierbare und kurzfristig recheneffiziente IT-Systeme nicht im umgesetzten Business Case sind. Um obige Szenarien umzusetzen, muss daher die Architektur der IT-Systeme zur Unterstützung bestehender und neuer (daten- und rechenintensiven) Geschäftsprozessen grundlegend überdacht werden.

Damit der Bekleidungseinzelhandel das Konsumentenverhalten ähnlich zu etablierten Onlineshops wie Amazon.de oder Zalando erheben und beeinflussen kann, sind deutliche Erweiterungen und Veränderungen der etablierten Unternehmensarchitekturen notwendig.

Ausgangspunkt ist die Erhebung von Daten zum Konsumentenverhalten. Sehr weit verbreitet in der Modebranche ist die Verwendung der Radio-Frequency-Identification [Fink02] (RFID). RFID ist eine Technologie zur Identifikation und Lokalisierung von Objekten mit Hilfe von elektromagnetischen Wellen. Das Lesegerät sendet dabei ein elektromagnetisches Wechselfeld aus. Gerät ein sogenannter RFID-Tag in dieses Wechselfeld, so nutzt dieser dazu, einen Chip mit Strom zu versorgen. Dieser Chip dekodiert die vom Lesegerät gesendeten Befehle und „sendet“ seinerseits Informationen wie seine Identifikationsnummer etc. durch Beeinflussung des vom Lesegerät ausgesandten Feldes. Damit ermöglicht diese Technologie kontaktlos Objekte zu identifizieren und bildet Grundlage für neue daten-getriebene Geschäftsprozesse. Rechtliche Grundlagen und Rahmenbedingungen sind in verschiedenen Einsatzsituationen im Einzelhandel bereits diskutiert [Müll04] [KnHa05] und je nach Rahmenbedingungen und Ausgestaltungen zulässig.

RFID-Erkennung erfolgt vor allem durch in Kleidung vernähte Chips [Wieh00] [Lang05] [Robe06]. RFID wird in der Bekleidungsbranche auch zur gesamten Optimierung der Supply-Chain eingesetzt werden [KoPe07]. So verwenden Unternehmen wie bspw. Gerry Weber [Wieh00] RFID zur Unterstützung von logistischen Geschäftsprozessen.

Durch Einsatz von RFID mit Hilfe von in Kleidung vernähte Chips ist eine, eindeutige Kundenidentifikation theoretisch möglich. In der Praxis gibt es jedoch etliche Einschränkungen. So müsste der Kunde immer mit mindestens einem durch RFID bekannten Kleidungsstück die Filialen betreten. Der eingenähte Chip dann zudem durch Waschen und ähnliche Prozesse beschädigt werden.

Eine Alternative stellt die Integration von RFID-Chips in Kundenkarten dar [BaYY10] [Robe06] [Mcbr11]. Der Kunde kann beim Beitreten der Filiale durch ein weitreichendes Lesegeräte (ca. 5-10m) identifiziert werden. Durch viele RFID Deckenantennen im Shop werden Kunden anhand der Kundenkarte getrackt. Dadurch kann jeder Aufenthalt in einem Produktbereich, wie in der folgenden Abbildung gezeigt, erfasst werden. Dazu wird seine Kunden-ID aus dem in der Kundenkarte enthaltenen RFID-Chip ausgelesen und mit den im CRM bzw. ERP-System gespeicherten Daten verknüpft.

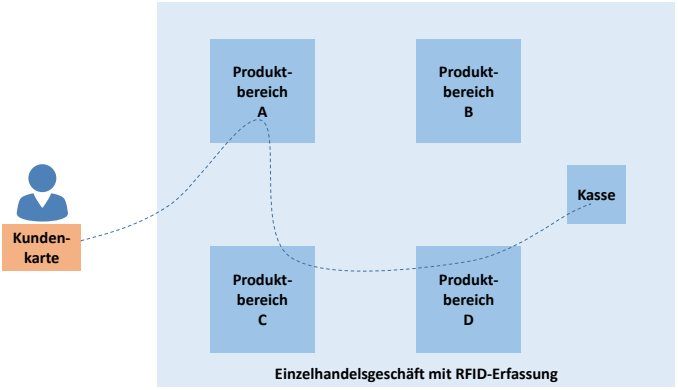


Abbildung 2: RFID-Kundenidentifikation und -tracking

Wenn der Kunde die Waren (physisch) evaluiert hat, kann er zur Kasse gehen oder die Filiale verlassen. Sein Tracking-Profil (4) [Mcbr11] wird über das RFID-System im CRM-System in den Kundenbewegungsdaten gespeichert werden. Dieser Speichervorgang ist je nach rechtlicher Ausgestaltung und Einbindung (sowie Mitwissenheit) des Kundenzulässig [Müll04] [KnHa05]. Alleine durch Auswertung der Verweildauern vor verschiedenen Produktgruppen kann ein intelligentes Category Management durchgeführt werden.

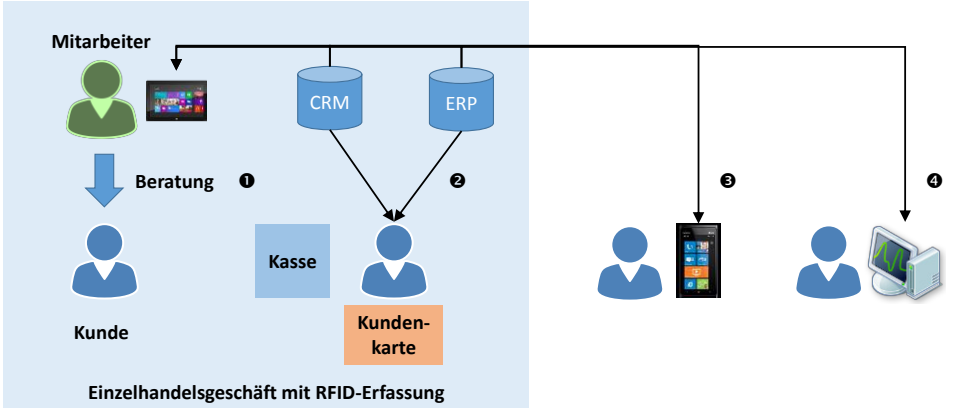


Abbildung 3: Datenzentrierte Einzelhandelsarchitektur

Je nach gewähltem Einsatzszenario sind unterschiedliche IT-Systeme in die jeweiligen Geschäftsprozesse integriert. Basierend auf der Erfassung der Kundeninformation lassen sich nun die folgenden Erweiterungen der klassischen Unternehmensarchitekturen im Einzelhandel vornehmen. Sie sind in Abbildung 3 dargestellt. Grundlage sind die in [Mcbr11] entwickelten Szenarien. Bereits vor dem Kauf können die Mitarbeiter des Unternehmens über die anwesenden Kunden und ihre mutmaßlichen Interessen informiert werden. Durch Integration von Tablets, Smartphones, Wearables oder Google Glass für Verkaufspersonal kann der Kunde bereits beim Betreten des Shops identifiziert und

Verkäufer benachrichtigt sowie mit relevanten Kundendaten versorgt werden. Der Verkäufer kann zudem weitere Angebote und Informationen über und für den Kunden erhalten. Hierdurch können die Mitarbeiter den Kunden individuell beraten (1). Die Kundenidentifikation an der Kasse kann genutzt werden, um kundenindividuelle Rabatte und Angebote (Cross-Selling) sowie eine persönliche Ansprache bereitzustellen (2). Die während des Besuchs des Kunden gesammelten Informationen können auch für Aftersales-Maßnahmen gesammelt werden. Um Aftersale-Potenziale zu ermöglichen werden auf Basis bestehender Kaufhistorie und spezifischen Produktedaten aus CRM- und ERP-System in BI-Systemen Data-Mining Analysen (wie bspw. Cluster- oder Assoziationsanalysen) durchgeführt [GlGC97]. Als Ergebnis stehen potenzielle Produkte und ggf. kundenindividuelle Preise fest, welche dem Kunden im nächsten Funktionsschritt via SMS (3) und E-Mail (3) übermittelt werden, auf Basis der nun im CRM-System hinterlegten Ergebnisse des Vorschrittes. Die beschriebene Situation der Erkennung und Dokumentation des Bewegungsprofils für ein intelligentes Category Management stellt ein weiteres Szenario dar (4).

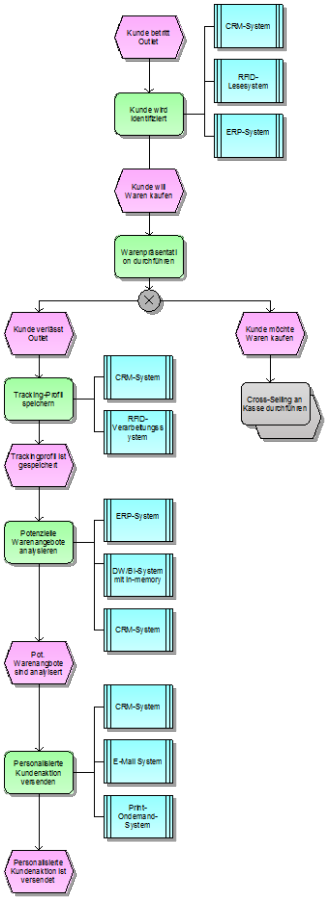


Abbildung 4: Beispielprozess Aftersales.

Am Beispiel des Geschäftsprozesses 3 „Aftersales“ sind die verschiedensten IT-Systeme mit unterschiedlichen Teilfunktionen bzw. Services beteiligt (siehe Abbildung 4). Wie dargestellt sind unterschiedlichste in der Regel nicht integrierte IT-Systeme an diesem Prozess beteiligt. Aufgrund der Vielzahl an zu verarbeitenden Daten (basierende auf den Analysen) ist weiterhin eine skalierbare und flexible Architektur nötig. Zur beschleunigten Verarbeitung von großen Datenmengen, unterschiedlichster Struktur bieten sich vorgestellte Technologien aus dem Big Data Kontext an.

Big Data hat ein hohes Maß an Aufmerksamkeit in Industrie und Forschung erhalten [LLSH11]. Die größten Nutzenpotentiale von Big Data bestehen bei deutschsprachigen Unternehmen vor allem in den Geschäftsbereichen der IT und Technologieentwicklung sowie des Marketings [SMMP14] [SMMP14] [SMMP14]. Es handelt sich aber nicht um eine spezifische Technologie oder Technologieplattform wie beispielsweise Hadoop [Whit12], sondern ist vielmehr eine durch eine Reihe von technologischen Fortschritten ermöglichte deutliche Ausdehnung der Entscheidungsunterstützung im betrieblichen Umfeld zu begreifen. Die Entscheidungsunterstützung in Unternehmen stützte sich lange Zeit in erster Linie auf strukturierte Daten aus internen Quellen. Diese stammen beispielsweise aus geschäftlichen Transaktionen und werden typischer durch ERP-Systeme verwaltet. Diese strukturierten Daten wurden dann durch Extraktion, Laden und Transformieren in ein Data Warehouse [ChDa97] übertragen. Hier standen diese für analytische Funktionen im Rahmen von Business Analytics [Dave10] oder Business Intelligence [KeBL13] zur Verfügung. Typisch ist hierbei eine eher beschreibende, vergangenheitsorientierte Sichtweise. Wegen des hohen Aufwandes und der langen Durchlaufzeiten wurden in erster Linie strategische Entscheidungsfindungsprozesse unterstützt. Bereits bei der Auswertung strukturierter Daten waren die klassischen Verfahren die im Kern auf einem zentralisierten Verarbeitungsansatz beruhen überfordert. Erst recht wurden diese Einschränkungen deutlich, als es um die Verarbeitung von semi- und unstrukturierten Daten geht, wie sie oft von externen Quellen stammen. Typisch sind hier Kundenkommentare die von einer Web-Site stammen und sich nicht an vordefinierte Strukturen halten. Um solche semi- und unstrukturierten Daten zu verarbeiten wurden eine Reihe von Technologien entwickelt, die sich meist durch eine deutliche Abweichung von den klassischen Systemarchitekturen auszeichnen, der Verwendung eines zentralisierten Datenbankansatzes und der Verwendung relationaler Schemata [Codd72]. Weitere Beispiele neben dem bereits erwähnten Hadoop sind NoSQL-Datenbanken [FHBB11].

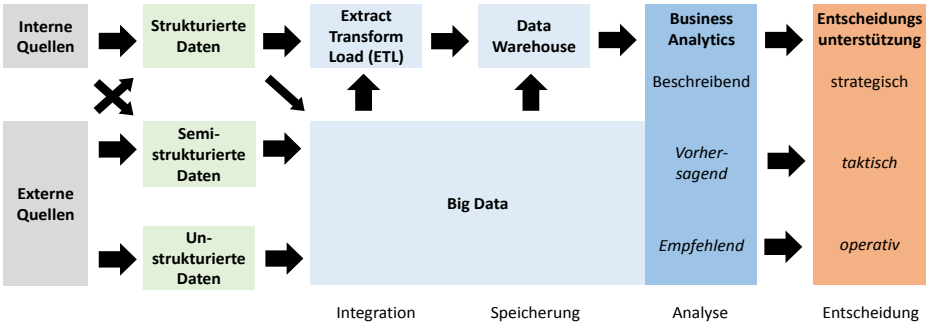


Abbildung 5: Big Data zur Entscheidungsunterstützung in Unternehmen

Durch eine hoch-parallele Architektur ist die Verarbeitung auch sehr großer Datenmengen möglich. Außerdem wird auf die Bildung von Datenzwischenprodukten weitgehend verzichtet. Es werden also nicht wie beim Data-Warehouse-Ansatz die Phasen Integration, Speicherung und Analyse strikt getrennt sondern es findet vielmehr eine starke Überdeckung der Phasen statt. Beides zusammen ermöglicht es im Kontext von Big Data in deutlich größerem Umfang vorhersagende oder sogar empfehlende Analysen zu erstellen. Die deutlich höhere Geschwindigkeit und die geringeren Kosten der Analyseerstellung ermöglichen zudem den Einsatz auch für taktische und operative Entscheidungsprozesse.

Besonders durch Nutzung von Big Data können folgende Vorteile im Vergleich zu klassischem Kundenloyalitätsprogrammen erzielt werden [Mcbr11]. Der Kunde kann individuell angesprochen werden und es ist möglich, spezielle kundenbasierte Rabatte am Point-of-Sale zu unterbreiten.

4 Gestaltungsempfehlungen für eine Enterprise-Architektur

Nach Becker und Schütte [BeSc04] stellt eine durchgehende Daten-, Funktion- und Prozessintegration im Handel eine wesentliche Grundlage funktionierender und wirtschaftlicher IT-Systeme dar. Dies kann ebenfalls durch eine SOA abgebildet werden, da Funktionen und Daten je nach Geschäftsprozesskomposition (und damit Geschäftsprozesse) miteinander integriert sind. Ebenfalls ändern sich häufig Anforderungen im Unternehmen und der Konsumenten (bspw. durch geänderte Compliance-Anforderungen oder techn. Neuerungen). Folglich muss die zugrundeliegende Architektur agil sein. Denn die IT-Anwendungslandschaft gilt als zentrale Voraussetzung für eine agile IT [TeRe12]. Diese Anforderung kann ebenfalls durch eine SOA abgebildet werden (bspw. durch Prinzip der losen Kopplung).

Eine weitere Anforderung ist die Skalierbarkeit der Services der SOA für die technische Umsetzung der vier Szenarien. Vor die kundenindividuelle Ansprache und Aftersales sind sehr rechenintensiv. Jedoch muss nicht zu jeder Zeiteinheit die nötige Rechenkapazität zur Verfügung stehen, da der Bedarf nicht zu jeder Zeit existiert und Kundenströme nicht gleichverteilt über den Tag erfolgen. Daher sollte ein Flexibilisierungsansatz dieser Services in Betracht gezogen werden. Die Bereitstellung dieser Services bspw. über eine Public-Cloud ermöglicht rechenintensive Services auszulagern und trotzdem agil die Geschäftsprozesse bereit zu stellen [ScMö13] [MKSM13]. So können rechenintensive Berechnungsprozesse von bspw. Assoziationsregeln oder Clusterungen in die Public-Cloud (daten-anonymisiert) ausgelagert werden [ScMö13].

Die Aufzeichnungen von kontinuierlichen Datenströmen (wie den RFID-Daten zur bspw. Bewegungsmustererkennung) sollte über eine Bandbreitenzuverlässige unabhängige Netzwerkinfrastruktur erfolgen, um das Antwortzeitverhalten anderer operativer System (wie bspw. Kassen-IT-Systeme) nicht direkt zu beeinflussen.

Zusammenfassend können für die Szenarien nachfolgende Kernanforderungen identifiziert und bewertet werden (Bewertungsskala: +: wenig wichtig; ++: wichtig; +++: äußert wichtig):

Szenario	Daten-integration	Funktions-integration	Prozess-integration	Rechen-intensive Funktionen	Skalier-barkeit	Antwort-zeit-verhalten	Agilität
1	+++	+++	+++	+++	+++	+++	+++
2	+++	+++	+++	+++	+++	+++	+++
3	+++	+++	+++	++	++	+	+++
4	+++	+++	+++	+	+++	+++	+

Tabelle 2: Anforderungen der Szenarien

Je nachdem, welches der vier Szenarien bzw. Geschäftsprozesse durchlaufen werden, sind unterschiedlichste IT-Systeme durch deren Services repräsentiert und müssen den Prozess korrekt abbilden. Im Rahmen dieser Service-orientierten Architektur können aufwendige Analyseprozesse genauso wie einfache Kundenbestellungsverbuchungen auf Daten aus dem RFID-System zurückgreifen und für unterschiedlichste Anwendungen nutzen.

Im Rahmen einer praktischen Implementierung ist eine derartige SOA beispielsweise im SAP-nahem Umfeld über die Netweaver-Plattform und externe (Web)-Services zur Datenanalyse bspw. über Amazon AWS oder Microsoft Azure [Tull13] realisierbar. Entgeltfreie Open-Source-Lösungen für kleine und mittlere Unternehmen zur Realisierung des Ansatzes sind ebenfalls möglich.

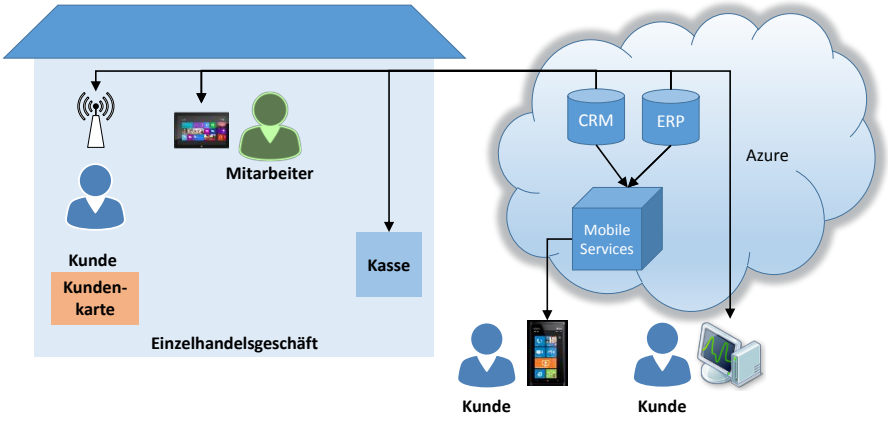


Abbildung 6: SOA auf Azure Basis als Implementierungsgrundlage

Eine Lösung als hybride Cloud [MeGr09] unter Zuhilfenahme von Microsoft Azure [Tull13] ist in der obigen Grafik skizziert. Im Ladengeschäft des Textileinzelhändlers

werden über RFID-Lesegeräte das Erscheinen, sowie der aktuelle Aufenthaltsort des Kunden mit seiner Kundenkarte erfasst. Diese Informationen werden mit den Informationen des CRM und ERP-Systems verknüpft, die in Azure betrieben werden. Auf diese Weise können auf die Tablet-PCs der Mitarbeiter Informationen über die gerade anwesenden Kunden weitergeleitet werden. Durch Verwendung von Software-Defined-Networking [BeWM13] ist es möglich, diese Tablet-PCs direkt mit den in Azure betriebenen CRM- und ERP-Systemen zu verbinden, ohne Rückgriff auf Mobilfunkverbindungen nehmen zu müssen. Beim Erscheinen des Kunden an der Kasse ist durch Erfassung seiner Kundenkarte und dementsprechenden Zugriff auf seine Daten in CRM und ERP-System eine persönliche Ansprache möglich. Darüber hinaus können dem Kunden noch vor Verlassen des Geschäfts Zusatzangebote (Cross-Selling) gemacht werden. Vor allem wenn der Kunde das Geschäft ohne Kauf verlassen hat, kann die cloud-basierte Anbindung mobiler Dienste zu Einsatz kommen. Für das Mobilgerät des Kunden können nun individuell abgestimmte Angebote generiert werden, um Verkaufspotentiale zu heben.

5 Fazit und Ausblick

Um im Wettbewerb mit Onlinehändlern besser zu bestehen, müssen Filialisten neue Wege der datengetriebenen Analyse und Steuerung des Konsumentenverhaltens gehen. Textileinzelhändler neue Unternehmensarchitekturen entwickeln, die eine echtzeitnahe Auswertung des Kundenverhaltens ermöglichen. Auf dieser Basis können dann gezielte Marketing-Aktionen initiiert werden. Zusätzlich wird so eine deutlich bessere Informationsversorgung des Vertriebs ermöglicht. Grundlage der hier vorgestellten datenzentrierten Unternehmensarchitektur ist die Erfassung einer RFID-basierten Kundenkarte, die es ermöglicht Kunden und ihr Verhalten in den Verkaufsräumen zu erfassen. Durch Anwendung von Technologien aus dem Big Data Kontext ist eine echtzeitnahe Auswertung der Information möglich. Umgesetzt werden kann die vorgestellte Architektur auf der Basis von service-orientierten Software-Systemen, wie beispielsweise Cloud-basierten Systemen.

Verschiedene Einsatzszenarien von RFID zur Kundenidentifikation und Analyse des Kaufverhaltens und ähnlicher Konsuminteressen über verschiedene Datensilos hinweg ermöglichen eine nachhaltige Implementierung dieses Ansatzes. Jedoch sind auch datenschutzrechtliche Vorschriften zur rechtssicheren Datenerhebung zu beachten und evtl. Akzeptanzproblemen von Kunden entgegen zu wirken (bspw. durch spezielle Rabattaktionen und kundenspezifische Kauferfolgserlebnisse). Rechtliche Rahmenbedingungen und Restriktionen zur RFID-Nutzung sind im Einzelfall zu ergründen und in Abstimmung mit dem Konsumenten (bzw. in seinem Mitwissen und mit seiner Erlaubnis) umzusetzen [Müll04] [KnHa05]. Zukünftige Forschung sollte konkrete

Implementierungen und Integration weiterer Datenquellen (beispielsweise aus sozialen Netzwerken) untersuchen.

Literaturverzeichnis

[BaYY10] Bayraktar, A. et al.: Implementation of RFID technology for the differentiation of loyalty programs. J. Relatsh. Mark. 9, 1, 30–42 (2010).

[BeHP12] Becker, J. et al.: 1 Controlling und Herausforderungen im Handel. Handb. Handel. 867 (2012).

[BeSc04] Becker, J., Schütte, R.: Handelsinformationssysteme. MI Wirtschaftsbuch (2004).

[BeWi04] Becker, J., Winkelmann, A.: Warenwirtschaft im 21. Jahrhundert-Bedeutung von IT im Handel. IT-Lösungen Für Den Handel HMD. 235, 5–14 (2004).

[BeWM13] Benmessaoud, N. et al.: Network virtualization and cloud computing. (2013).

[ChDa97] Chaudhuri, S., Dayal, U.: An overview of data warehousing and OLAP technology. ACM Sigmod Rec. 26, 1, 65–74 (1997).

[CLSF05] Curbera, F. et al.: Web Services Platform Architecture: SOAP, WSDL, WS-Policy, WS-Addressing, WS-BPEL, WS-Reliable Messaging and More. Prentice Hall PTR (2005).

[Codd72] Codd, E.F.: Relational completeness of data base sublanguages. IBM Corporation (1972).

[Dave10] Davenport, T.: The New World of "Business Analytics."Int. Inst. Anal. (2010).

[Derh00] derhandel.de: Textilhandel: P&C überzeugt Testkäufer - Der Handel, http://www.derhandel.de/news/unternehmen/pages/Textilhandel-P%26C-ueberzeugt-Testkaeufer-1280.html.

[Dest13] Destatis: Volkswirtschaftliche Gesamtrechnungen Private Konsumausgaben und Verfügbares Einkommen. (2013).

[FHBB11] Friedland, A. et al.: NoSQL: Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Carl Hanser Verlag GmbH & CO. KG (2011).

[Fink02] Finkenzeller, K.: RFID-handbuch. Hanser München (2002).

[GlGC97] Gluchowski, P. et al.: Management Support Systeme. Springer (1997).

[Hand13] Handelsdaten.de: Aktuelle Statistiken zum Textileinzelhandel in Deutschland. Handelsdaten.de (2013).

[HeSt11] Heinrich, L.J., Stelzer, D.: Informationsmanagement: Grundlagen, Aufgaben, Methoden. Oldenbourg, R (2011).

[KeBL13] Kemper, H.-G. et al.: An Integrated Business Intelligence Framework. In: Rausch, P. et al. (eds.) Business Intelligence and Performance Management. pp. 13–26 Springer London (2013).

[KeHa06] Keck, M., Hahn, M.: Integration der Vertriebswege: Herausforderung im dynamischen Retail Banking. Springer DE (2006).

[KnHa05] Knyrim, R., Haidinger, V.: RFID-Chips und Datenschutz. Österr. Recht Wirtsch. 23, 1, 2–6 (2005).

[KoPe07] Kolbe, H., Peseschk, S.: Wirtschaftlichkeitsanalyse des RFID-Einsatzes in einem Handelsunternehmen der Bekleidungsbranche. GI Jahrestagung (1). pp. 99–102 (2007).

[Lang05] Langheinrich, M.: Die Privatsphäre im Ubiquitous Computing—Datenschutzaspekte der RFID-Technologie. Das Internet der Dinge. pp. 329–362 Springer (2005).
[LLSH11] LaValle, S. et al.: Big data, analytics and the path from insights to value. MIT Sloan Manag. Rev. 52, 2, 21–32 (2011).
[Mcbr11] McBrearty, R.: The Future of Retail Customer Loyalty RFID Enables Breakthrough Shopping Experiences. Cisco Systems (2011).
[MeGr09] Mell, P., Grance, T.: The NIST Definition of Cloud Computing, http://csrc.nist.gov/groups/SNS/cloud-computing/.
[MKSM13] Möhring, M. et al.: Public-Cloud-Angebote: Kostenorientierte Entscheidungskriterien für kleine und mittlere Unternehmen. Controlling. 25, 11, 619–624 (2013).
[Müll04] Müller, J.: Ist das Auslesen von RFID-Tags zulässig. Datenschutz Datensicherheit DuD. 28, 5, 215 (2004).
[NuSc09] Nurcan, S., Schmidt, R.: Service Oriented Enterprise-Architecture for enterprise engineering introduction. Enterprise Distributed Object Computing Conference Workshops, 2009. EDOCW 2009. 13th. pp. 247–253 (2009).
[Robe06] Roberts, C.M.: Radio frequency identification (RFID). Comput. Secur. 25, 1, 18–26 (2006).
[ScMö13] Schmidt, R., Möhring, M.: Strategic alignment of Cloud-based Architectures for Big Data. Proceedings of the 17th IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW). , Vancouver, Canada (2013).
[SMMP14] Schmidt, R. et al.: Big Data as Strategic Enabler - Insights from Central European Enterprises. In: Abramowicz, W., Kokkinaki, A. (Hrsg.). 17th International Conference on Business Information Systems (BIS) (in press). Springer LNBIP. pp. 50–60 Springer Berlin Heidelberg, Larnaca, Cyprus (2014).
[TeRe12] Termer, D.-W.-I.F., von Rennenkampff, D.-W.-I.A.: Agile it-anwendungslandschaften als strategische unternehmensressource. HMD Prax. Wirtsch. 49, 2, 24–33 (2012).
[Tull13] Tulloch, M.: Introducing Windows Azure for It Professionals. Microsoft Press, c2013., Redmond, Wash. (2013).
[Whit12] White, T.: Hadoop: The definitive guide. O'Reilly Media (2012).
[Wieh00] Wiehr, H.: Gerry Weber rollt RFID aus: Ab 2010 ein RFID-Chip in jedem Kleidungsstück - CIO.de, http://www.cio.de/retailit/bestpractice/2216267/.

Adaptive Digitale Enterprise Architekturen für Big Data und Cloud-Systeme

Alfred Zimmermann[1], Kurt Sandkuhl[2], Rainer Schmidt[3],
Dierk Jugel[1], Matthias Wisotzki[2], Michael Möhring[3]

[1]Hochschule Reutlingen, Fakultät Informatik, Reutlingen
[2]Universität Rostock, Fakultät Informatik und Elektrotechnik, Rostock
[3] Hochschule München, Fakultät Informatik und Mathematik, München

alfred.zimmermann@reutlingen-university.de
kurt.sandkuhl@uni-rostock.de
rainer.schmidt@hm.edu
dierk.jugel@reutlingen-university.de
matthias.wissotzki@uni-rostock.de
michael.moehring@htw-aalen.de

Abstract: Big Data und Cloud Systeme werden zunehmend von mobilen, benutzerzentrierten und agil veränderbaren Informationssystemen im Kontext von digitalen sozialen Netzwerken genutzt. Metaphern aus der Biologie für lebendige und selbstheilende Systeme und Umgebungen liefern die Basis für intelligente adaptive Informationssysteme und für zugehörige serviceorientierte digitale Unternehmensarchitekturen. Wir berichten über unsere Forschungsarbeiten über Strukturen und Mechanismen adaptiver digitaler Unternehmensarchitekturen für die Entwicklung und Evolution von serviceorientierten Ökosystemen und deren Technologien wie Big Data, Services & Cloud Computing, Web Services und Semantikunterstützung. Für unsere aktuellen Forschungsarbeiten nutzen wir praxisrelevante SmartLife-Szenarien für die Entwicklung, Wartung und Evolution zukunftsgerechter serviceorientierter Informationssysteme. Diese Systeme nutzen eine stark wachesende Zahl externer und interner Services und fokussieren auf die Besonderheiten der Weiterentwicklung der Informationssysteme für integrierte Big Data und Cloud Kontexte. Unser Forschungsansatz beschäftigt sich mit der systematischen und ganzheitlichen Modellbildung adaptiver digitaler Unternehmensarchitekturen - gemäß standardisierter Referenzmodelle und auf Standards aufsetzenden Referenzarchitekturen, die für besondere Einsatzszenarien auch bei kleineren Anwendungskontexten oder an neue Kontexte einfacher adaptiert werden können. Um Semantik-gestützte Analysen zur Entscheidungsunterstützung von System- und Unternehmensarchitekten zu ermöglichen, erweitern wir unser bisheriges Referenzmodell für IT-Unternehmensarchitekturen ESARC – Enterprise Services Architecture Reference Cube – um agile Mechanismen der Adaption und Konsistenzbehandlung sowie die zugehörigen Metamodelle und Ontologien für Digitale Enterprise Architekturen um neue Aspekte wie Big Data und Cloud Kontexte.

1 Einführung

Informationen, Daten und Wissen sind heute zentrale Bestandteile unserer täglichen Aktivitäten. Soziale Netzwerke, mobile Geräte wie Smart Devices und intelligente Autos zeigen Aktuelle Beispiele einer allgegenwärtigen informationsorientierten Vision, die wir in unserem Forschungsszenario SmartLife nennen. Nach [Hs04] nennen wir die Sammlung von adaptiven Service-orientierten Unternehmensarchitekturen zusammen mit den zugehörigen Business- und Software-Services für zugehörige Informationssysteme ein digitales Ökosystem. In unserem Verständnis der Vision zu digitalen Ökosystemen [Ti14] sind selbstoptimierende Systeme mit selbstheilenden Eigenschaften von verteilten Service-orientierten Umgebungen mit anpassungsfähigen Service-orientierten Unternehmensarchitekturen fest verankert. Die Entwicklung solcher Anwendungen integriert Web-Services, Cloud Computing und Big Data Management, sowie spezifische Frameworks und Methoden der Semantik-Unterstützung für neue System- und Software-Architekturen [Es13] und [Am11].

Betrachten wir zum Beispiel den Fahrer eines künftigen intelligenten Autos. Ein Warnlicht auf dem Armaturenbrett zeigt während einer Fahrt auf der Autobahn eine Ausnahmesituation an. Er wählt "untersuchen" und die SmartLife-Anwendung des Autos sucht den Fehlercode und zeigt an, welche Fahrzeugkomponente die Warnung ausgelöst hat. Das System sucht dann mehrere Online-Datenbanken ab, welche Informationen der Autohersteller als auch die Service-Historien anderer Fahrzeuge zum Ausnahmefall der signalisierten Komponente analysiert. Die SmartLife-Anwendung fasst die Ergebnisse zusammen und empfiehlt einen sofortigen Austausch der Komponente, sowie mehrere Anbieter und überprüft deren Inventar. Der Fahrer hat die Wahl, einen Servicetermin in der Werkstatt gleich zu vereinbaren oder diese Komponente in einem Online-Shop zu erwerben und selber auszutauschen. Dieses Anwendungsszenario ist ein Beispiel für eine servicebasierte und semantisch erweiterte Anwendung für SmartLife, die auf neue adaptive Digitale Enterprise Architekturen, Graph-Analytik, Big Data Management, Services und Cloud-Computing aufsetzt.

Die komplex miteinander verwobenen technologischen und wirtschaftlichen Dimensionen von Systemen zur Unterstützung der Mobilität, von sozialen Netzwerken, von fallbasiertem Geschäftsprozessmanagement und Dienstleistungen, dem Rechnen mit großen Daten in Cloud-Umgebungen wirken sich unmittelbar auf anpassungsfähige Service-orientierte System- und Unternehmensarchitekturen für SmartLife-Ökosysteme aus. Während die fachliche Perspektive stärker auf intelligent bewertbare und besser steuerbare Ansätze und Strukturen der Geschäftstransformation setzt, ist die technologische Perspektive moderner Informationssysteme vielfältiger. Wichtig ist nicht nur die Fähigkeit leistungsgerechte, prozessintegrierte und serviceorientierte Informationssysteme effizient zu entwickeln, sondern diese Fähigkeit auch durch kostenoptimierte Betriebsperspektiven mit Cloud-Services und Big Data zu ergänzen. Leistungsstarke smarte Unternehmen erweitern gegenwärtig ihre Fähigkeiten, um systematisch ihr Business-Operating-Modell [Ro06] mittels digitaler System- und Unternehmensarchitekturen als Teil der architektonischen IT-Governance [We04] im Sinn optimierter Kosten und Leistungen systematisch zu bewältigen.

In unserer aktuellen Forschung erweitern wir unsere erste Version der ESARC - Enterprise Services Architecture Reference Cube [Zi12], [Zi13b] hin zu einer adaptiven Digitalen Enterprise Architektur für Big Data und Cloud Kontexte. ESARC ist eine EA-

Referenzarchitektur und formuliert einen ganzheitlichen Klassifikationsrahmen (Framework) für die Analyse (Assessments), Bewertung, Optimierung, Neugestaltung und Transformation von Service-orientierten Unternehmensarchitekturen, der zugehörigen Systeme, der Fachlichkeit und Technologien. Unternehmensarchitekturen für Services & Cloud Computing definieren heute ein Einsatzfeld, das sich auf die Erschließung von Potenzialen und das Management von Veränderungen fokussiert, die sich durch die zunehmende Verbreitung von Big Data, Cloud- und Service-basierten Architekturen ergeben. Hier besteht verstärkt der Bedarf der Koordination zwischen Strategieanforderungen, neuen Geschäftsanforderungen, technologischen Veränderungen und einer Landschaft von Projekten, Bebauungsplänen und Kostenoptimierungen.

Neue komplexere Produkte und Dienstleistungen erfordern eine effiziente und leistungsstarke EA-Funktion. Dafür passend ausgewählte und parametrierte EAM-Tools sollten durch ergänzende methodische Instrumente und ein geeignetes Wissen über Digitale Enterprise Architekturen bei allen fachlich und technologisch aufgestellten Beteiligten führen. Eine EA-Referenzarchitektur kann als Grundlage dieser neuen Positionierung von EAM in Unternehmen, aber auch als Handlungsgrundlage für die Weiterentwicklung von EAM-Werkzeugen und Methoden bei Herstellern und Partnern betrachtet werden. Eine derartige EA Referenzarchitektur muss konsequent auf die neuen Möglichkeiten von Big Data, Analytics und Architekturen für Services & Cloud Computing angepasst werden. Im vorliegenden Aufsatz wird der durch theoretische Vorarbeiten und praktische Szenarien erprobte Weg zu einer adaptiven Service-orientierten EA-Referenzarchitektur unter Berücksichtigung von Services & Cloud Computing beschrieben.

Unser Hauptbeitrag umfasst einen erweiterten Ansatz für EAM (Enterprise Architecture Management) durch die systematische Zusammensetzung und Integration von Architektur-Metamodellen, Ontologien, Views und Viewpoints für adaptive Digitale Enterprise Architekturen - in enger Ausrichtung auf Services & Cloud Computing und Big Data-Architekturen. Wir berichten über Erweiterungen unserer Forschungsarbeiten zu ESARC [Zi12], [Zi13b], [Es13b] und zur Semantik-gestützten Software-Wartung [Wi12], [Es13a] Evolution von Service-orientierten Anwendungen in Verbindung mit neuen Entwicklungen von EAM [Jo14] und Plattform- Ökosystemen [Ti14]. Wir integrieren gegenwärtig Perspektiven der adaptiven Software-Modelle [Sh11], [He06], [Sa06], [Yo02] und [Fe10] für flexibel veränderbare Unternehmensarchitekturen auch für KMUs und für die Unterstützung von Geschäftstransformationsprozessen durch EAM und erweitern diese Flexibilität um Aspekte der Datenkonsistenz [Bt11] und Big Data [Br13]. Zusätzlich zu dem Stand der Wissenschaft integrieren wir derzeit EAM-Metamodelle von Industriepartnern und aus EA-Tools.

Das folgende Kapitel 2 beschreibt unseren Ansatz für eine EA-Referenzarchitektur für SmartLife-Ökosysteme auf Basis des erweiterten ESARC. Kapitel 3 skizziert die synergetische Basis der EA-Modellintegration, grundlegende Adaptionsmechanismen und die Konsistenzbehandlung von EA-Modellen. Kapitel 4 beschäftigt sich mit Metamodellen und Ontologien als Grundlage einer entscheidungsunterstützenden formalen Semantik für die Analyse, Bewertung und Navigation zwischen den EA-Konzepten. Schließlich fassen wir im Kapitel 5 unsere Schlussfolgerungen, wesentliche EA-Trends und Pläne für die Zukunft zusammen.

2 Referenzarchitektur

In Bereichen, in denen Flexibilität und Agilität des Geschäfts flexibel unterstützt werden sollen, ist Services Computing [Zh07] der Ansatz der Wahl, um Fähigkeiten von verteilten Systemen zu organisieren und agil zu nutzen. Innovationsorientierte Unternehmen haben in den letzten Jahren Service-orientierten Architekturen eingeführt, um die Lücke zwischen Geschäft und IT zu schließen und gleichzeitig neue Potentiale wie Cloud Computing, Big Data, Analytics u. a. strategisch zu erschließen. Die Vorteile von Service-orientierten Architekturen [Ma06] wie Flexibilität, Prozessorientierung, Time-to-Market und agile Innovation sind prinzipiell anerkannt, jedoch leider noch nicht ganz gut beherrscht. Digitale Unternehmensarchitekturen oder klassisch EAM – Enterprise Architecture Management - adressieren die Gesamtsicht der Architekturen aus fachlicher und technologischer Sicht.

Der Kern unserer erweiterten EA-Referenzarchitektur ist ESARC - Enterprise Services Architecture Reference Cube [Zi12], [Zi13a], [Zi13b] (siehe Abbildung 1) präzisiert bestehende architektonische Standards und Frameworks für EAM - Enterprise Architecture Management [Ro06], [TG11] und [AM12] und erweitert diese Architekturstandards für Services Computing, Cloud Computing und Big Data. ESARC ist unser eigens entwickeltes Architektur-Referenzmodell für ein Service-orientiertes EAM, das einen ganzheitlichen Klassifikationsrahmen mit acht integrierten Architekturbereichen definiert. ESARC abstrahiert von einem konkreten Business-Szenario oder spezifischen Technologien, ist aber offen für konkrete architektonische Ausprägungen ausgelegt.

Das OASIS-Referenzmodell für Service Orientierte Architekturen [MK06] definiert einen abstrakten Rahmen für unsere Vorstellung von Referenzarchitekturen [Ba13], [Es08] und [OG11a]. Referenzmodelle sind konzeptionelle Modelle der funktionalen Zerlegung eines Systems in Modellelemente zusammen mit den Datenflüssen zwischen diesen Elementen. Das Referenzmodell für Service-orientierte Architektur von OASIS [MK06] definiert grundlegende generische Elemente und ihre Beziehungen einer Service-orientierten Architektur. Dieses Referenzmodell ist kein Standard, bietet aber eine gemeinsame Semantik für spezifischere Referenzarchitekturen. Referenzarchitekturen, in [Es08] und [OG11a], sind spezialisierte Modelle eines Referenzmodells. Referenzarchitekturen entstehen durch Abbildung von Referenzmodellen zu Software-Komponenten, die kooperativ die Funktionalität der Referenzmodelle implementieren. Oftmals wird diese Abbildung von Referenzmodellen zu Referenzarchitekturen durch Architekturmuster unterstützt. Referenzarchitekturen wenden darüber hinaus eine Sammlung von geeignet kombinierten Architekturstilen wie Client/Server, REST, Web-Services, u. a. an.

The Open Group Architecture Framework [TG11] und ArchiMate [AM12] liefern wesentliche Standards für erweiterte Service-orientierten Domänen (Abbildung 1) von ESARC [Zi13a], [Zi13b] wie: Architecture Governance, Architecture Management, Business & Information Architecture, Technology Architecture, Operation Architecture, Cloud Services Architecture und Security Architecture. ESARC bietet mit seinem Klassifikationsframework, Metamodellen und EA-Patterns eine kohärente Hilfe für zyklische Assessments, EA-Analytik, Optimierung und Neuausrichtung der multiperspektivischen Architekturqualität von adaptiven Digitalen Enterprise Architekturen.

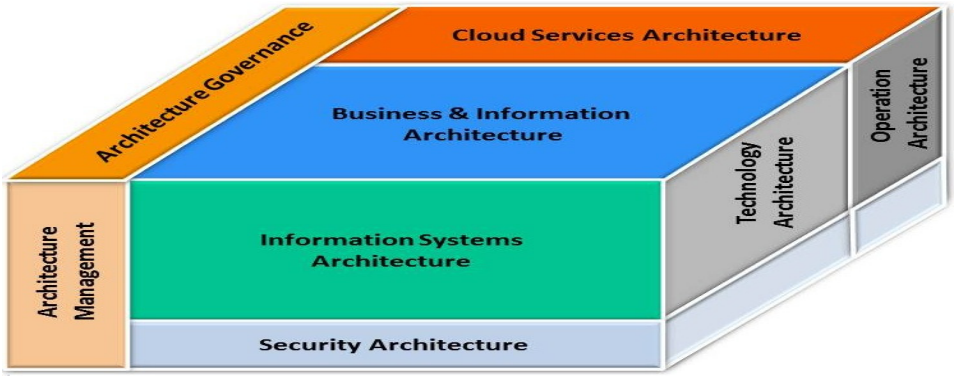


Abbildung 1: ESARC – Enterprise Services Architecture Reference Cube

Architecture Governance [Zi13a], [Zi13b] formuliert den Architektur-Governance-Zyklus [Ro06]. Er setzt den abstrakten Rahmen für konkrete Governance-Aktivitäten innerhalb des Unternehmens bzw. für die Entwicklung von Produktlinien durch grundlegende Management-Aktivitäten: Planen, Definieren, Umsetzen, Messen, Kontrollieren. Das zweite Ziel ist es, Compliance-Regeln für die Architektur-Governance zur Einhaltung interner und externer Standards zu setzen und zu überwachen. Unternehmensarchitekten und Software-Architekten verbindet ein anspruchsvoller Pfad, der von der Geschäfts- und IT-Strategie ausgeht und bis zur Definition und Überwachung der Architekturlandschaft von miteinander verknüpften Geschäftsbereichen, Produkten, Dienstleistungen, Geschäftsprozessen, Anwendungssystemen und Technologien führt. Architecture Governance setzt Regeln für die Übertragung von Verantwortung an handelnde Akteure, legt Strukturen und Verfahren für das Architecture Governance Board fest und definiert wesentliche Regeln für die Kommunikation. Folgende Begriffe für Architecture Governance wurden im Modell festgelegt: Service-Strategie, Lebenszyklus-Management für alle markanten Zustände von Architektur-Artefakten, Service-Sicherheit, Service-Tests und Überwachung, Serviceverträge, Service-Registries, Wiederverwendung von Services, Service-Verantwortlichkeiten, Service-Definition und Versionsverwaltung.

Die „Business and Information Reference Architecture – BIRA“ [Zi13a], [Zi13b] liefert die eindeutige Informationsquelle und einen umfassenden Wissensspeicher der fachlichen Konzepte, aus denen konkrete Unternehmensinitiativen und Systeme entwickeln werden. Operative fachliche und technologische Anforderungen müssen natürlicherweise noch durch jedes Umsetzungsprojekt ergänzt werden. Dieses Wissen ist modellbasiert und definiert ein integriertes Enterprise Business Modell, das auch Organisationsmodelle und Geschäftsprozesse umfasst. Die BIRA eröffnet eine Verbindung (inter-architectural dependency) zur Software-Architektur der Informationssysteme, zur Technology Architecture der IT-Infrastrukturen, zur Betriebsarchitektur, zur Cloud-Services Architecture und zur Sicherheitsarchitektur. Die BIRA ermöglicht ein abgestimmtes Business-IT Alignment durch mehrfach integrierte Modelle (intra-architectural dependency) und Angaben zur Business- und Informationsstrategie, beteiligte Organisationen und Einheiten, Geschäftsrollen, wesentliche Anforderungen ans Geschäft und für zugehörige Informationssysteme, wichtige Geschäftsprozesse, Geschäftsregeln, Geschäftsprodukte, Dienstleistungen und die damit verbundenen Steuerungs- und Kontrollinformationen.

Die „Information Systems Reference Architecture – ISRA“ [Zi13a], [Zi13b] ist die Software-Referenzarchitektur der Anwendung und enthält wesentliche anwendungsspezifische Typen von Services, die durch ein Schichtenmodell von aufeinander aufbauenden Schichten geordnet wurden. Die Kernfunktionalität der Domain-Services wurde mit Interaktionsservices und Geschäftsprozessservices der Kundenorganisation verbunden. Der Aufbau der ISRA (in Abbildung 2) berücksichtigt aktuelle Standards für Service-orientierte Referenzmodelle [MK06] und Referenzarchitekturen [Es08] und [AM12]. Um Big Data und Cloud Computing Technologien zu unterstützen, haben wir in Abbildung 2 einen einheitlichen Satz von 14 aufeinander aufbauenden Servicetypen kategorisiert.

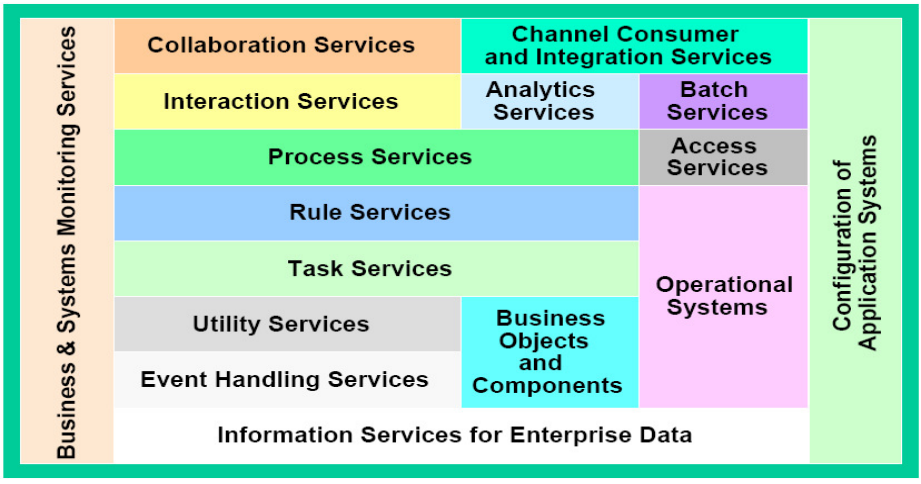


Abbildung 2: ESARC – Information Systems Reference Architecture

Informationsdienste für Unternehmensdaten [Zi13a], [Zi13b] sind datenzentrische elementare Komponenten, die den Zugang zu persistenten Geschäftsobjekten in der Regel über Datenbanken und andere Speicher realisieren. In der Nähe dieser Informationsdienste befinden sich Services zur Verwaltung der Datenkontexte, wie diese durch Grundmechanismen der Technologiearchitektur ermöglicht werden, wie beispielsweise Fehlerbehandlungsmechanismen, Transaktionsmechanismen für kurze und lange evtl. verteilte Transaktionen, Kompensationsdienste sowie weitere Mechanismen zur annähernden oder exakten Konsistenzbehandlung.

Process Services [Zi13a], [Zi13b] sind lang laufende Services, die Task Services und Information Services in Workflows zusammenbauen (orchestrieren), um die prozedurale Logik der Geschäftsprozesse zu implementieren. Prozessdienste können Prozessregeldienste (Process Rule Services) aktivieren, um den Teil der unstabilen (d. h. veränderlichen) kausalen Entscheidungslogiken von Gateways auszuklammern und damit einfacher veränderbare (agile) Geschäftsprozesse zu ermöglichen. Prozessdienste werden von Interaktionsdiensten oder spezifischen Diagnostik-Services oder Monitoring-Services aktiviert. Prozess-Services verwalten oft verteilten Daten- und Anwendungszustände indirekt durch die Aktivierung von Task Services und/oder Informationsdiensten. Wenn Prozessdienste in der interaktiven Workflows teilnehmen, müssen sie lange Transaktionen unterstützen und eventuelle Ausnahmen / Fehler durch Kompensationslogiken behandeln.

Big Data [Be13] und [Sc14] ist ein neueres Verarbeitungskonzept, um Sammlungen von sehr großen Datenmengen unterschiedlicher Herkunft und Ausgangsstruktur in einer hohen Geschwindigkeit zu verarbeiten. Vor allem im Bereich der Technologieentwicklung und IT sowie im Marketing und Vertrieb liegen demnach die größten Nutzenpotenziale. Zur Analyse großer Datenmengen unterschiedlichster Struktur im Rahmen von Big Data eignen sich traditionelle Verarbeitungswerkzeuge kaum. SmartLife Anwendungen erfordern spezielle Umgebungen für Big Data. SmartLife Anwendungen, wie das Szenario in diesem Papier zeigte, werden typischerweise mehrere Online-Datenbanken gleichzeitig abfragen. Um solche Anfragen zu behandeln werden dezentrale Parallelarchitekturen, die Daten über mehrere Verarbeitungseinheiten verteilen, benötigt. Big Data Bearbeitungsfunktionen wie MapReduce [De04] werden benötigt, um diese Art von Anfragen und Daten für SmartLife -Anwendungen parallel zu bearbeiten. MapReduce, welches auf dem Hadoop-Framework verwendet wird, zerstückelt die Abfragen auf kleine Segmente und verteilt sie auf parallelen Knoten zur Bearbeitung. Die Ergebnisse werden dann von den parallelen Knoten gesammelt, verdichtet und weiterbearbeitet, z. B. analysiert.

Cloud-Architekturen (Integrationskomponenten in Abbildung 3) befinden sind noch in Entwicklung und sind noch nicht so weit, ihr volles Potenzial bei der Integration von Enterprise Architekturen mit Services Computing und Cloud Computing zu erreichen. Die „Cloud Services Reference Architecture“ stützt sich auf eine Referenzmodell-basierte Synthese von aktuellen Standards und Referenzarchitekturen aus [Li11] , [Be11] und [CS09]. Die heutige Entwicklung von Technologien und Standards für Cloud Computing wächst sehr schnell und bietet eine zunehmend gut standardisierte Basis für Cloud-Produkte und neue Service-Angebote.

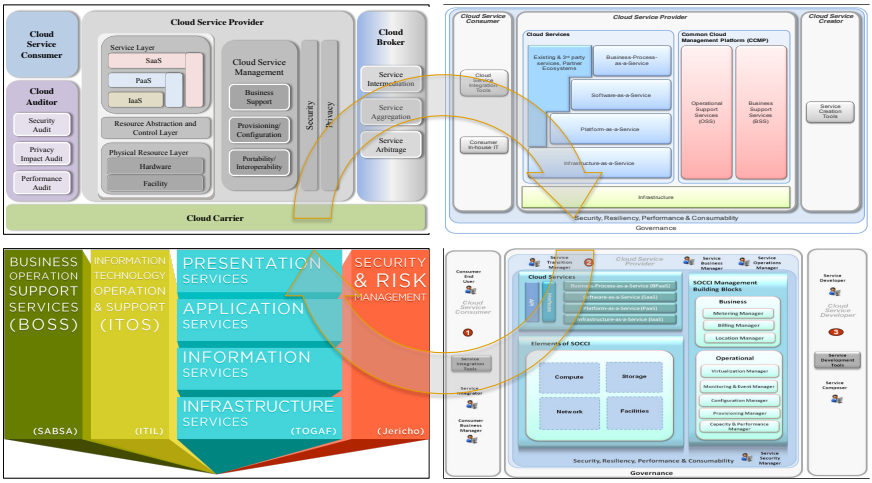


Abbildung 3: Integrating Cloud Computing Reference Architectures with SOCCI

Die NIST Cloud Computing Referenzarchitektur [Li11] definiert den Standard und damit das konzeptuelle Grundmodell für Cloud Computing aus der Perspektive folgender Rollen: Cloud Consumer, Cloud Provider, Cloud Auditor, und Cloud Broker. Der NIST-Standard definiert ferner folgende Bereitstellungsmodelle: Private Cloud, Community Cloud, Public Cloud, und Hybrid Cloud. Cloud Computing definiert wesentliche

Eigenschaften einer massiv parallelen virtualisierten Serverlandschaft: On-Demand-Self-Services, Broad Network Access, Resource Pooling, Rapid Elasticity, und Measured Services. Der grundlegende Teil der NIST-Referenzarchitektur formuliert folgende Cloud-Service-Modelle: IaaS - Infrastructure as a Service, PaaS - Platform as a Service , und SaaS - Software as a Service.

Einige Erweiterungen des NIST-Standards [Be11] liefern praktische Ergänzungen für die direktere Unterstützung moderner Business-Architekturen durch BPaaS - Business Process as a Service und ermöglichen durch ein durchgehendes Service-Konzept eine direkte Unterstützung von Service-orientierten Unternehmensarchitekturen. Die IBM Cloud Computing Reference Architecture erweitert die Standardisierung von NIST durch Best-Of-Industry Branchenwissen und Cloud-Produktspezifikationen auf Basis eigener Technologie-Stacks, Middleware, sowie Service-orientierte Entwicklungs- und Laufzeitplattformen [Be11]. Die IBM Cloud Computing -Referenzarchitektur [Be11] hat den grundlegenden NIST-Standard für Cloud Computing [Li11] mit der SOA-Referenzarchitektur [OG11a] der Open Group integriert: Alle Cloud- Services sind demnach SOA-Services, aber nicht alle SOA-Services sind auch Cloud-Services. Die Sicherheitsergänzungen aus den CSA Security Guidelines for Critical Areas of Focus in Cloud Computing [CS09] definiert eine auf Jericho-Sicherheitsmechanismen fokussierte Service-orientierte Referenzarchitektur für Cloud Computing und integriert die Managementperspektiven aus Standards wie ITIL und TOGAF [TG11].

Das SOCCI - Service-Oriented Cloud Computing Framework [OG11b] umfasst eine grundlegende Zusammenstellung für eine integrierte Reihe von Infrastrukturkomponenten IaaS für Cloud Computing. Im Grunde ist es die Synergie von Service-orientierten Architekturen und Cloud Computing Architekturen durch einen konsequenten As-a-Service - Mechanismus für alle Arten von Cloud-Services. Die grundlegenden Eigenschaften eines Service-orientierten Infrastruktur (SOI) sind: Business-Driven Infrastructure On-Demand, Operational Transparency, Service Measurement, und Consumer Provider Model. Das SOCCI - Service-Oriented Cloud Computing Framework ist die Erweiterung der in die SOA-Referenzarchitektur [OG11a] abgebildeten Service-orientierten Infrastruktur (SOI). Die SOI-Framework ist die Schicht über der Basisinfrastruktur und definiert wesentliche Elemente der SOCCI: Compute, Network, Storage, Facilities. SOCCI erweitert diese Grundelemente durch Basisbausteine (Building Blocks) zur flexiblen Unterstützung von Management und Betrieb von Cloud Infrastrukturen.

3 Adaptionsmechanismen

Die Integrationsmethode ESAMI – Enterprise Services Architecture Metamodel Integration – [Zi13a] wird gegenwärtig durch laufzeitaktive Adaptionsmechanismen erweitert. Um adaptive Enterprise Architekturen zu unterstützen, wurden in [Jo14] agil steuerbare Analyse-Metamodelle und zugehörige Architektur-Viewpoints eingeführt und auf folgende Qualitätskriterien zur Unterstützung der Agilität und Adaption von Architekturen ausgerichtet: Anpassbarkeit der Anwendungen, Genauigkeit der Daten, Verwendbarkeit der Anwendungssysteme, Service-Verfügbarkeit, Interoperabilität, Kosten und Nutzen. Plattform Ökosysteme werden in [Ti14] systematisch untersucht und mit den wichtigsten strategischen Treibern für die Systementwicklung und die Evolution der Ökosysteme verknüpft. Hauptkonzepte adaptiver Enterprise Architekturen für digitale

Ökosysteme basieren auf Mikroarchitekturen, die den Kontext von Internet-Systemen erfordern. Bevorzugte Mechanismen der Modularisierung von Systemen stützen sich auf Mechanismen zur Entkopplung von Subsystemen und auf die von Standardisierung von Schnittstellen. Architektur Governance-Modelle zeigen den Weg, um anpassungsfähige Ökosysteme zu ermöglichen und um die Evolution von Plattform und Systemen zusammenzuführen. Enterprise Architekturen für KMUs [Bl13] fokussieren auf sparsame und effiziente adaptive Funktionen der EA wie: Agilität, um die Ökosysteme an sich verändernde Umgebungen rasch und flexibel anzupassen, Strategie Planung und Entscheidungsunterstützung, kontinuierliches Ausrichtung des Geschäfts auf IT und umgekehrt (Business / IT-Alignment), Komplexitätsmanagement, Integration von Geschäftsprozessen, Vereinheitlichung der Daten und Datenintegration, Verknüpfung der Daten mit externen Partnern, sowie die kontinuierliche Wertsteigerung der IT. Der Schwerpunkt des ADaPPT [Sh11] EA-Ansatzes ist in erster Linie auf vier strategische EA Domain Elementen ausgerichtet: Menschen, Prozesse, Daten, und Technologien.

Um ein adäquates IT / Business Alignment für eine leistungsstarke Digitale Enterprise Architektur zu erreichen, sind drei Qualitätsperspektiven gemäß unserer Forschung wesentlich und daher durch geeignete Qualitätskriterien zu differenzieren: (i) IT-Systemeigenschaften: Leistung, Interoperabilität, Verfügbarkeit, Benutzerfreundlichkeit, Genauigkeit, Wartbarkeit und Eignung; (ii) Geschäftseigenschaften: Flexibilität, Effizienz, Wirksamkeit, Integration und Koordination, Entscheidungsunterstützung, Steuerung und Folgeaktivitäten, und Organisationskultur; und schließlich (iii) Governance-Qualitäten: Planung und Organisation, Akquisition und Implementation, Bereitstellung und Unterstützung, Monitoring und Evaluierung.

Aus der Sicht der Modellierung adaptiver Metamodelle für adaptive Digitale Enterprise Architekturen haben wir zusätzliche Anregungen aus dem „Adaptive Object Model Architectural Style“, von [Yo02] übernommen werden, um die Flexibilität der Systeme zu ermöglichen und Systemkonfiguration zur Laufzeit anpassen zu können. Dafür werden Geschäftsregeln explizit dargestellt und außerhalb des Programmcodes - und damit einfach veränderbar - abgelegt. Das Objektmodell im festen Code besteht lediglich aus einem Interpreter des Objektmodells der Nutzer. Das konfigurierbare Wissen ist außerhalb des Programms in den Regeln gespeichert. Dieser Semantik-gestützte Ansatz ermöglicht, dass Architekten und Entwickler – in ausbaubarere Form sogar Programme selbst - das adaptive Objektmodell flexibel und einfach zur Laufzeit anpassen können. Der Ansatz der „Core Patterns of Object-Oriented Meta-Architectures“ aus [Fe10] ermöglicht reflektive Architekturen, die in der Lage sind die Strukturen der Programme zur Laufzeit zu inspizieren und diese Strukturen und das Verhalten der Programme zur Laufzeit dynamisch anzupassen.

4 Schlussfolgerungen und Trends

Wir haben in diesem Beitrag über laufende Arbeiten zur Konzeption neuartiger adaptiver Digitaler Enterprise Architekturen für Big Data und Cloud Systeme berichtet. Dafür haben wir auf Basis unserer Vorarbeiten und des State of Art die Referenzarchitektur ESARC um adaptive Elemente erweitert, um insbesondere Transformationsprozesse für neue Geschäftsmodelle und Informationssysteme sowohl bei großen wie auch bei kleinsten Unternehmen agil durch EAM zu unterstützen. Dafür haben wir unsere Metamodell-basierte Integrationsmethode konsequent auf neue

Architekturstandards sowie auf den State of Art & Practice ausgerichtet und damit eine einfach zu handhabende Basis für weitere EA-Modellintegrationen geschaffen.

In unserer aktuellen Forschung legen wir ein großes Augenmerk auf die schnelle Anpassbarkeit der EA an neue Kontexte und erforschen daher wesentliche Adaptionsmechanismen für EA. Adaption bedeutet immer potentielle Inkonsistenz der Modelle und der EA-Information. Daher adressieren wir ebenfalls neue Aspekte der Konsistenzbehandlung und setzen hierfür Erfahrungen aus der Big Data Analytics und aus Vorarbeiten über Architekturen für Services & Cloud Computing ein. Der Weg zu intelligenten Systemen liefern uns Metamodelle und zugehörige Ontologien, um eine geeignete Form der EA-Wissensrepräsentation für neuartige und flexiblere Digitale Enterprise Architekturen zu ermöglichen.

Aus den skizzierten Lösungselementen folgen interessante weiterführende Themen für Service-orientierte EA-Referenzarchitekturen: Ausrichtung auf strategierelevante Kosten- und Optimierungsthemen, Unterstützung organisatorischer Neuordnungen und systematische Veränderungsprozesse, EAM für Innovationsmanagement und Risikomanagement, erweiterte Referenzmodelle und EA-Viewpoint-Models, Wissensrepräsentation und Semantikverarbeitung für EA, Visualisierung, Interaktion, Simulation und neue, schlanke und effiziente Verfahren und Mechanismen der Gestaltung, Nutzung und Entscheidungsunterstützung in EA-Kollaborationsprozessen.

Literatur

[Am11] Allemang, D.; Hendler, J.: Semantic Web for the Working Ontologist – Effective Modeling in RDFS and OWL. Morgan Kaufmann, 2011.

[AM12] ArchiMate 2.0 Specification. Open Group Standard, 2012.

[Ba13] Bass, C.; Clements, P.; Kazman, R.: Software Architecture in Practice. Addison Wesley, 2013.

[Be11] Behrendt, M.; Glaser, B.; Kopp, P.; Diekmann, R.; Breiter, G.; Pappe, S.; Kreger, H.; Arsanjani A.: Introduction and Architecture Overview – IBM Cloud Computing Reference Architecture 2.0. IBM 2011.

[Bl13] Van Belle, J-P.; Giqwa, L.: The Potential of Enterprise Architectureal Thinking for Small Enterprises: An Exploratory South African Study. International Journal of Advanced Research in Business, Vol. 1, No. 3, pp. 22-29, 2013-14.

[Br13] J. J. Berman: Principles of Big Data. Morgan Kaufmann, 2013.

[Bt11] Bertossi, L.: Database Repairing and Consistent Query Answering. Morgan & Claypool Publishers, 2011.

[CS09] CSA Cloud Security Alliance: Security Guidance for Critical Areas of Focus in Cloud Computing V2.1. 2009.

[De04] Dean, J.; Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. OSDI, pp. 1-13. 2004.

[DF09] DoDAF Architecture Framework. Version 2.0, Volume 1, Department of Defense USA, 28 May 2009.

[Em09] Emery, D. E.; Hilliard, R.: Every Architecture Description needs a Framework - Expressing Architecture Frameworks Using ISO/IEC 42010. IEEE/IFIP WICSA/ECSA 2009, pp. 31-40, 2009.

[Es08] Estefan, J. A.; Laskey, K.; McCabe, F. G.; Thornton, D.: OASIS Reference Architecture for Service Oriented Architecture. Version 1.0, OASIS Public Review Draft 1, 23 April, 2008.

[Es13a] El-Sheikh, E.; Reichherzer, T.; White, L.; Wilde, N.; Coffey, J.; Bagui, S.; Goehring, G.; Baskin, A.: Towards Enhanced Program Comprehension for Service Oriented Architecture (SOA) Systems. Journal of Software Engineering and Applications, Vol. 6 No. 9, 2013, pp. 435-445.

[Es13b] El-Sheikh, E.; Bagui, S.; Firesmith, D. G.; Petrov, I.; Wilde, N.; Zimmermann, A.: Towards Semantic-Supported SmartLife System Architectures for Big Data Services in the Cloud. Proceedings of the Fifth International Conference on Advanced Service Computing - Service Computation 2013, May 27 – June 1, Valencia, Spain, pp. 59-64, 2013.

[ES14] Essential Architecture Project. http://www.enterprise-architecture.org, last access: May, 9th, 2014.

[Fe10] Ferreira, H. S.; Correia, F. F.; Yoder, J.; Aguiar, A.: Core Patterns of Object-Oriented Meta-Architecture. ACM-PLoP, October 16-18, 2010, Reno/Tahoe, Nevada, USA, 2010.

[Go13] Goehring, G.; Reichherzer, T.; El-Sheikh, E.; Snider, D.; Wilde, N.; Bagui, S.; Coffey, J.; White, L.: A Knowledge-Based System Approach for Extracting Abstractions from Service Oriented Architecture Artifacts. IJARAI International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.3, pp. 44-52, 2013.

[He06] Hendrickx, W.; Gorissen, D.; Dhaene, T.: GRID Enabled Sequential Design and Adaptive Metamodeling. IEEE Winter Simulation Conference, pp. 872-881, 2006.

[Hs04] Heistacher, T.; Kurz, T.; Masuch, C.; Ferronato, P.; Vidal, M.; Corallo, A.; Briscoe, G.; Dini, P.: Pervasive Service Architecture for a Digital Business Ecosystem. ECOOP 2004, Oslo, Norway, June 14, 2004.

[Jo14] Johnson, P.; Lagerström, R.; Ekstedt, M.; Österlind, M.: IT Management with Enterprise Architecture. KTH Stockholm, Sweden, 2014.

[Li11] Liu, F.; Tong, J.; Mao, J.; Bohn, R.; Messina, J.; Badger, L.; Leaf, D.: NIST Cloud Computing Reference Architecture. US Dept. of Commerce NIST SP 500-292, 2011.

[Ma06] Marks, E. A.; Bell, M.: Service Oriented Architecture (SOA): a planning and implementation guide for business and technology. John Wiley & Sons, 2006.

[MF05] MODAF MOD Architectural Framework, Executive Summary, Version 1.0, 31 August, Ministry of Defense UK, 2005.

[MK06] MacKenzie, C. M.; Laskey, K.; McCabe, F.; Brown, P. F.; Metz, R.: OASIS Reference Model for Service Oriented Architecture 1.0. OASIS Standard, 12 October 2006.

[Mo13] Morgan, T. P.: Intel takes on all Hadoop disties to rule big data munching. The Register 2013. http://www.theregister.co.uk/2013/02/26/intel_rolls_up_hadoop_distribution/ last access: May, 9th, 2014.

[NF07] NATO Architecture Framework. Version 3. 2007.

[NG01] Noy, N. F.; McGuiness, D. L.: Ontology Development 101: A Guide to Creating Your First Ontology. Stanford University, USA, 2001.

[OG10] Open Group: Service-Oriented Architecture Ontology. The Open Group, 2006.

[OG11a] The Open Group: SOA Reference Architecture. Technical Standard, 2011, https://www2.opengroup.org/ogsys/jsp/publications/PublicationDetails.jsp?catalogno=c119, last access: May 9th, 2014.

[OG11b] Open Group Standard: Service-Oriented Cloud Computing Infrastructure (SOCCI) Framework. The Open Group 2011.

[Ro06] Ross, W.; Weill, P.; Robertson, D.: Enterprise Architecture as Strategy – Creating a Foundation for Business Execution. Harvard Business School Press, 2006.

[Sa06] Saat, J.; Franke, U.; Lagerström, R.; Ekstedt, M.: Enterprise Architecture Meta Models for IT/Business Alignment Situations. EDOC Conference 2010, Vitoria, Brazil, 2010.

[Sc14] Schmidt, R.; Möhring, M.; Maier, S.; Pietsch, J.; Härting, R. C.: Big Data as Strategic Enabler - Insights from Central European Enterprises. In: Abramowicz, W., Kokkinaki, A. (Hrsg.). 17th International Conference on Business Information Systems (BIS). Springer LNBIP. Larnaca, Cyprus. (forthcoming), 2014.

[Sh11] Shah, H.; Godler, P.: ADaPPT: Enterprise Architecture Thinking for Information Systems Development. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011.

[TG11] TOGAF The Open Group Architecture Framework. Version 9.1, The Open Group, 2011.

[Ti14] A. Tiwana, A.: Platform Ecosystems – Aligning Architecture Governance, and Strategy. Morgan Kaufmann, 2014.

[We04] Weill, P.; Ross, J.W.: IT Governance. Harvard Business School Press, 2004.

[Wi12] Wilde, N.; Coffey, J.; Reichherzer, T.; White, L.: Open SOALab: Case Study Artifacts for SOA Research and Education. Principles of Engineering Service-Oriented Systems, PESOS 2012, Zurich, Switzerland, pp. 59-60, June 4, 2012.

[Yo02] Yoder, J. W.; Johnson, R.: The Adaptive Object-Model Architectural Style. IFIP-WICSA Conference 2002, Kluwer B.V. pp. 3-27, 2002.

[Zh07] Zhao, J. L.; Tanniru, M.; Zhang, L. J.: Services computing as the foundation of enterprise agility - Overview of recent advances and introduction to the special issue. Information Systems Frontiers 9.1, 2007 pp. 1-8, 2007.

[Zi12] Zimmermann, A.; Zimmermann, G.: Enterprise Architecture Ontology for Services Computing. In: Proc. SERVICE COMPUTATION 2012, Nice – France – July 22-27, 2012, ISBN 978-1-61208-215-8, pp. 64-69, 2012.

[Zi13a] Zimmermann, A.; Sandkuhl, K.; Pretz, M.; Falkenthal, M.; Jugel, D.; Wißotzki, M.: Towards an Integrated Service-Oriented Reference Enterprise Architecture. European Software Engineering Conference ESEC 2013 and the ACM SIGSOFT Symposium on the Foundations of Software Engineering with WEA Workshop on Software Ecosystem Architectures, August 18-26, 2013, St. Petersburg, Russia, pp. 26-30, 2013.

[Zi13b] Zimmermann, A.; Pretz, M.; Zimmermann, G.; Firesmith, D. G.; Petrov, I.; El-Sheikh, E.: Towards Service-oriented Enterprise Architectures for Big Data Applications in the Cloud. 17th IEEE International EDOC Conference (EDOCW 2013): The Enterprise Computing Conference with SoEA4EE, 9-13 September 2013, Vancouver, BC, Canada, pp. 130-135, 2013.

Towards Data Supply Chains in Enterprise Architecture Management

Kurt Sandkuhl, Matthias Wißotzki

Chair Business Information Systems
The University of Rostock
Albert-Einstein-Str. 22
18059 Rostock, Germany
kurt.sandkuhl@uni-rostock.de
Matthias.wissotzki@uni-rostock.de

Abstract: This paper discusses the idea of transferring the concept of supply chains and value creation from material flows to data flows. The paper has a focus on enterprise architecture management (EAM) and the data required for EAM. Our proposal is to transfer the idea of supply chains from manufacturing to EAM by combining three different approaches (a) information demand patterns as means to express the information demand of organizational roles in EAM, (b) information logistics value chains as conceptual foundation for establishing the value creation steps required from the data supply chain, and (c) knowledge services as core elements of an architecture for EAM data supply chain. The paper presents work-in-progress which so far primarily stays on a conceptual level. The research method used primarily is argumentative-deductive.

1 Introduction

Since Porter's seminal work on value creation and value chains [Po08], it is common practice in many industrial areas to base organization structures and activity planning on value creation principles. In more complex industry areas, like automotive or telecommunication industries, even the supplier structure and the chain of supply is oriented towards value creation principles [Gu01]. This paper discusses the idea of transferring the concept of supply chains and value creation from material flows to data flows. The paper has a focus on enterprise architecture management (EAM) and the data required for EAM. The idea as such is inspired by work from information logistics, a research area focusing on demand-oriented information supply.

In the context of big data scenarios, demand-orientation is expected to contribute to solve or bypass some of the problems attached to big data. Big data challenges are commonly attributed to the three dimensions volume, variety and velocity [Zi12], i.e. large quantities of data have to be processed or analyzed nearly in real-time from data sources with varying structure. If the exact information demand and the value of the information for EAM are known, the volume of data to be analyzed probably is only

affected if data sources can be ruled out due to the demand, but the volume of data to be stored for later use definitively can be reduced.

Our proposal is to transfer the idea of supply chains from manufacturing to EAM by combining three different approaches:

- Information demand patterns as means to express the information demand of organizational roles in EAM
- Information logistics value chains as conceptual foundation for establishing the value creation steps required from the data supply chain
- Knowledge services as core elements of an architecture for EAM data supply chain

The paper presents work-in-progress which so far primarily stays on a conceptual level. We explore the transferability of value chain and information demand concepts from information logistics to data supply in enterprise architecture management. The research method used primarily is argumentative-deductive, i.e. we investigate existing theories and concepts regarding their applicability for the problem at hand and argue for the possibility to use them. Future work will have to include more design-oriented work, like, e.g., development of a prototype showing the feasibility and usefulness of the approach.

The remainder of this paper is structured as follows: section 2 summarizes the background for this paper. Section 3 presents our conceptual approach for data supply chain in EAM. Section 4 shows first elements of implementing the approach by using a typical EAM role as example. Section 5 summarizes the findings and gives an outlook.

2 Background

The background for this work stems from three different areas briefly discussed in this section: enterprise architecture management (2.1), supply chains and value creation (2.2) and information logistics (2.3).

2.1 Enterprise Architecture Management

Enterprises are complex, highly integrated systems comprised of processes, organizational units, information and supporting technologies, with multifaceted interdependencies and interrelationships across their boundaries [Ah12]. Enterprise Architecture Management (EAM) aims at providing an integrated view on all aspects of the organization under consideration with all dependencies of artifacts and information objects necessary for business performance. In this context, architecture is defined as a fundamental organization of a system embodied in its components, their relationships to each other and to the environment, and the principle guiding the organization's design and evolution [La09]. Enterprise Architecture (EA) is the formal declaration of the basic structures of an organization, its components and relations, as well as the processes used for development.

In this context, EAM provides an approach for a systematical development of the organization in accordance with its strategic vision [Ra11], yet its value depends on the organizational capabilities to perform EAM effectively. EAM consists of three major functions such as transforming, monitoring and planning. Transforming is the alignment of an enterprise architecture towards the target state of the respective enterprise architecture domains. Monitoring is the controlling that needs to be applied to ensure the intended progress and planning is the analysis of the current-state enterprise architecture that provides on-going information for specific questions that facilitate decision-making.

The need for efficient data supply in EAM was confirmed in a study regarding EAM challenges commonly perceived in industry [Wi13]. In particular the challenges that "*EAM is not just a task for enterprise architects […] but also for the whole enterprise*" and the "*lack of ability to express information demands*" confirm that there is a variety of information sources in enterprises providing important information for planning, monitoring and transformation tasks. Furthermore, many different organizational roles involved in EAM exist, each of them with distinct and different information demands.

2.2 Value Chains and Supply Chains

Supply chain and supply chain management (SCM) are terms frequently used in economics and production which have been thoroughly researched during the last decades. A multitude of textbooks and different approaches for structuring and describing supply chains and their management is available, like, e.g. [Cho04], [Cha07] and [Iva10]. Among the many definitions, the Council of SCM Professionals defines SCM as follows[1]: "*Supply Chain Management encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third-party service providers, and customers. […]*".

A supply chain as such consists according to [cho04] of all parties involved in fulfilling a customer request. The supply chain includes the manufacturer and suppliers, transporters, warehouses, retailers, and customers themselves. Within each organization, such as a supplier, the supply chain includes all functions involved in receiving and filling a customer request, like, e.g., new product development, marketing, operations, distribution, and customer service. Supply chains build upon the concept and paradigm of value chains. According to Porter [Po08] "*A value chain is a chain of activities that a firm operating in a specific industry performs in order to deliver a valuable product or service for the market.*" The idea of value chains is based on the process view of organizations, i.e. an organization is considered as a system, made up of subsystems each with inputs, transformation processes and outputs, all of them consuming resources, like money, labor, materials, equipment, administration and management.

Development of a value chain for data supply in EAM requires to not only consider data supply as chain of transformation processes but to also understand the value creation

[1] See http://cscmp.org/sites/default/files/user_uploads/resources/downloads/glossary-2013.pdf

connected to the different transformations. Transformation processes in this context are not necessarily data transformation processes, but also include transformation processes performed (manually) by organizational roles or by stakeholders outside the organization.

2.3 Information Logistics

The research field information logistics was established in the late 1990s and defined in [Sa07]. The main objective is optimized information provision and information flow, based on information content, time of delivery, location, presentation and quality. The information logistics field focuses on improving the information flow by applying logistic principles to information supply. Identifying and providing the right content is a core challenge of information logistics. During the last decade, many IT applications have been developed implementing the objective of information logistics. Some of the applications are services providing bad weather warnings, traffic information or personalized news, and solutions for businesses in different domains like WIND service (weather information on demand), Smart-Wear (location-based information supply for mobile users) [Sa07]. The concept of information demand is defined by [Lu07] as "…the constantly changing need for current, accurate, reliable, and integrated information to support (business) activities, whenever and where ever it is needed."

This definition implies a number of aspects that must be considered while analyzing information demand. Information demand should change as the task, roles and responsibilities, to which information demand is connected, change. The information should be relevant, current, accurate and reliable; otherwise it will contribute to information overflow. The information demand should be integrated with the business activities, as it is necessary to have a solid knowledge about the context in order to be aware of any changes of information demand that might happen. Whenever and where ever emphasize the importance of time and location while analyzing the information demand [Lu07]. A specific method for information demand analysis was developed and evaluated in a number of industrial projects [Lu11].

3 Approach for Data Supply Chain in EAM

This section introduces the proposed approach for establishing data supply chains in EAM. The basic idea behind our approach is to start from the information demand of the roles in enterprise architecture management and to focus on the value creation required for providing the information demanded by these roles. Traditional information system development approaches or software engineering approaches start from requirements and derive functional and non-functional requirements to be met by the IT-system which is supposed to provide required information. We argue that flexibility and adaptability in a data supply chain would benefit from first understanding the required value creation and in a second step investigating the required functionality and means to implement them. The value creation perspective allows for explicitly taking into account organizational and technical means to fulfill the information demand. This demand-

driven procedure is not focused on implementation issues and puts the need of the EAM roles in focus.

More concrete, our approach at this early stage of our work consists of two elements: (a) a defined process for eliciting the information demand and the value creation connected to this demand which has to be met by the data supply chain, and (b) an architecture for implementing the technical, IT-related part of the data supply. These two elements will be introduced in the next two sections.

3.1 Process for determining information demand and required value creation

The proposed process for defining a data supply chain in EAM starts from the roles in EAM and their tasks and responsibilities. We propose to start by either performing an information demand analysis according to the method published in [Lu11] or to use an information demand pattern for the role, if this already exists (step 1). Starting from the role's information demand description or information demand pattern, the information supply actions can be derived, which include the required value creation (step 2). The information supply actions can be used in order to determine IT-based services categories to be included in the supply chain (step 3). The binding of the IT-based services categories to concrete IT-components and processing sequences would be the final step (step 4). All steps are reflected by the architecture presented in section 3.1. Steps 1 to 3 will be described in more detail in the following.

Step 1: Information Demand Patterns
The general idea of information demand patterns is similar to most pattern developments in computer science [Sa11]: to capture knowledge about proven solutions in order to facilitate reuse of this knowledge. *In the context of data supply for EAM, information demand patterns should be developed for the typical roles in EAM (see section 4 for an example).* An information demand pattern can be defined as follows: *An information demand pattern addresses a recurring information demand problem that arises for specific roles and work situations in an enterprise, and presents a conceptual solution to it.*

Such a pattern consists of a number of essential parts describing the pattern:

- A statement about the *organisational context* where the pattern is useful. This statement usually identifies the application domain or the specific departments or functions in an organisation forming the context for pattern definition.
- *Problems* of a role that the pattern addresses. The tasks and responsibilities a certain role has are described in order to identify and discuss the challenges and problems, which this role usually faces in the defined organisational context.
- The *conceptual solution* that resolves the problem, which for information demand patterns includes three parts:
    - *Information demand* of the role, which is related to the tasks and responsibilities, is described as part of the pattern, i.e. the different parts of the information demand are identified

- *Quality criteria* for the different parts of the information demand include the general importance of the information demand part, the importance of receiving the part completely and with high accuracy, and the importance of timely or real-time information supply
- a *timeline* indicating the points in time when the different information parts should be available at the latest

- The *effects* that play in forming a solution. If the needed information part should not be available or arrive too late this might affect the possibility of the role to complete its task and responsibilities. The effects described in the pattern include
  - Potential economic consequences
  - Time/efficiency effects, i.e. whether the role will need more time for completing the task or will be less efficient
  - Effects on improving or reducing the quality of the work results
  - Effects with respect to the motivation of the role responsible
  - Learning and experience effects
  - Effects from a customer perspective

An example of an information demand pattern with the above structure is presented in section 4, table 1.

Step 2: Information Supply actions
The information supply actions define the value creation linked to the question: how to decide on the right way to provide the demanded information? Should the information be "pushed" to the person or "pulled" by the role? At what point in time is the value of the information optimal from the role's perspective? Should the information be aggregated with other related information?

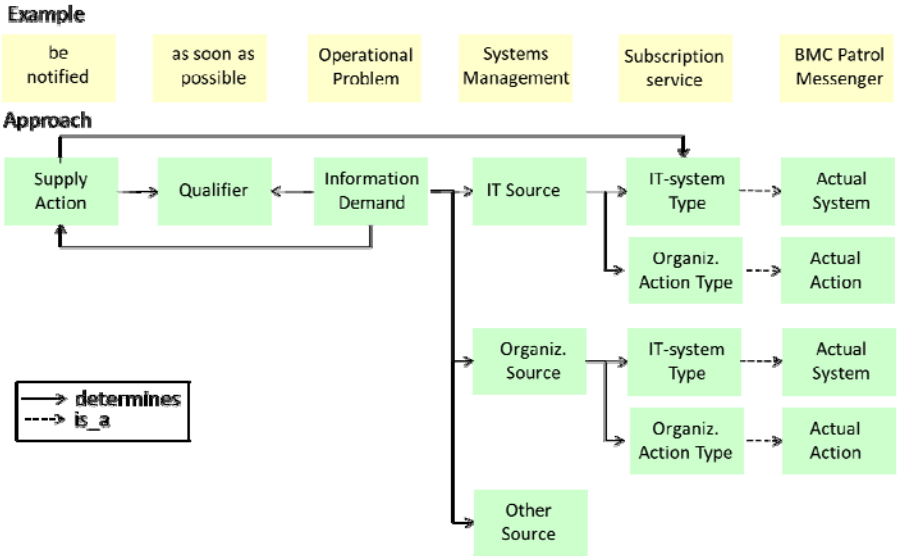


Figure 1. *Supply action and related concepts*

Figure 1 illustrates the main concepts in the area of information demand and information supply actions and their relations. The figure consists of an example (upper part, coloured yellow) and the actual approach (below the example, coloured green). The central concept of the approach obviously is the information demand, for example the "operational problems" of the role "application responsible". This information demand is characterized by a qualifier, which reflects the quality criteria contained in the information demand pattern. The example in figure 1 is "as soon as possible". The activity required to satisfy the information demand with the given qualifier we will call supply action. In figure 1 the supply action is to "be notified", i.e. in order to receive the information about an operational problem as early as possible, the role "application responsible" should be notified if new operational problems emerge.

Satisfying the information demand also requires to identify the source of information. This can be an IT based source (labelled "IT source" in figure 1), like a web site or an information system, or an organisational source (labelled "organiz. source"), like another role or organization in an enterprise, or another kind of source, like the personal network of the individual having the role or a community of practice. The example given in figure 1 shows the "systems management " as IT source for the operational problem information. For all supply actions, we potentially have both, technologies or organizational activities, which can be used to realize these actions, applied both on IT sources, organizational sources or other sources.

As a starting point, we propose to use the supply actions identified in several information logistics projects and summarized in [Kh10]. For brevity reasons, we will only include the IT-based supply actions, not the organizational ones:

- Receive: this supply action should be used if the most appropriate way to supply the information to a certain role is that this role "receives" this information, i.e. from the IT source or the organizational source the information is pushed to the role. This supply action typically is used with qualifiers indicating the timing of the supply, like to receive the information in real-time or receive it in case of a certain event. In comparison to the "be notified" action, the complete information is provided to the role.
- Be notified: this supply action means that the role receives a notification about how to get the demanded information. i.e. not the complete information is provided, but changes, updates or new information are indicated.
- Retrieve: this supply action should be used if the information required is not time-critical and not very dynamic. The user pulls this information when it is convenient for her or him.
- Monitor: this supply action should be used if changes in a specific information or information source are part of the information demand of a role. Monitoring can in practice be very similar to "be notified", but should be used for changes, not for new information.
- Look up: this supply action should be used if accuracy of an information is very important and the information is not used very often by a certain role. To "look up" the information at the place of origin on-demand is supposed to contribute to avoiding mistakes.

- Verify: this supply action is appropriate if the correctness and up-to-date-ness of an information has to be verified, i.e. the information basically is available. The source and way to retrieve it should be provided.
- Aggregate: in addition to receiving a certain information, some information demands even consist of aggregating related information into one "delivery". This supply action should be used if many small and related information items appear in a relatively short time frame and can be aggregated into a larger information package.

Step 3: Determine IT-services
The portfolio of IT-services applicable for implementing the supply actions can roughly be divided into tools for "pushing" information to the role having a certain information demand, and in tools requiring the role to "pull" this information. In general, there are usually standard off-the-shelf tools and solutions specifically developed for a certain application area. Examples for "pull" tools are: search engines, information retrieval systems, retrieval functionality in document management system, mediator based information systems, meta-information systems. Examples for "push" tools are news service, RSS feeds, e-mailing systems supporting mail lists, event-condition-action systems and subscription services based on information systems.

In EAM these IT-services will have to provide information from sources relevant for EAM, like EAM tools, systems management solutions, business activity monitors, incident managements systems or ERP-systems.

The example in figure 1 shows "subscription service" as IT-system type for implementing the supply action "be notified" and "BMC Patrol Messenger" as the actual system providing the information about "operational problems" from "systems management".

3.2 Architecture for Data Supply Solutions

The architecture for data supply solutions has to reflect the integration various information sources and the flexible, adaptive provision of information to organizational roles based on processing operations. Many different types of architectures have been proposed in research which would in principle be suitable for this purpose. Examples are context-based systems, publish/subscribe architectures or event-condition-action systems.

Our proposal is to base the data supply architecture on knowledge management architectures and, more specific, on knowledge services. Maier proposed an architecture with five layers depicted in figure 2. These layers are:

- Access services: authentication and authorization of the user via different access channels
- Personalization services: adaptation of user interface and content provided to the user's individual demands

- Knowledge services: core layer of the architecture including the four basic knowledge services search, publishing, collaboration and learning
- Integration services: integration of different information and knowledge sources including format and semantic integration, e.g. by using ontologies of structured vocabularies
- Infrastructure services: transparent provision of access to different information sources, which includes wrappers for the sources, homogenous access services, etc.

Maier's architecture needs adaptation for our purposes mainly in the architecture layers II and III, i.e. personalization and knowledge services. For personalization, we will use the results of the information demand analysis or information demand patterns. The information demand is uses to personalize the services to the role's specific information demands. Regarding the knowledge service, the supply actions will be integrated as specific additional services, which in turn will be implemented using the integration service layer. Most of the supply actions can be considered as contributions to the knowledge services "search" and "publishing": retrieve and lookup are based on search, be notified and receive are based on publishing, monitor and verify require a combination of search and publishing.

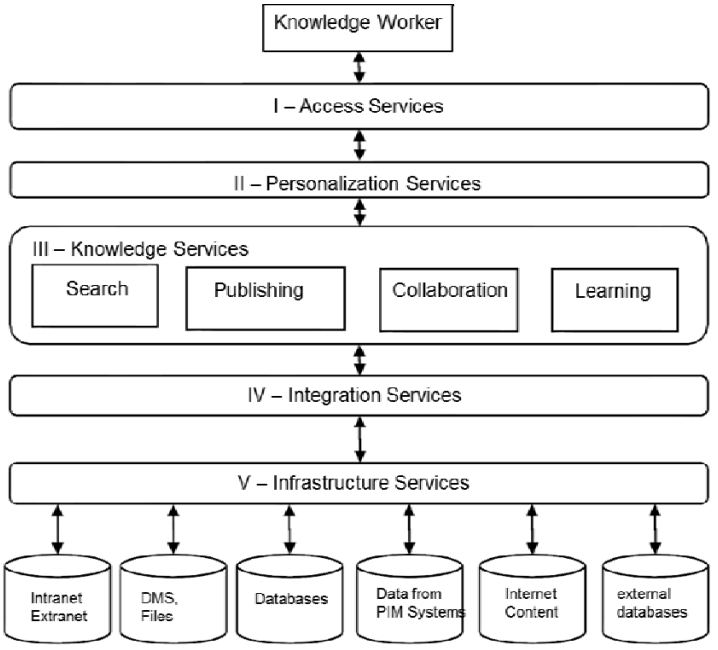


Figure 2. Adaptation of Maier's knowledge service architecture

4 Example: Data Supply for Application Owner

In order to illustrate our proposal and as an initial step of validation, we developed an information demand pattern for a typical role in EAM. Based on this pattern, we will perform the three steps introduced in section 3.1.

Step 1: As introduced in section 3.1, an information demand pattern consists of a number of essential parts used for describing the pattern: pattern name, organizational context, problems addressed, conceptual solution (consisting of information demand, quality criteria and timeline), and effects. For brevity reasons, we will limit the presentation of the information demand pattern on context, problem and the information demand element of the conceptual solution. The example considered is the role of an Application Owner. Table 1 shows an excerpt from the information demand pattern for this role.

Table 1: Excerpt from information demand pattern for Application Owner

Role: Application Owner in EAM
Context: the application owner role is considered in the context of EAM. The problems associated to the application owner role is: many different information sources in the organization potentially provide relevant information about the application, technical or organizational components relevant for the application, or business, technical and legal requirements important for the application from different organizational actors and at different not a-priori known points in time. To keep track of all information in time and act accordingly requires availability of all information to the application owner, which often is not implemented.
Tasks and Responsibility: An application owner is responsible for a defined IT-application in an organization during the application's complete lifetime and that the application meets the organizational requirements from an operational, functional and quality perspective. Among the tasks of the role are • Continuously capture business requirements and take them into account for further development of the application and during release planning • Make sure that the IT-infrastructure underlying the application meets the requirements of the application • Supervise quality and conformance with service level agreements of the application delivery • Responsibility for the application to be compliant to relevant policies and laws • Responsibility for the application to deliver the expected business value, which includes appropriate integration into business services
The information demand of the role consists of*:* • Changes in organization-internal and external policies, laws and regulations affecting the application • Severe operational problems which are recurring and not solved by first level and second level operational support • Planned changes in the IT-infrastructure possibly affecting the application, like

changes in the releases of operating systems, databases, server components or underlying services
- Strategic and tactical plans of the organization leading to new business requirements to the application
- Complaints and emerging new requirements from the users of the application
- Changes in the cost structures related to the application
- Changes in business process execution related to the application

Step 2: From the information demand, supply actions can be derived. Again, we present only an excerpt of the supply actions for the application owner's information demand pattern. The excerpt in table 2 focuses on severe operational problems and changes in business process execution.

Table 2: Supply actions for the information demand pattern in table 1 (excerpt)

Supply Action	**Qualifier**	**Information Source**	**Comment**
Information Demand: Severe operational Problems			
Be notified	As soon as problem occurs repeatedly	Incident management or trouble ticketing platform of the systems and application management responsible for operations of the application	It is essential to receive the information about recurring problems as soon as possible; the detailed problem description as such can be retrieved after assessing the severity. It is not always possible to subscribe "severe and recurring problems"
Information demand: change in business process execution			
Receive	In case the number of cases exceeds a given percentage of the overall process throughput	Business activity monitoring of the workflow platform underlying the application	If the business process usually is executed automatically, the number of cases to be handled manually is interesting for the role. If this number of "manual" cases grows, this indicates that potentially the business process has to be adjusted.

Step 3: Based on the supply actions, we have to determine the IT-services and how the value creation desired in the supply actions can be achieved. Table 2 already indicates which "information sources" are relevant for the information demand considered in the example, which in this example directly correspond to IT-services. In the following, we will focus on the different activities to be performed in order to achieve the value creation using the identified IT-services.

The example in table 2 identifies incident management and business activity monitoring as relevant sources. From both sources, information has to be used under certain conditions ("qualifier"). Thus, the implementation of the supply actions has to include to "subscribe" to certain information. However, the further processing steps are different. Regarding the supply action "be notified" from incident management, the processing could look as follows: subscribe to incident management information – adapt format of this information to standard – filter information according to qualifier – evaluate information with regards to relevance – send notification to role. For "receive" from business activity monitoring, the processing could be as follows: subscribe to business activity monitoring information – adapt format of this information to standard – store information to track record file – query track record for relevant set of records – create report according to qualifier – send report to role.

Both processing sequences contain similar and partly even identical steps, but also different steps. The similar / identical steps can be an indicator that the same (parameterizable) IT-service can be used which avoids multiple implementation of the same functionality. For the above example, this would, e.g., be "adapt format to standard". When analyzing the supply actions and required IT-services for all roles in EAM, we expect to find a large number of overlaps between the supply actions and high potential to reuse the same kind of IT-service. Thus, the design of a data supply chain consisting of a limited number of processing steps (IT-services) which are used many times would be possible, which is an advantage as compared to constructing information supply independently for different roles or separately for different information sources.

5 Summary and Future Work

The work presented in this paper has many limitations which have to be subject of future work. Only one example was considered to explain the approach and show the conceptual feasibility. The approach has not been fully elaborated; possibly, further adjustments in the reference architecture are required. An implementation of at least a prototypical solution is required in order to confirm feasibility. In particular for big data volumes we need adjustments on the integration and persistence service layers in the architecture discussed in section 3.2.

The most important future work, however, is in moving from the one example and single role focused view of the supply chain to a supply chain design for EAM in an organization in general. The real value of the supply chain idea is to consider data supply as management task and to manage the whole data supply chain in a holistic way. Principles and core processes from supply chain management in manufacturing

industries are important sources of inspiration for this future work. The supply chain management institute distinguishes eight different processes in SCM, all of them should be transferred to data supply chain management also: Customer Relationship Management Process, Supplier Relationship Management Process, Customer Service Management Process, Demand Management Process, Order Fulfillment Process, Manufacturing Flow Management Process, Product Development and Commercialization Process, and Returns Management Process.

Demand management, to take one example, is the SCM process that balances the customers' demand with the capabilities of the supply chain. With the right process in place, management can include forecasting, increase flexibility and reduce variability. In data supply chains, demand management would have to include the different role's demand, a view on potential demand changes triggered by organizational changes and a forecast on adaptation required in the data supply chain.

References

[Ah12] Ahlemann, F. (2012) Strategic enterprise architecture management: challenges, best practices, and future developments. Berlin, New York: Springer, 2012.

[Cha07] Charu Chandra, Janis Grabis: Supply Chain Configuration: Concepts, Solutions, and Applications. Springer Verlag, 2007.

[Ch04] Chopra, S. and P. Meindl (2004) Supply Chain Management. 2 ed. Upper Saddle River: Pearson Prentice Hall, 2004.

[Gu01] Gunasekaran, A. (2001). Agile Manufacturing: The 21st Century Competitive Strategy: The 21st Century Competitive Strategy. Elsevier.

[Iva10] Dmitry Ivanov,Boris Sokolov (2010) Adaptive Supply Chain Management. Springer Verlag, 2010.

[Kh10] Khan, M. T. and B. Khademhosseinieh (2010) The Supply side of Information Demand: Organisational and Technical Measures for demand oriented Information Supply. IADIS International Conference Information Systems 2010. ISBN: 978-972-8939-09-0.

[La09] Lankhorst, M. (2009) Enterprise architecture at work: modelling, communication, and analysis, 2nd ed. Dordrecht, New York: Springer, 2009.

[Lu07] Lundqvist, M., 2007. Information Demand and Use: Improving Information Flow within Small-scale Business Contexts, Licentiate of Engineering, Department of Computer and Information Science, Linköpings universitet.

[Lu11] Lundqvist, M.; Sandkuhl, K.; Seigerroth, U. (2011) Modelling Information Demand in an Enterprise Context: Method, Notation and Lessons Learned. International Journal Systems Modeling and Design, Vol. 2 (3), IGI Publishing, pp. 74-96, 2011.

[Ma07] Maier, R.: Services Supporting Knowledge Maturing in Small and Medium-Sized Enterprises. In G. Dhillon, B. Stahl and R. Baskerville (eds.): Information Systems – Creativity and Innovation in Small and Medium-Sized Enterprise. Springer Berlin Heidelberg. 301: 224-238, 2009.

[Po08] Porter, M. E. (2008). Competitive advantage: Creating and sustaining superior performance. Simon and Schuster.

[Ra11] Razavi, M.; Shams Aliee, F. and K. Badie (2011)An AHP-based approach toward enterprise architecture analysis based on enterprise architecture quality attributes. Knowl Inf Syst, vol. 28, no. 2, pp. 449–472, 2011.

[Sa07] Kurt Sandkuhl: Information Logistics in Networked Organizations: Selected Concepts and Applications. Enterprise Information Systems, 9th International Conference, ICEIS 2008. LNBIP, Springer.

[Sa09] Kurt Sandkuhl, Magnus Lundqvist, Ulf Seigerroth (2009) Information Supply Patterns and Technology & Activity Portfolio. InfoFLOW Deliverable D5, Version 1.1, March 2009, Jönköping University, Sweden.

[Sa11] Sandkuhl, Kurt (2011) Information Demand Patterns. Proc. PATTERNS 2011, The Third International Conferences on Pervasive Patterns and Applications, pp. 1-6. September 25-30, 2011; Rome, Italy. ISBN: 978-1-61208-158-8.

[Wi13] Matthias Wißotzki, Hasan Koç, Tino Weichert, Kurt Sandkuhl (2013) Development of an Enterprise Architecture Management Capability Catalog. Proceedings 12th International Conference on Perspectives in Business Informatics Research -, BIR 2013, Warsaw, Poland, September 23-25, 2013. LNBIP 158, pp. 112-126, Springer. ISBN 978-3-642-40822-9.

[Zi12] Zikopoulos, P. C.; Eaton, C. and P. Zikopoulos (2012) Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. Mcgraw-Hill Professional, 2012.

Performance-aware Application Distribution in the Cloud

Santiago Gómez Sáez, Vasilios Andrikopoulos, Frank Leymann
Institute of Architecture of Application Systems
Universtät Stuttgart,
Universitätsstraße 38, 70569 Stuttgart, Germany
{gomez-saez, andrikopoulos, leymann}@iaas.uni-stuttgart.de

Abstract: The emergence of Cloud computing and the improvement of resource management techniques have contributed to an increase in the number of application developers that are strong supporters of partially or completely migrating their application to a highly scalable and pay-per-use infrastructure. In this work in progress paper we begin the analysis on how to optimally distribute the application layers in the Cloud in order to adapt its topology to handle oscillating over time workloads. More specifically, through an empirical workload analysis and characterization we present our initial evaluation of an application persistence layer's performance under different deployment scenarios.

1 Introduction

With the increasing number of available Cloud services which allow to partially or completely deploy the application in the Cloud, it becomes possible to host only some of the application components off-premise (in the Cloud), e.g. its database, while the remaining of the application remains on-premise [ABLS13]. Our investigations concentrate on optimally distributing the application in the Cloud considering the workload that the underlying infrastructure must bear. Given that the application workload oscillates over time, we therefore aim to analyze how to optimally adapt the application deployment topology accordingly. In this work in progress paper we focus on the application persistence layer and discuss our initial performance analysis for application topology options using the TPC-H database benchmark workload and data as the basis. The contributions of this work can therefore be summarized as:

1. the analysis of the relationship between the application deployment topology, its workload oscillation over time, and the expected performance,
2. a workload characterization focusing on the application persistence layer by using the TPC-H benchmark, and
3. the generation and performance evaluation of artificial workloads for different deployment topologies of the application persistence layer.

The remainder of this paper is structured as follows: Section 2 summarizes relevant concepts and related works. Sections 3 and 4 present our experiments and discuss the most important

findings, respectively. Finally, Section 5 concludes with the future work.

2 Background

Application deployment and distribution tasks in a Cloud infrastructure require the fulfillment of preliminary tasks related to compliance, underlying resources description, application distribution arrangement, cost calculations, etc. In order to guarantee the expected application performance, such tasks should incorporate performance awareness. The migration of the different layers of an application to the Cloud is analyzed in [ABLS13], where multiple migration types are categorized and their corresponding adaptation needs are identified. In [LFM+11] a migration method and tool chain based on application model enrichment for optimally distributing the application across one or multiple Cloud providers is presented. In this work we target the challenge of optimizing the application layers' distribution in the Cloud based on its workload and expected performance. Moreover, this workload may oscillate over time due to external or internal parameters. For example, an online web store workload is increased at certain time periods, e.g. before the Christmas season, which may generate unnecessary monetary costs and/or performance degradation. This problem can be analyzed from two different perspectives: from the *Cloud Consumer* and the *Cloud Provider* objectives. On the one hand, the Cloud consumer aims to maximize the resource usage while minimizing the incurred monetary costs. On the other hand, the Cloud provider's goals are associated with the utilization of virtualization and multi-tenancy techniques to minimize operational costs while ensuring isolation between the loads generated by each Cloud consumer. Our goal, therefore, is to provide the necessary methodology and artifacts to analyze such workload oscillations over time and dynamically (re-)distribute the application layers towards bridging the gap produced by the existing conflict of interests between the Cloud consumer and the Cloud provider.

According to several investigations [BM11, GRCK07, JVS98, MMVP13, MMZVP13, WMG+10], two approaches for analyzing the application workload evolution can be identified: top-down and bottom-up. In the top-down approach, the application workload is characterized, and the application behavior model is derived before or during the deployment of the application. The importance of understanding the characteristics of the application workload to achieve efficiency is discussed in [JVS98], and hence a set of micro-architectural metrics are presented. In [BM11], an application workload specification language for creating synthetic workloads to evaluate Cloud applications is created. The combination of probability distribution fitting techniques and workload parameter analysis facilitates the building of the workload behavior model. However, a drawback of only following a top-down analysis approach is the inability of handling the workload evolution over time. Bottom-up approaches address this deficiency with the help of resource consumption monitoring techniques and performance metrics. Predicting infrastructure consumption needs based on the periodicity and similarity of occurrences over time workload demands is investigated in [GRCK07], where the cyclical aspect of multiple workloads in a large number of data centers allows to characterize, predict, and place workloads. Further resource prediction analysis can be driven using simulation approaches, such in [FFH12],

by simulating the monetary and performance cost of multiple Cloud deployment options (CDO). Towards satisfying Service Level Objectives (SLOs), in [WMG+10] the probabilistic relationship between resource consumption and application performance to model the response time distribution is analyzed. The analysis and generation of performance models for data-intensive workloads in public Clouds is addressed in [MMZVP13]. Such models can ease capacity management operations to predict the workload behavior and to determine the most cost-effective resource configuration [MMVP13].

Top-down and bottom-up application workload analysis approaches can be combined over time in order to proactively satisfy application demands by dynamically (re-)adapting its topology. In this work we first focus on analyzing the application workload at the persistence layer, and therefore we use the existing TPC-H benchmark[1] as the basis to generate application workloads that fit to different probability distributions, and analyze the consequent performance under different deployment topologies.

3 Experiments

3.1 Experimental Setup

We focus our experiments on emulating the behavior of an application which is built using the three layers pattern (*presentation*, *business logic*, and *data*, i.e. persistence) proposed in [F+02]. In a first step, we generate 1GB of representative application data using the TPC-H Benchmark. In a second step, we use Apache JMeter 2.9[2] as the application load driver to emulate the application business logic layer, using the automatically generated set of 23 TPC-H SQL queries as the load. We use the following infrastructures for the distribution of the application:

- an on-premise virtualized server on 4 CPUs Intel Xeon 2.53 GHz (2 virtual cores per CPU) with 8192KB cache, 4GB RAM, and running the Ubuntu 10.04 Linux OS and MySQL 5.1.72,
- an off-premise virtualized server (IaaS) hosted in the Flexiscale Infrastructure[3] consuming 8GB RAM, 4 CPUs AMD Opteron 2GHz with 512KB cache (2 virtual cores per CPU), and running the Ubuntu 10.04 Linux OS and MySQL 5.1.67,
- and an off-premise MySQL 5.1.69 db.m1.xlarge database instance (DBaaS) hosted on Amazon RDS[4].

We create three distribution scenarios, with the application data 1) in the MySQL on-premise, 2) the MySQL on the IaaS solution, and 3) on the DBaaS solution, while the load driver remains in all cases on-premise. The application persistence layer performance

[1] TPC-H Benchmark: `http://www.tpc.org/tpch/`
[2] Apache Jmeter: `http://jmeter.apache.org/`
[3] Flexiscale: `http://www.flexiscale.com/`
[4] Amazon RDS: `http://aws.amazon.com/rds/`

Table 1: Generated TPC-H Database Schema Analysis.

Table	Size (MB)	Columns	Access Count
customer	23.41	8	8
lineitem	614.06	16	18
nation	0.0023	4	9
orders	140.49	9	12
part	23.43	9	8
partsupp	113.77	5	5
region	0.0004	3	3
supplier	1.37	7	10

in each scenario is measured by normalizing the throughput (Req./s). The measured throughput is defined as the sum of the number of transaction per second of the application database engine, and the network latency introduced by the different deployment topologies of the application persistence layer. Such experiments have been driven across 10 rounds on average per day for a period of three weeks.

3.2 TPC-H Characterization

As previously discussed, top-down and bottom-up approaches can be combined to evaluate and analyze the application workload and behavior over time. For this purpose, using the first deployment scenario (on-premise) we analyze the database schema and initial workload of the application as summarized by Table 1 and Table 2. Table 1 shows the relationship between the database schema characteristics and the total number of accesses that the workload performs on each table.

Table 2: TPC-H Workload Analysis.

Query	Accessed Tables	Subqueries	Total Logical Evaluations	Throughput (Req./s)			Retrieved Data (B)
				On-Premise	IaaS	DBaaS	
$Q^{(1)}$	1	0	1	0.03425	0.03396	0.03817	538
$Q^{(2)}$	5	1	13	0.07927	0.14884	3.03260	15857
$Q^{(3)}$	3	0	5	0.08687	0.11733	0.31185	376
$Q^{(4)}$	2	1	5	0.53950	0.73922	0.94903	105
$Q^{(5)}$	6	0	9	0.01148	0.02014	0.33484	130
$Q^{(6)}$	1	0	4	0.20583	0.21355	0.28261	23
$Q^{(7)}$	5	1	11	0.03123	0.04782	0.20792	163
$Q^{(8)}$	7	1	11	0.97156	1.45380	0.18196	49
$Q^{(9)}$	6	1	8	0.05947	0.09123	0.05548	4764
$Q^{(10)}$	4	0	6	0.09168	0.11970	0.49834	3454
$Q^{(11)}$	3	1	6	2.59998	4.07134	0.26802	16069
$Q^{(12)}$	2	0	7	0.21147	0.22465	0.13981	71
$Q^{(13)}$	2	1	2	0.12771	5.32350	-	16
$Q^{(14)}$	2	0	3	0.03373	0.06017	0.29052	28
$Q^{(15)}$	1	0	2	201.533	22.2591	23.1152	9
$Q^{(16)}$	2	1	2	0.11346	0.11219	0.13471	120
$Q^{(17)}$	3	1	6	0.10931	0.19021	0.97148	648259
$Q^{(18)}$	2	1	5	0.98213	1.81212	-	25
$Q^{(19)}$	3	1	3	-	-	-	-
$Q^{(20)}$	2	0	25	4.05648	4.90228	0.17083	21
$Q^{(21)}$	5	2	8	3.02705	5.32847	-	8989
$Q^{(22)}$	4	2	13	0.01070	0.01734	0.06065	8944
$Q^{(23)}$	2	2	6	2.72083	3.30785	-	137
Avg.	3.17	0.73	7	9.89262	2.29976	1.72467	32188.5
Median	3	1	6	0.12058	0.20188	0.27531	133.5
σ	1.74	0.68	5.23	41.8356	2.67162	5.23164	137693.78

A secondary analysis consists of dissecting the set of queries which constitute the workload, and quantitatively analyzing their logical complexity, table joints, subqueries, etc. (Table 2). Throughput and retrieved data size measurements are considered as performance metrics, and therefore are a part of the bottom-up analysis approach. We combined both analysis approaches towards analyzing the relationship between the complexity of the workload queries and the performance of different application persistence deployment topologies. Towards this goal, queries are categorized by trimming the mid-range of the initial workload measured throughput and by comparing the total number of logical evaluations with respect to the remaining set of queries in the workload. Given the strong connection between the measured throughput and the resource consumption of the database engine in the TPC-H benchmark, the following categories are defined: compute high (CH), compute medium (CM), and compute low (CL).

3.3 Workload Generation and Evaluation

Subsequent to characterizing and grouping the application workload into the previously presented categories, workload generation and evaluation can take place. For this purpose, probabilistic workload generations techniques are supported, for example, by Malgen[5] and Rain [BLY+10]. The former supports the creation of large distributed data sets for parallel processing benchmarking, e.g. Hadoop, while the latter provides a flexible and adaptable workload generator framework for Cloud computing applications. However, such approaches do not consider the distribution of the application layers.

By using scripting techniques and the set of successfully executed TPC-H queries, we created multiple workloads for each of the categories we identified in the previous with size 1000 SQL queries each, using a different probability distribution for each workload. The workload probability distribution was derived by calculating its cumulative function and using the probability distribution fitting functionalities provided by EasyFit[6]. Table 3 summarizes the measured average throughput for each distribution. In the future, we plan to evaluate existing workload generation tools to incorporate support for generating multiple artificial workloads according to specified probability distributions considering the different deployment topologies of the application layers.

4 Analysis & Discussion

The empirical analysis depicted in Tables 1 and 2 drives the following conclusions with respect to the initial workload:

- the number of accesses on the different tables is not proportionally distributed across the benchmark queries,

[5] Malgen: `http://code.google.com/p/malgen/`
[6] EasyFit: `http://www.mathwave.com/`

Table 3: Workload and Application Data Distribution Evaluation Results.

Scenario	Category	Probability Distribution	Throughput (Req./s)
On-Premise	CL	Inv. Gaussian	0.27749
On-Premise	CM	Pareto	0.05888
On-Premise	CH	Lognormal	0.02696
DBaaS	CL	Log-Logaristic	0.45238
DBaaS	CM	Inv. Gaussian	0.19972
DBaaS	CH	Lognormal	0.10273

- the queries which constitute the initial workload are highly heterogeneous with respect to the logical complexity and to the retrieved data, and
- such heterogeneity is also observed when comparing the throughput fluctuation of the different queries among the analyzed scenarios.
- When deploying the database in the IaaS solution, the average performance of 85% of the successfully executed queries improves between 3% and 4078%. However, such maximum peak is due to the nature of query $Q^{(15)}$, an SQL *CREATE VIEW* statement. In average, the initial workload performance without considering the maximum and minimum peaks is improved in approximately 166%.
- When deploying the database in the DBaaS solution, the average performance of 70% of the successfully executed queries improves between 11% and 3725%. However, such maximum peak is related again to the nature of query $Q^{(15)}$, and the average performance without considering the maximum and minimum peaks improvement of the initial workload is 225%.
- There are queries whose performance is degraded when being executed off-premise, such as $Q^{(1)}$, $Q^{(15)}$, and $Q^{(16)}$ for the IaaS solution scenario, and such as $Q^{(8)}$, $Q^{(9)}$, $Q^{(11)}$, $Q^{(12)}$, $Q^{(15)}$, and $Q^{(20)}$ for the DBaaS solution scenario.

After analyzing the individual behavior of the queries that constitute the initial workload, we analyze the performance of the generated workloads under different deployment topologies of the application persistence layer. We can observe from the obtained results depicted in Table 3 that:

- the compute demand is indeed increased among the three different workload categories, and the throughput is reduced by 78% to 90% when executing the workload on-premise, and by 55% to 57% when executing the workload in a DBaaS solution, using the CL category as the baseline, and

- the overall performance is highly improved when executing the generated workloads in a DBaaS solution, observing an increase of 163%, 339%, and 381% for the CL, CM, and CH workloads, respectively.

It can be therefore concluded that not only different workload distributions perform in a different manner, but also that adapting the application deployment topology with respect to the workload demands significantly and proactively improves the application performance. We therefore observe a necessity of providing support for (re-)adapting the application topology, i.e. (re-)distributing its layers to optimally consume the required resources to satisfy the workload demands oscillations. With the help of workload characterization and generation techniques, probabilistic models, and prediction capabilities, the application can be proactively and optimally adapted to satisfy different workload demands. Focusing on the application persistence layer, the combination of data replication techniques with dynamic routing of requests may help to achieve an optimal performance. Dynamic routing of database requests has been previously studied and enabled by the CDASMix [GALS10] approach which allows transparent access to multiple databases without requiring adaptations to the other application layers. CDASMix can therefore facilitate the (re-)distribution of the persistence layer across on- and off-premise solutions.

5 Conclusions and Future Work

Through the experimental analysis discussed in this work in progress paper, we demonstrate that the different application layers can produce oscillating over time workloads which impact on the overall application performance. The application layers (re-)distribution is proposed as an approach to proactively ameliorate performance degradations. Focusing on (re-)distributing the application persistence layer by means of evaluating on- vs. off-premise deployment models, and analyzing, characterizing, and generating workloads with different characteristics, our experiments already show an enhanced performance when the persistence layer is deployed off-premise. This improvement however varies in each distribution scenario depending on the type of queries used and the workload distribution.

Future work focuses on analyzing the performance for further deployment topologies. In order to help application developers to derive its topology and to (re-)adapt and tune it over time, we also plan to extend existing application topology description approaches to include support to proactively meet fluctuating application performance demands.

6 Acknowledgments

This work is funded by the FP7 EU-FET project 600792 ALLOW Ensembles.

References

[ABLS13] Vasilios Andrikopoulos, Tobias Binz, Frank Leymann, and Steve Strauch. How to Adapt Applications for the Cloud Environment. *Computing*, 95(6):493–535, 2013.

[BLY+10] Aaron Beitch, Brandon Liu, Timothy Yung, Rean Griffith, Armando Fox, and David A. Patterson. Rain: A Workload Generation Toolkit for Cloud Computing Applications. Technical Report UCB/EECS-2010-14, University of California, Feb 2010.

[BM11] Arshdeep Bahga and Vijay Krishna Madisetti. Synthetic Workload Generation for Cloud Computing Applications. *JSEA*, 4:396–410, 2011.

[F+02] M. Fowler et al. *Patterns of Enterprise Application Architecture*. Addison-Wesley Professional, 2002.

[FFH12] Florian Fittkau, Sören Frey, and Wilhelm Hasselbring. CDOSim: Simulating cloud deployment options for software migration support. In *Proceedings of MESOCA'12*, pages 37–46. IEEE, 2012.

[GALS10] Santiago Gómez Sáez, Vasilios Andrikopoulos, Frank Leymann, and Steve Strauch. Evaluating Caching Strategies for Cloud Data Access using an Enterprise Service Bus. In *Proceedings of IC2E'14*, 2010. *(to appear)*.

[GRCK07] Daniel Gmach, Jerry Rolia, Ludmila Cherkasova, and Alfons Kemper. Workload Analysis and Demand Prediction of Enterprise Data Center Applications. In *Proceedings of IISWC'07*, pages 171–180, 2007.

[JVS98] Lizy Kurian John, Purnima Vasudevan, and Jyotsna Sabarinathan. Workload Characterization: Motivation, Goals and Methodology. In *Proceedings of WWC'98*, 1998.

[LFM+11] Frank Leymann, Christoph Fehling, Ralph Mietzner, Alexander Nowak, and Schahram Dustdar. Moving Applications to the Cloud: An Approach based on Application Model Enrichment. *IJCIS*, 20(3):307–356, October 2011.

[MMVP13] Rizwan Mian, Patrick Martin, and Jose Luis Vazquez-Poletti. Provisioning Data Analytic Workloads in a Cloud. *FGCS*, 29:1452–1458, 2013.

[MMZVP13] Rizwan Mian, Patrick Martin, Farhana Zulkernine, and Jose Luis Vazquez-Poletti. Towards Building Performance Models for Data-intensive Workloads in Public Clouds. In *Proceedings of ICPE'13*, 2013.

[WMG+10] Brian J. Watson, Manish Marwah, Daniel Gmach, Yuan Chen, Martin Arlitt, and Zhikui Wang. Probabilistic Performance Modeling of Virtualized Resource Allocation. In *Proceedings of ICAC'10*, 2010.

All links were last followed on October 20, 2008.

Einfluss von Cloud Services auf Auswahlverfahren für IT-Lösungen

Oliver Bossert
McKinsey & Company
Taunustor 2
60311 Frankfurt
oliver_bossert@
mckinsey.com

Michael Falkenthal
Fakultät Informatik
Hochschule Reutlingen
Alteburgstr. 150
72762 Reutlingen
michael.falkenthal@
reutlingen-university.de

Mark Mauerwerk
T-Systems International
Otto-Röhm-Str. 71C
64293 Darmstadt
mark.mauerwerk@
t-systems.com

Gunther Piller
Fachbereich Wirtschaft
Fachhochschule Mainz
Lucy-Hillebrand-Straße 2
55128 Mainz
gunther.piller@
fh-mainz.de

Abstract: Immer mehr Unternehmen setzen auf IT-Services aus der Cloud. Diese Entwicklung verändert auch die Auswahlverfahren für IT-Produkte und Dienstleistungen. Eine explorative Untersuchung zeigt dies beispielhaft für Entscheidungsprozesse und Kriterien zur Bewertung von Angeboten für betriebliche Anwendungssysteme sowie für IT-Plattformen und Anbieter auf.

1 Fragestellung und Vorgehensweise

Cloud Computing spielt in der Unternehmens-IT eine immer wichtigere Rolle. Typische Gründe hierfür sind ein erhöhter Automatisierungsgrad für das Management von IT-Infrastrukturen oder kurze Beschaffungszeiten von IT-Services. Aufgrund der wachsenden Bedeutung von Cloud-Angeboten, stellt sich die Frage, ob und ggf. wie existierende Auswahlprozesse für IT-Lösungen angepasst werden müssen.

Die vorliegende explorative Untersuchung analysiert erstmals Vorgehensweisen zur Auswahl von IT-Systemen und Infrastrukturen, die Cloud Services berücksichtigen und von bisher verwendeten Standardverfahren abweichen. Hierzu wurden exemplarisch mehrere Unternehmen zum Vorgehen bei der Bewertung von Software as a Service (SaaS), Platform as a Service (PaaS) und Infrastructure as a Service (IaaS) Angeboten befragt. Die hierbei extrahierten Ergebnisse können in nachfolgenden Untersuchungen quantitativ erhärtet werden. Trotz der zu diesem Zeitpunkt wissenschaftlich nicht zuverlässig gesicherten Resultate, erscheinen die gemachten Beobachtungen interessant genug, um hier zusammengefasst zu werden. Die von uns gemachten Beobachtungen liefern wertvolle Ansatzpunkte für weiterführende Analysen und für die Verbesserungen der Auswahlpraxis von IT-Lösungen, die auf Cloud Services beruhen.

Die Untersuchung fokussiert sich auf den operativen Auswahlprozess von Cloud-Lösungen für Softwareanwendungen aus den Anwendungsbereichen Customer Relationship Management (CRM) und Groupware/E-Mail sowie für IT-Plattformkomponenten. Hierzu wurden qualitative Interviews mit IT-Spezialisten und Managern verschiedener Großunternehmen durchgeführt. Die Ergebnisse der Befragung werden in den nachfolgenden Abschnitten beschrieben. Abschnitt 2 gibt einen allgemeinen Überblick über die Resultate. Grundsätzliche Überlegungen zu einem Einsatz von SaaS werden in Abschnitt 3 skizziert. In Abschnitt 4 werden veränderte

Kriterien für eine Bewertung von betrieblichen Anwendungen vorgestellt. Abschnitt 5 erweitert diese Diskussion für IT-Plattformen. Abschnitt 6 fasst die gemachten Beobachtungen zusammen.

2 Änderungen in Systemevaluation und Auswahl

Viele Anwenderunternehmen folgen bei der Auswahl passender IT-Lösungen und deren Anbietern typischen Standardverfahren. Der Abgleich funktionaler und nicht-funktionaler Anforderungen für IT-Infrastrukturkomponenten oder Informationssysteme läuft meist nach ähnlichen Mustern ab (siehe z.B. [BWV07], [Gr10]): Nach Projektvorbereitung und Ist-Analyse wird ein Soll-Konzept aufgestellt, das auf den Anforderungen der Auftraggeber basiert. Anschließend wird nach der Systemevaluation in mehreren Schritten der beste Anbieter ausgewählt: Marktsichtung und Vorauswahl auf Grundlage der Anforderungen, Systemtests einschließlich Detailanalyse, nachfolgend Verhandlungen und Abschluss. Unsere Untersuchung konzentriert sich auf den Ablauf der Systemevaluation und Auswahl. Die Ergebnisse der vorangehenden Phasen Projektvorbereitung, Ist-Analyse und Soll-Konzeption werden als Input für die Phase der Systemevaluation angenommen. So wird z.B. durch einen Abgleich der Ist-Analyse und der Soll-Konzeption der Handlungsraum aufgedeckt, der durch funktionale und nicht-funktionale Anforderungen an die neue IT-Lösung beschrieben werden kann. Die so identifizierten Anforderungen fließen anschließend als Kriterien zur Beurteilung von neuen Systemen in die Systemevaluation ein [BWV07].

Wie in den nachfolgenden Abschnitten gezeigt wird, verändern Cloud-Angebote (siehe z.B. [VHH12]) die bestehende Praxis der Systemevaluation. Unternehmen klären meist vor den Phasen Marktsichtung, Vorauswahl, Systemtest, Detailanalyse und Auswahlempfehlung, ob und in welchem Umfang ein Cloud-Ansatz wie SaaS, PaaS oder IaaS für die geplante Lösung überhaupt in Frage kommt. Dementsprechend wird eine zusätzliche Phase zur grundsätzlichen Beurteilung eines möglichen Einsatzes von Cloud Computing dem bisherigen Vorgehen vorangestellt. Zudem werden in den nachfolgenden, üblichen Phasen des Auswahlprozesses auch Cloud-spezifische Kriterien zur Beurteilung herangezogen. Abbildung 1 illustriert diese Änderungen, die in den folgenden Abschnitten detailliert werden.

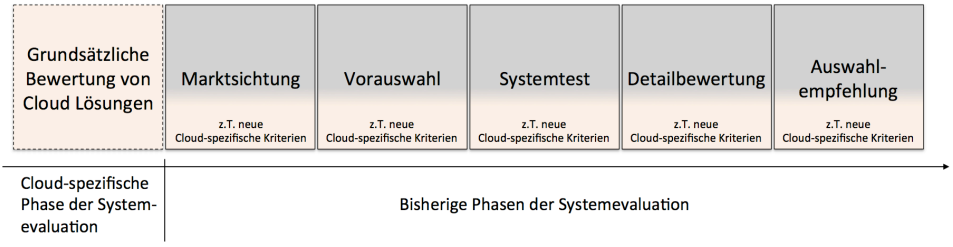


Abbildung 1: Cloud-spezifische Erweiterung des Vorgehens bei einer Systemevaluation

3 Grundsätzliche Bewertung von SaaS-Angeboten

Nur wenige der Befragten stellen Cloud-Angebote und klassische On-Premise-Lösungen gleichrangig nebeneinander und bewerten sie Schritt für Schritt. Vielmehr wird meist schon vorab entschieden, ob Cloud-Lösungen überhaupt in Betracht gezogen werden sollen. Entsprechende Kriterien werden hier beispielhaft für den Bereich SaaS beschrieben.

Grundsätzliche Entscheidungen für bzw. wider Cloud beruhen oft auf externen und internen Compliance- und Sicherheitsanforderungen – z.B. auf nationalen Gesetzen, die die Ablage sensitiver Daten in der Cloud einschränken. Interviews mit Unternehmen zeigen aber auch, dass eine pauschale Ablehnung von Cloud-Lösungen auf Grund solcher Anforderungen häufig differenziert hinterfragt wird. Bei genauerer Betrachtung sind oft auch Hybrid Cloud-Ansätze (siehe z.B. [Fe14]) möglich und sinnvoll. Ist beispielsweise die Speicherung von Rechnungsdaten in der Cloud untersagt, können Unternehmen diese zwar unternehmensintern ablegen und dennoch, z.B. in den Prozessen einer SaaS-Lösung für CRM verwenden. Auch Token-basierte Ansätze ermöglichen oft den Umgang mit strikten Vorgaben zur Datenspeicherung. Dabei werden Daten anonymisiert und zur Identifikation mit einem generierten eindeutigen Token versehen. Die anonymisierten Daten können dann in die Cloud übertragen und dort genutzt werden. Die Tokens verbleiben in unternehmensinternen On-Premise-Systemen und können so zur Identifikation der Daten herangezogen werden (siehe hierzu auch das Compliant Data Replication Pattern in [Fe14]). Ebenfalls beachtet werden nicht-funktionale Anforderungen, die für On-Premise-Lösungen meist kein Problem darstellen. So fordern einige Staaten, dass Finanzdaten auf physischer Ebene auditierbar sind. Das bedeutet für SaaS-Lösungen mitunter eine erhebliche Hürde, weil die großen Anbieter normalerweise keinen Zugang zu ihren Rechenzentren gewähren.

Darüber hinaus werden bei der Grundsatzentscheidung für oder gegen die Cloud völlig neue Aspekte wichtig. Ein Beispiel hierfür sind Risiken durch Imageverlust. So erscheint vielen Unternehmen eine Schlagzeile wie „Hacker laden vertrauliche Kundendaten der Firma XYZ aus der Cloud herunter“ weit schädlicher als eine Meldung über einen Einbruch in das firmeninterne IT-System. Aber auch die Frage, wie Cloud-Anwendungen in eine komplexe IT-Landschaft integriert werden sollen, ist nicht immer leicht zu klären. Beispiele aus Unternehmen zeigen, dass genauere Untersuchungen hierzu bei der Softwareauswahl oft zurückgestellt werden – oder die Unternehmen vertrauen den Zusicherungen von Implementierungsdienstleistern. Es erscheint aber sinnvoller, den Integrationsgrad schon zu Beginn des Auswahlprozesses zu klären. Denn je mehr die angestrebte Cloud-Lösung mit unterschiedlichen Systemen über Schnittstellen verbunden werden soll, desto höher ist im Allgemeinen die Hürde für ein SaaS-Angebot. Im Fall von CRM-Systemen kommen viele Unternehmen allerdings mit wenigen Schnittstellen aus. Eine möglicherweise schwierige Integration taugt also nicht per se als Argument gegen eine Cloud-Lösung.

Aus Architektursicht spielt schließlich auch eine Rolle, dass Cloud-Lösungen zu höherer Standardisierung beitragen können. Denn mehrere Erfahrungsberichte zeigen, dass beispielsweise Marketing und Vertrieb bereitwilliger einen durch SaaS vorgegebenen

höheren Standardisierungsgrad akzeptieren, wenn sie dadurch schnell eine flexibel skalierbare CRM-Lösung bekommen. Einige Unternehmen setzen Cloud-Lösungen deshalb bereits bewusst als Mittel ein, um strategische Konsolidierungsmaßnahmen rascher und mit höherer Akzeptanz umzusetzen.

4 Neue und veränderte Bewertungskriterien

Wenn eine SaaS-Lösung prinzipiell in Frage kommt, untersuchen Unternehmen die Anbieter in der nächsten Phase anhand der üblichen Kriterien: Übereinstimmung mit Geschäftsanforderungen, funktionale Bandbreite, Usability, Konsens mit Architekturprinzipien, etc. Bei einigen dieser Punkte ergeben sich allerdings für Cloud-Anwendungen neue Fragestellungen:

- Ist eine schnelle Umsetzung durch SaaS für uns ein Vor- oder Nachteil? Die kurze Zeit bis zum Rollout wird häufig als Argument für Cloud-Lösungen genannt. Denn diese Geschwindigkeit erlaubt es beispielsweise, CRM-Anwendungen in neuen, dynamischen Märkten zügig auszubauen und Vertriebsorganisationen flexibel einzubinden. In einigen Unternehmen ging dies aber so schnell, dass Mitarbeiter hinterher beklagten, es habe keine Zeit mehr gegeben, die anfänglichen Konzepte noch einmal anzupassen.

- Bekommen wir eine Komplettlösung? Mehrere Unternehmen betonen, wie wichtig es ist, ganzheitliche Lösungen zu evaluieren. Denn SaaS-Dienstleister bieten nicht immer alles aus einer Hand. So werden in China Anwendungen, Rechenzentren und Netzwerkservices oft nicht als Gesamtpaket bereitgestellt, sondern müssen von mehreren Dienstleistern bezogen werden. Zu einer CRM-Anwendung gehören auch Komponenten, die gewünschte Funktionalitäten abdecken, jedoch nicht Teil des SaaS-Standardprodukts sind (etwa eigenständige Call-Center-Lösungen). Zwar können solche Zusatzkomponenten später noch auf der Technologieplattform des Cloud-Anbieters entwickelt und bereitgestellt oder als On-Premise-Applikationen eingebunden werden. Aber das bedeutet in der Regel zusätzlichen Integrationsaufwand und höhere Betriebskosten. Auch die Integration der SaaS-Lösung in das Data Warehouse des Unternehmens oder Backup-Konzepte gehören zu einer Komplettlösung und sollten bei der Angebotsprüfung nicht übersehen werden.

- Wie innovativ ist der Anbieter? Für Endanwender erscheinen SaaS-Lösungen oft attraktiv, weil sie eine hohe Usability aufweisen und innovative Funktionalitäten schnell bereitstellen. So bot ein CRM- Dienstleister kürzlich seinen Kunden an, Sentiments aus aktuellen chinesischen und russischen sozialen Netzwerken in Neartime auszuwerten und in die Unternehmensprozesse einzubinden. Eine solche Fähigkeit zur Innovation haben Unternehmen bisher selten auf der Liste funktionaler und nicht-funktionaler Anforderungen an Anbieter von SaaS- und On-Premise-Systemen.

- Wie können interne Stakeholder von der Lösung überzeugt werden? In mehreren Fällen wurde die für den Nutzer attraktive Usability einer Groupware/E-Mail SaaS-Lösung verwendet, um das Change-Management für einen unternehmensweiten Ident-Service zu erleichtern.

- Brauchen wir die angebotenen Anpassungen und Erweiterungen? Wenn Unternehmen SaaS-Angebote mit funktionalen Anforderungen abgleichen, fällt meist auf, wie viele Optionen für Anpassungen und Erweiterungen die Anbieter sowie deren Implementierungs- und Softwarepartner offerieren (z.B. zu CRM-Analytics). Diese Optionen sollten die Unternehmen sorgfältig evaluieren, weil sie häufig im Widerspruch zu Standardisierungsbestrebungen oder einer gewünscht hohen Flexibilität stehen – und so dem eigentlichen Ziel der SaaS-Lösung zuwiderlaufen.

Bei den Interviews zeigte sich, dass Unternehmen für ein SaaS-Produkt in einem Anwendungssegment oft nur den jeweiligen Marktführer in Betracht ziehen. Dahinter steckte die Annahme, dass andere Anbieter wichtige Anforderungen bezüglich Sicherheit oder Compliance kaum genügen werden, wenn diese von den führenden Anbietern nicht erfüllt werden. Zudem sind Unternehmen meist nicht bereit, Pilotkunden kleinerer SaaS-Dienstleister zu werden, insbesondere weil die langfristige Marktfähigkeit dieser Anbieter noch nicht erwiesen ist.

Konsens aus den Interviews war zudem, dass maßgebende SaaS-Anbieter mit großen, dynamisch wachsenden Ökosystemen aus Systemintegratoren, unabhängigen Softwareherstellern (ISV) und weiteren Partnern punkten. Dadurch können sie Standardlösungen an die Anforderungen bestimmter Branchen und Regionen anpassen oder für Nischenmärkte erweitern. Zudem sichern die Netzwerke großer SaaS-Anbieter global verfügbare Ressourcen für Implementierungsprojekte, Rollout und Wartung. Viele SaaS-Innovationen kommen auch gar nicht vom Anbieter selbst, sondern von ISV oder aus dem Ökosystem des Anbieters. Unternehmen sollten allerdings darauf achten, dass solche Anpassungen, Erweiterungen und Innovationen schon in den Lizenzkosten berücksichtigt sind. Wer das versäumt, muss damit rechnen, später das vorgesehene Budget zu überschreiten.

Darüber hinaus gilt es zu berücksichtigen, dass Zusatzlösungen, die auf dem PaaS-Angebot eines Dienstleisters aufbauen, nicht immer auf dem gleichen technischen Stand sind, denn oft werden nicht die neuesten Plattformservices und Funktionen genutzt. Dies kann zu Komplikationen bei der Integration und beim Zusammenspiel verschiedener Lösungsbausteine führen. So wurde beispielsweise der Datenaustausch einer Zusatzanwendung für ein SaaS-CRM über ein proprietäres Protokoll gesteuert, obwohl die neueste PaaS-Version des Dienstleisters bereits standardisierte REST- und SOAP-Services zur Verfügung stellte. Erhöhte Integrations- und Betriebskosten waren die Folge.

5 Entscheidungskriterien für PaaS

Evaluieren Unternehmen PaaS als IT-Plattform, so lassen erste Interviews einen Unterschied zum SaaS-Fall vermuten: In einer ersten Phase der Analyse (siehe Abbildung 1) wird grundsätzlich entschieden, ob eine PaaS eingesetzt werden soll. Die nachfolgenden Schritte des Auswahlverfahrens konzentrieren sich dann häufig auf die Bewertung verschiedener PaaS-Angebote.

Die Trigger für die Einführung einer PaaS können vielfältig sein. Beispielsweise verspricht ihr Einsatz Vorteile beim Aufbau einer flexibel skalierbaren Plattform, die neue IT-Dienste besser unterstützen und deren Markteinführung wesentlich beschleunigen kann. Insbesondere lassen sich moderne DevOps-Prinzipien, wie kontinuierliche Delivery [JHDF11] und Integration, optimal umsetzten. In Bezug auf Virtualisierungsmöglichkeiten sind IaaS und PaaS vergleichbar, da PaaS auf IaaS aufbaut und somit den Virtualisierungs- und Automatisierungsgrad in Hinblick auf eine Industrialisierung der Lebenszyklen von Anwendungen erhöht (siehe hierzu u. a. das EU-Forschungsprojekt 4CaaSt zur Erstellung einer erweiterten PaaS-Cloud-Plattform [4C14]).

Folgende Fragestellungen fanden wir häufig bei unseren Analysen vor, wenn in Unternehmen initial über einen Einsatz von PaaS nachgedacht wird:

- Wie lässt sich der Automatisierungsgrad in der Softwareentwicklung erhöhen, um z.B. dynamisch skalierbare Umgebungen nahezu in Echtzeit bereit zu stellen? Mit PaaS können Anwendungsumgebungen in Form von vorkonfigurierten Anwendungscontainern, wie beispielsweise JavaEE Anwendungsservern, automatisch zusammen mit der Betriebssystemumgebung deployt werden. Idealerweise werden sämtliche Umgebungen von der Entwicklung über die Teststufen bis hin zur Produktion auf die gleiche Weise mit den gleichen Managementwerkzeugen aus identischen Quellen bereitgestellt. Darüber hinaus stehen Anwendungsmanagementdienste wie Health Monitoring, Resource Management, Skalierung und Deployment automatisiert zur Verfügung (siehe Elastic Platform in [Fe14]). Durch die Erhöhung des Automatisierungsgrads mit PaaS werden zudem Fehler bei der Umgebungsbereitstellung vermieden.

- Wie kann die Systemintegration verbessert werden? Die Kommunikation von Anwendungen untereinander erfolgt in einer PaaS über mitgelieferte Middleware-Komponenten - i.d.R. Messaging Systeme (siehe Message-Oriented Middleware in [Fe14]). Im Vergleich zu Installationen in oft unterschiedlichen Infrastrukturumgebungen entfallen innerhalb einer PaaS umgebungsspezifische Konfigurationen. Es werden modulare, lose gekoppelte Services erstellt, die sich gegenseitig konsumieren können. Nach außen lassen sich diese Dienste mithilfe der Prinzipien einer Service Orientierten Architektur (SOA) integrieren, um eine lose Kopplung zu bewahren (siehe [ARE10]).

- Wie lässt sich die Datenhaltung über Cloud-Ansätze optimieren? Bei dieser Fragestellung bedarf es einer sorgfältigen Evaluierung: Verwendet man Datendienste innerhalb eines virtuellen Containers, so lassen sich Datenbankzustände automatisch horizontal skalieren. Das führt allerdings zu einem Paradigmenwechsel in Bezug auf klassische Datenbank-zentrische Anwendungsarchitekturen, in welchen traditionell Datenbankzustände zentral in relationalen Datenbankmanagementsystemen (RDBMS) gehalten werden (siehe Stateless Component und Stateful Component in [Fe14]). Transaktionsmonitore und RDBMS gewähren dann die referentielle Integrität gemäß dem ACID-Prinzip. Eine Skalierung erfolgt aufwändig über die Infrastruktur. In verteilten Systemen bei einer horizontalen Verteilung von Datenbankzuständen kommt es zu einem Konflikt der unterschiedlichen Zugänglichkeitsaspekte, auch bekannt als CAP-Theorem [GL02]. Hier können immer nur zwei von drei Skalierungsaspekte gleichzeitig garantiert werden (Consistency vs. Availability vs. Partition Tolerance). Verschiedene NoSQL-Produkte unterstützen diese Aspekte mit unterschiedlichen Schwerpunkten. Die Architektur einer Cloud-Anwendung muss dann gemäß ihren Anforderungen ausbalanciert werden, um den benötigten CAP-Eigenschaften bei elastischen Skalierungen zu entsprechen [ACF12].

Stellt sich PaaS als vielversprechende Möglichkeit dar, zeigen Interviews mit Unternehmen, dass eine grundsätzliche Entscheidung für bzw. wider einer Einführung entsprechender Cloud Services oftmals wesentlich von organisatorischen Gegebenheiten abhängt. Stärker als die Auswahl einer Plattformtechnologie stehen Anpassungen der Arbeitsweise und eine bereichsübergreifende Zusammenarbeit bei der Einführung einer PaaS im Mittelpunkt der Diskussion. Differenzierende Faktoren sind weniger die Produktmerkmale der Plattformtechnologie und die mitgelieferten Standardframewoks, als vielmehr die vorhandenen Rahmenbedingungen innerhalb der eigenen Organisation. In diesem Zusammenhang wurden folgende Fragen hervorgehoben:

- Sind interne Strukturen und Schnittstellen zu IT-Infrastrukturkomponenten für eine vollständige Automatisierung hinreichend? Für die Implementierung einer PaaS ist die durchgängige Automatisierung der Deployment-Mechanismen des vollständigen Plattfom-Stacks erforderlich. Die betrifft insbesondere auch die Integration mit bestehender Abrechnungs- und Managementsoftware.

- Bieten bereits vorhandene Virtualisierungsansätze eine hinreichende Grundlage für eine weitere Dynamisierung? In der Regel führt der Weg zur Dynamisierung über die sukzessive Virtualisierung der Infrastruktur, die sich meist historisch entwickelt hat. In einem konkreten Fall wurden bestehende Ansätze als zu schwerfällig für weitere Dynamisierungsbestrebungen erachtet, was dann für die Implementierung eines neuen Cloud-Stacks sprach.

- Gibt es geeignete Sicherheitsstandards, die sich auch auf dynamisch skalierende Infrastrukturen anwenden lassen? Sicherheitsstandards sind noch oft auf statische Infrastrukturen ausgerichtet. Auch gibt es in Unternehmen mitunter wenige Erfahrungen mit virtuellen Entkopplungsmechanismen einer PaaS.

Daher ist die Zusammenarbeit mit den zuständigen Verantwortlichen für Sicherheit sorgfältig einzuplanen.

- Die Frage, ob mitgelieferten Anwendungsframeworks eine Rolle spielen, wurde als nachgelagertes Problem eingestuft.

Ist eine initiale Entscheidung für die Einführung einer PaaS gefallen, werden in nachfolgenden Phasen des Auswahlprozesses Detailfragen betrachtet. Beispiele hierfür sind:

- Welche Lifecycle-Services stehen zur Verfügung? Um den Industrialisierungs- und Automatisierungsgrad in einem Cloud-Stack zu erhöhen, bedarf es aufeinander abgestimmter Hardware- und Softwarebausteine. Dies betrifft insbesondere auch Lifecycle-Aspekte. Beispielsweise sollten alle abhängigen Plattformkomponenten in definierten Zyklen über sogenannte Matrix Updates zusammen gewartet werden.

- Wie kann eine optimale Integration von Hard- und Softwarebausteinen gesichert werden? Oft reicht es nicht, auf existierenden Virtualisierungsansätzen aufzubauen. Im Gegenteil, es besteht dann die Gefahr, dass sich durch zusätzliche PaaS-Komponenten die Komplexität der Gesamtlandschaft eher erhöht. Auch wird in vielen Unternehmen im Entwicklungs- und Produktionsbereich separat virtualisiert. Ferner können bestehende Lizenzmodelle zu dezidierten Umgebungen zwingen. Derzeit entwickelt sich ein Markt für vorintegrierte Technologiekomponenten (Converged Infrastructure, siehe z.B. [Be13]) die den Aufbau von Cloud-fähigen Rechenzentren erleichtern können. (Für einen systematischen Einstieg sowie Kriterien für die Auswahl geeigneter Cloud Service Modelle siehe z,B. [Ka14])

- Welche Vorteile bieten Open Source-Plattformen? Open Source-Lösungen basieren meist auf offenen Standards. Allerdings ist davon auszugehen, dass Modifikationen bzw. Ergänzungen notwendig sind. So müssen insbesondere Schnittstellen für eigene Authentifizierungsmechanismen und Self Service-Verfahren implementiert werden.

- Wodurch zeichnen sich kommerzielle Produkte aus? Hier ist insbesondere die Interoperabilität vorhandener APIs einer der wesentlichen Faktoren. Diese ist innerhalb des Stacks eines Lieferanten typischerweise gegeben, die Integration mit vorhandenen Management- und Abrechnungssystemen kann jedoch aufwändig werden.

Es lässt sich feststellen, dass PaaS kein Thema für eine einfache, vergleichende Plattformauswahl ist. Stattdessen wird meist in einem ersten Schritt geklärt, welche Möglichkeiten und Ziele für eine bestehende Organisation sinnvoll sind.

6 Zusammenfassung

Cloud-Services verändern die Auswahlverfahren für IT-Lösungen. Dies ist das Fazit der vorliegenden Momentaufnahme. Es zeigt sich, dass dem typischen Muster für die Evaluation von IT-Anwendungen und -Plattformen eine Phase vorgeschaltet wird, in der man grundsätzlich darüber entscheidet, ob Cloud-Services in Betracht kommen oder nicht. Nachfolgende Entscheidungsschritte enthalten dann zum Teil geänderte Bewertungskriterien, die auf die unterschiedlichsten Aspekte von Cloud-Angeboten eingehen.

Eine weiterführende Analyse zur Erhärtung und Detailierung der hier aufgezeigten ersten Beobachtungen erscheint sinnvoll. Sowohl für Anwendungsunternehmen als auch für Anbieter von IT-Dienstleistungen ist ein Verständnis der sich stetig ändernden Bewertungsprozesse und -kriterien von IT-Lösungen wichtig.

Literaturverzeichnis

[ACF12] Andrikopoulos, V.; Fehling, C.; Leymann, F.: Designing for CAP – The Effects of Design Decisions on the CAP Properties of Cloud-native Applications, Proceedings of the 2nd International Conference on Cloud Computing and Service Science, CLOSER, 2012; S. 365-374.

[ARE10] Simon, A.; Rischbeck, T.; Erl, T.: Modern ESB Architecture for SOA, Pearson Education Schweiz, 2010.

[Be13] Bernard, A.: Is Converged Infrastructure the Future of the Data Center?, CIO, 2013, http://www.cio.com/article/730423/Is_Converged_Infrastructure_the_Future_of_the_Data_Center_, abgerufen am 16.04.2014.

[BWV07] Becker, J.; Winkelmann, A.; Vering, O.: Softwareauswahl und -einführung in Industrie und Handel, Springer, Heidelberg, 2007; S. 61-108.

[GL02] Gilbert, S.; Lynch, N.: Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services, ACM SIGACT News 33, 2002; S. 51-59

[Gr10] Gronau, N.: Enterprise Resource Planning, Oldenbourg Wissenschaftsverlag, München, 2010.

[Fe14] Fehling, C. et al.: Cloud Computing Patterns: Fundamentals to Design, Build, and Manage Cloud Applications, Springer, Wien 2014.

[Ka14] Kavis, M.: Architecting the Cloud: Design Decisions for Cloud Computing Service Models (SaaS, PaaS, and IaaS), John Wiley & Sons, Hoboken, 2014.

[JHDF11] Humble, J; Farley, D.: Continous Delivery: reliable software releases through build, test, and deployment automation, Addison-Wesley, Boston, 2011.

[VHH12] Vossen, G.; Haselmann, G.; Hoeren, T.: Cloud Computing für Unternehmen, dpunkt.verlag, Heidelberg, 2012.

[4C14] EU-Forschungsprojekt 4CaaSt zur Erstellung einer erweiterten PaaS-Cloud-Plattform, http://www.4caast.eu/, abgerufen am 25.04.2014.

Cloud-gestützte Supply-Chain-Architekturen – Entwicklung eines Referenzmodells auf Basis einer systematischen Literaturanalyse

Andreas Jede, Prof. Dr. Frank Teuteberg

Fachgebiet Unternehmensrechnung und Wirtschaftsinformatik
Universität Osnabrück
Katharinenstraße 1
49074 Osnabrück
ajede@uni-osnabrueck.de
frank.teuteberg@uni-osnabrueck.de

Abstract: Die spezifischen Eigenschaften des Cloud Computing (CC) können Geschäftsprozesse im Supply Chain Management (SCM) unterstützen. Der vorliegende Beitrag bietet eine interdisziplinäre Literaturanalyse. Dabei werden 102 Publikationen, die den CC-Einsatz in SCM-Bereichen thematisieren, im Hinblick auf Problemfelder, IT-Artefakte sowie Erfolgs- und Einflussfaktoren analysiert. Die gewonnenen Erkenntnisse dienen der Konzeption eines Referenzmodells für Cloud-gestützte SCM-Architekturen. Insgesamt zeigen die Ergebnisse, dass CC und SCM zunehmend miteinander verzahnt werden. Die Praxis hat dieses Potenzial nur unvollständig erkannt und nutzt es nur ansatzweise. Aus Sicht der Wissenschaft fehlt es vor allem an theoretisch fundierten und empirischen Arbeiten.

1 Einleitung

Nach heutigem Verständnis ist die IT ein elementarer Erfolgsfaktor beim Streben der Unternehmen nach Effizienz entlang der gesamten Supply Chain (SC) [Ce12, SMW11]. Vor allem für die dezentrale Natur von globalen SCs bietet das CC-Paradigma signifikante Vorteile. Komplette Services aus gekapselten Funktionen, die skalierbar und elastisch sind, können ortsungebunden und jederzeit verfügbar bereitgestellt werden [Ba10]. So gewinnen digitale SC-Prozesse u.a. an Stabilität und Flexibilität [HWK12, Ce12].

Der aktuelle Stand in der CC-Forschung befindet sich jedoch in Theorie und Praxis im Frühstadium [Ma11]. Gleichwohl herrscht allgemeiner Konsens, dass CC sowohl eine informationstechnische als auch eine betriebswirtschaftliche Tragweite attestiert werden, erwähnt seien hier z.B. die Kosteneffekte [Ma11]. Derzeit sehen sich Unternehmen, die stets beide Aspekte bei Selektion, Adaption und Nutzung von CC im SCM berücksichtigen sollten, mit einem undurchsichtigen Gebilde von Service-Möglichkeiten konfrontiert [WRZ12]. Interdisziplinäre Handlungsrahmen für spezifische Unternehmensbereiche [HWK12], wie für das SCM [Le10], sind in der wissenschaftlichen Literatur rar. Ebenso

sind die Abhängigkeiten und Relationen zwischen SCM- und CC-Elementen weitestgehend unerforscht. Aus der Motivation heraus, den bisherigen Forschungsstand zusammenzuführen und weitere Forschungslücken aufzudecken, wurde im Rahmen dieses Beitrags eine systematische Literaturanalyse durchgeführt, die sich in der Schnittmenge der Disziplinen Wirtschaftsinformatik und SCM bewegt. Auf Basis dieser Analyse werden die identifizierten Erkenntnisse in ein Referenzmodell für CC-gestützte SCM-Prozesse überführt. Die dem Beitrag zugrunde liegenden Forschungsfragen lauten:

1) Wo liegen die aktuellen (Forschungs-) Schwerpunkte beim Einsatz von CC im SCM?
2) Welche Elemente sollten für ein Referenzmodell berücksichtigt werden, das Cloud-gestützte SCM-Prozesse adressiert?

Die Struktur der Arbeit ist wie folgt: Nach der Einleitung werden in Kapitel zwei die grundlegenden Begriffe erläutert. Das dritte Kapitel beschreibt das methodische Vorgehen der Literaturanalyse, deren Ergebnisse in Kapitel vier präsentiert werden. Danach wird in Kapitel fünf das Referenzmodell entwickelt. Die Arbeit schließt mit einem Fazit.

2 Begriffliche Grundlagen

Mentzer et al. [Me01] differenzieren beim SCM zwischen drei konstitutiven Verrichtungsebenen und definieren SCM wie folgt: „a set of … entities (organizations or individuals) directly involved in the upstream and downstream flows of 1) products, services, 2) finances, and/or 3) information from a source to a customer." Während Materialflussaktivitäten hierbei das Ziel verfolgen, sämtliche Prozessschritte der Wertschöpfungskette auf effiziente Materialflüsse auszurichten, behandeln Finanz- und Informationsflüsse die vertikale und horizontale Vernetzung von Prozessen. Speziell bei der Steuerung der Informationsflüsse steht die inner- und überbetriebliche, bedarfsorientierte Informationsversorgung im Zentrum, wobei die Herausforderung in der Selektion, Adaption und Nutzung von zweckmäßiger IT liegt.

CC kann die dritte Ebene unterstützen und wird vom „National Institute of Standards and Technology" wie folgt definiert [MG11]: „a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction". Dabei stellt CC keine neuartige Technologie dar, sondern ein neues Paradigma für IT-Prozesse [YBS08], indem es bekannte Einzeltechnologien konsequent verknüpft [Le10]. Ferner wird in der Literatur zwischen drei „as a Service"-Modellen unterschieden [HWK12]: Infrastructure, Platform und Software (IaaS, PaaS, SaaS). Desweiteren existieren vier CC-Ausprägungen [HWK12, Ma11]: Während beim „Public CC" ein externer Anbieter seine Services über das Internet zur Verfügung stellt, werden die Services beim „Private CC" unternehmensintern angeboten, was implizit einen notwendigen, eigenen CC-Betrieb voraussetzt. „Community CC" ist in Betracht zu ziehen, falls eine Gruppe von Unternehmen, die gemeinsame Interessen verfolgt und vergleichbare Sicherheitsstandards pflegt, den CC-Betrieb eigenständig kontrollieren will. Beim „Hybrid CC" (Kombination aus Private/Public CC) werden sensible Informationen intern abgegrenzt und unkritische Services und Daten in die Obhut eines externen Anbieters übertragen.

3 Methodisches Vorgehen

3.1 Systematische Literaturanalyse

Der vorliegende Beitrag basiert auf der Methode der systematischen Literaturanalyse, die sich in der wissenschaftlichen Forschung zweckdienlich zur Identifikation des aktuellen Forschungsstands erwiesen hat [Fe06]. Hierbei wird die selektierte Literatur qualitativ und quantitativ analysiert.[1] Ferner orientiert sich der Beitrag am Rahmenkonzept von Dibbern et al. [Di04]. Aus dem in Tabelle 1 dargestelltem Konzept leitet sich die gestalterische Umsetzung ab. Die Durchführung der ***qualitativen Analyse*** erfolgt in enger Anlehnung an das Fünf-Stufen-Modell von Fettke [Fe06]: 1) Problemformulierung, 2) Literatursuche, 3) -auswertung, 4) -analyse sowie 5) Interpretation. Dabei liegt der Schwerpunkt in der Zuweisung von Problemen, Inhalten und Erfolgsfaktoren eines jeden Artikels in ein vordefiniertes Raster. Hierbei ist hervorzuheben, dass ausgehend von der Problemdarstellung gemäß der Einleitung die Suche nach relevanten Arbeiten auf Basis von Keywords durchgeführt wurde (Fussnote 1 enthält den Link zur Keyword-Liste). Um die besten Publikationsorgane zu identifizieren, wurden drei vordefinierte Ranglisten herangezogen („WI“, WIRTSCHAFTSINFORMATIK; „AIS“, Association for Information Systems; „Transport, Logistics and SCM“, Universität Sydney). In Summe besteht die Auswahl aus 63 Zeitschriften (33 aus der WI und 30 aus dem SCM). Darüber hinaus erweitern die Konferenzen „ECIS“, „ICIS“ und „WI“ die Selektion, die laut der „Wissenschaftlichen Kommission für Wirtschaftsinformatik (WKWI)“ zu den drei größten Konferenzen zählen. Die Konferenz „AMCIS“ ist in der WKWI-Liste zwar nur an sechster Stelle geführt, jedoch besticht die Konferenz durch überproportional viele Beiträge im relevanten Themengebiet, sodass diese die Zeitschriftenauswahl komplettiert. Auf dieser Basis wurden inklusive Vorwärts- und Rückwärtssuche [WW02] 102 relevante Artikel identifiziert. Ziel der ***quantitativen Inhaltsanalyse*** ist die Identifikation detaillierter Zusammenhänge zwischen den Termini, die für CC und SCM relevant sind, sowie die Entlarvung verborgener Merkmale [Su01]. Hierfür wird die Software „WordStat“ von Provalis genutzt. Die 102 selektierten Artikel werden dabei in die Software kopiert und in einem Vier-Stufen-Schema bearbeitet. Während in Stufe eins die manuelle Normierung der Schreibweise und der Silbentrennung sowie die Entfernung aller Klammern erfolgen, beinhaltet Stufe zwei die systemgestützte Lemmatisierung, die Stammformreduktion, sowie das Exkludieren unnötiger Füllwörter. In den Stufen drei und vier dient die gewonnene Datenbasis der Analyse und der Interpretation.

Perspektive	Motiv	Forschungsfragen	Kapitel
Motivation	Warum?	Warum CC im SCM?	1
Fokus	Wie? Welche? Was?	Was sind die aktuell diskutierten Problemstellungen?	4.1
		Welche (IT-)Artefakte wurden bislang entwickelt?	4.1
		Wie sind die derzeitigen Zusammenhänge zwischen SCM und CC?	4.2
		Welche Elemente sind für ein Referenzmodell zu berücksichtigen?	5
Implikationen	Ergebnis	Was sind die Implikationen für Wissenschaft und Praxis?	6

Tabelle 1: Rahmenkonzept der Literaturanalyse

[1] Ausführliche Darstellungen zur Vorgehensweise und Methodik der qualitativen sowie der quantitativen Literaturanalyse sind unter (http://www.uwi.uni-osnabrueck.de/Informatik14_Anhang.pdf) verfügbar (siehe Anlage IIb und IIc).

3.2 Themenverwandte Veröffentlichungen

Im Rahmen der Literaturauswertung wurden lediglich fünf Artikel in den angeführten 63 Zeitschriften identifiziert, die ebenfalls eine Literaturanalyse in inhaltsähnlichen Themenfeldern abhandeln. Diese sind in der Tabelle 2 aufgeführt und nach den Kriterien Anzahl berücksichtigter Artikel („#“), „Forschungsfrage“ und „Ergebnisse“ gegliedert. Dieser Beitrag grenzt sich im Wesentlichen durch zwei Faktoren stark von den unten aufgeführten Veröffentlichungen ab. Nach Kenntnis der Autoren existiert bis dato keine Literaturanalyse, die konsequent das Themenfeld CC im SCM fokussiert. Ferner führen die Auswertungen zu einem ersten Referenzmodell für Cloud-gestützte SCM-Prozesse, welches systematisch auf Basis einer Literaturanalyse entwickelt wurde.

Ref.	#	Forschungsfrage	Ergebnisse
[HWK12]	60	Was sind die relevanten Aspekte im CC-Paradigma aus Sicht des nutzenden Unternehmens?	Literaturerkenntnisse werden in vier Dimensionen zusammengetragen: 1) Allgemeine CC-Eigenschaften 2) Determinanten, die die Adaption beeinflussen, 3) Notwendige staatliche Regelungen 4) Grds. Auswirkungen auf unternehmerische Geschäftstätigkeit
[GL08]	174	Welchen Einfluss haben Internet-basierte Technologien auf die Geschäftsprozesse im SCM?	Wettbewerbsvorteile für Unternehmen entstehen insbesondere beim elektronischen Bestellprozess, der Bestellerfüllungsquote sowie beim Informationsfluss
[STC12]	45	Was sind die Vorteile eines CC-basierten ERP-Systems?	Insbesondere kleine und mittlere Unternehmen profitieren aufgrund niedriger Investitionen und IT-Ressourcen vom CC-ERP-System
[FBW13]	28	Was ist der Einfluss von SaaS-basierten Geschäftsprozessen auf die operative Agilität?	„SaaS“-basierte Prozesse steigern zwar die operative Agilität, jedoch nicht unmittelbar, sondern z.B. durch verbesserte ERP- und Ablaufunterstützungssysteme
[Eg13]	17	Welche Auswirkungen haben neue IT-Systeme auf die horizontale Kollaboration in SCs?	Informationsaustausch und Transparenz der IT-Systeme sind die wichtigsten Faktoren für eine horizontale Kollaboration

Tabelle 2: Themenverwandte Veröffentlichungen

4 Ergebnisse der Literaturanalyse

4.1 Derzeitige Problemstellungen und die dazu entwickelten Artefakte

Beim Zusammentragen der Problemstellungen sind explizit nur die Inhalte aus den Arbeiten extrahiert worden, die den CC-Einsatz im SCM im Kern behandeln. In der nachstehenden Tabelle 3 sind die Schwerpunkte der Arbeiten und die dazugehörigen, bewiesenen Erfolgsfaktoren mit verwendeter Methode dargestellt. Ferner wurden die entwickelten Artefakte der selektierten Arbeiten untersucht. In dieser Arbeit werden substanzielle Bestandteile eines spezifischen Grundlagenwissens als Artefakte bezeichnet [HC10]. Die Struktur der Artefakte und die Definitionen orientieren sich dabei an March und Smith [MS95]. Tabelle 4 zeigt Beispiele von CC-Artefakten im SCM und die dazugehörigen Referenzen. Des Weiteren offenbart eine Unterteilung aller angewandten Forschungsmethoden und IT-Artefakte in das „konstruktionswissenschaftliche Paradigma“ und das „verhaltenswissenschaftliche Paradigma“ nach Wilde und Hess [WH07] einen deutlichen Überhang des erstgenannten Paradigmas.

Das Verhältnis ist ein Indiz für das frühe Stadium der Forschung im Themengebiet, da verhaltensorientierte Arbeiten das Vorhandensein von konstruierten Artefakten als Untersuchungsbasis für verhaltenswissenschaftliche Studien voraussetzen und vermehrt auf fortgeschrittenen Forschungsgebieten vorzufinden sind [MT09]. Um die Zusammenhänge auf dem untersuchten Forschungsgebiet noch besser nachvollziehen und beurteilen zu können, sollten unterrepräsentierte Methoden wie z.B. Referenzmodellierung oder Prototyping sowie Multi-Methoden-Ansätze künftig stärkere Berücksichtigung finden.

Problemstellung [Referenzen] **Mit der jeweiligen Problemstellung im Zusammenhang stehende, nachgewiesene Erfolgsfaktoren [Referenzen [Methoden]]**
Wie können Interoperabilität, Koordination, Kollaboration und Informationsaustausch entlang der SC durch CC verbessert werden? Wie kann die Transparenz erhöht werden? Wie sollte dabei die Integration diverser Enterprise-Applikationen konzipiert sein? [Be11b;Ce12;KR13;KDS10;SÖ13;RA10; SMW11;Eg13]
• Steigert die Transparenz der Hubs in der SC durch angemessene Standardsetzung [SMW11 [Fallstudie]]
Wie können intelligente Transportsysteme mittels CC verbessert werden? Wie sind sichere „Ad hoc vehicular clouds“ aufzubauen und wie sieht die Unternehmensanbindung aus? [Wa10;Ya13;EOY10;OKA11]
• Schützt vor Sicherheitsangriffen durch Schließen von Sicherheitslücken mittels konzipiertem Algorithmus, indem dieser die Authentifizierung sicherstellt [Ya13 [Fallstudie]]
Wie kann die Koordination von Frachttransport und Handling mittels CC optimiert werden? Wie sind Bearbeitungsaufträge für einen Daten-Layer in einer Zulieferkette zu gestalten? [LK11;LKS11;MDD12]
• Optimiert Handling- und Transportkoordination mittels eines Algorithmus [LK11 [Simulation]]
• Erhöht den Zugang zu externen Handling-Dienstleistern und beseitigt Engpässe [LKS11 [Simulation]]
• Erhöht die Auslastung der Ressourcen durch optimierte Verteilung der Aufträge [MDD12 [Experiment]]
Welche Möglichkeiten bietet CC für voneinander abhängige, übergreifende Produktionsprozesse? Wie können hier Reaktionsfähigkeit und Ressourcenauslastung in der SC gesteigert werden? [FBW13;HYZ13; QCY09]
Wie kann der elektr. Beschaffungsprozess in Netzwerken mittels CC unterstützt werden? Was sind spezifische Anforderungen in hybriden Absatz-Beschaffungsnetzen und -Verbindungen? [Be12;Sc12a;ST13;CTH13]
• Steigert den Automatisierungsgrad im mehrstufigen elektr. Bestellprozess [ST13 [Simulation]]
Wie ist ein geeignetes Informationssystem innerhalb der Unternehmensarchitektur aufzubauen, das RFID und CC sinnvoll entlang der Wertschöpfungskette integriert? [Zi10; Li12]
• Reduziert Verzögerungszeiten durch kürzere Netzwerkpfade und senkt durch Skalierung Infrastrukturkosten; Effekte in starker Abhängigkeit von der Anzahl der genutzten Datenzentren [Zi10 [Simulation]]
Unter welchen Voraussetzungen ist „Public/Private“ CC für ein SCM-lastiges Unternehmen sinnvoll? Bei welchen Prozessen im SCM sollte vom CC absehen / beim Eigenbetrieb verblieben werden? [MHQ13;Xi11]
Wie können Agilität und Flexibilität bei SC-Geschäftsprozessen gesteigert werden? [APF13;Le11]
• Erhöht die Agilität durch forcierten und koordinierten Wissensaustausch [Le11 [Fallstudie]]
Was ist die Erwartungshaltung und wie ist die Akzeptanz vor, während und nach der CC-Implementierung im SCM? Welchen Einfluss haben IT-Komplexität und Vielfalt auf die CC-Erwartung [ARD12;Au10]
Wie ist eine komplette CC-basierte „Business-Intelligence-Lösung“ für Warenströme zu konzipieren, die die Systeme „Sensoren-Technologie“ und „RFID“ vereint und mobil zur Verfügung stellt? [BK11;TW10]
Wie sind sichere CC-Geschäftsprozesse unternehmensintern und -übergreifend zu gewährleisten? Welche Auswirkungen haben die notwendigen Sicherheitsstandards auf die SC-Performance? [DG13,DCW11]
Welche Unterstützung bietet CC beim Wissensmanagement im SCM von Bildungseinrichtungen? [CW13a]
• Erhöht zum einen Lernmotivation und -zufriedenheit und stellt den Dozenten zum anderen ein effektives Lehrinstrument zur Verfügung [CW13a [Fallstudie]]
Wie kann das IT-Niveau bei Logistikdienstleistern mittels CC gesteigert werden? [LWC12;Do10;RTD11]
Was sind die wesentlichen Aspekte einer digitalen Strategie für SCM-Geschäftsprozesse? [Bh13;MK12]
Warum streben Unternehmen nach Datenkontrolle in unternehmensübergreifenden Informationssystemen und wieso ist dieses Bestreben für die Koordination der Geschäftsprozesse hinderlich? [CR13]
Welchen Einfluss haben Unsicherheiten bei Geschäftsprozessen, im Unternehmensumfeld und in der unternehmensübergreifenden Koordination auf die Implementierung von CC-Technologien? [Ce12]

Tabelle 3: Wesentliche Problemstellungen und Erfolgsfaktoren in der Literatur

Artefakt & Definition		Beispiele	Referenzen
Konstrukt	Verwendete Konzeptionierung zur Beschreibung eines Problems	Stärken	DH13;HG12; Ma11;Nu11
		Schwächen	Ma11; Nu11
		Chancen	BH11;Be11a;Cu12;FBW13;HWK12;Sc12b; Le10;Ma11,STC12;TS11;WRZ12;Bh13
		Risiken	CW13a;HWK12;Ma11;WRZ12;Ya13;BM13
Modell	Modellierung von Beziehungen zwischen Konstrukten	Konzeptionelle Modelle	Be11a;Be12;Be11b;DD13;Li12;Re12; Sc12a;Ta11;Fa13;Ne13;
		Analytische Modelle	APF13;BCM09;DCB10;MHQ13;MTZ12
		Architekturmodelle	DS13;FKB12;MDD12;Sc12b;TS11;Zi10
Methode	Abfolge von Aktionen zum Verrichten einer Tätigkeit	Rahmenkonzepte	Be11a;BH11;DS13;QCY09;Sc12a;ST13; VW12;WRZ12;Bh13;Sa13
		Hypothesenbildung	ARD12;Be09;BKH10;BH11;Ce12; GKW12;HG12;HB12;MK12;Nu11;RTD11; LPL13;CR13;GSR13;Ba13;LCC13;CTH13
		Algorithmen	FKB12;LK11;Ya13;ZTM11
		Richtlinien, Standards	BKH10;KR13;KDS10;LKS11;SMW11;ST13
Realisation	Realisierung eines Artefakts in seiner Umgebung	Netzwerkoptimierungssysteme	FKB12;ST13;Zi10;ZTM13
		Entscheidungsunterstützungs-systeme	DH13;FKB12;ZTM13
		Logistik-Planungssysteme	BK11;LKS11;KDS10
		Transportinformationssysteme	KK11;Wa10
		Distributionssysteme	MDD12;KDS10
		ERP-Systeme	Bo14;SD14;CW13a;CW13b

Tabelle 4: Entwickelte Artefakte

4.2 Wie sind die Zusammenhänge zwischen CC und SCM derzeit ausgestaltet?

Mittels der quantitativen Inhaltsanalyse können Worthäufigkeiten und Kookkurrenzen der Datenbasis bestimmt werden. Während sich die Worthäufigkeit aus der absoluten Summe ergibt, errechnet sich die Kookkurrenz in der Linguistik aus dem gemeinsamen Auftreten zweier Wörter (1:1) [TSK05, S.65-72]. Im Rahmen dieser Arbeit sollen Kookkurrenzen auf Paragraphenbasis ermittelt werden. Zur Berechnung dieser Interrelationen wurde der Jaccard-Index[2] verwendet. Aus diesen Daten können Dendrogramme abgeleitet werden, die die vorgegebenen Wörter in Verbindung mit Variablen abbilden (1:n). Die Ergebnisse können durch das Verfahren der multidimensionalen Skalierung (MDS) zu einer 2-D Karte visualisiert werden [Pr10]. Die Größe der Kreise korrespondiert mit der Worthäufigkeit. Die Distanz zwischen den Kreisen entspricht der Häufigkeit des gemeinsamen Auftretens von Wörtern. Während die Farben übergeordnete Gruppen markieren, repräsentieren die Linien besonders starke Abhängigkeiten zwischen den Kreisen. Obwohl alle selektierten Arbeiten im Überschneidungsbereich von CC und SCM liegen, ist eine klare Trennung der Bereiche in Abbildung 1 vorzufinden. Diese Tatsache führt zu der Feststellung, dass die Forscher beider Bereiche zwar Vorteile in einer Verknüpfung vermuten, da sie gerade diese Verknüpfung untersuchen, aber die Schwierigkeiten offensichtlich in der Internalisierung der Spezifikationen des Konterparts liegen. Wenig überraschend bestätigt die Karte den Fokus dieser Arbeit, indem eine Verbindung zwischen CC-Services und SCM durch ein Dreieck geschlagen wird, das aus Business (BUS), PROCESS und SYSTEM besteht.

[2] Weitere Details zum Vorgehen und zur Berechnung der Kookkurrenz sind unter (http://www.uwi.uni-osnabrueck.de/Informatik14_Anhang.pdf) verfügbar (siehe Anlage IIc).

Die Unterstützung durch CC bietet demnach Vorteile bei Faktoren wie Kosten, Flexibilität, Performance, Standardisierung, Informationsaustausch und Kollaboration (COST, FLEXIBL, PERFORM, STANDARD, SHAR, COLLABOR). Risiken hingegen werden der Datensicherheit und der -kontrolle beigemessen (DATA, SECUR, CONTROL). Ferner umkreisen vor allem technische Aspekte das CC-Paradigma (z. B. APPLIC, NETWORK). Diese werden im Zusammenhang mit SCM undifferenziert zum Terminus TECHNOLOGI aggregiert. Um die vermuteten Vorteile in Theorie und Praxis zu belegen und von diesen in realiter nachhaltig zu profitieren, sollten künftige Arbeiten die Zusammenhänge der Spezifika beider Bereiche stärker in den Fokus rücken. Hinsichtlich der Service-Modelle wird SaaS am häufigsten erwähnt. Zahlreiche Arbeiten der Stichprobe differenzieren zwar zwischen den Service-Modellen, richten ihre Forschung aber insbesondere auf SaaS aus. Das Übergewicht könnte damit begründet werden, dass SaaS für den Endnutzer i.d.R. als „Front-End“ dient [YBS08, HYZ13] und somit als einzig „sichtbarer“ CC-Kontakt. Grundsätzlich enthält die Stichprobe selten vergleichende Aussagen zu Servicemodellen und verbleibt in vielen Fällen rein argumentativ-deduktiv. Einzig Hoberg et al. [HWK12] verweisen kurz auf die „SaaS“-lastige CC-Literatur.

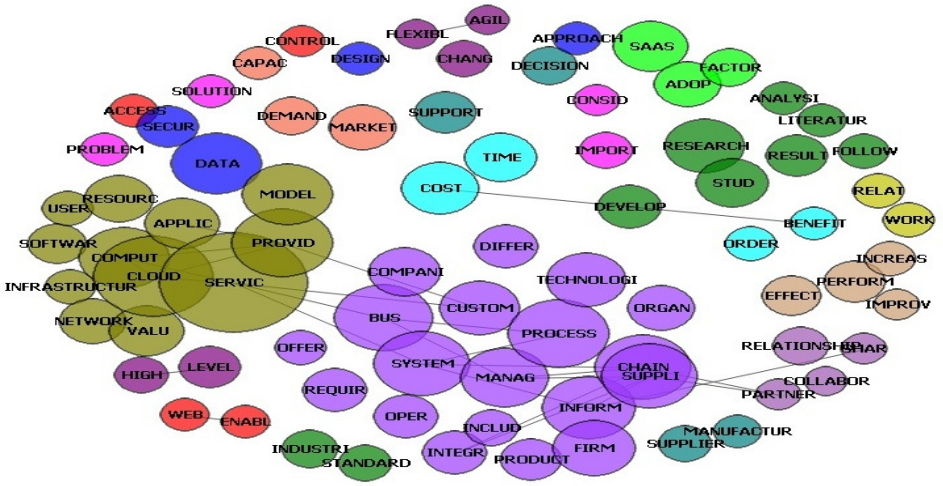


Abbildung 1: 2-D Karte der quantitativen Inhaltsanalyse

5 Referenzmodell für Cloud-gestützte SCM-Architekturen

Nach Kenntnis der Autoren existieren bislang keine systematisch auf Basis einer Literaturanalyse qualitativ und quantitativ hergeleitete Referenzmodelle für den CC-Einsatz im SCM. Auf diese Forschungslücke zielt das nachstehende Modell in Abbildung 2 ab. Die wichtigsten Erkenntnisse, Strukturen und Abhängigkeiten, die aus der qualitativen und der quantitativen Literaturanalyse gewonnen wurden, sind zu Elementen und zu Verbindungen zwischen den Elementen im Zusammenspiel beider Bereiche gebündelt worden. Aufgrund des möglichen tiefgreifenden Einflusses von CC auf die Prozesse eines Unternehmens bedarf es hierbei einer ganzheitlichen Unternehmensbetrachtung.

Ferner unterliegen dem Gesamtmodell einzelne Teilmodelle, die in grau gehalten sind. Innerhalb der Teilmodelle sind die Kookkurrenzen zweier Elemente in Klammern aufgeführt, um die derzeitige Relevanz des einen Elements für das andere zu verdeutlichen (je höher die Kookkurrenz desto höher die Relevanz). Aus Gründen der Nachvollziehbarkeit und der Transparenz hinsichtlich der Kookkurrenzen ist die Abbildung 2 in Originalsprache der Literaturauswahl dargestellt. Zunächst werden die acht zugrunde liegenden Teilmodelle in Kürze dargestellt: Für die Teilmodelle ***Strategy*** und ***Stakeholder*** wurde das „triple bottom line concept" [CR08] herangezogen, welches unterstellt, dass ökologische und soziale Verantwortung zu langfristigem und nachhaltigem Erfolg führen. Die definierte Strategie besteht aus Zielen, Aktionsplänen und Leistungsmessungen zur Zufriedenstellung interner und externer Stakeholder. Das ***Organization Model*** determiniert die organisatorischen Einheiten, die zur Durchführung eines Geschäftsprozesses inkludiert, exkludiert oder etabliert werden müssen. Das ***Process Model*** hingegen adressiert alle Geschäftsprozesse innerhalb der SC und nimmt eine zentrale Rolle im Gesamtmodell ein. Hierbei sind Qualität und Schnelligkeit insbesondere von der Effizienz der IT-Unterstützung abhängig [Ce12]. Im Referenzmodell werden die Geschäftsprozesse von den SC-Partnern direkt und indirekt (als strategiebestimmende Stakeholder) beeinflusst. Die Aufgabe des ***IT Architecture Model*** liegt einzig in der Unterstützung der Geschäftsprozesse. Vor dem Hintergrund von CC sind Schnittstellen und Konfigurationen zwischen dem eigenem Unternehmen, dem SC-Partner und dem CC-Anbieter von besonderer Bedeutung. Da der Gesamterfolg einer SC stark von den Interaktionen der SC-Partner abhängt, finden diese Beziehungen im Teilmodell ***SC Cooperation*** entsprechendes Gewicht. Darin sind diejenigen (indirekten) CC-Einflussfaktoren subsummiert, die auf wechselseitiger Abstimmung beruhen. Die in der 2-D Karte aufgeführten Faktoren INFORM SHAR und COLLABOR wurden auf Basis weiterer Untersuchungen in der Literatur zu den unmittelbaren CC-Vorteilen Koordination, Kollaboration und Wissensaustausch spezifiziert. Das ***CC Architecture Model*** zeigt die speziellen Eigenschaften und die zur Verfügung stehenden Ressourcen zur Unterstützung der Unternehmensarchitektur. Die fehlende Verbindung zum Geschäftsprozessmodell unterstreicht den indirekten Einfluss von CC auf die SCM-Geschäftsprozesse. Schließlich wird ein ***KPI-Model*** benötigt, das den organisatorischen Rollen unterliegt und die vier wichtigsten direkten Einflussfaktoren berücksichtigt (basierend auf der Kookkurrenz): Während die Faktoren Kosten und Sicherheit selbsterklärend sind, zielt der Faktor „IT Performance" auf den Leistungsbereich und die Funktionalität des CC-Angebots ab. Darunter fällt auch die Fähigkeit des Anbieters, mit Innovationen aufwarten zu können sowie Services den Kundenwünschen entsprechend zu optimieren. Unter Flexibilität sind Größen wie „Service Delivery Model", Ubiquität, Skalierbarkeit und Kompatibilität sowie Automatisierungsgrad subsummiert, womit die Anpassungsfähigkeit an veränderte Rahmenbedingungen abgebildet werden soll. Zur wirkungsvollen Anwendung des KPI-Modells wird beispielsweise eine Balance Scorecard vorgeschlagen [LPL13].

In Abbildung 2 ist eine hohe Kookkurrenz bei *Change* im Prozessmodell ersichtlich. Demnach propagiert die Literatur grundsätzliche Prozessveränderungen und diskutiert in diesem Zusammenhang die notwendigen Anforderungen. Aber dabei verbleibt die Literatur eher bei allgemeinen Empfehlungen. Evaluationen zu Subprozessen, Verbindungen oder Sequenzen, die sich besonders für spezielle Service Modelle eigenen, unterbleiben weitestgehend. Ferner wird die Schnittstellenkonfiguration überwiegend ignoriert.

Gerade diese Problematik stellt die Praxis vor elementare Herausforderung. Im Falle von strukturierten und standardisierten SCs ist die Verbindung weniger problematisch als bei reziproken, komplexen SCs mit dezentralen Verantwortlichkeiten [KD96]. Da die Literatur gerade bei komplexen SCs Vorteile durch CC vermutet [Ce12], sollte sich die Forschung künftig stärker den technischen Details widmen. Dieses Ergebnis wurde bereits in der 2-D Karte offensichtlich. Zusätzlich findet in der Literatur im Zusammenhang mit CC eine zu geringe Differenzierung der spezifischen Aufgaben der Informationsflüsse als auch der spezifischen Planungs- und Steuerungsaufgaben in der SC statt. Die hohe Kookkurrenz des „triple bottom line"-Konzepts in den Modellen „Stakeholder" und „Strategy" deutet auf die Validität des Konzepts im Gesamtkontext hin. Vorhandene IT-Ressourcen werden u.a. durch Skalierung und Virtualisierung effizient genutzt [MG11, Le10] und darüber hinaus ermöglicht CC die Bündelung und Lokalisierung der physischen Kapazitäten in umweltfreundlichen Umgebungen [VW12]. Aus Sicht der Anwender kann CC soziale Verbindungen vereinfachen und die Aufgabenkomplexität reduzieren, indem es ein Netzwerk um einen gemeinsamen Geschäftsprozess aufbaut [Ba13].

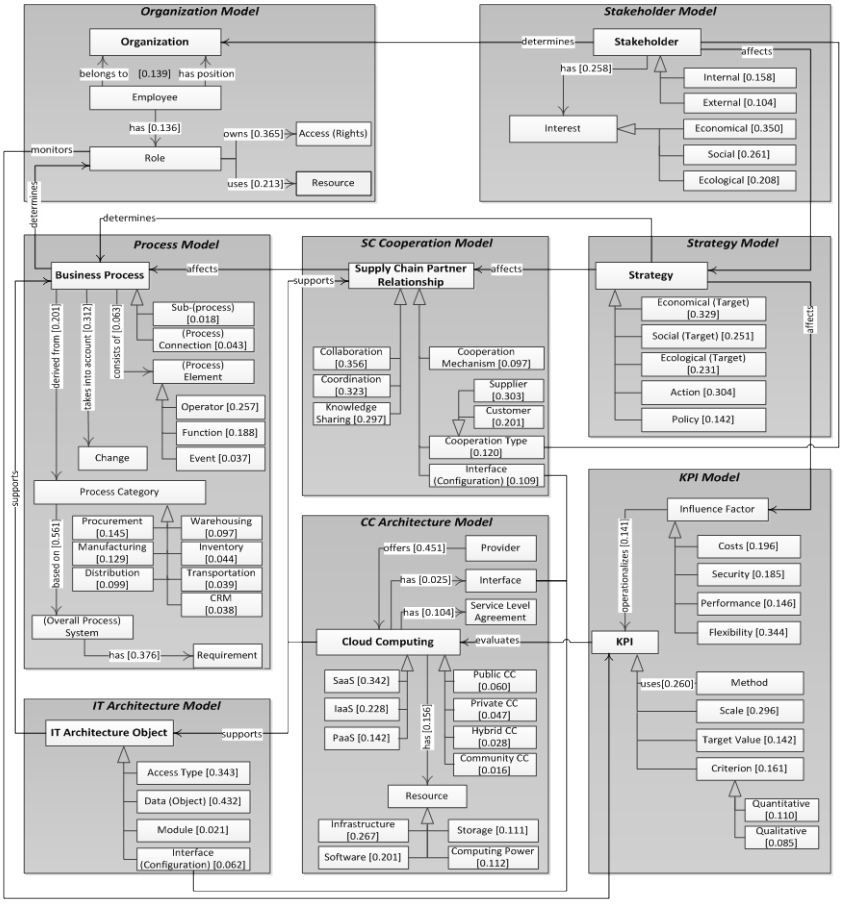


Abbildung 2: Referenzmodell für den CC-Einsatz in SCM-Architekturen

6 Fazit

6.1 Limitationen

Wie jede wissenschaftliche Arbeit, weist auch diese potentielle Limitationen auf. Demnach besteht die Möglichkeit, dass nicht alle relevanten Artikel in der Selektionsphase mittels Keywords gefiltert wurden. Die drei Ursachen hierfür können in der Unvollständigkeit der definierten Keywords, in alternativen Begriffsbezeichnungen in den Artikeln sowie in der Begrenzung auf vorab festgelegte Publikationsorgane liegen. Ferner erfordert die Kategorisierung der Artikel eine inhaltliche Prüfung und Bewertung, bei der eine Verzerrung durch Subjektivität nie vollends ausgeschlossen ist. Dennoch bietet diese Arbeit wichtige, neue Einblicke und deckt den aktuellen Forschungsstand auf. Um die Reliabilität der Resultate zu steigern, wurde der Analyseprozess im dritten Kapitel sorgfältig dokumentiert und Überprüfungen in den Kategorisierungen durchgeführt.

6.2 Implikationen für Wissenschaft und Praxis

Im Teilabschnitt 4.1 widmet sich die Arbeit den derzeitigen Problemstellungen. Themen wie Informationsaustausch, Koordination und Prozessoptimierung entlang der SC sowie die Frage nach der „Integration in die Unternehmensarchitektur“ treten vermehrt auf. Aus Sicht der Autoren ist für die weitere Theoriebildung eine Unterscheidung zwischen zwei Effizienzebenen hilfreich, die bislang eher willkürlich diskutiert werden: Streben nach Effizienz in der IT-Ressourcennutzung und Streben nach Effizienz in der Interoperabilität. Der Vorteil des CC durch ***effiziente IT-Ressourcennutzung*** ist nicht SCM-spezifisch sondern grundsätzlich für zahlreiche IT-Bereiche in der Unternehmensarchitektur denkbar. CC vereint dabei Ressourcen desselben Typs mit dem Ziel, diese zum einen optimal auszunutzen und damit zum anderen eine möglichst große Nachfrage zu bedienen [LKS11]. CC kann dabei Hardware und Software des nutzenden Unternehmens beispielsweise bei Produktionsprozessen, Lagerprozessen oder Transportprozessen unterstützen. Solange diese Unterstützung ohne Anbindung weiterer SC-Partner erfolgt, sind lediglich die direkten Einflussfaktoren zu berücksichtigen (siehe „KPI Model“). Der Vorteil durch ***Effizienz in der Interoperabilität*** hingegen ist SCM-spezifisch und wird im Referenzmodell im „SC Cooperation Model“ berücksichtigt. Da CC zwischen beliebig vielen SC-Partnern als Medium fungieren kann, welches über den Datenaustausch hinaus auch die kollaborative Datenanalyse und Prozessplanung sowie die Entscheidungsfindung unterstützt, handelt es sich hierbei letztlich um ***Wissensaustausch*** [Ce12, Le11]. Dass Interoperabilität Vorteile bietet, ist hinlänglich bekannt. Beispielsweise hat sich das EDI-System in der Praxis weitestgehend durchgesetzt. CC bedeutet einen tieferen und breiteren Eingriff in die Interoperabilität der SC als das EDI-System. Denn Daten werden nicht sequenziell von einem SC-Partner zum nächsten transferiert, parallel zum physischen Transport wie bei EDI, sondern stehen jedem SC-Partner jederzeit in Echtzeit an einer zentralen Stelle zur Verfügung. Die erhöhte ***Transparenz*** erlaubt das frühzeitige Beziehen von Informationen für Partner am Ende der SC. Wie zuvor erwähnt verstärkt sich dieser Vorteil bei langen und komplexen SCs.

Die Ergebnisse der Einflussfaktoren aus Abbildung 1 stellen die hohe Bedeutung einer möglichen ***Flexibilitätssteigerung*** heraus. Die IT-Nachfrage kann zeitnah und bedarfs-

orientiert gedeckt werden. Aufgrund des modularisierten Service-Ansatzes von CC können SCM-Prozesse konstruiert werden, die Unternehmen befähigen, schneller auf Veränderungen im allgemeinen Marktumfeld und in speziellen SC-Netzwerken zu reagieren [Ce12]. Ein weiterer Faktor zielt auf die ***finanziellen Vorteile*** ab. Vielfach wird das Argument eines Wechsels von Investitionen zu variablen Kosten und den damit verbundenen Vorteilen einer sinkenden Kapitalbindung angeführt [VW12]. Das Argument der „no-upfront-costs" [Ma11] darf aber nicht pauschal gelten, da es offensichtlich auf die Ausprägung „Public CC" abzielt. Ferner müssen unseres Erachtens die Abhängigkeiten zwischen den Faktoren „Flexibilität" und „Kosten" in Theorie und Praxis grundsätzlich vielschichtiger diskutiert werden. Da der Vorteil der Flexibilitätssteigerung vor allem für komplexe SCs gilt, erfordern gerade diese IT-Prozesse einen höheren internen Planungs- und Implementierungsaufwand, die den Kostenvorteil durch das „pay-per-use"-Modell überkompensieren können. Während erste Forschungsarbeiten auf den finanziellen Vorteil bei einfachen IT-Prozessen hindeuten [z.B. MDD12, Sc12a], fehlen gesicherte Erkenntnisse bezüglich der finanziellen Auswirkungen bei tiefgreifenden IT-Prozessen. Die Empfehlungen einiger Beiträge [z.B. MHQ13, STC12], komplexe CC-Systeme oder komplette CC-ERP-Systeme aus Kostengründen zu implementieren, ist unseres Erachtens daher zu voreilig, da diese Empfehlungen jeweils nur deduktiv hergeleitet wurden und eine empirische Untersuchung im realen Praxiseinsatz derzeit noch aussteht.

Beim Einflussfaktor „***IT-Sicherheit***" liegt die enorme Herausforderung für CC in den unterschiedlichen Sicherheitsstandards der SC-Partner. Eine Kooperation ist nur unter den jeweiligen Mindestvorrausetzungen vorteilhaft. Diesen entscheidenden Punkt umreißt nur eine Arbeit [OKA11], sodass hier Nachholbedarf gegeben ist. In der Praxis sollte dieser Aspekt bereits bei Selektion und Orchestrierung der CC-Module im unternehmensübergreifenden SC-Netzwerk Beachtung finden. In den Abschnitten 4.1 und 4.2 wird das frühzeitliche Stadium der Forschungsrichtung deutlich. Während in neueren Artikeln eine saubere Differenzierung zwischen den Service-Modellen herausgearbeitet wird, fehlt oftmals eine saubere Trennung bei der Ausprägung. Speziell für CC-Lösungen zwischen SC-Partnern bietet „***Community CC***" hohes Potential. Daher sollte sich die künftige Forschung noch stärker den Vor- und Nachteilen von CC-Varianten in spezifischen SC-Bereichen widmen (z.B. durch Erfolgsfaktorenforschung).

6.3 Ausblick

Die wissenschaftliche Literatur untersucht beim Bereitstellungsmodell „CC" bislang insbesondere technische Schwerpunkte und stellt vermehrt einen hohen Nachholbedarf von betriebswirtschaftlichen Denkansätzen im CC-Paradigma fest. Des Weiteren entdeckt die Forschung gerade erst den Nutzen von CC für spezifische Organisationsbereiche. Die beiden angeführten Feststellungen werden in dieser Arbeit bestätigt. Ferner weist die Literatur auf die Notwendigkeit einer „Cloud-Strategy" [Le10, MK12, Ma11] hin. Wie diese jedoch speziell für das SCM ausgestaltet werden soll, ist derzeit unklar.

Daher wird sich unser Fokus in der weiteren Forschung auf einen ***strategischen Handlungsrahmen*** richten, der beim Einsatz von ressourceneffizienten CC-Lösungen im SCM unterstützen soll. Wind et al. fordern in diesem Zusammenhang ein Cloud-Controlling [WRZ12]. Diese Anforderung soll in einem noch zu entwickelnden Handlungsrahmen die notwendige Beachtung finden.

Danksagung

Diese Arbeit ist im Rahmen des Projekts „IT-for-Green: Umwelt-, Energie- und Ressourcenmanagement mit BUIS 2.0“ entstanden. Das Projekt wird mit Mitteln des Europäischen Fonds für regionale Entwicklung gefördert (Fördernummer W/A III 80119242).

Literaturverzeichnis

Das Literaturverzeichnis enthält nur diejenigen im Beitrag zitierten Quellen, die nicht Gegenstand der Literaturanalyse sind. Alle anderen in der Arbeit zitierten Quellen konnten aufgrund der Seitenbegrenzung nicht beigefügt werden. Das Verzeichnis der 102 analysierten Quellen steht daher unter (http://www.uwi.uni-osnabrueck.de/Informatik14_Anhang.pdf) zum Abruf bereit (Anlage I).

[CR08] Carter C.R.; Rogers, D.S.: A framework of sustainable supply chain management: moving toward new theory, Int J Phys Distribution Logist Manag, 2008, 38(5), S. 360-387.

[Di04] Dibbern, J. et.al.: Information Systems Outsourcing: A Survey and Analysis of the Literature. ACM SIGMIS Database, 2004, 35(4), S. 6-102.

[Fe06] Fettke, P.: State-of-the-Art des State-of-the-Art: Eine Untersuchung der Forschungsmethode “Review” innerhalb der Wirtschaftsinformatik. Wirtschaftsinformatik, 2006 48(4), S. 257-266.

[Me01] Mentzer, J.T. et. al.: Defining Supply Chain Management. Journal of Business Logistics, 2001, 22(2), S. 1-25.

[HC10] Hevner, A.; Chatterjee, S.: Design Science Research in Information Systems. In Hevner, A; Chatterjee, S (Hrsg.): Design Research in Information Systems. S. 9-22, Springer US, Boston, 2010.

[KD96] Kumar, K.; Van Dissel, H.G.: Sustainable collaboration: managing conflict and cooperation in interorganizational systems, Management Information Systems Quarterly, 1996, 20(3), S. 279-300.

[MS95] March, S.T.; Smith, G.F.: Design and Natural Science Research on Information Technology. Decision Support Systems, 1995, 15(4), S. 251-266.

[MT09] Martens, B.; Teuteberg, F.: Why Risk Management Matters in IT Outsourcing - A Systematic Literature Review and Elements of a Research Agenda. Proceedings of the 17th European Conference on Information Systems (ECIS), 2009.

[MG11] Mell, P.; Grance, T.: The NIST definition of cloud computing”, available at: http://predeveloper.att.com/home/learn/enablingtechnologies/The_NIST_Definition_of_ Cloud_Computing.pdf (Abruf am 10. August 2013).

[Pr10] Provalis Research (2010). WordStat 6: Content Analysis Module for QDA Miner & SimStat. http://www.provalisresearch.com/Documents/WordStat6.pdf. Abruf am 13 September 2013.

[Su01] Sullivan, D.: Document Warehousing and Text Mining”, Wiley Computer Publishing, New York, NY, 2001.

[TSK05] Tan, P. N.; Steinbach, M.; Kumar, V.: Introduction to Data Mining. Addison-Wesley, Boston, MA, 2005.

[WW02] Webster, J.; Watson, R.T.: Analyzing the Past to Prepare for the Future: Writing a Literature Review. Management Information Systems Quarterly, 2002, 26(2), S. xiii–xxiii.

[WH07] Wilde, T.; Hess, T.: Forschungsmethoden der Wirtschaftsinformatik: Eine empirische Untersuchung. Wirtschaftsinformatik, 2007, 49(4), S. 280-287.

[YBS08] Youseff, L.; Butrico, M.; Da Silva, D.: Toward a unified ontology of cloud computing. Grid Computing Environments Workshop 2008 (GCE), S. 1-10.

Applications of Semantic Technologies

(AST 2014)

Applications of Semantic Technologies (AST 2014)[1]

Abstract: Semantic Technologies encode meanings explicitly and independent from concrete formats and application logic on the basis of formal languages. Applications of Semantic Technologies are currently being investigated in various fields, including Ambient Intelligence, Software Engineering, Cognitive Systems, Corporate Intranets, Knowledge Management, Bioinformatics, Robotics, eLearning and many more. Recent developments indicate that methods and tools provided by Semantic Technologies play a crucial role in these and other application areas for the foreseeable future, in particular for the processing and management of Big and Smart Data.

The international workshop on Applications of Semantic Technologies (AST) in its eighth edition brings together researchers, practitioners, and users from various fields to discuss goals, limitations, challenges, and real experiences. We accepted three submissions for publication and presentation at the workshop. The range of application areas covered by these papers is indeed diverse and shows how Semantic Technologies are applied in order to solve technical and domain-specific problems: integration of data from Linked Data sources into web applications, facility management with a focus on self-organisation, as well as adaptable and dynamic multi-agent systems. Semantic Technologies applied in these domains comprise Linked Data technologies, semantic widgets, ontologies, and semantic reasoning.

Organizing Committee: Stefan Zander (FZI Research Center for Information Technologies, Karlsruhe), Verónica Rivera-Pelayo (FZI Research Center for Information Technologies, Karlsruhe), Ignacio Traverso-Ribón (FZI Research Center for Information Technologies, Karlsruhe), Rudi Studer (Karlsruhe Institute of Technology, Karlsruhe), Jürgen Bock (FZI Research Center for Information Technologies, Karlsruhe)

Program Committee: Adila Alfa Krisnadhi (Wright State University Dayton Ohio, USA), Andreas Abecker (disy Informationssysteme GmbH, Germany), Bernhard Bauer (University of Augsburg, Germany), Bernhard Haslhofer (AIT- Austrian Institute of Technology), Bernhard Schandl (mySugr GmbH, gnowsis.com, Austria), Christoph Beierle (University of Hagen, Germany), Christoph Quix (Fraunhofer FIT, St. Augustin), Dragan Gasevic (Athabasca University, Canada), François-Paul Servant (Renault, France), Frithjof Dau (SAP AG, Dresden, Germany), Gabriele Kern-Isberner (University of Dortmund, Germany), Harald Kosch (University of Passau, Germany), Johannes Busse (Johannes Busse knowledge engineering), Jürgen Ziegler (University of Duisburg-Essen, Germany), Mark Hefke (CAS Software AG, Karlsruhe, Germany), Markus Luczak-Roesch (FU Berlin, Germany), Martin Junghans (Karlsruhe Institute of

[1] http://ast2014.fzi.de/

Technology, Germany), Robert Tolksdorf (FU Berlin, Germany), Rudi Studer (Karlsruhe Institute of Technology, Germany), Simone Braun (CAS Software AG, Karlsruhe, Germany), Stefan Schulte (TU Wien, Austria), Stephan Bloehdorn (IBM, Berlin, Germany), Stephan Grimm (Siemens AG, München, Germany), Thomas Fuhr (Georg Simon Ohm University of Applied Sciences Nuremberg, Germany), Thorsten Liebig (derivo GmbH, Ulm, Germany), Ulrike Lucke (University of Potsdam, Germany), Valentin Zacharias (codecentric AG, Germany)

SemwidgJS: A Semantic Widget Library for the Rapid Development of User Interfaces for Linked Open Data

Timo Stegemann, Jürgen Ziegler

University of Duisburg-Essen
Lotharstr. 65
47057 Duisburg
timo.stegemann@uni-due.de
juergen.ziegler@uni-due.de

Abstract: In this paper we introduce SemwidgJS, a JavaScript based library for displaying Linked Open Data via *Semantic Widgets*. SemwidgJS can be integrated in almost any website and handles the querying, processing and displaying of semantic data. SemwidgJS facilitates the retrieval of semantic data with its own simplified query language SemwidgQL, a Path Query Language that allows the traversal of the Linked Data graph. For complex queries, standard SPARQL can be used as well. We pursue two objectives with this development: first, we want to provide appropriate widgets for the full range of different data types in semantic databases, second, we aim at enabling Web content publishers who are not experts in semantic technologies to make use of these data in their ordinary Web development activities.

Introduction

The amount of semantically represented data on the open Web as well as within organizations has been growing significantly over the past few years. Semantic databases such as DBpedia[ABK+07] are continuously[1] collecting knowledge as interlinked RDF data, making them publicly available as Linked Open Data[2]. Large corporations, for example Google with its Knowledge Graph [Sin12] or the BBC with data sets for the FIFA World Cup 2010 and the Olympic Games 2012 [Ray12], have recently started to use internal semantic databases. While this data may be visible on standard websites, the technology exploiting the data is not open for external access. Although there is already an enormous amount of valuable information available, the usage of semantic data is still quite limited and has not yet made its way into wide-spread utilization either by the Web developer community or, even more so, by end users. Yet, due to their generic format, global interlinking and standardized query languages, semantic data and services form an excellent basis for building more flexible, generic applications and on-the-fly integrations of distributed data sets. A major issue for exploiting this potential is to significantly alleviate the development of applications based on Semantic Web data.

[1] `http://wiki.dbpedia.org/Changelog`
[2] `http://lod-cloud.net/`

To date, access to Linked (Open) Data usually takes place in two ways: first, through (Semantic) Web Browsers that allow a general, unfiltered display of information, or, second, through tools that are highly specialized for specific problems but limited in their wider applicability. With both types of tools, the interaction typically starts with the user entering a search string or the URI of a known semantic resource as an entry point into the semantic data graph. From there, the end user can then interactively explore related data. Alternatively, an automatic visualization of relevant data is generated.

The current techniques are a major obstacle for average bloggers, website authors, and other content publishers who may want to request Linked Open Data, integrate different sources, mix them with their own editorial content and publish them, similar to what BBC did with the sports datasets mentioned above. When using common blog-publishing services, users typically do not have the necessary access to the server back-end to integrate existing software that supports the usage of Linked Open Data.

To overcome some of these obstacles, we propose in this paper the concept of *Semantic (Data) Widgets* and describe an implementation that shields users from the complexity of the RDF query language (SPARQL[3]) and facilitate the construction of Web applications that use and integrate semantic data. We introduce SemwidgJS, a JavaScript based library for displaying Linked Open Data through Semantic Widgets. SemwidgJS can be integrated in almost any standard HTML webpage and handles the querying, processing and displaying of semantic data. While existing libraries sharing a similar goal only comprise widgets for information visualization purposes, SemwidgJS also features widgets for typical UI elements such as labels, links, and text input fields. To make querying semantic data easy, SemwidgJS supports its own query language – SemwidgQL which is used as an alternative to standard SPARQL (which can still be used additionally). SemwidgQL is a simple Path Query Language that allows the traversal of the Linked Data graph starting with a predefined resource. We aimed at keeping the query language as simple as possible to enable users who are not familiar with the techniques of the Semantic Web to use SemwidgJS.

The remainder of this paper is organized as follows: in the next section we review related work and the current state of research. Then we introduce SemwidgJS, subdivided into the SemwidgQL query language, the widget library, and its capability to act as a framework for custom widgets. Finally, we conclude with a discussion and give an outlook on future work.

Related Work

Several tools and frameworks for exploring and displaying information from the Semantic Web, in particular data from the Linked Open Data (LOD) cloud, have been described in the literature so far. Existing tools typically either allow to browse and visualize semantic data in a generic, standard format, as is the case for Semantic Web browsers [BLCC+06], or are highly specialized for solving a specific problem, for example, finding and visu-

[3] **http://www.w3.org/TR/rdf-sparql-query/**

alizing relations between Linked Data instances [HHL+09]. These tools are mostly very limited in terms of their ability to display semantic data in custom aggregations and layouts or mix them with ordinary Web content, which is the goal of our development.

Projects with similar objectives exist, using both client-side and server-side techniques. Consequently, they differ in the way they access data and how the type of visualization is defined. Sgvizler [Skj12] is a pure JavaScript libraries for information visualization utilizing Google Charts[4]. It queries Linked Open Data endpoints directly via SPARQL. Visualizations can easily be embedded in existing sites or systems (such as wikis or blogs) and are therefore usable for everyone. After including the JavaScript file Sgvizler widgets are declared directly within the HTML code. Chart type, SPARQL endpoint URL, and SPARQL query are defined by HTML5 data-* attributes. The chart is then rendered inside the page. Sgizler is based on RDF-Spark[5], which allows arbitrary visualizations by custom formatting functions programmed in JavaScript.

Exhibit[HKM07] also works on the client side, but does not utilize native semantic data from a SPARQL endpoint or an RDF file. Instead, data has to be serialized as JSON file and be embedded into the website that will contain the Exhibit widget. The widgets themselves are declared within the HTML code.

In contrast, LESS [ADD10] works on the server side. Templates are created in a Web application and saved to a server. Usable data is defined by a SPARQL query or by specification of the URI of a Semantic Web resource. Using the Smarty template engine for PHP, this data can be integrated into HTML code. Templates can be retrieved from the server via a URL and embedded into an existing website. Users hereby depend on both the SPARQL endpoint and the LESS server. The LESS server is no longer available online.

Paggr [Now09] allows the collaborative creation of widgets ("Sparqlets") in a wiki-like dashboard system. The system can query SPARQL endpoints, but also process RDF files directly. SPARQL queries can be generalized with placeholders and queries can be processed by means of a specific script language ("SPARQLScript"). However, using the widgets is restricted to the dashboard system.

Fresnel [PBKL06] is an RDF vocabulary for creating RDF style sheets. These style sheets consist of Lenses and Formats. Lenses select data, while Formats format the selected data. The way the data is displayed is determined by the tool, that utilizes Fresnel. To select data, either SPARQL or the Fresnel Selector Language (FSL)[6] can be used. FSL is a path language for selecting RDF data, similar to XPath[7], which operates on XML documents. However, FSL is not intended for directly querying data from a particular resource. The resource must first be selected by a rather sophisticated filter expression[8]. FSL is restricted to a single variable for its results. Therefore, result sets will always be one-dimensional lists. This is sufficient for simple UI elements, but it can be cumbersome to fill complex elements such as tables or charts that rely on multidimensional data.

[4] **https://developers.google.com/chart**
[5] **http://km.aifb.kit.edu/sites/spark/**
[6] **http://www.w3.org/2005/04/fresnel-info/fsl/**
[7] **http://www.w3.org/TR/xpath20/**
[8] `*[uri(.) = 'http://example.org/foo#bar']`

Another path language for selecting RDF data is LD Path [SBK+12], which became part of the Apache Marmotta project[9]. The alternative RDF serialization Notation3 (N3) already has its own syntax for path querying[10].

Outside the world of the Semantic Web, many JavaScript widget libraries and frameworks[11,12] are already available for normal Web content, which can be utilized to create Web UI elements dynamically and build Rich Internet Applications. Integrating semantic data, however, is mostly difficult or not possible at all.

None of the tools and languages discussed above fully meets the requirements of facilitating the integration of semantic data into conventional Web applications with Semantic Widgets that cover the whole range of data types available in LOD. While Sgvizler is integrateable into any website, it is limited to to the display of charts and users have be able to write SPARQL queries, which is one of the biggest obstacles. Server-side applications such as LESS add a potential single point of failure to the data retrieval process when used as a Web service. Installation of own server software is often impossible for average users or associated with high effort. Fresnel style sheets are written in RDF and therefore reserved for experts. In general, the existing tools require specific knowledge of semantic techologies and set a rather high entry threshold for the average Web developer.

While the different path-oriented query languages are fairly easy to use, they operate on schema level, which complicates their use as a query language for publishing website content. Operation on schema level implies that all data for a class or type is queried. Though typically websites show data for specific instances instead of lists of instances. Using these query languages on instance level requires additional filter expressions. Furthermore, FSL is limited to one-dimensional result lists, which can make it cumbersome to fill complex elements and can lead to a loss of context information. It is not sufficiently defined whether and how these query languages can be translated into SPARQL, which is indispensable for a client-side Web application accessing arbitrary semantic data sources.

SemwidgJS

SemwidgJS is a JavaScript based library for displaying *Semantic Widgets* that enriches websites with UI elements whose data is queried from SPARQL endpoints. SemwidgJS supports creation of simple UI elements such as labels and links, interactive elements for data input and selection that can manipulate other widgets, but also complex elements for displaying charts and maps that can base on further widget libraries.

While the data is queried from publicly available semantic databases, SemwidgJS otherwise works entirely on the client side. Thus no additional software is required. Widgets are declared directly within the HTML code and automatically processed after loading the website. For this purpose, only the SemwidgJS file must be included into the HTML

[9] **http://marmotta.apache.org/**
[10] **http://www.w3.org/TeamSubmission/n3/**
[11] **http://www.sencha.com/products/extjs/**
[12] **http://yuilibrary.com/**

document. To simplify data queries and thus be able to address a larger audience than comparable tools do, we developed SemwidgQL – a simple query language that is translated into SPARQL. If necessary, users can still express more sophisticated queries in SPARQL.

With SemwidgJS website authors can integrate simple data into their website. Furthermore, the internal event system and the ability to manipulate other widgets enables them to create complex Web applications. Changes which are made to widgets or the Document Object Model of a website, either by another widget or an external script will trigger an event the widgets can react on. In the simplest case the widget is reloaded, using the updated data.

In the further course, we present the SemwidgJS widget system in detail. Because the widgets can be used effectively only in combination with the SemwidgQL query language, we start by describing this language first. Then we introduce the widget library and describe how SemwidgJS can be used as a framework for creating own widgets.

SemwidgQL – Semantic Widget Query Language

SemwidgQL is a simple Path Query Language for Linked Data, starting with a predefined resource in the Linked Data graph. Linked Data is based on the RDF[13] data model. Information is represented in statements, called triples composed of subject, predicate and object. The object of a triple can be a literal such as a number or string, but also a separate Linked Data resource that is the subject of other triples.

SemwidgQL is translated into SPARQL – the standard query language for Linked Data – and can thus be used to query almost any public Linked Data endpoint. The intention behind SemwidgQL is not to replace or extend SPARQL. It is merely intended to simplify requesting instance data for users who are not familiar with the techniques of the Semantic Web.

SemwidgQL is limited to the traversal of the Linked Data Graph plus simple filter functions. The traversal is indicated by the dot notation, which is reminiscent of the syntax of object-oriented programming. Filters are enclosed in parentheses and refer to the previously defined property. Filters can be shortened by certain keywords.

Due to this straightforward syntax, an editor can easily provide assistance during the query creation process. Available properties of a given resource can be requested and be used for an auto-complete/-suggest feature. The same applies to the possible filter functions. We have developed an basic editor for creating SemwidgQL queries that offers the user such assistance[14].

The basic features of SemwidgQL are introduced below. All examples can be used to query data from DBpedia's public SPARQL endpoint.

Named Resources: SemwidgQL supports a naming mechanism for resources. Instead of fully qualified resource URLs or URLs prefixed by a namespace, a user can specify a

[13] **http://www.w3.org/TR/rdf-primer/**
[14] **http://semwidg.org/ql/editor/**

substitute name and use it within all SemwidgQL queries.

```
rome = http://dbpedia.org/resource/Rome
```

Path Traversal: The following SemwidgQL query requests all labels of the resource Rome. Below the corresponding translation into SPARQL is shown. In order to save space and increase readability, we replaced the fully qualified resource name of Rome by its prefixed name (dbpedia:Rome) in the SPARQL query.

```
rome.rdfs:label

SELECT DISTINCT
  (dbpedia:Rome AS ?uri) ?label
WHERE {
  dbpedia:Rome rdfs:label ?label .
}
```

Inverse Property Selection: Linked Data is represented by a directed graph. Therefore, it is possible that a resource is not aware of all connections that exist between itself and other resources. In order to obtain properties of resources that are linked unidirectionally to the initial resource, it is possible to invert the property selection by a prepended caret symbol (^). For example, the DBpedia database does not contain any links between a city and the people who were born in it. But it contains connections between people and their places of birth. To request a list of people who were born in Rome, the property selection must be inverted. The following query requests the labels of all individuals whose birthplace is Rome.

```
rome.^dbpedia-owl:birthPlace.rdfs:label

SELECT DISTINCT
  (dbpedia:Rome AS ?uri) ?birthPlace ?label
WHERE {
  ?birthPlace dbpedia-owl:birthPlace dbpedia:Rome .
  ?birthPlace rdfs:label ?label .
}
```

Property Filter: Object properties can be filtered by values of linked properties. The left-hand side of a filter expression specifies a property that references to the previously defined object property or resource in the query path. The rest corresponds to the usual SPARQL syntax for filter expressions. As a multilingual database DBpedia contains the name of Rome in different languages (Rome, Roma, Rom, 罗马 ...). One of the most common actions when querying multilingual texts for displaying purposes is consequently the filtering for a specific language. Therefore, we have a simple language filter integrated directly into our query language. The language filter is prefaced with the keyword `@lang`. The following query contains two filters and requests the English label of all soccer players who were born in Rome.

```
rome.^dbpedia-owl:birthPlace(rdf:type = dbpedia-owl:SoccerPlayer).
  rdfs:label(@lang = 'en')
```

```
SELECT DISTINCT
  (dbpedia:Rome AS ?uri) ?birthPlace ?label
WHERE {
  ?birthPlace dbpedia-owl:birthPlace dbpedia:Rome .
  ?birthPlace rdf:type ?filtertype .
  Filter (?filtertype = dbpedia-owl:SoccerPlayer) .
  ?birthPlace rdfs:label ?label .
  Filter (lang(?label) = '' || langMatches(lang(?label), 'en')) .
}
```

Multiple Property Selection: In some cases, certain properties are inseparable. Hence it is possible to select them simultaneously in a single query. Selection of multiple properties is only allowed as last part of a SemwidgQL query.

```
rome.[geo:lat, geo:long]
```

```
SELECT DISTINCT
  (dbpedia:Rome AS ?uri) ?lat ?long
WHERE {
  dbpedia:Rome geo:lat ?lat ;
               geo:long ?long .
}
```

The complete definition of SemwidgQL's syntax is avaliable online[15].

Semantic Widget Library

SemwidgJS's widget library covers simple widgets such as widgets for text display (SwLabel), lists (SwSimpleList) and Web links (SwLink), used only for presentation of semantic data, but also interactive elements for user input and manipulation of other widgets. Furthermore, complex elements exist for displaying charts and maps that can base on external widget libraries, for example Leaflet[16] to embed OpenStreetMap material.

Widgets are declared within a website's HTML code. A widget's properties are defined by HTML5 data-* attributes of standard HTML elements. While `div` elements should be the default host element of a widget, all other elements such as `h1` or `p` are also possible, as long as the element supports the widget's HTML code. The HTML code of a widgets is directly injected into its host element and therefore inherits its style properties.

SemwidgJS distinguishes between two classes of elements: Widgets (data-sw-widget) and configuration elements (data-sw-config). All elements are identified by the unique value of its ID attribute (data-swp-id). Configuration elements, just like widgets, are declared directly within the HTML code. There are five configuration elements for 1) resources, 2) endpoints, 3) namespaces, 4) proxies, and 5) mappings.

1 – 4 are used to configure semantic endpoint connections and the corresponding query

[15] **http://semwidg.org/ql/ebnf/**
[16] **http://leafletjs.com/**

in SPARQL or SemwidgQL format. An endpoint element stores, among others, the URL under which the SPARQL endpoint is accessible. A resource element stores a URI, which identifies a resource in a semantic database and the ID of an endpoint from which the data should be queried. When using SemwidgQL it is not necessary to enter the often quite long resource URI and SPARQL endpoint URL for every query but only to enter the ID of a resource element that is linked to an endpoint element.

The mapping element is used to replace data within an attribute of a widget. For example, a search string entered into a text input widget can change the value of a mapping element. Afterward all other widget attributes connected to the mapping element change accordingly.

In the further course, we present some example widgets and demonstrate the interoperation of the widget library and the SemwidgQL query language.

Configuration: Below a minimal example configuration for Semwidg JS is presented. First, the connection to the DBpedia SPARQL endpoint is configured. Second, the resource for the city of Rome is configured using its URI, and then linked to the previously created endpoint "dbpedia". The order and exact position within the HTML code is irrelevant for configuration elements. At the start SemwidgJS searches the entire HTML document and processes all elements in the required order.

```
<div data-sw-config="endpoint"
  data-swp-id="dbpedia"
  data-swp-url="http://dbpedia.org/sparql">
</div>

<div data-sw-config="resource"
  data-swp-id="city"
  data-swp-uri="http://dbpedia.org/resource/Rome"
  data-swp-endpoint="dbpedia">
</div>
```

Single Value Widgets: The configuration of widgets takes place in the same way as for configuration elements. Below a simple label widget is embedded into an `h1` element of a website to act as a heading. For this purpose the widget type (SwLabel) and a query must be specified. The widget is rendered directly within the `h1` element. In contrast to the configuration elements, the position is relevant. SemwidgQL queries are enclosed in single curly braces, while SPARQL queries are enclosed in double curly braces.

```
<h1 data-sw-widget="SwLabel"
  data-swp-value="{city.rdfs:label(@lang = 'en')}">
</h1>
```

Multiple Value Widgets: The link widget (SwLink) accepts several values in order to create a hyperlink. Simple text and queries – whether SemwidgQL or SPARQL – can be combined with each other.

```
<div data-sw-widget="SwLink"
  data-swp-value="{city.foaf:homepage}"
  data-swp-label="Website of {city.rdfs:label(@lang = 'en')}"
```

```
  data-swp-target="_blank">
</div>
```

Interactive Elements: Interactive elements such as text input fields are able to manipulate configuration elements and other widgets. For example, interactive elements can be used to change the URI of a resource element. Thus, the website can be built generically for all cities inside the database and the end users can select the city to which they wish to receive information. Following input widget (SwInputText) is connected with the previously defined resource element "city" and its attribute "(data-swp-) uri". Once the user confirms the input, the text is written to the corresponding attribute and the internal event system ensures that all related elements will be updated.

```
<div data-sw-widget="SwInputText"
  data-swp-id="resourceUriInput"
  data-swp-target="city.uri">
</div>
```

Mapping: In addition, another configuration element exists that can be used for the further generalization of widgets. The mapping element stores an arbitrary string that replaces the corresponding placeholders in attribute values. Again, the internal event system ensures that all related elements will be updated, after the value of a mapping element changed.

```
<div data-sw-config="mapping"
  data-swp-id="lang"
  data-swp-value="en">
</div>

<h3 data-sw-widget="SwLabel"
  data-swp-value="{city.rdfs:label(@lang = '$$lang$$')}">
</h3>
```

Example: Figure 1 show the composition of several Semantic Widgets in a website some of which have been described in the previous subsections. A complete running example plus its source code is available online[17].

Semantic Widget Framework

SemwidgJS does not only serve as a widget library, but also as a framework for building custom widgets. In most cases, widget authors only need to define the properties a widget should support (in addition to the standard properties inherited from the base widget class) and override the function that generates the HTML code. Getter and setter function for these properties are created automatically by the framework and contain all necessary event listener and dispatcher for proper cross-widget interoperation. Additional getter and setter functions for explicit SPARQL endpoint definition, selector functions, and query

[17] **http://semwidg.org/js/**

Figure 1: Page composed of several Semantic Widgets

results are also created automatically according to predetermined naming conventions. These functions can be utilized by widget authors at any time.

Although JavaScript can react to DOM events of a website, it does not have its own event system that can be utilized directly. Therefore, we implemented a simple event model that allows widgets to react to events of other objects, or conversely fire own events.

The automatic generation of functions ensures that all required events are fired (e.g. when setting a SPARQL result) and in return received by the respective controllers that are responsible for managing a widget's lifecycle. If necessary, these functions can be overwritten within the widget's code, for example, to check values before setting them or inject custom events.

All predefined SemwidgJS widgets are also constructed in this way. When defining custom properties, in addition to the property name the type and a property description is specified. Possible types are "rigid", "query" and "selector". The value of a rigid property is considered to be a simple text and will not be further processed by SemwidgJS, while query properties can contain simple text, a SemwidgQL query and a SPARQL query. Selector properties reference another SemwidgJS element by its ID and property name (id.property). This information should be used in the future by a widget editor to provide users information about these properties but also to be able to make assistance such as auto-complete/-suggest.

Discussion

The ongoing growth of the Semantic Web and Linked Open Data endpoints poses the problem of enabling non-expert users to interact with and make use of semantic data in an intuitive, yet flexible and powerful manner. Although there are already numerous tools and libraries that enable the use of semantic data within own websites, these tools and libraries are often cumbersome or highly limited in their functionality and reusability.

In this paper we have presented SemwidgJS, a library for Semantic Widgets that is easy to

extend and supports its own simple query language. With SemwidgJS, web design skills and fundamental knowledge of the Semantic Web are already sufficient to take advantage of Linked Open Data in own websites. Previous solutions required typically three different kind of roles or skills. RDF or SPARQL skills to query the data, programming skill to parse and process this data, and web design skill to integrate this processed data into the website. SemwidgJS supports the entire range of Web UI elements, from simple text output, through interactive input elements to complex charts and maps.

When building Semantic Widgets with SemwidgJS, the integration of semantic data can be defined on the instance level. A Semantic Web resource (instance) is defined explicitly and data is queried for this resource, for example the city of Rome. Included mapping and templating mechanisms also allow schema level integration.

According to a categorization of Semantic Web application [MA10] websites and Web applications created with SemwidgJS can presumably be categorized as visualization oriented. They are using extrinsic retrieval, are having a consuming information flow direction, and are semantically integrated.

Currently, we are planing an editor for the creation of complex templates or even complete Web applications based on SemwidgJS. The final result could then be exported as HTML code and be embedded into an existing website, a technique which is already known from many other websites and services (e.g. the embedding of Youtube videos[18]).

For users, it should be possible to drag and drop selected widgets into their workspace and configure them. Due to the explicit description of each widget's properties and the linear structure of the SemwidgQL query language, an editor can provide ideal support to the users in the creation and integration of *Semantic Widgets*.

References

[ABK+07] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg, 2007. Springer-Verlag.

[ADD10] Sören Auer, Raphael Doehring, and Sebastian Dietzold. LESS - Template-Based Syndication and Presentation of Linked Data. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *ESWC (2)*, volume 6089 of *Lecture Notes in Computer Science*, pages 211–224. Springer, 2010.

[BLCC+06] Tim Berners-Lee, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *In Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.

[HHL+09] Philipp Heim, Sebastian Hellmann, Jens Lehmann, Steffen Lohmann, and Timo Stegemann. RelFinder: Revealing Relationships in RDF Knowledge Bases. In *Proceed-*

[18] **https://developers.google.com/youtube/iframe_api_reference**

ings of the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009), pages 182–187, Berlin/Heidelberg, 2009. Springer.

[HKM07] David F. Huynh, David R. Karger, and Robert C. Miller. Exhibit: Lightweight Structured Data Publishing. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 737–746, New York, NY, USA, 2007. ACM.

[MA10] Michael Martin and Sören Auer. Categorisation of Semantic Web Applications. In *proceedings of the 4th International Conference on Advances in Semantic Processing (SEMAPRO2010) 25 October – 30 October, Florence, Italy*, October 2010.

[Now09] Benjamin Nowack. Paggr: Linked Data widgets and dashboards. *J. Web Sem.*, 7(4):272–277, 2009.

[PBKL06] Emmanuel Pietriga, Chris Bizer, David Karger, and Ryan Lee. Fresnel - A Browser-Independent Presentation Vocabulary for RDF. In *Lecture Notes in Computer Science (LNCS 4273), Proceedings of the 5th International Semantic Web Conference (ISWC 2006)*, pages 158–171. Springer, November 2006.

[Ray12] Jem Rayfield. Sports Refresh: Dynamic Semantic Publishing. Website, April 2012. Available online at **`http://www.bbc.co.uk/blogs/bbcinternet/2012/04/sports_dynamic_semantic.html`**; visited on March 24th 2014.

[SBK+12] Sebastian Schaffert, Christoph Bauer, Thomas Kurz, Fabian Dorschel, Dietmar Glachs, and Manuel Fernandez. The Linked Media Framework: Integrating and Interlinking Enterprise Media Content and Data. In *Proceedings of the 8th International Conference on Semantic Systems*, I-SEMANTICS '12, pages 25–32, New York, NY, USA, 2012. ACM.

[Sin12] Amit Singhal. Introducing the Knowledge Graph: things, not strings. Website, May 2012. Available online at **`http://googleblog.blogspot.de/2012/05/introducing-knowledge-graph-things-not.html`**; visited on March 26th 2014.

[Skj12] Martin G. Skjæveland. Sgvizler: A JavaScript Wrapper for Easy Visualization of SPARQL Result Sets. In *9th Extended Semantic Web Conference (ESWC2012)*, May 2012.

Facility Management in the Context of Self-Organisation

Eric Forkel[a], Janek Götze[a], Michael Clauß[b], Christan-Andreas Schumann[a]

[a]Fachgebiet Wirtschaftsinformatik
Westsächsische Hochschule Zwickau
Dr.-Friedrichs-Ring 2a
08056 Zwickau
eric.forkel@fh-zwickau.de
janek.götze@fh-zwickau.de
christian.schumann@fh-zwickau.de

[b]Fachgebiet Fabrikplanung und
Fabrikbetrieb
Technische Universität Chemnitz
Erfenschlager Str. 73
09126 Chemnitz
michael.clauss@mb.tu.chemnitz.de

Abstract: Nowadays there is a rise of complexity in most areas of planning, managing and executing modern production systems. One part is the facility management. Processes such as approval and maintenance in the production systems environment are important for modern companies. These processes generate enormous data during the lifecycle of a production system. The approach of self-organisation supports the handling of these data among others. In this position paper, a concept of a modern facility management with semantic technologies and augmented reality will be introduced at the beginning. Then a definition of three different conceptual use cases of self-organisation in the context of facility management will be illustrated. At the end, the adaption of contexts for supporting self-organized systems will be shown in this conceptual paper.

1 Introduction

Trends in production systems planning and management have lead to elaborated approaches for changeable and energy-efficient factories during the last years and decades. This development is increasingly being facilitated by the application of information technology [Cl13]. Digital and virtual models, tools and methods are of high relevance for the implementation of changeability in modern production systems [He03]. But more digital data does not necessarily lead to better systems and processes. First of all, complexity rises.

Just recently concepts such as the Digital Factory approach [Vd08] or the Internet of Things based on cyber-physical systems have been summarised by the Industry 4.0 initiative as an immediate challenge for the whole industrial landscape. This initiative is strongly focused on data. Internet and Web technologies are said to be widespread in factories in 10 to 15 years already. In the course of that, the visualisation of data,

assurance of semantic interoperability and user-specific information provision are the challenges to face [Sa11].

Typical aims of the digitalisation are an increased process control, improvement of planning quality, costs and time. The focus lies on the integrity, actuality and the reduction of redundancies in production data, the improved utilisation of planning knowledge and extensive support of communication and collaboration [BGW11].
The two fields of planning and management of production systems can hardly be separated anymore, since there are many interrelated, repetitive tasks connecting them. A basic requirement for projects in the Industry 4.0 is hereby an integrated data-management in the digital factory that reproduces the connection between the real and virtual world by synchronising the data between different working groups and tools during the whole lifecycle of a production system. The overall vision is to provide an integrated virtual model of the real production system in real-time [Vd09].

All in all, the integration of data that is spread across organisational borders and that has to be collected from many different data sources still leaves more questions than answers when it comes to the implementation of what was described above.

2 Concept of a Modern Facility Management with Semantic Technologies (and Augmented Reality)

In the so-called "FMstar" project (see http://www.fmstar.de) seven partners from science and industry develop a respective solution with focus on facility management (FM) processes such as approval and maintenance in the production systems environment. Hereby, an overall concept and corresponding prototype for the integration and presentation of planning data based on semantic technologies and visualization via augmented reality (AR) on mobile devices is being developed. The focus lies on building services engineering and factory maintenance planning applications that generate the most relevant data for the facility management processes in scope. They provide data with reference to 3d-models of the building structure. This data has to be integrated for presentation to the approval and maintenance staff and is supposed to support their work in different ways.

The basic idea for the integration of the data is the utilisation of semantic technologies as described by [BHL01] for the semantic web. Semantic searching within traditional, unstructured documents is just as semantic databases and webs still subject to intensive research [Ba13]. It is an important challenge to facilitate the patency of the data and a semantically correct mapping between the different tools [Vd09]. Hereby it's important to assure data correctness, consistency, integrity, actuality, security and avoidance of redundancy. One way to synchronize such distributed, partial models is using ontologies [Ha05]. Ontologies are a key concept when it comes to the improvement of information systems integration, communication and information- and knowledge-management in production systems planning and management [SL09]. In the project a basic ontology for

the relevant context is being developed. It describes the structure of the semantic database and is used for the mapping between the different models and files.

But besides the technical aspects of that, the question arises how to utilise the integrated data in the facility management processes. Part of the answer is that there is a fundamental paradigm shift going on from strongly structured approaches to more iterative, gradually detailing approaches with a particular consideration of the human/social perspective that puts the planning efforts under a self-organisational paradigm [S+06]. This empirical method is already relying on experiential knowledge, pragmatism and intuition of the involved people today. Also expert systems that support operative work in a self-organisational manner by providing accumulated know-how from previous utilisation in similar situations are a known concept in this context [BGW11].

As described in [Co14] for Big Data BI-projects, it is of substantial importance to use the gathered data, information and knowledge in the respective processes for a sustained improvement in the context at hand though offering guidance to workers in a self-organisational manner. This proves that self-organisation is a core concept in this context. However, the practical implementation of the before mentioned concepts is still far behind what is described in guidelines and scientific publications [AMR11]. So in the next sections different conceptual use cases for the self-organisational utilisation of the integrated data will be discussed in more detail for the facility management context.

3 Self-Organisation in Context of Facility Management

A part of the mentioned project FMstar is to develop conceptual use cases of self-organisation in context of facility management. Generally, self-organisation is the ability of systems to develop structures outside predetermined structures directly from the system [Pa82]. In context of the project self-organisation is defined as the ability of our prototype-system to generate new information structures from existing partial information of production system planning, building services and facility management. Hereby, information will be available new organized and new structured for maintenance and factory acceptance. This is, no less importantly, essential for context-based information.

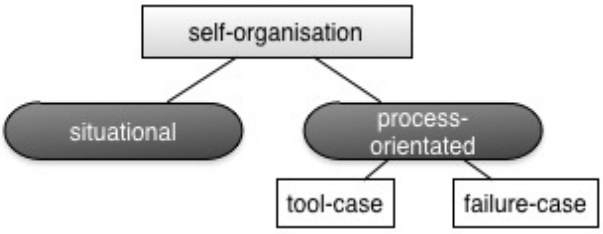


Figure 1: Conceptual use cases of self-organisation

Subsequently, three use cases of self-organisation in context of facility management (compare figure 1) will be introduced conceptually. In general, the use cases are distinguished between situational and process-oriented. On the one hand, situational means there are no predefined procedures for the technician. The technician must decide (especially the first steps) by oneself what to do. On the other hand, in a process-oriented case, there exist process operations to selected situations.

The conceptual approach of the classification and chronology of the cases is illustrated in figure 2. There is a general distinction between the socio-technical system and the information system. The socio-technical system handles with the descripted approach of self-organisation in this chapter. Hence, this exterior level system contains the technician with the mobile device and the information system. Contrary to this exterior level, the interior level – information system – describes the adaption of contexts. This concept will be submitted in chapter four of this paper. Following, the three conceptual cases are related to the exterior level.

In the first mentioned use case (situational) the technician shall be supported by the system. Therefor, the system will provide information specifically to the given situation of the technician. This information will depend on the qualification, competence or historical action with the tablet of the technician. This case is illustrated as “S1” in figure 2, which shows an overview of classification and chronology of the cases.

The next use case is the tool case, which is classified as process-oriented. The tool case scenario is relevant for the preparation of maintenance orders. Hereby, the system will put forward proposals for picking the tool case on the basis of upcoming orders (compare figure 2: “S2”). The risk that locally is a lack of tools, materials or special lubricants will be reduced.

Last but not least the case of failure is shown in “S3” in the figure 2. At this, on one side the system will analyse the failure case and support quick assistance for the planner. On the other side, the system will support the technician locally to fix the occurring failure quickly.

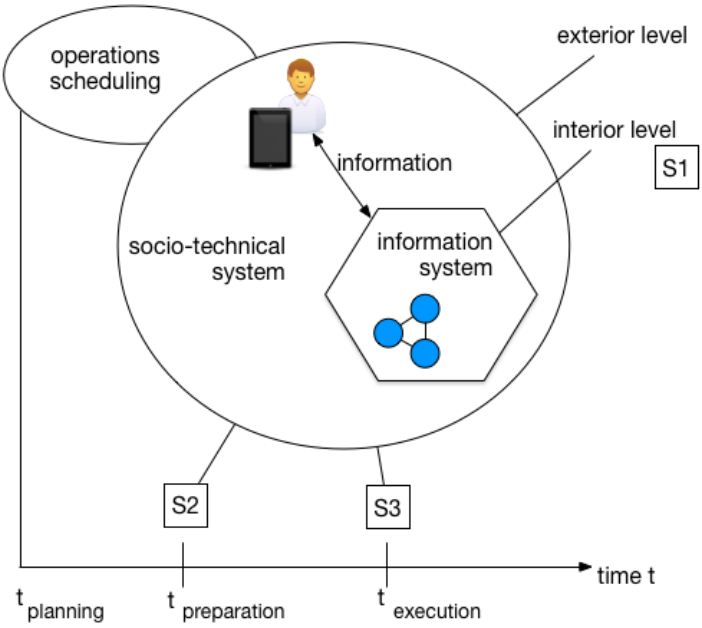


Figure 2: Overview of classification and chronology of the cases

4 Adaption of Contexts to Support Self-Organised Systems

One of the major thematic priorities of the project FMstar is the context-based filtering of the integrated data. The general objective is to avoid time-consuming and tedious search for information on the mobile assistance system to be developed. The context-based information supply is concerned with the question, how only relevant information can be provided to a FM technician. According to the Relevance Theory from Wilson/Sperber, relevance strongly depends on the reference of information to a specific user situation [WS02]. The situation of the user can be characterized by the context of the Human-Computer Interaction. According to Dey, context is „[…] any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves“ [De00].

The basic approach for utilizing the context for information supply within FM is to analyse the day-to-day operating processes of FM associates. This was based on a process and document analysis, workshops with industry partners and a small user survey [G+14]. As a result four context categories were identified (Figure 3: Context categories). The mobile assistance system being developed shall be capable to measure the respective context factor values of the users' situation and to interpret the context and intention of the Human Computer Interaction (HCI). Use cases utilize the context factors to filter or prioritize displayed information on the mobile device.

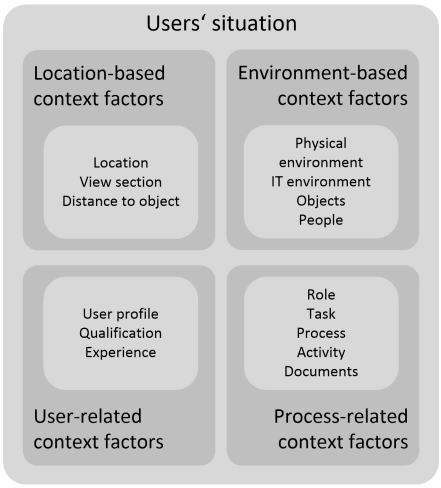


Figure 3: Context categories

In the following, our approach will be pointed out based on a use case regarding the main topic of the paper self-organisation. Figure 4 illustrates the differences between the learning of the socio-technical system and the information system itself. Context information can support both concepts. The following simple example (Figure 4: Use case with self-organisational effect) introduces how the information system can learn from the historical usage by the FM technician. The general assumption of the use case is that the relevance of information objects (e.g. attributes of a component) can be determined by the historical behaviour of the technician. If specific information objects were relevant in the past, they might be relevant in the present and future as well.

The objective of the use case is to shorten a given attribute list to a specific component by means of the context factor “historical relevance”. In the past the technician “used” the attribute 3 four-times and the attribute 1 two-times. In the current situation the list will be shorten to attribute 3 and 1. To achieve this functionality the application has to capture the usage and to transform it into a specific weight of the attribute. The TDB (Triple Data Base) assigns the respective weight to the attribute. The context sensor (e.g. service to provide a specific weight of an attribute) provides the weight of an attribute at runtime of the application. A context model interprets the value in the purpose of the use case (e.g. weight value into star category). The resulting self-learning effect of the software enables self-organisation because relevance of attributes can be considered without a predetermination of it in the planning phase. The relevance will be determined automatically by using the mobile assistance software.

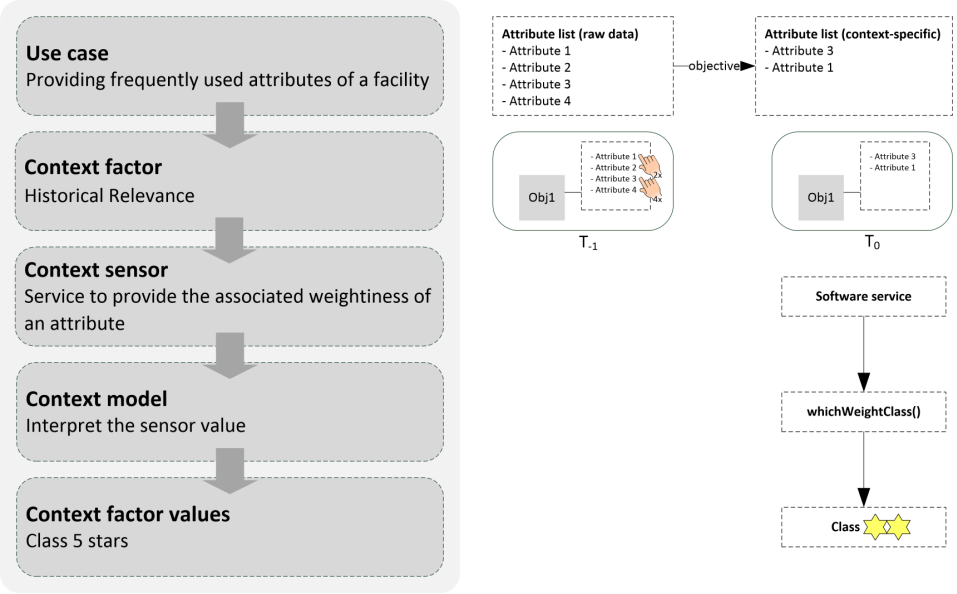


Figure 4: Use case with self-organisational effect

5 Conclusion

This paper showed a concept of a modern facility management with semantic technologies and different conceptual use cases of self-organisation in context of facility management. To support self-organized systems the adaption of contexts was described.

Finally, the mentioned use cases and the described approach of a facility management with semantic technologies and augmented reality will be continued to develop in the project “FMstar”. Currently, the project team develops conceptual use case definition. In the next steps the technical needs, implementation, evaluation and a case study will be occurred.

References

[AMR11] Anderl, R.; Mosch, C.; Rezaei, M.: Untersuchung der Notwendigkeit eines Fabrikdatenmanagements in der industriellen Praxis. In: Industrie Management, 27 (1). Gito Verlag, Berlin, 2011; S. 43-46.

[Ba13] Bast, H.: Semantische Suche. In: Informatik Spektrum, 36 (2). Springer-Verlag, Heidelberg, 2013; S. 136-143.

[BGW11] Bracht, U.; Geckler, D.; Wenzel, S.: Digitale Fabrik - Methoden und Praxisbeispiele. VDI-Buch. Springer Verlag, Berlin, Heidelberg, 2011.

[BHL01] Berners-Lee, T.; Hendler, J.; Lassila, O.: The Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. In: Scientific American, 284 (5). 2001; S. 34-43.

[Cl13] Clauß, M.: Methode zum Einsatz von Web 2.0-Werkzeugen in der Fabrikplanung. Dissertation, Chemnitz, 2013.

[Co14] Conrads, R.: In sieben Schritten zum erfolgreichen Big-Data-Projekt. In: Informatik Spektrum, 37 (2). Springer Verlag, Heidelberg, 2014; S. 127-131.

[De00] Dey, A. K.: Providing Architectural Support for Building Context-Aware Applications. Georgia Institute of Technology, Atlanta, 2000.

[G+14] Götze, J.; Ellmer, D.; Salzmann, P.; Schumann, C.-A.; Müller, E.; May, M.: FM Knowledge Modelling and Management by Means of Context Awareness and Augemented Reality. In: 13th EuroFM Research Symposium. 2014; S. 235-243.

[Ha05] Hahn, A.: Integration verteilter Produktmodelle durch Semantic-Web-Technologien. In: Wirtschaftsinformatik, 47 (4). Springer Gabler Verlag, Wiesbaden, 2005; S. 278-284.

[He03] Hernández Morales, R.: Systematik der Wandlungsfähigkeit in der Fabrikplanung. Fortschritt-Berichte VDI Reihe 16, Bd. 149. Springer VDI Verlag, Düsseldorf, 2003.

[Ko10] Koch, O.: Kontextorientierte Informationsversorgung in medizinischen Behandlungsprozessen. Vieweg+Teubner, Wiesbaden, 2010.

[Pa82] Patzak, G.; Systemtechnik - Planung komplexer innovativer Systeme. Grundlagen, Methoden, Techniken. Springer-Verlag, Berlin, Heidelberg, 1982.

[Sa11] Sauer, O. (2011): Informationstechnik in der Fabrik der Zukunft - Fabrik 4.0. Aktuelle Rahmenbedingungen, Stand der Technik und Forschungsbedarf. In: ZWF Zeitschrift für wirtschaftlichen Fabrikbetrieb, 106 (12). Hanser Verlag, München, 2011; S. 955-962.

[S+06] Schuh, G.; Gulden, A.; Gottschalk, S.; Kampker, A.: Komplexitätswissenschaft in der Fabrikplanung. In: wt Werkstatttechnik online 96 (4). Springer VDI Verlag, Düsseldorf, 2006; S. 167-170.

[SL09] Spath, D.; Lentes, J.: Ontologiegestützte kontinuierliche Produktionssystemplanung. In: wt Werkstatttechnik online, 99 (3). Springer VDI Verlag, Düsseldorf, 2009; S. 111-116.

[Vd08] VDI (Hrsg.): VDI 4499 - Digitale Fabrik: Grundlagen. Beuth Verlag, Berlin, 2008.

[Vd09] VDI (Hrsg.): VDI 4499 - Digitale Fabrik: Digitaler Fabrikbetrieb. Beuth Verlag, Berlin, 2009.

[WS02] Wilson, D.; Sperber D.: “Relecance Theory”, Horn, L.; Ward, G. L., Handbook of Pragmatics, Blackwell, Oxford; p. 251.

Semantic Service Manager – Enabling Semantic Web Technologies in Multi-Agent Systems

Nils Masuch[†], Christian Kuster[‡], and Sahin Albayrak[†‡]

[†]DAI-Labor, Technische Universität Berlin, Germany
Email: {firstname.lastname}@dai-labor.de
[‡]German-Turkish Advanced ICT Research Centre, Berlin, Germany
Email: {firstname.lastname}@gt-arc.com

Abstract: Computer systems are becoming more and more ubiquitous. Many of today's application areas do not fit for monolithic approaches, but for distributed intelligent systems. The multi-agent paradigm is an appropriate and therefore popular approach for building these systems. To interpret messages and provided services the agents need a common ontology. Solutions to integrate standards like OWL-S allowing for using semantic web services into agent frameworks have already been presented in recent years. However, they mostly focus on service matchmaking, neglecting the design process. As a result, semantic web services are not used in actual agent frameworks. In this paper, we introduce the Semantic Service Manager, a tool that enables for implementing semantic web services in the multi-agent framework JIAC V and therefore enables the dynamic integration of new services into complex systems.

1 Introduction

Due to the increasing complexity in today's software systems the distribution of their respective components has become a key attribute for successful software evolution. The reasons for that are manifold, however there are certain characteristics that are clearly advantageous compared to monolithic approaches. The modular software structure offers the possibility to easily reuse developed components in other systems. Further, the processor and memory load can be distributed over several hardware entities, in specific architectures even dynamically. Finally, the main advantage is the flexibility of a distributed system. The modular software structure namely allows to exchange specific modules by others, either by modifying the invocation code or by using an adaptable software solution that dynamically searches for appropriate alternatives.

However, an essential requirement that comes with distributed software is interoperability. On a basic level, this means that there have to be defined protocols on how information is exchanged. Going a level higher, the next step would be to have a common data type representation in order to be able to process the data. These can either be simple or complex data types. However, if we want to build intelligent systems, that are able to automatically understand the meaning of exchanged information and due to that being able to decide what next steps to do, we have the need of complex ontology representations and rules

that allow for detailed reasoning.

A popular approach for developing distributed systems is the multi-agent paradigm, in which multiple so-called software agents are either working collaboratively on a common goal or can be interpreted as opponents each of them trying to fulfil its own goals. Since one software agent usually is not able to solve a problem, the agents interact with each other and must thereby ideally behave autonomously and adaptable. However, this raises the challenge of understanding the intentions and functionalities of the other agents. As described before, the agents must have a common ontology and rule understanding for that. In current approaches, most of the agent functionalities are defined in proprietary languages making extensibility and interaction with other entities than software agents of a specific framework difficult.

The service-oriented computing (SOC) community developed well-established standards to describe and invoke functionalities, such as WSDL/SOAP and REST services. But furthermore, solutions for the semantic description of software services have been proposed in the last years. Amongst others, these are OWL-S [BHL+04], WSMO [WSM] and SAWSDL[1], which extend usual information by preconditions and effects and define input and output information in a more expressive ontology language, such as the Web Ontology Language (OWL) [MvH04].

Integrating standards like OWL-S or WSMO into the agent community were sighted theoretically but often discarded in practice. One of the reason is that many of the standards come with a complex syntactical structure and are therefore not easy to handle and an appropriate tool support is missing that is supporting the developers of multi-agent systems (MAS) by describing their functionalities semantically. It is our conviction that the integration of semantic web technologies into MAS will only be successful, if there is a comprehensive support. Therefore, in this paper we propose a software engineering component called Semantic Service Manager (SSM) that provides semi-automatic semantic annotation of agent functionalities for the multi-agent framework JIAC V [LKK+13]. The component enhances the services by the semantic service description ontology OWL-S, which is based on OWL and provides rule support in the Semantic Web Rule Language (SWRL) [HPSB+04]. Further, we show that with this tool support, we are able to develop software solutions, where agents can search for appropriate services with the help of a semantic service matchmaker, making a first step towards adaptable software agents that are also able to integrate external services that are using the same service expression language.

The rest of the paper is structured as follows. In chapter 2 we will give a detailed overview about multi-agent frameworks and their support for integration of semantic web technologies and discuss the current research status in the service oriented computing community. In chapter 3, we present our implementation, the SSM, and its integrated features. Chapter 4 describes how this tool can be used for the development of adaptable and dynamic multi-agent systems by showing a use case in the context of the multi-agent framework JIAC V. Finally we conclude and describe the next steps of our work.

[1] http://www.w3.org/2002/ws/sawsdl/

2 Related Work

Although the approach to combine the concept of semantic web services with agent technologies is becoming a more and more important research topic, the agent frameworks that are mostly used today don't offer any or inadequate support. In practice, extensions to known frameworks or framework-independent approaches are found.

Frameworks like *Jason* [BH04] or *3APL* [HDBVdHM99] are geared towards research and focus on cognitive concepts, especially planning. Communication is merely realized through speech acts to exchange knowledge between agents. The agents in these frameworks do not have the possibility to provide services to other agents.

Even though the agent framework *JACK* [Win05] is a more pragmatic and successful approach that is focused on industrial requirements, planning is still the main concept as it concentrates on the implementation of the BDI (Belief Desire Intention) architecture. Here, agents can communicate over simple speech acts by sending serialized Java objects. With a plug-in JACK is capable of FIPA speech acts, but as the authors do not see the demand for FIPA-compliance in industrial projects it isn't further integrated into the framework. There is no way to advertise services. JACK offers a plug-in that adds the concept of teams to the framework, allowing the agents that are grouped in teams to delegate tasks to sub teams. Using another plug-in agents can be accessed via the Java Servlet API to enable the implementation of web applications.

JADE (Java Agent DEvelopment Framework) [BPR99] as another often used agent framework employs services that are provided by agents. In every JADE system there is an agent called DF (Directory Facilitator) that holds the so-called yellow pages. This is a directory that contains provided services. These services are described by values such as *name*, *language*, *ontology* and self-defined properties of the service, therefore they cannot hold semantic content. The search in the running system is based on these values. However, extensions to JADE have emerged that focus on semantic web services. *SEMMAS (SEMantic web services and Multi-Agent System)* [GSMBVGFB08] is a framework based on the JADE platform that combines semantic web services and MAS. For the semantic annotations OWL-S, an ontology in OWL describing semantic web services, is used, and an external tool as editor. The authors focus on the service matchmaking, so the semantic markup of services is disregarded. *JADEX (JADE eXtension)* [BLP03] as another extension to JADE tries to bring JADE closer to the industry and integrates the BDI model into JADE. JADEX has a different approach for the implementation of services, namely *Contract Oriented Programming*. Services are Java classes that implement interfaces. The methods of these interfaces can be annotated with preconditions and effects, though, these expressions are limited to the parameter and results of the methods. JADEX provides rich tool support throughout the whole agent life cycle, but service descriptions have to be written in Java code. This poses a cumbersome design process. Furthermore, ontologies have to be transformed into JavaBeans with external editors or manually generated from scratch, decoupling them from the actual ontologies. This makes the use of JADEX inflexible. Nevertheless, JADEX is very close to the approach presented in this paper.

An agent-framework-independent approach is shown by McIlraith, Son and Zeng [MSZ01].

In their paper they use the agent programming language ConGolog [GLL00] in combination with DAML for semantic markup of web services. They use wrapper functions to translate the information given by web services at the html layer forth to DAML and back. This means there is no explicit support of the development of the services itself.

Independent from the research in agent technologies various research is done on semantic web services with the goal to combine both the semantic web [BLHL+01] and web services [BHM+04] to enable automatic service discovery by, e. g. agents. One concept to collect web services in a registry is UDDI (Universal Description, Discovery and Integration) [Coa00]. Although UDDI has not established itself as a global registry for web services as planned, the research in semantifying web services in UDDI shows up-to-date problems. Today, most web services only provide information about their functionality at a technical level via WSDL, but no semantics. Specifications like OWL-S were developed to fill this gap. Paolucci et al. [PKPS02] enrich UDDI with DAML-S, a predecessor to OWL-S. In their paper they show a mapping from DAML-S Profiles to UDDI records. Furthermore, they describe a matchmaker component. However, the approach doesn't provide any concept for the design and implementation of the semantic web services. Srinivasan et al. [SPS05] and Talantikite, Aissani and Boudjlida [TAB09] use OWL-S to semantify the web services in UDDI. Jaeger et al. [JRgL+05] show another approach, but their work only focuses on improving the matching algorithm.

Most of the presented papers merely focus on service matchmaking, but forget the design process, being just as important. But some focus on the semantic markup. The *OWL-S editor* [EDM+05] is a plug-in for Protégé, an open source ontology editor. With it a complete creation and editing of OWL-S descriptions is possible. Furthermore, the service behind the description can be tested via a GUI. Drawbacks of the OWL-S editor are that it cannot handle multiple ontologies, because of limitations of Protégé. There is no connection to a framework, meaning that the editor lacks usability. The authors also do not see the benefits of a Java-to-OWL transformation. They argue that most commonly the service is developed before the implementation in code. Another editor for OWL-S is the *OWL-S IDE* [SPS06]. It is a plug-in for Eclipse and, in contrary to the OWL-S editor, supports the generation of OWL-S skeletons out of Java code. However, this generation is limited to basic types due to the missing support of ontologies. Furthermore, it doesn't support preconditions nor effects.

As a result it can be stated that, though the agent community has gaining interest in the use of automatic discovery of services, most agent frameworks do not provide this functionality. Extensions to these frameworks as well as contributions from the SOA (service-oriented architecture) community show promising approaches, but merely focus on service matchmaking and mostly ignore the need for tools to implement semantic services. Existing tools are independent from frameworks, meaning a cumbersome work flow. However, adequate tools that integrate into the design process are essential for the acceptance in the industry.

3 The Semantic Service Manager

At the moment agent actions in JIAC V are implemented in plain Java code with the possibility to publish them under a name, simply represented by a string. This publication can be done via an annotation applied to the method that implements the action or via an explicit registration in the service directory of the agent node, the runtime environment for the agents. Hence, services can only be searched by their name, and even more, must be known by the agent at compile time, because such a representation doesn't provide any semantic information. The agent has no ability to understand actions provided in the system and is therefore not autonomous nor adaptive at all.

The *Semantic Service Manager (SSM)* is a tool aiming at the development process of agent-oriented software to enable the semi-automatic annotation of agent actions. The main goal of SSM is to semantify JIAC V actions so that they can be searched by semantic properties. This leads to MAS where agents are capable of adaptively using services. SSM focuses on the design of these semantic services and therefore provides a tool to create new services from scratch as well as add semantic information to existing JIAC V actions. Further, services can be loaded from an online source or a local file or be imported from an open Java file containing one or more agent actions. OWL-S is used as format for the service description since it is a standardised specification to describe semantic services that is based on OWL. The tool is capable of managing multiple services at once. An ontology management component holding loaded ontologies eases handling parameter types. To formulate preconditions and effects OWL-S allows for using different established rule languages. SSM implements a parser that supports SWRL. Finally, the semantic services can be saved as an OWL-S file. The designed services can later be searched by a semantic service matchmaker, which is not part of this paper. In the following sub sections the key concepts are explained in detail.

SSM is implemented as a plug-in for Eclipse and provides a view that integrates into the workbench. Figure 1 shows the workbench of an Eclipse instance with SSM view on the left and an open Java file on the right. The service description in SSM corresponds to the method in the Java file. The transformation from Java methods to service descriptions is further described in section 3.5. The view contains fields for the OWL-S file location, the name and the description of the service. Furthermore, the input and output parameters and preconditions can be chosen and effects for these parameters can be formulated. For the handling of ontologies *OWL API*[2] in version 3.4.4 is used.

3.1 Ontology Management

As OWL-S is based on OWL to reference the parameters and the relationships between them SSM provides an ontology management component. Using this component the user can browse through ontologies and see the different entities that are defined in them. For every entity properties that are defined on it are shown. Entities can then be selected

[2] http://owlapi.sourceforge.net/

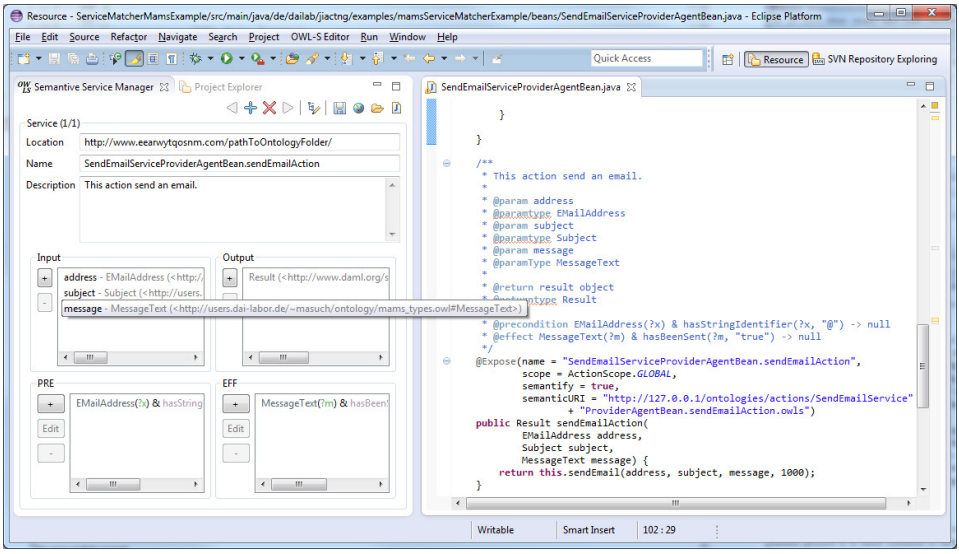


Figure 1: Semantic Service Manager: the service description (on the left) is derived from the Java method (on the right).

and used as parameters. To increase the usability and avoid unreferenced parameters they can only be selected from loaded ontologies. Figure 1 shows selected parameters for input and output. Here, the input parameter *message* refers to the ontology class *http://users.dai-labor.de/~masuch/ontology/mams_types.owl#MessageText*. New Ontologies are loaded into SSM either from an online or offline source and then are displayed in this component. Furthermore, ontologies that are defined in another format like Ecore (see 3.2) can be imported, too. The ontology storage is persistent and, thus, providing a database for OWL entities.

3.2 Ecore2OWL Transformation

In software engineering domain models are often used to describe a problem area. An open-source framework that has prevailed is the *Eclipse Modeling Framework (EMF)*[3]. In EMF domain models are built using the Ecore meta model, and the domain models are therefore called Ecore models. To integrate the possibility to use Ecore models SSM provides a *Ecore2OWL* transformation. Starting with an Ecore model rules are applied to the the model elements to transform them into elements of an OWL model. The transformation is realized using *ATL Transformation Language (ATL))*[4] in version 3.4.0 and based upon the implementation of Guillaume Hillairet[5]. The transformation produces a persis-

[3] http://www.eclipse.org/modeling/emf
[4] http://www.eclipse.org/atl/
[5] http://perso.univ-lr.fr/ghillair/projects/ecore2owl.zip

tent local OWL file that is loaded automatically into the ontology management component.

3.3 Semantic Service Description

To hold the information of the defined semantic services SSM implements a data structure *Semantic Service Description*, which abstracts from the format of the OWL-S specification and embodies only necessary fields to specify services. Then in the export process this service description is transformed into a valid OWL-S file. In SSM a service description consists of the following: A name, a human readable description and a location where the OWL-S file is stored. This location has to be a valid URL pointing to the ontology. Furthermore, it contains the input- and output parameters as well as preconditions and effects.

At the moment, preconditions can only be made about the input parameters, and effects must refer to the output parameters. This behaviour could be extended to a more general range of logic statements, though, possible benefits have yet to be elaborated.

3.4 SWRL Rule Parser

In OWL-S it is possible to formulate preconditions and effects for the services through SWRL. The language provides different concrete syntaxes to formulate rules as well as logic. SSM being a tool designed for human users, it features the informal so-called *Human Readable Syntax*. In that syntax, rules are of the form $antecedent \Rightarrow consequent$ where antecedent as well as consequent are conjunctions of properties written $p_1 \wedge \ldots \wedge p_n$. Variables in $p_{1\ldots n}$ are denoted by a leading question mark. In figure 1 the list for preconditions contains the rule $EMailAddress(?x) \wedge hasStringIdentifier(?x, \text{“@”}) \Rightarrow null$. Here, the consequent equals *null* meaning that there is no further logic to be reasoned. The example says that the variable x being of type *EMailAddress* has to include the string “@” as precondition. This requirement depicts a very simple check if the email address is valid.

For entering the SWRL rules a text dialogue can be used. Only properties that are defined for the input- and output parameters are available, avoiding errors due to chosen properties that are not allowed in the current context. The rules are checked by a parser that is implemented using *ANTLR*[6] in version 4.1.

3.5 Java2OWL-S Transformation

Although in a process planned from the beginning as a semantic service it is more common to develop the service before implementing the actual code, for us it has shown that in practice the development process often is reversed. For this reason we think that a se-

[6] http://www.antlr.org/

mantic annotation for already existing services is necessary for a complete support of the development process.

To keep the Java2OWL-S transformation simple, the Java file has to be opened in the editor (see figure 1). By starting the transformation all methods in the Java file that are annotated with "@Expose" are identified as JIAC V actions and therefore are selectable in a dialogue. For every selected method a separate service description is generated. As in Java annotations can have a list of key-value pairs in the transformation process it is checked whether a OWLS-S file already exists (*semanticURI* in figure 1). If this file exists and can be found it will be loaded into SSM, otherwise the service description will be generated from the Java method. For this purpose the types of the parameters and the return type are matched against the available classes from the ontology management component. The name is extracted from the Java annotation, and the preconditions and effects are imported from the Javadoc tags "@precondition" and "@effect" and checked by the SWRL parser. When changes are saved in the imported service description the Java method gets updated as well.

This Java2OWL-S transformation is generic enough to be used for every Java method. However, in our implementation it is limited to methods annotated with "@Expose" since in JIAC V agent actions are marked this way.

4 SSM in Practice

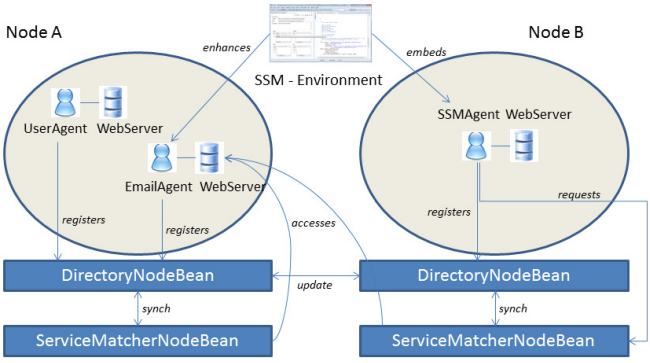


Figure 2: The SSM embedded in the JIAC V environment

4.1 Integration in JIAC V Methodology

One of the goals of SSM was its support and integration into the development methodology of the multi-agent framework JIAC V [LKK+13]. The framework itself comes up with a powerful discovery and messaging infrastructure, which allows to build up dis-

tributed systems even over network boundaries. In JIAC V, a distributed system is called platform. Each platform can contain multiple nodes (as seen in figure 2), which can run on several server. A node itself can host multiple agents, which can represent certain roles and provide different services. The system is flexible in a way that new agents can be deployed at runtime and are immediately able to communicate with other agents. Furthermore, nodes can implement so-called *Node Beans*, which provide functionalities for all agents on the node. Figure 2 shows a system with two nodes. Both nodes do provide two *Node Beans*, namely the *Directory Node Bean* and the *Service Matcher Node Bean*. The first one is responsible for the advertisement of services between the nodes and the second one is an implementation of an OWL-S based Service Matcher component called $SeMa^2$ [MHB+12]. In the figure, Node A contains two agents, the *UserAgent* and the *EMailAgent*. When an agent is being started it registers all its actions at the *Directory Node Bean*. If the action does contain a semantic service description it is being hosted by the agent on a web server and the address is communicated to the *Service Matcher Node Bean*. The Service Matcher then accesses this description and stores it in its repository of available services. On the other hand, the *Directory Node Bean* and the *Service Matcher Node Bean* of the other node are also being informed about the new services.

The SSM can now support the developer in two ways. On the one hand it can directly help annotating an action at design time. In this case the agent is not running and the service description is annotated as described in the previous chapter. The agent will load this description into its web server at start-up and will be available for the Semantic Service Matcher component. On the other hand, the SSM shall support the developer by finding an existing service that can be invoked. This option, we are currently working on, embeds its own agent, the SSMAgent. As the agent also has access to the service matcher component, which synchronises with all the other nodes via the *Directory Node Bean*, it can search for specific input, output, precondition or effect parameter. The Service Matcher will list all possible solutions and the developer can select the most appropriate one for integration.

4.2 Use Case for Semantic Service Descriptions for Agents

The overview of the SSM tool in figure 1 shows the semantic annotation of an email service of the *EMailAgent* shown in figure 2. The service has a precondition described in SWRL checking whether the mail address is valid. Further is has the effect that after invocation in successful cases the message will be sent to the recipient. Now, if the developer annotates its email service, another agent, for example the *UserAgent*, can either search via the SSM for a service at design time that delivers messages or it is also possible to integrate a search template into the code of the *UserAgent*. In this case the Semantic Service Matchmaker can reason on these information and as soon as the email service is online the service can be matched and invoked by the *UserAgent*.

5 Conclusion

Although the agent community has already revealed semantic web services as a promising concept to enable agents to understand intentions and functionalities of other agents, due to the lack of appropriate tools the potential of this concept is not used. The Semantic Service Manager (SSM) tries to fill this gap and and aims at the development of semantic web services in the agent framework JIAC V. It enables designing new services as well as semantifying existing agent actions. With the standard OWL-S actions can be described using ontologies and moreover preconditions and effects can be formulated in SWRL.

SSM is a promising tool for developing semantic web services in JIAC V as it allows for using ontologies in the context of agent actions to describe preconditions and effects at all. But even though the tool support for the design process already seems solid, many issues are undissolved. Currently, SSM lacks the distribution of the generated service descriptions in JIAC V. The concept to store and publish the service descriptions, as proposed in Section 4.1, has to be further elaborated and then implemented. An open question is here how to realize a lightweight web server component that is able to provide the OWL-S file.

Future work on SSM, besides the distribution of semantic service descriptions, can be the integration of the semantic service matchmaker at design time. This allows the developer to enter a service description into SSM and then let it search for services that match that description in running instances of JIAC V MAS. Additionally all running services can be visualised like in the ontology management component. This two solutions ease the design and gearing of new semantic services. Further transformations from WSDL to OWL-S and XSD to OWL are also possible next steps to achieve a greater support of different formats. Our goal is to combine the various different formats into OWL respectively OWL-S.

Acknowledgment

This work is partially funded by the German Federal Ministry of Education and Research under the funding reference numbers 01IS12049A and 01IS12049B.

References

[BH04] Rafael H Bordini and Jomi F Hübner. A Java-based agentSpeak interpreter used with saci for multi-agent distribution over the net, 2004.

[BHL+04] Mark Burstein, Jerry Hobbs, Ora Lassila, Drew Mcdermott, Sheila Mcilraith, Srini Narayanan, Massimo Paolucci, Bijan Parsia, Terry Payne, Evren Sirin, Naveen Srinivasan, and Katia Sycara. OWL-S: Semantic Markup for Web Services. Website, November 2004.

[BHM+04] David Booth, Hugo Haas, Francis McCabe, Eric Newcomer, Michael Champion, Chris Ferris, and David Orchard. Web services architecture. 2004.

[BLHL+01] Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.

[BLP03] Lars Braubach, Winfried Lamersdorf, and Alexander Pokahr. Jadex: Implementing a BDI-Infrastructure for JADE Agents, 2003.

[BPR99] Fabio Bellifemine, Agostino Poggi, and Giovanni Rimassa. JADE–A FIPA-compliant agent framework. In *Proceedings of PAAM*, volume 99, page 33. London, 1999.

[Coa00] UDDI Coalition. UDDI Technical White Paper, 2000.

[EDM+05] Daniel Elenius, Grit Denker, David Martin, Fred Gilham, John Khouri, Shahin Sadaati, and Rukman Senanayake. The OWL-S editor–a development tool for semantic web services. In *The Semantic Web: Research and Applications*, pages 78–92. Springer, 2005.

[GLL00] Giuseppe De Giacomo, Yves Lespérance, and Hector J. Levesque. ConGolog, a concurrent programming language based on the situation calculus. *Artificial Intelligence*, 121(12):109 – 169, 2000.

[GSMBVGFB08] Francisco García-Sánchez, Rodrigo Martínez-Béjar, Rafael Valencia-García, and Jesualdo T. Fernández-Breis. Knowledge Technologies-Based Multi-Agent System for Semantic Web Services Environments. In Leszek Rutkowski, Ryszard Tadeusiewicz, LotfiA. Zadeh, and JacekM. Zurada, editors, *Artificial Intelligence and Soft Computing ICAISC 2008*, volume 5097 of *Lecture Notes in Computer Science*, pages 1222–1233. Springer Berlin Heidelberg, 2008.

[HDBVdHM99] Koen V Hindriks, Frank S De Boer, Wiebe Van der Hoek, and John-Jules Ch Meyer. Agent programming in 3APL. *Autonomous Agents and Multi-Agent Systems*, 2(4):357–401, 1999.

[HPSB+04] Ian Horrocks, Peter F Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, Mike Dean, et al. SWRL: A semantic web rule language combining OWL and RuleML. *W3C Member submission*, 21:79, 2004.

[JRgL+05] Michael C. Jaeger, Gregor Rojec-goldmann, Christoph Liebetruth, Gero Mühl, and Kurt Geihs. Ranked Matching for Service Descriptions using OWL-S, 2005.

[LKK+13] Marco Lützenberger, Tobias Küster, Thomas Konnerth, Alexander Thiele, Nils Masuch, Axel Heßler, Michael Burkhardt, Jakob Tonn, Silvan Kaiser, and Jan Keiser. Engineering Industrial Multi-Agent Systems — The JIAC V Approach. In Massimo Cossentino, Amal El Fallah Seghrouchni, and Michael Winikoff, editors, *Proceedings of the 1st International Workshop on Engineering Multi-Agent Systems (EMAS 2013)*, pages 160–175, 2013.

[MHB+12] Nils Masuch, Benjamin Hirsch, Michael Burkhardt, Axel Heßler, and Sahin Albayrak. *Semantic Web Services: Advancement through Evaluation*, chapter SeMa^2 – a Hybrid Semantic Service Matching Approach. Springer-Verlag, Berlin / Heidelberg, May 2012.

[MSZ01] S.A. McIlraith, Tran Cao Son, and Honglei Zeng. Semantic Web services. *Intelligent Systems, IEEE*, 16(2):46–53, Mar 2001.

[MvH04] Deborah McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. Technical report, W3C, 2004. http://www.w3.org/TR/owl-features/.

[PKPS02] Massimo Paolucci, Takahiro Kawamura, Terry R. Payne, and Katia Sycara. Importing the Semantic Web in UDDI. In *Web Services, E-Business, and the Semantic Web*. Springer, 2002.

[SPS05] Naveen Srinivasan, Massimo Paolucci, and Katia Sycara. An efficient algorithm for OWL-S based semantic search in UDDI. In *Semantic Web Services and Web Process Composition*, pages 96–110. Springer, 2005.

[SPS06] Naveen Srinivasan, Massimo Paolucci, and Katia Sycara. Semantic Web Service Discovery in the OWL-S IDE. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 06*, HICSS '06, pages 109.2–, Washington, DC, USA, 2006. IEEE Computer Society.

[TAB09] Hassina Nacer Talantikite, Djamil Aissani, and Nacer Boudjlida. Semantic annotations for web services discovery and composition. *Computer Standards & Interfaces*, 31(6):1108 – 1117, 2009.

[Win05] Michael Winikoff. JACK(TM) intelligent agents: An industrial strength platform. In *Multi-Agent Programming*, pages 175–193. Springer, 2005.

[WSM] Web Service Modeling Ontology (WSMO). http://www.w3.org/Submission/WSMO/.

Recht und Technik: Datenschutz und Informationsmanagement im Diskurs

(RuT)

Recht und Technik: Datenschutz und Informationsmanagement im Diskurs (RuT)[1]

Abstract: Der Workshop will Informatiker, Techniker und Juristen zusammenbringen, die an Fragestellungen des Datenschutzes in und durch Technik arbeiten. Insbesondere waren Beiträge erwünscht, die anwendungsorientiertes Potential für interdisziplinären Diskurs und Zusammenarbeit bieten. Schwerpunktmäßig soll die Frage behandelt werden, wie Datenschutz durch Technik präzisiert und umgesetzt werden kann und welche Herausforderungen neue Technikentwicklungen für das Datenschutzrecht stellen.

Organisatoren: Prof. Dr. iur. Indra Spiecker genannt Döhmann, (LL.M. (Georgetown Univ.), Goethe-Universität Frankfurt), Prof. Dr. Christoph Sorge (Universität des Saarlandes)

Programmkomitee: Dr. Jens-Matthias Bohli (NEC Laboratories Europe), Thorsten Feldmann (JBB Rechtsanwälte Berlin), Prof. Dr. Rüdiger Grimm (Universität KoblenzLandau), Prof. Dr. Nils Gruschka (Fachhochschule Kiel), Prof. Dr. Gerrit Hornung (Universität Passau), Dr. Florian Kerschbaum (SAP Research), Jun.-Prof. Dr. Artus Krohn-Grimberghe (Universität Paderborn), Dr. Thomas Lapp (Dr. Lapp Rechtsanwälte und Nifis e.V.), PD Dr. Kai von Lewinski (Humboldt-Universität zu Berlin), Dr. Mario Lischka (acentrix), Prof. Dr. Frank Pallas (TU Berlin), Prof. Dr. Burkhard Schäfer (University of Edinburgh), Prof. Dr. Gernot Sydow (Justitiar Bistum Limburg und Universität Freiburg), Prof. Dr. Dirk Westhoff (Hochschule Furtwangen)

Der Workshop knüpft an die vorangegangene Veranstaltung auf der Jahrestagung 2013 der Gesellschaft für Informatik an. Auch in diesem Jahr konnten Beiträge sowohl von Juristen als auch von Informatikern gewonnen werden. Sie zeigen einmal mehr auf, dass viele Fragestellungen des Datenschutzes nicht mehr nur innerhalb einer Disziplin gelöst werden können und dass nur im Diskurs praktikable und rechtssichere Lösungen für anstehende Problemlagen gefunden werden können.

Den Anfang machen in diesem Jahr zwei Beiträge, die sich mit der datenschutzrechtlichen Einwilligung befassen, nach wie vor einem der zentralen Gestaltungsinstrumente von rechtlicher Seite. *Bräuchle* untersucht aus juristischer Sicht, wie eine solche Einwilligung in Smart-Metering-Systemen aussehen kann; *Gruschka* und *Jensen* stellen dar, wie Prozessmodellierung und Einwilligungsmanagement integriert werden können.

In den folgenden beiden Beiträgen zeigt sich, wie das Recht auf technische Entwicklungen reagiert. *Brummund* befasst sich mit datenschutzrechtlichen Risiken von Smartphone-Apps und geht dabei auch auf technische (Selbst-)Schutzmaßnahmen ein.

[1] http://www.cs.uni-paderborn.de/?rechtundtechnik2014

Das „Big Data“-Paradigma wird von *Kring* mit Blick auf den Konflikt zum Zweckbindungsgrundsatz beleuchtet.

Den Abschluss bilden zwei Beiträge mit technischem Schwerpunkt: *Huber et al*. stellen eine Architektur für datenschutzgerechte Videoüberwachung in einem Assisted-Living-Szenario dar, und Rodeck et al. präsentieren Werkzeuge zur Unterstützung von Datenschutzbeauftragten bei *Datenschutz*- und Sicherheits-Assessments.

Insgesamt spricht der Workshop damit eine Reihe von aktuellen Einzelfragen im IT-Bereich an (Videoüberwachung, vernetzte Systeme, Energiewende, Einwilligung, Kontrollinstrumente, Apps …), verbindet diese jedoch immer auch in die übergeordneten Fragestellungen (Big Data, Ausmaß an Selbstbestimmung, Prozesshaftigkeit, Effektivität von Recht und Technik).

Die datenschutzrechtliche Einwilligung in Smart Metering Systemen – Kollisionslagen zwischen Datenschutz- und Energiewirtschaftsrecht

Thomas Bräuchle, Ref. iur.

Karlsruher Institut für Technologie (KIT)
Zentrum für Angewandte Rechtswissenschaft (ZAR)
Forschungsgruppe Energieinformationsrecht und Neue Rechtsinformatik
Vincenz-Prießnitz-Straße 3
76131 Karlsruhe
thomas.braeuchle@kit.edu

Abstract: Die Energieversorgung der Zukunft und die Etablierung von intelligenten Stromnetzen (Smart Grids) soll mit der Umsetzung eines intelligenten Messwesens (Smart Metering) sicher und effizient ausgestaltet werden. Damit einher ging in der Vergangenheit die Umsetzung verschiedener europarechtlicher Vorgaben im Bereich des Energiewirtschaftsrechts. Vor allem die Einführung des Smart Metering in Deutschland und die damit verbundene Novellierung des Energiewirtschaftsgesetzes (EnWG) im Jahre 2011, wirft nach wie vor viele datenschutzrechtliche Fragen auf. Insbesondere werden die bereichsspezifischen Datenschutzregelungen des EnWG kontrovers diskutiert. Problematisch dabei ist, dass das Datenschutzrecht auf die wettbewerbsrechtlichen Regelungen des Energiemarktes trifft. In der Konsequenz können diese unterschiedlichen Rechtsmaterien neuartige Kollisionslagen hervorrufen, die geeignet sind den Schutzzweck des Datenschutzes zu konterkarieren.

1 Einleitung und Motivation

Unter dem Thema Big Data werden einerseits Chancen und Vorteile im Hinblick auf die Weiterentwicklung der digitalen Gesellschaft, andererseits aber auch potentielle Gefahren für die (Grund-)Rechte der betroffenen Personen kontrovers diskutiert.[1] Aus rechtlicher und technischer Sicht ergibt sich die Herausforderung, die Sammlung, Auswertung und Zurverfügungstellung strukturierter und unstrukturierter Daten aus unterschiedlichen Quellen zu bewältigen sowie einen Zweckmissbrauch der personenbezogenen Daten[2] zu verhindern.[3] Neben der Verwaltung und Auswertung datenschutzrechtlich relevanter Quellen, muss das vom Bundesverfassungsgericht

[1] Vgl. [We13a], S. 251 m.w.N.
[2] Siehe § 3 Abs. 1 Bundesdatenschutzgesetz (BDSG).
[3] [We13a], S. 251.

(BVerfG) im Jahre 1983 entwickelte Recht auf informationelle Selbstbestimmung[4] geschützt sowie gleichzeitig eine effektive Ausübung dieses Rechts gewährleistet werden. Die Probleme des klassischen Datenschutzrechts in Bezug auf Big Data sind in der Literatur schon mannigfaltig adressiert worden.[5] Insbesondere in Bezug auf soziale Netzwerke wird festgestellt, dass die Verbraucher als Teilnehmer solcher Plattformen nicht mehr nur passive Informationsnutzer, sondern vielmehr auch aktive Informationsanbieter seien.[6] Führt man sich dann vor Augen, dass eine datenschutzrechtliche Würdigung solcher Sachverhalte nach deutschem Recht im Wesentlichen (noch) nach dem Datenschutzkonzept des BDSG von 1990 durchzuführen ist, so kann die Kritik an einem veralteten nationalen Datenschutzrecht nicht ganz einfach von der Hand gewiesen werden. Auch die Europäische Datenschutz-Richtlinie 1995/46/EG basiert noch auf technischen Standards der 1970er Jahre, in denen Großrechnersysteme das Bild der automatisierten Datenverarbeitung der verantwortlichen Stellen prägten.[7] Aus rechtswissenschaftlicher Sicht ist zudem festzustellen, dass grundlegende Konzepte unterschiedlicher Regelungsmaterien einem stetig fortschreitenden Paradigmenwechsel unterliegen.[8] So weicht die gesellschaftliche und technische Entwicklung in vielen Bereichen immer stärker von der ursprünglichen Intention des Gesetzgebers zu den jeweiligen Regelungsmaterien ab. Problematisch an dieser Debatte im Kontext Big Data ist jedoch, dass das Auseinanderklaffen zwischen dem „legislatorischen status quo“ und den tatsächlich herrschenden Entwicklungen hauptsächlich auf „Internetsachverhalte“ insbesondere in Verbindung mit den Rechtsfragen der sozialen Netzwerke bezogen wird.[9] Der Ansatz des vorliegenden Beitrages ist daher, den durch Festlegungen der Bundesnetzagentur (BNetzA) regulierten[10] und daher „kontrollierbaren“ Energiemarkt in den Fokus zu nehmen und hieran die klassischen datenschutzrechtlichen Regelungen im Kontext vernetzter Infrastrukturen zu untersuchen. Das Energiewirtschaftsrecht ist zwar eine originär dem Wettbewerbsrecht zugeordnete Rechtsmaterie[11], jedoch handelt es sich bei den Festlegungen der BNetzA (noch) um weitgehend durch Verwaltungsverfahren geregelte Marktkommunikationsprozesse. Diese Prozesse werden nunmehr aufgrund europarechtlicher Vorgaben des Dritten Energiepakets der Europäischen Union[12] durch die Integration von Informations- und Kommunikationstechnologie (IKT) zu einer kommunikativ vernetzten intelligenten Infrastruktur ausgeweitet.

Der vorliegende Beitrag führt daher im ersten Schritt zunächst in die datenschutzrechtlichen Grundlagen des Energiewirtschaftsrechts ein (Kap. 2). Daran anschließend wird ein Referenzszenario gebildet, das beispielhaft Schutzlücken aus den Datenschutzregelungen des EnWG aufzeigen soll (Kap. 3). Schwerpunktmäßig wird

[4] [Bu83], S. 1 ff. (Volkszählung).
[5] Vgl. etwa [Wi14], S. 36.
[6] [Wi14], S. 36.
[7] [Wi14], S. 35.
[8] Vgl. dazu [Ra11], S. 3 f.
[9] [Wi14], S. 36.
[10] Beschlüsse der BNetzA sind Festlegungen nach § 29 EnWG, die nach [BGH08], S. 228 als Verwaltungsakte im Sinne der Allgemeinverfügung gemäß § 35 S. 2 VwVfG zu qualifizieren sind.
[11] Dazu [Th13], S. 10 Rn. 27.
[12] Siehe http://ec.europa.eu/energy/gas_electricity/legislation/third_legislative_package_de.htm [abgerufen am 18.06.2014].

dabei das Problem der sogenannten innerfamiliären Überwachung in den Fokus genommen und im Hinblick auf die datenschutzrechtlichen Einwilligungskonstellationen des Energiewirtschaftsrechts untersucht (Kap. 4). Anhand dieses Referenzszenarios sollen schließlich zentrale Herausforderungen aus rechtlicher und technischer Sicht erläutert werden, die sich aus der Kollisionslage zwischen dem Energiewirtschaftsrecht und den datenschutzrechtlichen Belangen ergeben (Kap. 5). Der Beitrag schließt mit einem Fazit und einem Ausblick (Kap. 6).

2 Datenschutz und Energiewirtschaftsrecht

2.1 Europarechtliche Vorgaben

Das Dritte Energiepaket der EU sieht neben der weiteren Liberalisierung der Strom- und Gasmärkte zwar auch die Stärkung der Verbraucherrechte vor, jedoch wurde unter Rückgriff auf die Rechtsgrundlage des ex-Art. 95 Abs. 1 EG-Vertrag[13] lediglich die Errichtung und das Funktionieren des europäischen Binnenmarktes motiviert und legitimiert. Insbesondere die Umsetzung der Elektrizitätsbinnenmarkt-RL 2009/72/EG in nationales Recht zeigt jedoch, dass der Rückgriff auf die Belange des Binnenmarktes unter dem Fokus der Versorgungssicherheit mit elektrischer Energie nicht alle Grundbedürfnisse und Verbraucherrechte, die es zu gewährleisten gilt, ausreichend berücksichtigt. Unter dem Gesichtspunkt der Steigerung der Energieeffizienz sollen nach Erwägungsgrund Nr. 11 der Elektrizitätsbinnenmarkt-RL unter anderem intelligente Messsysteme eingeführt werden, um den Elektrizitätsunternehmen eine Optimierungsmöglichkeit bezüglich des Stromverbrauchs zu gewährleisten. Diese Vorgabe steht auf nationalstaatlicher Ebene insbesondere mit der Rechtsprechung des BVerfG in Einklang, da die Sicherstellung der Energieversorgung „ein Gemeinschaftsinteresse höchsten Ranges“ ist[14], mithin „ein von der jeweiligen Politik des Gemeinwesens unabhängiges ‚absolutes‘ Gemeinschaftsgut“ darstellt.[15] Damit ordnet das BVerfG die Energieversorgung dem Bereich der Daseinsvorsorge zu, die „zur Sicherung einer menschenwürdigen Existenz unumgänglich“ ist.[16]

2.2 Novellierung des EnWG in Deutschland

Im Zuge der Umsetzung der europarechtlichen Vorgaben in nationales Recht, wurde das EnWG im Jahre 2011 novelliert.[17] Da jedoch nach den umzusetzenden Vorgaben der Elektrizitätsbinnenmarkt-RL die Einführung eines intelligenten Messwesens angestrebt werden soll, werden aus nationalrechtlicher Perspektive nunmehr Belange des Datenschutzes relevant. Denn in einem Energieinformationsnetz fallen regelmäßig personenbezogene Daten zwischen Verbrauchern, Energieversorgungsunternehmen und

[13] Nunmehr Art. 114 Abs. 1 des Vertrags über die Arbeitsweise der Europäischen Union (AEUV).
[14] [Bu71], S. 323 f. (Erdölbevorratung).
[15] [Bu71], S. 324 (Erdölbevorratung).
[16] [Bu84], S. 258; [Th13], S. 6, Rn. 15 m.w.N.
[17] BGBl. I 2011, 1554.

Messstellenbetreibern (MSB) an.[18] Um die Einführung eines intelligenten Messwesens entsprechend der verfassungsrechtlichen Vorgaben nach deutschem Recht konform umzusetzen, wurden daher in den §§ 21c ff. EnWG bereichsspezifische Neuregelungen zum Datenschutz eingefügt.[19] In der Begründung zum Entwurf des EnWG wird demgemäß neben der Förderung der modernen intelligenten Messsysteme[20] nunmehr explizit auch die Verankerung eines den Anforderungen von Datenschutz und Datensicherheit genügenden Smart Metering angeführt.[21] Dies ist insbesondere vor dem Hintergrund bedeutsam, dass der Gesetzgeber mit der Neuregelung des § 21c Abs. 1 EnWG erstmals auch für Letztverbraucher, die einen Jahresverbrauch von 6.000 Kilowattstunden erreichen, die Pflicht zur Nutzung von Informations- und Kommunikationstechnologien verankert. Mit der Einbaupflicht nach § 21c Abs. 1 EnWG ist daher die Grundlage für eine Gefährdung des Rechts auf informationelle Selbstbestimmung gesetzt, da hierdurch die Befugnis des Einzelnen, „grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen", eingeschränkt werden kann.[22] Aus diesem Grund komme dem Gesetzgeber die Pflicht im Sinne einer Folgenverantwortung zu, indem er dem Verbraucher damit korrespondierende wirksame Schutzmechanismen an die Hand gibt.[23]

3 Referenzszenario: Visualisierung des Energieverbrauchs

Aus der Perspektive des Energiewirtschaftsrechts sind die denkbaren Fallkonstellationen des Umgangs mit personenbezogener Daten aufgrund der datenschutzrechtlichen Zwecknormierungen in § 21g Abs. 1 EnWG abschließend vorgegeben.[24] Als Referenzszenario soll nachfolgend der Fall der Veranschaulichung des Energieverbrauchs zugrunde gelegt werden. Nach § 21g Abs. 1 Nr. 6 Alt. 2 EnWG darf „die Erhebung, Verarbeitung und Nutzung personenbezogener Daten aus dem Messsystem […] ausschließlich durch zum Datenumgang berechtigte Stellen erfolgen und auf Grund dieses Gesetzes nur, soweit dies erforderlich ist für […] die Umsetzung der Veranschaulichung des Energieverbrauchs […]." Diesem Szenario wird zudem die Annahme zugrunde gelegt, dass nach § 21c Abs. 1 lit. b) EnWG die Pflicht zum Einbau eines Messsystems besteht, da bei dem betroffenen Letztverbraucher[25] von einem Jahresverbrauch von größer als 6.000 Kilowattstunden ausgegangen wird. Im Rahmen des Zwecks der Veranschaulichung des Energieverbrauchs geht es also darum, den tatsächlichen Energieverbrauch und die tatsächliche Nutzungszeit nach den Funktionalitäten des § 21d Abs. 1 Hs. 2 EnWG widerzuspiegeln.

[18] [JRV11], S. 100.
[19] Siehe [BR11], S. 192 ff.; ausführlich hierzu [Ra11a], S. 831 ff. und [JRV11], S. 99 ff.
[20] Siehe Legaldefinition in § 21d Abs. 1 EnWG.
[21] [BR11], S. 192.
[22] [JRV11], S. 99.
[23] [Ra11a], S. 831.
[24] [Ra11a], S. 835 mit Verweis auf [BR11], S. 202.
[25] Siehe Legaldefinition in § 3 Nr. 27 EnWG.

3.1 Legislative Regelungsintention

Die Einführung der Veranschaulichung des Energieverbrauchs für Messsysteme geht auf die Elektrizitätsbinnenmarkt-RL zurück. Danach wird als Maßnahme zur Förderung der Energieeffizienz empfohlen, dass „Elektrizitätsunternehmen den Stromverbrauch optimieren, indem sie […] intelligente Messsysteme oder intelligente Netze einführen.“[26] Ferner wird in der Energieeffizienz-Richtlinie 2012/27/EU vorgeschrieben, dass in der Regel „individuelle Zähler“ zu verwenden sind, die „den tatsächlichen Energieverbrauch des Endkunden genau widerspiegeln und Informationen über die tatsächliche Nutzungszeit bereitstellen.“[27] Mit der Einführung von intelligenten Stromzählern sollte daher eine Sensibilisierung der Letztverbraucher bezüglich ihres Energieverbrauchs angestrebt werden. Dementsprechend wird in den Gesetzesmaterialien ausdrücklich das Ziel anvisiert, „Verbrauchern über die neue Technologie Energieeinsparungen und die Teilhabe am intelligenten Netz zu ermöglichen.“[28] Außerdem wird in der Gesetzesbegründung weiter ausgeführt, „Messsysteme können einen direkten und anschaulichen Zugang zu Verbrauchsdaten und variablen Tarifen eröffnen, die ein intelligentes Energiemanagement mit den Versorgungsbedürfnissen der Verbraucher verknüpfen, sowie das Verbrauchs- und Kostenbewusstsein stärken“.[29] Diese Vorgabe findet ihren gesetzlichen Niederschlag nunmehr in der Legaldefinition des Messsystems nach § 21d Abs. 1 EnWG anhand der Beschreibung der Funktionalität bezüglich des Widerspiegelns des tatsächlichen Energieverbrauchs und der tatsächlichen Nutzungszeit.

3.2 Technische Möglichkeiten der Verbrauchsvisualisierung

Konkrete Anforderungen an die Umsetzung der Verbrauchsvisualisierung sind nach dem Wortlaut der § 21g Abs. 1 Nr. 6 Alt. 2 i.V.m. § 21d Abs. 1 EnWG zunächst nicht ersichtlich. Jedoch beschreibt die BNetzA den Begriff des Widerspiegelns dahingehend, dass dem Anschlussnutzer die von der Messeinrichtung ermittelten Werte angezeigt werden müssen, indem eine Darstellungsform gewählt wird, die dem Anschlussnutzer die Informationen in angemessener Form visualisiert.[30] Unter einer geeigneten Visualisierung der feingranularen Verbrauchswerte wird dabei die Anzeige über eine lokale Anzeigeeinheit verstanden, die entweder direkt an den Verbrauchsgeräten oder aber über ein „Home-Display“ bzw. ein mobiles Endgerät eingesehen werden könnten.[31] In den Gesetzgebungsmaterialien zur Novellierung des EnWG aus dem Jahre 2011 wird darüber hinaus eine externe Veranschaulichung des Energieverbrauchs explizit in Erwägung gezogen. Danach ist es denkbar, dass eine „technische Ausführung“ gewählt wird, „die für eine Darstellung in einem Energieverbrauchsportal für den Verbraucher darauf angewiesen ist, dass beispielsweise ¼-Stunden-Werte an eine zentrale externe Servereinheit übertragen werden.“[32] Demnach könnten die individuellen Energieverbrauchsdaten eines Letztverbrauchers an einen externen Marktteilnehmer

[26] Elektrizitätsbinnenmarkt-RL 2009/72/EG, Erwägungsgrund Nr. 11.
[27] Siehe dazu Artikel 9 („Verbrauchserfassung“) der Richtlinie 2012/27/EU.
[28] [BR11], S. 192; [BT11], S. 76.
[29] Ebd.
[30] [BNA09], S. 3.
[31] [Ra11], S. 23.
[32] Funktionalitätsbeschreibung nach [BR11], S. 195.

(z.B. an den MSB oder Lieferanten, vgl. § 21g Abs. 2 EnWG) übermittelt und dort zur Visualisierung aufbereitet werden. Darüber hinaus liegt seit dem Jahre 2013 unter Bezugnahme auf die Ermächtigungsgrundlage des § 21i Abs. 1 Nr. 12 EnWG der Referentenentwurf einer Messsystemverordnung (MsysV-Entwurf) vor. Der MsysV-Entwurf beinhaltet nach § 3 Abs. 1 Nr. 2 Buchstabe a) bis e) Konkretisierungen bezüglich der verschiedenen Kategorien von anzuzeigenden Informationen, wie beispielsweise historische Energieverbrauchswerte, die tatsächliche Nutzungszeit oder abrechnungsrelevante Tarifinformationen. In Bezug auf eine externe Verbrauchsvisualisierung geht aus der Begründung zu § 3 Abs. 1 Nr. 2 MsysV-Entwurf hervor, dass Messsysteme neben den verschiedenen in Buchstabe a) bis e) genannten Anzeigeoptionen auch technisch in der Lage sein müssen, die Anzeigemodalität der externen Verbrauchsvisualisierung „über das Internet“ zu ermöglichen.[33] Damit würde die externe Verbrauchsvisualisierung als technischer Standard eingeführt werden.

3.3 Datenschutzrechtliche Einordnung der Verbrauchsvisualisierung

Im Falle einer lokalen Anzeige des Verbrauchs über eine Anzeigeeinheit werden noch keine Informationen über eine Kommunikationsschnittstelle der Messeinrichtung nach außen gegeben. Es ist daher davon auszugehen, dass die Messwerte des Letztverbrauchers lediglich geräteintern aufbereitet und auf der Anzeigeeinheit lokal angezeigt werden. Daher kann an diesem Punkt der Informationsverarbeitung noch nicht von einer Erhebung personenbezogener Daten nach § 21 g Abs. 1 EnWG gesprochen werden.[34] Entsprechend des Begriffsverständnisses der BNetzA bzw. der Gesetzgebungsmaterialien könnte die externe Verbrauchsvisualisierung über ein Webportal bzw. eine externe Servereinheit anders zu beurteilen sein. Die Energiedaten könnten in diesem Fall von der Messeinrichtung über ein Kommunikationsnetz im Wege der Fernauslesung an den MSB übermittelt werden. In dieser Konstellation ist jedoch fraglich, wie der innerfamiliäre Schutz vor einer Überwachung durch den sogenannten Haushaltsvorstand vermieden werden kann, da davon auszugehen ist, dass eine externe Verbrauchsvisualisierung den Zugriff auf aktuelle Verbrauchsdaten und Nutzungsprofile möglich macht. Die datenschutzrechtliche Problematik verschärft sich zudem dadurch, dass der Abruf der Verbrauchsdaten über einen externen Server jederzeit weltweit möglich wäre.[35] Hierzu führt die BNetzA lediglich aus, dass „[…] soweit dabei die Werte zur Kenntnis Dritter gelangen, es dazu des Einverständnisses des Anschlussnutzers bedarf […]“.[36] Fraglich ist daher, nach welchen Regelungen diese Einwilligung in Bezug auf die Kenntnisnahme Dritter (hier der externen Marktteilnehmer) zu beurteilen ist. Darüber hinaus stellt sich die Frage, ob eine datenschutzrechtliche Beurteilung dieses Szenarios sowohl das (Außen-)Verhältnis des vertragsgebundenen Letztverbrauchers (Haushaltsvorstand) gegenüber einem externen Marktteilnehmer sowie das innerfamiliäre Verhältnis zwischen dem Letztverbraucher als Haushaltsvorstand und den weiteren Haushaltsmitgliedern im Sinne einer angemessenen und umfassenden Realisierung der informationellen Selbstbestimmung berücksichtigt.

[33] [Bu13], Begründung S. 22.
[34] [Ra11], S. 23 f.
[35] [Ra11], S. 25.
[36] [BNA09], S. 4.

4 Innerfamiliäre Überwachung

Das Problem der sogenannten innerfamiliären Überwachung lässt sich beispielhaft an dem Vertragsverhältnis zwischen Letztverbraucher und MSB über die Durchführung des Messstellenbetriebs verdeutlichen (vgl. § 21b EnWG). Das Spannungsverhältnis ergibt sich aus der Möglichkeit einer extern erfolgenden Einsichtnahme in den Energieverbrauch des jeweiligen Haushalts über ein serverbasiertes Online-Portal des Anbieters. Folglich könnte der vertragsgebundene Letztverbraucher in einem Mehrpersonenhaushalt über die ihm (ausschließlich) zustehenden Zugangs- und Zugriffsdaten als Haushaltsvorstand jederzeit (weltweit) auf den entsprechenden Server mit den Energieverbrauchsdaten aus dem heimischen Messsystem und damit auf eventuell hoch aufgelöste Verbrauchsprofile der anwesenden Haushaltsmitglieder zugreifen. Demnach besteht die Gefahr einer unbemerkten Überwachung der Haushaltsmitglieder und damit einer möglichen Beeinträchtigung der informationellen Selbstbestimmung derselben.[37]

4.1 Zulässigkeit nach § 21g EnWG

Fraglich ist, ob diese Form der Verbrauchsvisualisierung einer rechtlichen Regelung unterliegt und insofern ein datenschutzrechtlicher Erlaubnistatbestand einschlägig ist. Da die Einsichtnahme über ein Online-Portal eine fernkommunikative Übermittlung der Daten aus dem Messsystem an den MSB als Anbieter erfordert, ist neben dem datenschutzrechtlichen Erlaubnistatbestand des § 21g Abs. 1 Nr. 6 Alt. 2 EnWG kumulativ eine Einwilligung in das sogenannte Fernmessen[38] nach § 21g Abs. 6 S. 5 EnWG notwendig.[39] Danach kann der Letztverbraucher in das Fernmessen einwilligen, sofern er „zuvor über den Verwendungszweck sowie über Art, Umfang und Zeitraum des Einsatzes unterrichtet worden ist“. Problematisch an dieser Konstellation ist allerdings, dass die Einwilligung nach § 21g Abs. 6 S. 5 EnWG nur auf den Letztverbraucher abstellt, dieser aber entsprechend der Begriffsbestimmung nach § 3 Nr. 25 EnWG Energie für den eigenen Verbrauch kauft. Daraus wird deutlich, dass im Rahmen der Einwilligung ins Fernmessen nicht die weiteren Haushaltsmitglieder adressiert sind, da diese mit dem zuständigen MSB keine Vertragsbeziehung innehaben.[40] Diese Argumentation wird zudem durch § 21b Abs. 2 EnWG gestützt, wonach nur der betroffene Anschlussnutzer vertraglichen Einfluss auf den Wechsel des MSB nehmen kann. Der Anschlussnutzer wird nach § 1 Abs. 3 der Niederspannungsanschlussverordnung (NAV) wiederum als Letztverbraucher definiert, „der im Rahmen eines Anschlussnutzungsverhältnisses einen Anschluss an das Niederspannungsnetz zur Entnahme von Elektrizität nutzt“, womit also ebenfalls der Haushaltsvorstand und nicht ein weiteres Haushaltsmitglied gemeint ist.[41] Hieraus lässt

[37] Im vorliegenden Szenario wird nicht die Möglichkeit betrachtet beispielsweise via VPN-Client per Fernzugriff auf das heimische Messsystem und den lokalen Speicher zuzugreifen.
[38] Ein Fernmessen liegt vor, wenn ferngesteuert oder -beobachtet Verbrauchsmessungen erfolgen, vgl. Verwaltungsvorschrift Nr. 29 zu § 29 DSG LSA.
[39] [LR14], § 21g Rn. 98; [We13], S. 311.
[40] [Ra11a], S. 837.
[41] [Ra11], S. 26.

sich ableiten, dass das interne Verhältnis zwischen Haushaltsmitgliedern und Haushaltsvorstand im Rahmen der Einwilligung nach § 21g Abs. 6 S. 5 EnWG keinen adäquaten Schutz der informationellen Selbstbestimmung der weiteren Haushaltsmitglieder beinhaltet. Es wird lediglich das Verhältnis zwischen dem Letztverbraucher und der zum Datenumgang berechtigen Stelle nach § 21g Abs. 2 EnWG betrachtet. Dadurch würde also der Umstand verkannt, dass es nicht nur um personenbezogene Daten des Letztverbrauchers als Anschlussnutzer, sondern darüber hinaus auch um schützenswerte personenbezogene Energieverbrauchsdaten der Haushaltsmitglieder gehen kann.

4.2 Privilegierung nach § 1 Abs. 2 Nr. 3 BDSG

Der Anwendungsbereich des Datenschutzrechts ist gemäß § 1 Abs. 2 Nr. 3 BDSG dann nicht eröffnet, wenn die Erhebung, Verarbeitung oder Nutzung personenbezogener Daten ausschließlich für persönliche oder familiäre Tätigkeiten erfolgt. Mit dieser Vorschrift soll der Bereich der persönlichen Lebensführung von der beruflichen und geschäftlichen Sphäre abgegrenzt und damit der im privaten Aktionskreis stattfindende Datenumgang privilegiert werden.[42] Die im privaten Umfeld anfallenden personenbezogenen Daten können dabei auf Endgeräten wie PC´s, Notebooks oder Smart Phones abgelegt werden, allerdings auch auf Servereinheiten.[43] Diese technischen Datenhaltungsmöglichkeiten bergen allerdings eine potenziell größere Gefährdungslage bezüglich der personenbezogenen Daten der weiteren Haushaltsmitglieder. Da insofern nicht exakt zwischen lokaler und externer Datenhaltung im privaten Bereich und den damit korrespondierenden Zugangs- und Zugriffsmöglichkeiten differenziert wird, könnte im vorliegenden Fall eine Privilegierung nach § 1 Abs. 2 Nr. 3 BDSG nach der Verkehrsauffassung[44] ausgeschlossen sein. Richtigerweise wird davon ausgegangen, dass die sich aus der Möglichkeit der innerfamiliären Überwachung ergebende Bedrohung für die informationelle Selbstbestimmung der weiteren Haushaltsmitglieder weitaus größer ist als in den von § 1 Abs. 2 Nr. 3 BDSG sonst üblichen Fällen.[45] So wird argumentiert, dass im Rahmen der Zweck-Mittel-Relation die vom Gesetzgeber angestrebte Effizienzsteigerung durch eine Verbrauchsvisualisierung per Fernzugriff auf die externen Daten des Messsystems[46] nicht notwendig erscheint, da auch über eine rein lokale Anzeige im Herrschaftsbereich der Wohnung der angestrebte Zweck ebenfalls erreicht werden könnte.[47] Zudem ergäbe sich bei einer externen Datenhaltung und -aufbereitung für die weiteren Haushaltsmitglieder ein gewisses Maß an Intransparenz bezüglich des Umgangs mit deren personenbezogenen Daten durch den Betreiber eines Webportals. Es drängt sich insbesondere die Frage auf, inwiefern ein Webportalbetreiber im Rahmen der Gestaltung und Aufbereitung der Visualisierungsmöglichkeiten Kenntnis über die Daten erlangt. Eine Verwendung personenbezogener Daten aus dem Messsystem zur Verbrauchsvisualisierung außerhalb der Sphäre aller Betroffenen wäre

[42] [Da11], § 1 Rn. 149.
[43] [Da11], § 1 Rn. 151.
[44] [Da11], § 1 Rn. 151.
[45] [Ra11a], S. 837.
[46] Vgl. [BR11], S. 195.
[47] [Ra11a], S. 837.

demgemäß nicht erforderlich und ist daher abzulehnen. Somit greift im vorliegenden Fall die Privilegierung nach § 1 Abs. 2 Nr. 3 BDSG nicht ein.

4.3 Vergleichbare Sachgestaltungen im Telekommunikationsrecht

Eine vergleichbare Fallkonstellation in Bezug auf die potentielle Überwachung im privaten Bereich liegt der Regelung des § 99 Abs. 1 Telekommunikationsgesetz (TKG) zugrunde. Über Einzelverbindungsnachweise könnte der Inhaber einer Telefonanlage als Anschlussnutzer das Kommunikationsverhalten der weiteren Haushaltsmitglieder überwachen.[48] Folglich sieht § 99 Abs. 1 S. 3 Hs. 1 TKG vor, dass bei Anschlüssen im Haushalt ein Einzelverbindungsnachweis nur dann zulässig ist, „wenn der Teilnehmer in Textform erklärt hat, dass er alle zum Haushalt gehörenden Mitbenutzer des Anschlusses darüber informiert hat und künftige Mitbenutzer unverzüglich darüber informieren wird, dass ihm die Verkehrsdaten zur Erteilung des Nachweises bekannt gegeben werden.“ Diese Regelung stellt damit eine Informationspflicht des Anschlussinhabers gegenüber den Mitbenutzern zu deren Schutz dar, die darüber hinaus dem Diensteanbieter in Textform (§ 126b BGB) zu erklären ist.[49] Eine vergleichbare Regelung als allgemeine verfahrensrechtliche Pflicht[50] fehlt jedoch im bereichsspezifischen Datenschutzrecht des EnWG und wäre im Falle der Erweiterung des Kreises der Einwilligungspflichtigen wohl auch nicht praktikabel.[51]

5 Rechtliche und technische Herausforderungen

Aus den vorangegangenen Untersuchungen ergibt sich daher, dass in Bezug auf die Sicherung der informationellen Selbstbestimmung der betroffenen Haushaltsmitglieder nach dem bisherigen Datenschutzkonzept des EnWG kein adäquates Schutzprogramm vorhanden ist. Daher wäre gemäß der Subsidiaritätsklausel nach § 1 Abs. 3 BDSG auf die allgemeinen Regeln des BDSG zurückzugreifen um die vorhandene Schutzlücke schließen zu können.[52] Folglich müsste auf eine Einwilligungslösung nach den Voraussetzungen des § 4a BDSG abgestellt werden. Hierbei ist allerdings zu beachten, dass die materiell-rechtlichen Datenschutzfragen nach § 21g Abs. 6 S. 1 EnWG in einer Rechtsverordnung nach § 21i Abs. 1 Nr. 4 EnWG geregelt werden sollen, in der insbesondere der Schutz der personenbezogenen Daten der an der Energieversorgung Beteiligten enthalten sein soll. Eine solche Rechtsverordnung bzw. ein dahingehender Referentenentwurf ist bislang nicht ersichtlich. Daher sollen nachfolgend rechtliche und technische Herausforderungen diskutiert werden, die in einer entsprechenden Regelung zu beachten wären.

[48] [Ra11], S. 27.
[49] [Kl13], § 99 Rn. 9f.
[50] [Ra11], S. 27.
[51] [Ra11a], S. 837.
[52] [Da11], § 1 Rn. 158.

5.1 Rechtsverordnung nach § 21g Abs. 6 S. 1 i.V.m. § 21i Abs. 1 Nr. 4 EnWG

Aufgrund der technologischen Veränderungen, die bereits im Zusammenhang mit der Einwilligung im Internet der Dienste diskutiert wurden, sind moderne Marktplattformen von hohen Transaktionsfrequenzen, verteilten Diensten und einer Vielzahl an beteiligten Akteuren gekennzeichnet.[53] In diesem Zusammenhang erscheint es fraglich, ob ein Rückgriff auf die Schriftform der Einwilligung nach § 4a S. 3 BDSG i.V.m. § 126 BGB (noch) geboten erscheint. Insbesondere bei elektronischen Datenübermittlungsprozessen besteht die Gefahr eines Medienbruchs, sodass insofern eine „andere Form" im Sinne des § 4a S. 3 BDSG angezeigt sein könnte.[54] Zwar besteht grundsätzlich die Möglichkeit die klassische Schriftform nach § 126a BGB zu ersetzen und eine qualifizierte elektronische Signatur nach dem Signaturgesetz (SigG) zu verwenden. Jedoch zeigt die sehr geringe Akzeptanz der qualifizierten elektronischen Signatur, dass das Festhalten hieran nicht praktikabel und sinnvoll ist.[55] Insofern erscheint es sachgerecht, entsprechend der Abwägung zwischen Verkehrsfähigkeit einerseits und der Umsetzung der datenschutzrechtlichen Warnfunktion andererseits[56], dem Vorbild des § 13 Abs. 2 TMG zu folgen. Somit könnte die elektronische Form als besonderer Umstand im Sinne des § 4a S. 3 BDSG bereichsspezifisch im EnWG bzw. der entsprechenden Rechtsverordnung geregelt werden.

5.2 Nachrichtenformate des Energiemarktes

Ferner ist zu berücksichtigen, dass datenschutzrechtlich relevante Datenflüsse sowohl zwischen dem Messsystem eines Letztverbrauchers (Datenquelle) und dem zuständigen MSB (Datensenke) verlaufen können, andererseits aber auch zwischen weiteren Akteuren des Energiemarktes (weitere Datensenken) ein Datenaustausch stattfindet. Die damit verbundenen Datenbedürfnisse ergeben sich aus den Vorgaben der BNetzA für die Kommunikationsprozesse des Energiemarktes.[57] Darin sind unter anderem die Datenformate für die konkreten Nachrichtentypen und den Datenaustausch verbindlich geregelt.[58] Demnach findet für den elektronischen Datenaustausch im Energiemarkt das Datenformat UN/EDIFACT Anwendung.[59] Fraglich ist jedoch wie prozessorientierter Datenschutz[60] im Energiesektor konkret realisiert werden kann, wenn unterschiedliche Nachrichtenformate bzw. -protokolle aufeinander treffen. So stellt sich insbesondere im Hinblick auf die Gewährleistung der datenschutzrechtlichen Integrität der zu übermittelnden Messdatensätze das Problem, dass unterschiedliche Nachrichtenformate in der Kommunikation zwischen Messsystem und MSB, sowie der Kommunikation zwischen MSB und weiteren externen Marktteilnehmern verwendet werden. Denn für die Kommunikation zwischen Messsystem und MSB ist nach den Vorgaben des BSI vorgesehen, dass die konkreten Datenstrukturen mit Hilfe von COSEM Interface-

[53] [RL11], S. 279.
[54] [RL11], S. 280; [Ra11a], S. 837.
[55] [RL11], S. 280 m.w.N.
[56] [LR14], § 21g Rn. 99; [RL11], S. 283.
[57] Siehe [BNA06] sowie [BNA10].
[58] [BNA06], S. 5 sowie [BNA10], S. 6.
[59] Siehe http://www.edi-energy.de/ [abgerufen am 18.06.2014].
[60] Vgl. [Ra11a], S. 834.

Klassen modelliert werden müssen.[61] Um die in den COSEM-Objekten repräsentierten Daten abbilden zu können, ist eine XML-Transfersyntax vorgesehen.[62] Für die Datenübermittlung zwischen externen Marktteilnehmern kommt jedoch nach UN/EDIFACT das Nachrichtenformat MSCONS zu Anwendung.[63] Sofern jedoch von Seiten des BSI eine XML-Transfersyntax für die Kommunikation in das Wide Area Network (WAN) des Energiemarktes zum Zwecke der weiteren Abwicklung der Geschäftsprozesse vorgesehen ist, liegt hier eine technisch bedingte Inkompatibilität zum Nachrichtenformat MSCONS vor: Eine seitens des BSI geforderte integritätssichernde Signatur[64] wird über eine XML-Transfersyntax beschrieben. Diese müsste jedoch zum Zwecke der weiteren Datenverarbeitung bzw. -übermittlung zunächst „aufgebrochen" werden und könnte sodann nicht mehr vollständig integritätswahrend auf das Nachrichtenformat MSCONS abgebildet und weiter kommuniziert werden. Somit kann eine verlustfreie Überführung der Signaturen nicht erfolgen und daher eine durchgehende Signaturintegrität nicht gewährleistet werden.[65]

6 Fazit und Ausblick

Mit der Umsetzung der Energiewende sind neben den rechtlichen Problemstellungen gleichfalls gesellschaftliche und technische Herausforderungen benannt. Klima- und umweltpolitische Ziele im Sinne einer nachhaltigen und sicheren Energieversorgung können mit Einschnitten in die informationelle Selbstbestimmung des Letztverbrauchers einhergehen. Es wird deutlich, dass die Anreicherung des Energiewirtschaftsrechts um Aspekte des Datenschutzes zu neuartigen Kollisionslagen führt und daher diese unterschiedlichen Regelungsregimes schonend miteinander in Einklang gebracht werden müssen. Den komplexen Kommunikationsprozessen des Energiemarktes wohnt die Gefahr inne, dass sich für den Letztverbraucher bezogen auf den Umgang mit seinen personenbezogenen Daten eine kritische Intransparenz ergeben kann. Die Fortentwicklung der Informations- und Kommunikationstechnologien im Energiesektor hängt damit entscheidend davon ab, wie das Vertrauen der Bürger in den Umgang mit den sie betreffenden personenbezogenen Daten erreicht werden kann. Insbesondere die datenschutzrechtliche Einwilligung als Instrument zur Absicherung der informationellen Selbstbestimmung des Betroffenen, steht im Kontext des Energiewirtschaftsrechts auf dem Prüfstand. Es ist ersichtlich, dass der Gesetzgeber bezüglich der Regelung von datenschutzrechtlichen Belangen im EnWG vor immensen technischen und rechtlichen Herausforderungen steht. Eine Auflösung dieser potentiellen Kollisionslagen sollte in der Rechtsverordnung nach § 21g Abs. 6 S. 1 i.V.m. § 21i Abs. 1 Nr. 4 EnWG in Bezug auf die materiell-rechtlichen Datenschutzregelungen ihren Niederschlag finden.

[61] [BSI13], S. 31 f.
[62] [BSI13], S. 32 i.V.m. [BSI13], Anlage II.
[63] Beschluss [BNA06], S. 5 sowie [BNA10], S. 6.
[64] Siehe [BSI13], S. 37 i.V.m. [BSI13], Anlage I.
[65] Vergleichend hierzu [PRW10], S. 409.

Literaturverzeichnis

[BNA06] Bundesnetzagentur: Festlegung einheitlicher Geschäftsprozesse und Datenformate zur Abwicklung der Belieferung von Kunden mit Elektrizität, Anlage zum Beschluss BK6-06-009 vom 11.07.2006 (konsolidierte Lesefassung gültig ab 01.04.2012).

[BNA09] Bundesnetzagentur: Konsultation eines Positionspapieres zu den Anforderungen an Messeinrichtungen im Sinne von § 21b Abs. 3a und 3b EnWG, BK6-09-170 vom 06.11.2009.

[BNA10] Bundesnetzagentur: Wechselprozesse im Messwesen, Anlage 1 zum Beschluss BK6-09-034/BK7-09-001 vom 09.09.2010 (konsolidierte Lesefassung gültig ab 01.04.2012).

[BR11] Bundesrat: Entwurf eines Gesetzes zur Neuregelung energiewirtschaftsrechtlicher Vorschriften, Drucksache 343/11 vom 06.06.2011.

[BSI13] Bundesamt für Sicherheit in der Informationstechnik: BSI TR-03109-1, Anforderungen an die Interoperabilität der Kommunikationseinheit eines intelligenten Messsystems. Version 1.0 vom 18.03.2013. Online unter: https://www.bsi.bund.de/DE/Themen/SmartMeter/smartmeter_node.html [18.06.2014].

[BT11] Deutscher Bundestag: Entwurf eines Gesetzes zur Neuregelung energiewirtschaftsrechtlicher Vorschriften, Drucksache 17/6072 vom 06.06.2011.

[Bu13] Referentenentwurf der Bundesregierung: Verordnung über technische Mindestanforderungen an den Einsatz intelligenter Messsysteme.

[Bu71] Bundesverfassungsgerichtentscheidungen (BVerfGE), 1971. In: Band 30, S. 292 – 336.

[Bu83] Bundesverfassungsgerichtentscheidungen (BVerfGE), 1983. In: Band 65, S. 1 – 71.

[Bu84] Bundesverfassungsgerichtentscheidungen (BVerfGE), 1984. In: Band 66, S. 248 – 259.

[BGH08] Bundesgerichtshof, 2008. In: Zeitschrift für Neues Energierecht 2008, S. 228 – 231.

[Da11] Dammann, U.: Kommentierung zu § 1 BDSG. In: Simitis (Hrsg.), BDSG, 7. Auflage 2011.

[JRV11] Jandt, S./Roßnagel, A./Volland, B.: Datenschutz für Smart Meter – Spezifische Neuregelungen im EnWG. In: Zeitschrift für Datenschutz 2011, S. 99 – 103.

[Kl13] Klesczewski, D.: Kommentierung zu § 99 TKG. In: Säcker (Hrsg.), Berliner Kommentar zum Telekommunikationsgesetz, 3. Auflage 2013.

[LR14] Lorenz, M./Raabe, O.: Kommentierung zu § 21g EnWG. In: Säcker (Hrsg.), Berliner Kommentar zum Energierecht, 3. Auflage 2014.

[PRW10] Pallas, F./Raabe, O./Weis, E.: Beweis- und eichrechtliche Aspekte der Elektromobilität. In: Computer und Recht 2010, S. 404 – 410.

[Ra11] Raabe, O./ Pallas, F./Weis, E./Lorenz, M./Boesche, K. V. (Hrsg.), Datenschutz in Smart Grids – Anmerkungen und Anregungen 2011.

[Ra11a] Raabe, O./Lorenz, M./Pallas, F./Weis, E.: Harmonisierung konträrer Kommunikationsmodelle im Datenschutzkonzept des EnWG - "Stern" trifft "Kette". In: Computer und Recht 2011, S. 831 – 840.

[RL11] Raabe, O./Lorenz, M.: Die datenschutzrechtliche Einwilligung im Internet der Dienste. In: Datenschutz und Datensicherheit 2011, S. 279 – 284.

[Th13] Theobald, C.: Grundlagen des deutschen Rechts der Energiewirtschaft. In: Schneider, J.-P./Theobald, C. (Hrsg.), Recht der Energiewirtschaft, 4. Auflage 2013, S. 1 – 48.

[We13] Weis, E./Pallas, F./Lorenz, M./Raabe, O.: Datenschutz und Elektromobilität. In: Boesche, K. V./Franz, O./Fest, C./Gaul, A. J. (Hrsg.), Berliner Handbuch zur Elektromobilität 2013, S. 297 – 323.

[We13a] Weichert, T.: Big Data und Datenschutz. In: Zeitschrift für Datenschutz 2013, S. 251 – 259.

[Wi14] Wiebe, A.: Datenschutz in Zeit von Web 2.0 und Big Data – dem Untergang geweiht oder auf dem Weg zum Immaterialgüterrecht? In: Zeitschrift für Informationsrecht 2014, S. 35 – 54.

Aligning User Consent Management and Service Process Modeling

Nils Gruschka
nils.gruschka@fh-kiel.de

Meiko Jensen
meje@mmmi.sdu.dk

Abstract: With evolving functionality of Internet services, the management of user consents becomes a complex challenge. As consents are a common means for establishing a legal basis for processing privacy-relevant user data, a sound consent management approach is required.

In this paper, we outline an approach for semi-automated generation of letter of consent documents, based on existing service implementation documentation. We illustrate the challenges of consent management in relation to service evolution, and we outline an integration of consent management into model-based process development systems.

1 Introduction

In times of continuous evolution of the Internet, with new features being added to online services on a daily basis, it becomes more and more challenging to maintain oversight on the different versions of services a provider has offered in its history. However, when it comes to court rulings, a particular user's consent to data processing can easily become a key parameter for determining law compliance of an online service, e.g. in terms of data protection regulations. With the emerging set of side-effects of online services (e.g. regarding surveillance, trading of personal data, or data theft), the future Internet is going to see a lot of lawsuits discussing what service a human individual has consented to, what scope that consent had, and how the service evolved after the consent was given. Hence, it becomes necessary to keep track of all forms of consent a service user has given explicitly, along with the scope of that consent. This challenge of *consent management*, which already emerged on a larger scale within the healthcare domain (cf. [HS13, Bon13, JIS11]), is about to become one of the most important concepts in the Future Internet.

In this paper, we introduce general requirements of *consent management systems* (cf. [Con10, epS, Tex12]), and elaborate their parameters and conditions. We continue to discuss the integration of consent management systems into state-of-the-art e-commerce suites, and we analyze different integration options for common process description languages like UML or BPMN. We describe the two main aspects of consent management (which are *consent documentation* and *consent information*), and we explore real-world applicability and future research directions within the domain of consent management systems.

2 Consent to Processing of Personal Data

One of the key necessities for providing Internet services within Europe is legal compliance. Among the multiple regulatory norms that define the European legal basis, the regulations for privacy and data protection [The95, The15] are among the more important ones. Basically, for most of these norms, it is strictly required by law to adhere to a certain set of conditions whenever a service provider processes any sort of data that directly relates to a human individual, i.e. is *personally-identifiable information*.

In this case, the particular service provider is required to identify the legal basis for such processing, which e.g. can be specific laws of the country the service provider operates within. However, one of the most common means to form such a valid legal basis for service operation is the explicit *consent* to data processing, given by the affected individual herself. Within certain boundaries, explicit consent of all affected individuals allows for almost any sort of processing of personally-identifiable information, including the option to forward personal data to other organizations, such as business partners.

A key aspect of such a consent, however, is that it must fulfill a certain set of conditions itself, in order to become a valid legal basis. Any form of legally valid consent must be

- given *freely*,
- given by an *informed* citizen,
- *documented* explicitly, and
- *revocable*.

Therein, *free* means that the user of a service had a free choice to give or not to give consent. She must not have given consent involuntarily, i.e. may not be punished for rejecting to give consent. It must be the individual's own choice, and her own decision to give consent to having her personal data being processed by the particular service. Any consent given under pressure becomes invalid, and thereby invalidates the legal basis for processing of affected personal data.

The term *informed* reflects that a citizen must know in detail about the conditions under which her personal data is going to be processed. Most importantly, this directly involves knowledge of the following:

- Who is the legal entity responsible for this particular service?
- What is the purpose of processing of a user's personal data?
- Which sort of personal data is to be used within the service?
- If personal data is stored, for how long is it maintained?
- To which other legal entities is the personal data forwarded?
- For what purposes is the personal data forwarded to other legal entities?

As can be seen, the definition of informed consent pre-assumes a lot of information to be provided by the service operator towards the users a priori. Typically, such information can be given by means of a *terms of use* document, which the individual user explicitly has to agree to, e.g. by clicking some "OK" button. Another option is the explicit definition of a *privacy policy* that contains all relevant information.

Documentation of given consent is a key requirement in order to prove the existence of consent at a later stage. If, for instance, a Data Protection Authority challenges the legitimacy of a certain sort of data processing, the scope and validity of a given consent is of critical importance. Hence, a service provider is urged to document the exact conditions under which the consent was given, and the scope it covers.

On *revocation* of consent, i.e. if a user changes her mind and does not allow any further processing of her personal data at the service provider, the service provider is obliged to respect and follow that decision within reasonable means. For instance, this implies to stop any ongoing process instances that handle personal data of the particular individual, and to prevent subsequent processes from starting when they are processing such personal data. Also, deletion of personal data (within the boundaries defined by other legal norms) must be performed on user request.

3 Consent Management

As can be seen, the concept of consent to personal data processing is quite complex, and bears a lot of challenges, especially for digital services provided over the Internet. In face of growing number of users, continuous changes to the implementation of a service, and ongoing alterations in the set of collaborating business partners, keeping an overview of all events related to user consent becomes a management issue. Within this challenge, the following characteristics become important for *consent management*:

- Each given consent must be documented in a durable, standardized way that allows for subsequent recall.
- For each given consent, the exact conditions of the service implementation for exactly the moment the consent was given (i.e. the *scope* of the consent) have to be documented as well.
- Whenever a user revokes consent, this information has to be documented as well, and must be linkable to the documentation of the consent that actually was revoked.
- Changes in the service implementation might lead to changes in the scope of consent. Hence, each change in the implementation must trigger a re-validation of the information given to users prior to consent.
- Changes in the set of business partners involved in processing of personal data also must be reflected in the user consent information.

As can be seen, the management of consent has two sides. On the one hand, all user decisions regarding consent have to be documented explicitly. On the other hand, for each of these decision events, the scope of consent, i.e. the information provided to the individual user, and the state of the service implementation at the time of the decision event, have to be documented as well.

Derived from these observations, it becomes evident that consent management consists of two parts: *User Consent Information* deals with the informational text provided to (existing or upcoming) users of a service, in order to gather an informed consent, as obliged by regulatory norms. Hence, this part deals with gathering information on the state of a service's technical and organizational implementation, and transforming that information into a representation that fulfills the property of *informedness* as required for consent by law. *User Consent Documentation* covers all tasks that result from a user's decisions regarding giving or revoking consent. This includes documentation of user's decisions, documentation of the state of the service at the time of decision, and implementation of all technical and organizational tasks that result from each type of user consent decision.

4 Consent Management vs. Service Management

Taking the characteristics of consent management as a basis, it becomes immediately clear that the challenge of consent management is directly linked to the challenge of service management. Every event in service management, such as the roll-out of a service update, must be considered in consent management as well, as it might affect the scope of future consents. Vice versa, every user-generated event in consent management, i.e. giving or revoking consents, may affect the service implementation, as it must e.g. trigger deletion of personal data, or shutdown of affected process instances.

For the technical aspects of service management, the state of the art today already covers a broad spectrum of techniques, such as explicit process modeling (e.g. by means of UML [RJB04], BPMN [OMG13], or others) and well-defined implementation change management processes (cf. e.g. [PM08]). However, those techniques are typically more focused on technical aspects of a service, such as implementation details, security properties, organizational aspects, or external event handling. To the best of our knowledge, as of today, there exists no relevant technological approach to align service and process management with consent management. This gap is to be addressed within the remainder of the paper.

The upper part of Figure 1 shows the typical workflow of software development procedure. A software developer:

a) performs a requirement analysis (not shown in the picture),

b) creates the software or process design (using modelling techniques),

c) uses the model as a basis for implementing the software in a programming language,

d) compiles the software and deploys the resulting service.

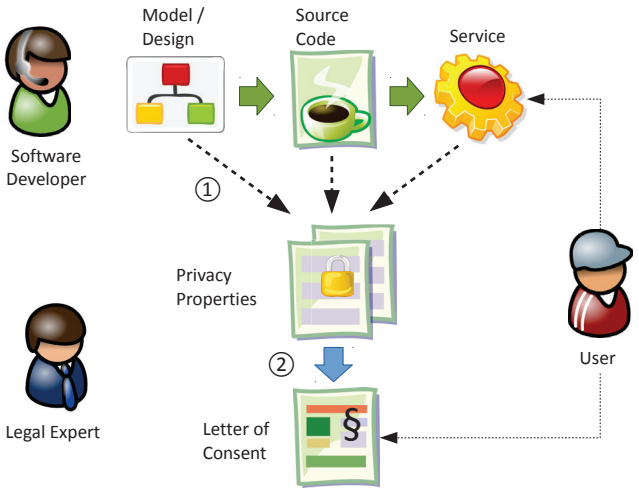


Figure 1: Architecture for consent management

In all these steps, information on processing of data is handled, including processing of private user information. Thus, we propose to extract all privacy-relevant information from the existing software design process (step 1). These privacy properties can be used as a basis for a legal expert to create a letter of consent (step 2). This letter is presented to the service user, who can read and agree to the privacy conditions of the service she is intending to use.

This combination of service management and consent management has a number of advantages (compared to current approaches), e.g. precise process information inside the letter of consent, better workflow for the legal experts, and fine-grained lifecycle management. Section 6 presents how privacy properties can be derived from the service design process. Before, we introduce an example service process, which will be used for illustrating our approach.

5 Example Process: Travel Agency

In the remainder of the paper we will use the example of a travel agency service. The service process is shown in Figure 2.

The user calls the travel agency service giving her name, address, date of birth and credit card number (plus the details on the travel booking like destination or car type which we will skip in the following). These values are stored locally in the travel agency's database. Further, the travel agency uses other services to fulfil the user's booking. After that (depending on the type of booking), either a flight is booked using an airline service or a car is booked using a car rental service or both in parallel. Finally, a billing service is called with billing address and credit card number as input parameters.

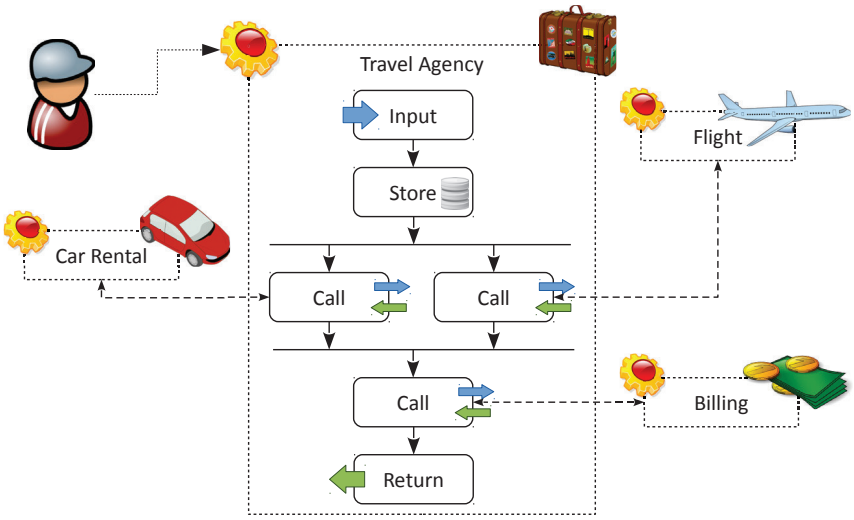


Figure 2: Process of a travel agency service

6 Generating Privacy Properties from Service Processes

6.1 Privacy Annotations

A letter of consent must include information on storage of personal data and its transfer to external institutions. In order to gather these information, the following steps inside a service process must be taken into account:

a) Which of the input parameters (i.e. data from the user) are privacy-relevant?

b) What processing steps are performed on the parameters? For what purpose?

c) Which data is stored by the service?

d) How long is the data stored?

e) Which data is used as input parameters when the service calls other services?

Item a) is obviously relevant. The letter of consent should only contain information on processing of personal information. However, it is not possible to perform this step automatically. For manual decision there are two possibilities. In the first approach, the generated privacy properties (from step 1 in Figure 1) contain processing information about all kind of data (personal and non-personal) and must be manually filtered by some legal expert when creating the letter of consent (step 2 in Figure 1).

The second possibility is to add *privacy metadata* to the service model, the source code, or other service descriptions, which holds additional information, e.g. about the criticality of

parameters. Using this metadata, the generator can create more precise privacy properties. The metadata can be created by the software developer or a privacy officer together with the legal expert. This possibility has some advantages in comparison to the first one. First, the privacy relevance can be best decided when looking directly at the model or code. Second, once the persistent metadata has been added, the generation of the (precise) privacy properties can be repeated, e.g. for a new service version. Finally, this approach creates privacy properties (in contrast to the general processing properties of the first approach), which can be used for exchange of privacy properties, or for dedicated service lifecycle management. In the following, we will use this approach of metadata annotation. For the date of birth input parameter from the travel agency example before the metadata might look like this:

```
input_data {
   variable = birthdate
   type = date
   privacy_relevance true
}
```

Item b) is relevant to watch the flow of personal data during the service processing, and to decide if personal data is stored or transmitted later on. During service processing, data might be copied from one variable to another, or data might be converted (e.g. date-of-birth to age, address to geographic coordinates). For the first case, methods from the discipline of *information flow analysis* [BC85] operating on the source code can be used. However, these method are costly and too fine-grained for this application. For example, in the case of data conversion, information flow analysis would always mark the resulting data as personal. However, such converted data might not be privacy-relevant anymore, e.g. if anonymization is applied in the conversion. Thus, also for item b) we decide for adding metadata manually to variables, in order to mark if they include personal data or not. This marking is only required for data which is stored (item c) or transmitted (item e).

As mentioned before, all data that is stored by the service (in a local database, not via a remote service) is enriched with metadata in the process model. This metadata includes if the data is personal, and which original parameters have been used to calculate the information. Taking the example of a date of birth as user parameter and the resulting age as data to be stored, this might look like this:

```
stored_data {
   variable = age
   type = integer
   derived_from = birthdate
   purpose = age_verification
   privacy_relevance = true
}
```

When using such annotations, no explicit information flow analysis of the process is required (cmp. item b), just the knowledge of the developer regarding the service semantics. Further, as the user is informed about all possible data processing steps there is no need to distinguish if for a certain runtime condition the car rental or the flight booking service

is called (or both in parallel). The overall privacy properties include the union of privacy properties from all process parts.

Item d) is very important, as the duration of data storage is a crucial factor of the informed consent. It is nearly impossible to derive this kind of information automatically from the service design or implementation. Thus, the only possibility is again to annotate the service. An extension to the previous example might look like this:

```
stored_data {
   ...
   storage_time = 3
   storage_time_unit = month
}
```

Item e) is the most relevant step inside a process, because the user needs to be informed if user data is not just processed inside the organisation, but also transmitted to an external partner. This information must include the relevant parameters, the original data, the parameters have been derived from, and information on the called service itself. If the called service is an internal service precise data processing information (like for the calling service) might exist and can be embedded or referenced. For example, if the privacy properties of an internal service shows, that it does not include any further service calls, the user knows, that his data does not leave the organisation. For external services this is usually not available. Thus, the service must be annotated with a least a semantic description of the service process. The letter of content at least contains the information, that the data leaves the organisation and the user can decide if she trusts the external organisation or not[1].

For the example of the travel agency service calling the (external) flight booking service the metadata might look like this (here: the country of residence is transmitted):

```
service {
   type = external
   description = Booking flights for various airlines
   privacy_info = N/A
   outgoing_data {
      variable = nationality
      type = string
      derived_from = address
      privacy_relevance = true
   }
}
```

[1] It must be noted that *internal* and *external* not necessarily reflect organisation borders. For example, internal might also be “same country” or “inside EU”.

6.2 Model-Based Annotation of Privacy Properties

Most of today's services originate from software implementations that have been re-iterated and improved over the years. In that case, documentation of the service's inner workings may exist, but much of the service descriptions are purely based on existing source code documentation. For more modern, model-driven process development suites (cf. [Sof, Ecl]), however, the core elements of a service implementation are specified in a graphical modeling suite, and transformed into source code stubs—or even executable process implementations—subsequently. Such process modeling suites exist for a broad set of process modeling languages.

For the purpose of generating a letter of consent, each of these process modeling suites may be used as a basis for a dedicated *consent management extension*. Such an extension would then integrate into the modeling suite, e.g. as a dedicated consent management view (cf. [Fej09, SWF+11]). On request, this extension can then ask for potential further input, and automatically generate a letter of consent document for subsequent utilization. The example in Figure 3 shows a sketch of a BPMN process model, as it could be created within any common BPMN modeling suite. By means of a dedicated extension, such a BPMN suite can then ask the model developer for input on the letter of consent. As is shown, this extension identified that a dedicated *billing address* data item is to be sent from the Travel Agency to the Billing Provider. Hence, it identifies this item as a potentially relevant privacy property of the overall process, and asks the model developer (in this case e.g. a lawyer) whether information on that data transfer is to be contained within the letter of consent or not. Similarly, by means of a sequence of dialogues, the extension can iterate over all data transfers, and also data storage points (as is shown in the Storage Provider's lane of Figure 3, represented by a database), and ask for inclusion into the letter of consent for each of these.

For subsequent re-use of the process model, it is recommended to store the lawyer's decisions along with the process model, so that each change in the process model can be aligned with a subsequent check of the set of stored privacy properties, to check whether new data transfers, storage locations, or other privacy-relevant parameters have been added. Then, the lawyer has to answer such dialogues only for the set of new and changed privacy properties, whereas all other information can be taken from the stored values. This way, the overall time to develop updates of a process (and of its letter of consent) can be minimized.

Obviously, the approach shown here for BPMN also works almost identically for other sources of process information, such as UML diagrams, WS-BPEL process descriptions, event-driven process chain models, even for semantic source code documentations, or any other sort of semantic service description.

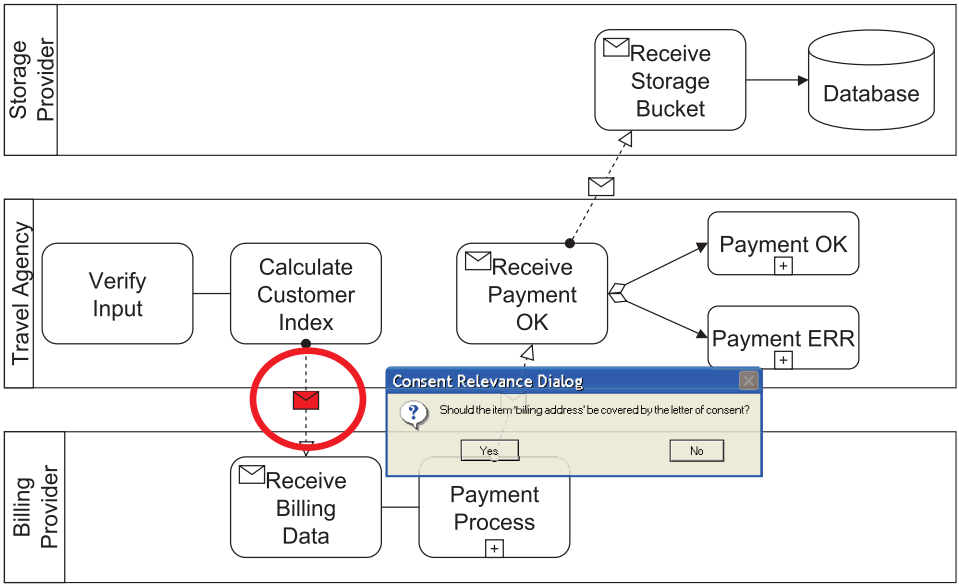


Figure 3: BPMN Process Model Example with Consent Relevance Dialog

6.3 Generating the Letter of Consent

Taking the data collected according to the techniques described above as a basis, it is now possible to automate (at least part of) the generation of a letter of consent document. Therefore, the deployment process iterates over the set of privacy properties, and identifies the information relevant for display towards the individual user. In the automated scenario, this task can completely be performed without human interaction, i.e. without consulting neither developers nor lawyers regarding validity of the generated letter of consent. However, it is not expected that such automation is mature enough yet to hold a thorough legal analysis. Therefore, we recommend the semi-automated setting (as described in Section 4), wherein a legal expert alters the automatically generated letter of consent stub according to specific legal requirements. Then, the user is presented with the manually altered version of the letter of consent. Nevertheless, both the automatically generated stub and the final version of the letter of consent are stored in the consent management system.

An example for how such a semi-automated letter of consent might be presented to the individual users is shown in Figure 4. As can be seen, each line in the letter of consent reflects one privacy property of the underlying process implementation. For example, the data forwarded to the external billing provider (as a separate organizational entity) is explicitly listed in the letter of consent, along with a detailed reference to the company's name (*ACME-Pay Corp.*), the exact data being forwarded (*name*, *billing address*, and *credit card information*), and the duration of storage at that external organization (*6 months*). Similarly, but more generic, the data forwarded to the external storage provider is covered by the 3rd line of the letter of consent. Here, the identity of the external or-

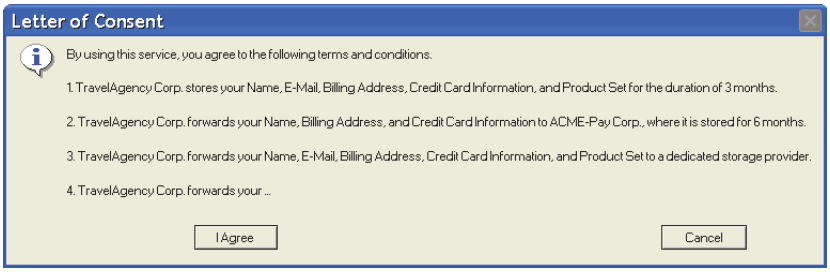


Figure 4: Letter of Consent Dialog Example

ganization was hidden (*a dedicated storage provider*), and there is no information given regarding storage times. This may have resulted from a management's decision at the TravelAgency Corp. not to disclose this information publicly, which was then annotated within the consent management extension of the BPMN modeling suite. Thus, for all future versions of the letter of consent, this information will always remain hidden, unless this option becomes explicitly changed within the process modeling suite.

7 Conclusions and Future Work

In this paper, we outlined potential approaches for aligning service management with consent management. Therein, we identified the two parts of consent management, which are *User Consent Information* and *User Consent Documentation*. We discussed the link between consent management and service development, and we outlined a viable way to integrate them into state-of-the-art service modelling tools.

As many of the observations discussed in this paper require a more in-depth analysis of validity, future work obviously consists in verifying these observations for real-world use cases, e.g. in the healthcare domain. However, we are confident to find most of our observations to be accurate.

Another obvious future work consists in implementation of a consent modelling tool as outlined in Section 6.2. The plan is to extend a typical, existing, common-to-market process modelling suite with a dedicated consent management extension along the outlines given in this paper. Once such a dedicated consent management extension is prototyped, it can be tried in real-world use cases to verify our finding's validity, and to lay the ground for subsequent work in terms of optimizations and formalizations of consent management in complex services of the Future Internet.

References

[BC85] Jean-Francois Bergeretti und Bernard A. Carré. Information-flow and Data-flow Analysis of While-programs. *ACM Trans. Program. Lang. Syst.*, 7(1):37–61, Januar 1985.

[Bon13] C.J. Bonnici. An extended conceptual model of consent for information systems. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, Seiten 149–154, June 2013.

[Con10] Joseph Conn. Govt., vendors show off consent-management tools. *Modern Healthcare, http://www.modernhealthcare.com/article/20100630/NEWS/100629932*, 2010.

[Ecl] Eclipse Foundation. Eclipse Modeling Framework. `http://www.eclipse.org/modeling/emf/`.

[epS] epSOS project. Technical solutions of patient consent. `http://www.epsos.eu/technical-background/systems-standards/technical-solutions-of-patient-consent.html`.

[Fej09] Sven Feja. An Approach for Semantic Checks of Process Models. In *BPSC*, Seiten 229–230, 2009.

[HS13] Susan D. Hosek und Susan G. Straus. *Patient Privacy, Consent, and Identity Management in Health Information Exchange*. RAND Corporation, 2013.

[JIS11] JISC Legal. Consent Management - Handling Personalisation Data Lawfully. *JISC Legal report*, 2011.

[OMG13] OMG Specification. Business Process Model and Notation (BPMN). Bericht, The Object Management Group, 2013.

[PM08] Robert A. Paton und James McCalman. *Change Management: A Guide to Effective Implementation*. SAGE Publications Ltd., 2008.

[RJB04] James Rumbaugh, Ivar Jacobson und Grady Booch. *Unified Modeling Language Reference Manual, The (2Nd Edition)*. Pearson Higher Education, 2004.

[Sof] Software AG. ARIS Business Process Platform. `http://www.softwareag.com/de/products/aris/default.asp`.

[SWF+11] Andreas Speck, Sren Witt, Sven Feja, Aneta Lotyzc und Elke Pulvermüller. Framework for Business Process Verification. In Witold Abramowicz, Hrsg., *Business Information Systems*, Jgg. 87 of *Lecture Notes in Business Information Processing*, Seiten 50–61. Springer Berlin Heidelberg, 2011.

[Tex12] Texas Health Services Authority. Texas Consent Management Services Whitepaper. `http://hietexas.org/component/docman/doc\_download/387-thsa-consent-management-services-white-paper`, 2012.

[The95] The European Parliament and the Council of the European Union. Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data, 1995.

[The15] The European Parliament. Personal data protection: processing and free movement of data (General Data Protection Regulation, upcoming), Expected for 2015.

Smartphones und Apps: Datenschutzrechtliche Risiken und deren Begrenzung

Anke Brummund

Goethe-Universität Frankfurt am Main
Lehrstuhl für Öffentliches Recht, Informationsrecht, Umweltrecht und Verwaltungswissenschaften
Grüneburgplatz 1, RuW 05
60323 Frankfurt a.M.
brummund@jur.uni-frankfurt.de

Abstract: Smartphones finden heutzutage zunehmend Verbreitung, insbesondere die Zahl der heruntergeladenen zusätzlichen mobilen Anwendungen - sogenannte Apps – steigt stetig. Durch diese Entwicklung kommt es vermehrt zu Risiken für personenbezogene Daten der Smartphone-Nutzer, müssen diese im Rahmen der Installation von Apps oft in umfassende Zugriffsrechte einwilligen oder gänzlich auf die Anwendung verzichten. Dieses Alles-oder-nichts-Prinzip wird einem wirksamen Datenschutz nicht gerecht. Immer häufiger werden zudem in Tests Sicherheitsmängel bei Apps festgestellt.

Im vorliegenden Beitrag werden deshalb zunächst die datenschutzrechtlichen Risiken für Smartphone-Nutzer dargestellt und sodann einige Möglichkeiten des verstärkten Selbstdatenschutzes, inklusive einer Bewertung ihrer rechtlichen Zulässigkeit, aufgezeigt. Vorgestellt werden hierzu drei Apps: der App Permission Watcher, SRT AppGuard, sowie XPrivacy. Der Fokus liegt vornehmlich auf den beiden Letzteren, die durch Modifikationen an datenschutzrechtlich bedenklichen Apps oder dem installierten Betriebssystem ein dynamisches Rechtemanagement herstellen. Wie derartige Modifikationen urheberrechtlich zu beurteilen sind, soll im Folgenden beleuchtet werden.

1 Einführung

Smartphones haben mittlerweile sowohl für die private als auch die geschäftliche Nutzung eine hohe Bedeutung erlangt. Gerade die Möglichkeit, dem Gerät durch zusätzliche mobile Anwendungen - sogenannte Apps - individuelle, auf die jeweiligen Bedürfnisse abgestimmte Nutzungsmöglichkeiten hinzuzufügen, machen deren Gebrauch nützlich und reizvoll.[1] Diese Anwendungen bergen jedoch auch ein hohes Risiko für personenbezogene Daten, erfordern die meisten doch eine Internetverbindung und die Einwilligung in umfassende Berechtigungen.

[1] [SM13], S. 303.

Mit einer Pressemitteilung[2] vom 05.03.2014 gab das Fraunhofer-Institut für Angewandte und Integrierte Sicherheit bekannt, dass es 10.000 Android-Apps auf Sicherheitsmängel getestet habe. 91% der untersuchten Apps sollen stets eine Internetverbindung angefordert haben. Eine Installation der betroffenen Anwendungen sei nur mit Zustimmung zur Internetnutzung möglich. Eine weitaus alarmierendere Feststellung war jedoch, dass 69% der untersuchten Apps die Daten unverschlüsselt sendeten.

Durch derartige Sicherheitsmängel und die Erteilung von Berechtigungen ist eine Vielzahl von personenbezogenen Daten betroffen wie z.B. Name, Kontodaten, Adressbuchdaten, Standortdaten, Fotos und deren Metadaten, Angaben in sozialen Netzwerken usw.[3]

Einige Beispiele für mögliche Berechtigungen der Apps sind:[4]

- der Zugriff auf die Kamera oder das Mikrofon;
  denkbar dadurch sind bspw. Ton-Mitschnitte des Nutzers
- der Zugriff auf Kontaktdaten im Adressbuch
- der Zugriff auf den Standort
- der Zugriff auf E-Mail-Dienste und SMS;
  denkbar dadurch sind bspw. das Lesen und Schreiben von E-Mails/SMS

Die Menge an gesammelten Daten ist erheblich und die Tendenz weiterer Sammlung stetig steigend. Wurden 2010 ca. 386 Millionen Apps innerhalb eines Jahres in Deutschland heruntergeladen, waren es 2011 bereits rund 962 Millionen.[5] Im Jahr 2012 wurden sogar 1,7 Milliarden heruntergeladene Apps gezählt.[6] Somit zeigt sich, dass Smartphone-Apps geeignet sind, eine Vielzahl von Daten zu sammeln. Aus datenschutzrechtlicher Sicht sind sie deshalb höchstbedenklich.

Zu beachten ist außerdem, dass die durch Apps gewonnenen Daten auch teilweise bei großen Unternehmen zusammengeführt werden können. So sorgte z.B. der Kauf des Instant Messengers WhatsApp durch das soziale Online-Netzwerk Facebook für Aufse-

[2] Fraunhofer-Institut für Angewandte und Integrierte Sicherheit, Pressemitteilung vom 05.03.2014: 10000 Apps und eine Menge Sorgen, abrufbar unter: http://www.aisec.fraunhofer.de/de/medien-und-presse/pressemitteilungen/2014/20140403_10000_apps.html, [abgerufen am 26.06.2014].
[3] [SM13], S. 304.
[4] vgl. Düsseldorfer Kreis, Orientierungshilfe zu den Datenschutzanforderungen an App-Entwickler und App-Anbieter, S. 19, abrufbar unter:
http://www.lda.bayern.de/lda/datenschutzaufsicht/lda_daten/Orientierungshilfe_Apps_2014.pdf, [abgerufen am 26.06.2014].
[5] BITKOM, Presseinformation vom 23.02.2012: Fast eine Milliarde App-Downloads allein in Deutschland, abrufbar unter:
http://www.bitkom.org/files/documents/BITKOM-Presseinfo_Markt_Apps_in_Deutschland_23_02_2012.pdf, [abgerufen am 26.06.2014].
[6] BITKOM, Presseinformation vom 08.05.2013: Umsatz mit Apps hat sich 2012 mehr als verdoppelt, abrufbar unter:
http://www.bitkom.org/files/documents/BITKOM_Presseinfo_Apps_werden_zum_Wirtschaftsfaktor_09_05_2013.pdf, [abgerufen am 26.06.2014].

hen.[7] Die Dimension wird deutlich, wenn man beachtet, dass es sich bei WhatsApp um eine der meistgenutzten Apps[8], und bei Facebook derzeit um das weltweit größte soziale Netzwerk handelt.

Für zusätzliche Brisanz sorgte auch die Meldung, dass Geheimdienste Daten mittels Smartphone-Apps wie z.B. dem Spiel „Angry Birds", sammeln.[9]

Aufgrund dieser datenschutzrechtlichen Risiken wird Smartphone-Nutzern häufig geraten, Apps nur aus vertrauenswürdigen Quellen zu installieren sowie deren Zugriffsrechte genau zu überprüfen. Greift die Anwendung auf Daten zu, die für die eigentliche Funktion nicht erforderlich sind, so soll man demnach auf die Installation verzichten. Dies geht z.B. aus einem Ratgeber des BMWi[10] aus dem Jahre 2011 hervor.

Unter dem Aspekt, dass die Mehrheit der Apps Berechtigungen einfordert und Daten erhebt, die für die eigentliche Anwendung nicht erforderlich sind, ist dieses Alles-oder-nichts-Prinzip eine nur unbefriedigende Lösung. Zudem werden viele vertrauliche Daten unbemerkt übermittelt, ohne dass der Smartphone-Nutzer davon Kenntnis erlangt.[11] Ein Verzicht auf die Installation ist dem Nutzer in solchen Fällen mangels Informiertheit erst gar nicht möglich.

Es besteht deshalb ein starker Bedarf an alternativen Lösungen. Die ledigliche Möglichkeit einer pauschalen Einwilligung in die Zugriffsrechte wird kritisiert und der Wunsch nach einer Stärkung des Selbstdatenschutzes durch individuelle Steuerungsmöglichkeiten laut.[12]

Zwischenzeitlich wurden diverse Anwendungen entwickelt, um problematische Apps aufzuspüren und gegebenenfalls anzupassen. Diese lassen sich im Wesentlichen in drei Kategorien einteilen: Anwendungen, die das Überwachen von Apps ermöglichen (2.1), solche, die kritische Apps selbst verändern (2.2), und Apps, die der Modifikation des Betriebssystems dienen (2.3), um sodann auf die kritische Anwendung einzuwirken. Diese Kategorien sollen im Folgenden an Hand jeweils einer konkreten Anwendung vorgestellt und deren urheberrechtliche Zulässigkeit mit Blick auf die erfolgten Modifikationen überprüft werden.

---

[7] Heise Online vom 19.02.1014, Facebook kauft WhatsApp, abrufbar unter: http://heise.de/-2118920, [abgerufen am 26.06.2014].
[8] vgl. [Kr12], S. 438.
[9] Süddeutsche vom 27.01.2014, Apps im Fokus von NSA und GCHQ – Angry Birds in Überwachungsmission, abrufbar unter: http://sz.de/1.1873548, [abgerufen am 26.06.2014].
[10] BMWi, Mobile Sicherheit – Ortung – Datenschutz, abrufbar unter: https://www.bsi-fuer-buerger.de/SharedDocs/Downloads/DE/BSIFB/Publikationen/extern/Mobile-Sicherheit-Ortung-Datenschutz.pdf?__blob=publicationFile, [abgerufen am 26.06.2014].
[11] [Ha11].
[12] Düsseldorfer Kreis, Beschluss vom 04/05. Mai 2011, Datenschutzgerechte Smartphone-Nutzung ermöglichen, S. 1 f., abrufbar unter: https://www.ldi.nrw.de/mainmenu_Service/submenu_Entschliessungsarchiv/Inhalt/Beschluesse_Duesseldorfer_Kreis/Inhalt/2011/Datenschutzgerechte_Smartphone-Nutzung_ermoeglichen_/Datenschutzgerechte_Smartphone-Nutzung_erm__glichen_.pdf, [abgerufen am 26.06.2014].

2 Schutzmöglichkeiten

2.1 Der App Permission Watcher

Bei dem App Permission Watcher (deutsch: App-Berechtigungswächter) handelt es sich um eine Anwendung, die es dem Smartphone-Nutzer ermöglicht, Zugriffsrechte von installierten Apps einzusehen und zu überprüfen.

Unter anderem ermöglicht die Anwendung folgende Funktionen:[13]

- die Anzeige verdächtiger Apps
- die Anzeige aller systemfremden Apps
- die Anzeige aller Berechtigungen und von welchen Apps diese genutzt werden
- die Möglichkeit bestimmte Apps als vertrauenswürdig zu markieren: diese werden danach nicht mehr als verdächtig gemeldet

Eine solche Anwendung ist zwar durchaus nützlich und wünschenswert, um Apps auf dem Smartphone regelmäßig zu überwachen, jedoch wird auch dadurch nicht das unbefriedigende Alles-oder-nichts-Prinzip beseitigt, da die Anwendung zwar die Analyse, nicht aber die Beschränkung von Zugriffsrechten einzelner Apps ermöglicht. Ein dynamisches Rechtemanagement wird so nicht hergestellt. Urheberrechtlich relevante Modifikationen finden nicht statt, gleichzeitig begrenzt dies jedoch auch die Effizienz des Selbstdatenschutzes mithilfe des App Permission Watchers.

2.2 Die Modifikation von Apps am Beispiel des SRT AppGuards

Der SRT AppGuard ist ein Beispiel für eine Anwendung, deren Funktion darin liegt, das Rechtemanagement von Apps zu überwachen und im Falle von Sicherheitsmängeln die jeweilige Anwendung zu deinstallieren, modifizieren und in neuer Version zu installieren. Die Kernaufgabe liegt also im Verändern der App selbst, das auf dem Smartphone installierte Betriebssystem bleibt unberührt.

Vorteil dieser Anwendung ist, dass es dem Benutzer so auf gewissermaßen einfache Art und Weise möglich ist, einzelne Berechtigungen durch Entfernen der jeweils daneben stehenden Häkchen zu entziehen und ein dynamisches Rechtemanagement herzustellen. Die unerwünschten Zugriffsrechte werden sodann blockiert. Dies birgt jedoch auch den Nachteil, dass es zu Funktionsproblemen und Abstürzen einer App kommen kann, sollte die Anwendung ohne die entzogenen Zugriffsrechte nicht mehr funktionsfähig sein.

Einige Wochen lang war der SRT AppGuard im Google Play Store zum Herunterladen bereitgestellt, wurde nach kurzer Zeit aber wieder entfernt und kann momentan nur über die Internetseite des Anbieters heruntergeladen werden.[14] Mögliche Ursache hierfür ist,

[13] App Permission Watcher, Informationen, abrufbar unter: http://www.apewatch.de/index_de.html, [abgerufen am 26.06.2014].
[14] Heise Online vom 05.07.2012, AppGuard: Kontrolle ungewollter App-Berechtigungen unter Android, abrufbar unter: http://heise.de/-1632532, [abgerufen am 26.06.2014].

dass Google die Anwendung als Verstoß gegen seine Google Play Nutzungsbedingungen[15] sah. Unter „6. Rechte und Einschränkungen“ werden darin Eigentumshinweise gegeben, nach denen es App-Käufern nicht gestattet ist, zu „versuchen, über Google Play erhaltene Inhalte zu verändern, einschließlich aller Veränderungen zum Zweck des Verbergens oder Änderns von Angaben zur Eigentümerschaft oder Herkunft von Inhalten“.

Eine wesentlich bedeutendere Frage ist jedoch, ob das Modifizieren der App eine Urheberrechtsverletzung darstellt.

2.2.1 Technische Grundlagen der Modifikation von Apps

Technisch erfolgt die Modifikation der Anwendung, indem dem Bytecode ein weiterer Programmcode – die sogenannte Sicherheitsbibliothek - hinzugefügt und die App anschließend neu installiert wird.[16]

Anwendungen für Android werden in der Regel in der Programmiersprache Java geschrieben.[17] Dieser zunächst erstellte Quellcode ist für Menschen lesbar, nicht jedoch von Computern bzw. Prozessoren ausführbar, da diese nur sogenannte Maschinencodes - Abfolgen aus Nullen und Einsen - ausführen können. Es bedarf deshalb der Übersetzung des Quellcodes, z.B. mittels javac. Diesen Vorgang bezeichnet man als Kompilieren.[18] Im Falle von Java besteht die Besonderheit darin, dass durch das Kompilieren nicht direkt der Maschinencode, sondern ein Zwischencode - der sogenannte Bytecode - erstellt wird. Dieser ist weder von Menschen lesbar, noch von Computern ausführbar.

In einem weiteren Schritt kann der Bytecode dann mittels eines Interpreters, z.B. Java Run Time, für den Prozessor in einen Maschinencode übersetzt und ausführbar gemacht werden.[19] Der Vorteil davon liegt in der Plattformunabhängigkeit, da der Bytecode auf allen Betriebssystem und jeder Hardware mit installierter Java Virtual Machine ausgeführt werden kann.

2.2.2 Urheberrechtliche Bewertung der Modifikation von Apps

Wird dem Zwischencode ein weiterer Programmcode hinzugefügt, um die Berechtigungen zu verändern, so könnte dem Urheber der App-Software eventuell ein Unterlassungs- und/oder Schadensersatzanspruch nach § 97 Abs. 1 UrhG zustehen. Zudem wäre ein Vernichtungsanspruch nach § 98 Abs. 1 UrhG denkbar.

---

[15] Google Play Nutzungsbedingungen, Stand vom 08.04.2014, abrufbar unter: https://play.google.com/intl/de_de/about/play-terms.html, [abgerufen am 26.06.2014].
[16] SRT AppGuard, Benutzerhandbuch, S. 18, abrufbar unter: http://apps.backes-srt.com/appguard/wp-content/uploads/sites/2/2014/03/AppGuard_Benutzerhandbuch.pdf, [abgerufen am 26.06.2014].
[17] [Ko12], S. 24 und S. 47.
[18] [LM06], S. 46.
[19] [LM06], S. 12.

2.2.2.1 Schutzgegenstand nach § 69a UrhG

Gemäß § 69a UrhG genießen auch Computerprogramme den Schutz des Urheberrechtsgesetzes. Der Begriff Computerprogramm wurde in § 1 (i) der Mustervorschriften für den Schutz von Computersoftware definiert als „Folge von Befehlen [...], die nach Aufnahme in einen maschinenlesbaren Träger fähig sind zu bewirken, dass eine Maschine mit informationsverarbeitenden Fähigkeiten eine bestimmte Form oder Aufgabe oder ein bestimmtes Ergebnis anzeigt, ausführt oder erzielt“.[20] Diese Definition wurde insoweit von der Literatur und Rechtsprechung übernommen.[21] Die Bundesregierung entschied sich aufgrund schneller technischer Entwicklungen gegen eine gesetzliche Definition.[22]

Der Bytecode ist eine Folge von Befehlen, die nach weiterer Übersetzung ein bestimmtes Ergebnis haben, nämlich die App ausführen. Es kommt für einen urheberrechtlichen Schutz nach § 69a UrhG weder auf die Programmiersprache an noch darauf, in welchem Code das Programm vorliegt. Es können auch Quellcodes, Maschinencodes, sowie der hier relevante Objektcode geschützt sein[23], obwohl dieser nur maschinell lesbar ist.[24] Der Bytecode ist dem Schutzgegenstand also grundsätzlich zuzuordnen.

Ob jeweils die Schutzvoraussetzungen von § 96a Abs. 3 UrhG vorliegen, nämlich eine ausreichende geistige Schöpfung des Urhebers, muss grundsätzlich im Einzelfall gesondert entschieden werden. Bei größeren Computerprogrammen kann jedoch in der Regel davon ausgegangen werden, da die Anforderungen an die Schöpfungshöhe seit Verabschiedung der EU-Richtlinie zum Computerprogrammschutz deutlich gesenkt wurden. Auch die Gesetzesbegründung weist ausdrücklich darauf hin, dass der Urheberrechtsschutz die Regel, die fehlende Schöpfungshöhe die Ausnahme sein soll.[25] Außerdem ist hinzuzufügen, dass durch Veränderungen des Bytecodes die Ausführung der gesamten Anwendung verändert werden kann.

2.2.2.2. Zustimmungsbedürftigkeit nach § 69c Nr. 2 UrhG

Kommt man deshalb zu dem Schluss, dass die vom SRT AppGuard modifizierten Apps von § 69a UrhG erfasst werden, so stellt sich ein beachtliches Problem für Smartphone-Nutzer im Rahmen ihres Selbstdatenschutzes. Bei der Modifikation handelt es sich um eine Umarbeitung der App, die nach § 69c Nr. 2 UrhG grundsätzlich zustimmungsbedürftig ist, da eine sonstige Programmänderung wie Erweiterung, Einschränkung und Ergänzung des Funktionsumfangs ausreicht, um eine Umarbeitung zu bejahen.[26] Das ausschließliche Recht dazu liegt lediglich beim Rechtsinhaber. Das ist aber nicht der Käufer der jeweiligen App, sondern der Urheber selbst.[27] Der App-Nutzer erwirbt durch den Kauf der jeweiligen Anwendung nicht das Urheberrecht an der selbigen, sondern nur

[20] Denkschrift über den Rechtsschutz der Datenverarbeitungssoftware, GRUR 1979, S. 306.
[21] [Sc11], S. 1086.
[22] BT-Drs. 12/4022, S. 9.
[23] [DS13], § 69a Rn. 12.
[24] [Sc13], Rn. 209.
[25] BT-Drs. 12/4022, S. 9.
[26] [DS13], § 69c Rn. 16.
[27] [Bo13], S. 722.

eine Nutzungslizenz. Eine Zustimmung wäre demnach erforderlich, die in der Regel aber fehlt.

Etwas anderes ergäbe sich nur, wenn es sich bei der App um eine Open Source Anwendung handelte. Anders als bei Closed Source-Software sind bei solcher Software Veränderungen und Weiterentwicklungen in der Regel gestattet.[28] Der Urheber verzichtet hierbei nicht vollständig auf seine Rechte, sondern es erfolgen die Übertragung eines nichtausschließlichen Nutzungsrechts für jedermann sowie die Überlassung des Quellcodes. Umarbeitungen sind gestattet, sofern sie ebenfalls nichtproprietär weitergegeben werden. Im Falle von Zuwiderhandlungen erlischt das nichtausschließliche Nutzungsrecht.[29] Ob eine App als Closed oder Open Source-Software konzipiert wurde und welchem Lizenzmodell sie unterliegt, muss im Einzelfall geprüft werden.[30] Festgestellt werden kann jedoch, dass proprietäre Software urheberrechtliche Probleme beim Selbstdatenschutz auslösen kann.

2.2.2.3 Keine Zustimmungsbedürftigkeit aufgrund von § 69d Abs. 1 UrhG

Um Smartphone-Nutzern einen besseren Schutz ihrer Daten zu ermöglichen, kam deshalb in der Literatur der Gedanke auf, ob § 69d Abs. 1 UrhG in diesem Fall eine Zustimmungsbedürftigkeit ausschließt.[31] Dazu müsste die Modifikation der Anwendung notwendig sein, um eine bestimmungsgemäße Benutzung der App zu ermöglichen.

Der Begriff „bestimmungsgemäße Benutzung“ unterliegt keiner gesetzlichen Definition. Was darunter zu verstehen ist, ist von den jeweiligen vertraglichen Vereinbarungen abhängig.[32] Erfordert eine App Berechtigungen, so wird der Käufer im Rahmen der Installation angehalten, in derartige Zugriffsrechte einzuwilligen. Bezüglich nichterforderlicher Berechtigungen gilt dies zunächst ebenso wie für Zugriffsrechte, die für die Funktionsfähigkeit der App zwingend erforderlich sind. Zu beachten ist aber auch, dass der App-Nutzer entweder in sämtliche Berechtigungen einwilligen oder gänzlich auf die App verzichten muss. Das entspricht dem heute gängigen Geschäftsmodell von „Leistung gegen Daten“.[33] Somit darf keinesfalls zweifelsfrei angenommen werden, dass der Zugriff auf kritische Daten stets vertraglich vereinbart wurde, denn eine Einwilligung bedarf zu ihrer Wirksamkeit unter anderem der Freiwilligkeit und Informiertheit. Besonders die Informiertheit ist aber problematisch, erfordert diese doch, dass der Nutzer im Voraus über den Zweck unterrichtet werden und er in der Regel genaue Kenntnis von den Umständen der Erhebung, Verarbeitung oder Nutzung der Daten erlangen muss.[34]

Bei Smartphone-Anwendungen werden derartige Kriterien in der Regel nicht erfüllt. In den meisten Fällen kommt man für nichterforderliche Berechtigungen deshalb zu dem Schluss, dass die Einwilligung als unwirksam anzusehen ist. In diesem Falle ist die ver-

[28] [DS13], § 69a Rn. 11.
[29] [DS13], § 69c Rn. 38.
[30] vgl. [BE13], S. 110 f.
[31] befürwortend: [Bo13], S. 725.
[32] [AG14], § 69d Rn. 5.
[33] [Bu10], S. 41.
[34] [BE13], S. 67.

tragliche Vereinbarung unwirksam und somit nicht vom bestimmungsgemäßen Gebrauch erfasst.

Fehlt es an einer Vereinbarung, muss sodann auf die übliche Nutzung abgestellt werden.[35] Zwar sind Smartphone-Apps in der Regel datenschutzrechtlich bedenklich, dies lässt jedoch nicht die Schlussfolgerung zu, dass die gewöhnliche Nutzung auch die rechtswidrige Erfassung von personenbezogenen Daten umfasst. Vielmehr wurde nur wirksam vereinbart, dass diejenigen Zugriffsrechte zulässig sind, die für die Funktionsfähigkeit der App erforderlich sind. Dieser Gedanke kann auch § 306 Abs. 1 BGB entnommen werden. Modifiziert der App-Nutzer nun mithilfe des SRT AppGuards die Zugriffsrechte auf das Erforderliche, so beschränkt er die Berechtigungen praktisch auf den bestimmungsgemäßen Gebrauch.[36] Von der Bedeutung des Begriffes wären derartige Modifikationen also durchaus erfasst.

Es stellt sich jedoch weiter die Frage, ob eine Beschränkung auf den bestimmungsgemäßen Gebrauch auch notwendig ist. Es reicht nicht aus, dass eine derartige Umarbeitung nützlich oder zweckmäßig scheint.[37] Stellt man jedoch fest, dass ohne eine Beschränkung der Zugriffsrechte auf das Erforderliche - mangels wirksamer Vereinbarung - bei der Nutzung der App stets eine Verletzung des Allgemeinen Persönlichkeitsrechts stattfindet und keine rechtmäßige Datennutzung vorliegen kann, so sind die Modifikationen keineswegs nur nützlich. Vielmehr besteht nur die Möglichkeit, diese Umarbeitungen vorzunehmen, soll der App-Entwickler Daten nicht rechtswidrig nutzen dürfen.

Um dem Selbstdatenschutz gerecht zu werden, erscheint es deshalb sachgerecht, § 69d Abs. 1 UrhG und den bestimmungsgemäßen Gebrauch weit auszulegen. Zwar gewähren die App-Anbieter ihre Anwendungen in der Regel nur deshalb kostenlos, um im Gegenzug die angefragten Daten zu erhalten, dieses Interesse überwiegt aber keineswegs das an dem Schutz der personenbezogenen Daten eines Smartphone-Nutzers.

2.2.3 Zwischenergebnis

Da nach § 69d Abs. 1 UrhG die Zustimmungsbedürftigkeit entfällt, sind Ansprüche nach den §§ 97 ff. UrhG also in der Regel nicht denkbar. Um Smartphone-Nutzern hier eine sichere Rechtslage gewährleisten zu können und sich nicht lediglich auf eine weite Auslegung stützen zu müssen, sollte eine ausdrückliche Regelung im Gesetz aufgenommen werden.

2.3 Die Modifikation von Betriebssystemen am Beispiel von XPrivacy

Bei XPrivacy handelt es sich ebenfalls um eine Anwendung, die dem Nutzer ein dynamisches Rechtemanagement ermöglicht. Der Smartphone-Nutzer kann auch hier mithilfe der Anwendung Berechtigungen von Apps beschränken, indem er sich deren Zugriffsrechte anzeigen lässt, um sodann durch Entfernen der jeweiligen Häkchen die Berechti-

[35] [AG14], § 69d Rn. 5.
[36] vgl. [Bo13], S. 725.
[37] [DS13], § 69d Rn. 11.

gung zu entziehen. Zudem gibt die Anwendung Auskunft darüber, welche Daten bereits von der App verwendet wurden. Als Folge der individuellen Einstellungen des Nutzers blockiert XPrivacy in den meisten Fällen jedoch nicht das Senden von Daten per se, sondern übermittelt entweder falsche Informationen wie z.B. einen anderen Standort oder leere Datensätze wie z.B. eine leere Kontaktliste, um dem Schutz personenbezogener Daten gerecht zu werden.[38] Hierin liegt ein Vorteil im Vergleich zum SRT AppGuard, da somit mangels blockierter Zugriffsrechte eine nahezu reibungslose Funktionsfähigkeit der App sichergestellt wird. Die Anwendung ist deshalb unter datenschutzrechtlichen Gesichtspunkten eine effektive Schutzmöglichkeit.

2.3.1 Technische Grundlagen der Modifikation von Betriebssystemen

Um XPrivacy nutzen zu können, bedarf es jedoch der vollen Zugriffs- und Schreibrechte auf dem Smartphone. Diese erlangt man via „Root“ bei Android bzw. „Jailbreak“ bei Apple iOS.[39] Hierunter versteht man die Modifikation des Betriebssystems, um sodann Administratorrechte zu erhalten. Das Rootkonto ist also ein Superuserkonto, das dem Nutzer auch ermöglicht, Kernfunktionen zu verändern. Bei „Root“ und „Jailbreak“ handelt es sich im Wesentlichen lediglich um verschiedene Begrifflichkeiten für dasselbe.[40] Der Nachteil besteht für den durchschnittlichen Smartphone-Nutzer jedoch in der Erforderlichkeit gewisser technischer Kenntnisse und der Entstehung von Risiken. So sind mobile Endgeräte nach dem „Rooten“ durchaus anfälliger für Malware, und es mangelt an Updates. Zudem können Schäden am Smartphone entstehen, sodass dessen Funktionsfähigkeit verlustig gehen kann („bricked device“[41]). Ein weiterer Nachteil liegt darin, dass viele Hersteller nach derartigen Eingriffen die Garantie versagen und man sich lediglich auf die normalen Gewährleistungsansprüche berufen kann.[42]

2.3.2 Urheberrechtliche Bewertung der Modifikation von Betriebssystemen

In beiden Fällen wird zunächst das Betriebssystem modifiziert, um sodann die betroffene Anwendung zu installieren. Dies wirft erneuert urheberrechtliche Fragestellungen auf, denn auch Firmware wie Android und Apple iOS ist von den §§ 69a ff. UrhG erfasst, sodass die Modifikation via „Root“ oder „Jailbreak“ eine Umarbeitung nach § 69c Nr. 2 UrhG darstellt.

Hier ist jedoch zwischen der Open Source-Software Android und der Closed Source-Software Apple iOS zu unterscheiden.

---

[38] vgl. Beschreibung des Google-Play-Stores, abrufbar unter: https://play.google.com/store/apps/details?id=biz.bokhorst.xprivacy.installer&hl=de, [abgerufen am 26.06.2014].
[39] Chip Online vom 09.12.2013, Android rooten: Vorteile und Nachteile, abrufbar unter: http://www.chip.de/news/Android-rooten-Vorteile-und-Nachteile_62681106.html, [abgerufen am 26.06.2014].
[40] [KK12], S. 100.
[41] [KK12], S. 99.
[42] [KK12], S. 101.

2.3.2.1 Android

Bei der Open Source-Software Android folgen aus der Umarbeitung keine besonderen urheberrechtlichen Probleme, da derartige Modifikationen von der Apache Software License, Version 2.0[43], erfasst sind.[44]

2.3.2.2 Apple iOS

Bei Apple iOS stellen sich derartige Umarbeitungen jedoch anders dar, da es sich hier um eine Closed Source-Software handelt. Es fehlt an einer Zustimmung zur Umarbeitung nach § 69c Nr. 2 UrhG. Hier kann auch keine Lösung über § 69d Abs. 1 UrhG getroffen werden, da der Smartphone-Nutzer durch „Jailbreak" nicht das Betriebssystem auf den bestimmungsgemäßen Gebrauch beschränken muss, um sich vor rechtswidriger Datennutzung durch die Firmware zu schützen. Vielmehr geht es beim „Jailbreak" erst in einem weiteren Schritt um den Selbstdatenschutz, nämlich sobald der Programmablauf der Anwendungen verändert wird. Die Veränderung des Betriebssystems iOS stellt demnach nach deutschem Recht eine Urheberrechtsverletzung dar, die Ansprüche nach §§ 97 ff. UrhG nach sich ziehen kann. Lediglich § 108b UrhG ist bei Modifikationen zum privaten Gebrauch nicht anwendbar.

In den USA wurde „Jailbreak" vor allem 2010 thematisiert und vorübergehend eine Ausnahmeregelung vom Digital Millenium Copyright Act beschlossen. Auch im US-amerikanischen Recht ist eine Umgehung von technischen Maßnahmen zum Urheberschutz grundsätzlich verboten (Section 1201(a)(1)(A) des Digital Millenium Copyright Act[45]). Gemäß Section 1201(a)(1)(C)[46] legt jedoch der Librarian of Congress, auf Empfehlung des Register of Copyrights, alle drei Jahre Ausnahmeregelungen für bestimmte Handlungen fest, sodass gewisse Umgehungen von Zugangsbeschränkungen für drei Jahre als legal anzusehen sind. Im Jahre 2010 wurde „Jailbreak" als eine solche Ausnahme aufgenommen.[47] Eine vergleichbare Diskussion fand in Deutschland bisher nicht in diesem Maße statt, sodass es an gleichwertigen Regelungen fehlt.

---

[43] Die Version 2.0 ist abrufbar unter: http://www.apache.org/licenses/LICENSE-2.0, [abgerufen am 26.06.2014].

[44] vgl. Android Licenses, abrufbar unter: https://source.android.com/source/licenses.html, [abgerufen am 26.06.2014].

[45] "§ 1201 (a) Violations Regarding Circumvention of Technological Measures. — (1)(A) No person shall circumvent a technological measure that effectively controls access to a work protected under this title. The prohibition contained in the preceding sentence shall take effect at the end of the 2-year period beginning on the date of the enactment of this chapter."

[46] "§ 1201 (a) Violations Regarding Circumvention of Technological Measures. — (1)(C) During the 2-year period described in subparagraph (A), and during each succeeding 3-year period, the Librarian of Congress, upon the recommendation of the Register of Copyrights, who shall consult with the Assistant Secretary for Communications and Information of the Department of Commerce and report and comment on his or her views in making such recommendation, shall make the determination in a rulemaking proceeding for purposes of subparagraph (B) of whether persons who are users of a copyrighted work are, or are likely to be in the succeeding 3-year period, adversely affected by the prohibition under subparagraph (A) in their ability to make noninfringing uses under this title of a particular class of copyrighted works."

[47] Federal Register, Vol. 75, No. 143/Tuesday, July 27, 2010/Rules and Regulations, III. B. "Computer programs that enable wireless telephone handsets to execute software applications, where circumvention is ac-

Abgesehen von den urheberrechtlichen Problemen der Modifikation von iOS selbst, stellt auch die Einwirkung des gerooteten Smartphones auf die Programmabläufe von Apps eine urheberrechtlich relevante Handlung dar, obwohl in diesem Fall keine Modifikation des Bytecodes erfolgt. Laut OLG Hamburg ist keine Veränderung der Programmsubstanz erforderlich.[48] Es genüge, wenn durch externe Befehle in die Ausführung und den Ablauf eingegriffen werde. Der Begriff Umarbeitung sei weit auszulegen.[49] Durch diese weite Auslegung werden Rechtslücken durch technische Neuerungen vermieden. So stellt auch das Ausführen von XPrivacy grundsätzlich eine zustimmungsbedürftige Handlung nach § 69c Nr. 2 UrhG dar. Es ist jedoch auch hier wieder § 69d Abs. 1 UrhG zu beachten, da in diesem Fall die Anwendung auf den bestimmungsgemäßen Gebrauch beschränkt wird, was bei rechtswidriger Datenerhebung notwendig ist. Hier kann auf die Ausführungen zum SRT AppGuard verwiesen werden.

2.3.3 Zwischenergebnis

Ansprüche nach §§ 97 ff. UrhG kommen also nur wegen der Umarbeitung der Firmware iOS in Betracht, nicht jedoch hinsichtlich der Einwirkung von XPrivacy auf den Ablauf der betroffenen App. Schadensersatzansprüche als Folge der Umarbeitung der Firmware sind allerdings wenig wahrscheinlich, da die zuverlässige Berechnung der Schadenshöhe meist nicht möglich sein dürfte.[50] Unterlassungs- und Vernichtungsansprüche sind hingegen denkbar.

3 Fazit und Ausblick

Zusammenfassend lässt sich feststellen, dass es zwar bereits Möglichkeiten gibt, den Selbstdatenschutz von Smartphone-Nutzern zu verbessern, diese jedoch zum einen für den durchschnittlichen Bürger nur schwer durchführbar sind, und zum anderen zu Urheberrechtsverletzungen führen können.

So erfordern die jeweiligen Mechanismen gewisse technische Kenntnisse, die nicht von jedermann erwartet werden können. Dies gilt insbesondere bei Anwendungen wie XPrivacy, die Modifikationen am Betriebssystem erfordern. Bei nicht ausreichender technischer Kenntnis kann es hier zu riskanten Beschädigungen kommen.

---

complished for the sole purpose of enabling interoperability of such applications, when they have been lawfully obtained, with computer programs on the telephone handset", S. 43828, abrufbar unter: http://www.copyright.gov/fedreg/2010/75fr43825.pdf, [abgerufen am 26.06.2014]; vgl. The Center for Internet and Society at Stanford Law School, Library of Congress: Fair use lets you jailbreak your iPhone, abrufbar unter: http://cyberlaw.stanford.edu/blog/2010/07/library-congress-fair-use-lets-you-jailbreak-your-iphone, [abgerufen am 26.06.2014].

[48] OLG Hamburg, Urteil vom 13.04.2012 – AZ 5 U 11/11, GRUR-RR 2013, S. 13.

[49] OLG Hamburg, Urteil vom 13.04.2012 – AZ 5 U 11/11, GRUR-RR 2013, S. 15.

[50] Gäfgen, Jailbreak, Root und Custom-Rom für Android und iOS legal oder nicht? Rechtslage kommentiert mit Fallbeispielen, PCGamesHardware.de vom 21.03.2012, abrufbar unter: http://www.pcgameshardware.de/Panorama-Thema-233992/Specials/Jailbreak-iphone-4s-ipad-3-Root-Custom-rom-873885/4/, [abgerufen am 26.06.2014].

Bezüglich Apple iOS ist zudem hinzuzufügen, dass derartige Modifikationen an der Firmware nach deutschem Recht eine Urheberrechtsverletzung darstellen.

Bei der Anwendung SRT AppGuard liegt der Vorteil zwar in den geringeren notwendigen Kenntnissen und der fehlenden Urheberrechtsverletzung hinsichtlich der Firmware, da keine Modifikation am Betriebssystem vorgenommen werden muss, sie blockiert jedoch die Zugriffsrechte per se und kann so zur vollständigen Funktionsunfähigkeit der modifizierten App führen. Dies entspricht nicht dem Ziel eines effektiven Selbstdatenschutzes. Zudem ist die Rechtslage bezüglich urheberrechtlicher Fragestellungen bei der Modifikation an datenschutzunfreundlichen Apps nicht ausreichend geklärt. Es bedarf deshalb einer gesetzlichen Klarstellung, dass derartige Umarbeitungen von § 69d UrhG als umfasst gelten, um Smartphone-Nutzern Rechtssicherheit zu verschaffen und die Motivation zum Selbstdatenschutz zu stärken.

Literaturverzeichnis

[AG14] Ahlberg, H.; Götting, H.-P.: Beck'scher Online-Kommentar Urheberrecht, München, Stand 01.02.2014.

[BE13] Baumgartner, U.; Ewald, K.: Apps und Recht, München, 2013.

[Bo13] Bodden, E.; Rasthofer, S.; Richter, P.; Roßnagel, A.: Schutzmaßnahmen gegen datenschutzunfreundliche Smartphone-Apps – Technische Möglichkeiten und rechtliche Zulässigkeit des Selbstdatenschutzes bei Apps, in: Datenschutz und Datensicherheit 2013, S. 720-725.

[Bu10] Buchner, B.: Die Einwilligung im Datenschutzrecht – vom Rechtfertigungsgrund zum Kommerzialisierungsinstrument, in: Datenschutz und Datensicherheit 2010, S. 39-43.

[DS13] Dreier, T.; Schulze, G.: Urheberrechtsgesetz: UrhG - Urheberrechtswahrnehmungsgesetz, Kunsturhebergesetz Kommentar, 4. Auflage, München, 2013.

[Ha11] Haag, N.: iPhone – Apps übermitteln unbemerkt vertrauliche Daten, MultiMedia und Recht-Aktuell 315048.

[KK12] Kersten, H.; Klett, G.: Mobile Device Management, Heidelberg, 2012.

[Ko12] Koppay, H.: Die Entwicklung und Vermarktung von Handy-Apps – Einstieg in die Welt der mobilen Applikationen, Hamburg, 2012.

[Kr12] Kremer, S.: Datenschutz bei Entwicklung und Nutzung von Apps für Smart Devices, in: Computer und Recht 2012, S. 438-446.

[LM06] Louis, D.; Müller, P.: Java 5 – Easy: Aufregend Programmieren, München, 2006.

[Sc11] Schwartmann, R.: Praxishandbuch Medien-, IT- und Urheberrecht, 2. Auflage, Heidelberg, 2011.

[Sc13] Schack, H.: Urheber- und Urhebervertragsrecht, 6. Auflage, Tübingen, 2013.

[SM13] Sachs, A.; Meder, M.: Datenschutzrechtliche Anforderungen an App-Anbieter, Prüfungen am Beispiel von Android-Apps, in: Zeitschrift für Datenschutz 2013, S. 303-308.

Big Data und der Grundsatz der Zweckbindung

Markus Kring, Ass. iur.

Zentrum für Angewandte Rechtswissenschaft (ZAR)
Karlsruher Institut für Technologie (KIT)
Vincenz-Prießnitz-Str. 3
76131 Karlsruhe
markus.kring@kit.edu

Abstract: Der Artikel definiert den Begriff Big Data und erläutert sodann den datenschutzrechtlichen Grundsatz der Zweckbindung. Am Beispiel der Erstellung und Auswertung von Kundenprofilen wird dargelegt, dass die Zweckbindung hiermit nur schwer zu vereinbaren ist. Auch die derzeit genannten Lösungsansätze vermögen dabei nicht weiterzuhelfen. Da der Entwurf der Datenschutzgrundverordnung der EU am Prinzip der Zweckbindung festhält, ist eine vertiefte Auseinandersetzung mit diesem Grundsatz nötig.

1 Einleitung

Oft wird festgestellt, dass Big Data-Anwendungen mit dem Grundsatz der Zweckbindung schwer zu vereinbaren[1] und hierdurch „tendenziell ausgeschlossen“[2] seien. Big Data sehe gerade keine Zwecksetzung und Zweckbindung vor.[3] Das Zweckbindungsprinzip gehöre auf den Prüfstand.[4] Teilweise wird sogar das gänzliche Streichen oder zumindest drastische Verändern des Prinzips gefordert, da es inkonsistent mit der heutigen Datennutzung sei.[5]

Im Folgenden soll der Frage nachgegangen werden, was sich hinter Big Data und dem Grundsatz der Zweckbindung verbirgt und inwiefern diese beiden miteinander vereinbar sind. Zur besseren Verständlichkeit soll hierbei auf das Anwendungsszenario des Erstellens von Kundenprofilen und deren Auswertung durch Big Data-Anwendungen abgestellt werden.

[1] [FF13], S. 303.
[2] [We13], S. 256.
[3] [Ro13], S. 564.
[4] [Hä14], Annex, Rn. 48.
[5] [CM13], S. 72.

2 Big Data

Eine allgemein anerkannte Definition des Begriffs Big Data gibt es bislang nicht.[6] Er wird aber zumeist mit den sogenannten drei „V“ beschrieben – Volume, Velocity und Variety. In manchen Definitionen werden diese Merkmale ergänzt durch Veracity und Value. Zur Konkretisierung wird nun auf die einzelnen Eigenschaften eingegangen.

Volume beschreibt große Datenmengen, die aufgrund des günstigen Speicherplatzes und einer einfachen Skalierbarkeit mittels Cloud Computing von Big Data-Anwendungen ausgewertet werden können. Herkömmliche Datenbank-Technologien können diese Menge an Daten hingegen nicht mehr sinnvoll auswerten.[7] Als Beispiel hierfür wird gerne darauf verwiesen, dass die weltweite Datenmenge exponentiell ansteigt. Es ist von einer Verdoppelung alle zwei Jahre die Rede.[8] Somit werden in Zukunft noch mehr Daten zur Verfügung stehen.

Velocity bezeichnet die Geschwindigkeit der Datenverarbeitung. Die Daten können mit Hilfe neuer Analyseverfahren unter Einsatz des Frameworks Hadoop, das auf dem Map Reduce Framework von Google basiert,[9] nahezu in Echtzeit ausgewertet werden.[10] Dies kann durch eine „divide-and-conquer-Strategie“ auf mehreren Rechenknoten gleichzeitig und unabhängig voneinander geschehen.[11]

Mit Variety wird die Vielzahl von Daten aus unterschiedlichen Quellen angesprochen.[12] In einem klassischen Data Warehouse erfolgte eine relationale Speicherung der Daten, d.h. beispielsweise in Tabellenform und daher konnten nur entsprechend strukturierte Daten dort gespeichert und verarbeitet werden.[13] Durch den Einsatz nicht relationaler Datenbanken (NoSQL: Not Only Structured Query Language)[14] ist aber nunmehr das Speichern und Auswerten unstrukturierter Daten, z.B. des Texts einer E-Mail, eines Fotos oder eines Tweets möglich.[15]

Das teilweise zusätzlich genannte Merkmal Veracity nimmt das Problem der Richtigkeit der Daten in den Blick.[16] So können die Daten veralten und deshalb nicht mehr der Realität entsprechen. Zudem kann es sich um von vornherein unrichtige Angaben handeln. Es geht also darum, die Genauigkeit der Daten möglichst korrekt einzuschätzen und dies in die Datenverarbeitung einfließen zu lassen.[17]

---

[6] Derzeit beschäftigen sich einige Standardisierungsorganisationen mit dem Versuch einer Definition: http://jtc1bigdatasg.nist.gov/home.php; http://bigdatawg.nist.gov/home.php [beide abgerufen am 22.04.2014].
[7] [FF13], S. 303; [We13a] S. 10.
[8] [Po13], S. 6.
[9] [Bi12], S. 27.
[10] [Ul13] S. 227.
[11] [Po13], S. 7.
[12] [Ul13], S. 227.
[13] [Sc03], Rn. 72.
[14] [Bi12], S. 96.
[15] [We13a], S. 10.
[16] [We13a], S. 10.
[17] [We13a], S. 10.

Zuletzt sei noch der Begriff Value definiert. Hiermit wird angesprochen, dass den Daten ein großer Wert innewohne, den es durch ihre Auswertung zu realisieren gelte. Dabei geht es vor allem um Effizienzsteigerungen durch den gezielteren Einsatz von Ressourcen.[18]

Zusammengefasst steht Big Data für die Auswertung riesiger, unterschiedlich strukturierter Datenbestände in nahezu Echtzeit.[19] Big Data-Anwendungen zeichnen sich insbesondere dadurch aus, dass unter Einsatz verschiedener Analyse-Algorithmen[20] nach bislang unbekannten Zusammenhängen gesucht wird.[21] Dabei kommen teils Verfahren zum Einsatz, die automatisiert Annahmen generieren und diese dann überprüfen.[22] So können Korrelationen festgestellt werden, ohne dass damit aber eine Aussage über Kausalitäten getroffen wäre.[23]

3 Kundenprofile (Profiling)

Die Auswertung von Daten ist insbesondere im Bereich des Marketings interessant. So sollen Kunden an das Unternehmen gebunden, ihr Abwandern verhindert und zugleich der Umsatz je Kunde maximiert werden.[24] Durch die Zusammenführung vieler Einzeldaten wird es möglich über die Einzelinformationen hinaus weitere Informationen zu gewinnen und so zum Beispiel das Kaufverhalten zu erkennen.[25] Im Folgenden sollen die datenschutzrechtlichen Rahmenbedingungen dieser Sammlung und Auswertung von Kundendaten betrachtet werden. Im Mittelpunkt steht dabei die Vereinbarkeit mit dem datenschutzrechtlichen Grundsatz der Zweckbindung.

4 Datenschutzrechtliche Rahmenbedingungen

Sofern personenbezogene Daten unter Einsatz von Datenverarbeitungsanlagen durch nicht-öffentliche Stellen erhoben, verarbeitet oder genutzt werden, findet das Bundesdatenschutzgesetz (BDSG) nach § 1 Abs. 2 Nr. 3 BDSG Anwendung. Das BDSG ist aber nur subsidiär anwendbar. Für den Bereich der Telemedien und der Telekommunikationsdienste gibt es bereichsspezifische Regelungen[26] mit einer strikten Zweckbindung,[27] auf die später kurz eingegangen wird. Das BDSG statuiert in § 4 Abs. 1 BDSG ein Verbot mit Erlaubnisvorbehalt:[28] Personenbezogene Daten dürfen nur erhoben, verarbeitet oder genutzt werden, wenn eine Rechtsvorschrift oder die Einwilligung des Betroffenen dies gestattet. Nach § 3 Abs. 1 BDSG sind personenbezogene Daten Einzelangaben über

[18] [Ul13], S. 228.
[19] [Bi12], S. 21.
[20] Einen Überblick verschiedener Analyseverfahren bietet [Bi14], S. 57 ff.
[21] [Ro13], S. 562.
[22] [Bü00], S. 12; [Sc03], Rn. 29; [We03], S. 119.
[23] [Ru13], S. 76.
[24] [Sc03], Rn. 3, 28.
[25] [Sc03], Rn. 35; [Wi00], S. 59.
[26] § 88 Abs. 3 S. 1 TKG, § 12 Abs. 2 TMG.
[27] [Si11], Rn. 7f.
[28] [GS12], Rn. 3.

persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person. Sofern es sich um rein produktbezogene Daten handelt, findet das BDSG daher keine Anwendung.

5 Grundsatz der Zweckbindung

Bereits im Volkszählungsurteil des Bundesverfassungsgerichts von 1983 wurde die Bedeutung der Zweckbindung hervorgehoben. Der Verwendungszweck müsse „bereichsspezifisch und präzise bestimmt“ und die Angaben für diesen Zweck geeignet und erforderlich sein.[29] Mitunter ist der Zweckbindungsgrundsatz als ein „Kernstück datenschutzrechtlicher Regelungen“[30] bezeichnet worden. Der Grundsatz der Zweckbindung hat dennoch im BDSG keine eigene Regelung erfahren.[31] Er findet aber in einer Reihe von Bestimmungen Erwähnung.[32] Auch auf europäischer Ebene wurde die Bedeutung der Zweckbindung erkannt. So schreibt Art. 6 Abs. 1 lit. b der Richtlinie 95/46/EG unter anderem vor, dass „… personenbezogene Daten für festgelegte eindeutige und rechtmäßige Zwecke erhoben und nicht in einer mit diesen Zweckbestimmungen nicht zu vereinbarenden Weise weiterverarbeitet werden.“

Es ist kritisiert worden, dass der Grundsatz der Zweckbindung sowohl in Bezug auf den Begriff des Zweckes als auch der daran anknüpfenden Bindung sehr offen formuliert und damit einer weiten Interpretation zugänglich sei.[33] Über die Bedeutung der Zweckbindung bestehe erhebliche Unklarheit.[34] Daher soll nun auf die beiden Elemente des Zweckbindungsgrundsatzes eingegangen werden: die Zweckfestlegung und die Zweckbindung.[35]

5.1 Zweckfestlegung

Sowohl im Falle einer Einwilligung als auch bei einer auf einer gesetzlichen Erlaubnisnorm beruhenden Datenerhebung ist dem Betroffenen der Zweck der Erhebung, Verarbeitung oder Nutzung mitzuteilen.[36] Dies können auch mehrere Zwecke sein.[37] Die Zwecksetzung muss bereits bei der Erhebung der Daten geschehen.[38]

Fraglich ist, welche Anforderungen an den festzulegenden Zweck zu stellen sind. Für die Konkretisierung bietet sich zunächst die Klärung der Wortbedeutung an. Dem Duden ist folgende Definition zu entnehmen: „etw. was jmd. mit einer Handlung beabsichtigt, zu

[29] [Bu84], S. 46.
[30] [LZ08], Rn. 102.
[31] [Ze03], Rn. 6.
[32] §§ 4, 4a, 4b, 4e, 28-32, 39, 40, Anlage zu § 9 BDSG.
[33] Europäische Kommission SEC 2012, 72 final, Annex 2 Impact Assessment, S. 25. http://ec.europa.eu/justice/data-protection/document/review2012/sec_2012_72_en.pdf [abgerufen am 22.04.2014].
[34] [Bu98], S. 223.
[35] [Ar13], S. 4; [Tr03], Rn. 36 ff.
[36] §§ 4 Abs. 3 S. 1 Nr. 2, 4a Abs. 1 S. 2, § 28 Abs. 1 S. 2, § 33 Abs. 1 S. 1 BDSG.
[37] [Ar13], S. 16; [LZ08], Rn. 103; [Ta13], § 4 Rn .75.
[38] [Ar13], S. 15.

bewirken, zu erreichen sucht; [Beweggrund u.] Ziel einer Handlung.“[39] Es handelt sich also um das Ziel einer Handlung. Anforderungen an die Konkretisierung des Ziels sind dem Begriff aber nicht zu entnehmen. Theoretisch wäre es daher denkbar den Zweck so umfassend zu formulieren, dass sich darunter alles fassen ließe. Beispielsweise könnte als Zweck angegeben werden: Auswertung der Daten durch Algorithmen zur Feststellung von Korrelationen. Bei einer derart weiten Zweckfestlegung kommt dieser aber keine eingrenzende Funktion zu.[40]

Wohl aus diesem Grunde sieht § 28 Abs. 1 S. 2 BDSG vor, dass die Zwecke konkret festzulegen sind. Diese Bestimmung wurde in Umsetzung der Richtlinie 95/46/EG eingefügt,[41] die vorsieht, dass die Zwecke eindeutig festgelegt werden. Auch insoweit im Gesetzestext keine Konkretisierung des Begriffs erfolgt, ist somit im Rahmen einer richtlinienkonformen Auslegung eine eindeutige Zweckbestimmung zu fordern.

Dieses Ergebnis wird auch durch eine teleologische Auslegung gestützt. Das BDSG dient dem Schutz des Persönlichkeitsrechts des Einzelnen.[42] Die informationelle Selbstbestimmung ist nur dann gewährleistet, wenn der Einzelne selbst bestimmen kann, wer, was, wann und bei welcher Gelegenheit über ihn weiß.[43] Gerade die Kenntnis des Zweckes ermöglicht dem Einzelnen, zu wissen, welche personenbezogenen Daten erhoben und verarbeitet werden und die Preisgabe von Daten selbst zu steuern.[44] Das Erfordernis einer informierten Einwilligung kann nur erfüllt werden, wenn der Betroffene präzise über den Zweck informiert wird. Die Angabe pauschaler Zwecke ist nicht ausreichend.[45] Des Weiteren ist die Zweckfestlegung die Voraussetzung für die Zweckbindung. Eine Zweckbindung ist nur bei eindeutig bestimmten Zwecken sinnvoll.[46] Vor allem ist aber nur hierdurch eine Kontrolle der Rechtmäßigkeit der Datenverarbeitung möglich, da die nötige Transparenz entsteht.[47] Gemäß dem Erforderlichkeitsprinzip dürfen nur jene Daten erhoben werden, die für den Zweck notwendig sind.[48] Ein wesentlicher Grundsatz[49] des Datenschutzrechts ist also nur bei einer eindeutigen Zweckfestlegung durchführbar. Aus dem Erforderlichkeitsprinzip folgt zudem, dass Daten, die für den Ursprungszweck nicht mehr benötigt werden, gelöscht werden müssen.[50]

Die Zweckfestlegung nimmt also eine zentrale Rolle im Datenschutzrecht ein. Ihrer Funktion kann sie aber nur gerecht werden, wenn der Zweck detailliert genug ist, um festzustellen, welche Art der Datenverwendung davon umfasst ist.[51] Eine Datenerhebung auf „Vorrat“ für nicht näher definierte Nutzungsmöglichkeiten ist damit nicht zu verein-

[39] [Du14], S. 2391.
[40] Vgl. [Ro06], S. 12. [Ro13] S. 565.
[41] [Si11], Rn. 2.
[42] § 1 Abs. 1 BDSG
[43] [Bu84], S. 43.
[44] [Ro13], S. 564; [Ze03], Rn. 4.
[45] [BMH13], § 4 Rn. 43.
[46] [Ho91], S. 127.
[47] [Si11], Rn. 43, 168; [Ta13], § 28 Rn. 111.
[48] [We13], S. 256.
[49] Ähnlich [GS12], § 13 Rn. 3: „prägendes Element des Datenschutzes“.
[50] [Ro13], S. 565; [Ze03], Rn. 43.
[51] [Ar13], S. 15.

baren.[52] Es kann gerade keine generelle Aussage dahingehend getroffen werden, dass eine „typische Darstellung des Sachverhalts“ in „generalisierender Form ausreichend sei.[53] Wie präzise der Zweck angeben werden muss, lässt sich nicht mit mathematischer Genauigkeit sagen, sondern hängt vielmehr von den konkreten Umständen des Einzelfalls ab.[54] Im Falle eines Kaufvertrags mag die allgemeine Formulierung „zur Durchführung des Kaufvertrags“ in der Tat genügen.[55] Die Erstellung von Kundenprofilen lässt sich darunter aber nicht fassen.[56] Jedenfalls darf der Grundsatz der Zweckbindung nicht durch einen allumfassenden Primärzweck ausgehebelt werden.[57] Unklarheiten gehen zu Lasten der verantwortlichen Stelle.[58]

Es ist umstritten, ob die Erstellung von Kundenprofilen für Werbezwecke und ihre Auswertung durch Big Data-Analyseverfahren den Anforderungen an die Zweckbestimmung genügt. Einerseits wird vertreten, dass die Angabe des Zwecks „Werbung“[59] oder der Hinweis, dass Daten „zu Zwecken der Profilerstellung oder Marktanalyse“[60] verwendet werden, genüge. Es könne nicht jeder Analysevorgang als Verwendungszweck aufgezählt werden.[61] Ausschließlich der „Endzweck“ der Verarbeitung sei für den Betroffenen entscheidend.[62] Von Werbezusendungen sei er nur minimal betroffen.[63] Andererseits wird die Angabe „zum Zwecke der Werbung“[64] oder „des Marketings“[65] als zu unbestimmt angesehen. Unter den Begriff des Marketings ließen sich viele Zwecke fassen von der Werbung für unterschiedlichste Produkte bis hin zu Produkt- und Dienstleistungsforschung.[66] Es bedürfe des Hinweises darauf, welche - zumindest der Art nach spezifizierte - Daten für diesen Zweck verarbeitet werden.[67]

Zwar mag es zutreffen, dass die Zusendung von Werbung den Betroffen nur geringfügig betrifft. Allerdings ist dies nur eine mögliche Aktion, die sich unter den Zweck „Werbung“ fassen lässt. So wären auch das Verschicken von E-Mails oder Telefonanrufe für diesen Zweck möglich. Des Weiteren ist damit nicht gesagt, welche Daten ausgewertet werden. Es ist auch nicht klar, ob die Daten mehrfach ausgewertet werden sollen. Sogar die dauerhafte Speicherung für ein Kundenprofil kann hierunter gefasst werden. Es ist daher zumindest – insbesondere für eine wirksame Einwilligung – zu fordern, dass konkret festgelegt wird, für welchen Zweck oder welche Zwecke, welche Daten ausgewertet werden. Sofern dies für das Verständnis der Zwecke erforderlich ist, bedarf es zudem

[52] [Si11], Rn. 38, 112; [We14a], Rn. 62.
[53] So aber [Pl13], Rn. 89, der dazu rät, den Zweck „möglichst weitläufig zu verfassen.“
[54] [Ar13], S. 16.
[55] So [Pl13], Rn. 89.
[56] [Bü00], S. 13; [Mö98], 566.
[57] [Da97], Rn. 7.
[58] [Da97], Rn. 13; [Si11], Rn. 42; [We03], S. 121.
[59] [Bu06], S. 1621 f.
[60] [Pl13], Rn. 57.
[61] [Pl13], Rn. 57.
[62] [Bu06], S. 1621; [Pl13], Rn. 57.
[63] [Bu06], S. 1621.
[64] [Bi12], S. 44; [Br97], S. 252; [Bü00], S. 15; [Ro13], S. 564.
[65] [Ba00], S. 8; [PP98], S. 146; [Sc03], Rn. 121.
[66] [PP98], S. 146.
[67] [Bü00], S. 15.

einer Information über die Art und Weise der Auswertung.[68] Die Angabe eines pauschalen Endzwecks wird dem nicht gerecht. Zuzustimmen ist dagegen der Auffassung, dass das konkrete Endergebnis der Auswertung noch nicht angeben werden braucht.[69] Denn dieses ergibt sich erst aus der Auswertung der Daten. Falls sich ein hinreichend konkreter Verarbeitungszweck aber erst aus dem Ergebnis der Datenanalyse ergibt, ist dies mit dem Prinzip der Zweckbindung nicht zu vereinbaren.[70] Eine Einwilligung ist dann nicht möglich.[71]

5.2 Zweckbindung

Ist ein Zweck für die Erhebung und Verwendung der Daten gesetzt, so ist dieser bei später folgenden Verwendungen grundsätzlich einzuhalten.[72] Durch die Zweckbindung wird die multifunktionale Verwendbarkeit der Daten bewusst eingeschränkt.[73] Dies ist insbesondere deshalb geboten, weil durch eine unkontrollierte Weitergabe von Daten der ursprüngliche Erhebungs- und Verwendungskontext verloren ginge.[74] Damit der Betroffene die Verwendung seiner Daten überblicken kann, muss die Zweckentfremdung eine eng auf den Einzelfall begrenzte Ausnahme bleiben.[75] Denn gerade im Falle einer nicht transparenten Zweckänderung droht diese Kontrolle verloren zu gehen.[76]

Wie der Zweckbindungsgrundsatz insgesamt, hat auch die Zweckbindung keine spezifische Regelung im BDSG erhalten.[77] Bei § 28 BDSG ergibt sie sich erst im Umkehrschluss aus den Absätzen 2, 3, 5 oder 8, die die Verwendung für einen anderen Zweck regeln.[78] Bei einer Einwilligung folgt sie daraus, dass diese sich nach § 4a Abs. 1 Satz 2 BDSG auf einen bestimmten Zweck bezieht.[79] Explizit kodifiziert wurde aber zum Schutz der Zweckbindung in S. 1 Nr. 8 der Anlage zu § 9 S. 1 BDSG die Verpflichtung zur Gewährleistung der getrennten Verarbeitung der zu unterschiedlichen Zwecken erhobenen Daten.

Für das Zusammenführen von Daten im Rahmen des Profilings stellt die Zweckbindung eine große Hürde dar. Denn durch das Zusammenführen von für spezifische Zwecke erhobenen Daten wird die jeweilige Zweckbestimmung verletzt und damit gegen die Zweckbindung verstoßen.[80]

---

[68] Ähnlich [Sc03], Rn. 123.
[69] [Ul13], S. 229.
[70] [Bü00], S. 15; [Sc03], Rn. 122.
[71] [Ba13], S. 16; [Sc03], Rn. 122.
[72] [Pl13], Rn. 89.; [So11], Rn. 42.
[73] [Ku99], S. 157.
[74] [Ho91], S. 21.
[75] [Ku99], S. 158.
[76] [Ta13], § 28 Rn. 111.
[77] [Ta13], § 28 Rn. 110.
[78] [Pl13], Rn. 88; [Ta13], § 28 Rn. 110.
[79] [Sc03], Rn. 118.
[80] [Ro13], S. 565; [Sc13], S. 225.

5.3 Durchbrechungen der Zweckbindung

Es gibt aber Ausnahmen vom Grundsatz der Zweckbindung. Wann eine Zweckänderung zulässig ist und ob hierdurch Big Data für Zwecke des Profilings zulässig ist, soll im Folgenden dargelegt werden.

Bei aufgrund einer Einwilligung erhobener Daten ist eine Zweckänderung nicht zulässig.[81] Es bedarf vielmehr einer erneuten Einwilligung für den neuen Zweck. Anders ist dies aber unter den Voraussetzungen der Erlaubnisnorm des § 28 BDSG. § 28 BDSG regelt die Datenerhebung und Verwendung für eigene Geschäftszwecke. Während Abs. 1 die Voraussetzungen einer Datenerhebung und Verwendung regelt, gestattet Abs. 2 die Übermittlung oder Nutzung für einen anderen Zweck als den Erhebungszweck. Er durchbricht also das Prinzip der Zweckbindung.[82] Nach § 28 Abs. 2 Nr. 1 i.V.m. § 28 Abs. 1 S. 1 Nr. 2 BDSG ist eine Nutzung für einen anderen Zweck zulässig, soweit dies zur Wahrung berechtigter Interessen der verantwortlichen Stelle erforderlich ist und kein Grund zu der Annahme besteht, dass das schutzwürdige Interesse des Betroffenen an dem Ausschluss der Nutzung überwiegt. Es ist also eine Abwägung der Interessen vorzunehmen.

Letztlich kommen bei § 28 Abs. 2 BDSG die gleichen Kriterien wie bei § 28 Abs. 1 BDSG zur Anwendung, weshalb die Regelung des Abs. 2 als überflüssig bezeichnet worden ist.[83] Der Aufbau einer Datenbank mit Kundendaten um diese später gezielt für die Kundenansprache auswerten zu können, stellt ein berechtigtes Interesse dar.[84] Allerdings richtet sich die Erstellung von Kundenprofilen ausschließlich zu Werbezwecken nach § 28 Abs. 3 BDSG,[85] der insofern lex specialis ist[86] und mit Ausnahme des Listenprivilegs[87] eine Einwilligung zur Voraussetzung macht. Zudem stellt das Zusammenführen von Daten zu einem Kundenprofil aufgrund des Kontextverlustes eine Veränderung der Daten und somit eine Verarbeitung und keine Nutzung dar.[88] Des Weiteren muss der nach der Zweckänderung zu verfolgende Zweck festgelegt werden, wobei zu beachten ist, dass es sich wie bei § 28 Abs. 1 BDSG um eigene Geschäftszwecke handeln muss.[89] Das Anlegen eines Kundenprofils und der Einsatz von Big Data sind daher nur zulässig, wenn die Anforderungen an die Zweckbestimmung erfüllt werden. Wie bereits dargelegt, werden die Anforderungen an die Zweckfestlegung aufgrund der Zweckoffenheit vieler Big Data-Analysen aber nicht erfüllt werden können. Auf die Möglichkeit der Zweckänderung nach § 28 Abs. 2 BDSG lässt sich der Einsatz von Big Data zu Zwecken des Profilings mangels Erfüllens der Tatbestandsvoraussetzungen somit nicht stützen.

---

[81] [Sc03], Rn. 118.
[82] [We14a], Rn. 69.
[83] [Si11], Rn. 170.
[84] [Ta13], § 28 Rn. 57; für das Profiling wird teilweise wegen des tiefen Eindringens in die Persönlichkeit vertreten, dass immer eine Einwilligung erforderlich sei, so [Ro13], S. 565; teils wird es wegen des Verbots des Erstellens totaler Persönlichkeitsbilder gänzlich für unzulässig gehalten, so [Wi00], S. 61 f.
[85] [Pl13], Rn. 56.
[86] [GS12], § 28 Rn. 42.
[87] Hierzu: [Pl13], Rn. 113 ff.
[88] [BMH13], § 3 Rn. 86; [GS12], § 3 Rn. 30.
[89] [Si11], Rn. 171.

6 Datenschutzgrundverordnung der EU

Im Kommissionsentwurf der Datenschutzgrundverordnung (DS-GVO)[90] ist der Zweckbindungsgrundsatz in Art. 5 lit. e DS-GVO enthalten. Demnach müssen personenbezogene Daten für genau festgelegte, eindeutige und rechtmäßige Zwecke erhoben werden und dürfen nicht in einer mit diesen Zwecken nicht zu vereinbarenden Weise weiterverarbeitet werden. Das entspricht der Definition in der Richtlinie 95/46/EG. Die entscheidende Frage wird daher die nach der „Vereinbarkeit" der Verarbeitung mit dem Erhebungszweck sein.[91] Da diese in jeden Einzelfall durch Abwägung zu bestimmen ist[92] und hierdurch Unsicherheiten entstehen können, ist daran Kritik geübt worden.[93]

Die DS-GVO sieht zugleich eine Ausnahme von der Zweckbindung vor. Es soll nach Art. 6 Abs. 4 DS-GVO zulässig sein Daten für einen mit dem Erhebungszweck nicht zu vereinbaren Zweck zu verarbeiten, wenn einer der Gründe des Art. 6 Abs. 1 lit. a-e vorliegt. Der Ausnahmetatbestand des Art. 6 Abs. 4 DS-GVO wurde als Herausforderung für das Zweckbindungsprinzip bezeichnet, da nun eine Datenverarbeitung gänzlich losgelöst vom ursprünglichen Zweck möglich sei[94] und eine Löschung des entsprechenden Absatzes gefordert.[95] Im Entwurf dem das EU-Parlament jüngst zugestimmt und damit zu seiner Grundlage für die weiteren Verhandlungen gemacht hat, ist der Ausnahmetatbestand des Art. 6 Abs. 4 DS-GVO nicht mehr enthalten.[96] Es bleibt also abzuwarten, in welcher Form die Zweckbindung tatsächlich verwirklicht werden wird.

7 Lösungsansätze

Da sowohl im TKG als auch im TMG eine strenge Zweckbindung vorgesehen ist, sollen nun deren Lösungsansätze betrachtet werden. Im TKG wird für die Nutzung von Standortdaten für Dienste mit Zusatznutzen, wie z.B. Navigationshilfen[97] in § 98 Abs. 1 TKG eine Einwilligung oder eine Anonymisierung[98] gefordert. Falls Big Data mit wirksam anonymisierten Daten durchgeführt wird, ist es datenschutzrechtlich unbedenklich.[99] Denn dann findet das Datenschutzrecht mangels Personenzugs der Daten keine Anwendung.[100] Eine wirksame Anonymisierung ist aber immer schwerer erreichbar, da im

[90] KOM (2012) 11 endgültig. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0011:FIN:DE:PDF. [abgerufen am 22.04.2014].
[91] Siehe hierzu [Ar13], S. 20 ff.
[92] [Ze03], Rn. 114.
[93] [PT13], S. 30.
[94] [Ar13], S. 11.
[95] [Ar13], S. 41.
[96] Legislative Entschließung des EU-Parlaments vom 12.03.2014: http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//DE&language=DE; inoffizielle konsolidierte Fassung der DS-GVO unter: http://www.janalbrecht.eu/fileadmin/material/Dokumente/DPR-Regulation-inofficial-consolidated-LIBE.pdf. [beide abgerufen am 22.04.2014].
[97] Erwägungsgrund Nr. 18 Richtlinie 2002/58/EG.
[98] § 3 Abs. 6 BDSG.
[99] [Ro13], S. 562 spricht insofern vom „Makro-Level".
[100] [We14], Rn. 49.

Rahmen von Big Data stets eine Vielzahl von Informationen verfügbar ist.[101] Es genügen oft wenige Angaben um eine Person eindeutig zu identifizieren. So sind 61 – 87 Prozent der US-amerikanischen Bevölkerung mit drei Merkmalen (Geschlecht, PLZ, Geburtsdatum) eindeutig zu identifizieren.[102] Zudem ist der Personenbezug für den gesamten Zeitraum der Speicherung zu prognostizieren.[103] Des Weiteren können bei der Auswertung der Daten neue Daten entstehen, die eine Reidentifizierung ermöglichen.[104] Es wird daher vertreten grundsätzlich von einem Personenbezug der Daten im Rahmen von Big Data-Anwendungen auszugehen, auch wenn diese anonymisiert worden sind.[105]

Geforscht wird derzeit daran, die Daten so zu zerlegen und zu vermischen, dass eine Zuordnung zu einer bestimmten Person nicht mehr möglich ist, eine Auswertung aber für statistische Zwecke vorgenommen werden kann.[106]

Diese Vorschläge helfen aber dann nicht weiter, wenn eine kundenspezifische Auswertung gewünscht ist. Im TMG wird deshalb auf die Pseudonymisierung[107] zurückgegriffen. Nach § 15 Abs. 3 TMG darf der Diensteanbieter für Zwecke der Werbung, der Marktforschung oder zur bedarfsgerechten Gestaltung der Telemedien Nutzungsprofile[108] unter Einsatz von Pseudonymen erstellen, wenn der Nutzer nicht widerspricht. Auch der Rückgriff auf eine Pseudonymisierung ist aber problematisch. Es bestehen die bei der Anonymisierung aufgezeigten Reidentifikationsprobleme. Sofern eine Reidentifizierung zwecks Kundenansprache erforderlich ist, z.B. für die Zusendung von Werbung, ist kein datenschutzrechtlicher Gewinn damit verbunden.[109]

Ein weiterer Vorschlag geht dahin, jedes einzelne Datum mit einem Zweck und Zugriffsrechten zu versehen, so dass bei der Auswertung nur die Daten herangezogen werden dürfen, die mit dem Zweck kompatibel sind.[110] In Betracht kommt dies nur bei einer nicht zweckoffenen Big Data-Analyse.[111] Es wurde bereits dargelegt, dass die Zwecke konkret zu benennen sind. Daher ist ein genereller Zweck nicht ausreichend. Ob mit konkreten Zwecken tatsächlich noch eine ausreichende Menge von Daten vorhanden ist, die dann im Rahmen von Big Data ausgewertet werden kann, ist sehr fraglich.

8 Fazit

Das Erstellen von Kundenprofilen und deren Auswertung durch Big Data ist mit dem Prinzip der Zweckbindung so gut wie nicht zu vereinbaren. Aufgrund der Notwendigkeit der Angabe eines konkreten Zwecks bei der Datenerhebung sind zweckoffene Big Data-

[101] Vgl. WP 29, 31.
[102] [Ba13], S. 15; [Ro13], S. 563;
[103] [Ro13], S. 565; [Ba13], S. 15 spricht insofern von einer „Zeitbombe."
[104] [Sc03], Rn. 60.
[105] [Ba13], S. 15; ähnlich [Ro13], S. 563.
[106] [Ro13], S. 566; [We13], S. 259.
[107] § 3 Abs. 6a BDSG.
[108] Zur Zulässigkeit solcher Nutzungsprofile ausführlich: [JL06], S. 316 ff.
[109] Vgl. [We13], S. 119.
[110] [We13], S. 120.
[111] [We13], S. 120.

Analysen mit personenbezogenen Daten unzulässig. Die derzeit vorgeschlagenen Lösungsansätze gehen fast ausschließlich in Richtung einer Anonymisierung. Aber diese ist mit einem gezielten Marketing, bei dem einzelne Kunden angesprochen werden sollen, nicht zu vereinbaren. So ist nicht verwunderlich, dass derzeit vielfach eine Abkehr vom Prinzip der Zweckbindung oder zumindest dessen Neukonzeption gefordert wird. Gleichwohl sieht der derzeitige Entwurf der DS-GVO ein Beibehalten dieses Grundsatzes vor. Daher erscheint eine vertiefte Auseinandersetzung mit dem Zweckbindungsgrundsatz und eine Konkretisierung seiner Anforderungen geboten.

Literaturverzeichnis

[Ar13] Article 29 Data protection working party: Opinion 03/2013 on purpose limitation. WP 203, 2. April 2013. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf [abgerufen am 22.04.2014].

[Ba00] Baeriswyl, B.: Data Mining und Data Warehousing: Kundendaten als Ware oder geschütztes Gut?. In: Recht der Datenverarbeitung 2000, S. 6-11.

[Ba13] Baeriswyl, B.: „Big Data“ ohne Datenschutz-Leitplanken. In: digma – Zeitschrift für Datenrecht und Informationssicherheit 2013, S. 14-17.

[BHM13] Bergmann/Möhrle/Herb, BDSG, Richard Boorberg Verlag, Stand April 2013.

[Bi12] BITKOM (Hrsg.): Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte, 2012.

[Bi14] BITKOM (Hrsg.): Big-Data-Technologien – Wissen für Entscheider, 2014.

[Br97] Breinlinger, A.: Datenschutzrechtliche Probleme bei Kunden- und Verbraucherbefragungen zu Marketingzwecken. In: Recht der Datenverarbeitung 1997, S. 247-253.

[Bü00] Büllesbach, A.: Datenschutz bei Data Warehouses und Data Mining. In: Computer und Recht 2000, S. 11-17.

[Bu84] Entscheidungen des Bundesverfassungsgerichts (BVerfGE). 65. Band. Mohr Siebeck Verlag, Tübingen 1984, S. 1-71.

[Bu98] Bull, H.: Europol und die Informationskultur. In: Lamnek/Tinnefeld (Hrsg.), Globalisierung und informationelle Rechtskultur in Europa, Nomos, Baden-Baden, 1998, S. 217-230.

[Bu06] Bull, H.: Zweifelsfragen um die informationelle Selbstbestimmung – Datenschutz als Datenaskese?. In: Neue Juristische Wochenschrift 2006, S. 1617-1624.

[CM13] Cate, H.: Mayer-Schönberger, V.: Notice and consent in a world of Big Data. In: International Data Privacy Law 2013, S. 67-73.

[Da97] Damann, U.: Kommentierung zu Art. 6 Richtlinie 95/46/EG. In: Damann/Simitis EG-Datenschutzrichtlinie, Nomos, Baden-Baden, 1997.

[Du14] Duden, Die deutsche Sprache, Wörterbuch in drei Bänden, Band 3 Q-ZZGL, Dudenverlag, 2014.

[FF13] Feiler, L.; Fina, S: Datenschutzrechtliche Schranken für Big Data. In: medien und recht 2013, S. 303-309.

[GS12] Gola/Schomerus, BDSG, C. H. Beck, München, 11. Auflage, 2012.

[Hä14] Härting, N.: Internetrecht. Verlag Dr. Otto Schmidt, Köln, 5. Auflage 2014.

[Ho91] Hoffmann, B.: Zweckbindung als Kernpunkt eines prozeduralen Datenschutzansatzes, Nomos, Baden-Baden 1991.

[JL06] Jandt, S.; Laue, P.: Voraussetzungen und Grenzen der Profilbildung bei Location Based Services. In: Kommunikation und Recht 2006, S. 316 – 322.

[Ku99] Kutscha, M.: Datenschutz durch Zweckbindung – ein Auslaufmodell?. In: Zeitschrift für Rechtspolitik 1999, S. 156 – 160.

[LZ08] Lübking, U.; Zilkens, M.: Datenschutz in der Kommunalverwaltung. Erich Schmidt Verlag, Berlin, 2008.
[Mö98] Möncke, U.: Data Warehouse – eine Herausforderung für den Datenschutz?. In: Datenschutz und Datensicherheit 1998, S. 561-569.
[Pl13] Plath, K.: Kommentierung zu § 28 BDSG. In: Plath (Hrsg.), BDSG, Verlag Dr. Otto Schmidt, Köln, 2013.
[Po13] Polzer, G.: Big Data – Eine Einführung. In: digma – Zeitschrift für Datenrecht und Informationssicherheit 2013, S. 6-9.
[PP98] Podlech, A.; Pfeifer M.: Die informationelle Selbstbestimmung im Spannungsverhältnis zu modernen Werbestrategien. In: Recht der Datenverarbeitung 1998, S. 139-154.
[PT13] Polonetsky J.; Omer T.: Privacy and Big Data: Making ends meet. In: Stanford Law Review Vol. 66, 2013 S. 25-33.
[Ro06] Roßnagel, A.: Datenschutz im 21. Jahrhundert. In: Aus Politik und Zeitgeschichte 2006, S. 9-15.
[Ro13] Roßnagel, A.: Big Data – Small Privacy? – Konzeptionelle Herausforderungen für das Datenschutzrecht. In: Zeitschrift für Datenschutz 2013, S. 562-567.
[Ru13] Rubinstein, I.: Big Data: The End of Privacy or a New Beginning?. In: International Data Privacy Law 2013, S. 74-87.
[Sc03] Scholz, P.: Datenschutz bei Data Warehousing und Data Mining. In: Roßnagel (Hrsg.), Handbuch Datenschutzrecht, C. H. Beck, München, 2003, S. 1833-1875.
[Sc13] Schaar, P.: Zwischen Big Data und Big Brother – zehn Jahre als Bundesbeauftragter für den Datenschutz und die Informationsfreiheit. In: Recht der Datenverarbeitung 2013, S. 223-227.
[Si11] Simitis, S.: Kommentierung zu § 28 BDSG. In: Simitis (Hrsg.), BDSG, Nomos, Baden-Baden, 7. Auflage 2011.
[So11] Sokol, B. Kommentierung zu § 4 BDSG. In: Simitis (Hrsg.), BDSG, Nomos, Baden-Baden, 7. Auflage 2011.
[Ta13] Taeger, J.: Kommentierung zu § 4 und § 28 BDSG. In: Taeger/Gabel (Hrsg.), BDSG, Deutscher Fachverlag, Frankfurt a. M., 2. Auflage 2013.
[Tr03] Trute, H.: Verfassungsrechtliche Grundlagen. In: Roßnagel (Hrsg.), Handbuch Datenschutzrecht, C. H. Beck, München, 2003, S. 156-187.
[Ul13] Ulmer, C.: BIG DATA – neue Geschäftsmodelle, neue Verantwortlichkeiten?. In: Recht der Datenverarbeitung 2013, S. 227-232.
[We03] Weichert, T.: Datenschutz als Verbraucherschutz. In: Recht der Datenverarbeitung 2003, S. 264-270.
[We13] Weichert, T.: Big Data und Datenschutz – Chancen und Risiken einer neuen Form der Datenanalyse. In: Zeitschrift für Datenschutz 2013, S. 251-259.
[We13a] Wespi, A.: Big Data: der nächste IT-Sicherheits-Trend?. In: digma – Zeitschrift für Datenrecht und Informationssicherheit 2013, S. 10-13.
[We14] Weichert, T.: Kommentierung zu § 3 BDSG. In: Däubler/Klebe/Wedde/Weichert, Bundesdatenschutzgesetz, Bund-Verlag, Frankfurt a. M., 4. Auflage, 2014.
[We14a] Wedde, P.: Kommentierung zu § 28 BDSG. In: Däubler/Klebe/Wedde/Weichert, Bundesdatenschutzgesetz, Bund-Verlag, Frankfurt a. M., 4. Auflage, 2014.
[Wi00] Wittig, P.: Die datenschutzrechtliche Problematik der Anfertigung von Persönlichkeitsprofilen zu Marketingzwecken. In: Recht der Datenverarbeitung 2000, S. 59-62.
[Ze03] Zezschwitz F.: Konzept der normativen Zweckbegrenzung. In: Roßnagel (Hrsg.), Handbuch Datenschutzrecht, C. H. Beck, München, 2003, S. 219-268.

A Provably Privacy Preserving Video Surveillance Architecture for an Assisted Living Community

Matthias Huber[2], Jörn Müller-Quade[1], Tobias Nilges[1], and Carolin Thal[1]

[1]Karlsruher Institut für Technologie,
{mueller-quade,nilges}@kit.edu,carolin.thal@student.kit.edu
[2]FZI Forschungszentrum Informatik, huber@fzi.de

Abstract: Video surveillance offers many advantages but also introduces privacy issues. We propose a video surveillance architecture that preserves privacy by providing different levels of anonymization. To this end, we model an assisted living scenario and show that our architecture can provide provable privacy under explicit assumptions while maintaining the utility of the architecture. This shows that video surveillance, tailored to specific scenarios, can be applied in a privacy preserving way.

1 Introduction

Video surveillance is increasingly common in every day life, for example in public places and buildings. While video surveillance is mostly used to improve the safety, these systems can also be used to compromise the privacy of individuals. To prevent misuse, laws have been established to provide transparency and set strong bounds on further use of the collected data. But laws cannot proactively prevent misuse, so privacy enhancing mechanisms can be applied to surveillance systems to enforce compliance with the law.

In literature, privacy enhancing mechanisms for video surveillance have been extensively studied. The proposed solutions, however, are mostly methods like obscuring faces [Bou05], relocating pixels [CC11], or different levels of anonymization [Sen+03; Sen+08]. These methods do not provide any privacy guarantees, i.e. it is not clear if and to which extend privacy is achieved. Therefore assumptions are necessary, e.g. that after application of method A the video data satisfies B, where B is a precise statement of the guarantee. It remains an open problem to provide reasonable privacy guarantees from methods.

In the case of database anonymization, where the model is far better defined, anonymization methods like k-anonymity [SS98] and l-diversity [KG06] stand in contrast to notions like differential privacy [Dwo06], that provide a clear guarantee. In [ABN08], the concept of k-anonymity is extended to obscure location information in moving objects databases. An individual only signalizes a location update if it traverses from one predefiend area to another. This is comparable to the map view that we describe in Section 4. For video data, it is unclear how to define a privacy metric which describes the information that is provided by a video, so apart from specific features that have to be hidden (e.g. height, color, gender), the quality of an anonymization seems to be an empirical measure.

Our contribution. We propose a system architecture for video surveillance of an assisted living community based on the NEST [MRV10] architecture. Here, an operator controls the surveillance system. The proposed system aids the operator to detect accidents and emergencies, while still providing reasonable privacy guarantees for the residents and employees. The system does not show any information about residents or employees if no one trespasses or needs help. Furthermore, we show that even when the system shows information about residents or employees, certain privacy guarantees are still fulfilled. These guarantees can improve the rate of consent for a surveillance system.

We make assumptions about the guarantees of the individual methods used in the system. In order to achieve a precise privacy guarantee for the overall system, we prove the privacy of the combination of these methods. Based on these guarantees, different views are provided for the operator, similar to [FS08]. In detail, the architecture consists of three levels. During normal operation, only guests are tracked on a *map*. In case of an (undetected) accident, a help gesture can be performed in front of a camera. The operator only sees an *indealized stick figure*. If an accident happens, the system will present the operator with a *silhouette view* of the video sequence of the accident.

The privacy has to hold against an operator who has detailed knowledge of the people that are surveilled. It is known that privacy methods do not compose in general, i.e. several anonymized datasets combined can still leak information about a single individual [NS08; SS98]. Thus, if we combine several methods in a complex architecture, the security crucially depends on the fact that the combination still preserves the privacy. We show that this is actually the case for our architecture.

2 Preliminaries

2.1 k-Anonymity

k-Anonymity is an anonymization method defined for tables. A table is a set of tuples $t = \{t_1, .., t_m\}$ with an associated set of attributes $A = \{A_1, .., A_n\}$, where every tuple represents an individual. k-anonymity classifies the attributes of a table into *identifier*, *quasi-identifier*, and *sensitive information*. In the following, w.l.o.g. we assume that a table contains only one sensitive information attribute.

Identifiers are attributes that directly allow for identifying individuals. A table that adheres k-anonymity may not contain such an attribute. A set of nonsensitive attributes $Q \subseteq A$ that linked with external data allows for the identification of at least one individual represented in the table is called quasi-identifier.

Definition 1 ([SS98]) *A table adheres k-anonymity if for each row, there are at least k-1 other rows with the same values of the quasi-identifier attributes. The set of all rows with the same attribute values of their quasi-identifiers are called an equivalence class.*

The idea of k-anonymity is that every sensitive statement in a k-anonymous table can relate

to at least k individuals, therefore, the privacy of a single individual is protected. Consider for example the table depicted in Figure 1.

		Non-Sensitive				Sensitive
		Zip Code	Age	Nationality		Condition
1		130**	< 30	*		Heart Disease
2		130**	< 30	*		Heart Disease
3		130**	< 30	*		Viral Infection
4		130**	< 30	*		Viral Infection
5		130**	3*	*		Cancer
6		130**	3*	*		Cancer
7		130**	3*	*		Cancer
8		130**	3*	*		Cancer

Figure 1: A 4-anonymous table (taken from [KG06]), with the quasi-identifier {*Zip Code, Age, Nationality*} and the sensitive information *Condition*. This table has two 4-buckets.

The notion k-anonymity has several drawbacks [Swe02; KG06; LLV07]. It does for example not guarantee diversity of the sensitive attribute value in an equivalence class as in the last equivalence class of Figure 1. Additionally, a composition of two k-anonymous tables may compromise the privacy of an individual [GKS08].

2.2 The Network Enabled Surveillance and Tracking System

This work is based on the Network Enabled Surveillance and Tracking (NEST) system [MRV10; VKB12]. The NEST camera system is a semi-automatic surveillance system. It can track individuals over multiple rooms and cameras, as well as support operators at evaluating situations. The NEST system can process incoming video data and extract features like an individuals height, hair color, sex, color information, and movement vector. These features are stored in a feature database and are fed to a model that is presented to the operator. After the feature extraction, the NEST system deletes the video data save for a very short buffer in order to protect the privacy of the surveyed individuals. It can alarm an operator about the occurrence of previously defined situations. Therefore, it reduces on the one hand the amount of data stored. On the other hand, it supports the operator by reducing the amount of data needed to be screened.

2.3 Legal Aspects of Video Surveillance in Germany

For the system proposed in this paper, there are requirements from German privacy regulations. They state that unless there is a legal basis or the individual involved consents, data related to it may not be collected, processed, stored or used otherwise. Video surveillance is treated as collecting personal data (§4 BDSG [Bun78]).

Video surveillance of public space is allowed, however, in order to enforce domestic au-

thority or other eligible interests with prior defined purposes [Zil07]. Involved individuals have to be informed about the surveillance and collected data has to be deleted after use.

For non public space there is no such regulation. In our scenario, we assume that the individuals have to consent to the video surveillance. In this case, the system proposed in this paper can improve the acceptance, since personal data only leaves the system under prior defined circumstances.

Another legal requirement is that the system may not make autonomous decisions (§6a BDSG [Bun78]). Such an autonomous decision could be alerting paramedics or even releasing private information. Therefore, the system needs an operator.

3 System Model

In this section we give an abstract representation of an assisted living community, from which we derive requirements for the database schema and for anonymization methods.

3.1 Scenario

We consider several real world settings which are easily extensible and can occur in an assisted living community.

- **Visits.** Visitors (both private and official) to the assisted living community have to be recognized.
- **Trespassing.** We assume areas in the assisted living community that are restricted to either staff or residents, e.g. recreation rooms for the staff.
- **Help gestures.** In case of an accident other individuals have the possibility to perform help gestures towards the camera in order to alarm the staff.
- **Accidents.** If no other individual is nearby when an accident (of a resident) happens, the staff shall be alarmed.
- **Emergency.** In case of an emergency, such as a fire, the surveillance system must assist the staff to evacuate every resident.

Due to the complexity of these settings we assume a significant amount of false positives for every setting. Therefore it is necessary to provide an operator with presumed emergencies to decide if action has to be taken. Additionally, as mentioned in Section 2.3, no surveillance system is allowed to act on its own and human interaction (in form of an operator) is necessary.

One way to provide the operator with the necessary information is to just replay the video of the situation which caused the system to evoke an alert. This, however, provides the operator with a lot of unnecessary additional information and even allows her to identify

the individual(s) on the video stream. It is also contrary to the principle of data minimization (§3a BDSG [Bun78]), because we can provide the operator with sufficient information without revealing the complete video stream.

3.2 Agents

We assume several types of individuals that take part in the assisted living scenario and thus are filmed by the video surveillance.

- **Residents.** Individuals in this group live in the assisted living community. Their privacy is paramount, since they are filmed (nearly) at all times.
- **Staff.** This group includes the nurses that take care of the residents as well as other individuals that are working permanently in the assisted living community.
- **Guests.** Here we consider all individuals that are either visiting residents or working only for a short period of time in the assisted living community, such as doctors.
- **Operator.** The operator uses the surveillance system and evaluates every alert of the system. We assume that the operator does not appear in the surveilled area.

Definition 2 *Let $\mathcal{U}$ be the set of all individuals that move inside the assisted living community. Further let* R *be the set of residents,* S *be the set of staff members and* G *the set of guests. We write $\mathcal{U}|_{\mathsf{X}}$ to denote a certain group $\mathsf{X} \in \{\mathsf{R}, \mathsf{S}, \mathsf{G}\}$.*

Considering that the surveillance system has to keep a record for each individual and needs to extract certain features to recognize the individual, we propose (w.r.t. the surveillance system described in Section 2.2) a database scheme $d_U = [\text{PID}, \text{Age}, \text{Size}, \text{Constitution}, \text{Colors}, \text{Role}]$ for this information.

3.3 Rooms

In our scenario, we distinguish between three different types of rooms. By the term room we do not necessarily mean a single room but an area that is covered by a camera and satisfies a certain criterion.

- **Rooms of residents.** This type summarizes all private rooms of the residents as well as recreation rooms and the kitchen.
- **Restricted areas.** Rooms such as recreation rooms for the staff and the pharmacy can be restricted. but also areas outside of the assisted living area can be declared, e.g. if residents with Alzheimer's disease may not leave the grounds.
- **Unobserved areas.** Due to the privacy of the residents, there are rooms that are not monitored, for example the sanitary facilities.

Definition 3 *Let $\mathcal{R}$ be the set of all rooms in the assisted living community.*

Since we have different types of rooms we also need to assign each room access rights. It is also important to know the alignment of the camera in each room (cf. Section 5.2.2), which leads to the relation $d_R = [\text{RID}, \text{Alignment}, \text{Access Rights}]$.

3.4 System View

The surveillance system internally needs to keep track of every individual in the assisted living community. Thus the view of the system $d_{AL} = [\text{PID}, \text{Age}, \text{Size}, \text{Constitution}, \text{Colors}, \text{Role}, \text{RID}]$ is a combination of d_R and d_U. Additionally, the buffered video sequences are kept to enable recognition of accidents or help gestures.

4 Methods for Anonymization of Video Data

It is not possible to completely characterize the extractable data of a video stream, so we will take a different approach. We state features that we want to be hidden from the operator in any case, and from the architecture we derive features that are necessarily visible to let the operator do her job.

In the following, we state the assumptions we need for the anonymized video and give examples for anonymization methods that are assumed to fulfill the assumptions. To make the assumptions explicit, we first need to introduce some notation. Let $X_{u,r}$ describe a video sequence that contains a user $u \in \mathcal{U}$ in a room $r \in \mathcal{R}$. We assume the anonymization mechanism to be deterministic functions f.

Definition 4 *Let f be an anonymization method for video sequences X. We call f a* k-anonymizer *iff for all $r, r' \in \mathcal{R}$ and for all $\mathcal{U}$ with $|\mathcal{U}| > n(f)$ it holds that for all $u \in \mathcal{U}$, there exists $v_i \in \mathcal{U}, 1 \leq i \leq k-1$, such that $f(X_{u,r})$ is indistinguishable from $f(X_{v_i,r'})$.*

By $n(f)$ we denote that depending on the anonymization method it might become necessary to have a larger set of individuals to compare to. The indistinguishability can only be based on empirical data, because no metric for the privacy of video data is known. Note also that the definition of an anonymizer implies that the rooms are pairwise indistinguishable as well.

Corollary 1 *Every k-anonymizer f implies k-anonymity on any feature set that can be extracted from video sequences anonymized with f.*

We propose three different anonymizers.

Map. The anonymizer f_M shows all *guests* of the assisted living community as dots on a map. In principle every individual is indistinguishable, i.e. $k = |\mathcal{U}_\mathsf{G}|$. All residents and the staff are not shown at all, and the guests cannot be identified by the dots (cf. Figure 3).

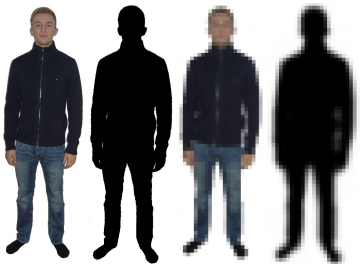
There are two attacks on this anonymization: the movement pattern of the depicted individuals, and associations of guests and residents. Movements patterns for guests are a small concern, and it might be possible to hide guests that regularly visit certain residents.

Avatar. The anonymizer f_A is based on an *idealized* stick figure on white background (cf. Figure 2(b)), which is extracted via gait recognition from the video stream. Apart from movements this method is assumed to remove most features of the originial individual. We assume that there are always atleast 2 individuals that are indistinguishable based on their gestures, i.e. $k = 2$.

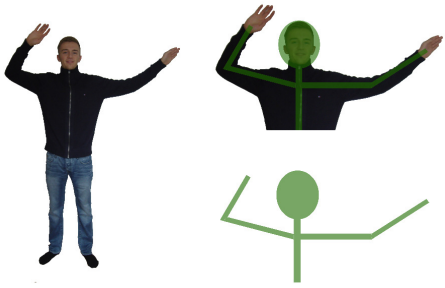
One might assume that gait recognition can be used by an adversary to uniquely identify each individual, similar to [HMH14], who use waving gestures for authentication. [HMH14] need very good cameras and use a non-idealized image, so their approach will probably fail. Additionally, several samples are necessary to train the distinguisher, and an identity has to be associated manually.

Silhouette. We propose to use a combination of silhouette and pixelization/soft-focus as anonymizer f_S to anonymize accidents. The silhouette of the individual is determined, and the environment cut out. The silhouette is filled and the contours are pixelated/put in soft-focus. This removes many of the body features, while it is still possible to observe an accident (cf. Figure 2(a)). To obtain the required anonymity the degree of pixelization can be adjusted. Still, Definition 4 requires that at least two individuals are indistinguishable given the anonymized video sequence.

Pose-estimation or gait recognition will not work on the anonymized video sequence, because people have to be positioned properly in front of a high resolution camera. Certain specific properties of elderly persons like walking aids have to be either anonymized, or another person must also have such a walking aid. This is the assumption we have to make, which in general implies that the amount of residents has to be fairly large.



(a) Plain picture with silhouette, soft-focus effect, and a combination of both anonymization methods applied to it.



(b) A stick figure replacing an individual that forms a help gesture with its arms.

Figure 2: Two anonymization methods for video data.

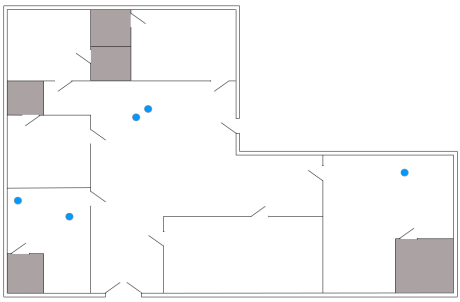


Figure 3: Example map with five individuals on it. Grey highlighted areas are unsupervised by camera. In our system, the maximal resolution for the position of agents are individual rooms.

5 Privacy Compliant Video Surveillance

Based on the model presented in Section 3 and the anonymization methods for video data presented in Section 4, we can now describe the behavior of the system. In Section 5.1, we describe the views the system presents to the operator under different circumstances. Based on this description, we provide a security analysis of the system in Section 5.2.

5.1 Behavior of the System

Depending on the situation, the operator only gets the information needed to assess it. The default view the system presents to the operator is the map view without individuals. In order to show accidents after they happened, the video stream is buffered in the video buffer (cf. Section 2.2) for a fixed time. Depending on the situation, the system presents additional information to the operator as follows:

- **Default.** The default view is the map view without any individuals.
- **Visitors.** If visitors are present, the system shows them as points on the map. Information about the position of visitors is necessary in order to prevent trespassing.
- **Trespassing of restricted areas.** If an individual trespasses to a restricted area, the operator needs to know the identity of the individual as well as her position. Therefore the trespassing individual is highlighted in the map view. Additionally the system presents the video stream of the trespassing individual to the operator.
- **Help gesture.** If the system recognizes a help gesture of an individual, it presents the operator the avatar view of the individual. If the operator verifies the help gesture, in order to assess the situation fully, the system presents the buffered video stream of the individual to the operator. If the operator recognizes a falsa alarm, the system switches back to default.

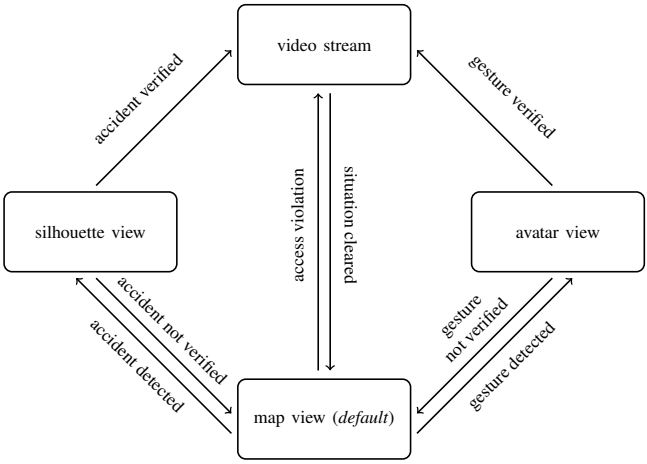


Figure 4: The behavior of the system. Depending on the situation, the system presents different views to the operator.

- **Accident.** If the system recognizes an accident of an individual, it presents the silhouette view of the accident to the operator. If the operator verifies the accident, the system presents the video stream, in order to support the operator assessing the situation fully and to react accordingly. If the operator does not verify the accident, the system switches back to default.
- **Emergencies.** Emergencies such as a fire are recognized by separate detectors and are issued manually. In case of an emergency, the system is informed about it. It then shows all camera streams to the operator.

In order to prevent misuse of false positives, every action of the operator that activates a plain video view is logged, adding a control instance for the operator.

5.2 Security Analysis

We will start our security analysis by first defining the adversary. Then we prove that the composition of views does not yield any more information than the individual views.

5.2.1 Adversary

Figure 5 depicts all possible attacks on the surveillance architecture. The attack on the connection between cameras and the server can be handled with standard techniques like encrypted connections and tamper-resilient cameras. We can also assume that the database server is placed in a secure location with proper access control.

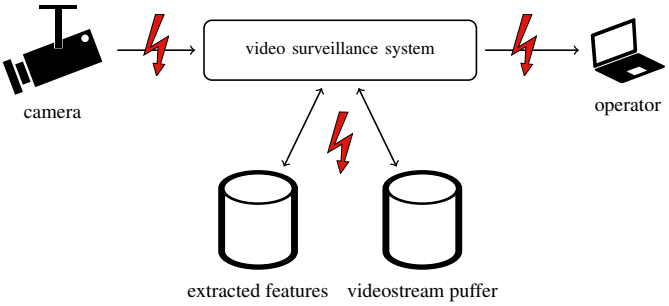


Figure 5: Possible weak points of the surveillance system that an adversary can attack.

The main point of attack that cannot be covered with standard techniques is the system view that the operator gets while supervising the system. She can try to correlate all the views to learn the identity of an individual in the assisted living community. We have to assume that the adversary has complete knowledge about all individuals, including clothes, behaviour and handicaps. Also, we assume that the operator stores everything that is depicted for later use.

5.2.2 Security Analysis

In Section 5.1 we defined four different views that an operator might see depending on the situation in the assisted living community. Obviously, the plain view of the video data directly identifies any individual. If plain video data is shown to the operator, however, either the individual wants to be (help gesture, accident, emergency) or needs to be (trespassing) identified. Thus we do not consider the unanonymized view in our security proof.

The remaining views of the operator are map view, avatar view and silhouette view. This leads to the following composition scenarios: (1) Map view $\leftrightarrow$ avatar view, (2) Avatar view $\leftrightarrow$ silhouette view and (3) Map view $\leftrightarrow$ silhouette view.

We will now prove that each of the above composition scenarios guarantees the privacy of any individual in the assisted living community. Please note, that due to the architecture of the video surveillance system, the operator $\mathcal{A}$ is either provided with the avatar view of a video sequence or a silhouette view but never with both views of the same video sequence.

Theorem 1 *Assuming that f_M, f_A and f_S are $|\mathcal{U}|$-, 2- and $|\mathcal{U}|_{\mathsf{G}}| + 2$-anonymizers respectively, as described in Section 4, an adversarial operator $\mathcal{A}$ cannot uniquely identify any individual in $\mathcal{U}$.*

The above theorem implies that any feature set extracted from the combined anonymized views will be 2-anonymous. This means that at least two individuals in the assisted living community are plausibly depicted in the anonymized view, independent of the information the operator might have gathered.

Proof 1 *Let $\mathcal{U}$ be the set of individuals present with $|\mathcal{U}| = n$, $\mathcal{U}|_{\mathsf{G}}$ be the set of guests present with $|\mathcal{U}|_{\mathsf{G}}| = g$, X_u be an video sequence of individual $u \in \mathcal{U}$, $d_M = f_M(d_{AL})$ be the map view (the RIDs of the guests), $d_A = f_A(X_u)$ be an avatar view of X_u, and $d_S = f_S(X'_{u'})$ be an silhouette view of $X'_{u'}$. Furthermore, let f_M and f_A provide n-anonymity and let f_S provide k-anonymity with $n > k \geq g + 2$. We show that the composition of (1) d_M and d_A, (2) d_A and d_S, and (3) d_M and d_S is 2-anonymous, respectively.*

The first two cases are straightforward: the anonymizer f_A provides 2-anonymity. Since the set of video sequences that are anonymized by f_A is disjunct from all video sequences that are anonymized by other anonymizers (due to the architecture, help gestures are only anonymized by f_A), there is no correlation between d_M and d_A respectively d_A and d_S.

Let $G_{f_S(X_u)}$ be the set of guests that could be represented by the silhouette view $f_S(X_u)$ and let $g' = |G_{f_S(X_u)}|$. Then, the composition of the map view and the silhouette view does not yield any advantage over $(k - g')$-anonymity:

The operator $\mathcal{A}$ might be able to extract the camera alignment from a silhouette view $f_S(X_u)$. In the worst case, all guests are in rooms with a different camera alignment. Since the map view provides the operator $\mathcal{A}$ with this information, he knows that the silhouette view can not be a representation of any guest. In this case, the composition of d_M and d_S provides $(k - g')$-anonymity.

Therefore, and since from $n > k \geq g' + 2$ it follows $k - g' \geq 2$, the surveillance system provides at least 2-anonymity. This concludes the proof.

It turns out that the camera positions are the only factor that contributes to the correlation between map view and silhouette view, so given that all cameras are in the same position for each room, we can give an even stronger guarantee.

Corollary 2 *Assuming that f_M, f_A and f_S are $|\mathcal{U}|$-, 2- and 2-anonymizers respectively, as described in Section 4, and the cameras position is the same for every camera, an adversarial operator $\mathcal{A}$ cannot uniquely identify any individual in $\mathcal{U}$.*

References

[ABN08] Osman Abul, Francesco Bonchi, and Mirco Nanni. "Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases". In: *ICDE*. 2008, pp. 376–385.

[Bou05] T. E. Boult. "PICO: Privacy through Invertible Cryptographic Obscuration". In: *Proceedings of the Computer Vision for Interactive and Intelligent Environment*. CVIIE '05. 2005, pp. 27–38. ISBN: 0-7695-2524-5.

[Bun78] Deutscher Bundestag. *Bundesdatenschutzgesetz*. `http://www.gesetze-im-internet.de/bundesrecht/bdsg_1990/gesamt.pdf`. accessed 06/04/2013. 1978.

[CC11] Janusz Cichowski and Andrzej Czyzewski. “Reversible video stream anonymization for video surveillance systems based on pixels relocation and watermarking”. In: *ICCV Workshops*. 2011, pp. 1971–1977.

[Dwo06] Cynthia Dwork. “Differential Privacy”. In: *ICALP (2)*. 2006, pp. 1–12.

[FS08] Sven Fleck and Wolfgang Straqer. “Smart Camera Based Monitoring System and Its Application to Assisted Living”. In: *Proceedings of the IEEE* 96.10 (2008), pp. 1698–1714.

[GKS08] Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. “Composition Attacks and Auxiliary Information in Data Privacy”. In: *KDD*. März 2008, pp. 265–273.

[HMH14] Eiji Hayashi, Manuel Maas, and Jason I. Hong. “Wave to me: user identification using body lengths and natural gestures”. In: *CHI*. 2014, pp. 3453–3462.

[KG06] Daniel Kifer and Johannes Gehrke. “l-Diversity: Privacy Beyond k-Anonymity”. In: *In ICDE*. 2006, p. 24.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-Closeness: Privacy Beyond k-Anonymity and l-Diversity”. In: *ICDE*. 2007, pp. 106–115.

[MRV10] Jürgen Moßgraber, Frank Reinert, and Hauke Vagts. “An Architecture for a Task-Oriented Surveillance System: A Service- and Event-Based Approach”. In: *Proceedings of the 2010 Fifth International Conference on Systems*. IEEE Computer Society, 2010, pp. 146–151.

[NS08] Arvind Narayanan and Vitaly Shmatikov. “Robust De-anonymization of Large Sparse Datasets”. In: *Proceedings of the 2008 IEEE Symposium on Security and Privacy*. SP ’08. 2008, pp. 111–125. ISBN: 978-0-7695-3168-7.

[Sen+03] Andrew Senior et al. “Blinkering Surveillance: Enabling Video Privacy through Computer Vision”. In: *IBM Research Report* 22886 (2003).

[Sen+08] Andrew W. Senior et al. “Enabling Video Privacy through Computer Vision.” In: *IEEE Security and Privacy* 3.3 (Aug. 18, 2008), pp. 50–57.

[SS98] Pierangela Samarati and Latanya Sweeney. “Generalizing Data to Provide Anonymity when Disclosing Information (Abstract)”. In: *PODS*. 1998, p. 188.

[Swe02] Latanya Sweeney. “k-Anonymity: A Model for Protecting Privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (2002), pp. 557–570.

[VKB12] Hauke Vagts, Erik Krempel, and Jürgen Beyerer. “User-Centric Protection and Privacy in Smart Surveillance Systems”. In: *Future Security*. 2012, pp. 237–248.

[Zil07] Martin Zilkens. “Videoüberwachung - Eine rechtliche Bestandsaufnahme”. In: *Datenschutz und Datensicherheit* 31.4 (2007), pp. 279–283.

Toolgestützte Assessments zu Datenschutz und Datensicherheit in kleinen und mittelständischen Unternehmen

Matthias Rodeck, Caroline Voigt, Arndt Schnütgen
BEL NET GmbH
matthias.rodeck@belnet.de, caroline.voigt@belnet.de,
arndt.schnuetgen@belnet.de

Ina Schiering, René Decker
Ostfalia Hochschule für angewandte Wissenschaften
i.schiering@ostfalia.de, rene.decker@ostfalia.de

Abstract: In kleinen und mittelständischen Unternehmen sind Datenschutz und Datensicherheit eine besondere Herausforderung. Zum einen sind diese Unternehmen geprägt durch dynamische Änderungen von Prozessen und IT-Services, zum anderen ist es notwendig einen Überblick zu erhalten und auf sich ändernde Rechtslagen zu reagieren. Das Management von Datenschutz und Datensicherheit basiert in diesen Unternehmen derzeit meist auf der Fachkunde des Datenschutzbeauftragten unterstützt durch Assessments, die sich auf Fragebögen oder auf Fragebögen basierende Tools stützen. Dieses Vorgehen ist aufwändig und wenig flexibel. Hier wird ein Ansatz vorgestellt, Datenschutzbeauftragte durch eine Kombination aus einem IT-Service zusammen mit einem mobilen Client zu unterstützen, die auf BPMN basierende abstrahierte Darstellungen von Prozessen in den Mittelpunkt von Assessments stellt.

Keywords: Datenschutz, Datensicherheit, BDSG, Assessment, Pattern, BPMN, IT-Service, mobiler Client

1 Einleitung

Um sich der Frage des Datenschutzes in Bezug auf kleine und mittelständische Unternehmen (KMU) zu nähern, ist es zunächst wichtig, die Struktur und Arbeitsweise solcher Unternehmen zu betrachten. Im Vergleich zu großen Unternehmen, ist ein wesentlicher Unterschied die größere Flexibilität. Bezogen auf die Produkte und Dienstleistungen, stellt diese Flexibilität genau den Marktvorteil dieser Unternehmen dar. Sie sind dadurch in der Lage, zeitnah auf Kundenbedürfnisse einzugehen oder kleine Veränderungen an Produkten vorzunehmen, wenn der Markt dahingehenden Bedarf signalisiert.

Die Flexibilität entsteht unter anderem dadurch, dass Entscheidungswege kurz sind. Führungskräfte des Unternehmens sind ansprechbar, Hierarchieebenen sind flach gehalten. Informationen werden aus diesem Grund leichter auch ohne Berechtigung zwischen den Abteilungen und Hierarchieebenen ausgetauscht.

Aus Sicht des Datenschutzes können genau diese Flexibilität und die kurzen Wege problematisch sein: So kann es zu Zugriffen auf personenbezogene Daten, wie z. B. Daten von Mitarbeitern oder Kunden kommen, weil ein Mitglied der Unternehmensleitung diese Daten einsehen möchte oder man sich unter Kollegen unbürokratisch hilft. Da es wenig schriftlich niedergelegte Richtlinien und Prozesse gibt, werden Entscheidungen wie der Einsatz von Videoüberwachung aufgrund von Diebstählen oder die Nutzung von komfortablen Cloud-Diensten wie Dropbox, Google Docs, GMail oder Siri getroffen, ohne die Auswirkung und den rechtlichen Rahmen in Bezug auf personenbezogene Daten zu berücksichtigen.

Ziel dieser Arbeit ist es, eine Tool-Unterstützung für Assessments im Bereich Datenschutz und Datensicherheit vorzustellen, die besonders die hier dargestellte Situation in KMU im Fokus hat. Bisher werden meist Fragebögen verwendet oder Tools, die im wesentlichen auf Fragebögen basieren. Ein solches Vorgehen ist wichtig, um eine möglichst vollständige Darstellung zu bieten. Besonders erfahrene Datenschutzbeauftragte weichen in der Praxis von diesen relativ starren Verfahren ab und entwickeln persönliche Vorgehensweisen.

Hier wird der Ansatz der Patterns vorgestellt. Sie bilden typische Ausprägungen von Verfahren für die Verarbeitung personenbezogener Daten und werden dargestellt als BPMN (Business Process Management Notation) Diagramme. Diese Patterns werden zusätzlich um Hinweise auf mögliche Schwachstellen innerhalb der einzelnen Verfahren ergänzt. Dabei wird versucht einen Kompromiss zu finden, zwischen Übersichtlichkeit, Transparenz für das Gespräch mit Ansprechpartnern und ausreichender Tiefe, so dass das Tool als roter Faden für Gespräche genutzt werden kann. Wichtiges Ziel dieses Vorgehens ist es, den Datenschutzbeauftragten mit Fachkunde in seiner Arbeit zu unterstützen. Zu dem Ergebnis des Assessments können anschließend innerhalb des Backends des IT-Services Maßnahmen zugeordnet werden. Unterstützt wird dies durch Hinweise zu Maßnahmen, die bereits in ähnlichen Situationen verwendet wurden. Damit ist dieser Ansatz geeignet qualitätsgesichertes Arbeiten zwischen mehreren Datenschutzbeauftragten zu ermöglichen und unterstützt besonders Personen mit z.B. technischem und juristischen Hintergrund bei der Zusammenarbeit. Der mobile Client und das Backend wurden als IT-Service innerhalb des Projekts „Datenschutz-Cloud“ entwickelt und werden derzeit bei einem Unternehmen aus dem Projektkonsortium getestet.

In Kapitel 2 werden kurz rechtliche Rahmenbedingungen für den Datenschutz in Deutschland und vorhandene Tools, sowie Vorgehensweisen für Datenschutz und Datensicherheit dargestellt. Im folgenden Kapitel 3 wird der Ansatz erläutert, typische Ausprägungen von Verfahren durch Patterns darzustellen. Anschließend wird in Kapitel 4 das Vorgehen bei einem Assessment mithilfe der Datenschutz-Cloud vorgestellt. Dann wird erläutert, wie die Sammlung von Patterns erstellt und weiterentwickelt werden kann (Kapitel 5). Außerdem wird die Architektur der Datenschutz Cloud dargestellt (Kapitel 6). Abschließend wird der hier vorgestellte Ansatz bewertet und ein Ausblick gegeben.

2 Hintergrund

Auch für kleine und mittelständische Unternehmen ist die Verarbeitung großer Datenmengen heutzutage Teil des Unternehmensalltags. Neben der Schaffung innerbetrieblicher Organisationsstrukturen zur effizienten und bedarfsgerechten Verarbeitung und Speicherung von Informationen, sehen sich die Unternehmen auch einer Reihe von rechtlichen Vorgaben gegenübergestellt. Im vorliegenden Kontext betrifft dies die Verarbeitung personenbezogener Daten.

Unter personenbezogenen Daten versteht man Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person, vgl. § 3 Abs. 1 Bundesdatenschutzgesetz (BDSG) [BDS]. Darunter fallen Informationen wie Name, Adresse, Kontonummer, Telefonnummer oder auch die IP-Adresse[1]. Einfachgesetzlich geregelt finden sich eine Vielzahl von Vorschriften zum Schutz personenbezogener Daten im BDSG, daneben z.B. auch im Telemediengesetz (TMG) und im Telekommunikationsgesetz (TKG). Für die Evaluation des hier vorgestellten Prototypen eines IT-Services, der Assessments im Bereich des Datenschutzes und der Datensicherheit unterstützt, konzentrieren wir uns auf das BDSG.

Zusammenfassend seien als Grundprinzipien des Datenschutzes basierend auf dem BDSG nach Bizer [Biz07] zu nennen:

- Rechtmäßigkeit der Verarbeitung personenbezogener Daten
- Einwilligung der betroffenen Personen
- Zweckbindung bei der Verarbeitung personenbezogener Daten
- Erforderlichkeit bezogen auf die für den Zweck notwendigen Daten und die Dauer der Speicherung
- Transparenz der Datenverarbeitung aus Sicht der Betroffenen
- Datensicherheit (s.u.)
- Kontrolle der Einhaltung der gesetzlichen Rahmenbedingungen durch Datenschutzbeauftragte und Aufsichtsbehörden

Besonders bzgl. der Datensicherheit werden in Anlage 1 zu § 9 BDSG Anforderungen für die Bereiche Zutrittsschutz, Zugangsschutz, Zugriffsschutz, Schutz bei der Übermittlung, Protokollierung, Kontrolle des Auftragsdatenverarbeiters, Verfügbarkeit der Daten und das Trennungsgebot formuliert.

Gemäß § 4d, § 4e des BDSG muss die Verarbeitung personenbezogener Daten dokumentiert werden. Dieses geschieht üblicherweise in einem Verfahrensverzeichnis, dass mindestens die Angaben nach § 4e enthält.

[1] Strittig, zum Streitstand siehe [LT12]

Bei der Realisierung des Datenschutzes in diesem dynamischen Umfeld werden als Basis derzeit häufig Fragebögen eingesetzt, wie z. B. der Baustein Datenschutz im IT-Grundschutz des Bundesamtes für Sicherheit in der Informationstechnik (BSI) [Bun13]. Wobei dabei ein vollständiges Informationssicherheits-Managementsystem (ISMS) im Vordergrund steht, das die Anforderungen gemäß ISO/IEC 27001 [ISO13] umfasst. Zu diesem sehr umfangreichen Vorgehensmodell gibt es die Variante ISIS12 [Gru13], die besonders kleine und mittelständische Unternehmen im Fokus hat. Sowohl für IT-Grundschutz, als auch für ISIS12, kann als Tool-Unterstützung zur Modellierung des Informationsverbunds und zum Ausfüllen der Fragebögen das Tool *verinice*[2] verwendet werden.

Des Weiteren werden fokussiert auf den Bereich des Datenschutzes Tools wie z. B. *2B Secure*[3] oder *privacyGUARD*[4] eingesetzt, die ebenfalls im Wesentlichen auf dem Ausfüllen von Fragebögen basieren.

Ein weiteres Beispiel für den Einsatz von Werkzeugen für die Beurteilung von Datenschutz und Datensicherheit ist das Tool CARiSMA [HWP+13]. Hier liegt der Fokus auf der Prüfung von Geschäftsprozessen in Cloud Services basierend auf dem Risikomodell der IT-Grundschutz Kataloge unter Verwendung von Ontologien. Alle diese Vorgehensweisen haben eine möglichst vollständige Abbildung des Bereichs im Fokus. Einen etwas anderen Ansatz verfolgen Privacy Impact Assessments (PIA), siehe z.B. [Off09], [fISB11] oder [WDH12]: Bei PIAs wird nicht abschließend ein fertiggestellter IT-Service überprüft oder ein Status erhoben, wie bei den bisher vorgestellten Ansätzen, sondern ein Projekt ab der Entstehung eingebettet in das Risikomanagement begleitet.

3 Patterns zu Datenschutz und Datensicherheit

Alle in Kapitel 2 vorgestellten Ansätze zu Assessments im Bereich Datenschutz und Datensicherheit für kleine und mittlere Unternehmen haben die Vollständigkeit der Darstellung im Fokus. Damit stellen sie gute Leitlinien für Personen mit wenig Erfahrung dar und ermöglichen zudem Vergleiche mit anderen Organisationen bzw. innerhalb der Organisation zu verschiedenen Zeitpunkten.

Der hier vorgestellten Ansatz hat zum Ziel, eine Darstellung zu finden, die dauerhaft als roter Faden in den Gesprächen des Assessments genutzt werden kann und dabei Übersicht und Transparenz fördert. Dazu wurde die Idee der Patterns genutzt. Dabei geht es darum, wiederverwendbare Lösungen für wiederkehrende Probleme zu finden. Dieses Konzept wurde zuerst von Alexander [AIS77] im Bereich der Architektur vorgeschlagen. In der Informatik wurden in vielen Bereichen Patterns entwickelt, z.B. Software Design Patterns von Gamma et al. [GHJV94], Security Patterns [YWM08] und Privacy Patterns [DG12].

Bezogen auf Assessments im Bereich Datenschutz, stellen die Verfahren wiederkehrende Punkte dar. Besonders bei kleinen und mittelständischen Unternehmen kommen zum größten Teil wiederkehrende Verfahren im Umgang mit Mitarbeiter- bzw. Kundendaten

[2] http://www.verinice.org/
[3] http://www.2b-advice.com/
[4] http://www.privacyguard.de/

vor, wie z.B. die Zeiterfassung, das Führen einer Personalakte oder der Umgang mit Bewerbungen. Zu den Workflows, die diesen Verfahren zugrunde liegen, werden auf BPMN (Business Process Management Notation) [Bus11] basierende, abstrahierte Darstellungen erarbeitet, welche die für die Beurteilung des Datenschutzes wichtigen Aspekte darstellen. Diese Modellierungen werden als *Patterns* bezeichnet. Zu jedem Verfahren kann es mehrere Ausprägungen geben, die durch Patterns beschrieben werden, wie z.B. Zeiterfassung realisiert durch elektronische Geräte oder papierbasierte Verfahren. Zusätzlich zu der Darstellung des Workflows werden Hinweise auf potentielle Schwachstellen ergänzt. Diese werden als Ellipsen dargestellt, die durch gestrichelte Linien mit Aktivitäten des Workflows verknüpft werden. Dabei wird nicht der volle Umfang der BPMN verwendet. Lediglich Start- und Endereignis, Aktivitäten, Gateways und ein Sequenzfluss zwischen den Elementen sind möglich.

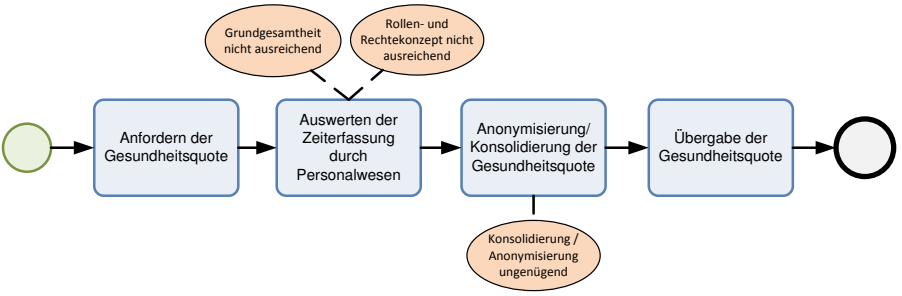


Abbildung 1: Controlling der Gesundheitsquote

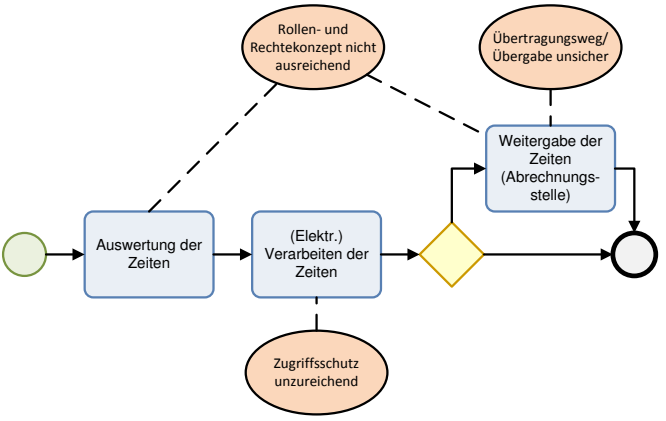


Abbildung 2: Zeiterfassung - elektronische Erfassung der Arbeitszeiten von Arbeitnehmern

Es ist wichtig zu beachten, dass es sowohl zu einer Aktivität mehrere potentielle Schwachstellen geben kann, als auch eine Schwachstelle, wie z.B. das unzureichende Rollen- und Rechtekonzept in Abbildung 2 bei mehreren Aktivitäten vorkommen kann.

Dadurch können Verfahren in Form von Patterns abgebildet werden. Weiterhin ist es notwendig im Bereich der technisch-organisatorischen Maßnahmen eine Reihe von typischen potentiellen Schwachstellen zu prüfen, die nicht direkt mit einzelnen Verfahren in Verbindung stehen. Zum Beispiel werden in Abbildung 3 potentielle Schwachstellen im Zusammenhang mit dem Zutritt zum Gelände, Gebäude oder kritischen Bereichen einer Organisation dargestellt.

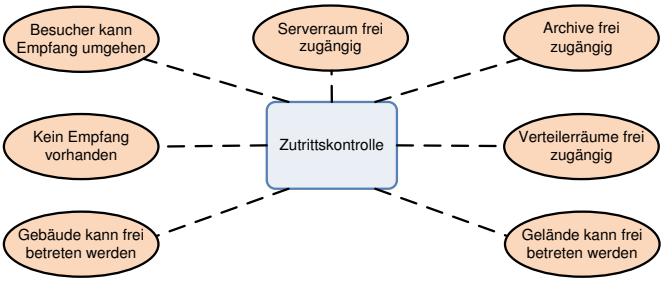


Abbildung 3: Potentielle Schwachstellen im Bereich Zutritt

Andere Bereiche mit ähnlichen Anforderungen sind z.B. Sicherung von Räumen, Verfügbarkeit, Zugriffsschutz. Hierbei wurde eine an BPMN-Diagramme angelehnte Darstellung gewählt, bei der das Thema in Form einer Aktivität in der Mitte steht und von potentiellen Schwachstellen umgeben wird. Gibt es mehr Schwachstellen, als in dieser Art darstellbar, werden die Schwachstellen thematisch gruppiert und auf mehrere Patterns verteilt.

Die Sammlung aller Patterns wird als *Patternpool* bezeichnet. Im Patternpool werden zusätzlich Schwachstellen und Maßnahmen bezogen auf Schwachstellen unabhängig von Patterns gespeichert, so dass eine Schwachstelle einheitlich in mehreren Patterns verwendet werden kann. In den Abbildungen 1 und 2 ist das z.B. „Rollen- und Rechtekonzept nicht ausreichend“. Bei Maßnahmen wird im Patternpool gespeichert, bei welchen Schwachstellen sie typischerweise verwendet werden.

4 Toolgestützte Assessments

4.1 Assessment auf Basis von Patterns mit mobilem Endgerät

Um Assessments durchzuführen, wird der gesamte Patternpool geordnet nach Ansprechpartnern, wie Geschäftsführung, Personalabteilung oder IT-Leitung und nach Bereichen (z.B. Verfahren) auf ein mobiles Engerät, beispielsweise ein Tablet, übertragen. Eine Übersicht zu der gesamten Architektur des IT-Services „Datenschutz-Cloud“finden Sie in Abbildung 7. Der mobile Client zur Unterstützung von Assessments wird dort auch als Frontend bezeichnet.

Nach Auswahl des Ansprechpartners, des Bereichs und eines konkreten Patterns, kann

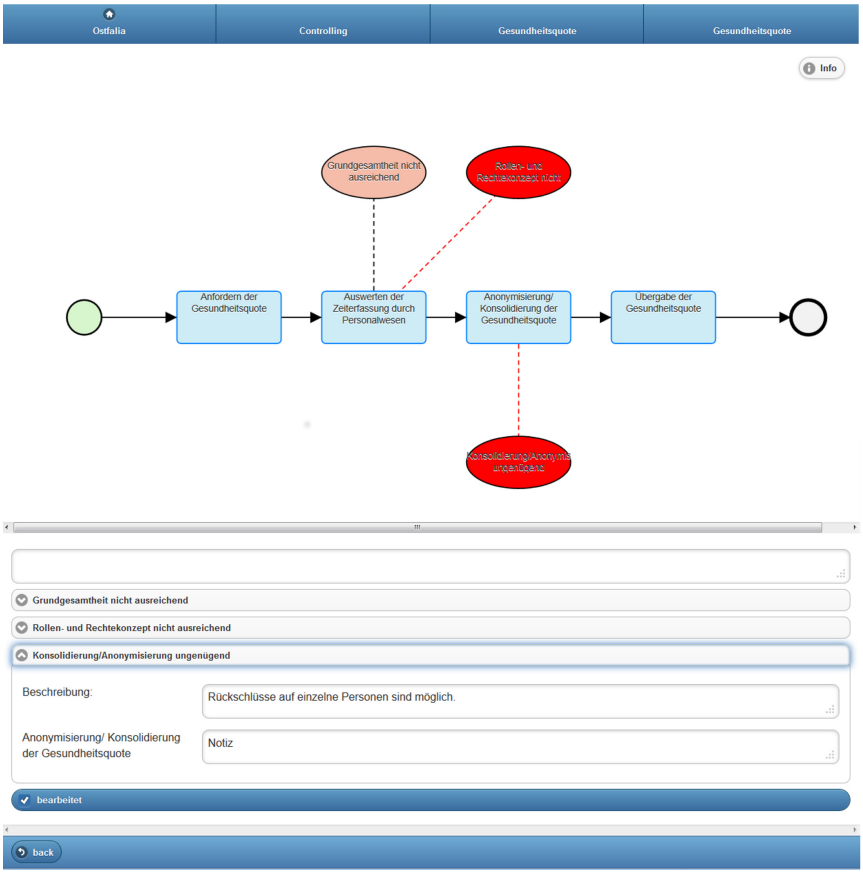


Abbildung 4: Pattern zum Verfahren „Gesundheitsquote“

man innerhalb des Patterns Schwachstellen in Verbindung mit Aktivitäten als zutreffend markieren. Diese werden nun rot dargestellt und man kann sowohl zum Pattern allgemein, als auch zu jeder Schwachstelle verbunden mit einer Aktivität Notizen in einem separaten Notizfeld machen (siehe Abb. 4). Der mobile Client gibt den Gesprächspartnern einen Überblick, welche Bereiche bereits bearbeitet wurden und welche Bereiche Schwachstellen enthalten und bietet damit einen roten Faden im Gespräch in Verbindung mit einer visuellen Unterstützung an, um Assessments vergleichbar durchzuführen.

4.2 Ergänzen von Maßnahmen zu Schwachstellen im Backend

Das Ergebnis des Assessments stellt den Ist-Zustand des Unternehmens aus datenschutzrechtlicher Sicht dar. Es besteht aus den verwendeten Patterns, die Verfahren im Unternehmen beschreiben oder bei denen technisch-organisatorische Maßnahmen geprüft wurden.

Zusätzlich wird gespeichert, wo Schwachstellen bei einem Pattern und einer Aktivität im Pattern im Assessment identifiziert wurden. Sowohl zum Pattern, als auch einem konkreten Vorkommen einer Schwachstelle können ergänzende Notizen erfasst werden.

Da Schwachstellen, die mehrfach in Patterns vorkommen im Backend durch einheitliche Elemente repräsentiert werden, kann das Ergebnis im Überblick als ungerichteter Graph dargestellt werden, mit Knoten die Patterns oder Schwachstellen darstellen. Damit ermöglicht man dem Datenschutzbeauftragten einen Überblick über den Ist-Zustand.

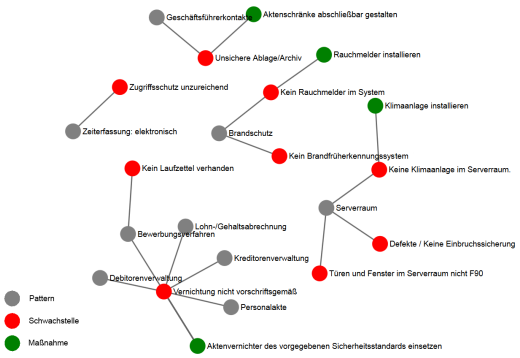


Abbildung 5: Strukturgraph - Ergebnis eines Assessments mit zugeordneten Maßnahmen

Über einen Web-Client im Backend des IT-Service können zu Schwachstellen Maßnahmen zugeordnet werden. Das Ergebnis des Assessments inklusive der zugeordneten Maßnahmen wird als *Strukturgraph* bezeichnet (siehe Abbildung 5). Dabei werden bei der Zuordnung von Maßnahmen Vorschläge angezeigt. Diese Vorschläge basieren darauf, dass bereits im Patternpool Maßnahmen angelegt und mit Schwachstellen verknüpft werden können. Die Reihenfolge der Anzeige basiert auf der Auswertung bereits abgeschlossener Strukturgraphen.

4.3 Regelmäßige Aktualisierung des Assessments

Durch die Flexibilität gerade von kleinen und mittleren Unternehmen, ist es notwendig in abgestimmten Intervallen das Assessment zu aktualisieren. Bei Assessments, die auf Fragebögen basieren, ist dieses aufwändig, da geprüft werden muss, auf welche der Fragen Änderungen im Unternehmen Auswirkungen haben.

Bei dem hier vorgestellten Ansatz von Assessments basierend auf Patterns, ist bereits ein Strukturgraph als Ergebnis eines vorangehenden Assessments vorhanden. Dieses Resultat des ersten Assessments oder der letzten Aktualisierung kann als Basis zusammen mit einer aktuellen Version des Patternpools auf den mobilen Client geladen werden. Nun kann ausgehend von den Patterns mit den Ansprechpartnern geklärt werden, in welchen Bereichen sich Änderungen oder Ergänzungen ergeben. Auch beim anschließenden Zuordnen von Maßnahmen kann auf den letzten Strukturgraphen zurückgegriffen werden, um einen Überblick über Änderungen zu gewinnen und Maßnahmen anzupassen.

5 Erstellen von Patterns im Patternpool

Der Patternpool kann über einen Web-Client im Backend ergänzt werden. Dadurch ist eine flexible Anpassung an neue Technologien oder neue Verfahren jederzeit möglich. Dazu können Schwachstellen und Maßnahmen direkt über einen Web-Client eingegeben und verknüpft werden. Die Erstellung von Patterns ist über einen Patterneditor möglich. Bei der Zuordnung von Schwachstellen kann ausschließlich auf im Patternpool bereits hinterlegte Schwachstellen zurückgegriffen werden. In Abbildung 6 wird der Patterneditor dargestellt.

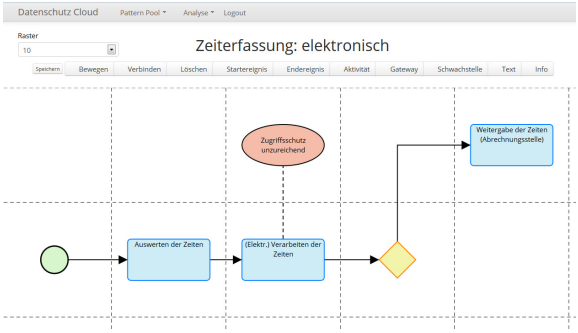


Abbildung 6: Erstellen von Patterns mit Patterneditor

6 Architektur des Gesamtsystems

Die Durchführung von Assessments und die Erstellung von Patterns werden wie folgt im Zusammenhang der gesamten Architektur realisiert: Zunächst kann mittels eines Web-Clients im Backend der Patternpool um Schwachstellen, Maßnahmen und Patterns ergänzt werden. Zur Durchführung von Assessments muss zunächst für das zu untersuchende Unternehmen ein Mandant im Backend erzeugt werden, falls das noch nicht geschehen ist. Dann wird der Patternpool auf den mobilen Client zur Durchführung des Assessments geladen. Anschließend an das Assessment, werden die Ergebnisse verknüpft an den Mandanten ins Backend geladen. Der mobile Client ist als HTML5-App realisiert, um einen plattformübergreifenden Einsatz zu ermöglichen.

Im Backend können zum Strukturgraphen Maßnahmen hinzugefügt werden. Um die Informationen bzgl. der Zuordnung von Maßnahmen als Hinweis für weitere Assessments nutzen zu können, werden aus dem Strukturgraphen sogenannte *Analysegraphen* extrahiert, die möglichst weitgehende Informationen zu der Situation erlauben sollen, in der eine Maßnahme angewendet wurde, ohne Rückschlüsse auf das Unternehmens zu ermöglichen, zu dem der Strukturgraph erhoben wurde.

Dazu werden aus dem Strukturgraphen Teilgraphen isoliert, die jeweils aus einer Schwachstelle mit allen Patterns und Aktivitäten gehören, in denen die Schwachstelle auftrat. Ein

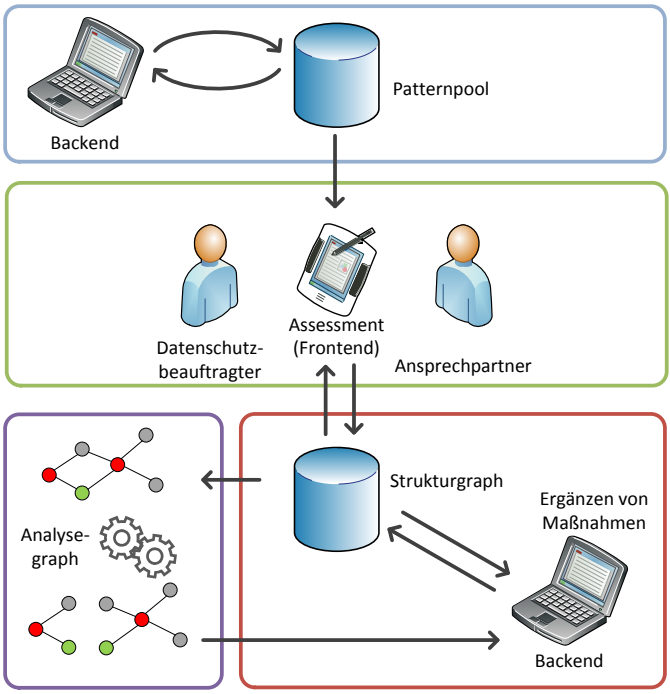


Abbildung 7: Übersicht IT Service Datenschutz-Cloud

Tupel bestehend aus Pattern, Aktivität, Schwachstelle und zugeordneter Maßnahme wird als *Kante* im Strukturgraph oder Analysegraph bezeichnet. Damit besteht ein Analysegraph aus einer Schwachstelle zusammen mit allen Kanten eines Strukturgraphen, in denen die Schwachstelle vorkommt.

Bei der Zuordnung von Maßnahmen zu Schwachstellen in einem neuen Strukturgraphen, werden nun die prinzipiell möglichen Maßnahmen gemäß eines Vergleichs der Umgebung der Schwachstelle im Strukturgraphen mit den Analysegraphen sortiert dargestellt. Dazu wurde ein Ähnlichkeitsmaß auf Analysegraphen definiert und ein darauf basierendes Ranking abgeleitet. Da Analysegraphen keinen Rückschluss auf andere Mandanten ermöglichen, können für das Ranking alle vorhandenen Analysegraphen verwendet werden. Faktoren die dabei eingehen, sind die Ähnlichkeit eines Analysegraphen mit der betrachteten Situation und das Entstehungsdatum des Analysegraphen. Bei der Aktualisierung von bestehenden Assessments, wird nun neben dem aktuellen Patternpool auch der bestehende Strukturgraph auf den mobilen Client geladen. Als Ergebnis wird jeweils ein aktueller Strukturgraph erzeugt, der zur Ergänzung von Maßnahmen und zur Archivierung ins Backend geladen wird.

7 Status und Abgrenzung

Der IT-Service, der im Projekt Datenschutz-Cloud entwickelt wurde, hat derzeit den Status eines Prototypen. Dieser Prototyp wurde von den Autoren anhand von abgeschlossenen Assessments evaluiert. Außerdem wurde besonders der mobile Client mehrfach in der Praxis eingesetzt und hat sich dort bereits bewährt. Dabei hat sich gezeigt, dass die gewählte Vorgehensweise die Zusammenarbeit und den Wissensaufbau zwischen Datenschutzbeauftragten mit stärker juristisch und technisch ausgeprägten Kompetenzen und damit qualitätsgesichertes Arbeiten in Teams fördert.

Der Aufbau der Patterns im Patternpool erfolgt durch den Industriepartner im Projekt nach Bedarf, um möglichst jeweils eine zeitnahe Erprobung von Patterns zu ermöglichen. Dabei wurden bisher Unternehmen untersucht, deren Geschäftsbeziehungen durch B2B (Business-to-Business) Geschäftsmodelle geprägt sind. Deshalb wurde bisher noch keine ausführliche Modellierung von Betroffenheitsrechten vorgenommen.

Aspekte, die besonders im Umgang mit personenbezogenen Daten von Kunden bei B2C (Business-to-Customer) kritisch sind und in dem hier vorgestellten Ansatz noch evaluiert werden müssen, sind z.B. die Zweckbindung mit verschiedenen Zugriffsrechten, Protokollierung auch von lesenden Zugriffen und die Klärung des Zugriffs auf Logs und Protokoll-Dateien. Hier lassen sich die Anforderungen der Betroffenen gut mithilfe von Datenschutz-Schutzzielen modellieren [BM12], [Pro12]. Dabei sollte im Weiteren evaluiert werden, wie sich Konflikte zwischen den Schutzzielen, wie z.B. zwischen Transparenz und Vertraulichkeit bei Zugriffen auf Logs und Protokolldateien, mittels Patterns modellieren lassen. Außerdem muss geklärt werden, inwieweit Schwachstellen, die sich aus der Anwendung von Maßnahmen ergeben, in diesem Ansatz darstellbar sind.

8 Zusammenfassung und Ausblick

Durch die Verwendung von Patterns als Basis von Assessments wird transparentes und qualitätsgesichertes Arbeiten im Bereich des Datenschutzes erleichtert. Datenschutzbeauftragte erhalten eine Unterstützung durch einen IT-Service, der auf ihrer bestehenden Fachkunde aufbaut.

Im nächsten Schritt soll der Patternpool weiter ausgebaut werden, um den Bereich B2C und branchenspezifische Verfahren stärker abzudecken und die in Kapitel 7 adressierten Punkte zu evaluieren. Außerdem sollen in einer Trainigsphase weitere Assessments durchgeführt werden, um die Struktur der Analysegraphen und das darauf basierende Ranking von Maßnahmen zu optimieren.

Danksagung: Dieses Vorhaben wird aus den Mitteln des Bundeswirtschaftsministeriums im Rahmen des Zentralen Innovationsprogramms Mittelstand (ZIM) unter den Förderkennzeichen KF3081801KM2 und KF2842903KM2 gefördert.

Literatur

[AIS77] Christopher Alexander, Sara Ishikawa und Murray Silverstein. A Pattern Language: Towns, Buildings, Construction (Center for Environmental Structure Series). 1977.

[BDS] Bundesdatenschutzgesetz in der Fassung der Bekanntmachung vom 14. Januar 2003 (BGBl. I S. 66), das zuletzt durch Artikel 1 des Gesetzes vom 14. August 2009 (BGBl. I S. 2814) geändert worden ist. http://www.gesetze-im-internet.de/bdsg_1990/BJNR029550990.html.

[Biz07] Johann Bizer. Sieben Goldene Regeln des Datenschutzes. *Datenschutz und Datensicherheit-DuD*, 31(5):350–356, 2007.

[BM12] Kirsten Bock und Sebastian Meissner. Datenschutz-Schutzziele im Recht. *Datenschutz und Datensicherheit-DuD*, 36(6):425–431, 2012.

[Bun13] Bundesamt für Sicherheit in der Informationstechnik (BSI). IT-Grundschutz-Kataloge, 2013. https://www.bsi.bund.de/DE/Themen/ITGrundschutz/ITGrundschutzKataloge/Inhalt/_content/kataloge.html.

[Bus11] Business Process Model and Notation (BPMN). Version 2.0. *Object Management Group specification*, 2011.

[DG12] Nick Doty und Mohit Gupta. Privacy Patterns, 2012. http://privacypatterns.org/.

[fISB11] Federal Office for Information Security (BSI). Privacy Impact Assessment Guideline for RFID Applications, 2011.

[GHJV94] Erich Gamma, Richard Helm, Ralph Johnson und John Vlissides. *Design patterns: elements of reusable object-oriented software*. Pearson Education, 1994.

[Gru13] Michael Gruber. ISIS12 - Informationssicherheit für mittelständische Unternehmen. In *D-A-C-H Security 2013, Nürnberg*, Seiten 275 – 282. syssec, 2013.

[HWP+13] Thorsten Humberg, Christian Wessel, Daniel Poggenpohl, Sven Wenzel, Thomas Ruhroth und Jan Jürjens. Ontology-Based Analysis of Compliance and Regulatory Requirements of Business Processes. In *Proceedings of the 3rd International Conference on Cloud Computing and Services Science (Closer 2013)*, Seiten 553–561. SciTePress, 2013.

[ISO13] ISO/IEC. ISO/IEC 27001 - Information technology - Security techniques - Information security management systems - Requirements, 2013.

[LT12] Patrick Lundevall und Tommy Tranvik. Was sind personenbezogene Daten? Die Kontroverse um IP-Adressen. *Zeitschrift für Datenschutz (ZD)*, 03004, 2012.

[Off09] Information Commissioners Office. Privacy Impact Assessment Handbook, 2009. http://www.ico.gov.uk/upload/documents/pia_handbook_html_v2/index.html.

[Pro12] Thomas Probst. Generische Schutzmaßnahmen für Datenschutz-Schutzziele. *Datenschutz und Datensicherheit*, 36(6):439–444, 2012.

[WDH12] David Wright und Paul De Hert. *Introduction to privacy impact assessment*. Springer, 2012.

[YWM08] Nobukazu Yoshioka, Hironori Washizaki und Katsuhisa Maruyama. A survey on security patterns. *Progress in Informatics*, 5(5):35–47, 2008.

Second European STAMP Workshop

(ESW 2014)

Second European STAMP Workshop (ESW 2014)[1]

Abstract: The objective of the second European STAMP workshop is to further establish a platform for the exchange of ideas, experiences, discussion, co-operation and dissemination in the application of STAMP and related methods. The workshop aims to bring together researchers and safety practitioners from different areas who are interested in improving their approaches to system safety and who are already trying STPA/CAST to share their experiences and provide direction for future research, and to encourage the use of STAMP techniques in real systems in industry. Furthermore, we would like to provide opportunities for informal group discussions included in the program to discuss the evaluation of the application of STAMP in academic and industry projects.

The main topics of the Workshop include, but are not limited to:

- Experiences using STPA and CAST
- Safety-guided design using STPA
- Accident analyses
- Certification and regulatory issues
- Evaluations and comparisons with traditional techniques
- Risk management and identifying leading indicators
- Using STPA/CAST for software-intensive systems
- Applications to security
- Tools, processes, and other support for analysis and design using STPA and CAST
- Safety management systems
- Management and adoption experiences or challenges

Workshop Chairs: Stefan Wagner (University of Stuttgart), Eckehard Schnieder (TU Braunschweig)

Programme Committee: Jens Braband (Siemens), Tim Kelly (University of York), Nancy Leveson (MIT), Brend Müller (Robert Bosch GmbH), Jörg Müller (Bombardier), Lorena Pelegrin (ILF), Eelco Vriezekolk (University of Twente)

[1] http://www.iste.uni-stuttgart.de/en/se/esw2014.html

An Integrated Hazard Identification Method for Socio-technical Systems based on STPA

Rui WANG, Wei ZHENG, Ci LIANG, Gilles MOTET

LAAS-CNRS
Université de Toulouse
7 avenue du colonel Roche
31077 Toulouse, France
rui.wang0309@gmail.com

National Engineering Research Center of Rail Transportation Operation and Control System
Beijing Jiaotong University
No.3 Shang Yuan Cun, Hai Dian District
100044 Beijing, China
wzheng1@bjtu.edu.cn

CASCO Signal Ltd.
6F Wanfengyihe Building, 12 Jia Yue Tan South Street
100045 Beijing, China
liangci321@126.com

LAAS-CNRS
Université de Toulouse
7 avenue du colonel Roche
31077 Toulouse, France
Gilles.Motet@insa-toulouse.fr

Abstract: The traditional hazard analysis approaches applied to the socio-technical system can not cover the complex organization structures, the interactions between systems and human behaviors, the interrelated factors among sub-systems and the safety culture of specific societies. This paper presents an integrated hazard identification methodology named BFM-STPA(STPA hazard identification Based on Formalization Model) based on formalization model, which can solve the above issues. Firstly, the hierarchical control structure models of the socio-technical system are built with Colored Petri Nets (CPN) due to its strong description ability and executable ability, which may also be regarded as the graphical system specification. Secondly, the hazards can be identified according to the CPN models following a series of guide conditions. Thirdly, the comprehensive contributing factors to the hazards will be found out by tracing the former states with in the reachable graph generated from CPN model. Finally, an integrated hazard log can be derived for further hazard analysis and safety-guided design. In this paper, the above method was applied in the Chinese Train Control System level 3 (CTCS-3).

And the process of hazard identification for the scenario of Temporary Speed Restriction issued was elaborated in detail. Compared with the hazard log generated by Hazard and Operability (HAZOP), the hazard log generated by BFM-STPA covered not only the subsystem failures, but also the deviation of interactions among subsystems from design intent, human errors and socio-technical drawbacks related to the CTCS-3.

Keywords: Railway Transportation; Hazard Identification; Colour Petri Net; STAMP;

1 Background

The train control system is a typical safety-critical system in terms of the high operating speed and thousands of passengers on board. In recent years, the major accidents of trains configured with CTCS still occurred from time to time, like the 4-28 Jiaoji railway accident and 7-23 Yongwen railway accident. According to the investigation reports of the major accidents from Chinese Government, these two railway accidents were assigned as the accidents mainly caused by staff negligence after Temporary Speed Restriction (TSR) order loss or component failures [Ouyang02] [Dong03]. And even for the high speed train whose operation speed is over 300km/h (186mph), the accidents occasionally occurred due to the organizational flaws and human errors. All the personnel related to the above accidents were either subjected to correspondent legal sanctions or given internal punishment. Though fair punishment to people who violates the operation rules is certainly important, merely punishing the relative crews will not solve the existing potential defects in technical aspects and organizational management[Leveson05]. Therefore, it is an emergent task to find out a method, which has the ability to deal with this kind of complex system, to identify the potential hazards hidden in the complex system before the system is designed and put into operations. And the safety requirements obtained from the hazard log are utilized to instruct the safety-guided design, which ensures the safety of the system to a large extent.

In this paper, we proposed an innovative hazard identification method for complex system according to STPA, that is, STPA hazard identification Based on Formalization Model(BFM-STPA).The BFM-STPA combined the STPA hazard identification approach with the formalization method of Colored Petri Nets (CPN) to establish system control structure models, identify hazards, and generate hazard log. The distinct features of this method are listed:

1) Taking the advantage of the excellent description capability of CPN, we established the safety models of CTCS Level 3 (CTCS-3) in hierarchical structure. In this case, the control structure model was more explicit and more apposite to be analyzed as a substitute for a real physical system.

3) A comprehensive hazard log in regard to the TSR issued scenario of CTCS-3 had been generated, including the organizational and technical hazardous states. Moreover, the interactive factors among internal subsystems and external interferences could be identified in this method effectively.

4) We also proposed a formalized approach to conduct the hazard analysis, figuring out the causes leading to hazards. Namely, with the help of the reachable state graph derived from the CPN model, the causal factors of a hazard could be found out.

2 Introduction and Application of BFM-STPA

The ascendency of STAMP is the comprehensive consideration of all kinds of contributing factors to accident. However, the classic STAMP control structure model suffers from the lack of performance on formal description of specific system behaviors, the process models of controllers and the deficiency of data structure. Hence, an optimization method, the BFM-STPA, has been creatively put forward in this paper to supplement the drawbacks of modeling means of the STAMP. This method provides a series of hierarchical control structure models to describe the STS and facilitates analyzing the social and organizational flaws. Specifically, the systematic control structure models are composed of Socio-Technical Control Structure Model (ST-CSM), Operating Scenario Control Structure Model (OS-CSM), and Controller Process Model (CPM). Figure 1 illustrates the multi-layer structure of the three kinds of CPN models. CPN helps to improve the ability of modeling in a large scale. The multi-layer models use the different types of variables (that is the colored variable) to describe the commands and feedback messages between controllers of each level, and utilize the ML language to constrain the interactive conditions.

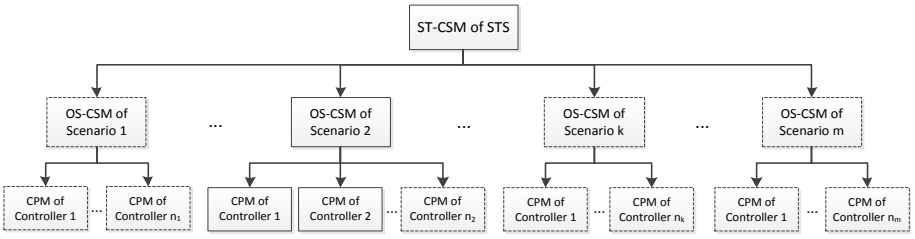


Figure 1: The multi-layer structure of CPN models

The procedure of the BFM-STPA hazard identification method is shown in Figure 2. There are 5 steps to implement this analysis work. The specific steps of performing the hazard identification on STS with BFM-STPA are described as follows:

1) Establishing the three-layer hierarchical control structure models in accordance with the functional requirements and system requirements of the CTCS-3;

2) According to the fundamental hazards and safety constraints on the system level, identifying the potential inadequate control actions that lead to hazardous states with the given guide conditions. The inadequate control actions or the situation of breaking the safety constraints could be caused by the occurrence of the following guide conditions:

F1. Required control actions (for safety) are not provided;

F2. Incorrect or unsafe control commands are given;

F3. Potentially correct control commands are provided at the wrong time (too early or too late), or in wrong sequence;

F4. Control is stopped too soon or applied too long.

3) Generating the SSRG of the CPN model with CPN tools;

4) Based on the SSRG and control structure models, figuring out the ways how the hazardous control actions identified in Step 3) could happen in the following 3 paths.

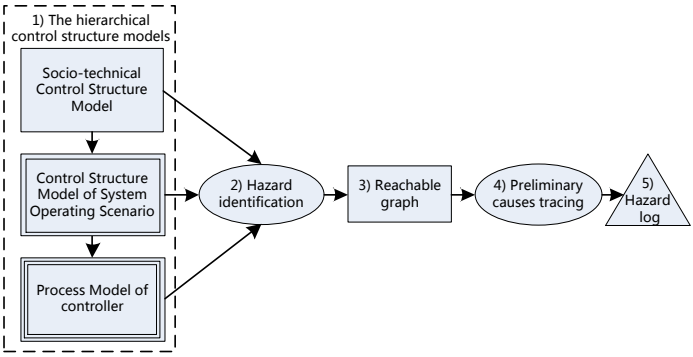


Figure 2: Frame of BFM-STPA hazard identification method

a. Tracing the possible causes that lead to the hazards with SSRG.

Ⅰ. Finding out the reachable state in which the control action carries out, based on the specific tokens in places shown in SSRG;

Ⅱ. Tracing back to the previous states and searching for the causes. Basically, the errors causing the hazards can be categorized into system error and stochastic error. Here, it is more appropriate to divide the human error and the organizational flaw from the system error so as to pay more attentions to these two types of errors. Thus, the system error mainly refers to the dysfunction of software, and the stochastic error is component failure of which the distribution is generally in concordance with the Bathtub curve [Norros06].

b. Identifying the potential causes resulting in unsafe control actions with the hierarchical control structure models. For each hazardous control actions, all parts of the corresponding control loops have to be examined; there may be conflicts and potential coordination problems that need to be identified, when multiple controllers are responsible for the same component or the same safety constraint.

c. Speculating how the control actions could degrade over time, and adding the preventive measures, including:

Ⅰ. Setting up adjustment mechanism of organizational structure to assure implying the safety constraints where some changes appear;

Ⅱ．Implementing the performance audit；

The performance audit is the prerequisite for operating audit and control. Only in this way, can the safety constraints be implied.

Ⅲ．Tracing the abnormal situations found in the accident analyses;

5) Generating the comprehensive hazard log which should be suitable for being maintained in following stages of system life cycle.

Due to the complexity of CTCS-3 system, we apply BFM-STPA methodology to the scenario of temporary speed restriction issued for instance. To analysis the result of this method, we compared the BFM-STPA with the method of HAZOP (Table 1) against the same scenario of TSR Issued in ETCS-2. For there was no published work that conducts the HAZOP on CTCS-3 system and the ETCS-2 system had the exactly same operating mechanism with the CTCS-3, we made this comparison between these two system. BFM-STPA identified 49 hazards in total, including 6 hazards related to socio-technical factor and 9 hazards involving human factor. However, 14 hazards related to this scenario were identified by method of HAZOP[ESROG07] involving only 1 hazards of human factor. Therefore, BFM-STPA hazard identification presented in this paper not only covers socio-technical factors and human factors, but also contains more comprehensive technical factors.

Furthermore, the HAZOP requires great quantity of HAZOP sessions and expertise to build models and identify the hazards. However, the BFM-STPA demands some personal experience to establish the control structure models and less systematic knowledge to identify the inadequate control actions and causal factors due to the enough information contained in the models. And it makes the process of system hazard identification standardized, and reduces the dependence on experience of experts.

Table 1 Result of contrasting BFM-STPA method with HAZOP method

Method	**#Socio-technical Factor**	**#Human Factor**	**#Technical Factor**	**#Hazards in Total**
BFM-STPA	6	9	34	49
HAZOP	0	1	13	14

3 Conclusion

This paper presented the hazard identification method of BFM-STPA mainly for STS and its application. And the incorporation of CPN mode, which helped to substitute the original control structure model in STPA, largely enriched the systematic features that could be modeled before. As we all know, the more details of a system were describe in models, the more valuable hazards with significant guidance to system design could be identified. Besides, though the accident causation model of classic STAMP focused on

the social part of contributing factors that leading to mishaps, the proposed method in this paper involved a large number of technical elements as well. Furthermore, taking the advantage of the CPN tools, the reachable graph could also be utilized to analyze the causes of a certain hazard. During the process of analyzing CTCS-3 in this paper, we obtained the following experience of how to effectively utilize this method.

On one hand, it should be emphasized that the CPN models have to involve as many details of system as possible on the basis of analysis requirement. Hence, the control structure models can become the graphical specification of the system. Although the complexity of models requires more expertise, it is worthy of figuring it out in this stage and decompose immense STS in order to avoid the large-scale workshops of the hazards identification conducted afterwards.

On the other hand, the CPMs should be analyzed with the guide conditions as well. For we find that the guide conditions limit the hazards in the layer of the analyzed control actions, the only possibility to identify the detailed hazards exists in studying more control actions in CPMs, where there are lots of functions and interactive messages.

In addition, this method is highly recommended to be widely applied in the other STS and the future work will be the risk analysis based on the hazards identified by the BFM-SPTA.

References

[Ale04] Ale B. J. M., Bellamy L. J., Cooke R. M., et al., 2006. Towards a causal model for air transport safety—an ongoing research project. Safety Science, 44(8): 657-673.

[ASKCI01] ASKCI, 2013, China's Railway Market Forecast and Industry Research Report (2013-2017). http://www.askci.com.

[Ouyang02] Ouyang M., Hong L., Yu M. H., et al., 2010. STAMP-based analysis on the railway accident and accident spreading: Taking the China–Jiaoji railway accident for example. Safety Science, 48(5): 544-555.

[Dong03] Dong A., 2012, Application of CAST and STPA to railroad safety in China. mit.edu.

[Ale04] Ale B. J. M., Bellamy L. J., Cooke R. M., et al., 2006. Towards a causal model for air transport safety—an ongoing research project. Safety Science, 44(8): 657-673.

[Leveson05] Leveson, N.G., 2011. Engineering a Safer World: Systems Thinking Applied to Safety (Engineering systems). MIT Press, Cambridge, MA.

[Norros06] Norros, L., 1996. System disturbances as springboard for development of operators' expertise. Cognition and communication at work, 159-176.

[ERTMS07]ERTMS Safety Requirements and Objectives Group (ESROG), 2000. ERTMS Scope, Boundary and Hazards, ESROG-SBH-01, Issue 1 Draft 3.

Using STAMP to Develop Leading Indicators

Nancy G. Leveson

Aeronautics and Astronautics Dept.
MIT
77 Massachusetts Ave., Room 33-334
Cambridge, MA 02139
leveson@mit.edu

Abstract: Ths paper describes an approach to using STMP and STPA to derive leading indicators of increasing risk.

There are always warning signs before a major accident, but these signs may only be noticeable or interpretable as a leading indicator in hindsight. Before an accident, such "weak signals" are often perceived only as noise. The problem then becomes how to distinguish the important signals from the noise. Defining effective leading indicators is a way to accomplish this goal by providing specific clues that can be monitored.

A lot of effort has been spent on trying to identify leading indicators, particularly in the petrochemical industry. Much of the past effort has involved finding a set of generally applicable metrics or signals that presage an accident. Examples of such identified leading indicators are quality and backlog of maintenance, inspection, and corrective action; minor incidents such as leaks or spills, equipment failure rates, and so on. Some depend on surveys about employee culture and beliefs, with the underlying assumption that all or most accidents are caused by employee misbehavior, and include as leading indicators such culture aspects as safety awareness, mutual trust, empowerment, and promotion of safety [Am12]. A large number of proposals for leading indicators outside the petrochemical industry focus on occupational safety rather than system safety, and some are simply a listing of potential hazards, such as lack of safety training; whether there is a lock-out, tag-out policy or a stop-work policy; and whether there are medical facilities on site [HSE03]. In fact, the BP Grangemouth Major Incident Investigation Report suggested that industries may have a false sense of safety performance due to their focus on managing personal safety rates rather than process safety[1] [HSE03].

As a result of major accidents in the chemical industry, a concerted and long-term effort has been devoted to identifying leading indicators of risk. Most assume that accidents are caused by component failures and that likelihood of failures should be used to reduce the scope of the search for leading indicators despite the fact that likelihood may often be unknown, and the practice may result in overlooking low likelihood events.

[1] While the term "system safety" is common in most industries, the same thing called "process safety" in the process industries.

Early attempts to develop process safety performance metrics (leading indicators) date from the mid-1900's, but attempts accelerated after the Grangemouth report recommended that "companies should develop key performance indicators for major hazards and ensure that process safety performance is monitored [HSE03]. A series of documents have been issued since that time by the American Institute of Chemical Engineering (AICE), the OECD, the U.K. Health and Safety Executive, the U.S. Occupational Safety and Health Agency (OSHA), and the American Petroleum Institute (API). Most of these standards recommend that the identification of leading indicators start from the hazard analysis, but they assume that accidents are caused by a linear chain of events and do not address indirect interactions and complex systemic factors in accidents [Kh12, Le04]. Almost all recommend that the likelihood of failures should be used to reduce the scope of the search for leading indicators despite the fact that likelihood may often be unknown and the practice may result in overlooking low likelihood events.

There is commonly a belief—or perhaps, hope—that a small number of such "leading indicators" can identify an increase in risk of an accident. While some general indicators may be useful, large amounts of effort over decades has not provided much progress [Kh12]. The lack of progress may be a sign that such general, industry-wide indicators do not exist or will not be particularly effective in identifying increasing risk. An alternative, which is the focus of this paper, is to identify leading indicators that are specific to the system being monitored.

This paper proposes an approach to identifying and monitoring system-specific leading indicators and provides guidance in designing a risk management structure to use such indicators effectively. It extends the suggestions in the series of standards by the chemical industry, basing them on a different, more comprehensive model of accident causation and, perhaps more important, provides a formal foundation for identifying leading indicators.

The approach is based on the STAMP model of accident causation and tools that have been designed to build on that model [Le04. Le12]. The basic proposal is to base leading indicators and risk management on our assumptions about why accidents occur, which is the essence of STAMP.

The idea of assumptions being the basis for identifying leading indicators is not original but has been proposed for more general risk management programs. RAND developed the methodology of assumption-based planning (ABP) primarily to assist U.S. Army clients with mid- and long-term defense planning and to reduce uncertainty and manage risk [De02].

The process involved uses assumptions and their vulnerability as the basis for identifying leading indicators rather than the classic probabilistic risk methods. The STAMP/STPA process for safety-guided design and hazard analysis provides the framework for a structured leading indicator identification process. System hazards are first identified and used to derive the safety constraints and system safety requirements. Hazards are categorized, if necessary, with respect to potential worst-case severity and

vulnerability. The functional safety control structure is designed with safety responsibilities identified for each component, where these control responsibilities are traceable to the system safety constraints. Once the safety control structure is created, STPA is used to identify unsafe control actions and their causes. Elimination of the causes should be first attempted and, if elimination is not possible, to mitigate and control them.

During this process, the information necessary to identify the assumptions being made during design and development is recorded and used to plan operations, to design the data and feedback that must be collected, and to design the overall leading indicator management program. The relative importance of the leading indicators (the consequences of not detecting something) and potential action plans upon their failure is also determinable by the ranking of the severity of the hazards to which they are traceable.

A serious problem that has been well studied in risk assessment is the difficulty in overcoming heuristic biases [e.g., KST82, SFL80]. A structured approach to overcoming common biases that lead to flawed risk assessment can be defined and a process for identifying, monitoring [DFI13] and managing a leading indicators program on this assumption-based approach to risk management can be developed using STPA. Managing such a program should include a process for updating leading indicators as experience is gained and as both the system and its environment change over time.

References

[Am02] American Bureau of Shipping, Safety Culture and Leading Indicators of Safety, January 2012, Houston TX.

[HSE03] The BP Grangemouth Major Incident Investigation Report, UK Health and Safety Executive, August 2003.

[De02] James Dewar, *Assumption-Based Planning*, Cambridge University Press, 2002.

[DFI13] Ioannis M. Dokas, John Feehan, and Syed Imran. EWaSAP: An Early Warning Sign Identification Approach Based on a Systemic Hazard Analysis, Safety Science, Vol. 58, Oct. 2013, pages 11-26.

[KST82] D. Kahneman, P. Slovic, and A. Tversky. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press, 1982

[Kh12] Ibrahim Khawaji, Developing System-Based Leading Indicators for Proactive Risk Management in the Chemical Processing Industry, SM Thesis, MIT, 2012.

[Le04] Nancy Leveson, A New Accident Model for Engineering Safer Systems, *SafetyScience*, 42(4):237-27, 2004.

[Le05] Nancy Leveson, *Engineering a Safer World: Applying Systems Thinking to Safety*, MIT Press, 2012.

[SFL80] P. Slovic, B. Fischhoff, and S. Lichtenstein, Facts and fears: Understanding perceived risk, *Societal Risk Assessment*, 181-216, 1980

Model-based Concept Development and Safety Driven Design

Cody H. Fleming

Department of Aeronautics & Astronautics
Massachusetts Institute of Technology
77 Massachusetts Avenue, 33-407C
Cambridge, MA 02139 USA
chf44@mit.edu

Abstract: This paper outlines a new approach for safety-driven concept development based on the systems-theoretic accident model and process. Compared to traditional approaches, this model captures more types of accident causes prevalent in modern systems. In addition, the model emphasizes functional behavior in addition to physical behavior, which makes it a promising candidate for use during early system development stages. The new approach represents an extension of the accident causality model that allows stakeholders to systematically develop a model of a concept and then query the model to identify potential vulnerabilities and mitigation strategies.

1 Introduction

As aerospace systems become increasingly complex and the roles of human operators and autonomous software continue to evolve, traditional safety-related analytical methods are becoming inadequate. Traditional hazard analysis tools are based on an accident causality model that does not capture many of the complex behaviors found in modern engineered systems. Additionally, these traditional approaches are most effective during the late stages of system development, when detailed design information is available. However, system safety cannot be assured by discovering problems at these late stages and adding patches or expensive updates to the design. Rather, safety must be designed into the system from its very conception. The primary barrier to achieving this objective is the lack of effectiveness of the existing analytical tools during early concept development.

This research proposes a new technique, which is based on a more powerful model of accident causality that can capture behaviors that are prevalent in these complex, software-intensive systems. The proposed approach builds on the accident causality model, called systems-theoretic accident model and process (STAMP), extending the process so that it can be applied systematically during the early, concept development stages of systems engineering.

The goals are to (1) develop rigorous, systematic tools for the analysis of future concepts in order to identify vulnerabilities and undocumented assumptions, and (2) extend these tools to assist stakeholders in the development of concepts using a safety-driven approach. This

work first develops a methodology for hazard analysis of a concept using control theory to generate a model of the concept of operations. Formal, systems-theoretic concepts such as hierarchy, emergence, communication, and coordination are used to analyze the model and identify vulnerabilities in the concept.

This model-based approach represents a significant departure from the state of the art, where a concept is defined, developed, documented, and analyzed according to a control theoretic model rather than free form, natural language text. In this thesis, the power of the proposed approach will be demonstrated on a real concept currently being developed by the United States Federal Aviation Administration, in conjunction with the Single European Sky (SESAR) program.

2 Section heading

This research is different from—represents a contribution to—existing STAMP/STPA approaches and the general systems engineering literature for several reasons. First, systematic and rigorous methods are typically not used in the development and analysis at the conceptual stage. Secondly, model-based systems engineering approaches do not (explicity) consider hazardous behavior or help stakeholders to identify potential safety-related vulnerabilities.

The focus of this research requires further explanation of some of the general characteristics that are often present during concept development. The primary artifact of concept development, the ConOps (see for example [INC11, Kap10, IEE98]), often consists of natural language text and/or low fidelity graphical depictions of work- and information-flows. ConOps should contain some reference to stakeholder goals and system-level requirements, but they rarely contain specific design requirements. Because a ConOps is developed long before the system becomes operational, it typically includes (often implicit) assumptions about the future that are not necessarily true when the document is being developed. For example, a ConOps may contain assumptions about future technologies that do not yet exist. Finally, often in practice a ConOps is developed by committee, with disparate members who have different goals and perspectives [JPD11]. All of these characteristics make it difficult to systematically analyze a Concept of Operations and in particular to rigorously, systematically:

1. identify *missing information* or *undocumented assumptions* that will be required for effective operation of the system;
2. identify *inconsistent* or *conflicting* information within a concept that may lead to hazardous behavior;
3. identify where *more specific* operational concepts are required to understand safety- and functionally-related behavior of the system;

Model-Based System Engineering methods and tools have been proposed in order to ensure consistency in requirements and testing efforts, develop specifications, define inter-

face requirements, develop test plans, and intellectually manage complexity. Elements of an effective model-based system engineering model include clear and unambiguous language, behavior expressed in relationships that represent the "structure" of the system, a system representation built from the language and structure, and the ability to express the representation from different views [LZ11].

Model-Based System Engineering methods represent the functional relationships among system components – the "structure" – with a generic set of guide words. For example, in Systems Modeling Language (SysML), structural relationships include Sequence, Select, Multiple-Exit, Iterate, Loop, Concurrency, or Replicate [Del13]. There are many other languages (e.g. Object-Process Methodology [Dor02, Sod02]) and representations (e.g. Design Structure Matrix [EB12]) used to represent entities, their attributes, and the relationships between them [LZ11].

Rather than directly using the control flaws and causal factors from STPA Step 2 (see [Lev12], page 223), this approach first examines the most *basic* functions of each entity in the control loop. That is, what is required of each entity in the control loop for effective, safe system behavior? What are the responsibilities of the controller, actuator, controlled process, and sensor? How do these entities interact with each other, with the environment, and with other control loops? For the purposes of meeting document submission requirements, consider just two of the four basic elements in a control loop.

The Controller:

- Creates, generates, or modifies control actions based on algorithm or procedure and perceived model of system
- Processes inputs from sensors to form and update process model
- Processes inputs from external sources to form and update process model
- Transmits instructions or status to other controllers

The Sensor:

- Transmits continuous dynamic state measurements to controller (i.e. measures the behavior of controlled process via continuous or semi-continuous [digital] data)
- Transmits binary or discretized state data to controller (i.e. measures behavior of process relative to thresholds; has algorithm built-in but no control authority)
- Synthesizes and integrates measurement data

This information can be built into a template that analysts and stakeholders use when developing, analyzing, and discussing a concept of operations. In fact, this information can be formalized into a formal, mathematical model that can be rigorously queried to ensure completeness and consistency. Such a formalization will be shown in later work.

The roles of the controller, actuator, controlled process, and sensor, and their interactions with the environment and other control loops can be summarized with 13 generic keywords

or tagwords. Figure 1 depicts these tagwords in the familiar control loop format. With a proper accounting of these 13 items, the control loop can achieve the four necessary conditions[1] [Ash57] of process control and adequately interact with its environment, other processes, and other controllers. In other words, these keywords are necessary to ensure that a control loop is controllable and coordinable with other controlled processes.

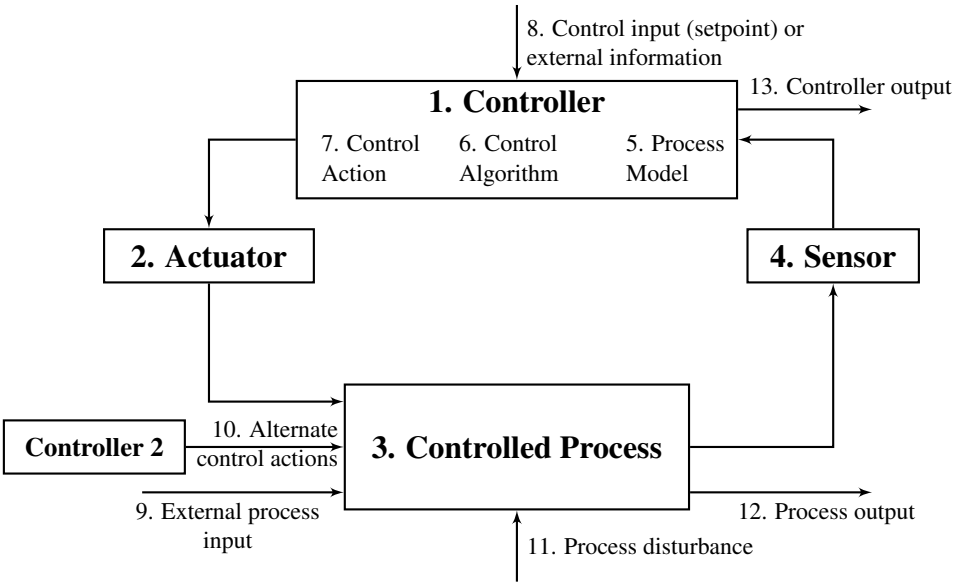


Figure 1: Control Loop with generic STAMP entities

The information in Figure 1 and the above lists (Controller, Sensor but also Actuator and Controlled Process) can then be used to systematically parse and query the natural language description or graphical depiction in a concept of operations. The resulting model and database are easy to interrogate and visualize. These qualities help the analyst to check for internal inconsistencies and/or missing information that may result in unsatisfied control conditions, and also to check for inconsistencies across the system hierarchy.

3 Conclusions

The workshop presentation will further show how these concept models can be interrogated systematically in order to identify vulnerabilities and suggest potential architectural alternatives. The presentation will use a real-world application in air traffic management being developed jointly in the United States and Europe called Trajectory-Based Operations (TBO).

[1] (1) Goal, (2) Action, (3) Model, and (4) Observability conditions are required for effective process control.

References

[Ash57] W Ross Ashby. *An Introduction to Cybernetics*. Chapman & Hall Ltd., 1957.

[Del13] Lenny Delligatti. *SysML Distilled: A Brief Guide to the Systems Modeling Language*. Pearson Education, 2013.

[Dor02] Dov Dori. *Object Process Methodology: A Holistic Systems Paradigm; with CD-ROM*. Springer, 2002.

[EB12] Steven D Eppinger and Tyson R Browning. *Design structure matrix methods and applications*. The MIT Press, 2012.

[IEE98] IEEE. IEEE Std 1362—Guide for Information Technology—System Definition—Concept of Operations Document. Technical report, Institute of Electrical and Electronics Engineers, Inc., 3 Park Avenue, New York, New York 10016-5997 USA, 1998.

[INC11] SE INCOSE. INCOSE Systems Engineering Handbook v. 3.2. 2. Technical report, INCOSE-TP-2003-002-03.2. 2. October, 2011.

[JPD11] JPDO. JPDO Trajectory-Based Operations (TBO) Study Team Report. Technical report, Joint Planning and Development Office, 2011.

[Kap10] Stephen J Kapurch. *NASA Systems Engineering Handbook*. DIANE Publishing, 2010.

[Lev12] Nancy G. Leveson. *Engineering a Safer World*. MIT Press, 2012.

[LZ11] David Long and Scott Zane. Language: The Systems Model Is Language-Based. In A Primer For Model-Based Systems Engineering (2nd ed.), 37. Technical report, Vitech Corporation, Blacksburg, VA, 2011.

[Sod02] Nathan R Soderborg. *Representing systems through object-process methodology and axiomatic design*. PhD thesis, Massachusetts Institute of Technology, 2002.

Application of STAMP to Improve the Evaluation of Safety Management Systems

Robert J. de Boer, Raymon van der Maarel

Aviation Academy
Amsterdam University of Applied Science
Weesperzijde 190, 1097DZ, Amsterdam
The Netherlands
rj.de.boer@hva.nl
Raymon.van.der.Maarel@hva.nl

1 Background

Although Aviation safety is continuously improving and has reached an all-time low in 2013 [FM 14], [ICAO 13], further efforts are still required to maintain absolute safety levels in periods of growth, and to improve safety levels in those niches of the air transportation system that have poor safety performance. One development to support continuous improvements in safety is the compulsory implementation of Safety Management Systems (SMS) [ICAO 13b], [CAA 14]. The adoption of Safety Management Systems to warrant aviation safety differs significantly from conventional safety assurance. Safety Management Systems acknowledge that any human endeavour cannot be completely free of hazards and risks, including those due to human activities and errors. An appropriate balance needs to be maintained between production and protection (cf. [H 14]). Safety is an "emergent", dynamic property of the complex, open aviation system so that safety risks needs to be continuously assessed and mitigated. The objective of the SMS is to maintain safety risks "under an appropriate level of control" [ICAO 13b]. In contrast, conventional safety assurance focuses on compliance to rules.

The change from a compliancy-based to a risk-based safety management requires a modification in oversight from compliance oversight to performance oversight [W 13]. However, evaluation of safety performance is not enough: although improved safety performance is a stated objective, SMS allow companies to achieve "an acceptable level of safety while balancing the allocation of resources between production and protection" [ICAO 13b]. Companies will always have to comply with a minimum set of regulations, but company management has some leeway in deciding what they consider to be an acceptable level of safety and how they allocate resources to mitigate risks. Therefore, the oversight of SMS cannot be entirely based on objective or numerical safety performance. It is therefore interesting to investigate whether and how the European aviation regulatory system is coping with this new challenge. For instance, one development is the guidance material for the national regulatory authorities to assess the implementation of SMS at the company level. The Safety Management International Collaboration Group (SM-ICG, consisting of representatives from industry and regulatory authorities) has generated such

guidance for the oversight of SMS, in the form of an evaluation tool for SMS [SMICG 12].

STAMP is a relatively new hazard analysis method based on system thinking rather than on component thinking [L 11], [L 13]. It is based on the principle that the sum of the safety of each component of a system is not necessarily equal to the safety of the whole system. One will need to look for conflicting interactions or safety constraints among components in order to guarantee safety. In a system, components could be physical components as well as organisations or persons within organisations. Applying STAMP to management systems is even more unique and has barely done before (cf. [dB 14]). There is a need for more knowledge about this hazard analysis method and its practical application to supervisory socio-technical systems.

2. Purpose

To identify the value of applying STAMP to the European aviation regulatory system in identifying risks that have not previously been identified for the change from a compliancy-based to a performance-based oversight. The scope of the current study is limited to aviation maintenance as it is vulnerable for complex and inefficient regulatory systems since maintenance companies are relatively small. Besides, the decreasing number of aircraft crashes resulting in an increasing level of in-flight safety forces the industry to pay more attention to on-ground hazards [OSZ 13].

3. Approach

The current study has applied a modified five-step STPA methodology [L 13], [dB 14]: (1) Hazards and safety requirements; (2) Functional control structure; (3) Control actions; (4) Allocation of safety requirements to components; and (5) Control loop effectiveness. The results of the analysis is compared to the currently known hazards described in the most recent EASA annual report [EASA 12]: (1) Inadequate provision of resources in relation to existing and new tasks (extension of the remit) of the Agency; (2) Unaddressed safety issues at the time of certification, validation, approval of flight conditions resulting in a crash of an aircraft. (3) Unaddressed safety issues resulting in a catastrophic consequences due to non-issuance of an airworthiness directive or insufficient corrective actions, inadequate compliance time specified by an Airworthiness Directive; and (4) Inadequate occurrence reporting system and ineffective processing of occurrence reports discrediting continuing airworthiness of products and leading to preventable accidents or serious occurrences.

4. Results

The main goal of the European aviation regulatory hierarchy system applied to maintenance service providers is to create an environment in which legislation is drawn

up and reformed according to the latest knowledge, occurrences and developments; and this legislation is provided to the industry in such a way that aviation maintenance is carried out properly and timely in order to maintain or reduce the number of incidents, accidents and losses caused by technical failures. Taking into account this goal, a generic hazard has been established which conflicts with the aforementioned goal. The system-wide hazard is that maintenance is carried out in such a way that the aircraft does not function as intended and crashes or causes economic or environmental impact as a result of at least one technical failure. This system-wide hazard can be split up in more specific and better understandable hazards, which apply to different areas of the system. These hazards are: (1) Regulations are not prescribed resulting in unsafe operations of maintenance organisations on its turn resulting in incidents or accidents; (2) Maintenance organisations do not comply with prescribed regulations resulting in unsafe operations of maintenance organisations on its turn resulting in incidents or accidents; (3) Risks, other than the risks mitigated by prescribing regulations, are not identified and thus not mitigated resulting in unsafe operations of maintenance organisations on its turn resulting in incidents or accidents; and (4) Feedback on the functioning of the regulatory system is not provided resulting in a not completely effective regulatory system in terms of preventing the three aforementioned hazards from happening resulting in unsafe operations of maintenance organisations on its turn resulting in incidents or accidents. Each of these is hazards is translated into constraints iteratively in discussions with EASA, see Van der Maarel [vdM 14] for a full account.

Figure 1: Control structure for the European aviation maintenance regulatory system

In the European aviation maintenance regulatory hierarchy, organisations are involved which (1) carry out maintenance and have to comply with the prescribed regulations, (2) national authorities which enforce regulations to maintenance organisations and audit compliancy and (3) European authorities which draw and adapt regulations and enforce them to Member States. The control structure for the European aviation maintenance regulatory system is shown in Figure 1.

Susequently, it is established what control actions exist in the control structure. This is based on Part-M and Part-145 regulations together with documents from several European Institutions. The first constraint is: 'Risk control in the form of regulations shall be in place at any level'. This constraint can be translated basally into the control action 'command to make, enforce or comply with the regulations'. The second constraint is: 'A system which properly oversees and enforces regulations shall be in place'. This constraint can be translated basally into the control action 'command to provide information related to auditing compliance or to perform audits to check compliance of organisations'. The second constraint is: 'A system which properly oversees and enforces regulations shall be in place'. This constraint can be translated basally into the control action 'command to provide information related to auditing compliance or to perform audits to check compliance of organisations'. The fourth constraint is: 'Feedback on the functioning of the system shall be provided'. This constraint can be translated basally into the control action 'command to provide feedback related to the functioning of the system or other relevant input'.

5 Conclusion

Three potential hazards have been identified in the capability of the regulatory hierarchy to evaluate the SMS effectiveness. EASA has no appropriate powers to directly address national aviation authorities. The EASA can either provide guidance material or they have to issue through the Commission which is seen as a heavy and time-consuming measure. Furthermore the Commission has political interests and opinions which could influence the decision-making process, whereas EASA has the most knowledge in the aviation regulatory work field and is politically neutral.

There is no proper risk-management capability assessment tool. Now, competent authorities mainly have to audit compliancy which is considered as a 'black and white' process. After the implementation of SMS, authorities will be required to audit risk-management competences of the maintenance service provider, which is harder and more time-consuming to do. Ongoing discussion is expected about the acceptable level of risk and an acceptable level of safety with the introduction of Safety Management Systems. It is not defined neither measurable what exactly is considered as an acceptable level of safety or as an acceptable level of risk.

Due to the financial crisis, governments have less leeway in their budgets. Whereas previously a compromise between safety and economics was never been made in the disadvantage of safety, now we see more well-considered and tight allocation of (financial) resources. The law prescribes the government to designate a competent authority and provide it with sufficient resources, which is a subjective requirement. Taking into account that SMS are required to be implemented in various kinds of aviation organizations, which in turn requires more work for the competent authority in initial approval and oversight, this leads to a potential hazard.

These findings are quite different to the currently known hazards described in the most recent EASA annual report [EASA 12]. We conclude that the STAMP analysis is a useful

addition to contemporary safety analysis tools, and look forward to further studies to identify its strengths and weaknesses. In particular, we expect to be able to study its value in comparison to currently available guidance material for the national regulatory authorities to assess the implementation of SMS at the company level [SMICG 12].

References

[CAA 14] CAA-UK, "SMS Implementation Dates," 2012. [Online]. Available: https://www.caa.co.uk/default.aspx?pageid=11558. [Accessed: 22-Apr-2014].

[dB 14] R. J. de Boer, "Using STAMP to Predict Safety: Application at a Ground Service Provider," Saf. Sci., accepted, 2014.

[EASA 12] EASA, "Annual Report 2012," Cologne, Germany, 2012.

[FM 14] D. Fouda and I. Maragakis, "Significant improvements for Global Aviation Safety in 2013." EASA, Cologne, Germany, 2014.

[ICAO 13] International Civil Aviation Organization, "Safety Report," Montreal, 2013.

[ICAO 13b] International Civil Aviation Organization, "Safety Management Manual (SMM)," Montreal, 2013.

[H 14] E. Hollnagel, "The ETTO principle as ETTOing – or Occam ' s Razor redux," 2012.

[L 11] N. G. Leveson, Engineering a safer world: Systems thinking applied to safety. Cambridge, MA: MIT Press, 2011.

[L 13] N. G. Leveson, "An STPA Primer," Cambridge, MA, 2013.

[vdM 14] R. van der Maarel, "Using STAMP to Analyse the Effectiveness of the European Regulatory System," Amsterdam University of Applied Sciences, 2014.

[OSZ 13] C. V. J. Oster, J. S. Strong, and C. K. Zorn, "Analyzing aviation safety: Problems, challenges, opportunities," Res. Transp. Econ., vol. 43, pp. 148–164, 2013. [SMICG 12] Safety Management International Collaboration Group, "Evaluation Tool," Cologne, Germany, 2012.

[W 13] C. Wassink, "The Netherlands - Implementation SSP & SMS."

Utilising STAMP to define the First-into-Man system for medicines

Dr Brian Edwards

Vice President Pharmacovigilance and Drug Safety

Alliance Clinical Research Excellence and Safety

brian.edwards@ndareg.com

Abstract: The (so-called) pharmaceutical safety system is struggling. Part of the problem is failure to define the complexity of the system and design processes which are evidenced based and both adaptable and flexible for those who work in them. The size of the challenge is huge and global. This is why ACRES (Alliance Clinical Research Excellence and Safety) was established as a not-for-profit international organisation to develop a global system for clinical trials, responsibly conducted according to the highest standards of safety, quality and efficiency as this currently does not exist. When defining the safety system for clinical research into medicines where do we start? ACRES suggests examining and defining the current safety system for how medicines move from the laboratory into man applying the STAMP model. This has never been done which is why the influence of some stakeholders on safety remains unknown. The aim is to build on what we know about the scientific evidence, current legislation and regulatory guidance underpinning pharmacological and regulatory science in pharmaceuticals. We aim to build a multisciplinary team including safety engineers This includes working with pharmaceutical professionals with different expertise and safety engineers with experience in applying STAMP from both the EU and US to map out the system using STAMP.

1 Introduction to ACRES and its rationale

The Alliance for Clinical Research Excellence and Safety (ACRES) aims to transform the world of clinical research so clinical trials are more responsibly conducted according to the highest standards of safety, quality and efficiency. ACRES is a not-for-profit organization whose leadership is comprised of experts in health policy, pharmaceutical

development and safety, academia, research sites, patient advocates and the ethics community. No such organisation exits with the intent of better defining and implementing a global system for clinical research.In order to effectively manage risk, it is essential to have an agreed definition of what the ideal system should be for the safe development and life-cycle management of medicines and yet this does not exist. The system for developing medicines has never been mapped out systematically to define all risks and hazards. Identifying all stakeholders who impact risk management has never been performed. We were influenced by an analysis of the US post-licensing system perfomed using Systems Theoretic Accident Model and Processes (STAMP) methodology [LCT13]. The STAMP model may offer an opportunity to do just this in an international setting. This because the guiding principles of Good Clinical Practice (GCP) underpins all intervention a research with medicines in humans [EMA12], [ICH 96], [FDA 13], [DoH 01], [FDA 06], (MHRA 13]. We propose as a first step, the examination of the early phases of drug development to determine if the implementation of improvement tools from organisational science (such as STAMP) can make a significant impact on safety. By STAMP treating safety breakdown as a control problem, not a failure problem, we can more logically design constraints on component behavior and interactions. This approach requires applying additional evidence from human factors research and safety culture so that we can better define how all the different players in clinical research interact and control safety within the system.

1.1 Why is such an approach needed?

Currently clinical reseach is based on application of regulations to pharmaceutical science:

• Too often regulation for clinical research are reactions to safety issues rather than being prospectively designed based on evidence

• Regulations stress compliance to systems and processes rather than prioritizing the avoidance of potential harm to human subjects.

• Regulations are added by different regulators without a holistic view of how they impact the system

• There have been no known attempts to apply organisational science techniques such as human factors with little attention to human performance analysis to identify and mitigate risk.

• There is no agreed Clinical Development Accident terminology.

• There are concerns about less than rigorous methodology and compliance in less regulated territories

• There are concerns about volunteer remuneration, lack of informed consent, informed consent under duress, coercion

• Drug development decisions are highly influenced by economics: market forces, commercial considerations, available capital. As a result, clinical development is often an exercise in risk aversion rather than risk management; as a result, many promising therapies are never developed.

In effect, we do not know what the system for moving medicines from bench into man looks like and what the systemic factors are for managing risk and there is no standard conduct for administering products for the first time into Man. Expertise outside of Biomedical R&D is crucial to success to creating a safety culture in our business sector.

2 Outline of the project

This novel ACRES project will examine First-into-Man (FIM) Clinical Research including early phase clinical trials (Proof of concept programs) to:

• Identify current system components, strengths and improvements that can be made in in the safety activities of research at this phase.

• Review the regulatory framework, role and relationships in this phase of clinical development, adapt to a systems model and identify potential changes that may be required.

Various approaches and methodologies may be used in examining the system of clinical research in the FIM phase to identify its core issues and strengths as well as identify who are the stakeholders who infuence safety. This will enable us to better utilize available safety engineering approaches, such as STAMP, in the examination and evaluation of system strengths and weaknesses, as well as their root causes which often relate the Human Factor. This way safety contraints can then be designed to design a more logical and scientifically satisfying system for managing risk.

2.1 Objectives and results which might be expected

2.1.1. Complete a fully defined system 'map' of the various components of the FIM Clinical Research including early phase clinical trials.

2.1.2. Apply available approaches, including STAMP, to examine this part of the overall clinical research system to determine:

a. System strengths including identifying who are the influences on safety

b. System failures defined in depth in accordance to the STAMP methodology

c. Define what system accidents need to be prevented

d. Propose system changes

e. Identify expected or possible difficulties with applying STAMP and develop solutions, if possible. During the workshop expected and possible difficulties in implementing the plan and applying STAMP to address the systems deficiencies will be discussed.

Expected difficulties by designing a theoretical system rather than one based on real life data (which would be commercially sensitive and unlikely to be made available) are as follows:

• Lack of information or inaccurate information especially based on what we would like to happen. Indeed, we have to assume that all in the system are agreed on the objectivies of clinical research into medicine

• Variability within a process often is not defined or agreed

• Intrinsic pharmaceutical uncertainty in developing a new medicine making it difficult to define safety control actions

• There may be multiple process controllers

• Impact of inaction and organisational silence on managing risk is uncertain

• The environment in which controllers operate may be variable and difficult to define

3. What is required for this project?

To conduct a project of this type will require various types of expertise and a range of of professionals as follows:

• Project Manager (pharmaceutical industry preferred)

• 10-12 Varied pharmaceutical professionals with experience in early phase research and safety for workshop with two moderators

• Lead STAMP Expert(s) for conduct of STAMP methodology and approach

• Graduate students to conduct interviews and document review

4. References

[LCT 13] Leveson, N. Couturier, M. Thomas, J. Dierks, M. Wierz, D. Psaty, B. Finkelstein, S. Applying System Engineering to Pharmaceutical Safety Journal of Healthcare Engineering Vol. 3; No. 3, 2012 pp. 391-414

[EMA 12] Reflection paper on ethical and GCP aspects of clinical trials of medicinal products for human use conducted outside of the EU/EEA and submitted in marketing authorisation applications to the EU Regulatory Authorities EMA/121340/2011 16 April 2012

[ICH 96] Guideline For Good Clinical Practice ICH 6(R1) Current Step 4 version 10 June 1996

[FDA 13] Oversight of Clinical Investigations — A Risk-Based Approach to Monitoring U.S. Department of Health and Human Services Food and Drug Administration August 2013

[DoH 01] The revised Declaration of Helsinki: interpreting and implementing ethical principles in biomedical research International Journal of Pharmaceutical Medicine 2001, 15:131–143

[FDA 06] Exploratory IND Studies Food and Drug Administration Center for Drug Evaluation and Research (CDER) January 2006

[MHRA 13] Clinical trials for medicinal products: Phase 1 Accreditation Scheme v.2 October 2013
http://www.mhra.gov.uk/Howweregulate/Medicines/Medicinesregulatorynews/CON203 3112

Applying STAMP to safety standards of mowing robots

Mitka Eleftheria

Electrical and Computer Engineering
Democritus University of Thrace
University Campus, Xanthi
67100 Greece
em3933@ee.duth.gr

Abstract: The use of mowing robots in anthropic environment is increasing, leading to the call for specific safety standards. This is a challenging endeavour, however, because the user must usually be placed in the robot's workspace and the residents must frequently interact with the robot. Although specific safety standards for mowing robots do not yet exist, there are several mowing device standards and well-established principles of risk analysis and safety design that can and should be applied. This paper presents an overview of safety constraints for mowing robots, starting with a discussion of high-level system safety requirements, followed by methods for risk assessment (or hazard analysis) and a presentation of STAMP (Systems-Theoretic Accident Modelling and Processes) as a safety strategy.

1 Introduction

Mowing robot is a mechatronic system used in a domestic environment. Robotic mower is constructed to cut grass in gardens at any day and time without human intervention. It is usually small, compact, noiseless and light enough to be transported manually. This home-used appliance could be adjusted to cut in different locations. Throughout its task, the apparatus cut the grass surrounded by a perimeter cable. Its scope is to cut wide locations, preferably at daytime according to the battery charging cycle. When it bangs with an obstacle such as branches, small stones, wall or fence, it bounces and goes along a different direction. The lawn-mower executes an asymmetrical movement pattern that is not recurred depending on its sensing of growth of grass.

Since particular safety guidelines for mowing robots do not exist, the authors suggest that their safety-guided design should be based on standards that developed in areas of agriculture and garden equipment. In addition, the methodologies for safety for agricultural equipment [AA09], the standard for safety of household and similar electrical appliances [EN94] providing safety requirements for domestic environment may be useful in establishing standards for these robots. Every manufacturer should follow the basic instructions summarized in the basic safety standards such as safety of machinery [EN91], safety distances and danger zones [EN92], emergency stop requirements [EN92], safety of household and similar electrical appliances [BS97], Safety Specifications for Commercial Turf Care Equipment [AO04] and safe features

according to Safety Specifications for Turf Care Equipment – Power Lawn Mowers, Lawn and Garden Tractors [AO03]. Establishing standards for controlling hazards in anthropic environment involves the entire control system and modelling the socio-technical diagram that will enforce the primary constraints in design level. Ineffective control is a consequence of lack of standards or from standards that are not enhanced adequately. The author below apply STAMP in order to build a robust base with safety requirements regarding mowing robots.

2 Applying STAMP to standards

2.1 Identify the accidents

The first step in any safety analysis is to define which types of accidents needs to be considered. Indeed, the manufacturer that develops this kind of robotic system has to take account of the accidents occurred. The accident model that derives from this definition does not indicate a particular causal factor [Le09]. In other words, in a systems-theoretic view of safety, accidents could be recognized, by categorizing the set of safety requirements that were ignored and answering why the control efforts were inaccurate in enforcing them. In addition, an accident definition in the design of a mowing robot is:

A1. Explosion of the device caused by storing the device in an improper place against manufacturer's specifications.
A2. The mower is moving near an operating rotisserie or barbecue where vapor may reach a flame or a spark causing an explosion of the device.
A3. The user is changing the operating program leading to a harmful injury.
A4. An abnormal high speed movement is not blocked by the stop button damaging property or robot's equipment.
A5. The perimeter wire is cut leading to breakdown of the robot.
A6. The mower chassis has burnt signs or signs of corrosion causing skin/eyes damage.
A7. The drive motors are damaged due to severe overloading of the wheel motors and overheating.
A8. A pet or children could be pinned under the robot if the alarm of the robot doesn't warn about restart or scheduled programming operation.
A9. The mower is mowing outside the yard and not across the perimeter wire hitting someone.
A10. A resident is mowing via manual control over an obstacle and the object is thrown at speeds greater than 200 mph and 50 feet or more, causing death or injuries ranging from blindness to severe bruising.
A11. The collision detection system of the device is not functioning leading to trapping points.
A12. The equipment of the lawn mower starts vibrating abnormally.
A13. Water and other liquids are sprayed directed to the robot with a garden or a liquid spray hose causing internal electronic damage of electrical connections.
A14. Inability of the device to follow the tasking path due to incorrectly placed perimeter wire/ too much slope/ overheated power supply unit/ incorrect wheel transmission/ bumper pressed during warm up/ bumper pressed for more than 2 sec during manual mowing/ bumper pressed while departing from the charging station.
A15. A resident could be injured by contact with the blades.
A16. A child may fall off and be seriously injured if he rides on an operating robot.
A17. A resident that uses the robot via remote control hits a person or a pet by pulling the mower backwards without looking.
A18. Severe injury could occur if hair, fingers or clothing of resident are caught in the exposed mechanism of the robot due to indented opening of the cover.

2.2 Assign a level of severity

With aim to arrange the severity of the above accidents, the precedence of the levels of severity is considered necessary. Severity is described as the seriousness of the hazard. The overall safety policy for arranging a level of severity for an accident is that all

accidents leading to death or injury must be restricted from the design level [Le95].

Level 1
A1-1: The collision detection system of the device is not functioning.
A1-2: A resident is mowing over an obstacle.
A1-3: The mower is mowing outside the yard hitting someone.
A1-4: The user is changing the operating program leading to a harmful injury.
A1-5: Vapor may reach a flame or a spark causing an explosion of the device.
A1-6: Explosion of the device caused by storing the device in an improper place.
A1-7: Rotating blades catches part of resident's body.
A1-8: Coming in contact with any exposed mechanism part of the robot.
A1-9: A child rides on an operating robot.
A1-10: A resident is pulling the mower backward without paying attention.
Level 2:
A2-1: A pet or children could be pinned under the robot.
A2-2: The mower chassis has burnt signs or signs of corrosion.
Level 3:
A3-1: An abnormal high speed movement is not blocked by the stop button.
A3-2: Inability of the device to follow the tasking path.
A3-3: Water and other liquids causing damage of electrical connections.
A3-4: The mowing equipment starts vibrating abnormally.
A3-5: The drive motors are damaged due to severe overloading of the wheel motors.
A3-6: The robot ceases its operation due to cut of the perimeter wire.

2.3 Identify the High-Level System Hazards

On top of the measures taken by the manufacturer to avoid the risk of hazards that robots in general might cause to the user, the product should also be evaluated against the following list of possible dangers [MGKM12]:

- Mechanical Hazard (Crushing, Shearing, Cutting or severing, Entanglement, Impact, High pressure fluid injection, Shape, Acceleration/deceleration, inadequate mechanical strengths, Mass and velocity, Potential elements or elastic elements).
- Electrical Hazard (Contacts of persons with live parts, Breakdown, Leakage current, Electrostatic phenomena, Thermal radiation)
- Hazard generated by vibration (White-finger disease, Neurological, osteo-articular disorders)
- Hazard generated by radiation (electromagnetic fields, infra-red light, visible light and ultra-violet, light Laser radiation, X and γ rays, α and β rays, electron or ion beams, neutrons, ionizing or non-ionizing radiation)
- Thermal Hazard (Burns and scalds)
- Hazards generated by neglecting ergonomic principles in machine (physiological and psycho-physiological effects, human errors)
- Slipping, tripping and falling hazards.
- Hazard generated by materials and substances (ingestion, inhalation of fluids, gases, mists, fumes, fibers, dusts or aerosols (harmful, toxic, corrosive, teratogenic, carcinogenic, mutagenic, irritant or sensitizing effect, biological hazards))
- Environmental Hazards (temperature, wind, snow, lightning, vapor, explosive or flammable atmospheres) [ISO10].

For practical reasons, a restricted set of high level system hazards need to be identified considering safety as an emergent property and not a component property, as a freedom of accidents or losses and combining potentially hazardous conditions with accidents identified at first step [Le09]. Beginning with an extensive set of system hazards, usually result in a confused hazard analysis even in the case of the most simple systems. The preliminary hazards might be defined as [MGKM12]:

H1: Environmental hazards [A1, A2]
H2: Electrical hazards [A5, A8]
H3: Mechanical hazards [A4, A7, A10, A12, A14, A15]
H4: Falling and tripping hazards to residents [A9, A11, A16, A17]
H5: Hazards generated by materials and substances [A6]
H6: Hazards generated by neglecting ergonomic principles in machine [A3, A13, A18]

2.4 Identifying Safety design constraints and safety requirements for the hazards

After the hazard and system identification, the next important step is to define the system-level safety requirements and design constraints that are essential to protect from incidents (Table 1). Hazard identification is considered as a top-down procedure that must take into account the harmonization of the system, eliminating or controlling a hazard that is considered possible to allow a loss to take place [Le09]. According to

STAMP, the constraints may differ in the design level [Le12].

Table I. Design constraints and safety requirements such as user instructions, user information and observations for some of the mowing robot's hazards identified above

HAZARD	SAFETY DESIGN CONSTRAINTS
Explosion of the device due to improper storing leading to environmental hazard.	a. Remove the power plug from the charging station in case of storing the robot. b. The mower shall be equipped with a manual controller so that the user can simply move it from the grass back to a storage place when not in use. User shall not store the robotic equipment in an overheated and enclosed space, such as a garage or shed, where vapour may reach a flame or a spark or a source of extreme heat or on a hill. c. Do not leave containers with grass cuttings in the storage or charging area of the device. d. The appliance should be stored in a sheltered and dry place with good ventilation and lightning conditions. e. To reduce the fire hazard, keep the robot, charging station and storage area free of grass, leaves, or excessive grease.
The mower explodes due to flame or spark leading to environmental hazard.	a. Always keep the mower away from water, heat sources (stoves, radiators, open flames, water heater etc.), and hazardous chemicals to avoid electrocution, overheating or chemical burn hazard. b. Human–robot environment (chemical, thermal conditions). If gasses, liquids or combustible substances are contained inside the robot, the designer should ensure that any increase of temperature will not cause burn injury.
The user is changing the operating program leading to a harmful injury leading to hazard caused by neglecting ergonomic principles in machine.	a. Settings that could be stored in ROM might include: manual user options, user preferences (sound, wire position, language, learn edge, blades replaced, anti-theft), charging options (weekly program, entry points, auto depart), safety tests, information (total time, run time), temperature, software version, charging voltage, charging temperature, entry points (set, default), child guard, lock settings, alarm function (disable, enable), rain sensor (restart, disabled, pause), auto setup (enable, disable), sound, date and distance format, time (set real time clock), signal type, lock keyboard, scan width, zones setup, first time calibration and ground clearance, follow loop, garden shape (open, normal, complex). b. Robot with operating units shall use only ROM to prevent from changing the operating program. c. Faulty programming may result in anti-theft alarm don't stop beeping or not operating, incorrect set of clock, incorrect work time, not enough work hours, not completely mowed the secondary area, the remote control don't work.
An abnormal high speed movement is not blocked by the stop button leading to mechanical hazard.	a. The software of the robot should bear a self-checking program that manages suspicious movements. Unless such a program exists, the user shall not be able to reach a control to stop a disoriented robot in proper time (as in industrial robots). Suspicious movements are mowing out of the mowing area due to a disconnection of the wire, work less that being programmed, doesn't cross over the wire, doesn't complete the edge, doesn't following the plot, doesn't working at the time scheduled, doesn't position correctly in the charging station, doesn't mow properly around the flowerbeds. b. Frequent stops and restarts require additional battery energy. c. User should try to control continuously the robot when operating it via remote control.
The perimeter wire is cut leading to breakdown of the robot leading to electrical hazard.	a. In the user's guide, the user of mower is informed that he should periodically inspect blades for foreign material or debris, calibration for interference, loop signal from the wire for interference from metallic objects (perimeter fence, reinforcement bar), the shield of wires for damaging, the battery for leading of corrosive liquid, the covering hood for opening, to clean base-station with a damp cloth, to confirm good contacts between station wires and contacts and that power supply is plugged at main power receptacle and that the coiled cord is properly placed in the holder, to check the reliable connection of the power cable with the wire connector.
The mower chassis has burnt signs or signs of corrosion leading to hazard generated by materials and substances.	a. Frequently clean the charging station and the robotic contacts using only a damp cloth. b. Use only a damp, soft sponge or cloth and a dry brush to wipe the outer surfaces after scraping. c. User shall not use harsh or abrasive cleaning solutions. d. If the chassis remains dirty, a soap or washing-up liquid might be necessary. e. User shall not wash the inherent components so as not damage electric and electronic elements since mower is non-waterproof.
The drive motors are damaged due to severe overloading of the wheel motors and overheating.	a. User shall not cut grass on slippery conditions where the traction is reduced and it may block the discharge or dry/scorched lawn. b. User shall inspect the drive wheel and remove grass residuals or other object. c. If the motor overheats, then there are two ways of unblock. If the parameters fall within the first range, the robot will execute manoeuvres to unblock the blade. If the over-current is below the allowable limit, the robot will stop and signal an error. Motors have operated with over-current for too long or some objects are accumulated or clogged around the blades.
A pet or children could be pinned under the robot if the alarm of the robot doesn't warn about restart or scheduled programming operation leading to electrical hazard.	a. Before turning on the mower, user should make sure that small children, pets or obstructions are far away of the yard where the robot is operating. b. If a dog hits a mower or put its paws, nose, or mouth-parts under the edge of the chassis, the possibility of injury increases. c. The restart procedure for lawn mower should require returning it manually to charging position, entering a restart password, checking the functionality and reliability of sensors by the robot software itself, releasing stored energy and troubleshooting any software bug caused by the emergency stop, diagnostics software detects the cause of any malfunction that lead to the stop. d. Improper restart might happen after an incorrect restart position, after wrong reminder counter of blades after the replacement, a temporary power loss, a prolonged activity. e. The robot may restart its operation elsewhere especially near obstacles (fallen branches, forgotten objects) due to severe load on wheel drive motors, due to unknown fault, due temporary power loss, after a long period of inactivity or after the replacement of the blades. f. Residents should be informed by the user for the user-programmed automatic weekly scheduled day and time. g. Control access of small children before mow in reverse, up and down via manual control. User should ensure that children younger than 6 are indoors, at least 100 feet away from the mowing field the robot are

		programmed to mow at all times and under the watchful care of a responsible individual.
The mower is mowing outside the yard leading to falling and tripping hazards	a. b.	User shall not use an extension cord to increase this distance from power receptacle. This event may be due to the perimeter wire being routed in the wrong direction around an obstacle, which blocks the signal. User shall pay extra attention if the area around the mowing field is populated or congested. Always look for traffic when mowing near roadways, walks or gravel drives.
A resident is mowing via manual control over an obstacle and the object is thrown at speeds greater than 200 mph and 50 feet leading to mechanical hazard.	a. b. c. d. e.	The area where the robot will execute must be well defined and operational flaws detected must be mitigated. Such operational contingencies include: inability to determine location in the lawn; inoperative sensors; not accessible charging/ base/ docking station; charging station not charging due to not active charging process; sensor failure; high temperature of an actuator or a wheel/ electrical blade/ drive motor; low battery due to inability to detect the charging station or wrong search settings or cutting long and dense grass, and so on. A wheel over the edge or an edge caving in could cause unexpected rollover. Obstacles that could lead to a collision, such as trees or bushes higher than 15 cm, do not have to be delimited by the perimeter cable. Mower will bang and bounce when it collides with this type of obstacle creating an island around them. User shall not operate the robot for trimming hedges, for mowing vegetation other than grass, for leaf or debris collection, or on grounds other than mowing field, for pulling or carrying loads. User should not mow over gravel, stones or hard, immovable objects like pipes, rocks or sidewalk edges or objects like blind corners, shrubs, trees with holes or flower around it that might obscure vision.
The collision detection system of the device is not functioning leading to trapping points.	a. b. c.	A subsystem that accomplishes the whole range management of collision detection through a virtual force field. This system allocates a force vector to obstacles across the mowing field. This force is controlled by the system using the signals obtained from sensory devices. Obstacles such as trees, flower beds, fountains or bushes higher than 15 cm shall not have to be delimited by the perimeter wire. Mower will bang and bounce when it collides with this type of obstacle creating an island around them. The perimeter wire shall not be laid around an object that robot can be allowed to collide with, like a fence or a dense hedge.
The equipment of lawn mower starts vibrating abnormally leading to mechanical hazard.	a. b. c. d.	In case of abnormal vibrations, user should stop the robot, remove the power plug and visually check for any damage of the blade or mowing chassis and search directly for the cause. The cause might be unstable blade disc or worn out blades. Vibration is usually a sign of malfunction. Remove the power plug from mower, in case that it begins trembling irregularly. User shall check the blades and screws and replace them if they are damaged. The malfunction that leads to continuous vibration has to be repaired by service experts.
Liquids are sprayed directed to the robot leading to hazard caused by neglecting ergonomic principles in machine.	a. b. c.	User or residents shall not spray water towards the robot, the charging station or the panel of the station. Never use the charger or charge the device in grasses with dampness or when wet contact is expected. User must not, under all circumstances, remove, bend, and cut, fit, weld electrical or electronic parts inside the chassis of the robot.
Inability of the device to follow the tasking path leading to mechanical hazard.	a. b. c. d.	User and residents should be aware of the automatic departure, the mowing path, the scheduled work time on an automatic weekly program for the season and the case of sudden overturn on slopes in order to detect when the device does not follow the operating path. Mowing robots might be used from physically and visually impaired persons and this affects the control design. Navigation control on mowers requires a cable to delimit the area to be trimmed. The manufacturer should inform the buyer of mower that he/she should keep all guards, shields, switches, safety devices such as front and rear bumpers, thermistors (overheat protection), manual controller buttons and lift sensors in place and properly connected. Inspect to verify that these protective devices are appropriately installed and operating correctly. User ought to be advised not to remove the protective guards that need to be used with the mower.
Exposed mechanism leading to hazard caused by neglecting ergonomic principles in machine.	a. b.	The mowing covering hood shall not collect grass residuals after mowing damp or wet lawn. It shall be inspected and maintained regularly for foreign material using a damp cloth or another similar tool [MM12]. The mechanism of the mowing robot shall be protected with plastic cover to prevent from bumping or changing it [MM12].
Pulling the mower backwards without looking leading to falling and tripping hazards.	a. b. c.	Online tutorials and help menus shall contain the appropriate instructions, so that users shall have direct access to information on how to operate the robot. A built-in electronic hardware control system and/or safety operational software shall be selected to force the robot to shut itself down in an emergency. Mowing robot shall be equipped with an emergency stop switch on the manual controller that ceases the rotation of blades and wheels within seconds [MM12].
A child rides on an operating robot leading to falling and tripping hazards.	a. b. c.	Warning signs shall be established to protect residents who may consider that they can ceaselessly be reckless with the operating robot. Specific responsibilities concerning safety shall be assigned to an adult user. The robot shall be equipped with a specific audio or visual signal, easily recognizable by everyone, to let people know whether it is on or off. Use a frequency that is not within the range of noise frequencies, in case of an audio signal [MM12].
Contact with the sharp rotating blades leading to mechanical hazard.	a. b. c.	Manual should advice the user to present dramatic description on the hazardous behaviors of mistreating the robot and warn correctly concerning the appropriate safety features. The replacement of not well-maintained or worn out blades shall not require routine or extraordinary maintenance more than once a year. User's manual indicates that every maintenance, service, replacement or inspection of worn or damaged parts should be carried out by service experts [MM12].

Conclusion

Since specific safety standards for mowing robots are not available, the author here proposed that they should be carried out based on the fundamental user / designer requirements in the domain of robotics safety. Moreover, the author proposed the implementation of STAMP on establishing standards that links system requirements with accidents, in order to ensure safety on such an emerging technology and to spread the presence of such robots in every-day life. A system is considered safe when firstly meets the basic standards of safety, which are defined from the national directives and regulations, and secondly all the safety requirements that had been pointed out by the authors, are fulfilled. These requirements could serve as a base on building standards in order to increase commercialization and acceptance of these robots by the customers.

References

[AA09] ANSI/ASAE S318: Safety for Agricultural Equipment, 2009.

[AO03] ANSI/OPEI B71.1: Safety Specifications for Turf Care Equipment – Power Lawn Mowers, Lawn and Garden Tractors, 2003.

[AO04] ANSI/OPEI B71.4: Commercial Turf Care Equipment - Safety Specifications, 2004.

[BS97] BS 3456: Part 2 Safety of household and similar electrical appliances - Section 2.42 Battery-operated lawnmowers, 1997.

[EN94] EN 60335: Part 1 - Safety of household and similar electrical appliances, 1994.

[EN91] EN 292: Parts 1 and 2 Safety of Machinery - Basic concepts, general principles for design, 1991.

[EN92] EN 294: Safety of Machinery - Safety distances to prevent danger zones being reached by the upper limbs, 1992.

[EN92] EN 418: Safety of Machinery - Emergency stop equipment, functional aspects – Principles for design, 1992.

[Le12] Leveson, N.: Engineering a Safer World: Systems Thinking Applied to Safety, The MIT Press, 2012.

[Le95] Leveson, N.: Safeware, System Safety and Computers, Addison-Wesley, 1995.

[Le09] Leveson, N.: Applying systems thinking to analyze and learn from events, Safety Science, Vol. 49, pp. 55–64, 2009.

[MGKM12] Mitka, E.; Gasteratos A.; Kyriakoulis N.; Mouroutsos G.S.: Safety certification requirements for domestic robots, Safety Science, 2012.

[MM12]Mouroutsos G.S., Mitka, E.: Safety-guided design concerning standardization's requirements of mowing robots, In proc. of International Conference APMS 2012: Advances In Production Management Systems, Rhodes, Greece, 2012.

[ISO10] ISO 12100: Safety of machinery —general principles for design, risk assessment and risk reduction, 2010.

Applying STAMP/STPA to Human Safety System for Four Wheel Drive Power-train

Yasuhiko Kawabe, **Tatsuya Yanagisawa**

UNIVANCE CORPORATION
2418 Washizu, Kosai-city, Shizuoka-Pref., 431-0494 Japan
yasuhiko.kawabe@champ.uvc.co.jp
tatsuya.yanagisawa@champ.uvc.co.jp

Abstract: We have been using FT or FMEA…etc to design our new products till now. In this study, we find that the new vehicle product development can be more efficient if we adopt STAMP/STPA, which shows the complicated relationship between human and machine and feedback the hazard in an early development stage.

1 Background

After the Great East Japan Earthquake happening on March 3rd, 2011, our company kept pressing forward with the evaluation and the preventive measures of Safety Risk. In particular, among the automobile drive parts, there are many parts needing gas carburization and quenching, which requires flammable gas. In order not to cause explosion or fire while the accident happens, we pushed forward the risk assessment and confirmed the effectiveness of using STAMP/STPA method. And we reported our work at the 2014 STAMP Conference. [Mn14]

Taking this occasion, UNIVANCE CORPORATION attempts to apply STAMP/STPA on the 4WD systematic safety planning and embed the safety management in the future products at the developing stage. Here, we are going to introduce how we build the scheme of development in this paper.

UNIVANCE CORPORATION's self-developed transfer case is equipped in the "GTR," which is the flagship 4WD vehicle of NISSAN Motors. Until now, it has also developed the 4WD systems of other vehicle manufacturers.

Different from ISO26262 [ISO11], which regulates the minimum safety of the individual components and became a great issue of concern recently, we adopted the idea of STAMP/STPA, which focuses on the safety of the whole system including the complicated relationship between human and machine.

We focus on the concept of "human factors" within STAMP/STPA method, in order to consider "the way of developing 4WD driving system within the vehicle automatic control function" and "the system safety of product development" at early stage of the development.

2 Applying STAMP/STPA

2.1 The characteristics of STAMP/STPA and this study

The STAMP/STPA is introduced by MIT Professor Nancy Leveson, the writer of " Engineering a Safer World." [Ln11] This methodology proves that even though there is no defect in the individual sub-systems or components, the defect will still occur while the whole system is built up by them.

Under the circumstance of the road traffic network nowadays, to drive the vehicle by automatic control completely is still under study. For the vehicles equipped with 4WD system, times of driving in the natural circumstance are more than the other vehicles. For example, although the high speed 4WD vehicles are supposed to run on the pavement, most drivers expect them to run in the complicated road conditions into which many natural conditions are added into, such as snowy roads or frozen roads…etc. Therefore, we think the product development which includes human recognition is necessary during the early stage of the product planning. And then, we move ahead on our study based on the cases which is related to human factors and were presented at the 2014 STAMP Conference. In this study, we attempt to explore this concept in depth.

First of all, STAMP/STPA has the following characteristics.

- Focusing on the system-level issues: considering the interactions between the controllers, but not only on the malfunctions of the individual components.

- Considering the accident is not caused only by the malfunction of the component, but caused by that the safety constraint toward "Unsafe Control Action" which leads to dangerous status is not practiced.

- Emphasizing that safety constraint should be recognized and embedded in the design, but not only the malfunction countermeasures, such as fault tolerance design.

- The cause of the accidents due to the software or human factors is that the process which the controller's supposition is inconsistent with the actual process.

- The unsafe control actions and the hazard factors can be analyzed by Guide Words.

Among these characteristics of STAMP/STPA method, we focus on the 4th one, and integrated STAMP/STPA into the thinking at the early stage of development.

2.2 Analyzing methodology which integrate the human behavior into the STAMP/STPA

The STAMP/STPA is Practiced according to the following steps.

- Recognition of the hazard and the high-level safety requirement.
- Building the Control Structure Diagram to control the hazard.(Step 0)

- Recognizing the hazard scenario caused by impropriate control action. (Step 1)
- Recognizing the potential cause that leads to hazard scenario. (Step 2)

In this study, in order to add in the human error factors, we refer to the Dr. Jens Rasmussen's "Skills, Rules, Knowledge (SRK) framework" [Rj90] which is mentioned in the presentation of Hoshino from Japan Manned Space Systems Corporation (JAMMS) at the 2014 STAMP/STPA Conference. This model shows the rule or mechanism which works in the process of recognition. It presents the real-life problems with all kinds of factors which correlated to one another complicatedly in the cognitive science.

Rasmussen states that when human practices a certain behavior, from a prospective of consciousness, that behavior is also being automated at the same time. And, the automation will be effected by the following 3 recognition levels.

- Skill-based behavior: Performance is smooth, automated, and consists of highly integrated patterns of behaviour in most skill-based control. [Rj90]
  ⇒it is practiced without intention, and sometimes a reflex movement.
- Rule-based behavior: Characterised by the use of rules and procedures to select a course of action in a familiar work situation. [Rj90]
  ⇒we can also shift this behavior to skill-based behavior by repeating practicing.
- Knowledge-based behavior: By identifying the environment, building up the psychological model and finding the countermeasures from the knowledge which one has already had.
  ⇒needed at complicated and ambiguous environment.

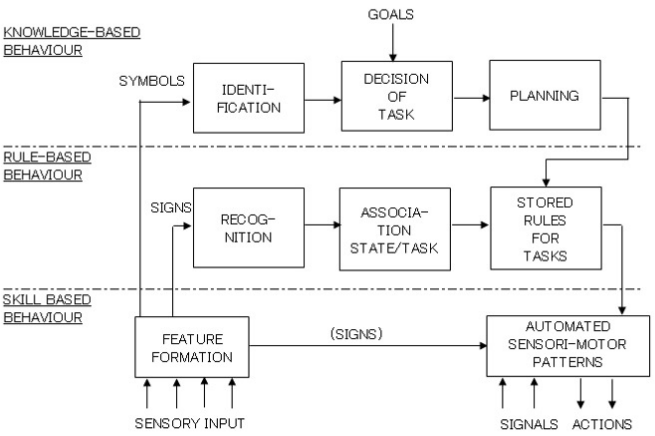


Figure 1 · The relation of the 3 Behavior Layers

The advantage of the model showed in the figure above is that it makes it possible to recognize or realize the interface behavior of the user and to find out the factors which cause human error. And then, it makes it possible to prevent the human error by improving the interface.

According to the statement of Hoshino from JAMSS(Japan Manned Space Systems Corporation), the process of human recognizing behavior can be divided into the following 4 items:

- Detection
- Identification
- Decision
- Action

And then, he divides every error pattern in each level defined by Rasmussen into 4 items.

Table 1 shows how the Human Mental Model analysis is conducted.

Layer	Process			
	①Detection	②Identification	③Decision	④Action
(1)SKILL-BASED BEHAVIOR				
(Ⅱ)RULE-BASED BEHAVIOR				
(Ⅲ)KNOWLEDGE-BASED BEHAVIOR				

Table 1. The idea of Human Mental Model analysis

3 Result of analysis

3.1 Applying to product developing

We suppose that if we can find out the conditions precedent that cause the system of 4WD vehicle to collide with the human behavior, then we will be able to build up a system which can automatically avoid the accident effectively. Under this assumption, we try to find out the collision between human and the system by the following condition which happens in 2 or 3 seconds.

Usually, it is necessary to consider the oncoming vehicle. However, because the purpose of this study is about the product development, we assume that there is no oncoming car. (One way, no car coming from the opposite direction)

① The start point is on a dry pavement. The vehicle can speed-up without difficulty.

② Suddenly, the snowy road with a sharp corner is confirmed.

③ During the corner, it is confirmed that in the front, there is an extremely sharp corner which concaves oppositely.

④ After passing the crowning point (clipping point), cornering out.

⑤ After that, in order to get into the next sharp corner, the driver turns the steering wheel to the end.

We suppose that if the driver is not a professional driver, the movement of the vehicle will spin out because of understeer.

This might be a rear case in US, but it exits in Europe or Asia where it snows. Figure 2 illustrates the situation above.

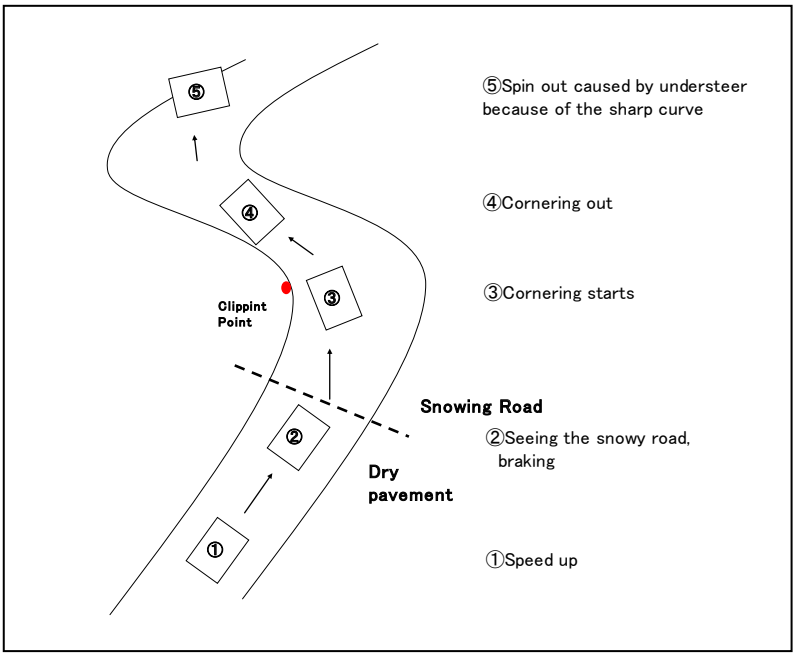


Figure 2 The situation of the road and driving conditions

Under this kind of situation, the following systematic control is necessary.

① Start moving and speed up: 4WD system reaches the fastest situation and the front-back torque distribution is under control.

② Braking: Optimizing the braking force in order to prevent the lock of the wheels.

③ Clipping point: Distributing the front-back torque in order to maximize the steering, acceleration, braking balance, and cornering force of the tires…etc. and increasing the stability.

④ Cornering out: Recovering the distribution of the front-back torque gradually and maintaining the balance of the vihecle.

⑤ Spin out caused by understeer: Entering the sharp curve again, and acting with the same logic as ③.

This kind of systematic control is programmed according to the driving features or experiences of the professional drivers. However, under the situation ⑤, it is too difficult for a normal driver. Therefore, we consider that the collision will occur between the systematic control and the driver's behavior here.

3.2 Analysis by STAMP/STPA

As we have mentioned above, what we want to talk about is "the secon sharp cuver on the snowy road." And we use STAMP/STPA to analyzed this situation.

3.2.1 Step 0 Control Structure

We have made the control structure diagram, considering the datum among the components which are related to the hazards.

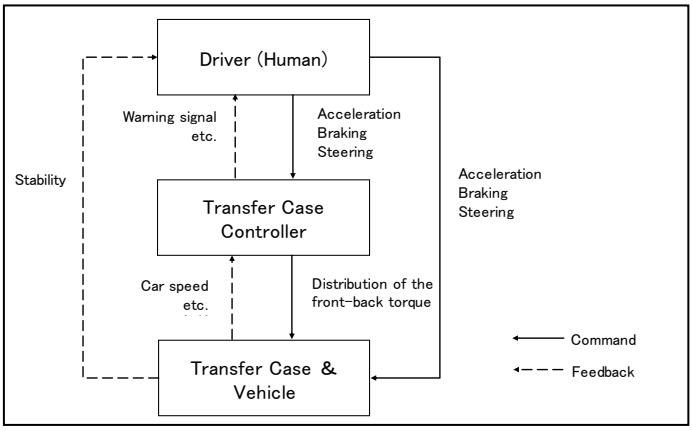


Figure 3 Control Structure Diagram

Here, we take the driver as one component and take the "transfer case controller" and the "transfer case & vehicle" which receive the command from the driver as the other two components. The commands from the driver are acceleration, brake, and steering…etc. While the transfer case controller receive the commands, it will compute the distribution of the front-back torque, and then send it to the transfer case & vehicle. After the transfer case & vehicle receive the commands from the driver and from the transfer case controller, it will practice the acceleration, brake, or turning, and provide the driver stability.

3.2.2 Step 1 Hazard senario analysis based on recognizing unsafe control action

We summarized the the general sequence of every command below.

#	Event / Command	from	to	Description
1	Acceleration	Driver	vehicle	Speed up
2	Acceleration	Driver	Transfer case controller	Compute the front-back torque
3	Steering	Driver	vehicle	Turing
4	Steering	Driver	Transfer case controller	Compute the front-back torque
5	Distributing the front-back torque	Transfer case controller	Transfer case & vehicle	Distributing the torque into front and back, increasing the connering force, in order to optimize the movement

Table 2 The general sequence of every command during a sharp curve

Once receiving the command #1, acceleration, and command #3, steering, the vehicle will speed up and turn in the same time. And, by command #2, acceleration, and command #4, steering, it will reach situation #5, in which the distribution of the front-back torque is worked out so that the "speed up & turning" becomes stable because of the distribution of the torque. Although all the commands are writen sequetially above, all of them happen in just 2 or 3 seconds.

Here, we analyze the unsafe control action of every command by aaplying the 4 guide words.

#	Event/Command	not provided	incorrectly provided	provided too early/late	stopped too soon
1	acceleration to vehicle	No hazard	become unstable, and then spin out (UCA1)	No hazard	No hazard
2	acceleration to controller	cannot compute the torque, but no hazard	cannot compute the torque, but no hazard	No hazard	cannot compute the torque, but no hazard
3	steering to vehicle	cannot steer correctly, and understeer + spin out (UCA2)	cannot steer correctly, and understeer + spin out (UCA2)	cannot steer correctly, and understeer + spin out (UCA2)	cannot steer correctly, and understeer + spin out (UCA2)
4	steering to controller	cannot compute the torque, but no hazard	cannot compute the torque, but no hazard	No hazard	cannot compute the torque, but no hazard
5	distributing the front-back torque	cannot increase the cornering force, and understeer + spin out (UCA3)	cannot increase the cornering force, and understeer + spin out (UCA3)	cannot increase the cornering force, and understeer + spin out (UCA3)	cannot increase the cornering force, and understeer + spin out (UCA3)

Table 3 Unsafe Control Action analysis

As shown in table 3, there are three kinds of unsafe control action.

UCA1: Becoming unstable, and then spin out. When it is time to speed up, if the steering or brake commands are incorrectly provided, the hazard occurs.

UCA2: Cannot steer correctly, and become understeer and spin out. If the adequate command (steering) is not provided at the adequate timing, the hazard occurs.

UCA3: Cannot increase the cornering force, and become understeer and spin out. If the torque of front and back is not distributed at the adequate timing, the hazard occurs.

3.2.3 Step 2 Analyzing the hazard factors by Control Loop

Among the unsafe control actions we have derived, we next focus on the UCA2 and practice the Control Loop analysis, because we want to focus on human behavior in this study.

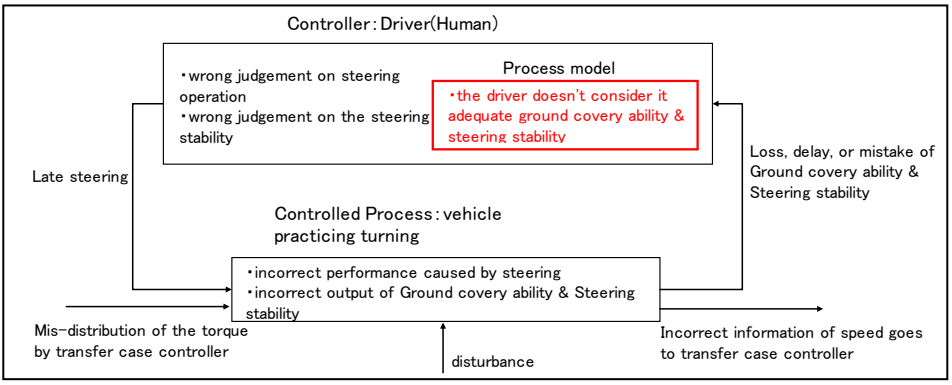


Figure 4 The analysis of hazard factors by control loop diagram

As shown in Figure 4, many hazard factors are found. According to the general process of STAMP/STPA, the safety countermeasures should be made for every hazard factor. However, because we focus on human behavior in this study, we only discuss the part that is relative to the driver(human). That is, we will analyze why the driver doesn't consider it adequate ground covery ability & steering stability.

3.2.4 Human Mental Model analysis

We use the Human Mental Model, which is mentioned in 2.2, to analyze.

Layer	Process			
	①Detection	②Identification	③Decision	④Action
(1)SKILL-BASED BEHAVIOR	(1a)mis-consider it as a gentle curve (1b)cannot judge if it is snow or ice road	N/A	N/A	(4a)speed up incorrectly (4b)fail to turn the steering wheel
(Ⅱ)RULE-BASED BEHAVIOR	(1a)mis-consider it as a gentle curve (1b)cannot judge if it is snow or ice road	(2a)cannot figure out that he should turn the steering wheel to the end and press the acceleratinopedal	(3a)because of fear, the driver judges that he should press the brake pedal and return the steering wheel	(4a)speed up incorrectly (4b)fail to turn the steering wheel
(Ⅲ)KNOWLEDGE-BASED BEHAVIOR	(1a)mis-consider it as a gentle curve (1b)cannot judge if it is snow or ice road	(2b)press the brake pedal because of the snowy road	(3a)because of fear, the driver judges that he should press the brake pedal and return the steering wheel	(4a)speed up incorrectly (4b)fail to turn the steering wheel

Table 4 the result of Human Mental Model analysis

If the driver is not a professional driver, he will take the (Ⅱ) RULE-BASED BEHAVIOR or (Ⅲ) KNOWLEDGE-BASED BEHAVIOR. And the process that may lead to the hazard will fall on Identification or Decision.

Each process which is necessary for the system control is as following.

A. Detection: Recognize there is a sharp curve again.

B. Identification: Identify that he should turn the steering wheel to the extreme and press the acceleration pedal.

C. Decision: Even the speed is high, “B” still works out.

D. Action: Turn the steering wheel to the extreme and press the acceleration pedal.

And then, we try to make the countermeasure for each situation.

ID	Causal Factor	Safety Constraints
(2a)	Cannot figure out that he should turn the steering wheel to the end and press the acceleratinopedal	For general drivers, it is difficult to complete the process in such a short time. Therefore, ESC(Electronic Stability Control) should be equipped.
(2b)	Press the brake pedal because of the snowy road	
(3a)	Because of fear, the driver judges that he should press the brake pedal and return the steering wheel	

Table 5 Countermeasure

For the general drivers, it is difficult to practice the process that the system does. However, due to the safety reason, we cannot just modify the behavior of the driver directly, but should try to prevent the event. Therefore, we decide to equip the vehicle with ESC (Electronic Stability Control) to monitor the yaw rate gyro and ensure the running stability.

Although we conclude with the countermeasure above, we are also considering the “auto-drive” concept that makes everyone be able to carry out the process that the systematic control attempt to pratice.

3.3 Safety countermeasure based on the method in the past

Naturally, the safety countermeasures are adopted in the development of vehicles, including GT-R etc. In the vehicle indutry, QFD (Quality Function Deployment) [IK98] [ONF98], including FT or FMEA, is applied. About the “spin out caused by understeer” issue which we discussed about this time, the safety countermeasure has already been made by the method in the past. Especially, the result of ESC equipment, which is derived from STAMP/STPA, is completely the same to the one which is derived from the method in the past.

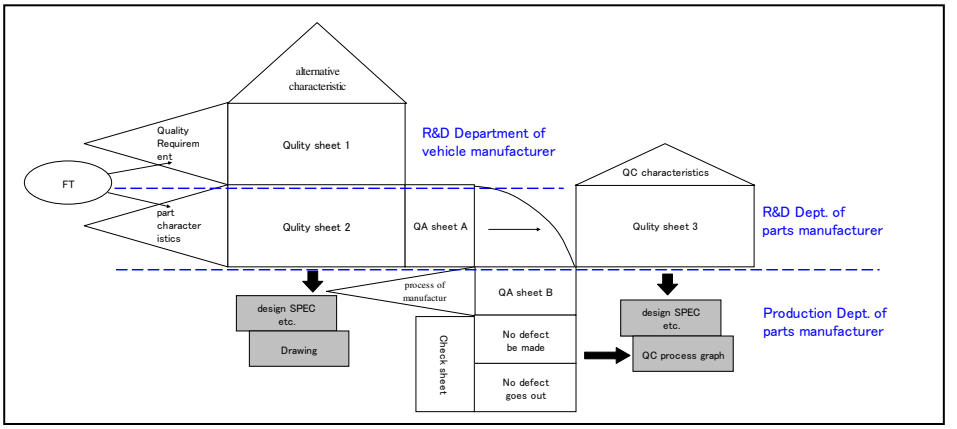


Figure 5 The procedure of QC based on QFD

By adopting the procedure shown in Figure 5, the design department receives the request from the customer and make instructions. Then, the production department or the suppliers receive that instructions and function smoothly to meet the customer's needs. Here, the R&D Dept. of vehicle manufacturer issues the quality sheet 1, and the parts manufacturers make quality sheet 2, and then QA sheet A, QA sheet B…etc. The vehicle's unique sales point, or the safety design concept is defined in the quality sheet 1. Here, we want to focuss on the quality sheet 1 because it acts as the STAMP/STPA analysis we performed in this paper.

Alternative characteristic / Quality required					characteristic	Initial		Durability				Appearance				
								heat		quiet						
						××××	××××	××××	××××	××××	××××	××××	××××	××××		
					value	○○kg	○○kg	○○kg	○○kg	○○kg	○○kg	△△	△△	△△		
product name	function	point (Quality)	Quality required (1)	quality reqired(2) detail		A	A	A	A	A	A	B	B	B		
Part A	To decorate ××	comfort	clean	the WAX powder won't stay on it									○	○		
		safety	hard to get hurt	won't get hurt even touching by hands	A							◎				
				nail won't get caught on when car was	A								○	○		
				clother won't get caught on	A								○	○		

Table 6 Sample of quality sheet 1

Table 6 shows a sample of qulity sheet, which is usually applied on product development, and the designer designs according to the detail value derived from the quality sheet or FT.

About the theme, spin out caused by understeer, the vertical axis should be:

- Characteristic(quality): safety, comfort

- Quality reqired(1): hard to get hurt, easy to drive
- Quality reqired(2): stable when acceleration or braking at the sharp curve

  stable when lane changing at the sharp curve

And the horizontal axis will be ESC and the alternative characteristics of the ECU.

This way, we can still consider the human level factors and conduct the safety design. However, if the person in charge is not well experienced, there may be miss in the FT graph, and it may take much time on breakdown all the detail characteristics.

4 Result and Observation

By STAMP/STPA, which can analyze the system including human behavior, we become able to find out the hazard faster than the methodology in the past. In the future, we attempt to stand on a higher safety point of view as the vehicle manufacturer to develop safest product. And, we also attempt to make people realise the importance of STAMP/STPA and push forward the standardization through these kind of product developing activities.

References

[Mn14] Morishita N.: Applying STAMP/STPA to Analyze the Cause of the Unexpected Fire Happening at the Heat Treatment Process, 2014 STAMP Workshop, 2014.

[ISO11] ISO26262-1:2011(en) Road vehicles-Functional Safety, ISO, 2011

[Ln11] Leveson, N.: Engineering a Safer World, Massachusetts Institute of Technology, 2011.

[Rj90] Rasmussen J.: Mental models and the control of action in complex environments. In D. Ackermann, D. & M.J. Tauber (Eds.). Mental Models and Human-Computer Interaction 1 (pp. 41-46). North-Holland: Elsevier Science Publishers. ISBN 0-444-88453-X

[IK98] Isobe, T.; Katagiri, S.: Quality Function Development concept for quality assurance, Advanced QFD Technology for Value Creation, Union of Japanese Scientists and Engineers, JUSE, 1998

[ONF98] Ooishi, T.; Negishi, M.; Furukawa, M.: Applied to the transmission of automotive and industrial vehicles, Advanced QFD Technology for Value Creation, Union of Japanese Scientists and Engineers, JUSE, 1998

Evaluating Systems with Multiple Processes Using STPA: A Case Study in a Medical Intensive Care Unit

Aubrey Samost

Engineering Systems Division
Massachusetts Institute of Technology
77 Massachusetts Ave.
Cambridge, MA, USA 0139
asamost@mit.edu

Abstract: This paper evaluates the use of STPA in a heatlhcare setting specifically a complex setting involving controllers participating in multiple processes at one time. This is a scenario that has not been explored with STPA before but is the status quo in most healthcare systems. In an intensive care unit, the process of providing care to patients is comprised of many processes including drug delivery, lab draws, radiological imaging, bedside procedures, physical therapy, and far more. The interactions of these procedures support patient care but also increase risk in healthcare delivery in ways that need to be considered.

1 Introduction

Risk analysis in healthcare today commonly involves the application of Failure Mode Effects Analysis to various processes, including prescribing and running intravenous drug infusions [Ap04] and drawing lab samples [WRY06]. The approach to performing an FMEA in a healthcare setting is relatively constant regardless of the process being analyzed. The authors map the process of interest and analyze each step for severity of failure, likelihood of failure, and the ability of the system to detect the failure before harm occurs. They use these measures to locate risky steps in the process to target for corrective action. However, besides issues with the underlying assumptions in FMEA, highlighted in [Le11], healthcare is faced with another issue in using FMEA for risk assessments. These analyses in healthcare look only at one process at a time. In the Intensive Care Unit, for example, there can be many processes occurring simultaneously – medications are being ordered and given, labs are being drawn and sent for analysis, images are being taken, and invasive procedures are being performed. To do an analysis of an ICU as a whole using FMEA would involve performing a separate analysis on each of these processes. However, even this would provide an incomplete assessment of the risk of care delivery in this unit. Each FMEA assumes that the process is carried out independent of every other process. These processes though are all a part of the larger ICU system and have the potential to interact. This creates a compelling case for using a

hazard analysis technique that can look at the system as a whole. This work seeks to use the Systems-Theoretic Process Analysis as a way of integrating these multiple processes into one hazard analysis.

2 Methods

This work was all performed in the setting of an 8-bed Medical Intensive Care Unit at a large academic medical center. The Institutional Review Board of the Beth Israel Deaconess Medical Center granted this work approval as protocol #2014P-000020. Data collection was completed by observation of the workflows of nurses and physicians during weekday hours as well as through unstructured interviews. These observations were used to perform STPA as outlined in [Le11].

3 Selected Results

3.1 Identifying Accidents and Hazards

The major hazards for an intensive care unit include not treating the patient appropriately for their diagnosis or giving them an iatrogenic injury or infection that was not present when they were admitted. These correspond to the possible accidents, which include death or disability due to the disease, the treatment, or an acquired injury or infection.

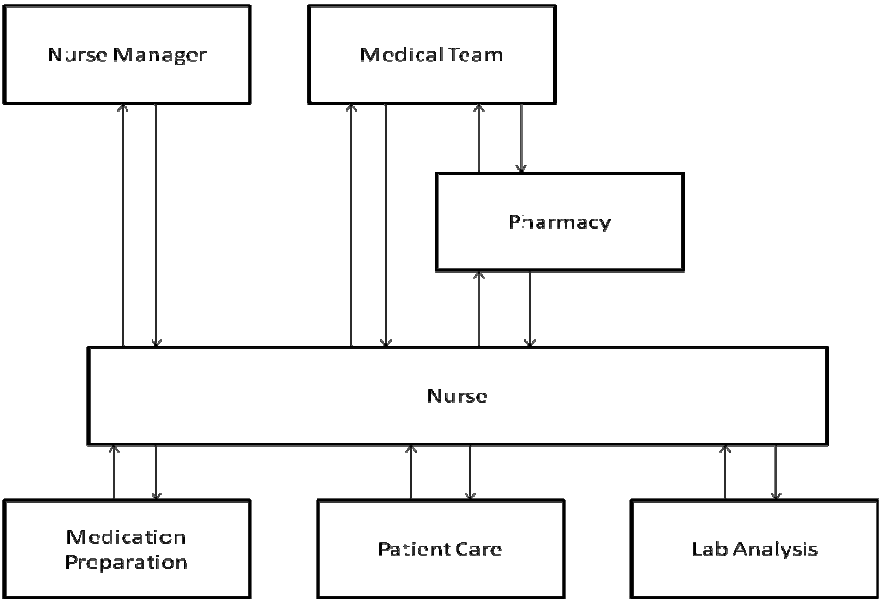


Figure 1: Simplified (Top-Level) Hierarchical Control Structure for ICU

3.2 Step 1: Unsafe Control Actions

The nurse (RN) has nine unique control actions from five different processes: lab testing, medication administration, patient positioning and other direct patient care, infection prevention, and patient monitoring. Using Step 1 tables 25 unsafe control actions were identified from these 9 control actions.

3.3 Step 2: Causal Factors

One interesting unsafe control action was the nurse mislabeling a lab specimen, either with the wrong patient identifiers or the wrong details about the tests ordered. In this ICU the process was to draw the sample and then print a label at a central printer which was affixed to the tube and sent to the lab. A step 2 scenario was the delayed operation of printing and labeling the specimen leading to the nurse forgetting potentially which patient the specimen came from. This could occur very easily in drawing a drug trough level, which is drawn just before the next dose of that medication is administered. A common scenario involves the nurse drawing the lab specimen, putting it down, and administering the drug. Administering the drug can take several minutes during which the nurse may not recall that the specimen needs to be labeled, leading to the potential for mislabeling the test tube. This is an example where a safe control action, administering the medication, can contribute to an unsafe control action, mislabeling or not labeling the lab specimen.

4 Discussion

The scenario where one process interacts in a risky manner with another process is not an uncommon scenario in a complex setting like an ICU. The FMEA completed on lab draws identified a failure mode of not putting the time on the label with a root cause listed as „time of collection not recorded“ [WRY06]. Because of this as the root cause the recommended control was to add more training. This analysis looked at blood draws as an independent process, so they never identified the clash found with STPA, which as a causal scenario would lead to implementation of very different controls to promote patient safety. This ability to handle multiple processes in one analysis is an important strength of STPA because these interactions are clinically important. STPA performs the analysis at the level of the controller, not at the level of the step in the process, allowing this flexibility in analysis and consideration of these interactions.

References

[Ap04] Apkon, M. et. al.: Design of a Safer Approach to Intravenous Drug Infusions: Failure Modes Effects Analysis. Quality and Safety in Health Care, 2004; S. 265-271.

[Le11] Leveson, N.: Engineering a Safer World, MIT Press, 2011.

[WRY06] Wagar, E.A., Raz, M., Yasin, B.: Patient Safety Partnerships Projects in the Clinical Laboratory. American Journal of Clinical Pathology, 2006; S. S53-S60.

Developing systematic procedures to identify causal factors in Systems Theoretic Process Analysis

John Thomas

Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139

Abstract: Although recent work has developed systematic procedures and formal methods to help identify unsafe control actions for Systems-Theoretic Process Analysis (STPA) Step 1, the identification of causal factors in STPA Step 2 has received less attention. This research develops additional guidance and procedures that can be used to identify causal factors and build scenarios in STPA Step 2.

1. Introduction

Systems-Theoretic Process Analysis (STPA) is a new hazard analysis technique that addresses many growing causes of accidents including requirements flaws, design errors, complex human behavior, and dysfunctional component interactions as well as traditional component failures [Le12]. Systematic procedures have been defined to identify unsafe control actions in STPA, but the current process for identifying causal factors and scenarios in STPA Step 2 has been less structured and guided mainly by a generic causal factor diagram. Although this process has proven useful in existing applications to date, the results often include substantial repetition and overlap in terms of the identified causal factors. These results indicate that there is an opportunity for improvement if a structured process can be defined to reduce repetition in the analysis while improving consistency among users and ensuring that no factors are overlooked.

2. Case Study

In this research, new procedures are developed and applied to an automotive shift-by-wire example. The shift-by-wire system replaces the traditional shift lever and mechanical cable with electronic actuators and a shift computer. Most manufacturers are now providing or developing similar systems that also utilize the shift computer to incorporate additional functionality and logic to either prevent certain operator shift commands or to automatically issue additional shift commands.

Previous work [Su14] has already identified unsafe control actions and safety constraints for this system using STPA Step 1. STPA Step 2 has also been performed on this system using the traditional approach, although only for a limited scope of the system. In this research, new Step 2 procedures are developed and applied to the same system to enable evaluations of the effectiveness and efficiency of the approaches.

3. Developing a structured process

The proposed new process begins with the safety constraints produced by STPA Step 1. It is hypothesized that beginning with the safety constraints may lead to a more holistic process than beginning separately with unsafe control actions and independently considering actions not followed. However, both approaches are evaluated and considered in this research.

The new procedures structure the STPA Step 2 analysis by first identifying the conditions in the safety constraints from STPA Step 1. For example, an identified safety constraint could be "Train doors must not be opened while train is moving". The two embedded conditions are "train doors are open" and "train is moving". Many of the causal factors in STPA Step 2 will relate to these or similar conditions identified in other safety constraints. For example, process model flaws, inadequate feedback, control algorithm flaws, and the ultimate behavior of the controlled process all relate to the conditions in the safety constraints that must be enforced. The current overlap and inefficiency in STPA Step 2 results may be due to the fact that a single causal factor such as missing position feedback can affect many safety constraints (and unsafe control actions) that reference the same conditions in the system. For example, incorrect feedback that the train is stopped when in fact it is moving could lead to a violation of many safety constraints that all involve the condition "train is moving". By identifying these conditions and using them to structure the Step 2 analysis, there is an opportunity to potentially reduce or avoid repetition while providing additional guidance and procedures for this part of the analysis. There is also a potential to improve the way Step 2 results are documented, which currently involves significant repetition as many of the same or similar causes are listed separately for every unsafe control action.

References

[Le12] Leveson, N.: Engineering a Safer World, MIT Press, 2012

[Su14] Sundaram, P.; Vernacchia, M.; Thomas, J.; Placke, S.: Application of STPA to a Shift by Wire System, MIT STAMP Workshop, 2014

Investigating Safety and Cybersecurity Design Tradespace for Manned-Unmanned Aerial Systems Integration Using Systems Theoretic Process Analysis

Kip Johnson[1], Nancy Leveson[2]

Department of Aeronautics and Astronautics
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, Massachusetts 02139
1. Graduate Student, johnskip@mit.edu
2. Professor, leveson@mit.edu

Abstract: Safety and cybersecurity are two emergent behaviors critical to mission success of Unmanned Aerial Vehicles (UAVs). This paper presents the concepts driving the initial research investigation of a hazard and vulnerability analysis for the integration of manned-unmanned aerial systems (UAS) using System-Theoretic Process Analysis (STPA). The motivation for this analysis is the need to engineer safe and secure UAS integration into the National Airspace System, or any system where manned and UASs will operate together. In contrast to more traditional safety and vulnerability assessments, this research aims to develop systems engineering methods and processes that are beneficial for initial design space exploration and developing system design requirements from the beginning of systems engineering phases. It is expected that research results will provide system architecture and design engineers the system framework and requirements necessary to engineer safe and secure integrated flight operations.

1 Introduction

The integration of manned-unmanned aerial systems is of critical and timely importance, especially for UAV integration into the United States (U.S.) National Airspace System (NAS). Interest and pressure is mounting for access to the NAS, and the Federal Aviation Administration (FAA) is responding. The FAA is under a U.S. congressional mandated deadline for UAV integration, due by the end of September 2015. [USC12] As highlighted by the FAA's 2013 Civil UAS Roadmap,

> "Ultimately, UAS must be integrated into the NAS without reducing existing capacity, decreasing safety, negatively impacting current operators, or increasing the risk to airspace users or persons and property on the ground any more than the integration of comparable new and novel technologies." [FAA13a]

The motivation for this research is to develop UAS integration system-level safety and security requirements to help guide and intellectually manage the engineering and

development of this large complex sociotechnical system (CSS). It is imperative to have methods for bringing safety and cybersecurity to the mission concept and requirements development systems engineering stages as the mission is often one and the same with safety and security. Integrating manned-UAS operations is one of these safe and secure essential missions.

2 Systems Engineering Methods

2.1 Background

Traditional methods are being used assure safety for UAS integration in the NAS, based on reliability theory and underlying chain of events accident models. The Radio Technical Commission for Aeronautics (RTCA) steering committee (SC) 228 was charged by the FAA to develop the Minimum Operational Performance Standards for UASs, specifically Detect and Avoid (DAA) and Command and Control (C2) technologies. Based on the 2013 Sense and Avoid report, the community concluded that a risk ratio method would provide the "most effective" Target Level of Safety. [FAA13b] The C2 safety standards are in development and currently are envisioned to rely on a probability concept, the required communications performance metric. [RTC14]

This approach alone, we argue, is not the best suited for this large CSS. There is little to no historical flight integration data, human interaction is integral, and the system is heavily software dependent—all challenges for traditional safety and hazard analyses.

The safety effort is further complicated by the shear enormity and complexity of the task and the ability of decision makers to adequately comprehend the system, let alone define the safety requirements driving the system's regulatory and technical design efforts. [Di13] [GAO12] Efforts across the globe to integrate UAV operations began in the late 1990s and early 2000s within government, academia, and industry stakeholders. [De04] Accommodation of UAVs for the Next Generation transportation system was enacted into U.S. law through the Vision 100-Century of Aviation Reauthorization Act, 2003. [El12] The efforts continue today in 2014, over a decade after they began. Designing for safe and secure UAV integration operations is a difficult and challenging endeavor.

2.2 The Case for a Systems Approach to Design

A systems approach is fundamentally about two paired concepts. [Ch93] First is the concept of systems hierarchy and emergence. Safety and cybersecurity are emergent properties that cannot adequately be handled through decomposition. Second, there needs to be communication and control between the system levels. The idea of control is the foundation for recent developments in systems safety concepts, more specifically STAMP (Systems-Theoretic Accident Model and Processes). [Le11] Safety can be analyzed as a control problem, and it is hypothesized that the STPA (System-Theoretic

Process Analysis) method can be applied and extended to the design and development of safe and cybersecure integrated manned-UAS flight operations.

3 A New Approach for Qualitative Systems Analysis in Design

As conceptualized by STAMP, safety is an emergent or a systems level behavior. The analysis of such requires a systems approach, one that analyzes component interactions. The application of STAMP-STPA to another emergent property, cybersecurity, is currently being developed by researchers at the Massachusetts Institute of Technology. [YoLe14] For UAS integration, both safety and cybersecurity are both critical emergent properties.

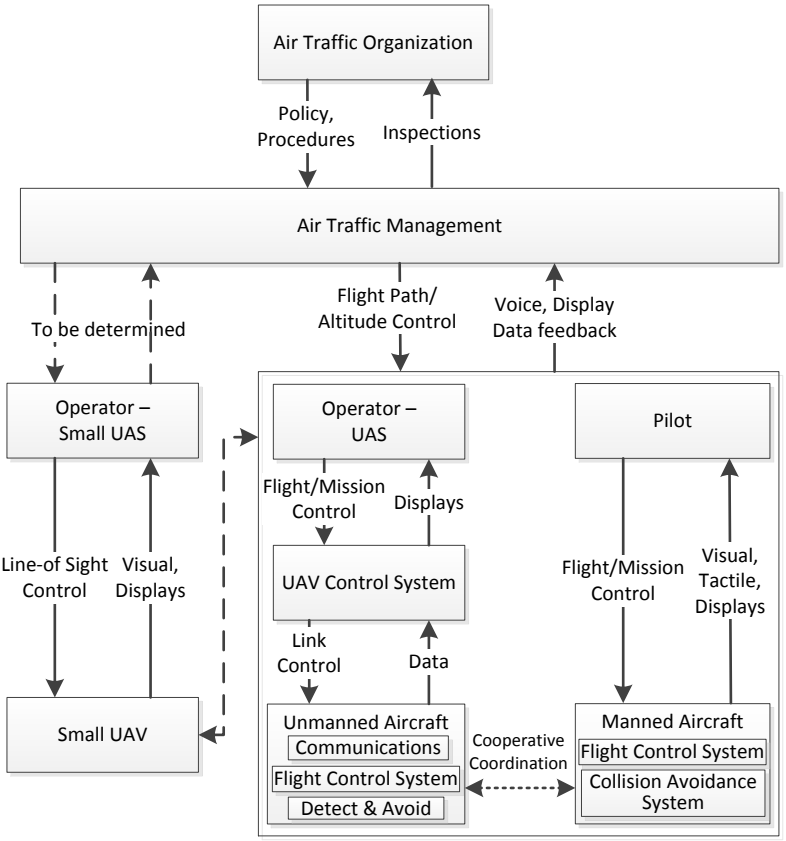


Figure 1: Representative Manned-Unmanned Aerial System Safety Control Structure

Figure 1 is a representative STAMP safety control structure modeling the manned-UAS integration system for flight operations in a controlled environment. The dotted lines represent current unknown controls, feedback, and information coordination. The Detect

and Avoid, and C2 systems are the perspectives of interest, but not independent from the more complete complex sociotechnical system represented in Figure 1.

Inspired by multidisciplinary systems design optimization methods, we are investigating the design space, potential tradespace, and concept of a potential qualitative Pareto tradeoff between safety and cybersecurity for manned-UAS integrated systems. A design is optimal, or on the Pareto frontier, if improving one design objective must diminish the other. [PaWi00] Demonstrating true Pareto optimality is not the goal here, nor is it possible. Rather, what can be learned about the design space, potential tradespace, and relationships during an STPA-Safety Driven Design analysis in pursuit of a qualitative Pareto?

The research aims to answer what is the qualitative system design tradespace for safety and cybersecurity, and can the concept of an approximate qualitative Pareto tradeoff be demonstrated on manned-UAS integration to benefit system design? The authors will investigate framework and methodology to characterize the design space using STPA for the hazard and vulnerability systems engineering analysis.

4 Conclusions

The accompanying presentation at the Second European STAMP Workshop (ESW2014) will present the initial hazard and vulnerability STPA analyses conducted on manned-UAS integration enabling technologies and concepts, both UAS DAA and C2. Research findings on the development of multi-objective systems engineering design methods to analyze and evaluate system emergent properties will also be discussed.

It is hypothesized that the application and extension of STPA will be successful for both safety and cybersecurity alone. Systems theoretic design methods and processes to characterize the bi-objective design space should have immediate impact on the development of requirements and standards for engineering UAV integration complex sociotechnical systems. In addition, the research more broadly develops STPA-SDD (Safety-Driven Design) by integrating safety and security requirements earlier into the systems engineering design and analysis processes. [Le11]

Unmanned aircraft integration is a complex and difficult endeavor. A systems-theoretic approach to the design and analysis of manned-UAS integration systems should prove beneficial to the realization of safe and secure integrated flight operations.

Acknowledgments

This work is sponsored by the Department of the Air Force under Contract #FA8721-05-C-0002. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Massachusetts Institute of Technology Lincoln Laboratory, United States Air Force, Department of Defense, or the U.S. Government.

References

[USC12] United States Congress, FAA Modernization and Reform Act of 2012. 2012.

[FAA13a] Federal Aviation Administration, “Integration of Civil Unmanned Aircraft Systems (UAS) in the National Airspace System (NAS) Roadmap,” 2013.

[FAA13b] Federal Aviation Administration, “Sense and Avoid (SAA) for Unmanned Aircraft Systems (UAS). Second Caucus Workshop Report,” 2013.

[RTC14] RTCA SC-228, “Command and Control (C2) Data Link White Paper,” 2014.

[Di13] G. L. Dillingham, “Government Accountability Office Report to Congress-Unmanned Aircraft Systems: Continued Coordination, Operational Data, and Performance Standards Needed to Guide Research and Development,” 2013.

[GAO12] U.S. Government Accountability Office, “Report to Congress - Unmanned Aircraft Systems: Measuring Progress and Addressing Potential Privacy Concerns Would Facilitate Integration into the National Airspace System,” 2012.

[De04] M. T. DeGarmo, “Issues Concerning Integration of Unmanned Aerial Vehicles in Civil Airspace,” 2004.

[El12] B. Elias, “Pilotless Drones: Background and Considerations for Congress Regarding Unmanned Aircraft Operations in the National Airspace System,” 2012.

[Ch93] P. Checkland, Systems Thinking, Systems Practice. Chichester: John Wiley & Sons, Inc., 1993.

[Le11] N. Leveson, Engineering a Safer World: Systems Thinking Applied to Safety. Cambridge: The MIT Press, 2011.

[YoLe14] W. Young and N. G. Leveson, “An integrated approach to safety and security based on systems theory,” Commun. ACM, vol. 57, no. 2, pp. 31–35, Feb. 2014.

[PaWi00] P. Y. Papalambros and D. J. Wilde, Principles of Optimal Design: Modeling and Computation, Second. Cambridge: Cambridge University Press, 2000.

Use of STPA in digital instrumentation and control systems of nuclear power plants

Rejzek Martin, Christian Hilbes

Institute of Applied Mathematics and Physics
Zurich University of Applied Sciences
Technikumstrasse 9
8401 Winterthur, Switzerland
martin.rejzek@zhaw.ch
christian.hilbes@zhaw.ch

Abstract: Nuclear power plant operators increasingly face the task of replacing their instrumentation and control systems with modern (digital) systems. In this work the "System Theoretic Process Analysis" (STPA) risk analysis method was adapted and amended to enable it to be used in digital instrumentation and control systems.

1 Background Information

Nuclear power plant operators increasingly face the task of replacing their instrumentation and control (I&C) systems with modern systems to ensure their availability, reliability and safety in the future as well. Replacement of these systems typically features simultaneous transition from primarily analog systems to software-based, digital systems.

The "System Theoretic Process Analysis" (STPA) risk analysis method specifically investigates risks which are generated by functional interaction between the control units present in the system as well as risks caused by component failure [Le11]. As a result, STPA is suitable for analysis of software-based and dynamic systems for which it is indeed typical that system failures occur without actual component failure. Modern digital I&C systems belong to this category of systems.

2 Applying STPA to digital insturmentation and control systems

In collaboration with swissnuclear and the Gösgen nuclear power plant, the STPA method was adapted and amended to enable it to be used in digital I&C systems. The actual implementation was demonstrated and discussed on the basis of a case study. Among others, the following aspects formed the focus of the work:

Representation of the system as a hierarchical control structure is a basic prerequisite for carrying out STPA analysis. One of the first questions is therefore how a hierarchical

STPA control structure can be developed efficiently for I&C systems, and how the specific properties can be handled (for example intensive use of the redundancy safety pattern).

While STPA Step 1 is meant to be predominantly based on a functional point of view of the system, STPA Step 2 can be used to perform an analysis based on both a functional and a physical representation of the system. This shift in point of view is explicitely implemented in our procedure, allowing for an easier integration of classical, component based analysis results into the overall effort.

3 Conclusions

STPA is one of several methods which can be used for analysis of nuclear power plant systems. Optimum benefit is generated when the various methods can be combined in suitable fashion. The adapted and amended process of the STPA method was thus designed to allow interfaces to other methods to be realized and, for example, to enable the causes of hazards which have already been established during the course of fault tree analyses to be incorporated in STPA.

References

[Le11] Leveson, N.: Engineering a safer world: Systems thinking applied to safety; MIT Press, 2011.

Future Search Engines - Big Data und semantische Technologien in Backend und Frontend von Suchmaschinen

(FoRESEE)

FoRESEE: Future Search Engines 2014*

Abstract: The workshop 'FoRESEE 2014: Big Data and Semantic Technologies in Backend and Frontend of Search Engines' aims at analyzing and integrating vast amounts of data and semantic technologies in search engines. This workshop will discuss state-of-the-art approaches and methodologies regarding industrial as well as research interests.

1 Focus: Big Data and Search (Backend)

Search Engines are a prime example of Big Data technology. In addition to processing huge amounts of data, their storage, analysis and management are central issues of Big Data.

Various crawling, mining, linking and integration methods form the basis of every modern search technology. Nowadays, search engines exceed classic keyword-based document retrieval by leveraging semantic technologies such as question answering, user context inception and conversational interaction. Furthermore, semantic datasets such as the Linked Open Data Cloud are freely available, supporting knowledge acquisition at scale.

Within this workshop we will explore ontologies as novel way to structure search. Furthermore, we will discuss architectures of semantic search engines dealing with large amounts of domain-specific data.

2 Focus: Human-Semantic Interaction and Human-Computer Interaction (Frontend)

Visually well represented search results are an important factor for the success of a search engine. Next to presenting large amount of data to the user, we focus on the integration of semantic technologies in Web-Interfaces. Traditional non-semantic search engines demand different interfaces.

Data from knowledge bases, such as the Google Knowledge Graph, needs special representation. In particular, the semantic component of the query needs to be explicitly depicted. Semantic entities have to be conveyed in an adequate way by a smart interaction design. Resulting user data generated inside the browser can be intercepted and be

*https://foresee.unister.de/

further utilized to optimize the search interaction. Client-side data comprises mouse cursor movement, page visits and clicks on certain results. For instance, this data enables improving ranking results iteratively. Additionally, the increasing vast amount of data requires stream-based technologies to handle resulting information efficiently.

This branch of research is represented by two submissions in our workshop. Haag et al. present *Aspect Grid*, which is a novel visualization technique for results based on important aspects of the search domain. Both the aspects (e.g., "price" for restaurants) and the recognized sentiments (e.g., cheap, medium, expensive) are represented within a grid. The user can then filter results based on the presented information (e.g., only cheap restaurants). Keck et al. discuss novel visual representations of search results for specific domains such as travel and financial product search. Among others, they present an inspiration-based search interface that incorporates semantics through expressive images (e.g., "sunny" travel destinations); and an interface based on parallel coordinates for faceted product search within multi-dimensional data sets.

Head of Workshop: Dr. Andreas Both (Unister GmbH), Prof. Dr.-Ing. Martin Gaedke (Technische Universität Chemnitz), Prof. Dr.-Ing. habil. Rainer Groh (Technische Universität Dresden), Junior-Prof. Dr. Matthias Hagen (Bauhaus-Universität Weimar)

Organizer: Michael Röder, MSc (Universität Leipzig), Maximilian Speicher, MSc ETH Inf.-Ing. (Technische Universität Chemnitz), Ricardo Usbeck, MSc (Universität Leipzig)

Programme Committee: Prof. Dr. Sören Auer (Universität Bonn & Fraunhofer IAIS), Dr. Bernd Eickmann (Unister GmbH), Dr. Christiane Lemke (Unister GmbH), Dipl.-Bioinform. Martin Nettling (Martin-Luther-Universität Halle-Wittenberg), Dipl.-Wirt.-Inf. Mathias Priebe (Unister GmbH), Prof. Dr. Benno Stein (Bauhaus-Universität Weimar), Dipl.-Medieninf. Martin Voigt (Technische Universität Dresden), Dr. Matthias Wauer (Unister GmbH), Prof. Dr. Abraham Bernstein (Universität Zürich)

Aspect Grid: A Visualization for Iteratively Refining Aspect-Based Queries on Document Collections

Florian Haag, Qi Han, Markus John, Thomas Ertl

Institute for Visualization and Interactive Systems, University of Stuttgart
Universitätsstraße 38 · 70569 Stuttgart · Germany
{ florian.haag, qi.han, markus.john, thomas.ertl }@vis.uni-stuttgart.de

Abstract: Search results comprising of unstructured documents are traditionally displayed as plain lists, whereas the formulation of complex queries can be aided by comprehensive visualizations that highlight relevant attributes of the search results. We propose a visualization concept that incorporates a sentiment analysis of aspects mentioned in examined documents. Our concept provides a means of iteratively refining the query based upon preliminary search results. After an abstract description of the concept, we elaborate on its usage in a practical example of filtering customer reviews in order to choose a restaurant based on user-defined preferences.

1 Introduction

Traditionally, search engines expect an enumeration of features being looked for—such as keywords that must appear in a document—and results are displayed as a purely sequential list of references to the found documents. This generic way of searching is suitable for the web with its diverse kinds of resources. However, users who look for some particular information often require more specific queries and start using advanced operators [AKG10]. In addition, they are interested only in particular types of web resources—for instance, customers looking for a hotel offer wants to find only web resources that contain information on hotel rates, they do not want to find artistic images, phone book entries or recipes. Accordingly, search engines can offer means for defining complex queries and improve the overview of search results by taking advantage of the special traits of a given resource kind [Do08].

Often, such querying systems rely on structured information available along with the search results. This is particularly true in query systems for the Semantic Web [St12], but can also be observed with other systems, such as image search engines that examine metadata [Ye03]. While users can examine the search results to manually refine their query, a direct means of extracting further query restrictions from an intermediary result set and incorporating them into the query is often missing. This is complicated by the fact that documents of the same type, but from different data sources, may feature incompatible schemas of structured information, if any, leaving search engines with the unstructured textual part of the resource as the only guaranteed common feature. We propose Aspect

Grid, a novel approach that comprises of a natural language processing component extracting relevant information from documents, combined with a user interface that provides an overview of the active query criteria, as well as options to directly refine the query based upon the extracted information.

2 Related Work

Results are displayed in a way specific to their type by domain-specific systems, such as search engines specialized on summarizing news articles [Ra05], as well as by domain-aware multi-purpose systems that apply different output rules to a variety of search result types [Bo10]. There have been attempts to structure these results, by creating clusters of similar subsets [Hi11], by providing an overview of various aspects of each search result [Iw12], by embedding search results into other visualizations such as maps or timelines, and by combining several types of search results in a compound view [Bo11]. Moreover, the search results can be organized into nodes of a topic-map [MP13] or be directly displayed as a graph [RTB11].

The expressiveness of the queries for which search results are retrieved can be increased by using faceted approaches that apply individual filter criteria to different aspects of the dataset. As large amounts of data are available in the form of unstructured text documents, the language processing community has been researching methods for the automatical extraction of structured information from unstructured data for a long time [MRS08]. Among these, sentiment analysis has received much attention recently [Li12]. Its goals are the detection and extraction of subjective information from documents, especially reviews, using supervised or unsupervised methods [PL08, BE10]. Systems that summarize and visualize sentiment information from a set of documents have been successfully implemented [Ma11].

While there are various visualization concepts that provide an overview of the facets of search results, as well as of the information extracted from unstructured documents, few of them are combined with a visual representation of the active query. Again, techniques from the Semantic Web tend to support that kind of view by integrating the search results into a query graph [St12, He09]. Many others, however, display results merely in a separate area that has no direct graphical connection to the visualization of the query [Ru08]. Only few specialized systems incorporate a direct means of iteratively refining the query based on the current result set [Ko09], or deriving a more precise query from intermediary results [Ha06].

In order to allow for a high degree of expressivity and intuitively show the currently active query, search parameters and their mutual connections—such as the logical operators *and* and *or*—, the query expression can be visualized. A variety of approaches for query visualization has been proposed, based upon a multitude of paradigms such as Venn diagrams [Jo98], flowcharts [Sh94, Ha06], graphs [Ru08, St12], or custom visual elements [Sp93]. While they successfully indicate the active search parameters and are to some extent tightly integrated with the search results, they usually do not provide a direct

feedback channel for query refinement based on the results; the query has to be updated manually by the user after looking at the results.

3 The Aspect Grid

In this work, we present the Aspect Grid approach—a visualization technique for filtering collections of text documents that integrates search parameters and results with the possibility of refining the query based upon text excerpts from search results. Two sketches of different states of the visualization are presented in Figure 1. Based upon a collection

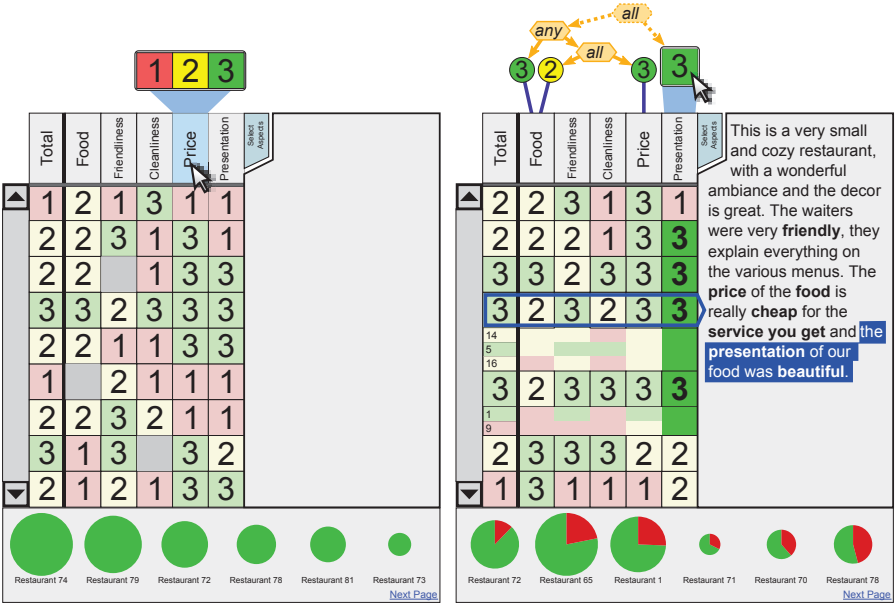


Figure 1: Two depictions of the Aspect Grid visualization being used for filtering restaurant reviews: Left: All reviews are displayed as no filters have been set up yet. The user is about to select a restriction for the aspect *Price*. Right: Several filter criteria have been set up and combined with *any* and *all* operators. The user has opened one review in the preview area and selected a statement in that review that positively mentions the *Presentation* aspect. An according filter is suggested, and as the mouse cursor hovers over it, any rows with suitable values are highlighted by intensifying the background color. The suggested filter is automatically linked with the *all* operator.

of text documents from a given domain, our approach first determines important aspects in the domain. While this is currently done by using a domain-specific prepared list of aspects, in the future we want to use more advanced techniques to automatically extract the important aspects [BE10]. Once knowing which aspects are relevant for the current domain, our system scans each document for mentions of any of the aspects. It then further calculates the sentiment scores expressed towards the detected aspects using a sentiment

lexicon-based method, as described by Liu [Li12]. We choose this method because it works without a domain-specific training dataset. We assign colors to sentiment scores to make it easier for users to navigate the results. Our sentiment color scores are one for negative (red), two for neutral (yellow), three for positive (green) and for no finding we use a empty fields (gray). When a domain-specific dataset annotated with aspects and sentiment scores is available, the accuracy of our approach can be further improved using supervised method.

Both the aspects and the recognized sentiments are reflected in the visualization. The central visual part of the Aspect Grid approach is a grid whose columns correspond with the various aspects for which statements were found in the texts. At first, each document that belongs to the general category of texts being examined occupies one row in the grid. Documents that were excluded by the filter criteria are clustered based on their aspects and reduced in size. The query to the underlying search engine is expressed graphically by depicting the restrictions for each aspect right above the respective column. In order to define different combinations (*all*, *any*) of aspect restrictions, these restrictions are visually linked among each other.

The third part of the visualization is an area that can be used to display the contents of single search results. As opposed to mere preview displays [Ru08], parts of the text in that area can be selected and used as a basis for additional query restrictions. The algorithm determines which aspects are most likely addressed in the selected text range, and what statements are made about them in the selected search result. It then suggests to add additional filters to obtain only documents that express the same sentiment about the mentioned aspects.

An area at the bottom of the visualization displays a pie chart indicating how many documents of the total set match the filter criteria. If the documents of the total set can be clearly divided into distinct subtopics, there will be several pie charts that can be compared to each other. Figure 1 is shown with the exemplary use case of filtering restaurant reviews, so the bottom area shows one pie chart per restaurant (in an actual implementation, restaurant names would be shown instead of placeholders). The radius gives a hint about the total relative number of reviews per restaurant, and the pie charts are first sorted by the percentage of reviews matching the filter criteria, then by the number of reviews.

4 Example of Application

In the following, we present a usage scenario that demonstrates the usefulness and applicability of Aspect Grid. Alice wants to find a suitable restaurant for seeing a friend. By entering the first few search terms such as *restaurant*, *recommendation* and the name of the city, the search engine recognizes that the user is interested only in restaurant reviews. Therefore, the Aspect Grid visualization is loaded with aspects related to restaurant reviews. The user starts to check some of the lines in the table and notices that some items need to be filtered out. She notes that it is important for her that the food has left a positive impression on the reviewers, or that reviewers had a neutral impression of the food, in

which case they should have judged the price as positive ("Either the food must be extraordinary, or if it isn't, it should at least not cost much."). She adds these three filters—the state right before adding the filter on the *Price* aspect is depicted in Figure 1 (left)—and connects them with the appropriate *any* and *all* operators. After selecting and reading some of the filtered reviews, she discovers a statement about the presentation of the food in one of the reviews and realizes that she would appreciate that aspect, as well. Therefore, she highlights the statement in the review text, whereupon a positive judgment of the *Presentation* aspect is automatically suggested as a new filter as depicted in Figure 1 (right). She adds the filter by clicking on the suggestion. The bottom area of the Aspect Grid display now shows a short list of restaurants, ordered by the ratio of reviews that matched her filters compared to the reviews that did not match her filters. Thus, based upon her custom aspect-related review filters, she picks one of the first few restaurants.

While this usage scenario shows a specific application example for Aspect Grid, we are confident that our approach is useful for iteratively refining filters on any collection of text documents from which different aspects can be extracted.

5 Conclusion and Future Work

We have presented an interactive visualization concept that allows for the filtering of documents based upon aspects that were extracted from unstructured text. Different from the related works, our approach incorporates the unstructured documents from preliminary result sets into the visualization in a way that they can be used as a basis for suggestions on additional filter restrictions. This way the user can build the desired filter more intuitively and efficiently and thereby achieves his or her intended goal faster.

We are planning to implement a prototype of the visualization and evaluate the usefulness of the approach in a user study. Possible extensions of the concept that may be recognized as necessary in user experiments include an enrichment of the data extracted from unstructured text with structured data, as well as an implicit training method for enhancing the aspect recognition precision in documents on the fly.

References

[AKG10] Aula, A.; Khan, R. M.; Guan, Z. How Does Search Behavior Change as Search Becomes More Difficult? In *Proc. CHI '10*, pages 35–44, New York, NY, USA, 2010. ACM.

[BE10] Brody, S.; Elhadad, N. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *NAACL HLT '10*, pages 804–812, Stroudsburg, PA, USA, 2010. ACL.

[Bo10] Bozzon, A. et al. Liquid Query: Multi-domain Exploratory Search on the Web. In *Proc. WWW '10*, pages 161–170, New York, NY, USA, 2010. ACM.

[Bo11] Bozzon, A. et al. Visualization of Multi-Domain Ranked Data. In *Search Computing*, volume 6585 of *LNCS*, pages 53–69. Springer, Berlin/Heidelberg, Germany, 2011.

[Do08] Dork, M. et al. VisGets: Coordinated Visualizations for Web-Based Information Exploration and Discovery. *IEEE TVCG*, 14(6):1205–1212, 2008.

[Ha06] Hansaki, T. et al. FindFlow: Visual Interface for Information Search Based on Intermediate Results. In *Proc. APVis '06*, pages 147–152, Darlinghurst, Australia, 2006. ACS.

[He09] Heim, P. et al. RelFinder: Revealing Relationships in RDF Knowledge Bases. In *Semantic Multimedia*, volume 5887 of *LNCS*, pages 182–187. Springer, Berlin/Heidelberg, Germany, 2009.

[Hi11] Hindle, A. et al. Clustering Web Video Search Results Based on Integration of Multiple Features. *World Wide Web*, 14(1):53–73, 2011.

[Iw12] Iwata, M. et al. AspecTiles: Tile-Based Visualization of Diversified Web Search Results. In *Proc. SIGIR '12*, pages 85–94, New York, NY, USA, 2012. ACM.

[Jo98] Jones, S. Graphical Query Specification and Dynamic Result Previews for a Digital Library. In *Proc. UIST '98*, pages 143–151, New York, NY, USA, 1998. ACM.

[Ko09] Koch, S. et al. Iterative Integration of Visual Insights During Patent Search and Analysis. In *VAST 2009*, pages 203–210, 2009.

[Li12] Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Ma11] Marcus, A. et al. Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration. In *Proc. CHI '11*, pages 227–236, New York, NY, USA, 2011. ACM.

[MP13] Mirylenka, D.; Passerini, A. ScienScan—An Efficient Visualization and Browsing Tool for Academic Search. In *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *LNCS*, pages 667–671. Springer, Berlin/Heidelberg, Germany, 2013.

[MRS08] Manning, C. D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*, volume 1. Cambridge University Press, Cambridge, UK, 2008.

[PL08] Pang, B.; Lee, L. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Ra05] Radev, D. et al. NewsInEssence: Summarizing Online News Topics. *Commun. ACM*, 48(10):95–98, 2005.

[RTB11] Rástočný, K.; Tvarožek, M.; Bieliková, M. Supporting Search Result Browsing and Exploration via Cluster-Based Views and Zoom-Based Navigation. In *Proc. WI-IAT '11*, volume 3, pages 297–300, 2011.

[Ru08] Russell, A. et al. NITELIGHT: A Graphical Tool for Semantic Query Construction. In *Proc. SWUI '08*, volume 543. CEUR-WS, 2008.

[Sh94] Shneiderman, B. Dynamic Queries for Visual Information Seeking. *IEEE Software*, 11(6):70–77, 1994.

[Sp93] Spoerri, A. InfoCrystal: A Visual Tool for Information Retrieval & Management. In *Proc. CIKM '93*, pages 11–20, New York, NY, USA, 1993. ACM.

[St12] Stab, C. et al. Semantics Visualization for Fostering Search Result Comprehension. In *The Semantic Web: Research and Applications*, volume 7295 of *LNCS*, pages 633–646. Springer, Berlin/Heidelberg, Germany, 2012.

[Ye03] Yee, K.-P. et al. Faceted Metadata for Image Search and Browsing. In *Proc. CHI '03*, pages 401–408, New York, NY, USA, 2003. ACM.

Architecture and applications of the Parallel Distributed Image Search Engine (ParaDISE)

Roger Schaer, Dimitrios Markonis, Henning Müller

HES–SO, Business Information Systems,
TechnoArk 3 3960 Sierre, Switzerland
roger.schaer@hevs.ch
dimitrios.markonis@hevs.ch
henning.mueller@hevs.ch

Abstract: Medical images are important to physicians for diagnosis and treatment planning. The number of images produced in medical institutions bas been increasing rapidly over the past years. Making these imaging data available and allowing medical professionals to perform retrieval based on visual characteristics of images is the challenge that content–based image retrieval systems are tackling. In this paper we propose a system called ParaDISE (Parallel Distributed Image Search Engine). It allows indexing and retrieving images using both visual features as well as text. It is designed to scale well with big data sets, and it should be easily extensible with new features. The architecture of the system is described and its use in medical applications is presented. An open source release of the system is in progress.

1 Introduction

Medical images are a precious tool in aiding doctors to diagnose and treat a variety of diseases. [Doi05]. The amount of images produced by hospitals has been growing at an almost exponential rate over the past 20 years (as an example, the Radiology Department of the University Hospitals of Geneva produced over 200'000 images per day in 2011).

Accessing this huge amount of data is a challenge that CBIR (Content–based image retrieval) systems attempt to deal with, by allowing users to search for images based solely on their visual content (example images or image regions), without requiring the input of textual keywords. Studies show that CBIR systems can assist radiologists in diagnosis, particularly radiologists with little experience [ABWM$^+$03, LADT09]. Novel image retrieval algorithms have been published [LADT09, MMBG04].

In recent years, retrieval systems have become increasingly complex, improving text search by using semantics, as well as utilizing modern visual representations such as Bag–of–Visual–Words [SZ03] and Vectors of Locally Aggregated Descriptors (VLAD) [JDS10]. Additionally, multimodal search (text combined with CBIR) is becoming more common, driving the need for interoperability and expandability in retrieval systems.

This paper presents the Parallel Distributed Image Search Engine (ParaDISE), an image

retrieval system that combines text with CBIR techniques. The system was designed to be scalable, meaning it should be able to handle ever–growing datasets and large image collections that are available on the Internet. It was built in a modular fashion, consisting of several independent components, thus increasing the flexibility and distributability of the system. The third main aspect of ParaDISE is interoperability, which is achieved by the use of standard protocols and data formats in order to facilitate the integration of ParaDISE with other retrieval systems.

The main driving force behind the development of ParaDISE is KHRESMOI[1], a European project focused on medical information analysis and retrieval [HBGM11]. The goal of KHRESMOI is to develop a multilingual, multimodal search system for biomedical information and documents.

2 Related Work

Several projects have already been realized in the field of information retrieval in the medical domain and made systems available as open source. Among them is the Viper project [SMM+01], the outcome of which was the GNU Image–Finding Tool (GIFT), a CBIR system that enables users to perform "Query By Example" search operations and improve the quality of results using relevance feedback. ParaDISE uses more recent and sophisticated features than GIFT, which only uses visual information for querying and relevance feedback.

Another noteworthy project is Lire (Lucene Image Retrieval) [LC08], a library based on the Lucene text retrieval software, which provides image retrieval using various visual features related to color and texture. The system uses purely visual search and provides little support for several state–of–the–art representations (such as spatial pyramid matching or bag–of–colors), indexing parallelization or flexible index structuring.

Flexible Image Retrieval Engine (FIRE) is yet another example of a CBIR system [DKN08], also used in medical image retrieval evaluation apart from other applications. Being developed before 2007, the system does not support state–of–the–art mid–level representations (such as vectors of locally aggregated descriptors). No parallelization schema is mentioned for indexing large scale datasets, either.

Additionally, there are also several web applications available for medical image retrieval, which use various retrieval techniques, including CBIR and semantic data. GoldMiner[2] is a text–based radiology image search engine which allows filtering results based on imaging modality. RadMiner[3] is a search tool in radiology reports using semantic information. Yottalook[4] is a web–based radiology image search system which supports searching across multiple types of sources. These three systems do not currently use any visual information for searching. OpenI [DFAST12] allows search based on text and on image examples.

[1] http://khresmoi.eu (accessed on 19-06-2014)
[2] http://goldminer.arrs.org (accessed on 19-06-2014)
[3] http://www.averbis.de/public/download/RadMiner_brochure_en.pdf (accessed on 19-06-2014)
[4] http://yottalook.com (accessed on 19-06-2014)

ParaDISE is geared towards a higher level of abstraction, allowing users to configure and expand every aspect of the system (search, storage, indexing, fusion, etc.). It contains many state–of–the–art descriptors and allows for parallelization in indexing and retrieval.

3 System Description

Backend The various features provided by the ParaDISE backend are distributed among four main components, shown below: The **Extractor** extracts local visual features from an image, such as Scale Invariant Feature Transform (SIFT) [Low04]. The **Descriptor** creates a vector that serves as a global representation of an image, such as Bag–of–Visual–Words. The **Storer** accesses visual indices (read/write operations). ParaDISE supports several systems to store and access the images' visual descriptors, from simple comma–separated values (CSV) files to database systems such as MySQL or Apache CouchDB. The **Fusor** combines the results of multiple query images, or multiple features of the same image. It is also designed for integrating ParaDISE with other systems.

These four basic components are integrated into two composite components (Indexer and Seeker) to provide the two main retrieval features: indexing and similarity search. The Indexer can perform serial as well as parallel indexing using Apache Hadoop [MSE+12], allowing the distribution of the indexing workload across several computers and thus reducing the overall time needed to perform the task, as well as making it easier to scale up to bigger datasets (by simply adding nodes to the Hadoop cluster). It supports building Approximate Nearest Neighbor (ANN) search indices and weighting the indices. The Seeker can use various distance and similarity metrics for the comparison of the vectors. It also supports multiple queries and relevance feedback techniques. Additionally, the Seeker takes advantage of the ANN indices created by the Indexer, helping to keep retrieval time short in spite of very large datasets (by restricting the search to a subset of images).

ParaDISE has been used in the ImageCLEF [5] 2013 challenge, evaluating image retrieval performance on a dataset containing images from the biomedical literature [GSdHMS+13].

Frontend The main features of ParaDISE are exposed in a web service layer built on top of the ParaDISE backend. The web services are based on the principles of the Representational state transfer (REST) architectural style for web applications. They use standard Hypertext Transfer Protocol (HTTP) requests, making interoperability and integration with client applications (both web–based and desktop clients) easily achievable. The JavaScript Object Notation (JSON) data exchange format is used for communication. Additional features such as the text search are also provided by the web service layer. The architecture of the web services is explained below and shown in Figure 1. The **Caption web service** is responsible for search by keyword and uses the Lucene search engine. The **Visual web service** uses the two composite components from the ParaDISE backend (Indexer & Seeker) to provide content–based image retrieval and indexing. The **Fusion web service** is responsible for combining results from different sources (text search, image search) into

[5] http://www.imageclef.org/ (accessed on 19-06-2014)

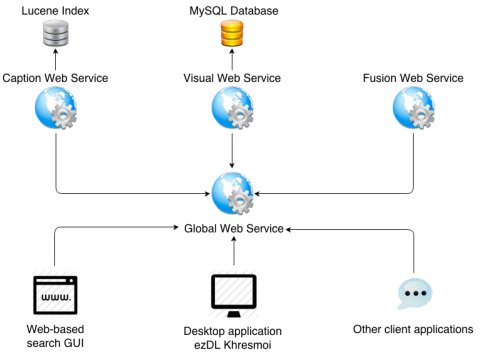


Figure 1: Overview of the ParaDISE web service layer architecture.

a single list (based on a given fusion rule), using the Fusor component from the ParaDISE backend. The **Global web service** acts as a facade for client applications. It hides the complexity of calling the underlying web services, for example by providing a general "search images" method, which can use text, visual information, or both.

4 Applications

KHRESMOI ParaDISE is used in the KHRESMOI subsystem that accesses images in the open access medical literature. This part of KHRESMOI supports text queries, queries by image example and the combination of these two for accessing the image data. It allows filtering by modality (by applying automatic modality classification methods) and is interconnected with other tools, such as the semantic search and 3D image retrieval components of Khresmoi. It supports searching into subfigures of multi–panel figures with the help of automatic compound figure detection and separation methods [CMFR+13]. ParaDISE is used as the backbone of this subsystem. Indexing a dataset of 1.7 million PubMed Central images with the parallel computing mode of ParaDISE took approximately 6 hours.

Shambala and Leap Motion™ Whereas ezDL, used in the Khresmoi system, is a comprehensive interface, Shambala is an alternative search interface mainly targeted towards simple navigation. It is built on 3 principles : it should be easy to use (not overwhelming the user with options and parameters), interactive (allow the user to drag&drop images, update search results in real–time) and modern (use of HTML5 and Leap Motion™ technology[6]). Leap Motion allows the user to interact with the interface using only hand and finger gestures, without needing to use a mouse or keyboard. More details about gesture interaction with Shambala can be found in [WSMM14].

[6]http://leapmotion.com (accessed on 19-06-2014)

SearchParadise and Google Glass™ Another application prototype which uses ParaDISE, called SearchParadise, was created for the Google Glass™, a wearable computing device developed by Google[7]. The application allows the user to perform a search query based on a photo taken by the Glass camera and see the results on the display located in front of the right eye. This can allow physicians to get additional information on situations or lesions that they have and does not interrupt the contact with the patient.

5 Conclusions

Medical image retrieval can potentially benefit from novel techniques such as CBIR. Taking into account the characteristics of the domain, CBIR systems need to provide scalable and interoperable solutions. The ability to evaluate different image representations and to add new visual features needs to be supported, as well. ParaDISE, the image retrieval system described in this paper, provides a flexible solution for image indexing and similarity search operations on large datasets. It can be easily integrated with other systems and extended with new features. It can be used for very diverse client applications, including desktop applications, web–based interfaces as well as emerging technologies such as the Google Glass™. Future steps for the system include using a distributed computing framework for online operations (retrieval) that is currently only implemented for the offline indexing steps. Releasing the source code to the public is currently under way to encourage the addition of new features by third parties.

6 Acknowledgments

This work was supported by the EU 7th Framework Program in the context of the Khresmoi project (FP7 grant 257528).

References

[ABWM+03] Alex M. Aisen, Lynn S. Broderick, Helen Winer-Muram, Carla E. Brodley, Avinash C. Kak, Christina Pavlopoulou, Jennifer Dy, Chi-Ren Shyu, and Alan Marchiori. Automated Storage and Retrieval of Thin–Section CT Images to Assist Diagnosis: System Description and Preliminary Assessment. *Radiology*, 228(1):265–270, July 2003.

[CMFR+13] Ajad Chhatkuli, Dimitrios Markonis, Antonio Foncubierta-Rodríguez, Fabrice Meriaudeau, and Henning Müller. Separating compound figures in journal articles to allow for subfigure classification. In *SPIE Medical Imaging*, 2013.

[DFAST12] Dina Demner-Fushman, Sameer Antani, Matthew S. Simpson, and George R. Thoma. Design and Development of a Multimodal Biomedical Information Re-

[7] http://www.google.com/glass (accessed on 19-06-2014)

trieval System. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

[DKN08] Thomas Deselaers, Daniel Keysers, and Hermann Ney. Features for image retrieval: an experimental comparison. *Information Retrieval*, 11(2):77–107, April 2008.

[Doi05] K Doi. Current status and future potential of computer–aided diagnosis in medical imaging. *British Journal of Radiology*, 78:3–19, 2005.

[GSdHMS+13] Alba García Seco de Herrera, Dimitrios Markonis, Roger Schaer, Ivan Eggel, and Henning Müller. The medGIFT Group in ImageCLEFmed 2013. In *Working Notes of CLEF 2013 (Cross Language Evaluation Forum)*, September 2013.

[HBGM11] Allan Hanbury, Célia Boyer, Manfred Gschwandtner, and Henning Müller. KHRESMOI: Towards a Multi–Lingual Search and Access System for Biomedical Information. In *Med–e–Tel, Luxembourg, 2011*, pages 412–416, 2011.

[JDS10] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3304 – 3311, June 2010.

[LADT09] L. Rodney Long, Sameer Antani, Thomas M. Deserno, and George R. Thoma. Content-Based Image Retrieval in Medicine: Retrospective Assessment, State of the Art, and Future Directions. *International Journal of Healthcare Information Systems and Informatics*, 4(1):1–16, January 2009.

[LC08] Mathias Lux and Savvas A. Chatzichristofis. Lire: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1085–1088, October 2008.

[Low04] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[MMBG04] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content–based image retrieval systems in medicine–clinical benefits and future directions. *International Journal of Medical Informatics*, 73(1):1–23, 2004.

[MSE+12] Dimitrios Markonis, Roger Schaer, Ivan Eggel, Henning Müller, and Adrien Depeursinge. Using MapReduce for Large–scale Medical Image Analysis. In *2nd IEEE Conference on Healthcare Informatics, Imaging and Systems Biology (HISB)*, September 2012.

[SMM+01] David McG. Squire, Henning Müller, Wolfgang Müller, Stéphane Marchand-Maillet, and Thierry Pun. Design and Evaluation of a content–based image retrieval system. In *Design & Management of Multimedia Information Systems: Opportunities & Challenges*, chapter 7, pages 125–151. Idea Group Publishing, London, 2001.

[SZ03] Josef Sivic and Andrew Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 1470–1477, Washington, DC, USA, 2003. IEEE Computer Society.

[WSMM14] Antoine Widmer, Roger Schaer, Dimitrios Markonis, and Henning Müller. Gesture interaction for content–based medical image retrieval. In *ICMR*, 2014.

Search Ontology, a new approach towards Semantic Search

Alexandr Uciteli[1] Christoph Goller[2] Patryk Burek[1]
Sebastian Siemoleit[1] Breno Faria[2] Halyna Galanzina[2] Timo Weiland[3]
Doreen Drechsler-Hake[3] Wolfram Bartussek[2,3]
Heinrich Herre[1]

[1]Institute for Medical Informatics, Statistics and Epidemiology (IMISE)
Leipzig University
alexander.uciteli@imise.uni-leipzig.de

[2]IntraFind Software AG [3]novineon CRO & Consulting Ltd
christoph.goller@intrafind.de

Abstract:
We present an innovative system for semantic search, based on an ontology, called Search Ontology. The Search Ontology contains search terms, for which synonymous labels can be defined, and search concepts, specified by rules, that determine how search terms are combined with abstract NEAR or Boolean operators to describe corresponding concepts in documents. A search query can be generated from the ontological specification and executed on an information retrieval system such as Lucene[1] afterwards. This approach has the advantage that the user can create powerful and complex queries by ontological specifications only, with minimal effort and without knowing the query syntax. The ontology itself is easily adaptable, extensible and reusable. No information contained in the ontology is used while preprocessing and indexing the documents, since the ontology is being constantly expanded by users of the system and changes in the ontology should not trigger new indexing and analysis for the whole document collection. The system is intended for domain experts, e.g., patent examiners or experts in the field of post-market surveillance of medical devices.

1 Introduction

The need to retraceably gather documents from heterogeneous sources in order to substantiate answers to very specific questions is a recurring scenario in many domains, e.g., patent examination and medical post-market surveillance (PMS). The quality of the search results depends on several aspects, including the possibility to specify powerful queries. The central component of the expounded search method is an ontology, called Search Ontology (SO), that allows the expert to formally specify domain concepts, search terms associated to the domain, and rules describing domain concepts. Furthermore, a prototypical software tool has been developed which generates Lucene queries from information contained in the ontology. These queries can be used by a suitable search engine (Solr[2],

[1]http://lucene.apache.org/
[2]http://lucene.apache.org/solr/

Elasticsearch[3]). This method supports an autonomous construction of complex queries by a domain expert and allows an easy adaptability of the ontology and the reusability of its parts for recurring tasks.

The ideas, presented in this paper[4], have been developed within the BMBF[5]-supported project OntoVigilance[6]. The aim of this project is the development of a tool which supports a systematic gathering of clinical safety data of medical devices. Each manufacturer is obliged to set up a system of PMS. The primary purpose is protection of the patients safety and development of safer medical devices for public health. However, the retrieval of suitable information for a rule-consistent PMS is hampered by highly heterogeneous data sources and a whole amount of data sets to be evaluated.

2 Detailed Description of the Approach

2.1 Search Ontology (SO)

The SO is developed in Web Ontology Language (OWL)[7]. The categories of the ontology (see **Categories**) are represented in OWL by classes, the relations (see **Relations**) - by properties and the rules (see **Rules**) - by property restrictions. The SO can be used for information retrieval in any domain. For that, it must be extended by the corresponding domain ontology. In the OntoVigilance project an ontology for the post-market surveillance and vigilance domain (PVO) is developed. The PVO enables the identification of available scientific data on safety, complications and efficacy of a medical device in the field of endoscopy. For example, recurrence of bleeding (re-bleeding) after an unsuccessful endoscopic intervention due to device failure or user errors is a life threatening complication. A faulty designed endoscopic tool, an improper handling or unsuitable anatomical location for the endoscopic device may lead to failure in hemostasis. By identifying failure modes and root causes of device failures and user errors, the design or instructions for use given for such medical devices may be improved to further prevent such complications.

We use the top level ontology GFO (General Formal Ontology, [Her10]) as a framework for the foundation of the Search Ontologys components. GFO provides, among others, an ontology of categories. The instances of the category GFO:Concept are concepts, e.g., the concept Hemostasis, whereas the instances of the category GFO:Symbolic_Structure are abstract strings, e.g., the abstract string “Hemostasis”, the instances of which are tokens, written, e.g., on a sheet of paper.

[3] `http://www.elasticsearch.org/`

[4] We present work in progress. Only parts of the ideas described are already implemented. The ontologies and tools will be licensed after completion of the project.

[5] Federal Ministry of Education and Research; `http://www.bmbf.de`

[6] `http://ontovigilance.org/`

[7] `http://www.w3.org/TR/owl2-overview/`

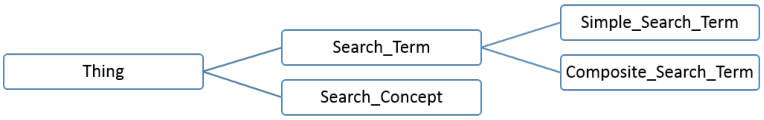


Figure 1: Categories of the Search Ontology

Categories The SO includes - on the top level - two categories, called Search_Term and Search_Concept. The category Search_Concept is a subcategory of GFO:Concept and the category Search_Term is a subcategory of GFO:Symbolic_Structure. The category Search_Term has two subcategories: Simple_Search_Term and Composite_Search_Term (see Fig. 1). The category Simple_Search_Term is instantiated by single words, while the instances of Composite_Search_Term are composed of the instances of Simple_Search_Term by using a relation has_part (see **Relations**), according to certain rules (see **Rules**). The domain experts may define various subcategories of Simple_Search_Term resp. Composite_Search_Term which can be used to structure the search terms of the domain. Examples of subcategories of Simple_Search_Term from the PVO (see Fig. 2) are Hemostasis_Adjective (having instances: inadequate, incomplete), Hemostasis_Noun (with instance: hemostasis) and Rebleeding_Noun (with instance: rebleeding). An example of a subcategory of Composite_Search_Term is Rebleeding_Phrase. The instances of this category are combinations of an instance of Hemostasis_Adjective and an instance of Hemostasis_Noun (e.g., "inadequate hemostasis", "incomplete hemostasis"). To any instance of the category Simple_Search_Term may be associated various labels. By such an association terms may be denoted by several synonymous notations resp. notations in different languages. The category Search_Concept is the root node of the subtree of domain concepts. For any domain concept a search rule can be specified (see **Rules**). Examples of concepts of the PVO are Complication or its subcategory Rebleeding.

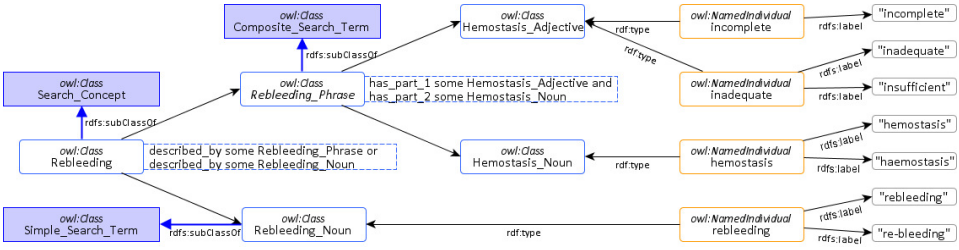


Figure 2: Example

Relations The SO includes three relations: has_part, max_distance and described_by. The first two relations connect instances of Search_Term, whereas the relation described_by connects instances of Search_Concept with instances of Search_Term. By using the relation has_part instances of Composite_Search_Term can be assembled from instances of Simple_Search_Term. The relation max_distance describes the maximal number of words which are allowed to occur between the simple terms within the composite term. The

relation described_by specifies a connection between the domain concepts and the combinations of terms describing them.

Rules For the specification of rules we use property restrictions and associate them by owl:equivalentClass with the category (represented by owl:Class) for which the rule should apply. The SO includes three types of rules which are based on the mentioned three relations: 1. the rules for assembling composite terms from simple terms are based on the relation has_part and are associated with the subcategories of Composite_Search_Term (e.g., has_part_1 some Hemostasis_Adjective and has_part_2 some Hemostasis_Noun); 2. the rules for determining the maximal distance between simple terms within the composite term are based on the relation max_distance and are associated with the subcategories of Composite_Search_Term (e.g., max_distance value 2); 3. the rules for specification of term combinations describing a domain concept are based on the relation described_by and are associated with the subcategories of Search_Concept (e.g., described_by some Rebleeding_Phrase or described_by some Rebleeding_Noun).

2.2 Queries and OntoQueryBuilder (OQB)

We developed the OntoQueryBuilder, a software component which automatically generates Lucene queries, using the information contained in the SO. In the present stage of the SO and OQB, following parts of the Lucene query syntax are supported: single terms, simple term groups, proximity term groups, Boolean operators and brackets. We sketch the working method of the OQB. The queries are generated for all domain concepts (subclasses of Search_Concept, e.g., Rebleeding) which contain a property restriction, based on the property described_by (see Fig. 2). The Boolean operators are taken from the property restriction. Then we distinguish two cases. If one of the categories contained in the property restriction is a subcategory of Simple_Search_Term (e.g., Rebleeding_Noun) the labels of the instances must be simply connected by OR (rebleeding OR re-bleeding). If the category is a subcategory of Composite_Search_Term (e.g., Rebleeding_Phrase) in the first step the property restriction of the category, based on has_part, must be interpreted and contained subcategories of Simple_Search_Term must be identified. From the labels of the instances of these categories (Hemostasis_Adjective and Hemostasis_Noun) all combinations (possibly by using the order) must be assembled and connected by OR (“inadequate hemostasis” OR “incomplete hemostasis” OR “insufficient hemostasis” OR “inadequate haemostasis” OR . . .). In case of two categories, the instances of each category have altogether, e.g., 10 labels, 100 combinations must be assembled. These combinations can be automatically generated, which shows the benefit of this solution. If the maximal distance between the single terms is defined, the query will be adapted, accordingly (“inadequate hemostasis”~2).

3 Future Work

When compared to the traditional approach of using the ontology at index time [WD10], using the ontology at search time imposes some inflexibility on the information extraction. Even though this loss is more than compensated by the flexibility gain of not having to reindex on every ontology change, it makes sense to explore ways of transferring back parts of the information extraction to index time. Currently SO and OQB support basic full text queries with Boolean and NEAR operators. We plan to use information extraction technology (entity extraction, dependency parsing) to extract SO-independent entities and general relations from text. The results will be used to produce an enriched full text index, which in turn will be the basis for more precise query operations. We already implemented such an approach for named entities such as person names, organizations and locations and plan to extend this idea to storing general dependency relations.

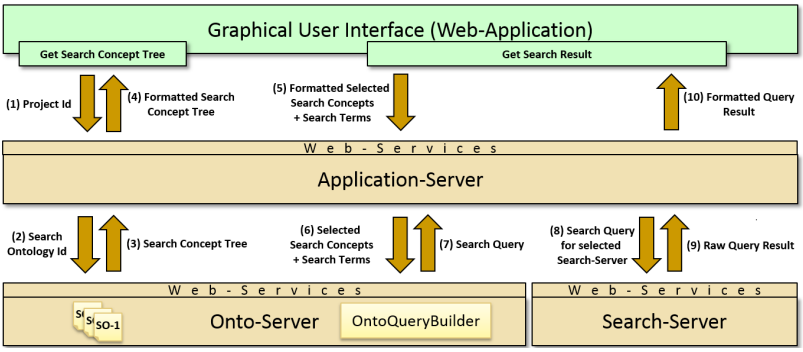


Figure 3: Prototypical Application

Moreover we are working on the prototypical implementation of the complete system as web application (see Fig. 3).

4 Related Work

There are various approaches which use ontologies or related/similar conceptions in full text search. [WD10] presents a survey of current ontology based information extraction systems, which all have in common the fact, that the ontology is used only in the text-processing step, and that the query answering system is decoupled from and unaware of the ontology. GoPubMed ([DDKS04]) is closely related to our work because it employs ontologies to search over medical domains, i.e., it also retrieves documents from PubMed. In contrast to our approach, GoPubMed is based on the idea that the ontology reflects the given documents and thus force the use of given taxonomies inherent to the system. Therefore, a user is not able to gain a higher precision determined by his subjective needs. Our approach solves this problem with the use of Search Ontology. Its inherent methodology simplifies the definition of search rules by domain experts. These rules enable the user to

send powerful and complex queries which only have to be implicitly formulated.

Our approach of storing semantic information in a full text index is related to the approach used by Sindice[8]. However Sindice stores RDF triples in the index, while we use information extraction technology to identify potential entities and relations in documents and store the results in a full text index. No information from the ontology is used during indexing. In this way we do not have to change the index when the ontology changes.

5 Conclusion

The introduced Search Ontology provides a promising way for specifying and re-using complex search queries. It is mainly intended for domain experts such as patent examiners or retrieval experts in the field of post-market surveillance of medical devices. Advantages of this approach are: 1. the modular design of the Search Ontology which can be easily adapted and extended as well as its relatively simple structure, so that the Search Ontology is applicable by non-ontologist domain experts without excessive efforts and deeper knowledge of the query syntax, and 2. the fact that this approach accommodates constantly evolving ontologies, without the need to re-index the document collection on ontology changes. The rationale of the introduced ontology promises a feasible application and a sufficient flexibility to be used in a real-world application. The realization of the OntoVigilance project revealed that the domain experts may use the SO and are able to develop the corresponding domain ontologies. The first prototype of the OQB was tested and proved to generate correct Lucene queries.

6 Acknowledgment

This work was supported by the BMBF-project OntoVigilance, FK: 01-S12038D.

References

[DDKS04] Ralph Delfs, Andreas Doms, Er Kozlenkov, and Michael Schroeder. GoPubMed: ontology-based literature search applied to GeneOntology and PubMed. In *In Proceedings of German Bioinformatics Conference. LNBI*, pages 169–178. Springer, 2004.

[Her10] Heinrich Herre. General Formal Ontology (GFO) : A Foundational Ontology for Conceptual Modelling. In Roberto Poli and Leo Obrst, editors, *Theory and Applications of Ontology*, volume 2. Springer, Berlin, 2010.

[WD10] Daya C. Wimalasuriya and Dejing Dou. Ontology-based Information Extraction: An Introduction and a Survey of Current Approaches. *J. Inf. Sci.*, 36(3):306–323, June 2010.

[8] `http://sindice.com/`

Semantik-gestützte Analyse von und Suche in Kundenspezifikationen im Maschinenbau

Martin Voigt, Daniel Hladky

Ontos AG
Mittelstrasse 24, CH-2560 Nidau
martin.voigt,daniel.hladky@ontos.com

Abstract: Die gezielte Suche von Informationen in großen Dokumentenmengen ist eine der wesentlichen Herausforderungen der heutigen Zeit. In diesem Papier wird beschrieben, wie wir die Analyse von und Suche in mehrsprachigen Kundenspezifikationen in einem aktuellen Kundenprojekt im Maschinenbau realisiert haben. Im Rahmen der Dokumentenanalyse kommen computerlinguistische und semantische Technologien zum Einsatz. Basis für die Suche bildet das Paradigma des Faceted Browsing.

1 Motivation und Problemstellung

In nahezu allen Unternehmensbereichen wachsen heutzutage die Datenmengen drastisch an, so dass es für den Nutzer zunehmend schwieriger wird, einen Überblick zu behalten bzw. die darin enthalten Informationen schnell und zuverlässig aufzufinden. Diese Herausforderungen wurden insbesondere auch im noch laufenden Kundenprojekt mit der AviComp Controls GmbH (ACC), einem Engineering-Spezialist für Steuerungslösungen für rotierende Maschinen, deutlich. In diesem Unternehmen muss der Vertrieb effizient bis zu 100 Spezifikationen je Anfrage von internationalen Kunden hinsichtlich der Passgenauigkeit auf das eigene Leistungsspektrum analysieren, Unklarheiten identifizieren und Angebote unterbreiten. Hierbei müssen teilweise auch bereits abgegebene Angebote oder hauseigene Spezifikationen aus dem gesamten Dokumentenstamm (> 100.000) hinzugezogen und analysiert werden. Die zeitaufwändige Informationssuche wurde bisher manuell mit der betriebssystemeigenen Suche oder mit Standardsoftware, wie Adobe Acrobat, durchgeführt. Als ein weiteres Problemfeld konnte die hohe Einstiegshürde für neue Vertriebsmitarbeiter identifiziert werden, da sie genau wissen müssen, welche (technischen) Konzepte bei der Prüfung von Spezifikationen relevant sind und welche nicht. Erfahrene Vertriebsingenieure bauen hier auf ihr angesammeltes, implizites Wissen, sind jedoch auch nicht davor gewappnet, die Prüfung bestimmter Fakten zu versäumen.

Um die genannten Problemstellungen zu lösen, wurde die Ontos Linked Data Information Workbench[1] (LDIW) entsprechend der Anforderungen konfiguriert und erweitert. Die drei wesentlichen Beiträge zur Unterstützung der Arbeit des Vertriebes von ACC sind:

[1] `http://www.ontos.com/products/ontosldiw/` (23.06.2014)

- Automatische Analyse, Verlinkung und Indexierung multilingualer (Text-) Dokumente, u. a. in Englisch, Deutsch und Russisch, in verschiedensten Formaten, wie etwa Microsoft Word und Excel, PDF, Bilder, ...
- Unterstützung der multilingualen Wissensstrukturierung von relevanten Konzepten
- Benutzerfreundliches, effizientes User Interface zur Suche von Dokumenten, deren Relationen untereinander sowie von wesentlichen Konzepten in den Spezifikationen

Im Folgenden wird das Konzept und die technische Lösung vorgestellt, ehe im Anschluss auf die gewonnenen Erkenntnisse und das bisherige Kundenfeedback sowie die weitere Entwicklung der Software eingegangen wird.

2 Konzept und technische Lösung

Basis für die technische Lösung ist die Ontos LDIW, deren Bestandteile für das Kundenprojekt in Abbildung 1 überblicksartig angezeigt und in den folgenden Abschnitten kurz erläutert werden. Der Zugang für die Nutzer erfolgt in einer Portal-artigen Webanwendung, die einen Workflow über die verschiedenen Arbeitsschritte vom Hochladen der Dokumente und der Informationsextraktion bis hin zur eigentlichen Suche bietet.

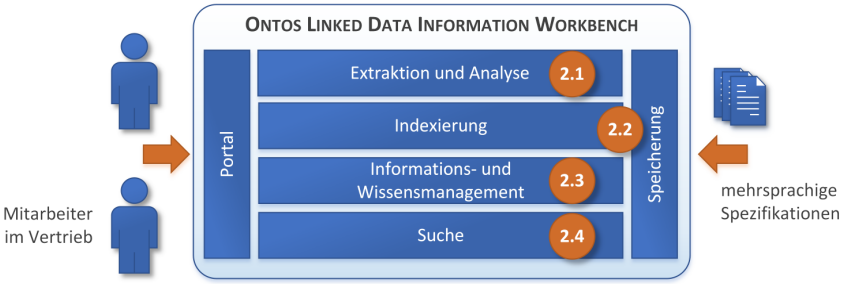


Abbildung 1: Ontos LDIW im Überblick mit Verweisen auf die beschreibenden Abschnitte

2.1 Extraktion und Datenanalyse

Die Verarbeitung der Spezifikationen läuft in parallelen Threads, wobei in jedem ein 3-stufiges Verfahren angewandt wird: 1) Homogenisierung und Textextraktion, 2) Vorverarbeitung und 3) Konzepterkennung. Nachfolgend werden diese näher erläutert.

Alle von den Vertriebsmitarbeitern bereitgestellten Dokumente werden zunächst in ein gemeinsames Format gebracht, um die Dokumente auch in Zukunft mit einem einheitlichen, verbreiteten Werkzeug öffnen und betrachten zu können. Hierbei fiel die Wahl auf das PDF-Format. Existieren Informationen nur in Bildern, was etwa durch den Scan alter Dokumente (z. B. Schaltpläne) bzw. aufgrund technischer Zeichnungen sehr oft der Fall ist, werden aus diesen mittels der freien OCR-Software tesseract-ocr[2] maschinenlesbare

[2] `https://code.google.com/p/tesseract-ocr/` (23.06.2014)

Texte extrahiert und diese ebenfalls als PDF bereitgestellt. Basierend auf dem homogenen Datenbestand werden anschließend mittels Apache Tika[3] die Texte seitenweise aus den Dokumenten gelesen.

Diese werden in zwei Schritten vorverarbeitet, um später eine möglichst genaue Analyse zu ermöglichen: Zunächst wird eine seitenweise Spracherkennung durchgeführt, da sich die Sprache nicht nur von Dokument zu Dokument sondern auch innerhalb einer Spezifikation ändern kann. Ein weiteres, in einem Vorverarbeitungsschritt zu lösendes Problem stellt die Entfernung von wiederkehrenden Texten, wie z. B. Kopf- oder Fußzeilen oder Kommentaren dar. Sind in diesen relevante Konzepte enthalten, würde deren Analyse ein Informationsrauschen erzeugen und so die Qualität der Suche verschlechtern.

Im letzten Schritt erfolgt die Erkennung der relevanten Konzepte in den vorverarbeiteten Texten. Grundlage ist ein multilingualer SKOS[4]-Thesaurus, dessen Erstellung in Abschnitt 2.3 näher erläutert wird. Aus diesem werden alle Bezeichner der definierten Konzepte in der Sprache des zu analysierenden Textes gelesen. Um ein effizientes Matching der Namen mit dem Text zu ermöglichen, wird der Aho-Corasick String Matching Algorithmus [AC75] eingesetzt.

2.2 Speicherung und Indexierung

Im Anschluss an die Datenanalyse erfolgt zunächst die Speicherung der beim Hochladen definierten Metainformationen, wie z. B. Name des zugehörigen Projektes oder die Nummer und Version des Dokumentes. Aber es werden auch die extrahierten Informationen, bspw. Sprache oder URIs der Konzepte, persistiert. Als Speicherlösung kommt der Ontos-eigene Triple Store OntoQUAD[5] zum Einsatz, der u. a. in Benchmarks aber auch in der Praxis seine Leistungsfähigkeit beweisen konnte [PPD+13]. Die Verwendung von RDF als Datenmodell und RDFS als Schemasprache bringt eine Vielzahl von Vorteilen mit sich, wie etwa eine leichte Erweiterbarkeit des Modells in späteren Projektphasen oder eine automatische Inferenz der Instanzdaten.

Um eine schnelle Suche in und Filterung von Dokumenten zu gewährleisten, wird die Apache Solr[6] Suchplattform eingesetzt. Sie bietet u. a. Volltextsuche, Faceted Browsing, oder die Indexierung von PDF-Dokumenten. Diese werden zusammen mit den auch in OntoQUAD gespeicherten Metainformationen hochgeladen, indexiert und stehen quasi in Echtzeit zur Suche bereit. Solr wurde durch das lucene-skos-Plug-In[7] erweitert, welches das Feature der Labelexpansion implementiert. Hierbei werden die Bezeichner des SKOS-Thesaurus erneut eingelesen und bei der Volltextsuche genutzt. Diese bringt bspw. den Vorteil, englischsprachige Begriffe auch in russischsprachigen Dokumenten zu finden. Weiterhin kommt bei der Indexierung das Semantic Vectors Package[8] zur Anwendung, was verschiedene Algorithmen, wie das Latent Semantic Indexing, beinhaltet. Diese be-

[3] http://tika.apache.org/ (23.06.2014)
[4] http://www.w3.org/2004/02/skos/ (23.06.2014)
[5] http://www.ontos.com/products/ontoquad/ (23.06.2014)
[6] http://lucene.apache.org/solr/ (23.06.2014)
[7] https://github.com/behas/lucene-SKOS (23.06.2014)
[8] https://code.google.com/p/semanticvectors/ (23.06.2014)

rechnen Vektorwortmodelle, mit denen Ähnlichkeiten zwischen Dokumenten anhand von gewählten Wörtern berechnet werden können.

2.3 Informations- und Wissensmanagement

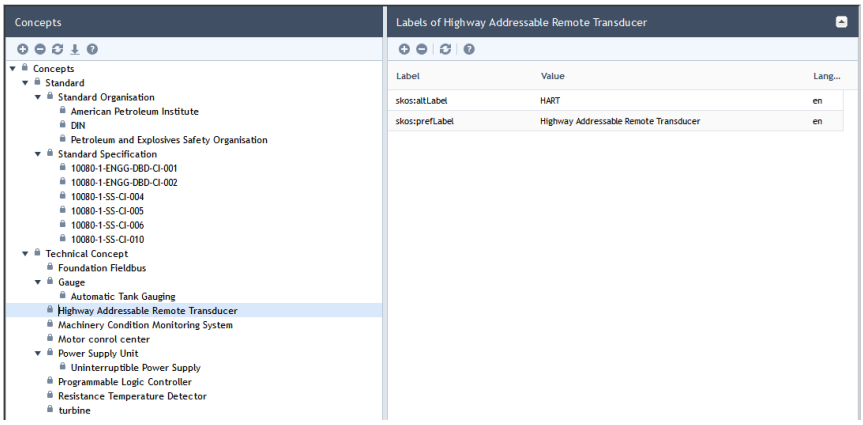


Abbildung 2: Beispielhafter, SKOS-basierter Thesaurus entwickelt in MiniDix

Für das Wissensmanagement kommt der webbasierte, mehrsprachige Ontologie-Editor OntoDix[9] zur Verwendung, der für die Erstellung von SKOS-basierten Thesauri deutlich vereinfacht wurde, um so auch eine höheren Akzeptanzgrad bei Nutzern ohne Semantic-Web-Erfahrung zu erzeugen. Wie Bild 2 zeigt, besteht er im Wesentlichen aus zwei Bereichen. Im linken Widget kann man neue Konzepte erstellen, existierende in der Hierarchie umsortieren oder auch entfernen. Auf der rechten Seite kann man für das jeweils gewählte Konzept die Bezeichner verwalten, wobei in bevorzugte (*skos:prefLabel*) und in alternative (*skos:altLabel*) Bezeichner unterschieden wird. Weiterhin können diese hier für unterschiedliche Sprachen definiert werden, was bei der späteren Suche den Mitarbeitern mit fehlenden Sprachkenntnissen zugutekommt.

2.4 Suche

Wie bereits genannt, wurde Apache Solr in die LDIW integriert und als Suchmaschine genutzt. Da die Anpassung von Solr an die Anforderungen des Projektes leicht zu realisieren waren, lag der Implementierungsschwerpunkt vor allem im Suchinterface, das neben einer Volltextsuche eine leichte Filterung über die Metadaten basierend auf dem Paradigma des Faceted Browsing [YSLH03] bieten soll. Um dem Kunden frühzeitig ein erstes User Interface bereitstellen zu können, das die wesentlichen Funktionen beinhaltet, wurde auf das in Solr integrierte, Template-basierte Framework Apache Velocity[10] zurückgegriffen. Bild 3 zeigt die Benutzerschnittstelle, die in vier eingefärbte Bereiche untergliedert ist: (1)

[9] `http://dix.ontos.ru/dix/` (23.06.2014)
[10] `http://velocity.apache.org/` (23.06.2014)

Feld zur Volltextsuche mit autocomplete-Funktion, (2) eine Auflistung der zur Filterung genutzten Facettenwerte, (3) Facetten und deren Facettenwerte sowie (4) eine Ergebnisliste. Letztere zeigt die Suchtreffer samt des von Solr automatisch aus den PDFs extrahierten Texts, der für die Volltextsuche genutzt wird, und relevante Metainformation, wie die gefundenen semantischen Konzepte, an. Suchtreffer werden automatisch hervorgehoben und helfen dem Nutzer bei der Bewertung der Ergebnisse. Hinsichtlich des User Interfaces konnten wir feststellen, dass die Suche schnell (Filteranfragen über ca. 2500 Testdokumente liegen im Schnitt bei ca. 40ms) und flüssig durchgeführt werden kann.

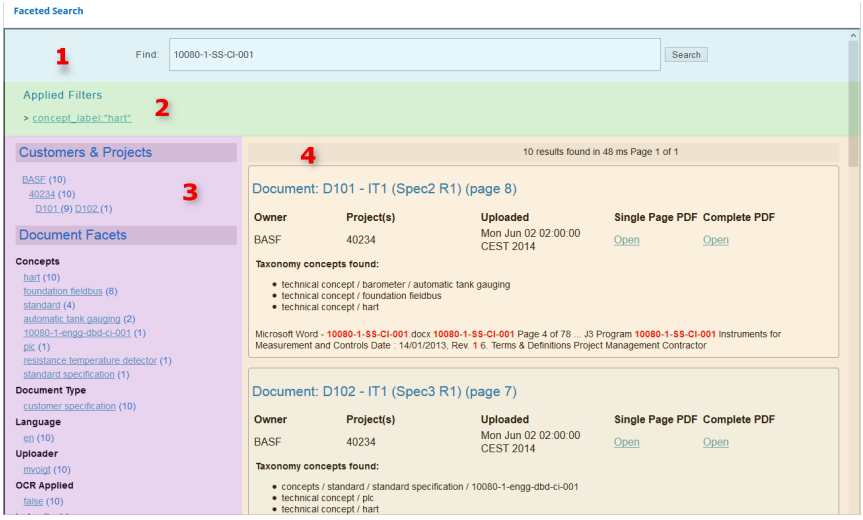


Abbildung 3: Facettierte Suche nach relevanten Dokumenten

3 Erfahrungen und geplante Weiterentwicklungen

Für die Umsetzung des ACC-Projektes wurde ein agiles, iteratives Vorgehen gewählt, so dass frühzeitig Prototypen vorliegen mit denen auch die Mitarbeiter des Kunden experimentieren und so hilfreiches Feedback geben können. Trotz anfänglicher Skepsis, die bspw. im Mehraufwand durch das Hochladen von Dokumenten und in der Einarbeitung in das neue Werkzeug begründet ist, wurde den Testern schnell der Mehrwehrt der neuen Suchlösung deutlich.

Bei der Umsetzung zeigten sich jedoch drei Probleme, denen im Vorfeld keine große Bedeutung beigemessen wurde, deren Behebung jedoch viel Zeit in Anspruch nahm:

Qualität der Spezifikationen Nach den ersten Tests der Verarbeitungspipeline mit realen Spezifikationen stellte sich heraus, dass deren textuelle Inhalte sich stark in der Qualität unterscheiden und so auch die Qualität der späteren Suche beeinflussen. So beinhalten manche Dokumente nur Verweise auf andere und sind somit weniger wichtig, wiederkehrende Kopf- und Fußzeilen oder kopierte Textpassagen sorgen für Rauschen, Sprachen werden im Dokument oder gar in einer Seite vermischt.

Erstellung einer Wissensbasis Die Entwicklung einer semantischen Wissensbasis stellte ebenfalls eine große Hürde dar. So fällt es auch sehr erfahrenen Ingenieuren schwer, das implizite Wissen zu externalisieren. Hierzu gehört einerseits die Identifikation wesentlicher Konzepte aber andererseits auch deren Strukturierung. Zur Unterstützung des Prozesses wurde der Ontologie-Editor OntoDix auf das Notwendigste reduziert. Weiterhin wurde ein einführender Workshop zum Thema Wissensstrukturierung mit Mitarbeitern durchgeführt.

Entwicklung des Suchinterfaces Letztlich wurde angenommen, dass die facettierte Suche schnell zu realisieren ist. Wie oben beschrieben, bot sich Apache Velocity sehr gut an. Jedoch wurde bei Tests mit Mitarbeitern und Echtdaten sofort deutlich, dass eine individuellere Funktionalität notwendig ist. So wird bspw. die Menge der Metainformationen, aus denen die Facettenwerte für die Filterung extrahiert werden, so groß sein, dass sie nicht über einfache Listen abbildbar sind. Weiterhin wird auch eine direkte Vorschau der Suchergebnisse im PDF benötigt, so dass die Ingenieure den Kontext besser erfassen können. Diese und weitere clientseitge Funktionen wie die Integration des Latent Semantic Indexing lassen sich durch die fehlende AJAX-Unterstützung in Velocity schwierig oder nicht realisieren.

Die zuletzt genannte Problemstellung stellt auch den aktuellen Arbeitsschwerpunkt dar. Auf Basis des Ajax-Solr-Frameworks[11] wird eine Widget-basierte, leicht konfigurier- aber auch erweiterbare Suchoberfläche entwickelt, die Apache Velocity ersetzen wird. Nach deren Fertigstellung wird dieses User Interface und das Gesamtsystem einem breiteren, längeren Nutzertest sowie einem Lasttest mit großen Datenmengen unterzogen.

Weitere künftige Erweiterungen in der LDIW sind die Konzeption und Realisierung von Mechanismen zur Disambiguierung von gefundenen semantischen Konzepten sowie die Integration von Crawling-Methoden, um automatisch Ausschreibungen aus dem Web dem System zuzuführen. Und es wird auch eine Integration der Daten aus dem ERP sowie des CRM erwogen, was durch die Nutzung von W3C-Standards und semantischen Technologien gut realisierbar ist. Die Kundendaten sollen durch im Web verfügbare Nachrichten mittels der genannten Crawling-Mechanismen extrahiert und angereichert werden.

Literatur

[AC75] Alfred V. Aho und Margaret J. Corasick. Efficient String Matching: An Aid to Bibliographic Search. *Commun. ACM*, 18(6):333–340, Juni 1975.

[PPD+13] Alexander Potocki, Anton Polukhin, Grigory Drobyazko, Daniel Hladky, Victor Klintsov und Jörg Unbehauen. OntoQuad: Native High-Speed RDF DBMS for Semantic Web. In *Knowledge Engineering and the Semantic Web*, Jgg. 394 of *Communications in Computer and Information Science*, Seiten 117–131. Springer Berlin Heidelberg, 2013.

[YSLH03] Ka-Ping Yee, Kirsten Swearingen, Kevin Li und Marti Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '03, Seiten 401–408, New York, NY, USA, 2003. ACM.

[11] `https://github.com/evolvingweb/ajax-solr` (23.06.2014)

Visual Innovations for Product Search Interfaces

Mandy Keck[1], Martin Herrmann[1], Dana Henkens[2], Severin Taranko[2],
Viet Nguyen[3], Fred Funke[3], Steffi Schattenberg[3], Andreas Both[3], Rainer Groh[1]

[1]Chair of Media Design, Technische Universität Dresden, 01062 Dresden
{mandy.keck, martin.herrmann, rainer.groh}@tu-dresden.de
[2]queo GmbH, Tharandter Str. 13, 01159 Dresden
{d.henkens, s.taranko}@queo-group.com
[3]Unister GmbH, Barfußgäßchen 11, 04109 Leipzig
{viet.nguyen, fred.funke, steffi.schattenberg, andreas.both}@unister-gmbh.de

Abstract: As the 'gatekeepers' of the internet, search engines provide access to a huge amount of data. Nowadays, many search engines are able to interpret user queries and automatically suggest additions or corrections. Despite these technological advances, the search paradigm itself remains stable for most engines. Especially for users with little knowledge of the search domain or a very vague idea of what they are looking for, it is difficult to formulate a query with the established keyword search paradigm. Faceted Search is one of the techniques that facilitate the process of query formulation by offering a structured and classified overview of the domain. In this paper, we give an overview of search paradigms besides keyword-based techniques and will present alternative visual approaches for product search.

1 Introduction

In the last decade search engines have become the 'gatekeepers' for the contents of the internet. Since the first search engine 'Archie' in early 1990[1], which was able to search files in remote public FTP directories with an 8 character search string, the scope of functions supporting search activities have grown enormously to compete with the tremendously rising amount of data. Every day, 2.5 quintillion bytes of data are created, so much that 90 percent of the data in the world today has been created in the last two years alone[2]. Mobile devices, social networks and real-time information are driving forces for this developement. Hence, the size of items, the variety of sources as well as the overall amount of data will continue to grow. With this development, accessing relevant information is an increasingly difficult problem [CHJR+13].

To address this problem, many websources (e.g. DBpedia[3]) have started to publish structured data on the web according to Linked Data principles. The so called 'Web of Data' contains billions of RDF triples describing relations between entities and connecting dif-

[1]http://people.lis.illinois.edu/~chip/projects/timeline/1990archie.htm, last accessed: 23.06.2014
[2]http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, last accessed: 23.06.2014
[3]http://dbpedia.org, last accessed: 23.06.2014

ferent data sources [BBBE11]. This large knowledge base (open linked data) can be used to structure and contextualize unstructured data like emails or images and facilitate real time decision making. However, without an appropriate interface, it is hard to formulate queries with complex RDF query languages like SPARQL.

Appropriate visual query systems (VQS), focusing on expressiveness and usability, are needed to transform a request of a non-technical user into a valid query for the underlying database [SGJR+13]. Especially when the user does not have a precise idea of his desired search result and is not familiar with the search domain, the system should provide a vocabulary to prevent incomplete or ambiguous queries.

In this paper, we focus on users with a vague information need in product search scenarios. We discuss different visual approaches and interaction strategies to enable the user to put his desires in concrete terms. Furthermore, we classify these approaches and emphasize the use of contextualized and structured data to improve the search experience and the quality of results.

2 Search Paradigms

Lookup and Exploratory Search are search activities that can be distinguished in the context of web search [Mar06]. Lookup describes the most basic kind of search tasks such as fact retrieval, known item search, and question answering. Exploratory Search is a more complex process that usually starts with a vague information need and therefore requires multiple iterations of learning, investigation, and reformulation of the search query.

Our goal to support product search based on vague ideas of the possible result can also be described as Motive-based Search [KHB+13]. Motive-based Search refines Explorative Search by specifying the motive and the aim of the process to find a suitable product. A motive can be defined as the reason for a search as well as particular conditions like how much a product should cost.

Thus, search interfaces are required that support the user to express his information need based on these reasons and conditions. Therefore, two search paradigms are well established in the web search world: Direct Search and Navigational Search [Tun06]. Direct Search allows users to simply write their queries in a text box and became enormously popular with web search engines, such as Google and Yahoo! Search. Text boxes and search forms are well-suited for lookup-scenarios, in which the user has a concrete idea of the desired product (e.g. looking for a flight to London). However, in motive-based search scenarios, it is difficult to transform a possibly vague information need into a specific query. In contrast, Navigational search systems provide guidance through the use of a taxonomy [Tun06]. Common web interfaces often use hierarchical structures to browse the information space in a predetermined order. To allow multiple access points for the search, the principle of Faceted Browsing provides an alternative that is based on a multidimensional taxonomy [ST09]. It enables users to explore the information space by progressively refining their choices in each dimension [RR11].

The described search functionalities are based on the principle of reduction where the

results from the set of all resources are progressivley narrowed down to the desired set of answers. Clemmer and Davies define the term 'general-to-specific' (GtS) to refer to query interfaces that allow the user to specify abstract criteria for a result set [CD11]. By contrast, they define 'specific-to-general' (StG) to refer to interfaces that explicitely support starting with an example and generalzing it to find similar examples. This interface paradigm is based on expansion where aspects of a concrete example are progressively generalized to find other result that match the pattern [CD11].

In the following, some examples for these search paradigms that match to our motive-based search scenario are described. For instance, the Faceted Browsers Elastic Lists [SM07] and Flamenco [YSLH03] can be classified according to the GtS-Paradigm and Navigational Search. Both support the construction of complex search queries by selecting values of the data facets, thereby restricting the result set iteratively and allowing to browse through the information collection. Elastic lists offer a visual interface that allows additional analytical tasks (e.g. the size of each facet value highlights predominant values) [SM07]. Compared with this, Flamenco provide a textual interface with additional search functionalities such as Keyword Search and Similarity Search. Hence, it combines all of the mentioned paradigms: GtS, StG, Navigational Search, and Direct Search. Faceted Browsers are well suited for the exploration of structured data. For semi-structured data, Similarity Search is an effective way to start or continue a search. Similarity searches or search by example approaches aim to find objects with attributes as close as possible to a reference object and can be classified according to the StG paradigm. Whereas Flamenco and the fotosharing service Flickr[4] allow to find items with similar properties based on the selection of tags (Navigational Search), other search by example approaches solve query formulation problems with multimedia queries such as last.fm[5], retrievr[6], google image search[7] and picalike[8]. Last.fm is a music discovery service that gives personalised recommendations based on the music the user is listening to or looking for and thus combine Direct and Navigational Search. Retrievr enables the user to define a vague search query by using sketches. Google image search and picalike are search by example approaches that allow a similarity search based on pictures and belong to direct search interfaces. Picalike also incorporates a product search that detects products similar to the reference object. The results can be filtered by various criteria such as brand and price. Hence, the user can switch to a search paradigm using Navigational Search.

Both the GtS and StG paradigms are well-suited for motive-based search scenarios to narrow down or broaden the scope. Combined, they could become a powerful tool to support the user during an exploratory search process with frequently changing information needs.

[4] https://www.flickr.com/photos/tags/search/clusters/, last accessed: 23.06.2014
[5] http://www.last.fm, last accessed: 23.06.2014
[6] http://labs.systemone.at/retrievr, last accessed: 23.06.2014
[7] http://images.google.de/, last accessed: 23.06.2014
[8] http://www.picalike.com/products/similarity-search.php, last accessed: 23.06.2014

3 Visual Exploration Systems for Product Searches

In this chapter, we present three product search interfaces different from conventional form-based interfaces with text boxes, radio buttons and drop down menus which are developed in our research project "VISEA - Visual Search Interfaces"[9]. Each of the suggested approaches support a particular search strategy to find the desired product in a very large database with structured or semistructured data. Further on, we consider different search domains such as travel search and financial products. Since both domains entail different user requirements and expectations, the resulting interface concepts support emotional or analytical decisions.

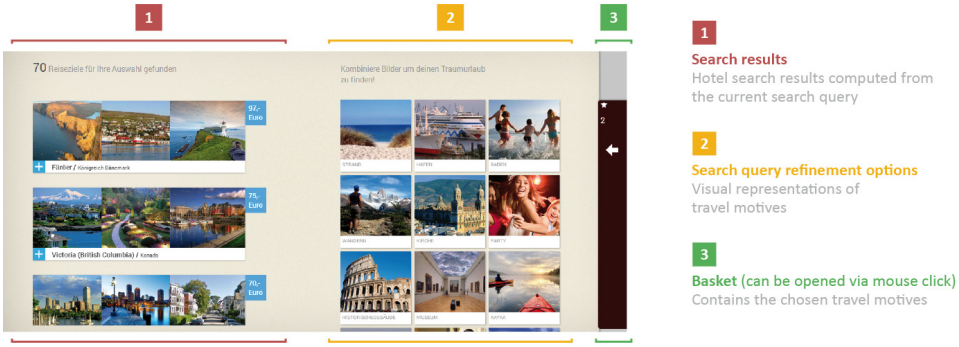


Figure 1: getInspired: interface with visual concepts for travel searches

3.1 Recommendation-based Search Approach

Figure 1 shows the interface 'getInspired' for travel search. We designed this interface to support and trigger emotionally driven decisions. Instead of a direct query on the attributes of the result set, the user communicates his preferences to the system through a selection of concepts represented by expressive images. Based on the previous decision of the user, the system decides which concepts are presented in the next refinement step. For each step, the interface provides the best concepts in terms of a fast reduction of the result set. In order to implement this interface, the results need to be tagged with concepts from a domain topology that serves as knowledge base. After each refinement step, the most applicable results are presented on the left side of the interface. These results can be explored or further specified by selecting more appealing concepts on the right side. The interface focuses on easy self-explanatory interactions and the use on mobile and touch interfaces. An initial user study with 29 participants and 12 different search-tasks was conducted to compare the recommendation-based approach to a strict navigational concept selection. The study indicates that the time and interactions needed to solve a given search task can be improved significantly. In order to validate these claims in real conditions, a more

[9] http://visea-projekt.de

extensive knowledge base as well as an automatic method for the classification of images are needed. These improvements are necessary to extend the users possibilities to express his needs and the systems ability to respond to users actions.

3.2 Analytical Approach using Parallel Coordinates

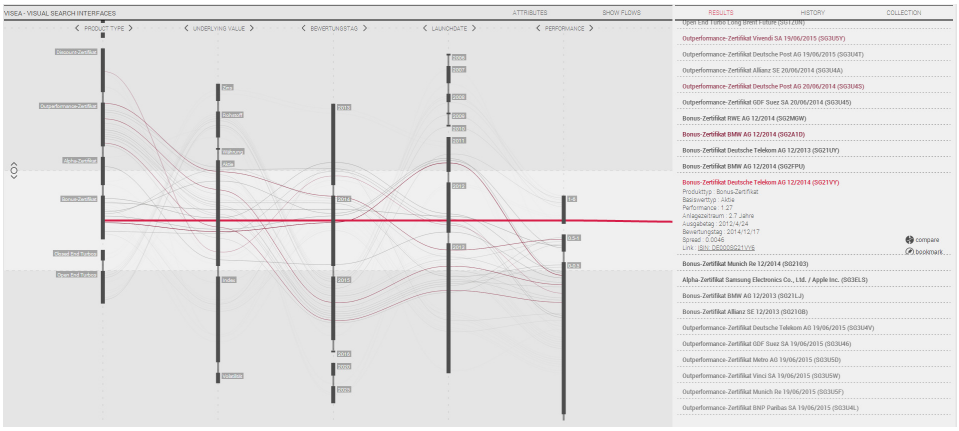


Figure 2: Interface concept based on Parallel Coordinates: the Similarity Search centers the selected product; a fuzzy filter (light grey) reduces the amount of polylines to focus on similar products

The interface presented in Figure 2 [KHB+14] is designed for an analytical aproach and elaborates ideas of the product search interfaces from the Product Explorer [ROF12]. It combines principles of Faceted Browsing with the visualization method of parallel coordinates [ID90] to support additional analytical tasks (cp. Elastic Lists in section 2). Each parallel axis represents one facet of the multi-dimensional data set and is spatially separated in container elements representing the facet values. The height of each container, in reference to the overall height of the axis, shows the distribution of data items within the facets and allows the analysis of predominant values. A product is visualized by a polyline that intersects each axis at the appropriate value. In addition to the analytical benefits of the parallel coordinate visualization presented in [ROF12], this interface supports facet filters (GtS) as well as a Similarity Search (StG). Multiple filters can be activated on one or different axes to create complex filters as known from Faceted Browsing to narrow the scope. With the Similarity Search approach, the user starts his search by selecting a product of interest matching his expectations in one or more characteristics. The function automatically centers all intersections of the selected product to the median line of the visualization. Hence, products with similar properties are arranged close to the center and can be easily identified. The interface is developed to explore financial data. In this use case, the user needs to get an overview of variety, range, and distribution of the provided dataset and their attributes. Consequently, the interface has a strong analytical focus and combines the StG and GtS paradigms to explore the information space.

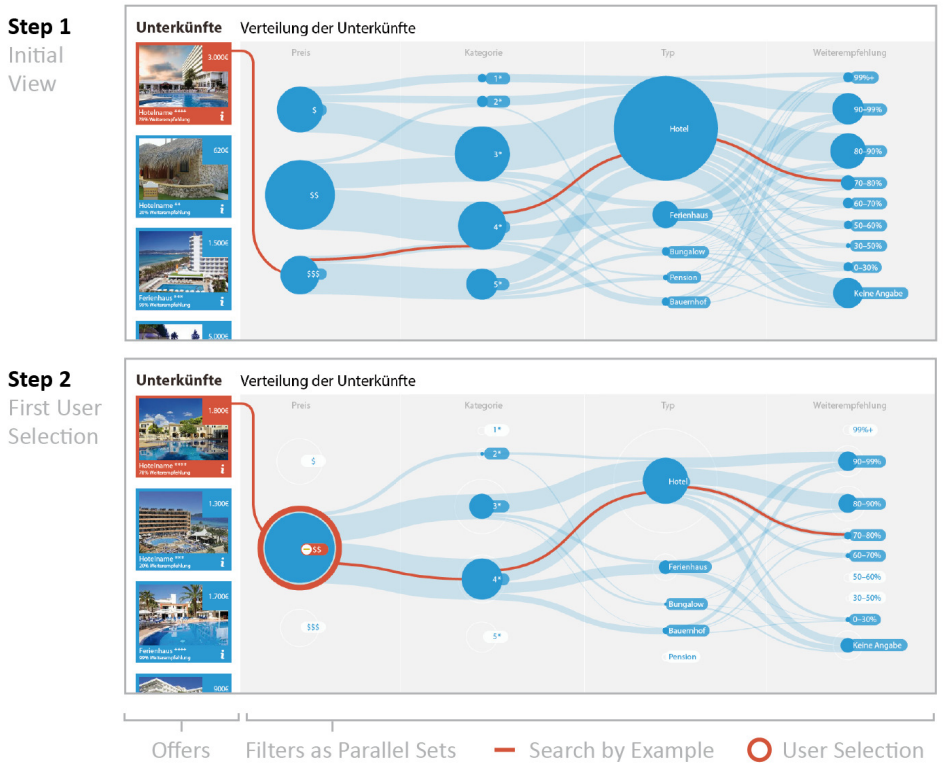


Figure 3: Faceted Browsing with Parallel Sets

As a variation of the parallel coordinates, parallel sets [KBH06] present sets of products instead of single polylines. Therefore, the containers are connected by polygonal streams representing the logical conjunction of the adjacent containers. The size of these streams indicate the frequency of items included in the conjunction. Figure 3 shows an interface designed for travel search based on the principle of parallel sets. Facets of each attribute are depicted as circles. The distribution of characteristic within an attribute is mapped to the size giving the user an impression on the most common values. In comparison to the parallel coordinates, parallel sets are independent to the number of visualized products and are a good possibility to gain insights in the overall structure of the dataset.

The user tests conducted in [KHB+14] have shown that parallel sets are faster and easier to understand with regard to faceted searches and analysis tasks, whereas parallel coordinates have advantages in comparison tasks and similarity-based searches.

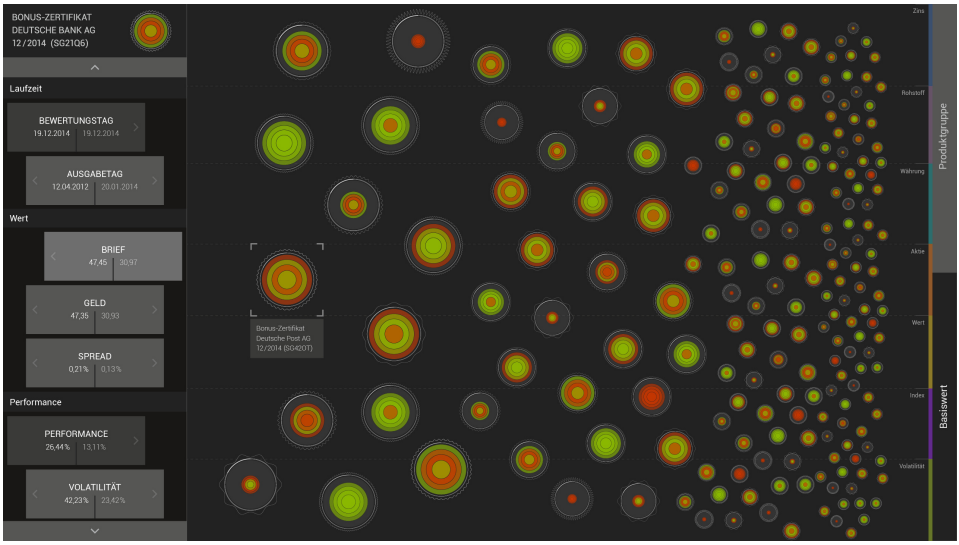


Figure 4: Visual Similarity Search: glyphs on the right side represent similar products, the attributes of the reference product (left side) can be selected or deselected to focus on properties of interest

3.3 Search by Example Approach

A visual approach for a similarity search in a financial product scenario is shown in Figure 4. Each circular glyph represents a single product and depicts its most important features, like performance over time and investment term. The horizontal positioning is used to express the similarity to a chosen reference object. Similar objects are displayed larger to show details (semantic zooming). The less significant results are smaller to give the user an overview of the data distribution. The global similarity calculation is determined by the ordinal attributes selected on the left side of the interface. The vertical positioning can be used to cluster products in categories aligned on the right side of the interface. The interface is designed for touch devices and a limited amount of items and attributes. An early prototypical implementation suggests that this kind of interface is well suited for the paradigm of search by example (StG) and the browsing of a large structured or semi structured dataset. Whilst the overall design of the interface may remain the same for any kind of structured dataset, the iconic representation should be adapted to the use case. With a more sophisticated version of our current prototype, we hope to create a flexible and playful way to find a suitable product and gain insights in distribution and structure in the complex area of financial products.

4 Conclusion

The presented interface concepts incorporate several search paradigms and strategies. The GtS paradigm is well-suited to narrow down large result sets of a known domain with the help of filters (Parallel Coordinates or Parallel Sets) or visual concepts (getInspired Interface). However, without a certain knowledge of the search domain, the user is often not able to specify abstract criteria to exploit the information space. Especially in the case of financial search, novices do not know which attributes such as performance or investment term would lead to a suitable financial product. For this use case, the StG paradigm (Search by example) is a promising approach. Here the user can start with one product to discover the meaning of the included product features and then generalize the interesting features to find similar results.

All three concepts can be classified into two groups: recommendation-based and analytical approaches. The recommendation-based approach (getInspired) hides the complex data structure of the search system and leads to a small result set in a few steps. The analytical approach (Parallel Coordinates or Parallel Sets) provides a deep insight into the data structure and enables an evaluation of the chosen product in comparison to the remaining information space. The presented search by example approach combines both strategies by offering products similar to the reference item, but also represents smaller glyphs to give an overview of the data distribution. The suggested usage of these strategies depends on the use case as well as the preferences and previous experience of the user. Whereas the recommendation-based approach allows to find a suitable product in very short time, the analytical approach requires more time and effort to explore and analyze the data set, but can lead to better and more trustworthy results.

The proposed examples indicate that the search experience and speed in product search scenarios can be significantly improved by visually appealing interfaces. Therefore, we suggest further research in this area. Future extensions need to focus on the backend infrastructure and the linking with sophisticated knowledge sources. An improved semantic evaluation of the users query and a better understanding of his current needs is the pathway towards an intelligent and user-adapted interface. Another challenge is to ensure the reliability of results and establish visual search interfaces as credible source of information and decision making in a big data environment with a huge variety of different information sources and formats. To this end, search tools have to find a good balance between deep insights in the underlying data sources and a user-friendly, easily comprehensible interface.

5 Acknowledgments

This work has been supported by the European Union and the Free State Saxony through the European Regional Development Fund (ERDF). The research presented in this article has been conducted in cooperation of the Chair of Media Design - Technische Universität in Dresden, Unister GmbH from Leipzig and queo GmbH from Dresden, Germany.

References

[BBBE11] Christian Bizer, Peter A. Boncz, Michael L. Brodie, and Orri Erling. The meaningful use of big data: four perspectives - four challenges. *SIGMOD Record*, 40(4):56–60, 2011.

[CD11] Aaron Clemmer and Stephen Davies. Smeagol: A âœSpecific-to-Generalâ Semantic Web Query Interface Paradigm for Novices. In Abdelkader Hameurlain, StephenW. Liddle, Klaus-Dieter Schewe, and Xiaofang Zhou, editors, *Database and Expert Systems Applications*, volume 6860 of *Lecture Notes in Computer Science*, pages 288–302. Springer Berlin Heidelberg, 2011.

[CHJR+13] Diego Calvanese, Ian Horrocks, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Michael Meier, Mariano Rodriguez-Muro, and Dmitriy Zheleznyakov. On Rewriting and Answering Queries in OBDA Systems for Big Data (Short Paper). In *OWL Experiences and Directions Workshop (OWLED)*, 2013.

[ID90] Alfred Inselberg and Bernard Dimsdale. Parallel Coordinates: A Tool for Visualizing Multi-dimensional Geometry. In *Proceedings of the 1st Conference on Visualization '90*, VIS '90, pages 361–378, Los Alamitos, CA, USA, 1990. IEEE Computer Society Press.

[KBH06] R. Kosara, F. Bendix, and H. Hauser. Parallel Sets: interactive exploration and visual analysis of categorical data. *Visualization and Computer Graphics, IEEE Transactions on*, 12(4):558–568, July 2006.

[KHB+13] Mandy Keck, Martin Herrmann, Andreas Both, Ricardo Gaertner, and Rainer Groh. Improving Motive-Based Search: Utilization of Vague Feelings and Ideas in the Process of Information Seeking. In Norbert Streitz and Constantine Stephanidis, editors, *Distributed, Ambient, and Pervasive Interactions*, volume 8028 of *Lecture Notes in Computer Science*, pages 439–448. Springer Berlin Heidelberg, 2013.

[KHB+14] Mandy Keck, Martin Herrmann, Andreas Both, Dana Henkens, and Rainer Groh. Exploring Similarity: Improving Product Search with Parallel Coordinates. In *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*, volume 8522 of *Lecture Notes in Computer Science*, pages 160–171. Springer, 2014.

[Mar06] Gary Marchionini. Exploratory Search: From Finding to Understanding. *Communications of the ACM*, 49(4):41–46, April 2006.

[ROF12] Patrick Riehmann, Jens Opolka, and Bernd Froehlich. The product explorer: decision making with ease. In Genny Tortora, Stefano Levialdi, and Maurizio Tucci, editors, *AVI*, pages 423–432. ACM, 2012.

[RR11] Tony Russell-Rose. Designing the Search Experience. In Pedro Campos, T. C. Nicholas Graham, Joaquim A. Jorge, Nuno Jardim Nunes, Philippe A. Palanque, and Marco Winckler, editors, *INTERACT (4)*, volume 6949 of *Lecture Notes in Computer Science*, pages 702–703. Springer, 2011.

[SGJR+13] Ahmet Soylu, Martin Giese, Ernesto Jimenez-Ruiz, Evgeny Kharlamov, Dmitry Zheleznyakov, and Ian Horrocks. OptiqueVQS – Towards an Ontology-based Visual Query System for Big Data. In *International Conference on Management of Emergent Digital EcoSystems (MEDES 2013)*. ACM, 2013.

[SM07] Moritz Stefaner and Boris Muller. Elastic Lists for Facet Browsers. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, DEXA '07, pages 217–221, Washington, DC, USA, 2007. IEEE Computer Society.

[ST09] Giovanni Maria Sacco and Yannis Tzitzikas. *Dynamic Taxonomies and Faceted Search Theory, Practice, and Experience.* Springer, New York, 2009.

[Tun06] D. Tunkelang. Dynamic category sets: An approach for faceted search. *ACM SIGIR*, 6, 2006.

[YSLH03] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted Metadata for Image Search and Browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 401–408, New York, NY, USA, 2003. ACM.

Data Management in the Cloud

(DMC2014)

2. Workshop Data Management in the Cloud[1] und Herbsttreffen der GI-Fachgruppe Datenbanken

Abstract: Der 2. Workshop Data Management in the Cloud, organisiert durch den GI-Arbeitskreis Datenmanagement in der Cloud (DMC) fand am 22. und 23. September 2014 im Rahmen der 44. Jahrestagung der Gesellschaft für Informatik in Stuttgart statt. In den Workshop integriert war auch das traditionelle Herbsttreffen der Fachgruppe Datenbanken der Gesellschaft für Informatik.

Angelehnt an das Leitthema der Jahrestagung, Big Data – Komplexität meistern, adressierte der Workshop die Komplexität des Managements und der Nutzung großer Datenmengen mit einem besonderen Schwerpunkt auf Cloud-Infrastrukturen. Die Beherrschung großer Datenmengen ist heute eine Notwendigkeit quer durch alle Branchen und alle Bereiche des täglichen Lebens. Cloud-Infrastrukturen spielen bei der Entstehung, der Verwaltung, Nutzung und Analyse der Daten eine immer wichtigere Rolle. Im Blickpunkt des Workshops standen daher sowohl Technologien und Infrastrukturen für das skalierbare Datenmanagement als auch etablierte und neue Anwendungsbereiche, in denen das effiziente Management großer Datenmengen eine zentrale Rolle spielt.

Das Programm umfasste neben begutachteten Vorträgen der Universitäten Hamburg, Magdeburg und Stuttgart zu den genannten Themen auch die Vorstellung neuer Kompetenzzentren im Bereich Big Data. Erhard Rahm (Uni Leipzig) stellte das Competence Center for Scalable Data Services and Solutions in Dresden und Leipzig vor, Volker Markl (TU Berlin) ging in seinem Vortrag auf die Vision des Berlin Big Data Center ein und Harald Schöning (Software AG) berichtete über das Smart Data Innovation Lab. Eingeladene Vorträge aus Industrie und Forschung rundeten das Programm ab. Stefanie Scherzinger (Hochschule Regensburg) berichtete über die Datenmodellierung in NoSQL-Datenbanken, Torsten Steinbach (IBM) stellte IBM's Analytic Warehouse Cloud Service vor während Privacy-protecting Data Management in Cloud-Infrastrukturen im Zentrum eines Vortrags von Andreas Müller, Sebastian Hudert und Victor Fäßler (TWT GmbH) stand.

Leipzig, Ilmenau, Stuttgart, im Juli 2014

Andreas Thor, Deutsche Telekom Hochschule für Telekommunikation Leipzig
Kai-Uwe Sattler, Technische Universität Ilmenau
Holger Schwarz, Universität Stuttgart

[1] http://dmc.informatik.uni-stuttgart.de/ws2prog.html

NoSQL OLTP Benchmarking: A Survey

Steffen Friedrich, Wolfram Wingerath, Felix Gessert, Norbert Ritter

Databases and Information Systems Group
University of Hamburg
{friedrich, wingerath, gessert, ritter}@informatik.uni-hamburg.de

Abstract: In recent years, various distributed NoSQL datastores have been developed that offer horizontal scalability and higher availability than traditional relational databases, but fewer querying options and only reduced consistency guarantees. The diversity of the design space makes it difficult to understand the performance implications of individual system designs. Existing benchmarking tools measure some relevant aspects, but do not capture all of them. In this paper, we give an overview of the state-of-the-art in NoSQL OLTP benchmarking, identify missing features as well as open challenges and point towards possible solutions.

1 Introduction

Traditional relational database management systems (*RDBMSs*) were designed to be the one solution to all data management and storage problems and thus they all share a broad foundation: They provide the full SQL spectrum (including joins) and are not highly available as they follow strongly consistent ACID semantics. Hence, these systems can be compared reasonably well by benchmarks (e.g. TPC) that quantify their performance solely along dimensions like throughput and request latency under realistic OLTP workload patterns.

During the last decade, however, a myriad of distributed NoSQL datastores have emerged that promise to fill the growing gap between what highly distributed (web) applications require and what traditional RDBMSs can provide by sacrificing consistency guarantees and querying options (hence the name) in favour of high availability and scalability. The design space is vast and the dependencies between desirable properties are complex. Even though high-level abstractions such as the CAP Theorem [Bre00, GL02] or the PACELC model [Aba12] can help to understand the basic trade-offs that are involved, the performance implications of individual system designs are often not obvious, so that the optimal data storage solution for an application cannot be chosen by simply comparing candidate specifications. Representative benchmarking results are not available, though, as current approaches do not capture all the relevant aspects and employ operation mixes as workloads that only loosely correspond to real-world scenarios.

In this paper, we survey the in our opinion most relevant NoSQL OLTP benchmarks, point out missing features, discuss open challenges and possible ways to approach them.

In Section 2, we survey current efforts in evaluating NoSQL database systems according to request latency and throughput, availability and consistency in its various forms. We briefly present a novel approach to benchmarking consistency in Section 3 and give a summary of open challenges and a conclusion in Section 4.

2 Comparing Distributed Datastores

Data in distributed datastores are spread (sharded) across several nodes (servers). In order to prevent data loss and to preserve availability in the presence of errors, data are also replicated across several nodes. But sharding and replication have severe implications on performance: While replication can increase read performance as the same data items are available on several machines, it can also increase write latency (synchronous replication) or give rise to stale reads or conflicting writes (asynchronous replication). Furthermore, sharding introduces an overhead to operations that span data items on different nodes; operations such as joins over datasets on different nodes become infeasible and global invariants are hard to maintain. As no single system can achieve all desirable properties at once, each NoSQL datastore is optimised for a specific property set.
In the remainder of this section, we discuss different dimensions along which NoSQL datastores can be evaluated and compared. First, we cover the most basic form of benchmarking that translates to measuring raw performance in terms of request latency and operational throughput, concentrating on YCSB as the most popular and widely accepted benchmarking framework in the field of NoSQL OLTP benchmarking. We then consider related work on the quantification of availability. Finally, we discuss different notions of consistency in distributed databases and examine approaches towards their measurement.

2.1 Request Latency and Operational Throughput

The de facto standard for NoSQL OLTP benchmarking is the *Yahoo Cloud Serving Benchmark (YCSB)* framework which was published in 2010 [CST+10]. Primarily targeting key-value store functionality, YCSB is a highly generic general-purpose OLTP benchmark that can easily be extended to support additional databases as it only requires the implementation of a simple CRUD interface (read, scan, insert, update, delete). In contrast to other benchmarks such as TPC-C that evaluate system performance under realistic workloads, YCSB is built around the idea of examining different characteristics of a datastore using distinct architectural *tiers*: A tested database is stressed with CRUD operations by the YCSB Client to measure throughput, operations per second and request latency under high load (*performance tier*). YCSB workloads do not model real-world applications, but are customisable mixes of read/write operations that represent entire application classes: Random distributions[1] determine which operations are performed, which records are read/written, how frequently they are read/written and other workload details. The records that are read and written consist of a number of (random) ASCII string attributes; the number

[1] YCSB supports uniform, multinomial and variants of Zipfian distributions.

of columns and the size of each value can be configured. Scalability and elasticity can be quantified by measuring the increase in performance as machines are added to the system (*scalability tier*).
The generic workloads and the simple CRUD interface make YCSB suitable to explore general trade-offs in a wide class of storage systems. But due to this simplicity, YCSB does not account for functionality beyond that of simple key-value stores and abstracts from many differences such as data models. Seeing that many NoSQL systems actually are highly specialised for particular distributed large-scale applications, a performance evaluation for specific real-world tasks appears desirable. Furthermore, the single-node client poses a scalability problem for benchmarking in large-scale environments as it cannot saturate large distributed database systems under test. Additional tiers for characteristics like availability, replication or fault tolerance were proposed in the original contribution, but not implemented in YCSB.

BG [BG13] is another benchmarking framework which, in contrast to YCSB, is still actively developed. It models actions in a social networking application and therefore provides a richer conceptual model and requires the implementation of a more complex interface than YCSB for operations such as listing all friends of a given user or returning the top-k posts on a user's profile. In addition to throughput and latency measurements as in YCSB, BG supports SLA (service level agreement) conformance checks: The *Social Action Rating (SoAR)* represents the highest throughput that can be sustained without violating a given SLA.

2.2 Availability

A partition-tolerant available system can always accept read and write requests by clients and will eventually return a meaningful response, i.e. not an error message. However, anomalies such as network partitions or server crashes occur on a daily basis in large distributed environments and therefore data have to be replicated across different failure domains, so that operation can be sustained in spite of such anomalies. Since availability is of paramount importance for many applications, information on the performance impact of different replication strategies during normal operation or in failure scenarios are valuable.

The *Under Pressure Benchmark (UPB)* [FMdA+13] aims at quantifying and comparing the availability of different distributed database systems by measuring the operational throughput under three different configurations: with replication turned off and no failing nodes, with replication turned on and no failing nodes and with replication turned on in the face of node failures. In every setting, the system under test is given some time to stabilise (warming period) before measurements are made. In order to generate a heavier workload than is possible with a single client, UPB uses several independent YCSB clients in parallel and aggregates the results.
While it does manage to quantify the impact of replication on steady-state performance during normal operation and after node failures, the UPB does not quantify availability well, in our opinion, because the measurements do not reflect performance problems *directly after* node failure.

In a 2013 evaluation, Engber et al. [NE13, Eng13] measure the *immediate* impact of node failures on operational throughput and therefore availability. Similarly to the UPB, they aggregate the results of several independent YCSB clients after a short warming period. As they induce node failure *during* measurement, the authors are able to observe real-time behaviour such as downtime during failover, performance drops caused by load balancing and, more generally, database performance before the system stabilises again.

2.3 Consistency

The CAP Theorem advertises that strong consistency is unachievable for a highly available system in the face of a partition, the underlying notion of strong consistency being *linearisability* which translates to the guarantee that reads and writes are always executed atomically and are sequentially consistent (linearisable [HW90]). In simple words, strong consistency guarantees that all clients have the same view on the data at all times. Some systems sacrifice availability to remain strongly consistent while others employ available, but only eventually consistent models: In an *eventually consistent (EC)* system, all replicas of a data item are guaranteed to converge in the absence of updates and partitions, i.e. they will reach an identical state at some point in the future. The order in which individual writes are applied, though, is arbitrary and dependent writes may become visible out of order. However, as demonstrated by current research [LFK+11, BGH+13], it is possible to simultaneously achieve EC and *causal consistency (CC)* at the cost of increased staleness and additional complexity in comparison to plain EC. Informally, any two operations on a data item in a CC system are executed in the order they are received, if one of them (directly or transitively) causally depends on the other.[2] EC and CC are *data-centric consistency* models that have their focus on the internal state of the storage system and synchronisation between replicas, whereas inconsistencies that can be observed by clients are captured by *client-centric consistency* models: *Monotonic Read Consistency (MRC)* mandates that a client that has observed version n of a data item will never observe this particular data item in a version less than n. *Monotonic Writes Consistency (MWC)* requires that two updates issued by the same client are executed in the order that they arrive at the storage system and *Write Follows Read Consistency (WFRC)*[3] guarantees that a write operation following a read of object x in version n is never applied to replicas with versions of x less than n. Under *Read Your Writes Consistency (RYWC)*, a client that has written version n of a data item will never observe this particular data item in a version less than n.
While strongly consistent systems display no inconsistencies with respect to neither ordering nor staleness and therefore provide all of the above-mentioned client-centric guarantees, many EC systems provide none[4]. Thus, it is hard to determine whether the advantages of using an EC datastore outweigh the disadvantages of potentially stale or conflicting data

[2] An operation o_1 is directly causally dependent on another operation o_2, if (1) the storage system receives o_1 before o_2 and both operations are issued by the same client or if (2) o_1 is an update and o_2 is a read that returns the result of o_1.

[3] Example: WFRC guarantees that a reply to a posting can never be observed before the initial posting.

[4] MRC, MWC and WFRC can be provided in an eventually consistent system.

in a particular application scenario.

In the following, we discuss ways to compare different distributed databases with respect to the ordering and staleness properties they exhibit. We also cover the notion of *transactional consistency* known from ACID databases which differs from distributed replica consistency and efforts in verifying consistency after system partitions.

2.3.1 Staleness

Given any highly available EC system, it is impossible to provide strict bounds on how strongly the different replicas diverge or how long it will take them to reach agreement once they are out of sync. However, various efforts have been made to predict or measure staleness for different systems.

With the *Probabilistically Bounded Staleness (PBS)* prediction model, Bailis et al. [BVF+12] estimate the expected bounds on staleness in Dynamo-style datastores with respect to both versions and wall clock time on the basis of write propagation and read messaging delays. They introduce the two metrics t-*visibility* for the probability of observing a write t time units after it returned and k-*staleness* for the probability of reading one of the last k versions of a data item and then further combine them both in $\langle k, t \rangle$-*staleness consistency* to encapsulate the probability of reading one of the k latest versions of a value, given the latest value was written at least t time units ago.
On the downside, PBS does not treat the system as a black box, but instead requires internal knowledge of the storage system (delays) and abstracts from implementation details; as anti-entropy protocols such as read-repair or the use of Merkle trees are not taken into account, actually observed staleness may be less severe than predicted by PBS. Furthermore, it is only applicable to a specific class of distributed (Dynamo-style) databases.

In [WFZ+11], Wada et al. describe an experimental setup to measure time-based staleness in different cloud databases. The authors describe different experimental configurations that employ one reader and one writer which are hosted on the same virtual machine (VM), on different VMs in the same datacenter and on VMs in different datacenters: The writer periodically writes its local time to a particular data item and the reader repeatedly retrieves this item; on discovery of a new timestamp, the reader computes the observed staleness window as the difference between its own local time and the observed timestamp.
As the authors do not mention whether or how the local clocks of writer and reader are synchronised, we assume there to be no synchronisation. Under this assumption, all experiments where writer and reader do not share the same machine are profoundly flawed, because they might contain arbitrary measurement errors. The significance of the remaining experiments appears limited as well since writer and reader might experience reduced or no staleness at all, e.g. because they are routed to the same replicas.

To address this issue, Bermbach et al. [BT11, BT14] propose an extension to the approach of Wada et al. that employs multiple readers: One writer periodically writes the current local timestamp and a version number, while each of several readers repeatedly polls the storage system and logs its local read timestamp, the observed write timestamp and the observed version number. In a subsequent analysis of all reader logs, t-visibility and k-

staleness are approximated under the assumption of perfect synchronisation of all client clocks. In addition to their consistency measurement component, they also run a YCSB workload with one YCSB client to saturate the storage system.
Similar to Wada et al., Bermbach et al. also rely on clock synchronisation through the cloud provider and thus significantly diminish the reliability of their results: As they assess themselves [BZS14], using local timestamps from different readers with potentially drifting clocks may lead to dubious results.

In contrast to the studies discussed above, Golab and Rahman et al. [RGA+12, GLS11] from HP Labs aim to measure the observed consistency under a given workload instead of designing a workload to derive consistency properties. They criticise that other measurement techniques introduce artificial operations (repeated reads) that tend to disrupt the workload, stress the system under test considerably and thus may distort the results. They formally define Δ-atomicity as a time-based consistency property that informally requires a read operation to return a value that is at most Δ time units stale. In other words, a system provides Δ-atomicity, if every value becomes visible during the first Δ time units after[5] the acknowledgement of its write. Figure 1 illustrates the idea of how to compute the value Δ for a given database history: Several read and write operations are executed on data item x over time where the entirety of all operations regarding version i of x is referred to as zone Z_i. For each pair of zones that belong to the same data item and overlap in terms of time, a value χ is computed that corresponds to the width of the respective staleness windows. Δ is the maximum of all χ values, i.e. the largest observed staleness window. The goal of the authors' experiments is to quantify the consistency that is provided by a distributed datastore in terms of Δ-atomicity. The required information are collected by extended YCSB clients that log timing information for each operation.
Even though the workload design differs from the approaches discussed before, the actual measurements are very similar. While the precision of the described approach also depends on clock synchronisation, the authors state the error margin to be "around 1ms".

As part of the SLA conformance check, *BG* also quantifies the amount of stale and otherwise inconsistent data that are observed during experiments. A BG workload is executed by several distributed clients each of which operates on a logical partition of the dataset. Every client is aware of the initial state of the data and each update operation and therefore stale reads can be detected through log analysis.
While the different BG clients do not interact and therefore do not require clock synchronisation and do not suffer from communication delays between reader and writer, staleness and other anomalies may not be observable in some scenarios, as both read and writer share the same physical machine. Given that reader and writer in real-world social networking applications are often globally distributed, this appears as a major drawback in terms of realism.

An approach that does not rely on clock synchronisation is implemented in **YCSB++** [PPR+11] where multiple clients are coordinated via one central ZooKeeper server [HKJ+10]. Most notably, YCSB++ measures time-based staleness with multiple clients working together in a producer-consumer pattern as illustrated in Figure 2: Initially, the consumer client subscribes (1.) to be notified by ZooKeeper as soon as a specific object is

[5] For $\Delta > 0$, Δ-atomicity is strictly weaker than atomicity which demands immediate visibility for any write.

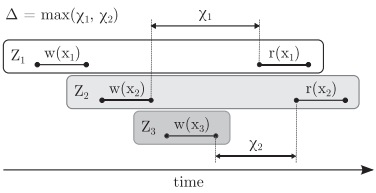


Figure 1: Example computation of Δ as the maximum of all χ values similar to [RGA+12, Figure 1]. Note that there is no χ value between Z_1 and Z_3 as the largest staleness window regarding x_1 is determined by χ_1.

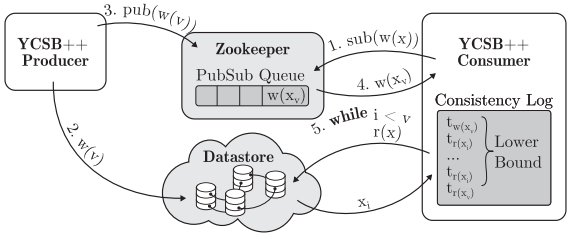


Figure 2: YCSB++ benchmark coordination and consistency measurement with Zookeeper as described in [PPR+11] .

updated by the producer. After the producer client has inserted or updated (2.) a record, it publishes (3.) the write to the ZooKeeper PubSub queue. The consumer client is notified (4.) via ZooKeeper and then repeatedly requests (5.) the updated record, until the new version is returned.
To provide a lower bound for the actual inconsistency window, the delay between the first attempt and the first successful attempt to read a record is measured by the consumer. To keep the impact of client coordination on the experimental results minimal, consistency measurements are performed on only 1% of all objects.

2.3.2 Ordering Guarantees

Knowledge about the order in which writes become visible can be derived from knowledge about stale reads. In [BVF+12], Bailis et al. capture the probability of *MRC* in the PBS prediction model. Both Wada et al. [WFZ+11] and Bermbach et al. [BT14] compute MRC violations on the basis of the log files generated during their staleness measurements. In addition, Bermbach et al. also examine their experiment logs for violations of *MWC* and *RYWC*. To our knowledge, no method for the quantification of *WFRC* violations has been proposed so far.

2.3.3 Transactions

Recently, a wealth of scalable transactional datastores have been proposed and implemented. Examples of these systems are large-scale distributed systems like Megastore, Spanner, Percolator and F1 from Google, key-value stores enhanced by atomic multi-key transactions like COPS, Granole, G-Store and Hyperdex, as well as more general-purpose commit protocols and transaction managers like MDCC or Omid.

Bailis et al. examine the design space for systems that provide ACID guarantees and high availability at the same time in [BDF+13]. They identify Read Committed as an isolation level that is often used in traditional single-node database systems and can be achieved in

a highly available transactional datastore. Furthermore, the authors state that lost updates and write skew as well as concurrent updates and potentially unbounded staleness can never be prevented in a highly available system.

To fill the gap between classic SQL-based transactional benchmarks (e.g. TPC-C) and simple cloud service benchmarks, **YCSB+T** [DFN+14] was proposed. It adds to the four tiers defined in the original YCSB contribution (performance, scalability, availability and replication) by introducing two new tiers *transactional overhead* and *consistency*. The *Transactional overhead* tier measures the latency of transactional operations (read, scan, insert, update, delete, read-and-modify) and transaction demarcation (start, abort, commit). To achieve this, a so-called *Closed Economy Workload (CEW)* is defined. It simulates bank account transactions in a closed system where money neither enters nor exits. This workload executes operations similar to YCSB, but wrapped in a single transactional context. For instance, `doTransactionalReadModifyWrite` reads two account records, transfers some money from one to the other and writes both records back. To achieve this, the central contract between YCSB and the database, the *DB interface*, is enhanced by (optional) transactional methods. The amount of concurrent transactions is determined by YCSB's *threads* parameter which defines how many threads execute the workload in parallel. To ensure that no anomalies (e.g. lost updates) occur during workload execution, a *validation phase* is executed after the transaction (consistency tier). The application-defined validation method takes the database state as an input and calculates an *anomaly score*. For the CEW, this score is simply defined as the difference between the initial and final sum of all accounts normalized by the amount of executed operations. This score has a major problem: Not only does it not detect dirty reads and non-repeatable reads, it also catches only a fraction of all lost updates, as lost updates can easily occur without any changes to the sum of all account balances.
The evaluation of YCSB+T demonstrates the usage for one particular system, but lacks a comparison of different transactional datastores. Benchmarking different scale-out transactional systems remains an important open issue in this field. YCSB+T furthermore does not detect transaction anomalies, it is limited to verifying state-based consistency constraints. While measuring anomalies (i.e. isolation level compliance) has been studied for single-server scenarios [FGA09], deriving a scheme for distributed transactional datastores thus remains an open issue. Another open challenge is the inclusion of the *availability* and *replication* tiers for transactional benchmarking. To decouple concurrency from transaction sizes and amounts, it would also be desirable to construct a benchmark which allows for more complex configurations than a one-to-one mapping between threads and transactions.

2.3.4 Consistency in the Face of Partitions

Informally, partition tolerance is the ability of a system to sustain operation in the presence of message loss between the nodes. As partitions are generally unavoidable, partition tolerance is guaranteed by virtually all NoSQL databases.
While availability in failure scenarios has already been addressed by several studies, only little work has been done on whether NoSQL databases keep their promise to remain con-

sistent during and after partitions. In his *Call Me Maybe* blog post series[6], Kyle Kingsbury examines several widely used distributed systems and their behaviour during network partitions, revealing that many do not comply with their marketing claims and actually lose data where they should not. To this end, he presents *Jepsen*[7], an open-source project that facilitates injecting failures and running tests in a virtualised distributed environment. Through his work, Kingsbury uncovers misunderstandings and implementation errors in existing database systems.

3 Our Approach

To provide strict bounds on staleness, we are working on an approach that takes up the idea of YCSB++ to provide a lower bound for the staleness window and extends it by measuring an upper bound. The basic concept of our approach is illustrated in Figure 3:

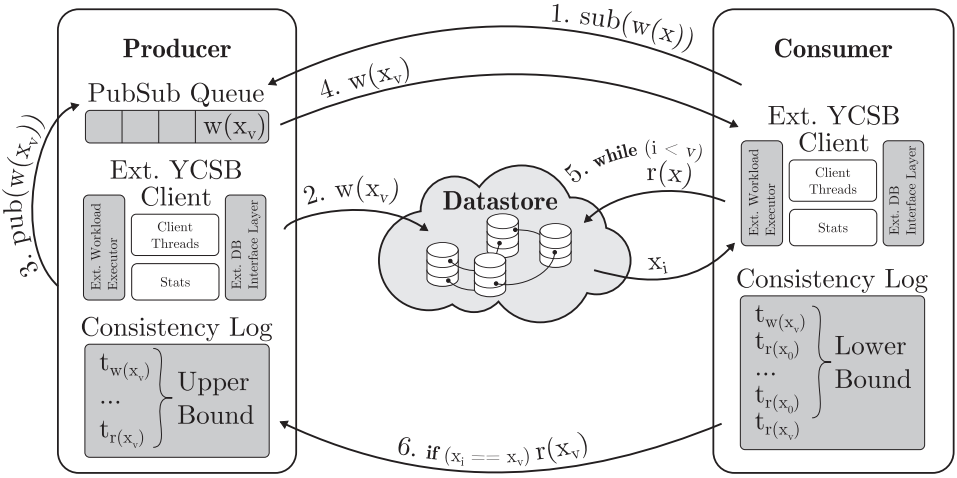


Figure 3: Our multi-client consistency measuring approach employs decentralized PubSub queuing to measure lower and upper bounds for staleness.

Similar to the YCSB++ approach, our experimental setup comprises at least one producer and one consumer client as well as, of course, one datastore. One physical machine can take the role of the producer for one data item and the role of the consumer for another. Furthermore, several consumer clients can be used to measure the staleness window of the same data item. Instead of handling communication through centralised messaging (ZooKeeper in the YCSB++ approach), we use decentralized client-side messaging. Prior to the start of an experiment, a consumer subscribes (1.) to be notified as soon as a specific data item is updated by the producer. The producer writes (2.) the data item and publishes (3.) the write to its local queue, so that every consumer is notified (4.) to repeatedly read (5.) the item, until the new value is observed. If a consumer reads the new version of a

[6] *Call Me Maybe* blog post series: `http://aphyr.com/tags/jepsen`.
[7] Jepsen: `https://github.com/aphyr/jepsen`

stale value, he notifies the producer (6.). This allows us to compute an upper bound for the staleness window of this particular read operation as the difference between the producer's timestamps directly after the consumer notification (6.) and directly before the write (2.). Apart from preventing contention, reducing latency and simplifying the process of setting up an experiment in comparison to YCSB++, using local messaging queues eliminates network latency altogether, if producer and consumer share the same physical machine. Furthermore, our approach does not rely on clock synchronisation and provides bounds for the actual inconsistency that a client experienced. It should be noted, though, that a valid upper bound can only be yielded under the assumption of MRC; in the absence of MRC guarantees, the consumer client can request (.5) the data item for a longer time in order to increase the probability of observing MRC violations, if they exist.

4 Conclusion and Open Challenges

NoSQL OLTP benchmarking is an active research topic and the boundaries of what is achievable in the field of distributed databases are being probed by both scientists and practitioners. Arguably, the most popular and most widely accepted OLTP benchmark for NoSQL databases is YCSB which facilitates measuring operational throughput and request latency for generic CRUD workload mixes. YCSB's obvious strong points are an easy-to-implement database interface, easy-to-use design and easy-to-extend architecture. On the other hand, its highly generic design can also be seen as a drawback since it does not account for functionality beyond simple CRUD and therefore does not capture the performance of datastores well that offer more sophisticated operations. It is also possible to model real-world applications like BG does, but this restricts applicability.

Several aspects of availability such as the effect of replication on steady-state performance or system performance during node failure and recovery have already been addressed in experimental evaluation. The differences between the available replication strategies and their respective configurations, however, have not been studied to full extent as far as we are aware.

As data-centric consistency cannot be measured without internal knowledge of the storage system, only client-centric approaches seem viable for generic benchmarking frameworks. Client-centric consistency in distributed systems can be measured along two dimensions: Staleness describes the time during which an acknowledged write is not applied and ordering refers to the order in which writes become visible to clients.
An inherent issue with benchmarking consistency in a distributed system, however, is that distributed measurement is also subject to imprecision through network latency and clock synchronisation. Several approaches to deal with these issues can be found in the literature: Having reader and writer of the data item under consideration share the same machine eliminates the communication delay, but also may lead to unrealistic results. Distributing reader and writer over separate physical machines, in contrast, necessitates clock synchronisation or leads to latency values that either include or ignore network delays. One rationale in workload design is to measure only observed inconsistency, while the other is to introduce artificial operations to increase the precision of the results. An interesting

question is whether the latter approach can actually lead to an observer effect.
To our knowledge, only little work has been done on the quantification of ordering guarantee violations in distributed databases. Until recently, the transactional consistency provided by a datastore has not been addressed by experimental evaluation. YCSB+T is an important first step towards transactional benchmarking of scale-out datastores and provides a useful basis for quantifying the overhead introduced by wrapping CRUD operations in a transaction. However, the inclusion of the availability and replication tiers for transactional benchmarking which have already been proposed in the original YCSB paper remain open issues.

Ongoing work shows that aspects of performance like availability and consistency can be quantified, but individual experiments only cover small parts of the vast space of possible experiment configurations. For example, the impact of different failure scenarios or (wide-area) replication strategies on both availability and consistency still need to be examined. Further, the integration of existing approaches into a comprehensive and widely applicable benchmarking suite is an important goal of future work.

References

[Aba12] D. Abadi. Consistency Tradeoffs in Modern Distributed Database System Design: CAP is Only Part of the Story. *Computer*, 45(2):37–42, Feb 2012.

[BDF+13] Peter Bailis, Aaron Davidson, Alan Fekete, Ali Ghodsi, et al. Highly Available Transactions: Virtues and Limitations. *PVLDB*, 7(3):181–192, 2013.

[BG13] Sumita Barahmand and Shahram Ghandeharizadeh. BG A Benchmark to Evaluate Interactive Social Networking Actions. In *CIDR*, 2013.

[BGH+13] Peter Bailis, Ali Ghodsi, Joseph M. Hellerstein, et al. Bolt-on Causal Consistency. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 761–772, New York, NY, USA, 2013. ACM.

[Bre00] Eric A. Brewer. Towards Robust Distributed Systems., 2000.

[BT11] David Bermbach and Stefan Tai. Eventual Consistency: How Soon is Eventual? An Evaluation of Amazon S3's Consistency Behavior. In *Proceedings of the 6th Workshop on Middleware for Service Oriented Computing*, MW4SOC '11, pages 1:1–1:6, New York, NY, USA, 2011. ACM.

[BT14] David Bermbach and Stefan Tai. Benchmarking Eventual Consistency: Lessons Learned from Long-Term Experimental Studies. In *Proceedings of the 2nd IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2014. Best Paper Runner Up Award.

[BVF+12] Peter Bailis, Shivaram Venkataraman, Michael J. Franklin, et al. Probabilistically bounded staleness for practical partial quorums. *Proc. VLDB Endow.*, 5(8):776–787, April 2012.

[BZS14] David Bermbach, Liang Zhao, and Sherif Sakr. Towards Comprehensive Measurement of Consistency Guarantees for Cloud-Hosted Data Storage Services. In *Performance Characterization and Benchmarking*, volume 8391 of *Lecture Notes in Computer Science*, pages 32–47. Springer International Publishing, 2014.

[CST+10] Brian F. Cooper, Adam Silberstein, Erwin Tam, et al. Benchmarking cloud serving systems with YCSB. In *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, pages 143–154, New York, NY, USA, 2010. ACM.

[DFN+14] Akon Dey, Alan Fekete, Raghunath Nambiar, et al. YCSB+T: Benchmarking Web-scale Transactional Databases. In *Proceedings of International Workshop on Cloud Data Management (CloudDB'14)*, Chicago, USA, 2014.

[Eng13] Ben Engber. How to Compare NoSQL Databases: Determining True Performance and Recoverability Metrics For Real-World Use Cases. Presentation at NoSQL matters 2013, 2013.

[FGA09] Alan Fekete, Shirley N. Goldrei, and Jorge Pérez Asenjo. Quantifying Isolation Anomalies. *Proc. VLDB Endow.*, 2(1):467–478, August 2009.

[FMdA+13] Alessandro Gustavo Fior, Jorge Augusto Meira, Eduardo Cunha de Almeida, et al. Under Pressure Benchmark for DDBMS Availability. *JIDM*, 4(3):266–278, 2013.

[GL02] Seth Gilbert and Nancy Lynch. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant web services. *SIGACT News*, 33(2):51–59, June 2002.

[GLS11] Wojciech Golab, Xiaozhou Li, and Mehul A. Shah. Analyzing Consistency Properties for Fun and Profit. In *Proceedings of the 30th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, PODC '11, pages 197–206, New York, NY, USA, 2011. ACM.

[HKJ+10] Patrick Hunt, Mahadev Konar, Flavio P. Junqueira, et al. ZooKeeper: wait-free coordination for internet-scale systems. In *Proceedings of the 2010 USENIX conference on USENIX annual technical conference*, USENIXATC'10, pages 11–11, Berkeley, CA, USA, 2010. USENIX Association.

[HW90] Maurice P. Herlihy and Jeannette M. Wing. Linearizability: a correctness condition for concurrent objects. *ACM Trans. Program. Lang. Syst.*, 12(3):463–492, July 1990.

[LFK+11] Wyatt Lloyd, Michael J. Freedman, Michael Kaminsky, et al. Don't Settle for Eventual: Scalable Causal Consistency for Wide-area Storage with COPS. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, SOSP '11, pages 401–416, New York, NY, USA, 2011. ACM.

[NE13] Denis Nelubin and Ben Engber. NoSQL Failover Characteristics: Aerospike, Cassandra, Couchbase, MongoDB. Technical report, Thumbtack Technology, 25 Broadway, Floor 9, New York, 2013.

[PPR+11] Swapnil Patil, Milo Polte, Kai Ren, Wittawat Tantisiriroj, Lin Xiao, Julio López, Garth Gibson, Adam Fuchs, and Billie Rinaldi. YCSB++: benchmarking and performance debugging advanced features in scalable table stores. In *Proceedings of the 2nd ACM Symposium on Cloud Computing*, SOCC '11, pages 9:1–9:14, New York, NY, USA, 2011. ACM.

[RGA+12] Muntasir Raihan Rahman, Wojciech Golab, Alvin AuYoung, et al. Toward a Principled Framework for Benchmarking Consistency. In *Proceedings of the Eighth USENIX Conference on Hot Topics in System Dependability*, HotDep'12, pages 8–8, Berkeley, CA, USA, 2012. USENIX Association.

[WFZ+11] Hiroshi Wada, Alan Fekete, Liang Zhao, Kevin Lee, and Anna Liu. Data Consistency Properties and the Trade-offs in Commercial Cloud Storage: the Consumers' Perspective. In *CIDR'11*, pages 134–143, 2011.

Cost-Aware Query Optimization during Cloud-Based Complex Event Processing

Andreas Meister, Sebastian Breß, Gunter Saake
firstname.lastname@ovgu.de
University of Magdeburg, Germany

Abstract: *Complex Event Processing* describes the problem of timely and continuous processing of event streams. The load of Complex Event Processing systems can vary (e.g., event rates). Static resource provision leads to higher monetary costs because enough resources have to be provided to efficiently handle peak loads. Therefore, most of the time the resources will not be fully utilized.

One way to achieve scalable processing and elastical resource allocation fitting varying requirements is to use *Cloud Computing*. Properties of Cloud Computing are the pay-as-you-go-payment model and high availability. These properties can be used in Complex Event Processing systems to minimize the monetary costs of systems while satisfying Service Level Agreements. Complex Event Processing systems must continuously optimize the event processing to adapt to varying loads without violation of Service Level Agreements. To guarantee efficiency, the optimization cost must be considered, leading to cost savings without violating the Service Level Agreements.

In this work, we discuss factors, which should be considered during the optimization of cloud-based Complex Event Processing systems that use the pause-train-resume strategy to migrate operators. Furthermore, we propose heuristics to estimate the cost of these factors. In our experiments, the cost could be decreased by 15 % by using a cost-aware optimizer. This proofs that the costs of cloud-based Complex Event Processing systems can be further decreased if optimization is cost-aware.

1 Introduction

Complex Event Processing (CEP), the problem of timely and continuous processing of event streams, becomes increasingly important (e.g., in stock markets [BDG07] or financial applications [ScZ05]). The different application fields have in common that the system loads vary during runtime. Additionally, certain characteristics agreed in *Service Level Agreements* (SLAs) must be guaranteed (e.g., event processing time). CEP systems can use *Cloud Computing* to provide resources elastically (e.g., adding or removing additional nodes) to adapt to changing workloads and to save resources (and hence, money) in case of low system load. Based on the pay-as-you-go-payment of Cloud Computing, elastical resource provision leads to reduced costs, because only the needed amount of resources will be provided. To minimize system costs, a continuous optimization is needed to adapt the CEP system to the current requirements. During the optimization, it might be necessary to migrate operators form over- or underloaded nodes to other nodes of the system.

Problem statement. Operator migration can lead to additional costs. Therefore, the query optimizer should be cost-aware, meaning that the optimizer should consider the cost

and effects of operator migrations during optimization. In this paper we investigate the following research question:

> How much can we decrease the overall cost of a cloud-based CEP system using a cost-aware query optimizer considering the cost of the query optimization?

In this paper, we contribute an overview of effects of operator migrations and identify factors usable to calculate the effects. Furthermore, we present cost models for the effects of operator migrations in the used CEP systems. To evaluate the cost models, an existing optimization framework [Rö12] will be extended by integrating the cost models into the optimization process. Our experiments show that we can save up to 15% monetary costs while increasing robustness of CEP systems w.r.t. Service Level Agreements (SLAs).
The remainder of this work is structured as follows. In Section 2, we briefly describe the background information. We specify our cost models to estimate the effects of operator migrations in Section 3. In Section 4, we discuss our evaluation. We provide an overview of related work in Section 5. In Section 6, we present a conclusion and future work.

2 Background

In this section, we will provide an overview of cloud-based CEP, efficient CEP query optimization, and the process of an operator migration.

2.1 Cloud-based Complex Event Processing

CEP describes the problem of timely, continuous processing of event streams [Luc01]. Goal of CEP is to gain additional information based on single events by searching for known or unknown patterns in event streams. The gained information are used for decision making or starting of new processes (e.g., locking of a credit card by a detected fraud).
The processing of events is based on defined queries. In the context of CEP systems, a query processes events until it is stopped by the system. Therefore, the query execution time is potentially unbounded. Since queries in CEP systems are long running, the probability that the requirements change are high (e.g., event rate of streams, etc.) [MSHR02].
Systems that use a static resource provision cannot adapt to changes in the workload, which can lead to under- or overutilization of the system. *Overutilization* leads to higher query execution times and may lead to inaccurate results, if not all events of the event streams can be processed (load shedding). *Underutilization* leads to inefficient resource usage, because not all available resources are fully occupied, leading to an increased end user fee.
Typically an elastic resource provision (e.g., *Cloud Computing*) can be used to solve the problem of over- and underutilizations and to minimize system costs. Using Cloud Computing, resources can be automatically added or removed during the runtime of the system. Since in cloud-based CEP systems resources are charged according to the usage, the efficient resource usage can lead to lower overall monetary costs and hence, lower costs for the end user. The system has to adapt the usage of the resources according to the current requirements of the system to minimize the overall cost.
In order to minimize the cost of the system an efficient optimization is needed to adapt the resources to current requirements of the CEP system while avoiding SLAS violations.

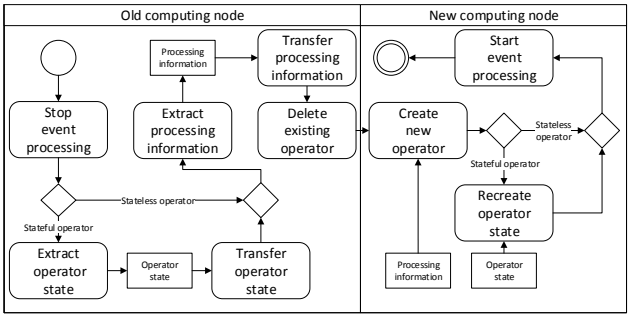


Figure 1: Migration process for stateful and stateless operators

2.2 Efficient query optimization for Complex Event Processing

The goal of query optimization is to ensure the efficiency of query processing. In cloud-based CEP systems, one essential part of the query optimization is the operator placement. The operator placement determines a assignment of existing operators to computing nodes based on the current placement and requirements of the operators (e.g., CPU load).
Since the effects of an optimization can lead to increased costs [BB05] (e.g., additional resource consumption or SLA violations), several aspects must be considered in order to provide an efficient CEP optimizer besides the efficient determination of a new placement: the estimated cost of an optimization, the point in time, when we start the optimization, and the time duration between optimizations.

Cost estimation. In order to provide a cost-aware optimization, an approach is needed that estimates the cost of an optimization. To guarantee the accuracy of the estimation, the cost calculation must consider the properties of the query processing and the operator migration of the used CEP system.

Optimization start time. The optimization causes additional resource consumption. Respectively, the optimization should only be triggered, if the system properties or workload have changed, and the query processing must be adapted to improve the overall efficiency of the system. An approach is needed to estimate, when the query optimization should be started.

Duration between optimizations. Since the optimization cost must be amortized during the runtime, the cost considerations of an optimization should include the duration of the optimized query execution. We cannot consider the remaining execution time, but the time to the next optimization, because the query execution time is potentially unbounded.

There are already suitable approaches to identify a point in time to start the optimization [KD98] and to determine the time duration between optimizations [JHJ+10]. Therefore, we focus on the cost estimation of operator migrations in the remainder of this paper.

2.3 Migration process

To remove underutilized nodes or use added nodes to release overloaded nodes, operators must be migrated between nodes. The decision which operators should be migrated will

be decided by the optimization framework. Hereby, the optimization framework has to determine three factors: (1) a node, from which an operator will be migrated, (2) the operator, which will be migrated, and (3) the target node, where the operator will be migrated to. The process of the operator migration depends on the used system. The considered cloud-based CEP system uses the pause-drain-resume strategy [ZRH04]. Based on the operator type, different steps must be performed to migrate an operator (see Figure 1). The operators can be divided into two groups: stateless and stateful operators. For stateless operators, the following steps must be performed: First, the event processing must be stopped. Second, the processing information has to be extracted and transferred. Third, the existing operator must be deleted on the old node and the new operator must be initialized on the new node. Finally, the event processing must be restarted.
For stateful operators, the state of an operator must be moved additionally. The state of an operator consists of one to n sliding windows based on the number of input streams. Since not all events of an unlimited event stream can be stored, only a selection of the event stream will be stored in a sliding window. The selection is based on a maximal storing time or a maximal number of events.
To migrate the state of an operator, the state must be extracted, transferred and recreated. We illustrate this in Figure 1. To extract the state, the events stored in sliding windows must be read and stored in a transfer format. The recreation of the state requires the processing of all events stored in the extracted sliding windows.
Notably, the described process is based on the used cloud-based CEP system and can vary in other systems. However, in all elastic CEP systems, an operator migration has to be performed, and hence, the basic problem remains the same.

3 Cost-estimation of operator migrations

To estimate the cost of operator migrations during query optimization, several different aspects must be considered. In this section, we describe what effects the operator migration has and specify corresponding cost-models.

3.1 Effects of operator migrations

In order to determine possible effects and costs of the query optimization, we conducted a literature review. To the best of our knowledge, there exists no work in the context of cloud-based CEP about an optimizer that is aware of the optimization cost itself. Since the application field of operator migration in cloud-based CEP systems is similar to virtual machine migration, we adapt existing approaches (see section 5) and apply them for cloud-based CEP systems. We identified three effects of an operator migration in the used CEP system:

Resource consumption. In cloud-based CEP systems, the cost calculation is based on the resource consumption. The operator migration consumes additional resources in order to transfer the operator information, leading to a higher fee for the end-user.

Lost revenue. During the operator migration the event processing must be stopped. Since stopping event processing means that no service can be provided to the costumer, potential revenues are decreased.

Penalty charge. In systems with SLAs, different quality features can be agreed on with the customer (e.g., event processing time). These quality features can be violated during an operator migration because event processing is paused. Therefore, the operator migration may cause additional penalty charges.

The operator migration also influences other aspects, such as the user satisfaction. Since these effects cannot be calculated, they are omitted. In the following sections, we will describe how the different effects of operator migrations can be estimated.

3.2 Resource consumption

The operator movement consumes different kinds of resources. Relevant resources in CEP systems are used CPU time, memory capacity, and network bandwidth capacity. To simplify the following considerations, we do not consider the consumed resources, but the resulting costs. Since the costs of different resources can vary, different constants are used to estimate the cost of the CPU time (mc_{cpu}), memory capacity (mc_{mem}) and network bandwidth capacity (mc_{net}). For stateful operators, the resource consumption for the migration of the state must also be considered. The resource consumption of a state migration is depending on the state size ws and differs from the resource consumption of the migration of the processing information. Therefore, other constants (sc_{cpu}, sc_{mem}, sc_{net}) must be used to estimate the cost for the resource consumption. Since more than one operator can be migrated in one optimization step, the overall cost for the resource consumption mc_d is calculated as sum of all operator migrations:

$$mc_d = \sum_{i=1}^{n} (mc_{cpu,i} + mc_{mem,i} + mc_{net,i}) + \sum_{j=1}^{m} [mc_{cpu,j} + mc_{mem,j} + mc_{net,j} + ws_j \cdot (sc_{cpu} + sc_{mem} + sc_{net})] \quad (1)$$

3.3 Duration of an operator migration

We consider mainly two aspects, to estimate the time of an operator migration mt_o. First, the time to migrate the processing information mt_{oc} (e.g., select or join conditions). Second, the time to migrate the operator state. Since the steps and therefore the time for transferring the processing information is identical for each operator type, we will use a time constant. The time to migrate the state is mainly based on the state size ws, the time to extract and recreate the state st, and the available network bandwidth bw.

$$mt_o = mt_{oc} + ws \cdot st + \frac{ws}{bw} \quad (2)$$

Since the operators cannot be migrated in parallel due to limitations of the used cloud-based CEP system, the time of migrating all operators of a query, is the sum of all migrated

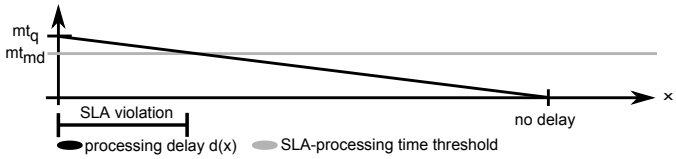


Figure 2: Processing delay $d(x)$ based on the x-th processed event and query migration time mt_q considering the SLAs mt_{md}.

operators ($mt_{o,1}, ..., mt_{o,n}$). Since the used system is a distributed system and only one operator at a time can be migrated, an additional delay mt_d should be considered.

$$mt_q = \sum_{i=1}^{n} (mt_{o,i} + mt_{d,i}) \tag{3}$$

The delay is the time of all operator migrations before the considered operator. Therefore, the delay depends on the order of the operator migration. Since the order cannot be determined in the real system, the worst case will be used for all operators. The worst case for an operator is to be migrated last. In this case, the delay is the sum of the migration time of all other operators of the same node.

$$mt_d = \sum_{i=1}^{k} mt_{oc,k} \tag{4}$$

3.4 Lost revenues and penalty charges

In contrast to a revenue loss, penalty charges can also occur without an optimization (e.g., if a host is overloaded). For this work, we consider only penalty charges caused by the optimization. Lost revenues and penalty charges are both caused by the stop of event processing during the operator migration.
The stop of event processing introduces an additional delay during event processing. The delay is caused by the events stored during the stop of event processing. The storing of events is necessary to guarantee an accurate and correct result of the query processing. In order to avoid the dropping of events (load shedding), enough capacity must be provided to store events. For simplicity, we assume that enough capacity will be provided, and no events must be dropped. Otherwise, the dropping of events must also be considered in the effects of an optimization (e.g., example through a penalty fee for every dropped event).
The number of stored events can be calculated by the event rate of the streams er and the time for the operator migration of the corresponding query mt_q. The behavior of the delay is shown in Figure 2. We assume that after the optimization the delay is decreasing, because enough resources can be provided in a cloud-based CEP system. If enough resources are provided, no node is overloaded. Therefore, more events can be processed than arriving in the system. The processing time of the stored events depends on the processing time of the operators. The delay depends on the order of arrival, because the events will be processed according to that order. The maximal delay is the query migration time mt_q. From this delay, the time of arrival must be subtracted by the event rate and the position of the input stream, $er \cdot x$. Since arriving events are processed in FIFO style, the processing time of

preceding events have to be added., $pt \cdot (x-1)$. The delay of the x-th event can therefore calculated by:

$$d(x) = mt_q - \frac{1}{er} \cdot x + pt \cdot (x-1) \tag{5}$$

The maximal delay mt_{md} of the events without penalty charges is the difference of the maximal processing time lt agreed with the customer and the current query latency ql.

$$mt_{md} = lt - ql \tag{6}$$

Combining equation 5 and 6 leads to:

$$mt_{md} = mt_q - \frac{1}{er} \cdot x + pt \cdot (x-1) \tag{7}$$

This equation can be rearranged to calculate the number of events which have to arrive in order to guarantee that the system response time is smaller than the maximal delay mt_{md}:

$$x = \frac{mt_{md} - mt_q + pt}{-\frac{1}{er} + pt} \tag{8}$$

The number of events can be used to calculate the SLAs-violation time mt_v based on the event rate:

$$mt_v = \frac{x}{er} \tag{9}$$

Inserting equation 8 and rearranging leads to the following equation, which computes the SLA violation time depending on all major factors:

$$mt_v = \frac{mt_{md} - mt_q + pt}{er \cdot \left(-\frac{1}{er} + pt\right)} \tag{10}$$

In general, the processing time between operators differ. A higher event processing time of operators will lead to a higher delay. Since only the maximal delay is of interest, the operator with the maximal processing time will be used to calculate mt_v.
The cost of lost revenues and penalties can be calculated as follows:

$$mc_i = \begin{cases} mt_v \cdot mc_p & \text{if } mt_v > mt_q \\ mt_q \cdot mc_p & \text{otherwise} \end{cases} \tag{11}$$

In case no latency threshold is defined or violated, only the lost revenues during the query migration is considered. Otherwise, the additional time of SLAs-violation is considered. Hereby, mc_p can be a fixed price or a percentage of the cost of the affected query.
The interesting effect is that the system can dynamically decide to violate SLAs or allocate additional resources depending on the caused costs. The overall migration cost considered by the optimization should include the cost for the consumed resources mc_d, the penalty charges, and the lost revenues. Since the cost should be amortized during runtime, the ratio between the estimated costs and periods between the optimizations should be considered during the optimization.

4 Evaluation

In order to answer our research question – *how much can we decrease the overall cost of a cloud-based CEP system using a cost-aware query optimizer considering the cost of the query optimization* – described in the previous sections, we conducted a series of experiments. Hereby, we examined different scenarios with different requirements (query patterns and event rates) based on a real application scenario.

4.1 Experiment Design

There are three different scenarios based on different combination of query patterns and event rates:

- **Scenario 1:** query pattern: linear, event rate: variable
- **Scenario 2:** query pattern: real, event rate: fixed
- **Scenario 3:** query pattern: real, event rate: variable

Since for all queries the same structure (event source, selection, aggregation, and event sink) is used, the load of the system is based on query count. In the linear query pattern, the query count is continuously increased from 1 to a maximum of 33 queries. After reaching the maximum of queries, the query count is constant. In contrast, the real query pattern dynamically changes the query count from 1 to 33 queries. Similar to the query pattern, the event rates can either be fixed ($500\frac{\text{events}}{s}$) or variable (100-1000 $\frac{\text{events}}{s}$). The used events, the real query pattern, and the variable event rates are based on information extracted from the Frankfurt stock exchange market representing real work loads.
In every scenario, five different optimization strategies are evaluated:

- **Baseline:** Optimum
  Considers migration cost: no
- **Variant 1:** Number of considered system states: one
  Considers migration cost: no
- **Variant 1+:** Number of considered system states: one
  Considers migration cost: yes
- **Variant 2:** Number of considered system states: many
  Considers migration cost: no
- **Variant 2+:** Number of considered system states: many
  Considers migration cost: yes

Since the optimization strategies are not the main focus of this work, we refer the interested reader to Rödiger et al. [Rö12].
To have a baseline of the optimization, the search space of the optimization is restricted to a small search space where an optimal solution can be efficiently determined. Hereby, the baseline solution just considers the cost of the current system and the optimized systems ignoring the migration cost. The difference between Baseline and Variant 1 is that Variant 1 uses Recursive Random Search [Rö12] to improve the efficiency of the system. In contrast, Variant 2 analyses several past system states. Variant 1 and Variant 2 were both adapted to consider the operator migration cost (+) using our described cost model (see Section 3).

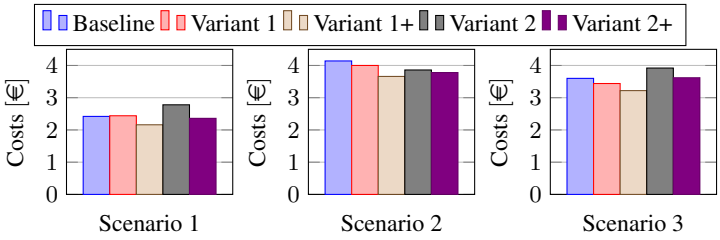


Figure 3: Experiment costs

4.2 Measurement

The experiments were performed on one computing node for measurement and optimization (Intel Core i7 870, 16 GB memory) and four virtual machines (AMD Opteron 6128 HE, 4 GB memory) for the event processing.
Since the load of the system cannot be recreated exactly, and the migration of operators is based on heuristic approaches, a variability of different runs with the same configuration (combination of scenario and optimization strategy) exists. Therefore, every configuration of each experiment is conducted five times. All results for the same configuration are aggregated using the average.
Hereby, the costs and time for the processing of a certain amount of events are measured. The costs are calculated based on the number of used nodes used in fixed time windows. The number of considered time windows n_{tw} is determined by the execution time of the experiment et and the duration of a time dtw

$$n_{tw} = \frac{et}{dtw} \tag{12}$$

For every time window the maximum number of used nodes mn will be determined and charged according to the used pricing model [Mey12] and costs for one node during the time window nc, independent of the real usage of the nodes in the time window.

$$costs = \sum_{i=0}^{n_{tw}} (nc_i \cdot mn_i) \tag{13}$$

The cost calculation is based on commercial cost models (e.g, Amazon E2[1]).

4.3 Evaluation Results

The costs (see Equation 13) for the different scenarios are displayed in Figure 3. In Scenario 1, compared to the baseline, the experiment costs can be reduced if the cost for the operator migration is considered during the optimization. The experiment cost were reduced by 2 % (Variant 2+) to 11 % (Variant 1+). When the different variants with and without the considerations of the migration costs are compared, the consideration of migration costs reduces the costs by 11 % (Variant 1) to 15 % (Variant 2).
In Scenario 2, compared to the baseline, the experiment cost can be reduced if the cost for the operator migration is considered during the optimization. The experiment cost

[1] http://aws.amazon.com/ec2

were reduced by 8 % (Variant 2+) to 11 % (Variant 1+). When the different variants with and without the considerations of the migration costs are compared, the consideration of migration costs reduces the costs by 2 % (Variant 2) to 8 % (Variant 1). Different to the other experiments, variant 2+ only achieves a similar result compared to the baseline in Scenario 3. For variant 1+ the experiment cost were reduced by 10 % compared to the baseline. When the different variants with and without the considerations of the migration costs are compared, the consideration of migration costs reduces the experiments costs by 2 % (Variant 2) to 8 % (Variant 1).

4.4 Discussion

The results of the different experiments show that the monetary costs of the system can be reduced up to 15% if the cost of the optimization is considered during the optimization. The main reason for this is that the number of operator migration is reduced. Hereby, operation migrations are prevented if the gain of efficiency is used up by the cost of the migration. An operator migration will therefore only be executed if it is necessary, for example if a host is overloaded, or if the gain in efficiency exceeds the migration costs. Since the operator migrations are restricted to beneficial migrations, the system needs less time to adapt the query processing. The faster adaption of the query processing leads to an improved efficiency of the overall system resulting in lower monetary cost. The costs of the baseline compared to Variant 1 is unexpected. Normally, the baseline should have better costs in all experiments. Since in two experiments, Variant 1 is better than the Baseline, it seems that the number of experiments is too low to void the effects of the described variety of workload. A gain of 15% is here substantial, because of the potential unbounded runtime of CEP queries. Also the consideration of SLAs during the optimization is leading to a robust optimization, which reduces the probability of SLA violations.

5 Related work

Migration Cost. In the field of virtual machine migration, several relevant effects of a migration were identified. To guarantee the efficiency of the system, the migration itself must be efficient [GSF11]. In the field of virtual machine migration for example the energy [Str12], bandwidth [ASR+10], and memory consumption [WZ11] are relevant. Furthermore, response time [QZW+12], migration costs [SSSS10], and lost revenues [ZZSB13] should be considered.
Since the requirements of virtual machine migration and operator movement in cloud-based CEP systems differs, the existing approaches must be adopted.
For example, the resource consumption differs. In the context of operator movements in CEP systems normally only CPU, main memory, and network traffic are relevant, while the used disk space is negligible. This does not hold in the context of virtual machine migration, where the migration of the disk is the main problem. Also the energy consumption is not relevant in this work, because cloud providers only charge for the used/ reserved amount of CPU, memory, and disk capacity and not for the consumed energy.
Response time, migration cost, and lost revenues are examples for aspects, which hold for both application fields, operator movement in CEP systems and virtual machine migration.

Elastic CEP. Several different approaches exists to provide elasticity for CEP. Hereby, different levels of elasticity can be differentiate. Elasticity can be provided on operator, query, or engine level [Hei11]. By providing elasticity on operator level, instances of an operator process can be added or removed [SAG+09]. If this approach is transferred to the query level, whole queries or subqueries can be replicated and performed in parallel [GJPPMV10]. If statefull operators are processed in parallel, the different instances have to synchronize the global state of the operator leading to an communication overhead. This communication overhead is restricting the system parallelism. Stateless operators can be easily parallized, since no synchronization is needed. If elasticity is provided on engine level, new computing nodes can be added or removed during runtime to migrate existing or execute new operator processes [SHLD11]. A combination of the different levels of elasticity is possible. Nevertheless, no approach is known to estimate operator migration costs or to include operator migration costs of CEP systems into the query optimization.

Operator migration. Our work considers operator migration processes based on the pause-drain-resume-strategy [ZRH04]. There exists also other strategies to migrate operators. For example, new operator processes can be initiated, and executed in parallel to the old operator processes until all needed information (e.g., operator state) from the new and the old operator process match [WB10]. Although the stated considerations must be adapted if another migration strategy is used, the general problem that operator migrations create costs and influence the efficiency of the system still exists. Therefore, the operator migration costs should considered during optimization whatever migration strategy is used.

6 Conclusion

In this paper, we investigated the research problem of cost-aware optimization in cloud-based CEP-systems. We identified factors (e.g., resource consumption) that need to be considered in the optimization process. Based on a cloud-based CEP prototype, cost models are presented estimating the identified factors. To provide a proof of concept, three experiments were conducted based on workloads of a real application scenario of a stock exchange. The results of the experiment show that the monetary costs of the system can be decreased by up to 15 %, when migration costs are considered during optimization.
In future work, a comprehensive statistical analysis is needed to assess the presented cost models conclusively. Since operator migrations do not only influence the efficiency of the system but also other factors (e.g., average query latency), further examinations are needed studying the influence of a cost-aware optimization of the query processing.

References

[ASR+10] Sherif Akoush, Ripduman Sohan, Andrew Rice, Andrew W. Moore, and Andy Hopper. Predicting the Performance of Virtual Machine Migration. MASCOTS, pages 37–46. IEEE, 2010.

[BB05] Shivnath Babu and Pedro Bizarro. Adaptive Query Processing in the Looking Glass. CIDR, pages 238–249, 2005.

[BDG07] Alistair Barros, Gero Decker, and Alexander Grosskopf. Complex Events in Business Processes. In *BIS*, volume 4439 of *LNCS*, pages 29–40. Springer, 2007.

[GJPPMV10] V. Gulisano, R. Jimenez-Peris, M. Patino-Martinez, and P. Valduriez. StreamCloud: A Large Scale Data Streaming System. ICDCS, pages 126–137. IEEE, 2010.

[GSF11] Pablo Graubner, Matthias Schmidt, and Bernd Freisleben. Energy-Efficient Management of Virtual Machines in Eucalyptus. IEEE CLOUD, pages 243–250, 2011.

[Hei11] Thomas Heinze. Elastic Complex Event Processing. MDS, pages 4:1–4:6. ACM, 2011.

[JHJ+10] Gueyoung Jung, Matti A. Hiltunen, Kaustubh R. Joshi, Richard D. Schlichting, and Calton Pu. Mistral: Dynamically Managing Power, Performance, and Adaptation Cost in Cloud Infrastructures. ICDCS, pages 62–73. IEEE, 2010.

[KD98] Navin Kabra and David J. DeWitt. Efficient Mid-Query Re-Optimization of Sub-Optimal Query Execution Plans. *SIGMOD Rec.*, 27(2):106–117, 1998.

[Luc01] David C. Luckham. *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley, 2001.

[Mey12] Patrick Meyer. Spezifizierung, Visualisierung und Bewertung von Eigenschaften monetärer Kosten für Complex Event Processing in der Cloud. Master thesis, Technical University Munich, 2012.

[MSHR02] Samuel Madden, Mehul Shah, Joseph M. Hellerstein, and Vijayshankar Raman. Continuously Adaptive Continuous Queries over Streams. SIGMOD, pages 49–60. ACM, 2002.

[QZW+12] Xiulei Qin, Wenbo Zhang, Wei Wang, Jun Wei, Xin Zhao, and Tao Huang. Optimizing Data Migration for Cloud-Based Key-Value Stores. CIKM, pages 2204–2208. ACM, 2012.

[Rö12] Lars Rödiger. Kosteneffizientes Cloud-basiertes Complex Event Processing. Diploma thesis, Technical University Dresden, 2012.

[SAG+09] S. Schneider, H. Andrade, B. Gedik, A. Biem, and Kun-Lung Wu. Elastic Scaling of Data Parallel Operators in Stream Processing. IPDPS, pages 1–12. IEEE, 2009.

[ScZ05] Michael Stonebraker, Uğur Çetintemel, and Stan Zdonik. The 8 Requirements of Real-Time Stream Processing. *SIGMOD Rec.*, 34(4):42–47, 2005.

[SHLD11] B. Satzger, W. Hummer, P. Leitner, and S. Dustdar. Esc: Towards an Elastic Stream Computing Platform for the Cloud. IEEE CLOUD, pages 348–355, 2011.

[SSSS10] Upendra Sharma, Prashant Shenoy, Sambit Sahu, and Anees Shaikh. Kingfisher: A System for Elastic Cost-aware Provisioning in the Cloud. Technical Report UM-CS-2010-005, University of Massachusetts Amherst, 2010.

[Str12] Anja Strunk. Costs of Virtual Machine Live Migration: A Survey. IEEE SERVICES, pages 323–329, 2012.

[WB10] B. Wolf and I. Behrens. On-the-Fly Adaptation of Data Stream Queries. ISORC, pages 175–179. IEEE, 2010.

[WZ11] Yangyang Wu and Ming Zhao. Performance Modeling of Virtual Machine Live Migration. IEEE CLOUD, pages 492–499, 2011.

[ZRH04] Yali Zhu, Elke A. Rundensteiner, and George T. Heineman. Dynamic Plan Migration for Continuous Queries over Data Streams. SIGMOD, pages 431–442. ACM, 2004.

[ZZSB13] Mohamed Faten Zhani, Qi Zhang, Gwendal Simon, and Raouf Boutaba. Dynamic Migration-Aware Virtual Data Center Embedding for Clouds. IEEE IM, 2013.

ScaDS Dresden/Leipzig: Ein serviceorientiertes Kompetenzzentrum für Big Data

Erhard Rahm, Wolfgang E. Nagel
rahm@informatik.uni-leipzig.de, wolfgang.nagel@tu-dresden.de

Das BMBF hat im Rahmen seiner Förderinitiative zu Big Data die Einrichtung von zwei Kompetenzzentren in Deutschland angekündigt, ScaDS Dresden/Leipzig und BBDC (Berlin Big Data Center). Das "Competence Center for Scalable Data Services and Solutions Dresden/Leipzig" (ScaDS Dresden/Leipzig) (www.scads.de) bündelt die Methodenkompetenz der TU Dresden und der Univ. Leipzig zur ganzheitlichen Adressierung von Big Data Herausforderungen in unterschiedlichen wissenschaftlichen und wirtschaftlichen Anwendungsbereichen. Profilbestimmende Forschungsschwerpunkte liegen in den Gebieten der Datenintegration, der Wissensextraktion sowie der visuellen Analyse. Anforderungen der Datensicherheit und des Datenschutzes werden umfassend berücksichtigt und adressiert.

In einem serviceorientierten, modularen Ansatz werden in ScaDS Dresden/Leipzig Anwendungen nach Gemeinsamkeiten und Unterschieden im Hinblick auf ihre Verarbeitungsmerkmale und Anforderungen in Aufgabenklassen aufgeteilt. Pro Anwendungsklasse soll in einem iterativen Prozess ein umfassendes Konzept für Big Data Services entwickelt und diese als interdisziplinäre Lösungen branchen- und fächerübergreifend zur Verfügung gestellt werden. Erste Anwendungspartner kommen aus den an ScaDS Dresden/Leipzig beteiligten Instituten und aus assoziierten Einrichtungen und Unternehmen. Das Kompetenzzentrum ist darüber hinaus offen für vielfältige Kooperationen, insbesondere zur Nutzung und Weiterentwicklung von Technologien und Service-Konzepten. Zu diesem Zwecke dienen auch regelmäßige Veranstaltungen für Anwender und Wissenschaftler sowie ein internationales Gastprogramm.

Breaking the Chains: On Declarative Analysis and Independence in the Big Data Era

Volker Markl

Database Systems and Information Management
Technische Universität Berlin, Germany
volker.markl@tu-berlin.de

Abstract: Data management research, systems, and technologies have drastically improved the availability of data analysis capabilities, particularly for non-experts, due in part to low-entry barriers and reduced ownership costs (e.g., for data management infrastructures and applications). Major reasons for the widespread success of database systems and today's multi-billion dollar data management market include data independence, separating physical representation and storage from the actual information, and declarative languages, separating the program specification from its intended execution environment. In contrast, today's big data solutions do not offer data independence and declarative specification. As a result, big data technologies are mostly employed in newly-established companies with IT-savvy employees or in large well-established companies with big IT departments. We argue that current big data solutions will continue to fall short of widespread adoption, due to usability problems, despite the fact that in-situ data analytics technologies achieve a good degree of schema independence. In particular, we consider the lack of a declarative specification to be a major roadblock, contributing to the scarcity in available data scientists available and limiting the application of big data to the IT-savvy industries. In particular, data scientists currently have to spend a lot of time on tuning their data analysis programs for specific data characteristics and a specific execution environment. We believe that the research community needs to bring the powerful concepts of declarative specification to current data analysis systems, in order to achieve the broad big data technology adoption and effectively deliver the promise that novel big data technologies offer. In addition, we will present the vision of the Berlin Big Data Center (BBDC) with respect to combining machine learning and data management. A 4-page paper extended abstract of this presentation has appeared at PVLDB Volume 7.

Volker Markl is a Full Professor and Chair of the Database Systems and Information Management (DIMA) group at the Technische Universitat Berlin (TU Berlin) as well as an adjunct status-only professor at the University of Toronto. Earlier in his career, Dr. Markl lead a research group at FORWISS, the Bavarian Research Center for Knowledge-based Systems in Munich, Germany, and was a Research Staff member & Project Leader at the IBM Almaden Research Center in San Jose, California, USA. Dr. Markl has published numerous research papers on indexing, query optimization, lightweight information integration, and scalable data processing. He holds 7 patents, has transferred technology into several commercial products, and advises several companies and startups. He has been

speaker and principal investigator of the Stratosphere research project that resulted in the "Apache Flink" big data analytics system. Dr. Markl currently serves as the secretary of the VLDB Endowment and was recently elected as one of Germany's "digital minds" (Digitale Köpfe) by the German Informatics Society (GI).

Smart Data Innovation Lab - eine Forschungsplattform mit realen Daten

Dr. Harald Schöning

Head of Research
Software AG
Uhlandstr. 12
64297 Darmstadt
Harald.Schoening@softwareag.com

Abstract: Das Smart Data Innovation Lab ist eine nationale Forschungsplattform, die der Wissenschaft Zugang zu realen Industriedaten für Forschungszwecke auf einer leistungsfähigen IT-Infrastruktur verschafft und einen schnellen Ergebnistransfer zurück in die Industrie ermöglicht.

1 Hintergrund

Das Thema „Big Data“ hat in den letzten Jahren an Aufmerksamkeit gewonnen, nicht nur durch Fortschritte in Algorithmen und verfügbaren Softwarelösungen, sondern vor allem durch die rasant anwachsende Verfügbarkeit von Daten. Neben dem wachsenden Verkehr auf sozialen Medien ist die verstärkte Verbreitung von Sensoren und anderer vernetzer Objekte aus dem „Internet of Things“ auch im industriellen Einsatz ein wichtiger Treiber dieser Entwicklung. Mit der Verfügbarkeit dieser Daten geht der Bedarf einher, sie nutzbringend zu aggregieren und auszuwerten. Neue Anwendungen in Industrie, Städten, Medizin, Mobilität usw. werden möglich. Viele Forschergruppen haben den Bedarf erkannt und forschen an neuen Algorithmen und Lösungen. Insbesondere die Zusammenführung von Daten aus mehreren Quellen verspricht neue Erkenntnisse. Oft steht die Forschung allerdings vor dem Problem, dass sie ihre Ergebnisse nicht in realen Anwendungsszenarien mit realen Daten validieren kann, da diese Daten von deren Inhabern nicht frei zur Verfügung gestellt werden. Die Gründe hierfür sind vielfältig. So können solche Daten personenbezogene Informationen enthalten oder Rückschlüsse auf Produktionsverfahren und –geheimnisse zulassen. Andererseits sind die Dateninhaber aber durchaus daran interessiert, aus ihren Daten unter Mitwirkung von Forschern neue Erkenntnisse zu gewinnen.

2 Das Smart Data Innovation Lab

Hier setzt das Smart Data Innovation Lab (SDIL, sdil.de) an, das auf Initiative der AG6 des IT-Gipfels entstand. Es stellt eine nationale Forschungsplattform dar, durch die die

Wissenschaft Zugang zu realen Industriedaten für Forschungszwecke auf einer leistungsfähigen IT-Infrastruktur erhalten kann und ein schneller Ergebnistransfer zurück in die Industrie ermöglicht wird. Forschungsprojekte sollen insbesondere von der Zusammenführung verschiedener Datenquellen (Wissenschaft, Verwaltung, Industrie, kommerzielle Datenquellen) profitieren. Besonderes Augenmerk liegt auf geeigneten Nutzungsbedingungen und Sicherheitsmaßnahmen, die Risiken wie den Missbrauch sensitiver Daten, Streitigkeiten bezüglich des Eigentums an abgeleiteten Daten und Forschungsergebnissen, Rechteverletzungen bezüglich eingelagerter Daten oder angebundener kommerzieller Datenquellen minimieren.
Partner im SDIL sind Dateninhaber, Forschungseinrichtungen und Software- und Hardwarehersteller. In z.Z. vier inhaltlichen Clustern (Innovation Communities, s. Abb. 1) werden unter gemeinsamer Leitung jeweils eines Unternehmens und einer Forschungseinrichtung die Akteure eines Themengebiets vernetzt. Eine eigene Arbeitsgruppe *Data Curation* ist für Datenpflege, -beschreibung und –verwaltung zuständig. Daten können dabei entweder auf der Plattform gespeichert sein oder dort nur registriert sein. In diesem Fall werden sie vom Dateninhaber direkt nach gesonderter Vereinbarung mit den jeweiligen Forschern verfügbar gemacht. SDIL fungiert dabei als Moderator. Die Arbeitsgruppe *Recht* entwirft entsprechende Nutzungsvereinbarungen.
Um eine leistungsfähige IT-Plattform zu gewährleisten, stellen verschiedene Software-Anbieter ihre Produkte zu Verfügung. Der Betrieb wird vom KIT sichergestellt, dazu haben Hardware-Partner Rechnerkapazität bereitgestellt. Abbildung 1 zeigt die Organisationsstruktur des SDIL. Forschungseinrichtungen können sich über den SDIL-Koordinator auf Plattform- und Datenverwendung bewerben.

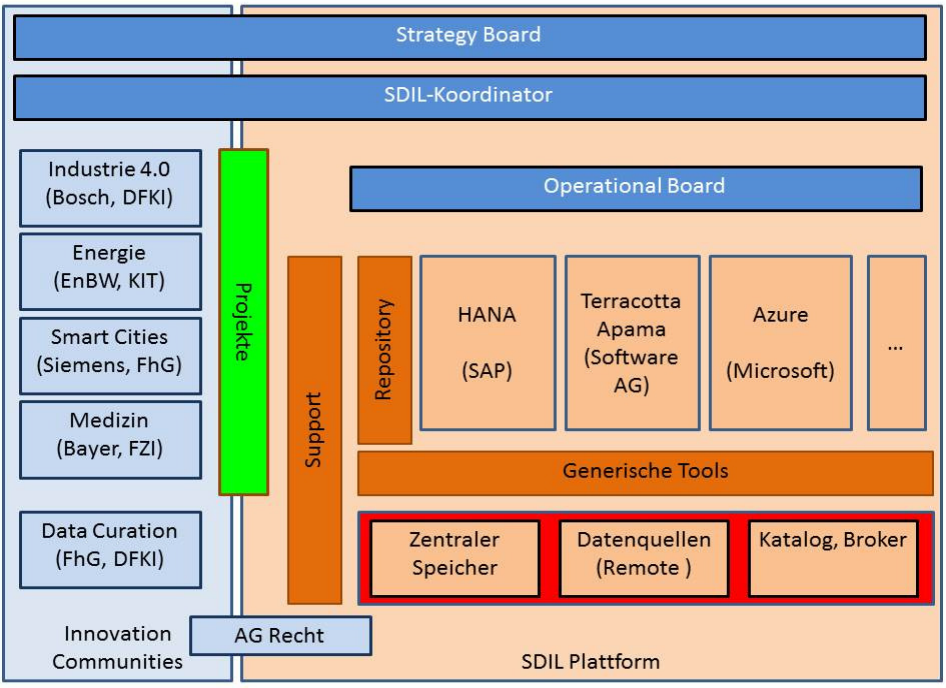


Abbildung 1: organisatorische Struktur des Smart Data Innovation Labs

Towards a Scalable and Unified REST API for Cloud Data Stores

Felix Gessert, Steffen Friedrich, Wolfram Wingerath,
Michael Schaarschmidt, Norbert Ritter

Database and Information Systems Group
University of Hamburg
{gessert, friedrich, wingerath, xschaars, ritter}@informatik.uni-hamburg.de

Abstract: In the last years, many database-as-a-service (DBaaS) systems have started to offer their functionalities through REST APIs. Examples are record stores like DynamoDB and Azure Tables, object stores such as Amazon S3 as well as many NoSQL database systems, for instance Riak, CouchDB and ElasticSearch. Yet today, there has been no systematic effort on deriving a unified REST interface which takes into account the different data models, schemas, consistency concepts, transactions, access-control mechanisms and query languages to expose cloud data stores through a common interface without restricting their functionality. This work motivates the design of such a REST API as well as the challenges of providing it in an extensible, scalable and highly-available fashion. To this end, we propose the REST middleware ORESTES that consists of an independently scalable tier of HTTP servers that map the unified REST API to aggregate-oriented (NoSQL) data stores. It extracts a wide range of DBaaS concerns (e.g. schema management and access control) and provides them in a modular, database-independent fashion at the middleware-level. To tackle the latency problem of cloud-based web applications we introduce the *Bloom filter-bounded staleness* cache consistency algorithm. It leverages the global web caching infrastructure for geo-replication to allow consistent low latency reads. We furthermore show the first steps towards a *Polyglot Persistence Mediator* that exploits the decoupling of the REST API from the data store to route data and operations based on SLAs.

1 Introduction

The success of the database-as-a-service model has brought a wide variety of commercial and research systems. Though many have been extensively studied, there is a lack of taxonomy. Throughout this work, we will distinguish between *managed database services*, *proprietary database services* and *Backend-as-a-Service* (BaaS). In a managed database service a DBaaS provider offers a cloud-deployed DBMS and automates operational tasks such as provisioning, multi-tenancy, backups, security, access control, elasticity, scaling, performance tuning, failover and replication [CJP+11]. Proprietary database services build on newly designed database systems (e.g. Google DataStore based on Megastore [SVS+13]) or integrate different databases to a polyglot persistence environment to offer them through provider-specific protocols and APIs. The Backend-as-a-Service model enhances the DBaaS model by adding abstractions for backend concerns of mobile appli-

cations and web sites (e.g. authentication, push notifications, data validation and assets).

DBaaS systems differ in the degree to which they provide automation of operational tasks, the underlying data store, pricing models, service level agreements (SLAs), multi-tenancy strategies and, most importantly, their interfaces. While Proprietary database services offer provider-specific interfaces, managed database services offer database-specific protocols that were not designed for cloud environments. In this paper, we will focus on *aggregate-oriented* NoSQL services. As described by Fowler et al. [SF12] this encompasses NoSQL databases that have a (potentially denormalized) aggregate as their primary unit of access: *rows* in wide-column and record stores, *documents* in document stores and *key-value pairs* in key-value stores. This aligns well with the resource-oriented model of REST. The key observation and motivation for this work is that aggregate-oriented DBaaS systems can be greatly enhanced by providing them through a single unified and scalable REST API. This is attractive as clients can be reused, data models be shared and applications be migrated. The contributions of this paper are threefold:

1. **Unified REST API**. We propose a Database- and Backend-as-a-Service REST API (see Figure 1) that captures functional concepts (transactions, queries, commutative updates, etc.) across a wide range of systems.
2. **Modular Middleware**. We introduce the modular ORESTES middleware approach, where backend dabatase systems can be exposed through the unified REST API by plugging them into reusable modules (e.g. for partial updates). This yields an architecture where core concepts like authentication and schema management only need to be implemented once and then can be reused for many database systems.
3. **Low latency through consistent web caching**. As achieving low latency is crucial for any user-facing application, we propose the *Bloomfilter-bounded staleness* algorithm to drastically reduce latency through web caching.

The paper is structured as follows. Section 2 outlines the requirements and overall concept before presenting the proposed architecture. In Section 3 an evaluation is given, followed by a discussion of networking aspects relevant for the applicability of REST in the cloud database setting. Section 4 overviews related work and section 5 presents conclusions.

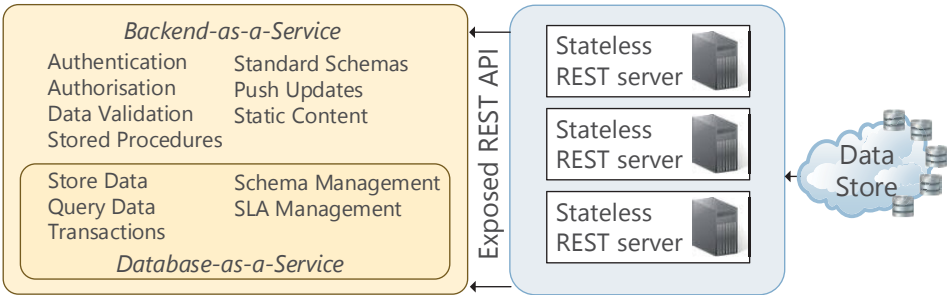


Figure 1: Vision: a scalable REST tier for unmodified, aggregate-oriented data stores.

2 Concept and REST middleware architecture

To expose existing data stores as a Backend/Database-as-a-service without prior modification, the ORESTES middleware and its *unified REST API* have to be powerful enough to expose all possible capabilities of the underlying database (e.g. conditional updates) without compromising its nonfunctional properties (e.g. scalability of data volume or linearizable consistency). We distinguish between two types of modules in the middleware. *Data Modules* express the mapping of data operations in the REST API to the underlying database (CRUD operations, queries, indexing, system configuration). Data Modules have to be implemented for each database that is to be exposed through the unified REST API. *Default Modules* on the other hand implement abstractions that can be provided by default on top of databases through data modules and middleware services (authentication, authorization, data validation, stored procedures, push notifications, transactions, schema management, SLA management, elastic scaling). Default modules can be overwritten to leverage existing native capabilities (e.g. table-level ACLs for authorization).

Today, web performance is governed almost completely by latency [Gri13]. Numerous studies by Amazon, AOL, Google, Bing, Akamai and others demonstrated the severe implications of latency[1]. For this reason, web sites are increasingly built as Single-Page Applications (SPAs) to eliminate the overhead of server-side page generation, client redraws and asset refetching. However, data requests remain extremely latency-critical. We solve the latency problem through web caching of database objects. This is a novel approach, as the web caching model is based on a-priori cache expiration times which necessitates a cache coherence mechanism. In this work we propose a dual strategy: *cache invalidations* perform asynchronous purging of server-controlled web caches (reverse-proxies and CDN nodes) and *Bloomfilter-bounded Staleness* ships probabilistic summaries of recently changed objects to clients that can then revalidate (bypass) stale cached copies (in client caches, forward proxies and ISP caches).

Architecture. Figure 2 shows the ORESTES middleware architecture. Clients that can either be pure DBaaS clients (application servers) or BaaS clients (browsers and mobile devices) use the unified REST API supported by persistence APIs (JDO/Java and JSPA/-JavaScript). At the server-side, HTTP servers wrap a (potentially distributed) data store which currently can be VOD, db4o (OODBMSs), Redis (in-memory key-value store) or MongoDB (document store). Through the stateless design of the servers, latency and throughput is only bounded by the saturated database system as the middleware can scale horizontally. Objects delivered over REST/HTTP are transparently cached for a fixed duration. To allow server-side stored procedures and data validation through `before` and `after` triggers, every server runs a Node.js process which executes registered client code written in the same JSPA browsers run. This server-side code execution is essential for BaaS when business logic should not be exposed and validation of user input is needed.

REST API. The unified REST API is composed of different modules, as shown in Figure 3a. It is specified through a new REST specification format similar to routing lan-

[1] Amazon found that with every 100ms of additional page load time, revenue decreases by 1%. AOL discovered that all users in the 10%-quantile of fastest users visit 7.5 pages, slower users only 5. When increasing load time of search results at Google by 500ms, traffic decreases by 20%. See: http://velocityconf.com/velocity2009

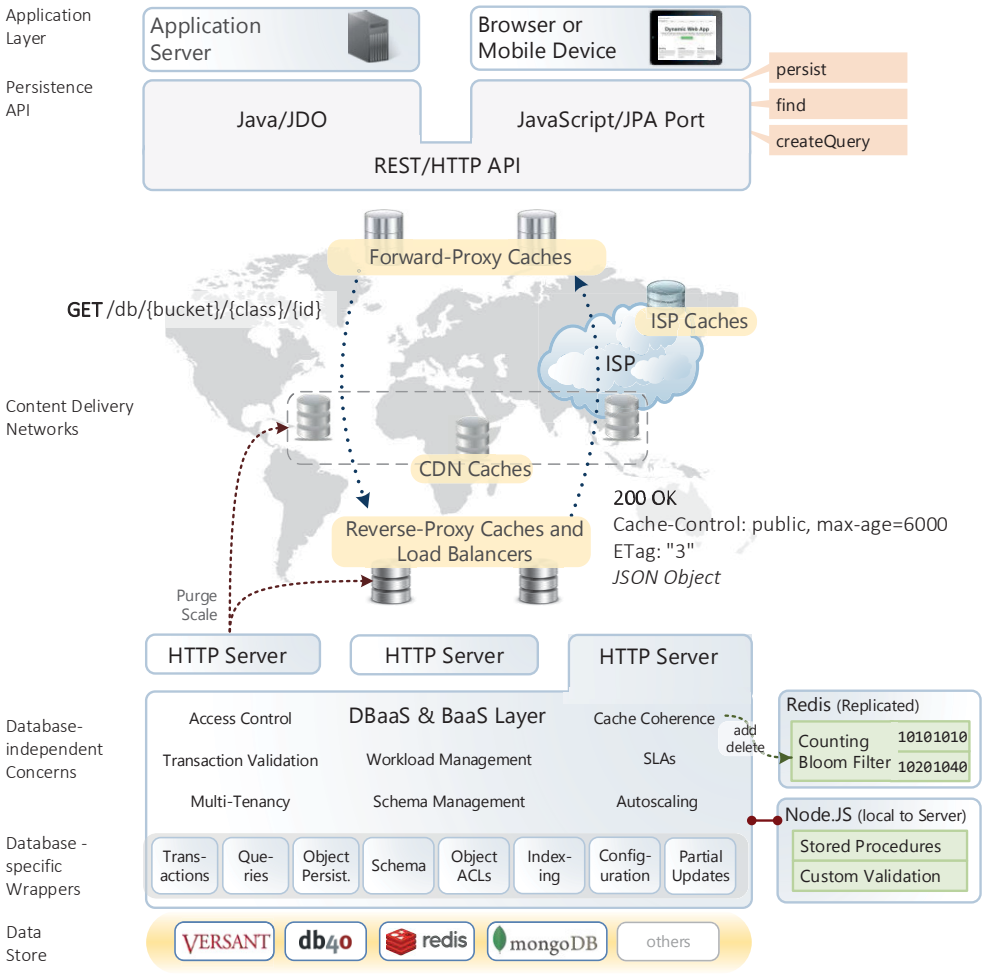
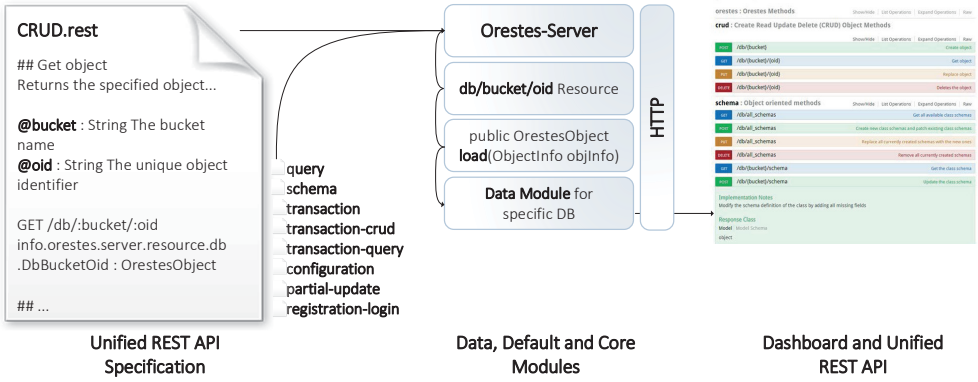


Figure 2: ORESTES middleware architecture with an exemplary object request.

guages in MVC web frameworks (e.g. Play) but enhanced to also capture descriptions and types of parameters and return values. It describes the effects of HTTP methods on resources identified by URI patterns. The REST specification is loaded by the ORESTES HTTP servers for validation, conversion (e.g. from JSON to a schema object) and to generate an interactive REST API documentation on the server's dashboard. The operations declared in the REST specification correspond with one or more methods in the data and default modules that in turn communicate with the underlying database via its specific drivers and protocols. For an end-to-end example, consider a web application in which a user loads his profile. The call is done through the JavaScript JSPA persistence API: `entityManager.find(id)`. It employs the unified REST API: `GET`

`db/profiles/id`. Suppose the ORESTES middleware wraps a MongoDB cluster. After parsing and checking the request, the server will call the CRUD data module's `load` method for MongoDB. The data module will issue a `db.find()` query to MongoDB that returns the document. Based on the schema defined for `Profile`, the server returns the requested object. The fully-typed profile object returned by JSPA can afterwards be displayed, for instance by feeding it into a template of an SPA framework like Angular.js or Backbone. Figure 3b provides examples of other REST API methods.



(a) REST specification lanaguage and its interpretation.

Request	Response	Explanation
`POST /db/:bucket` `JSON-Object`	Created object including assigned object id (oid) and version number	Creates a new object.
`PUT /db/:bucket/:oid` `JSON-Object`	Created/replaced object	Replaces or creates an object using an object id. The request can be conditioned on a version.
`GET /db/:bucket/:oid`	Database object	Fetches or revalidates an object. Request can be answered by web caches.
`GET /db/:bucket?query` `&start=0&count=-1`	List of matching ids	Executes a DB-specific ad-hoc query.
`GET /db/all_schemas`	All schemas for all classes	Retrieves all schemas.
`POST /transaction`	ID and URL of the transaction	Starts a new transaction.
`POST /db/:bucket/:oid/:field`	Success or validation failure	Performs a partial update (e.g. counter increase).

(b) Example requests from different modules.

Figure 3: Composition of the unified REST API through modules.

Schema Management. ORESTES employs explicit object-oriented schemas as a default module. Schema-free databases (Redis, MongoDB) thus get "bolt-on" rich schemas, whereas schema-aware databases (db4o, Versant) can expose their own schemas. We think that explicit schemas are an advantage as they allow type-checking and validation to prevent data corruption. An ORESTES-schema is a mapping of field names to types. A type can be *primitive* (String, Boolean, Integer, Float, Date, GeoPoint), a *reference* (object Id), a *collection* (Set, Map, List), an *embedded object* (defined by a schema) or *JSON* (Array, Object). Through the JSON types the schema grants all freedoms of schema-free data stores. However, this lack of structure is usually unnecessary as schemas can be asynchronously evolved. Every ORESTES server holds all schemas in memory and broad-

casts client-initiated schema changes to all other servers. There are two kinds of schema changes: *Safe Changes* (adding fields, changing field types to a parent type) are commutative, associative and idempotent and thus can be safely transferred asynchronously. As safe changes are non-destructive, they can be lazily applied. *Unsafe Changes* (deleting and renaming fields, changing field types to a non-parent type) have side-effects and need to be broadcasted in a coordinated, blocking fashion using 2PC.

Access Control. To extend a DBaaS to a BaaS, user and access control are required as clients cannot be trusted. In ORESTES this is achieved by role-based access-control (RBAC) and default schemas for common BaaS use cases: users, groups, installations, messages, social posts, in-app purchases etc. These BaaS schemas have special semantics. For instance, a user instance automatically gets created when a user registers via OAuth or OpenId and logins are checked against the user database object. Registration and login can thus be provided as database-independent default modules. On successful login, an access token is generated, which uses an enhanced scheme of Fu et al. [FSSF01]: $expires = t \,\&\, data = u \,\&\, digest = HMAC_k(exp = t \,\&\, data = u)$ where $u = id(user) \,\&\, \{id(role) \mid user \in role\}$ and $k = tenant_id \,\&\, server_secret$, i.e. the token contains the user's id and his roles and a signature (a HMAC for performance reasons). Based on this token the servers check *schema-level ACLs* and *object-level ACLs*. An ACL grants or denies access to users and roles for certain operations: *read* and *write* on object-level and *read*, *query*, *create*, *delete*, *extend schema*, *subclass schema* on schema-level. In contrast to schema-level ACLs, object-level ACLs need a specific data module for mapping. In MongoDB, for instance, object level ACLs are enforced by conditioning data operations over a per-object, indexed ACL field containing allowed and denied users and groups. As the access token is self-contained, no shared server state is required. Even reverse-proxy caches and CDNs can perform ACL checks for read operations if they support sufficiently expressive configuration languages (e.g. Varnish VCLs). To enable this, a merged view of schema- and object-level ACLs is attached to objects, which are used by these caches and stripped from objects before sending them to clients. To execute critical business logic, server-side JSPA scripts can be registered as stored procedures, before- or after-handlers. Before- and after-handlers are called with the operation and schema they are defined for (e.g. to validate user input), whereas stored procedures are explicitly invoked over the REST API (e.g. to submit an order). Through Node.js, JSPA can likewise be used on the client and server. REST API calls are handled via inter-process communication between the Java-based ORESTES servers and the scripts in Node.js.

Polyglot Persistence. Making use of the loose coupling provided by the unified REST API, we introduce a *Polyglot Persistence Mediator* (PPM), which automatically routes data to different backend data stores. Decision-making is based on a per-schema SLA, which combines different functional and nonfunctional requirements. The prototype supports a decision between MongoDB and Redis with MongoDB as the default principal storage facility. The SLA allows for simple annotations like saving an attribute in a certain datastructure (e.g. priority queue or hash-set) for fast updates and special queries (e.g. top-k) as well as for more refined, parameter based requirements. For instance, a hotspot-field could be annotated with a certain maximum write latency. The persistence mediator takes the schema annotations to make a runtime decision for the appropriate backend, based

on a scoring of available backends for the demanded SLA. Similarly, fields can be annotated with particular boundaries on latency, availability and replication factors. Cross-data store materialization is handled through parallel background processes in each ORESTES server, which guarantee a strict upper time limit for each objects materialization. This new approach enables applications to make use of polyglot persistence on a declarative basis. Though this is ongoing work, we present some promising preliminary results in Section 3.

Consistent Caching. If read access is not restricted by ACLs, objects are returned as publicly cachable, i.e. carrying a `Cache-Control:public,max-age=ttl` header, so intermediate caches save object copies for `ttl` seconds. We define 4 consistency levels:

Read-Any (RA) clients may receive any previously cached version of an object; staleness is only bounded by the expiration time `ttl`.

Read-Newest (RN) explicitly fetches the newest object version using an HTTP cache revalidation request, i.e. through refreshing intermediaries.

Transactional (TA) clients read-sets and write-sets are validated and checked for stale reads at commit time. This approach is cache-compatible and conceptually similar to the strategy employed for optimistic transactions in Google's F1 [SVS+13] and the Omid system [GJK+14] and have been described for ORESTES in [GBR14].

Bloom-Filter-Bounded Staleness (BFB) By loading a Bloom filter of recent changes, clients are guaranteed to see only object versions that are at least as recent as the database state by the time the Bloom filter was generated.

Read-Any has the strongest latency benefit – in the best case, the object is resident in the client cache and even a round-trip to a content delivery network (CDN) is cheap (in the order of 20ms). *Read-Newest* guarantees freshness at the cost of a full round-trip to the cloud data-center. As shown in Figure 2, web caches are transparently leveraged at HTTP level. While client-, ISP- and forward-proxy caches cannot be invalidated by the server, CDNs and reverse-proxies support explicit invalidations. ORESTES uses a plug-in mechanism to propagate updates in their respective purging protocols. SLAs provided by CDNs differ significantly, for instance Amazon Cloudfront rate-limits total invalidations and takes more than 10 minutes to apply them, whereas other CDNs (e.g. Fastly) grant arbitrary invalidation frequencies and instant application. This has to be taken into account when coupling the REST API to a CDN as it strongly effects cache-hit ratios and staleness.

BFB solves the following fundamental problem: if objects are cached for a predefined duration `ttl`, ad-hoc changes will result in stale cache reads. BFB introduces the new idea of transferring the cache invalidation task from the server to the client. To achieve this, the server tracks all objects that have been updated in the last `ttl` second in a shared Counting Bloom Filter. It represents the set S of potentially stale objects o_i, i.e. all objects whose logical write timestamp $WTS(o_i)$ lies less than `ttl` seconds in the past and is not equal to their creation timestamp $CTS(o_i)$: $S = \{o_i \in db | now() - ttl > UTS(o_i) \wedge WTS(o_i) \neq CTS(o_i)\}$. Upon connection, the flat Bloom filter is piggybacked to the client. Before retrieving an object over the REST API, the client performs an $O(1)$ hash-lookup in the Bloom filter to decide whether the object should be fetched normally, i.e. allowing cache-hits or if a HTTP revalidation should be attached to the request, i.e. letting caches validate the freshness by the server through an `If-None-Match:version-number` condition. The benefit of this scheme is that any HTTP-compliant cache is implicitly used by

the unified REST API, yielding the largest conceivable geo-replication system without any up-front deployment and with a strong guarantee – the age of the Bloom filter is a strict lower time bound for staleness.

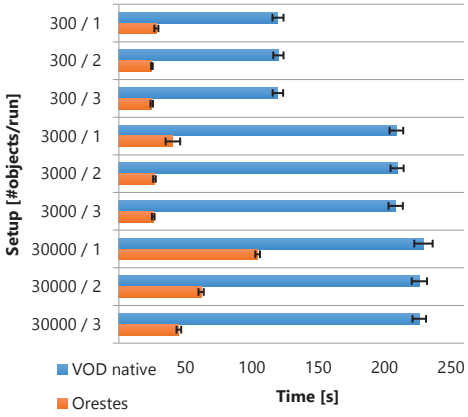
The Bloom filter is very efficient for representing the objects in S, since due to its probabilistic nature it is very small. A Bloom filter with m bits using k hash functions to store n distinct objects, has a false-positive rate $f \approx (1 - exp(-kn/m))^k$. This allows to trade off false-positives (revalidations instead of cached reads) against size, i.e. the overhead of fetching the Bloom filter. For instance, with $f = 0.01$, each object requires 9.6 bits. So if 1000 individual objects were changed in the last `ttl` seconds, only 1.17 KB would need to be transferred. At the server-side, a Counting Bloom filter is employed to support object removal. As every update operation leads to an addition in the filter, it has to support high update rates as well as fast retrieval of the flat Bloom filter. As shown in Figure 2, this is achieved by storing the Counting Bloom filter as an in-memory bit vector (in Redis) using a scheme where each bit-position b_i is represented by a materialized Bloom filter bit and the respective counter: $b_i = (count > 0, count)$. The materialized bits are stored contiguously, i.e. to fetch the whole Bloom filter it can be directly read from main memory. To remove objects from the filter that have not changed in the last `ttl` seconds, each server maintains a priority queue of objects inserted into the filter. As time advances the top elements from priority queue are successively evicted from the filter. The Counting Bloom filter can also be employed to estimate the update frequency of objects to compute their optimal `ttl` value. The enabling technique, known as *spectral Bloom filters*, performs cardinality estimation of an object using its counter values [BM02]. This estimation divided by the current `ttl` yields the update frequency. Our open-source Bloom filter framework, including a detailed statistical and performance analysis, can be found online[2]. Bloom filter-bounded staleness is ongoing work. We are currently working on a sharded Counting Bloom filter as a *Commutative Replicated Data-Type* (CRDT) to also serve it in a highly available eventually consistent manner, as well as statistical framework for the optimal choice of Bloom filter parameters and caching durations.

3 Evaluation

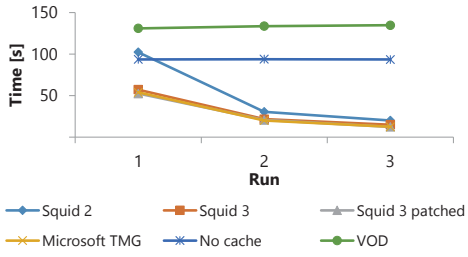
We conducted several experiments to evaluate the performance implications of the REST/HTTP layer and the effects of caching. Figure 4a shows the results of a setup where 50 client VMs using a forward-proxy cache and ORESTES with VOD are separated by a network latency of $165ms \pm 2ms$ using the Amazon EC2 regions Ireland and California. The clients concurrently execute a microbenchmark modeled after a social networking scenario (500 operations, read/write ratio 90%/10%, navigational access with sporadic queries). There are three consecutive runs for different sizes of the database comparing ORESTES exposing VOD against the native binary VOD protocol. The graph shows that the average execution time is heavily reduced with ORESTES as a consequence of its caching approach. Figure 4b reports the results of the same microbenchmark for a setup of hardware machines using a single client, different caches (both in Hamburg) access-

[2] github.com/DivineTraube/Orestes-Bloomfilter

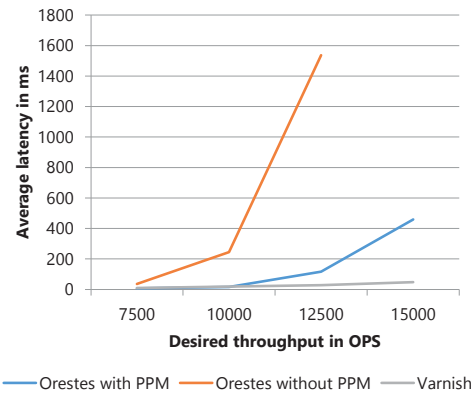
ing ORESTES/VOD (California). The results show that the large performance advantage is consistent across different web caches. This is a consequence of ORESTES relying on standard HTTP caching. Please note that both experiments only consider forward-proxy caching. In our current work, we are tackling the problem of also considering client caches, CDNs and reverse-proxies, which are likely to yield even more drastic results.



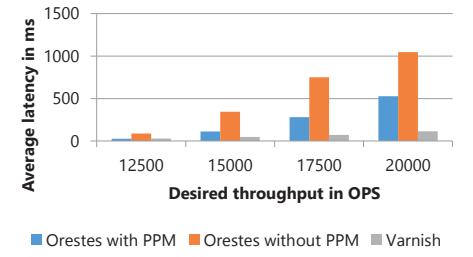
(a) Read-heavy cloud scenario.



(b) Behaviour of different web caches.



(c) Cloud-based PPM Evaluation.



(d) Effects of increased DB performance.

Figure 4: Evaluation of ORESTES, its caching behaviour and the PPM.

We also report preliminary results for the Polylgot Persistence Mediator. The evaluation considers a web site/app displaying top-k elements based on a counting metric in content objects (e.g. impressions for a news article). The schema SLA specifies that counter increases and top-k queries should both be performed with small latency. The PPM in ORESTES, which is backed by MongoDB and Redis, routes counter increases to a Redis *sorted set* mapping counters to object ids. The PPM materializes counters from Redis to MongoDB in 1 minute intervals so they can also be accessed through MongoDB-queries (for other queries than top-k). The experiment shown in Figure 4c is performed on Amazon EC2 using a c3.4xlarge instance for both the client and ORESTES server and a weaker m1.large instance for each Redis and MongoDB. MongoDB is attached to an EBS storage volume with 1000 provisioned IOPS and is accessed using the `acknowledged` write concern that does not guarantee persistence. Figure 4c shows the throughput of performing

100k counter updates comparing ORESTES with PPM to ORESTES with MongoDB. This is compared to the HTTP performance "gold standard" by replacing the ORESTES server by a Varnish reverse-proxy serving a static response from memory. The graph shows that the PPM is able to handle much higher throughput with lower latency by using the SLA to route counter updates to Redis. It is also evident that the overhead of the ORESTES layer is very low – despite additional database communication latency remains comparable to Varnish up to 2/3 of the maximum throughput. Figure 4d shows a consistent result when instance sizes of the database machines are increased to c3.4xlarge.

HTTP enhancements. Though HTTP is widely used for cloud data management, it has restrictions. The evolving HTTP/2.0 standard will address some performance problems of HTTP/1.1 through multiplexing, binary encoding, header compression and server push. However, it leaves HTTP semantics unchanged. We propose two semantic improvements to HTTP for REST-based cloud data management: *variable-precision timestamps* and *multipart caching*. Timestamp-based versioning schemes of underlying databases cannot be exposed as HTTP timestamps – their precision is limited to seconds. This lack of precision forces REST APIs to expose database timestamps as opaque ETags. The cost is that ordering semantics are lost and HTTP conditional requests (e.g. If-Modififed-Since) are not possible. In recent NoSQL databases timestamp-based versioning is common, for instance MongoDB's ids are prefixed with a 32-bit timestamp, HBase uses ms-timestamps for cell-level-versioning and Cassandra generates μs-timestamps for its last-write-wins concurrency. Therefore we suggest to extend the *HTTP-date* type defined in RFC 2616 [FGM+99] to also allow UTC timestamps (ISO 8601) that support arbitrary second fractions: `HTTP-date = rfc1123-date | rfc850-date | asctime-date |` **`ISO-8601-date`**. A second suggested improvement is *multipart caching*. Multipart messages (RFC 1341) are used in HTTP to send a collection of resources in bulk (e.g. query result sets). However, these resources do not carry individual metadata, only encoding information. We propose a `multipart/resources` type allowing Cache-Control headers as well as canonical URIs. This eliminates additional round-trips and makes GET request profit from previous bulk replies.

Networking improvements. We also found that cache-aware REST APIs for cloud databases lead to two problems: *falsely non-persistent HTTP methods* and *temporary TCP deadlocks*. We encountered non-persistent methods in the Squid proxy cache which is arguably the best-researched web cache [Nag04] and heavily used by large-scale web sites such as Wikipedia and Flickr. In Squid, persistent TCP connections [FGM+99] were handled incorrectly as depicted in Figure 5, where the server connection is terminated after just one request. This misbehavior was caused by falsely treating requests containing a body as non-idempotent, instead of only POST requests. Since the default rule for non-idempotent messages was to open a new connection as a precaution, this caused at least one costly additional round-trip (TCP handshake) for POST/PUT/DELETE requests. We reported and corrected this issue which lead to significant performance improvements. A second issue with an even higher performance impact is still present in all current versions of Squid, Microsoft TMG and possibly other caches: a temporary TCP deadlock of 200-500ms (depending on the OS). It is caused by an interference of the *Nagle algorithm* and the *delayed ACK algorithm* which are defined in the TCP protocol standard to increase

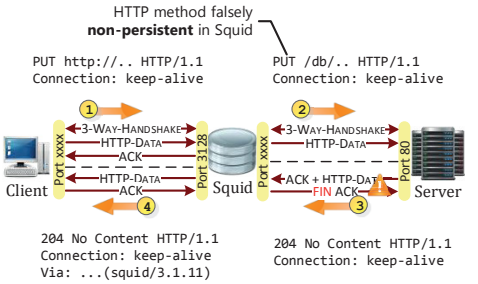


Figure 5: Non-persistent HTTP methods.

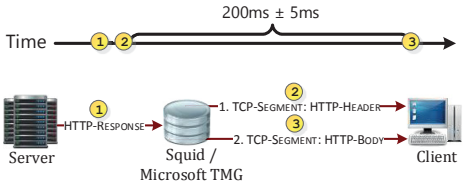


Figure 6: Temporary TCP deadlock.

the effectiveness of TCP buffer management. Figure 6 illustrates the problem: **(1)** When the server returns the HTTP response, the web cache processes the HTTP header and body and passes them to the system's socket interface for forwarding in two consecutive steps. In modern operating systems, the Nagle algorithm is enabled by default and follows the intuition of buffering data until new ACKs arrive or a complete TCP segment can be filled. **(2)** As there is no outstanding ACK from the client, the HTTP header will be sent immediately. The client receives the first packet but does not instantly acknowledge it, as its socket's delayed ACK algorithm tries to send ACKs piggybacked with data or for every second full-sized segment it receives. **(3)** As HTTP is a request/response protocol, the client has no data to send, so the ACK is delayed until a timeout of 200ms (depending on the OS) occurs. The web cache's Nagle algorithm then detects the incoming ACK and sends the queued HTTP body segment. Affected systems (e.g. Squid and TMG) should amend the problem by buffering data in the application or turning off the Nagle algorithm. The temporary deadlock is detrimental to the performance of REST APIs, in particular for latency-sensitive cloud database workloads. Until the problem is addressed in the TCP specification, we suggest a temporary solution: small HTTP messages can be filled with whitespace characters to reach the full segment size that triggers immediate propagation.

4 Related Work

REST APIs for DBaaS systems have been tackled from different directions. Google's GData and Microsoft's OData are two approaches for a standard HTTP CRUD API that are used by some of their respective cloud services. Many commercial DBaaS systems offer custom REST APIs tailored for one particular database (e.g. DynamoDB, Cloudant). A first theoretic attempt for a unified DBaaS REST API has been made by Haselman et al. [HTV10] for RDBMSs. Unlike the Orestes REST API, these approaches cannot make use of caching and do not offer a scalable middleware for exposing arbitrary aggregate-oriented data stores, nor can they be used in a BaaS setting. Two older approaches for reducing database access latency are DBCache and DBProxy [APTP03]. Both require a full-fledged RDBMS as a proxy. Bloom filters are widely used for content summaries [BM02] (e.g. in Summary Cache [Nag04]). To our best knowledge, though, they have

never applied to client-driven cache consistency. A different approach to low latency is geo-replication which can be either synchronous (e.g. in MDCC, Spanner, MegaStore [SVS+13]) or asynchronous (e.g. BigTable/HBase, Cassandra, Riak). Unlike the web caching approach, this requires database-specific protocols and deployments.

5 Conclusions

In this paper, we present a unified modular REST API that can expose aggregate-oriented data stores as a Database- and Backend-as-a-Service. The ORESTES middleware provides the REST API in a scalable, extensible fashion allowing data stores to easily plug-in and reuse common functionality like schema management, authentication, access control, caching and transactions. To solve the latency problem of the modern web, we proposed the BFB cache consistency scheme that offloads revalidations to clients using an efficiently managed Bloom filter representing a sliding window of recently changed objects. We also present the first steps towards a Polyglot Persistence Mediator that routes data to different storage backends based on declarative SLAs. The evaluation shows that the unified REST/HTTP API achieves high throughput and very low latency.

References

[APTP03] K. Amiri, S. Park, R. Tewari, and S. Padmanabhan. DBProxy: A dynamic data cache for Web applications. In *Proceedings of the ICDE*, page 821–831, 2003.

[BM02] A. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey. In *Internet Math.*, 2002.

[CJP+11] Carlo Curino, Evan PC Jones, Raluca Ada Popa, Nirmesh Malviya, et al. Relational cloud: A database-as-a-service for the cloud. In *Proc. of CIDR*, 2011.

[FGM+99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and other. RFC 2616. 1999.

[FSSF01] Kevin Fu, Emil Sit, Kendra Smith, and Nick Feamster. The Dos and Don'ts of Client Authentication on the Web. In *USENIX Security Symposium*, page 251–268, 2001.

[GBR14] Felix Gessert, Florian Bücklers, and Norbert Ritter. ORESTES: a Scalable Database-as-a-Service Architecture for Low Latency. In *CloudDB*, 2014.

[GJK+14] Ferro Daniel Gómez, Flavio Junqueira, Ivan Kelly, Benjamin Reed, et al. Omid: Lock-free Transactional Support for Distributed Data Stores. In *ICDE*, 2014.

[Gri13] Ilya Grigorik. *High performance browser networking*. O'Reilly Media, 2013.

[HTV10] T. Haselmann, G. Thies, and G. Vossen. Looking into a REST-Based Universal API for Database-as-a-Service Systems. In *CEC*, page 17–24, 2010.

[Nag04] S. V. Nagaraj. *Web caching and its applications*, volume 772. Springer, 2004.

[SF12] Pramod J. Sadalage and Martin Fowler. *NoSQL distilled*. Pearson Education, 2012.

[SVS+13] Jeff Shute, Radek Vingralek, Bart Samwel, Ben Handy, et al. F1: A distributed SQL database that scales. *Proceedings of the VLDB Endowment*, 6(11):1068–1079, 2013.

Datenmanagement in der Cloud für den Bereich Simulationen und Wissenschaftliches Rechnen

Peter Reimann, Tim Waizenegger, Matthias Wieland und Holger Schwarz

Institut für Parallele und Verteilte Systeme, Universität Stuttgart
Vorname.Nachname@ipvs.uni-stuttgart.de

Abstract: Für Organisationen, die Simulationen nicht als ihr Kerngeschäft verstehen und sie daher eher sporadisch durchführen, lohnt sich der Betrieb einer eigenen Recheninfrastruktur nur selten. Dies betrifft z. B. kleine und mittlere Unternehmen sowie einige wissenschaftliche Institutionen. Besserung können öffentliche Cloud-Infrastrukturen als Plattform für die Ausführung von Simulationen verschaffen. Das Datenmanagement in der Cloud ist aber speziell für den Bereich Simulationen noch weitgehend unerforscht. In diesem Beitrag identifizieren wir daher noch offene Fragestellungen bzgl. des Datenmanagements von Simulationen in der Cloud. Dies betrifft vor allem die Datenbereitstellung und inwieweit nutzer- und simulationsspezifische Anforderungen an das Datenmanagement in der Cloud eingehalten werden können. Wir untersuchen Technologien, welche sich diesen Fragestellungen widmen, und diskutieren, ob und wie sie in der Cloud sowie für Simulationen einsetzbar sind. Weiterhin skizzieren wir wichtige zukünftige Forschungsthemen.

1 Einleitung

Einige Organisationen, z. B. kleine und mittlere Unternehmen oder manche wissenschaftliche Institutionen, führen Simulationen eher sporadisch durch, sodass sich der Betrieb einer eigenen Recheninfrastruktur nur selten lohnt. Eine Alternative sind öffentliche Cloud-Infrastrukturen als Plattform für die Ausführung von Simulationen. Ein Kernaspekt von Cloud-Infrastrukturen ist, dass Nutzer der Cloud die für ihre Anwendungen benötigten Ressourcen nicht explizit vorhalten müssen, sondern sie dynamisch und schnell vom Cloud-Anbieter bereitgestellt bekommen. Ebenso können die Ressourcen bedarfsgerecht und elastisch an die individuell benötigte Leistungsfähigkeit angepasst werden, von sehr kleinen Berechnungen bis hin zu komplexen Anwendungen mit hoher Parallelität. Gerade wenn Simulationsberechnungen nur sporadisch durchgeführt werden, bieten diese bedarfsgerechten, dynamischen und kostengünstigen Provisionierungsmöglichkeiten einen erheblichen Vorteil gegenüber einer eigenen Recheninfrastruktur.

Cloud-Anbieter stellen ihre Dienste in verschiedenen Abstraktionsstufen bereit, die in Infrastructure-, Platform- und Software-as-a-Service eingeteilt werden (IaaS, PaaS, SaaS) [MG11]. Auf der untersten Ebene bieten sie mit IaaS-Diensten virtualisierte Hardware wie Rechenleistung, Speicher und Netzwerk an, die universell zum Betrieb beliebiger Anwendungen dienen. In PaaS-Angeboten werden Plattform- oder Middleware-Dienste zur

Verfügung gestellt, z. B. Datenbankdienste oder Plattformen für Web-Anwendungen. Die Kategorie SaaS umfasst vollständige (Anwender-)Software, deren Betrieb und Wartung der Cloud-Anbieter übernimmt. Zur Provisionierung und Verwaltung solcher Dienste gibt es bereits eine Reihe von Lösungen. Ein repräsentatives Beispiel ist der OASIS-Standard *Topology and Orchestration Specification for Cloud Applications* (TOSCA) [OA13], der auch schon im Bereich von Simulationen erfolgreich Einzug gehalten hat [VHKL13]. Die spezifischen Fragestellungen, die sich Cloud-Nutzer bzgl. des Datenmanagements ihrer Simulationsanwendungen in der Cloud stellen, blieben bisher aber weitgehend unbeachtet:

- Wie können Nutzer Eingabedaten für eine Simulation in der Cloud bereitstellen, und wie werden ihnen die Ergebnisdaten wieder zur Verfügung gestellt?
- Wie können Nutzer sicherstellen, dass alle ihre individuellen nicht-funktionalen Anforderungen an das Datenmanagement in der Cloud eingehalten werden?

Bleiben diese Fragen unbeantwortet, kann sich dies negativ auf die Nutzerakzeptanz von Cloud-Diensten für Simulationen auswirken. Dieser Beitrag liefert dazu folgende Mehrwerte und ist wie folgt aufgebaut:

Abschnitt 2 leitet – anhand einer Beispielsimulation von Strukturänderungen in Knochen [Kr13] – die wichtigsten Anforderungen her, die aus Nutzersicht für das Datenmanagement in Simulationen beachtet werden müssen. In Abschnitt 3 wird der Einsatz des OASIS-Standards TOSCA für die Bereitstellung von Simulationssoftware beschrieben. Abschnitt 4 diskutiert die erste offene Fragestellung bzgl. der Bereitstellung von Eingabedaten und der Rückgabe der Ergebnisdaten von Simulationsabläufen. Zudem wird ein Vorschlag für die Umsetzung dieser Datenbereitstellung und Datenrückgabe vorgestellt, der auf dem SIMPL-Rahmenwerk (*SimTech, Information Managemenent, Processes, and Languages*) basiert [RS13, Re11]. Abschnitt 5 widmet sich Technologien zur Einhaltung der in Abschnitt 2 identifizierten Nutzeranforderungen. Dabei diskutieren wir, inwieweit diese Technologien in der Cloud einsetzbar sind, und skizzieren wichtige zukünftige Forschungsthemen. Abschnitt 6 fasst abschließend die wichtigsten Aspekte zusammen.

2 Datenmanagement in Simulationen

Die Untersuchung komplexer Probleme erfordert häufig die Kopplung von Simulationsmodellen verschiedener wissenschaftlicher Anwendungsgebiete. Als Beispiel betrachten wir eine Simulation von zeitabhängigen Strukturänderungen in Knochen, die z. B. bei Heilungsprozessen nach Knochenbrüchen relevant ist [Kr13]. Abb. 1 zeigt, wie diese Simulation Modelle aus den Anwendungsgebieten Biomechanik und Systembiologie koppelt. Das biomechanische Simulationsmodell beschreibt das Verhalten von Knochen auf einer makroskopischen Gewebeebene. Das auf der Finite-Elemente-Methode (FEM) basierende Pandas-Rahmenwerk[1] bietet hierzu eine numerische Implementierung. Es wandelt auf Knochen wirkende externe Belastungen in eine charakteristische Lösung für die interne

[1] http://www.mechbau.uni-stuttgart.de/pandas/index.php

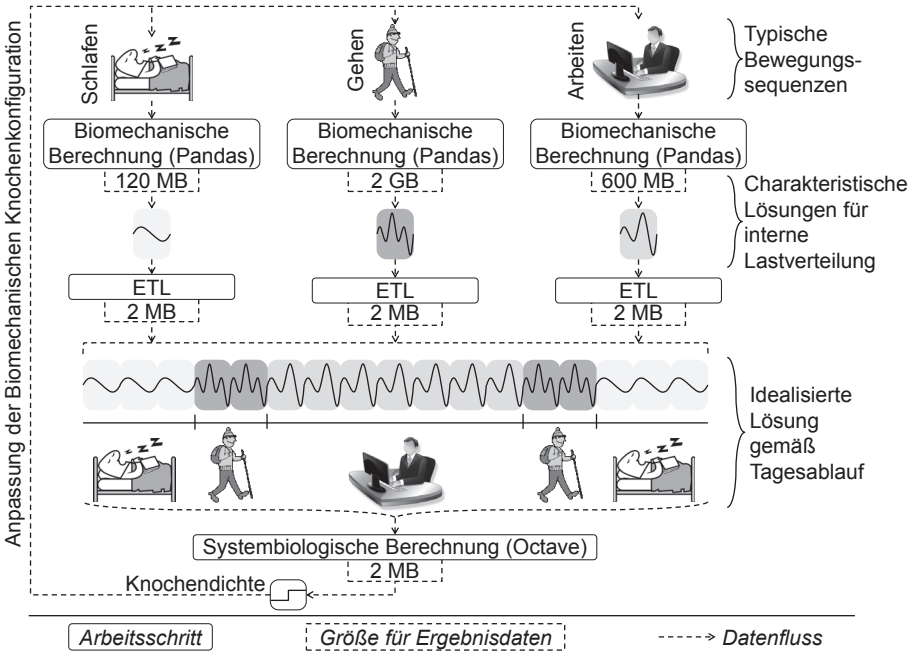


Abbildung 1: Prozess für die gekoppelte Simulation von Strukturänderungen in Knochen

Lastverteilung innerhalb des Knochengewebes um. Dies erfolgt jeweils unabhängig für typische Bewegungssequenzen einer Person. Als Ergebnis speichert Pandas pro Punkt des FE-Gitters und pro Zeitschritt bis zu 20 Variablen in einer Datenbank. Abhängig von der Anzahl der Gitterpunkte und der Anzahl der Zeitschritte ergibt dies ein Datenvolumen zwischen 100 MB und mehreren GB für jede Bewegungssequenz.

Die von Pandas berechneten Lösungen berücksichtigen allerdings keine zellularen Reaktionen im Knochengewebe. Genau hier kommt die systembiologische Simulation ins Spiel, die die Bildung oder den Abbau des Gewebes auf Basis der Interaktion von Zellen bestimmt. Dies kann mithilfe der Rechenumgebung GNU Octave[2] umgesetzt werden. Vorbereitende ETL-Prozesse (Extraktion, Transformation, Laden) filtern passende Daten aus der Datenbank von Pandas und aggregieren sie anschließend über alle Zeitschritte. Die einzelnen charakteristischen Lösungen von Pandas werden zudem zu einer idealisierten Lösung zusammengefasst, die dem approximierten Tagesablauf der Person entspricht. Diese Lösung wird schließlich in CSV-basierten Dateien (*comma-separated values*) gespeichert, sodass Octave sie einlesen kann. Die systembiologische Simulation rechnet die genaue Knochendichte nach Ablauf eines Tages aus. Diese wird genutzt, um die Konfiguration des biomechanischen Modells anzupassen. Der gesamte Prozess wiederholt sich, bis die vom Anwender vorgegebene Anzahl an Tagen berechnet wurde.

[2] http://www.gnu.org/software/octave/

2.1 Anforderungen an das Datenmanagement

Für das Datenmanagement in Simulationen ergeben sich eine Reihe spezifischer Anforderungen, welche für die an den Simulationsergebnissen interessierten Nutzer wichtig sind. Um diese Anforderungen zu identifizieren, haben wir mehrere Szenarien aus der Literatur (siehe z. B. [SR09, TDG07]), verschiedene Simulationssoftware wie Pandas oder Octave sowie reale Simulationen analysiert. Neben der in Abschnitt 2 vorgestellten Simulation gehören hierzu die von Fehr et al. und Rommel et al. beschriebenen Beispiele [FE11, RK11]. Nachfolgend diskutieren wir die aus Nutzersicht wichtigsten Anforderungen.

Datensicherheit: Im Unternehmenskontext beschreiben Daten häufig schutzwürdiges geistiges Eigentum. Weiterhin müssen zahlreiche gesetzliche Bestimmungen eingehalten werden, welche die Offenlegung von Informationen regulieren. Wissenschaftliche Institutionen möchten z. B. Forschungsergebnisse nicht vor einer Publikation der Arbeit offenlegen oder im Fall von Studien die persönlichen Daten der Teilnehmer schützen. Dies alles macht adäquate *Vorkehrungen zur Datensicherheit* notwendig.

Datenqualität: Mathematische Simulationsmodelle sind lediglich Abbildungen der realen Probleme, die untersucht werden sollen. Die numerische Implementierung dieser Modelle bringt häufig weitere Approximationen und damit Ungenauigkeiten mit ein. Dies impliziert hohe Anforderungen an die *Qualität von Simulationsergebnissen* [Re14].

Effizienz und Optimierung: Die Gesamtgröße der in einzelnen Simulationsläufen involvierten Daten kann zwischen wenigen 100 KB und einigen TB liegen. Die Größe und Komplexität der Daten sind entscheidende Faktoren für die Ausführungszeit von Simulationen. Dies führt zwangsläufig zu Anforderungen bzgl. der *Effizienz* des Datenmanagements und bzgl. der Unterstützung entsprechender *Optimierungsmöglichkeiten* [Vr07].

Reproduzierbarkeit und Nachvollziehbarkeit: Ein weiterer wichtiger Aspekt ist die Sicherstellung der *Reproduzierbarkeit und Nachvollziehbarkeit einer Simulation und ihrer Ergebnisse* [Fr08]. In Bezug auf das Datenmanagement muss hierbei z. B. dafür Sorge getragen werden, dass sämtliche Ergebnisdaten einer Simulation nie verloren gehen und sie dem Nutzer nach den Berechnungen wieder zur Verfügung gestellt werden. Gleiches gilt auch für Log- oder Provenance-Informationen [CVu12, Fr08], die während der Simulation gesammelt werden. Provenance-Informationen beschreiben den detaillierten Simulationsablauf, z. B. die Herkunft der Eingabedaten und wie diese weiterverarbeitet werden.

3 Softwarebereitstellung in der Cloud

Die manuelle Installation und Konfiguration einer Simulationssoftware in einer Cloud ist sehr zeitaufwändig. Um diese Softwarebereitstellung zu automatisieren, kann der OASIS-Standard TOSCA eingesetzt werden [OA13]. TOSCA erlaubt es, beliebige Anwendungen mit geringem Aufwand interoperabel in verschiedenen Cloud-Umgebungen bereitzustellen. Den Kern von TOSCA bildet eine Modellierungssprache, welche die Definition einer Anwendung und ihres Aufbaus in einer sog. Diensttopologie ermöglicht. Diese Topologie besteht aus Anwendungskomponenten und deren Beziehungen. Des weiteren können

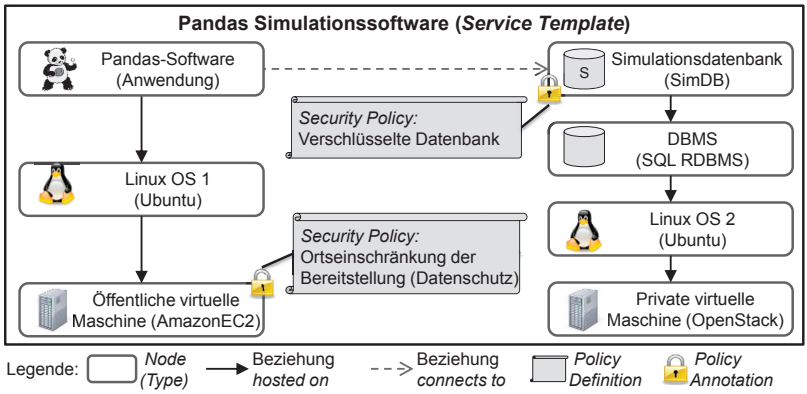


Abbildung 2: Beispiel einer TOSCA Diensttopologie für die Simulationssoftware Pandas

Deployment- und Management-Pläne als Workflows erstellt werden, um die Installation und die Verwaltung einer Anwendung zu automatisieren. Nachfolgend illustrieren wir beispielhaft, wie TOSCA zur Bereitstellung der in Abschnitt 2 betrachteten Simulationssoftware Pandas verwendet werden kann.

3.1 Bereitstellung der Simulationssoftware Pandas

Abb. 2 stellt die mit TOSCA definierte Topologie der Pandas-Software dar. Die Topologie besteht aus zwei Server-Stacks: einer für die Simulationssoftware (links) und einer für die Datenhaltung (rechts). Der Simulationssoftware-Stack wird auf einer virtuellen Maschine in einer öffentlichen Cloud-Umgebung betrieben. Dies bietet einen Kostenvorteil, da die Nutzer keine eigene Infrastruktur einrichten und verwalten müssen, sondern diese nur bedarfsgerecht anmieten, wenn sie tatsächlich eine Simulation durchführen wollen. Weiterhin kann die Infrastruktur jederzeit skaliert werden, je nachdem wie rechenintensiv die Simulation ist. Auf dieser öffentlichen virtuellen Maschine ist ein Linux-Betriebssystem installiert (dargestellt durch die Beziehung *hostet-on*), auf welchem wiederum die Pandas-Software läuft. Die Pandas-Software ist mit dem Datenbank-Stack über eine Beziehung *connects-to* verbunden, d. h. die beteiligten Komponenten kommunizieren über ein Netzwerk. Im Gegensatz zum Software-Stack nutzt der Datenbank-Stack eine private virtuelle Maschine, welche z. B. auf einem OpenStack-System des Nutzers betrieben wird. Dadurch kann die Datensicherheit auf hohem Niveau gehalten werden, da die Datenspeicherung nicht öffentlich erfolgt. Auf dieser virtuellen Maschine ist in einem Linux-Betriebssystem ein SQL-Datenbanksystem installiert, das die Simulationsdatenbank verwaltet.

Prinzipiell bietet TOSCA noch vielfältigere Modellierungsmöglichkeiten. So können beispielsweise an jeder Anwendungskomponente und an der gesamten Topologie über Policies frei definierbare Anforderungen annotiert werden, welche bei der Installation, Konfiguration und während des Betriebs der Anwendung eingehalten werden müssen. Im betrachteten

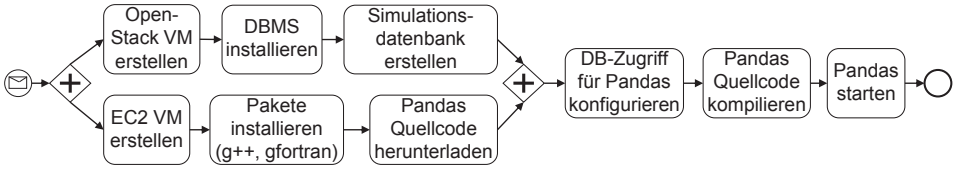


Abbildung 3: TOSCA Deployment-Plan für Pandas

Anwendungsfall wird als Sicherheitsanforderung z. B. der Ort eingeschränkt, an dem der Server für den linken Software-Stack läuft. Weiterhin definiert die Topologie, dass die Simulationsdatenbank ihre Daten verschlüsseln muss.

Um eine mit TOSCA definierte Anwendung in der Cloud aufzusetzen, können verschiedene Laufzeitumgebungen für TOSCA eingesetzt werden. Ein Beispiel ist OpenTOSCA, das als Eingabe ein *Cloud Service Archive* (CSAR) benötigt, welches zusätzlich zur Diensttopologie alle für die Anwendung nötigen Artefakte und Softwarepakete enthält [Bi13].

3.2 Plan-basierte Bereitstellung und Verwaltung

Ein weiterer zentraler Teil von TOSCA sind Pläne: *Deployment-Pläne* nehmen die Installation von Anwendungen vor, während *Management-Pläne* Anwendungen zu deren Laufzeit verwalten und überwachen. Abb. 3 illustriert einen Deployment-Plan für die Simulationssoftware Pandas. Dieser besteht aus zwei parallel ablaufenden Pfaden, welche die beiden in Abb. 2 gezeigten Stacks aufsetzen. Im oberen Pfad erstellt der Plan zunächst eine virtuelle Maschine auf dem privaten OpenStack-System, installiert darauf ein Datenbanksystem und erzeugt die Simulationsdatenbank. Unten erstellt er analog eine virtuelle Maschine auf Amazon EC2, installiert die für Pandas notwendigen Linux-Pakete und lädt den Pandas-Quellcode von einem Repository herunter. Danach wird die parallele Verarbeitung synchronisiert, damit die Verknüpfung der beiden Stacks über die Konfiguration des DB-Zugriffs erfolgen kann. Weiterhin muss der Pandas-Quellcode kompiliert und Pandas anschließend gestartet werden, sodass es bereit ist, die nachfolgenden Simulationsberechnungen durchzuführen. Dies kann durch Management-Pläne erfolgen, was im nächsten Abschnitt anhand der Datenbereitstellung und Datenrückgabe näher erläutert wird.

4 Datenbereitstellung in der Cloud

Eine wesentliche Herausforderung an die Datenbereitstellung und Datenrückgabe in der Cloud ergibt sich aus dem Wunsch, Simulationen sowie die heterogenen Datenlandschaften aus verschiedenen Anwendungsgebieten zu koppeln (siehe Abschnitt 2). Um eine nahtlose Kopplung zu ermöglichen, muss eine entsprechende Lösung hinreichend *generisch* sein. Dies bedeutet, sie muss insbesondere alle von den verschiedenen Anwendungsgebieten

und ihrer Nutzer benötigten Datenformate und Datenmanagementoperationen unterstützen. Ein weiterer wichtiger Aspekt ist, dass die an den Simulationsergebnissen interessierten Wissenschaftler ihre Simulationen häufig selbst implementieren und dabei auch einen Großteil des Datenmanagements umsetzen müssen. Um Wissenschaftler dabei nicht mit komplexen Implementierungsdetails dieses Datenmanagements zu überfordern, sollte eine Lösung für die Datenbereitstellung eine geeignete *Abstraktionsunterstützung* bieten [RS13].

Da TOSCA eine hohe Modellierungsmächtigkeit für die Softwarebereitstellung bietet, ist zur Erreichung einer generischen Lösung eine Integration der Datenbereitstellung in TOSCA-Plänen ratsam. Bisher kann die Datenbereitstellung in TOSCA-Plänen nur durch proprietäre Services umgesetzt werden, welche wiederum keine generische Lösung darstellen [OA13]. In diesem Abschnitt diskutieren wir zunächst mögliche Alternativen für die Umsetzung der Datenbereitstellung. Als Ergebnis dieser Diskussion stellt sich schließlich heraus, dass das SIMPL-Rahmenwerk (*SimTech, Information Managemenent, Processes, and Languages*) den oben beschriebenen Herausforderungen an eine generische Lösung und an eine geeignete Abstraktionsunterstützung am besten begegnet [RS13, Re11]. Folglich erläutern wir abschließend, wie SIMPL mit TOSCA integriert werden kann.

4.1 Diskussion von Alternativen für die Datenbereitstellung

Standard-Tools für die Definitionen und Ausführung von ETL-Prozessen, wie z. B. die von IBM[3] , Pentaho[4] und Talend[5] angebotenen Lösungen, bieten zwar vielfältige und daher in Simulationen generisch einsetzbare Möglichkeiten zur Umsetzung von Datenmanagementoperationen. Vertrauen Wissenschaftler aber ausschließlich diesen ETL-Tools, müssen sie umfangreiche Implementierungsdetails spezifizieren. Zum Beispiel betrifft dies die in Abschnitt 2 angesprochenen ETL-Prozesse zum Filtern, Aggregieren und Transformieren der Ergebnisdaten von Pandas. Hierbei müssen Wissenschaftler komplexe SQL-Anfragen definieren oder sogar Skript- bzw. Programmiersprachen zur Implementierung des Datenmanagements einsetzen [RS13]. Wissenschaftler besitzen zwar eine hohe Expertise in ihrem Anwendungsbereich der Simulation, weisen aber i. d. R. eingeschränkte Fähigkeiten bzgl. dieser Anfrage- oder Programmiersprachen auf. Daher bieten solche ETL-Tools keine für Wissenschaftler adäquate Abstraktionsunterstützung.

Eine Alternative zur Umsetzung der Datenbereitstellung in TOSCA-Plänen bieten die Workflow-Produkte von IBM, Microsoft und Oracle mit sog. SQL-Aktivitäten [Vr08]. Eine solche Aktivität beinhaltet eine SQL-Anweisung und sendet diese Anweisung bei der Ausführung der Aktivität an ein externes Datenbanksystem. Die Einschränkung auf SQL sorgt aber dafür, dass diese Lösung in vielen Simulationen nicht einsetzbar ist, insbesondere wenn wie bei der in Abschnitt 2 beschriebenen Simulation auch Zugriffe auf Dateien nötig werden. Das wissenschaftliche Workflow-System Kepler bietet ähnliche Aktivitäten für den Zugriff auf SQL-Datenbanksysteme, Dateisysteme und Sensornetze [Lu06]. Allerdings gibt es für jeden Typ von Datenressource und sogar für verschiedene Typen von Datenmana-

[3] http://www.ibm.com/software/data/integration/info_server/
[4] http://www.pentaho.de/
[5] http://de.talend.com/

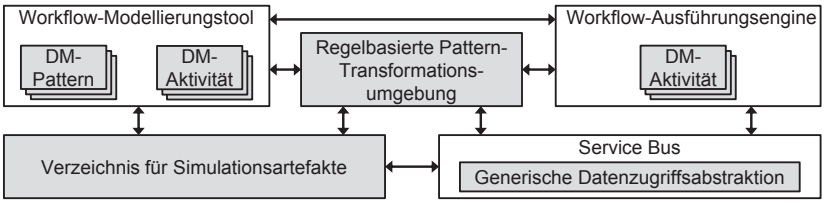


Abbildung 4: Zentrale Komponenten des SIMPL-Rahmenwerks, vgl. [RS13, Re11]

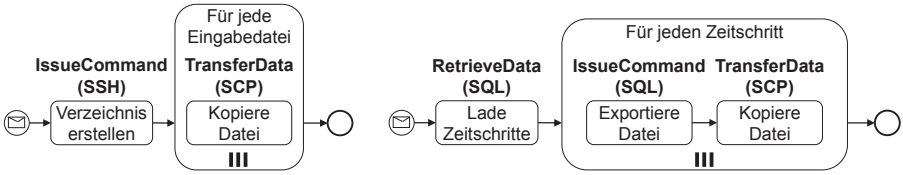
gementoperationen ganz unterschiedliche proprietäre Aktivitätstypen. Die Entwickler von Kepler haben selbst eingeräumt, dass dies keine generische Lösung darstellt und dass die große Anzahl an Aktivitätstypen die Workflowmodellierer sogar überfordern kann [Ba10].

Im Gegensatz zu den bisher diskutierten Ansätzen bietet das SIMPL-Rahmenwerk eine generische sowie einfach zu benutzende Lösung für die Datenbereitstellung [RS13, Re11]. Abb. 4 zeigt, wie SIMPL die Architektur eines Workflowsystems erweitert. Der Service Bus stellt Workflowmodellierern über eine *generische Datenzugriffsabstraktion* einen einheitlichen Zugriff auf beliebige externe Datenressourcen zur Verfügung [Re11]. Das Workflow-Modellierungstool und die Workflow-Ausführungsengine bieten zudem eine Unterstützung für ebenso generische Datenmanagementaktivitäten (*DM-Aktivitäten*). Die Aktivitäten können beliebigen externen Datenressourcen zugeordnet werden und beliebige Befehle in deren Befehlssprachen beinhalten – also nicht nur SQL-Anweisungen, sondern z. B. auch Shell-Kommandos für den Zugriff auf Dateien. Als weiterführende Abstraktionsunterstützung und als Alleinstellungsmerkmal gegenüber anderen Ansätzen ermöglicht eine Erweiterung des Workflow-Modellierungstools die Nutzung von Datenmanagementpatterns (*DM-Patterns*) als Workflow-Bausteine [RS13]. Jedes Pattern fasst mehrere feingranulare Workflow-Schritte zusammen, was bereits die Anzahl der für Workflowmodellierer sichtbaren Schritte reduziert. Weiterhin müssen diese nur wenige abstrakte Parameterwerte für die Patterns festlegen, anstatt komplexe Implementierungsdetails zu spezifizieren. Die *regelbasierte Patterntransformationsumgebung* bildet solche parametrisierten Patterns mithilfe von Metadaten aus dem *Verzeichnis für Simulationsartefakte* auf ausführbare Workflowfragmente oder Services ab [RS13].

4.2 Integration von SIMPL mit TOSCA

Um die Datenbereitstellung und Datenrückgabe für Simulationen in der Cloud umzusetzen, können die SIMPL DM-Aktivitäten und die Patterns in TOSCA-Pläne integriert werden. Dazu muss lediglich die Workflowausführungsengine in OpenTOSCA [Bi13] um die DM-Aktivitäten erweitert werden. Alle anderen Komponenten des SIMPL-Rahmenwerks können über ihre Service-Schnittstellen nahtlos an OpenTOSCA angebunden werden.

Abb. 5 zeigt die mit SIMPL DM-Aktivitäten angereicherten Management-Pläne für die Datenbereitstellung bzw. Datenrückgabe der Simulationssoftware Pandas. Der Plan für



(a) Datenbereitstellung (b) Datenrückgabe

Abbildung 5: TOSCA-Pläne mit SIMPL-Aktivitäten zur Datenbereitstellung und -rückgabe für Pandas

die Datenbereitstellung wird nach dem in Abb. 3 gezeigten Deployment-Plan ausgeführt. Er erzeugt zunächst über eine IssueCommand-Aktivität ein Verzeichnis auf der virtuellen Maschine von Pandas, in das später die Eingabedateien kopiert werden können. Dazu setzt die Aktivität über eine SSH-Schnittstelle ein entsprechendes Shell-Kommando an die virtuelle Maschine von Pandas ab. Anschließend kopiert der Plan über eine parallele Schleife und über eine darin eingebettete TransferData-Aktivität jede relevante Eingabedatei mithilfe eines Secure Copy (SCP) Befehls in das zuvor erzeugte Verzeichnis. Der Plan für die Datenrückgabe wird nach den Simulationsberechnungen ausgeführt. Eine gängige Praxis ist, für jeden Zeitschritt der biomechanischen Simulation eine Datei aus der Pandas-Datenbank zu exportieren und diese Datei auf einen anderen Rechner zu kopieren. Dort kann z. B. eine entsprechende Zeitreihe visualisiert oder analysiert werden. Der Plan lädt zunächst über eine RetrieveData-Aktivität mit eingebetteter SQL-Anweisung eine Liste aller Zeitschritte aus der Datenbank. Die nachfolgende Schleife iteriert über diese Liste und exportiert für jeden Zeitschritt die passende Datei mithilfe einer IssueCommand-Aktivität und einer darin eingebetteten SQL Export-Anweisung. Schließlich kopiert eine TransferData-Aktivität diese Datei auf den Zielrechner, womit die Datenrückgabe abgeschlossen ist.

5 Einhaltung nicht-funktionaler Anforderungen

Die formale Beschreibung von Anforderungen wird in TOSCA über Policies abgedeckt (siehe Abschnitt 3). Verwenden Cloud-Nutzer TOSCA, können sie mit dem Policy4TOSCA-Rahmenwerk beliebige Anforderungen definieren, insbesondere auch die in Abschnitt 2.1 diskutierten [Wa13]. Die Policies werden dabei entweder global an einer gesamten Dienst-topologie oder lokal an einzelne ihrer Komponenten angehängt. Anschließend müssen die über Policies definierten Anforderungen mit geeigneten technischen Mitteln umgesetzt werden, was für jede der in Abschnitt 2.1 identifizierten Anforderungen sehr unterschiedlich erfolgen kann. Daher betrachten wir im Folgenden vor allem diese technische Umsetzung der Policies. Für die einzelnen in Abschnitt 2.1 identifizierten Anforderungen diskutieren wir hierzu jeweils den aktuellen Forschungsstand rund um Simulationen und/oder Cloud und skizzieren einige offene Forschungsthemen.

Datensicherheit: Die Einhaltung der von den Policies vorgegebenen Anforderungen an Datensicherheit wird durch die genutzte TOSCA-Laufzeitumgebung umgesetzt. Dies kann

von verschiedenen Umgebungen auf unterschiedliche Art erfolgen. Für Policies, die lokal an einzelne Dienstkomponenten angehängt sind, wählt OpenTOSCA entsprechende Implementierungen und Konfigurationen der Komponenten aus, für welche die Policies gelten [Bi13]. Für global an eine Diensttopologie annotierte Policies ändert OpenTOSCA die Deployment-Pläne, sodass die Policies zur Deployment-Zeit beachtet werden.

Datenqualität: Reiter et al. beschreiben ein Rahmenwerk zur Definition und Evaluierung von Datenqualitätsanforderungen in Simulationsworkflows [Re14]. Workflowmodellierer können ihre Anforderungen an die Datenqualität über Policies direkt in Workflows definieren, z. B. an Workflowaktivitäten oder bei Kontrollflussentscheidungen. Die Evaluation der Datenqualität kann entweder im Workflowsystem erfolgen oder auf Ebene der Dienste, welche die Daten liefern bzw. verarbeiten. Um die Datenqualität zu verbessern, können z. B. diese Dienste ausgetauscht oder die Parametrisierung der Dienstaufrufe geändert werden [Re14]. Die von Reiter et al. betrachteten Workflows entsprechen in TOSCA den Plänen (siehe Abschnitt 3.2). Im Policy4TOSCA-Rahmenwerk werden Policies nicht wie bei Reiter et al. in Workflows bzw. Plänen definiert, sondern in den Diensttopologien. Hier muss in Zukunft untersucht werden, welche der beiden Varianten für die Datenqualität von Simulationen in der Cloud am besten geeignet ist. Die Evaluation der Datenqualität sowie die Maßnahmen zu deren Verbesserung erfolgen hingegen analog zu OpenTOSCA – auf Ebene der Workflows bzw. Pläne oder auf Ebene der Dienstkomponenten.

Effizienz und Optimierung: Um die *Effizienz* von Cloud-Anwendungen zu steigern, können insbesondere Optimierungen der TOSCA Diensttopologien und der Pläne erfolgen. Sollten z. B. die Effizienzanforderungen an die Datenhaltung der Software Pandas sehr hoch sein, so kann die in Abb. 2 gezeigte Diensttopologie geändert werden. Der Datenbank-Stack könnte in einer öffentlichen Cloud und als effizienter, skalierbarer und kostengünstiger Datenbankdienst umgesetzt werden. Allerdings muss dabei mit negativen Auswirkungen auf die Datensicherheit gerechnet werden. Hierbei müssen ganzheitliche Methoden entwickelt werden, welche flexibel auf solche sich konkurrierenden Anforderungen reagieren können. Vrhovnik et al. schlagen zudem einen Ansatz zur *Optimierung* von Workflows mit eingebetteten SQL-Anweisungen vor [Vr07]. Solch ein Ansatz kann prinzipiell auch auf die in Abschnitt 4.2 beschriebenen, mit SIMPL-Aktivitäten angereicherten TOSCA-Pläne für die Datenbereitstellung und Datenrückgabe angewandt werden. Als ein zukünftiges Forschungsthema muss die Einschränkung auf SQL allerdings überwunden werden, insbesondere da bei vielen Simulationen auch Zugriffe auf Dateien nötig sind.

Reproduzierbarkeit und Nachvollziehbarkeit: Da das Datenmanagement in der Cloud im Fokus dieses Beitrags steht, gehen wir für die *Reproduzierbarkeit* davon aus, dass Simulationsberechnungen selbst bei gleicher Eingabe immer das gleiche Ergebnis liefern. Darüber hinaus muss dafür gesorgt werden, dass die Softwarebereitstellung, die Datenbereitstellung und die Berechnungsabläufe erneut durchführbar sind. In der Laufzeitumgebung OpenTOSCA enthält ein CSAR-Archiv eine vollständige Beschreibung der Software sowie alle Pläne, die zum Deployment der Software, zur Datenbereitstellung und für die Berechnungsabläufe benötigt werden [Bi13]. Die auf diese Weise beschriebenen Prozessschritte einer Simulation können daher jederzeit wiederholt werden. Die Eingabedaten selbst werden allerdings nicht in einem CSAR-Archiv gehalten. Sie müssen für eine erneute Durchführung einer Simulation vom Nutzer vorgehalten oder bei der vorherigen Durchführung in der Cloud

abgelegt werden. Ein offenes Thema in diesem Kontext ist die Langzeitarchivierung der großen Eingabe- und Ergebnisdaten einer Simulation. Die *Nachvollziehbarkeit* einer Simulation kann über Log- oder Provenance-Informationen sichergestellt werden [CVu12, Fr08]. Cuevas-Vicenttín et al. zeigen hierbei die Notwendigkeit auf, unterschiedliche Provenance-Informationen aus verschiedenen Quellen in der Cloud zu integrieren, um sie so Wissenschaftlern besser zugänglich zu machen [CVu12]. Als Basis können bereits verfügbare Systeme verwendet und an die Bedürfnisse von Simulationen angepasst werden. Beispiele sind Systeme für die Kontrollpflicht der Auftragsdatenverarbeitung in der Cloud [Se13].

6 Zusammenfassung und Ausblick

Speziell für Organisationen, die Simulationen eher sporadisch durchführen, geht der Trend für die Ausführung dieser Simulationen zu Cloud-Umgebungen. Die Hauptgründe sind die bedarfsgerechten, dynamischen und automatisierbaren Provisionierungsmöglichkeiten für benötigte Ressourcen. Die Bereitstellung von Software in der Cloud wird bereits von verschiedenen Systemen und Standards, wie z.B. TOSCA unterstützt. Die Fragestellungen bzgl. des Datenmanagements von Simulationen in der Cloud, blieben bisher aber weitgehend unbeachtet. In diesem Beitrag untersuchten wir zum Einen verschiedene Möglichkeiten für die Datenbereitstellung und Datenrückgabe in der Cloud. Wir schlugen hierfür das generische und einfach zu benutzende SIMPL-Rahmenwerk vor [RS13, Re11] und zeigten, wie es mit TOSCA integriert werden kann. Zum Anderen untersuchten wir verschiedene Technologien für die Definition und Einhaltung simulationsspezifischer Nutzeranforderungen. Das Policy4TOSCA-Rahmenwerk kann hierbei als gute Basis dienen, um nicht-funktionale Anforderungen zu definieren [Wa13]. Bezüglich der Einhaltung dieser Anforderungen diskutierten wir zudem verschiedene offene Forschungsthemen. In diesem Rahmen wichtige Themen sind insbesondere die in Abschnitt 5 diskutierten Optimierungsmöglichkeiten sowie die dort skizzierte Verwaltung und Integration von Provenance-Informationen.

Danksagung: Die Autoren danken der Deutschen Forschungsgemeinschaft (DFG) sowie dem Bundesministerium für Wirtschaft (BMWi) für die Förderung des Projekts im Rahmen des Exzellenzclusters Simulation Technology bzw. im Rahmen von CloudCycle (01MD11023). Außerdem danken wir Dimka Karastoyanova und Frank Leymann sowie den Mitgliedern ihrer Forschungsgruppe für ihre hilfreiche Kooperation.

Literatur

[Ba10] D. Barseghian et al. Workflows and Extensions to the Kepler Scientific Workflow System to Support Environmental Sensor Data Access and Analysis. *Ecological Informatics*, 5(1), 2010.

[Bi13] T. Binz et al. OpenTOSCA – A Runtime for TOSCA-based Cloud Applications. In *Tagungsband der 11. International Conference on Service Oriented Computing (ICSOC)*, Berlin, Deutschland, 2013.

[CVu12] V. Cuevas-Vicenttín et al. Scientific Workflows and Provenance: Introduction and Research Opportunities. *Datenbank-Spektrum*, 12(3), 2012.

[FE11] J. Fehr und P. Eberhard. Simulation Process of Flexible Multibody Systems with Non-modal Model Order Reduction Techniques. *Multibody System Dynamics*, 25(3), 2011.

[Fr08] J. Freire et al. Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3), 2008.

[Kr13] R. Krause et al. Scientific Workflows for Bone Remodelling Simulations. *Applied Mathematics and Mechanics*, 13(1), 2013.

[Lu06] B. Ludäscher et al. Scientific Workflow Management and the Kepler System. *Concurrency and Computation: Practice and Experience*, 18(10), 2006.

[MG11] P. Mell und T. Grance. The NIST Definition of Cloud Computing, 2011.

[OA13] OASIS. *Topology and Orchestration Specification for Cloud Applications Version 1.0*, Mai 2013.

[Re11] P. Reimann et al. SIMPL - A Framework for Accessing External Data in Simulation Workflows. In GI, Hrsg., *Datenbanksysteme für Business, Technologie und Web (BTW)*, Kaiserslautern, Deutschland, 2011.

[Re14] M. Reiter et al. Quality of Data Driven Simulation Workflows. *Journal of Systems Integration*, 5(1), 2014.

[RK11] J. B. Rommel und J. Kästner. The Fragmentation-Recombination Mechanism of the Enzyme Glutamate Mutase Studied by QM/MM Simulations. *Journal of the American Chemical Society*, 26(133), 2011.

[RS13] P. Reimann und H. Schwarz. Datenmanagementpatterns in Simulationsworkflows. In GI, Hrsg., *Datenbanksysteme für Business, Technologie und Web (BTW)*, Magdeburg, Deutschland, 2013.

[Se13] A. Selzer. Die Kontrollpflicht nach § 11 Abs. 2 Satz 4 BDSG im Zeitalter des Cloud Computing. *Datenschutz und Datensicherheit - DuD*, 37(4), 2013.

[SR09] A. Shoshani und D. Rotem. *Scientific Data Management: Challenges, Technology, and Deployment*. Computational Science Series. Chapman & Hall, 2009.

[TDG07] I. Taylor, E. Deelman und D. Gannon. *Workflows for e-Science - Scientific Workflows for Grids*. Springer, London, UK, 2007.

[VHKL13] K. Vukojevic-Haupt, D. Karastoyanova und F. Leymann. On-demand Provisioning of Infrastructure, Middleware and Services for Simulation Workflows. In *Tagungsband der 6. IEEE International Conference on Service Oriented Computing and Applications*, Kauai, HI, USA, 2013.

[Vr07] M. Vrhovnik et al. An Approach to Optimize Data Processing in Business Processes. In *Tagungsband der 33. International Conference on Very Large Data Bases (VLDB)*, Wien, Österreich, 2007.

[Vr08] M. Vrhovnik et al. An Overview of SQL Support in Workflow Products. In *Tagungsband der 24. International Conference on Data Engineering*, Cancùn, México, 2008.

[Wa13] T. Waizenegger et al. Policy4TOSCA: A Policy-Aware Cloud Service Provisioning Approach to Enable Secure Cloud Computing. In *Tagungsband der 3. International Conference on Secure Virtual Infrastructures (DOA-Trusted Cloud'13)*, Graz, Österreich, 2013.

Datenmodellierung in der Anwendungsentwicklung mit NoSQL-Datenbanken

Stefanie Scherzinger
OTH Regensburg
stefanie.scherzinger@oth-regensburg.de

Abstract: NoSQL-Datenbanken sind gerade in der Webentwicklung zunehmend beliebt. Oft sind es die großen Datenmengen, die es zu verwalten gilt, mitunter sind diese Systeme aber auch wegen ihrer Schema-Flexibilität für agile Entwicklungsteams interessant. Indem viele NoSQL-Datenbanken keine Unterstützung für die Definition, Einhaltung und Wartung eines globalen Schemas bieten, verlagern sich klassische Aufgaben des Datenbankmanagementsystems in die Anwendungssoftware. Dieser Vortrag gibt einen Überblick über konkrete Herausforderungen, die sich in der Praxis beim Entwurf eines Datenmodells für *Key-Value*- und Dokumenten-Datenbanken ergeben. Dazu zählen eine Modellierung, die atomare Updates ermöglicht, das Vermeiden von *Hot-Spot*-Datenobjekten, wie sie durch hochfrequente, parallele Schreibzugriffe gegen dasselbe Objekt verursacht werden, sowie Strategien zum Umgang mit kontinuierlicher Schema-Evolution. Der Vortrag zeigt auf, dass gerade die Datenbank-Community mit ihrem Erfahrungsschatz im Schema-Management und ihrem breiten Fundus an formalen Methoden hier einen wertvollen Beitrag leisten kann.

Während der letzten Jahrzehnte wurde die Architektur von Desktop- und Webanwendungen nahezu ausnahmslos von relationaler Datenbanktechnologie dominiert. Es gibt jedoch Sparten, in denen NoSQL-Datenbanksysteme eine attraktive Alternative darstellen. Unter dem Begriff "NoSQL" tummelt sich mittlerweile ein bunter Zoo von Systemen. So sehr sich die einzelnen Produkte voneinander unterscheiden, so zeichnen sie sich typischerweise in den "3 *V*s" aus: *Volume*, die Fähigkeit, auf große Datenmengen zu skalieren, *Velocity*, die Fähigkeit, laufend neu hinzukommende Datensätze zu verarbeiten, und *Variety*, die Fähigkeit, mit stark heterogenen Datenstrukturen zu arbeiten.

Unabhängig davon, welches "*V*" letztlich bei der Entscheidung für eine NoSQL-Datenbank maßgeblich ist, stellen sich in der Anwendungsentwicklung gegen diese Systeme Herausforderungen, auf die das grundständige Informatikstudium oft nur ungenügend vorbereitet [ST14b]. Dieser Vortrag gibt einen Überblick über die gängigen *Pain Points* in der Datenmodellierung bei der Anwendungsentwicklung gegen NoSQL-Datenbanksysteme, wie etwa *Key-Value*- und Dokumenten-Datenbanken:

- Transaktionen mit ACID-Verhalten sind in Systemen, die darauf ausgelegt sind, auf zehntausenden physischen Knoten verteilt zu laufen, nicht uneingeschränkt möglich. Naiv entwickelte Anwendungen zeigen daher oft *Eventual Consistency*-Effekte, welche zu erheblichen Irritationen bei den Anwendern führen können. So sind eben vollzogene Änderungen an Daten, z.B. nach Ausfüllen eines Formulars, womöglich nicht unmittelbar sichtbar.

Um Änderungen an mehreren Objekten atomar durchführen zu können, sind Entwickler mitunter gezwungen, an sich logisch eigenständige Objekte gemeinsam auf ein persistentes Objekt abzubilden [SF12]. Dies läuft entgegen der in der Datenbankvorlesung vermittelten Strategie, die Daten möglichst zu normalisieren. Manche Systeme erlauben es auch, logisch eigenständige Objekte 1:1 auf persistente Objekte abzubilden und diese dann zu gruppieren. Die Objekte in solchen *Entity Gruppen* werden physisch im verteilten System kollokiert, was atomare Änderungen und strikt konsistenten Zugriff innerhalb der Gruppe ermöglicht (vgl. [San12, BBC$^+$11]). Somit ist der Gestaltungsspielraum bei der Datenmodellierung systemspezifisch.

- Das häufige Ändern einzelner sog. *Hot-Spot*-Objekte stellt für hochverteilte Systeme typischerweise ein Problem dar. Naiv entwickelte Anwendungen leiden daher ausgerechnet während Lastspitzen unter Laufzeitfehlern und fehlgeschlagenen Schreibzugriffen. Hier spielt die Datenmodellierung eine tragende Rolle, wenn solche Effekte vermieden werden sollen [SDAIDF13, ST14a].
- Indem viele NoSQL-Systeme kein festes Schema fordern, erlauben sie eine flexiblere Handhabung heterogener Daten bei Schema-Evolution. Langfristig erschwert eine hohe Heterogenität in den persistenten Objekten jedoch die Wartbarkeit der Anwendung. In der Praxis behilft man sich mit Migrations-Skripten (d.h. "Custom Code"), um die Datenbasis *eager* zu migrieren. Alternativ werden Object-Mapper eingesetzt, die einzelne persistente Objekte beim Laden in die Anwendung transformieren. Dadurch wird die Datenbasis inkrementell bzw. *lazy* migriert. Ob *eager* oder *lazy*, diese Ansätze sind bisher bloße Notbehelfe, fehlt es doch an einem theoretisch fundierten Rahmenwerk, das ein systematisches Testen und eine saubere Migrationssemantik überhaupt erst ermöglicht [KSS14, SKS13].

Diese *NoSQL Pain Points* zeigen insbesondere die enge Verflechtung zwischen logischem und physischem Datenbankentwurf auf, die in den meisten NoSQL-Datenbanken gegeben ist. So führt ein ungünstig gewähltes Datenmodell dazu, dass ACID-Eigenschaften bei Änderungen nicht gewährleistet werden können, oder Anwendungen unter hoher Schreiblast nicht skalieren. Es gilt zudem, die Entwickler von der bestehenden Notwendigkeit zu befreien, sich mit den Eigenheiten der jeweiligen NoSQL-Datenbank eingehend beschäftigen zu müssen. Konkret ist ein neues Ökosystem an Werkzeugen zu entwickeln, die um die proprietären Funktionalitäten dieser Systeme wissen und die Entwickler aktiv darin unterstützen, ein geeignetes Datenmodell zu entwerfen.

Dieser Vortrag zeigt auf, welchen wertvollen Beitrag die Datenbank-Community bei der Entwicklung solcher Werkzeuge für die Software-Technik leisten kann.

Biographie: Stefanie Scherzinger ist seit 2012 Professorin an der OTH Regensburg. Sie promovierte 2008 an der *Universität des Saarlandes*, mit Zwischenstationen an der *TU Wien* und der *Cornell University*. Danach erwarb sie Industrieerfahrung als Softwareentwicklerin bei IBM und Google. Aus dieser Zeit stammen mehrere erfolgreiche Patentanmeldungen, sowie einschlägige Erfahrungen in der professionellen Anwendungsentwicklung mit NoSQL-Datenbanken. Ihre aktuelle Forschung widmet sich Problemen im Datenmanagement, die aus ihrer Berufspraxis motiviert sind.

Literatur

[BBC+11] Jason Baker, Chris Bond, James C. Corbett, JJ Furman et al. Megastore: Providing Scalable, Highly Available Storage for Interactive Services. In *Proc. CIDR*, 2011.

[KSS14] Meike Klettke, Stefanie Scherzinger und Uta Störl. Datenbanken ohne Schema? *Datenbank-Spektrum*, 14(2), 2014.

[San12] Dan Sanderson. *Programming Google App Engine*. O'Reilly Media, Inc., 2. Auflage, 2012.

[SDAIDF13] Stefanie Scherzinger, Eduardo Cunha De Almeida, Felipe Ickert und Marcos Didonet Del Fabro. On the Necessity of Model Checking NoSQL Database Schemas when Building SaaS Applications. In *Proc. TTC 2013*, 2013.

[SF12] Pramod J. Sadalage und Martin Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley Professional, 1. Auflage, 2012.

[SKS13] Stefanie Scherzinger, Meike Klettke und Uta Störl. Managing Schema Evolution in NoSQL Data Stores. In *Proc. DBPL*, 2013.

[ST14a] Stefanie Scherzinger und Andreas Thor. AutoShard - Declaratively Managing Hot Spot Data Objects in NoSQL Data Stores. In *Proc. WebDB*, 2014.

[ST14b] Stefanie Scherzinger und Andreas Thor. Cloud Technologien in der Hochschullehre. *Datenbank-Spektrum*, 14(2), 2014.

IBM's Analytic Warehouse Cloud Service

Torsten Steinbach, Michael Wurst

IBM Germany Research & Development
Schoenaicher Strasse 220
71032 Boeblingen
torsten@de.ibm.com
mwurst@de.ibm.com

Abstract:

Moving data warehousing into the cloud is a major opportunity for enterprises. It is also a challenge for traditional database management technology. Some of the novel requirements for a successful cloud data warehouse include simplicity (no tuning), seamless scalability, security, efficient data movement and integration with third-party tools. IBM addresses these challenges through its analytics warehouse as a service offering that is powered by IBM's BLU technology.

This presentation walks you through the major aspects of this service starting by diving into the cloud service hosting platforms SoftLayer and BlueMix and how they abstract the operation of IBM's data warehouse service from the consumption by users such as application developers or data analysis. Certain duties and operations that are traditionally performed by the database administrators on client side are now performed behind the scenes and contribute to the value that the service provides. The presentation will also show how application developers instantiate and use the warehouse service without much overhead, which allows them to focus in their actual task of developing the warehouse application. Special focus is given on advanced analytics workloads. We will show how a tight integration of the statistical language R into the overall service offering enables users to run even complex analytics efficiently close to the data in a cloud environment.

The presentation will be accompanied by life demonstration of all these capabilities.

Privacy-protecting Data Management in distributed Cloud Infrastructures

Andreas Müller, Sebastian Hudert, Victor Fäßler

TWT GmbH Science & Innovation
Ernsthaldenstraße 17
70565 Stuttgart
{firstname.lastnameg}@twt-gmbh.de

Abstract: Today's cloud infrastructure platforms such as Amazon Web Services (AWS) allow to deploy and run complex services without having to worry about scalability, reliability and general maintenance tasks. However, with distributed services running on multiple instances and on potentially untrusted nodes, the protection of data and in particular the preservation of privacy has become a huge challenge: On the one hand, data storage must be achieved in a way that an attacker with access to the file system, or in some use cases also the cloud operator, is not able to retrieve critical information such as personal data. On the other hand, the combination of individual data chunks to privacy critical information at runtime must be prevented when creating distributed applications. For example, a commercial location-based service needs information about the location of a device, as well as personal data of the owner for billing issues. Today, both of these details (location and personal data) are kept at the service although a combination of the two is not necessarily needed: a location-based service can operate without knowledge about personal data of the owner and a related billing service without the exact location. This talk focuses on such scenarios and presents ongoing work that is carried out in the BMWi-funded research project Shared E-Fleet.

The research project Shared E-Fleet develops a cloud-based e-mobility platform for electric fleets in corporate environments. The main idea is to share electric vehicles between different companies to increase their efficiency and to minimize costs. For example, a technology park consisting of many individual companies may operate a fleet and host the Shared E-Fleet cloud infrastructure allowing companies to book and use vehicles in a flexible way. One of the challenges here is that on the one hand different companies share a common fleet of electric vehicles through a common platform, while on the other hand the companies might be competitors. Thus, the separation of data storage, as well as the careful aggregation of data stored in the platform is needed.

The Shared E-Fleet architecture is composed of multiple web-services, each one dedicated to a specific functionality either related to the management and booking of the fleet or to services needed at runtime. Thus, each of these services collects and stores data to operate properly: general management services hold static information about the customers and vehicles (e.g. company names, user accounts, addresses and billing

information). Additionally, on-board units (OBUs) in the vehicles periodically push status information (e.g. the current location, state of charge, error messages) to a dynamic data pool. Other services then rely on the combination of static and dynamic data to provide the essential Shared E-Fleet services: a booking service needs to know about the customers, as well as about the current electric range of available vehicles in order to make sure to assign the correct vehicle to a requested booking. Similar combinations are needed for route optimization, charging strategies and also billing. Thus, our goal is to ensure data reduction and data economy when combining such data to create new services.

To achieve this goal we have developed a concept for the data management of privacy critical data running on a private cloud infrastructure (in our case a combination of Openstack and WSO2). The exchange of data between the individual web-services is done via a trust center service that maintains authorization policies for the access and for the combination of data. After state of the art mutual authentication, authorization for the specific attributes takes place and only if the trust center policy approves the requested combination of data based on the token of the requester, the information is aggregated and forwarded.

Additionally, a more sophisticated concept, e.g. for deploying the architecture on public clouds, is envisioned. This includes the encrypted storage of data, as well as the integration of Trusted Computing technologies to ensure that only nodes that run a valid and approved configuration of the operating system and applications are allowed to access data on a specific host.

Dynamik und Compliance in IT-gestützten Geschäftsprozessen

(DyCo 2014)

Dynamik und Compliance in IT-gestützten Geschäftsprozessen (DyCo 2014)[1]

Abstract: Änderungen im Geschäftsumfeld, ein steigender Marktdruck und hohe Kundenerwartungen führen zu einem Bedarf an Quasi-Echtzeit-Steuerung von Geschäftsprozessen. Dynamische Geschäftsprozesse sind notwendig, welche transparent sind und sich den Gegebenheiten anpassen oder anpassen lassen. Verschiedene Methoden unterstützen die Steuerung von Geschäftsprozessen in automatisierten Ausführungsumgebungen – allerdings sind auch andere Lösungen denkbar, welche im Zusammenspiel mit der heterogenen, gewachsenen IT-Infrastruktur des Unternehmens interagieren oder auf andere Prozessbeschreibungsmethoden setzen, wie deklarative Prozesse.

Ein wichtiger Treiber von Veränderung ist Compliance – die Rechtskonformität von IT Systemen sowie den unterstützenden Geschäftsprozessen. Dies ist für viele Unternehmen heutzutage eine bereichsübergreifende Herausforderung. Die relevante Anzahl verschiedener Regelwerke wächst – nationale und internationale Gesetze, branchenspezifische Vorgaben, und IT-Sicherheitsrichtlinien. Änderungen in den Regelungen und Widersprüche bringen neue Umsetzungsanforderungen. Compliance muss in dieser dynamischen Umgebung unter sich ändernden Randbedingungen sichergestellt werden – wenn sich beispielsweise Geschäftsprozesse ändern oder die IT auf eine Cloud-basierte Infrastruktur migriert wird.

Um diesen Herausforderungen zu begegnen, ist ein flexibles Zusammenspiel von Maßnahmen auf allen Unternehmensebenen notwendig – von der Geschäftsprozessebene bis hin zur Umsetzung in IT. Weiterhin ist ein Querschnitt von Funktionen im Unternehmen betroffen – es müssen Experten für rechtliche Regelwerke, für die IT-Umgebung, für die Geschäftsprozesse, für das Controlling, etc. eigebunden werden, um eine ganzheitliche Sicht und Dokumentation von Compliance zu ermöglichen. Ohne Unterstützung durch IT wird das in Zukunft – bei kürzeren Änderungszyklen und einer wachsenden Anzahl von Regeln und Regelungen – nicht mehr manuell handhabbar sein. Der Workshop zielt darauf ab, eine ganzheitliche Betrachtung der Dynamik, der Herausforderungen für Compliance, und Lösungen von Compliance im Kontext von Geschäftsprozessen und IT über die Unternehmensebenen und die Technologieebenen hinweg zu ermöglichen. Sowohl Ergebnisse aus der Forschung als auch Erkenntnisse aus der Anwendung sollen verbunden werden – besonders im Fokus steht die Verbindung von Compliance und Dynamik, aber auch Beiträge aus einem der beiden Themen sind willkommen.

[1] http://www.iaas.uni-stuttgart.de/dyco/

Workshop-Organisatoren: Prof. Dr. Leymann (IAAS Universität Stuttgart, geschäftsführender Direktor), apl. Prof. Dr.-Ing. Weisbecker (Fraunhofer IAO und IAT Universität Stuttgart, Stellvertretende Institutsleiterin Fraunhofer IAO und IAT Stuttgart)

Programmkomitee: Tom Baeyens (Effektif GmbH), Dr. Cristina Cabanillas (Wirtschaftsuniversität Wien), Dr. Gero Decker (Signavio GmbH), Jens Drawehn (Fraunhofer IAO), Dr. Sebastian Hudert (TWT GmbH), Prof. Dr. Jan Jürjens (Universität Dortmund), David Knuplesch (Universität Ulm), Falko Kötter (Universität Stuttgart), Prof. Dr. Jan Mendling (Wirtschaftsuniversität Wien), Prof. Dr. Manfred Reichert (Universität Ulm), Prof. Dr. Stefanie Rinderle-Ma (Universität Wien), David Schumm (TWT GmbH), Dr. Daniel Wutke (W&W Informatik GmbH)

Unterstützung von dynamischen Compliance Prozessen durch Business Process Management Plattformen

Daniel Burgwinkel
Neuhausstrasse 30
CH-4057 Basel
daniel.burgwinkel@outlook.com

Abstract: Die Einhaltung von gesetzlichen, regulatorischen und unternehmensinternen Richtlinien spielt insbesondere in der Pharmabranche und im Bankwesen eine wichtige Rolle. Bei der Einführung von Complianceprozessen können Business Process Managementsysteme eine wirksame Unterstützung bieten. Anhand von zwei Fallbeispielen wird analysiert, welche Faktoren die Dynamik der Complianceprozesse beeinflussen und welche Herausforderungen bei der Einführung von BPM-Systemen im Kontext Compliance bestehen.

1. Einleitung

Die Umsetzung von neuen Compliance-Vorgaben stellt Unternehmen vor die Herausforderung ihre Geschäftsprozesse anzupassen und strenger zu kontrollieren. Gleichzeitig verändern sich Unternehmen dynamisch durch Reorganisationen, Übernahmen und Abspaltungen.

Bei der Implementierung der Complianceprozesse können Business Process Managementsysteme die Umsetzung wirksam unterstützen. Am Beispiel von zwei ausgewählten Beispielen aus der Pharma- und Finanzbranche sollen die Anforderungen an Flexibilität und Dynamik der Prozesse untersucht werden.

2. Unterstützung der Compliance mit Hilfe von IT-Systemen

Im Kontext Compliance und IT können folgende Themenbereiche unterschieden werden:

Compliance-Vorgabe	Ziel	Rolle von Business Process Management (BPM)
Gesetzliche Vorgaben und Regularien für Entwicklung und Betrieb von IT-Systemen (z.B. GxP in der Pharmaindustrie)	IT-System entspricht gesetzlichen und regulatorischen Auflagen	• BPM-System muss Vorgaben erfüllen
Compliance-Vorgabe für ausgewählte Geschäftsprozesse (z.B. Sunshine Act in der Pharmaindustrie)	Geschäftsprozess wird ordnungsgemäss ausgeführt und ist nachvollziehbar	• Einsatz von BPM für Prozessautomatisierung und Umsetzung von Geschäftsregeln. • Vorgangssteuerung und Unterstützung der Mitarbeiter (Human-Centric Workflow) • Prozessdokumentation
Unternehmensweites Compliance Management System	Alle Mitarbeiter handeln rechtskonform und ethisch	• Einsatz von BPM für Review- und Genehmigungsworkflows, Nachvollziehbarkeit und Dokumentation

Tabelle 1: Compliance-Vorgaben und Einfluss auf IT Systeme

2.1 Compliance-Vorgaben für Entwicklung und Betrieb von IT-Systemen

Im Bereich der IT-Compliance muss eine Compliance-Vorgabe als Anforderung an ein IT-System umgesetzt werden [SW08]. So muss beispielsweise in der Pharmaindustrie das ausgewählte IT-System im Produktionsbereich die GxP (Good Manufacturing Practices) Kriterien erfüllen, u.a. Sicherheitsanforderungen an Login, Audittrail und Archivierung. Insbesondere für IT-Systeme, welche finanzielle Geschäftstransaktionen durchführen, besteht eine Vielzahl von gesetzlichen Auflagen.

2.2. Compliance-Vorgaben für Geschäftsprozesse

Von der IT-Compliance zu unterscheiden sind ausgewählte Compliance-Vorgaben, welche als Geschäftsregeln (Businessrules) im IT-System abgebildet werden können. Verschieden Forschungsprojekte untersuchen Möglichkeiten und neue Ansätze bei der Umsetzung von Compliance-Vorgaben durch Business Rules [KF13].

Eine Compliance Vorgabe kann auch zu einer neuen Funktion führen, welches das IT-System erbringen muss, z.B. durch eine neue Regulierung muss ein Bericht erstellt und an eine Behörde geliefert werden. Oder eine Compliance-Vorgabe erfordert, dass ein Geschäftsprozess in bestimmter Weise ausgeführt und die Aktivitäten dokumentiert werden, z.B. in der Finanzindustrie die Aufklärung eines Kunden über die Risiken eines Anlageproduktes.

Ein Beispiel für eine neue Compliance Regel in der Pharmaindustrie ist die Vorschrift zum „Physician Financial Transparency Reports (Sunshine Act)“ [AMA13]. Pharmaunternehmen müssen in den USA alle finanziellen Transaktionen zu Medizinern in einem jährlichen Bericht an die Behörde liefern. Um diese neue Compliance-Vorgabe zu erfüllen, muss das Unternehmen analysieren, welche Geschäftsprozesse betroffen sind und welcher Änderungsbedarf besteht. Das Unternehmen muss die Finanz- und IT Systeme anpassen, um diesen neuen Bericht für Behörden zu erstellen. Falls die Compliance-Vorgabe die Einhaltung finanzieller Obergrenzen verlangt, müssen die Geschäftsprozesse so gestaltet werden, dass die Einhaltung sichergestellt ist. Zudem müssen Mitarbeiter geschult werden, um die neuen Vorgaben einzuhalten.

2.3 Prozessunterstützung in einem unternehmensweiten Compliance Management System

Insbesondere globale Unternehmen haben interne Compliance-Organisationen aufgebaut und Compliance Management Systeme eingeführt, um die notwendigen Richtlinien einzuhalten und den Stakeholdern Rechenschaft abzulegen. Das Ziel eines Compliance Management Systems wird vom TÜV Rheinland wie folgt beschrieben „systematisch die Voraussetzungen in der Organisation dafür zu schaffen, dass Verstöße gegen Compliance-Anforderungen vermieden bzw. wesentlich erschwert und eingetretene Verstöße erkannt und behandelt werden können“ [TÜV11].

Im Rahmen eines unternehmensweiten Compliance Management Systems muss sichergestellt werden, dass sich Mitarbeiter rechtskonform und ethisch verhalten. Beispielsweise dokumentiert Siemens im öffentlich zugänglichen „Siemens Business Conduct Guidelines – Prinzipien und Regeln für das Verhalten aller Mitarbeiter“ das sich alle Mitarbeiter zu rechtmäßiges Verhalten, gegenseitiger Respekt, Ehrlichkeit und Integrität verpflichten und das das Unternehmen dies durch geeignete Maßnahmen sicherstellt [Si09]. Dies erfordert eine Umsetzung in den verschiedensten Geschäftsprozessen und IT-Systemen und den aktiven Einbezug der Mitarbeiter, wie z.B. bei der Korruptionsbekämpfung. Zudem kann es je nach Unternehmen dazu führen neue Prozesse und IT-Systeme einzuführen, wie z.B. E-Learningsysteme für das Compliance Training oder E-Discovery Systeme, welche die Nachforschungen im Fall von Rechtstreitigkeiten unterstützen.

3. Einsatz von BPM zur Unterstützung der Compliance

Business Process Management Systeme kommen in vielen Unternehmen zum Einsatz und werden vom Verband EABPM definiert als „System aufeinander abgestimmter (interoperabler), spezialisierter IT-Anwendungen, die ein breites Spektrum aller notwendigen Funktionen zur Modellierung und Simulation, Analyse, Entwicklung und zum Betrieb von BPM-Lösungen unterstützen“ [EA09].

In verschiedenen Publikationen werden die Einsatzszenarien und Best-Practices von BPM erörtert [Ko11], [SN11]. Business Process Management Systeme können zur Sicherstellung der Compliance in verschiedenen Einsatzgebieten genutzt werden.

- **Fokus Prozessautomation**: Einsatz einer BPM-Applikation zur Abbildung von Geschäftsregeln, welche auf Compliance Vorgaben beruhen. Geschäftsregeln (Business Rules) werden definiert als „interne oder extern gesetzte Richtlinien oder Vorschriften, die das Verhalten der Beteiligten in einem Unternehmen festlegen und damit Verhaltensspielräume einengen oder bestimmte Verhaltensweisen erzwingen“ [EA09]. Das Unternehmen kann den Nachweis der Compliance erbringen, indem es aufgezeigt, dass das IT-System anhand der definierten Regeln den Prozess unterstützt und das Risiko des Fehlverhalten eines Mitarbeiters minimiert.

- **Fokus Human-Centric Workflow / Case Management**: Eine BPM-Applikation kann zur generellen Prozessunterstützung der Mitarbeiter eingesetzt werden, z.B. Dokumente werden zur Genehmigung weitergeleitet und strukturiert in digitalen Akten abgelegt. Das System bietet den Nachweis, dass Geschäftsprozesse ordnungsgemäß durchgeführt worden, z.B. dass alle Genehmigungen erfolgt sind und die relevanten Informationen strukturiert abgelegt wurden.

- **Fokus Zusammenführung von Daten**: Einsatz eines BPM-Systems zur Vernetzung von bisher isolierten Anwendungen, um Compliance Vorgaben zur erfüllen, z.B. einen Bericht über finanzielle Transaktionen zu generieren, welcher Daten aus verschiedenen Quellen zusammengeführt.

Insbesondere globale Organisation bauen unternehmensweite BPM-Plattformen auf, um die verschiedenen Anwendungsgebiete mit einer einheitlichen Technologie abzudecken. Folgende Abbildung zeigt ein Beispiel für den schematischen Aufbau einer Plattform auf der mehrere BPM-Systeme realisiert sind:

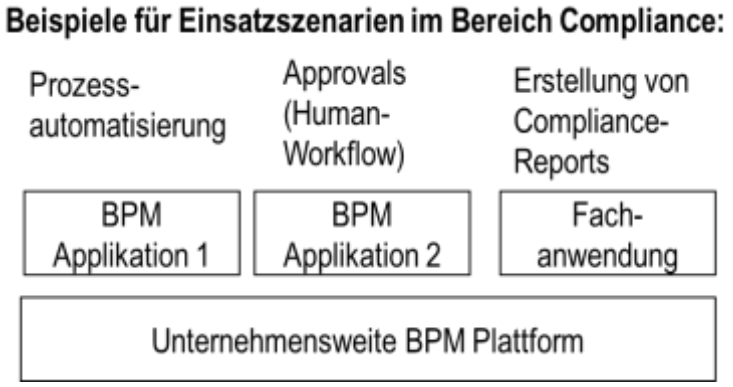


Abbildung 2: Beispiel einer unternehmensweiten BPM Plattform

4. Fallbeispiele

4.1 BPM in einem globales Pharmaunternehmen

Als Beispiel für ein BPM-System, welches direkt Complianceprozesse unterstützt soll folgender Fall angeführt werden. Ein globales Pharmaunternehmen ist in mehr als 140 Ländern vertreten und beschäftigt über hunderttausend Mitarbeiter. Das Unternehmen ist mit verschiedenen Divisionen und Produktbereichen erfolgreich am Markt vertreten. Um ausgewählte Bereiche der Compliance zu unterstützen wurde eine globale BPM-Applikation eingeführt, welche u.a. die Genehmigung von Marketingunterlagen und Veranstaltungen nach regulatorischen und ethischen Gesichtspunkten unterstützt [Bu12].

Diese BPM-Applikation kann als dynamisch und komplex bezeichnet werden. Basierend auf einem generischen Prozessmodell wurden die Genehmigungsprozesse definiert und für jede Geschäftseinheit in jedem Land konfiguriert. Da die Prozesse auf die rechtliche und organisatorische Situation jedes Landes und jeder Division angepasst werden mussten ergab sich eine erhöhte Komplexität. Durch interne Reorganisationen und die Marktdynamik, z.B. Kauf und Verkauf von Geschäftseinheiten,

müssen die Prozesse kontinuierlich angepasst werden.

4.2 BPM in einen globalen Finanzinstitut

Für den Einsatz von human-centric Workflows auf Basis eines BPM System soll ein Beispiel aus dem Private Banking angeführt werden. Das BPM System unterstützt die Prozesse vom Kundenberater in der Bankfiliale zur Auftragsabwicklung im Backoffice.

Ausgewählte Prozesse unterstützen die Zusammenarbeit mit der internen Complianceabteilung. Obwohl in einer Bank viele Aufträge direkt in Transaktionssystem

und Fachapplikationen eingegeben werden, besteht ein hoher Bedarf Aufträge zu bündeln und je nach Kundensegment durch verschiedene Geschäftseinheiten bearbeiten zu lassen, z.B. Aufträge von Geschäftskunden für Kreditkarten werden durch den Kundenberater erfasst, aber durch die Kreditkarteneinheit abgewickelt. In diesem Beispiel nutzen mehr als 10 Geschäftseinheiten aus verschiedenen Privat- und Geschäftskundensegmente jeweils bis zu 200 Prozesse. Neben direkten Complianceprozessen, z.B. Kundenüberprüfungen und Nachforschungen, stellt der human-centric Workflow sicher, dass ein Nachweis über die ordnungsgemäße Abwicklung der Aufträge erfolgt.

5. Dynamik und Komplexität der Compliance Prozesse

Basierend auf den Erfahrungen bei der Einführung der beiden beschriebenen Systeme sollen im Folgenden die Anforderungen an die Komplexität und Dynamik der Prozesse beschrieben werden und Entscheidungshilfen für den Einsatz von BPM-Systemen im Bereich der Compliance hergeleitet werden. Um beurteilen zu können, ob in einem Unternehmen der Einsatz einer BPM-Plattform für die Compliance sinnvoll ist sollten folgende Aspekte analysiert werden.

5.1 Dynamik und Komplexität der Compliance-Vorgaben

Folgende Fragestellungen helfen die Dynamik der Compliance-Vorgaben für das ausgewählte Gebiet einzuschätzen:

- Um welche **Compliance-Vorgaben** handelt es sich und welche Stakeholder sind betroffen (Mitarbeiter, Partner, Kunden)? Welche der Stakeholder sind in den Prozess involviert und welche Aktivitäten müssen durchgeführt werden (z.B. Genehmigungsprozesse)?

- Welche **Objekte** stehen im Mittelpunkt der Compliance-Vorgaben und wie soll der Nachweis der Compliance erbracht werden? Müssen Dokumente auf ihre Korrektheit überprüft werden oder muss abgeklärt werden, ob für einen Kunden bestimmte Produkte zulässig sind?

- Wie sind die externen Compliance-Vorgaben in unternehmensinterne Richtlinien umgesetzt? Gibt es **Richtlinien**, welche Grundsätze auf abstraktem Level beschreiben und konkrete Arbeitsanweisungen (Standard Operating Procedures) für die Umsetzung? Können unternehmensintern globale Standards etabliert werden oder müssen für jedes Land eigene Vorgaben entwickelt werden?

- Wie schnell ändern sich die Compliance-Vorgaben bzw. wie schnell müssen neue Compliance-Vorgaben global berücksichtigt werden? Oft tritt der Fall ein, dass

Regulierungen eines Landes, z.B. USA, von anderen Ländern in ähnlicher Art und Weise übernommen werden.

- Gibt es Abhängigkeiten zu anderen Compliance-Vorgaben die berücksichtigt werden müssen, z.B. Datenschutzrechte von Kunden, rechtliche Vorgaben bzgl. Standort der Datenhaltung?

5.2 Dynamik und Komplexität der Geschäftsprozesse und IT-Systeme

Die Geschwindigkeit bei der Umsetzung der Compliance-Vorgaben wird massgeblich durch die Dynamik und Komplexität der Geschäftsprozesse und dem Aufwand für die Anpassung der IT-Systeme beeinflusst. Bei den Geschäftsprozessen treibt die Anzahl und der Änderungsbedarf der durch die Compliancevorgabe betroffenen Prozesse die Komplexität. Bei den IT-Systemen ist die Anzahl und und Änderungsbedarf der betroffenen Applikationen massgeblich. Folgende Analysepunkte helfen den Aufwand abzuschätzen:

- Welche Geschäftsprozesse und IT-Systeme sind von den Compliance-Vorgaben betroffen? Compliance Vorgaben können für spezifische Prozesse bestehen (z.B. Wertpapierhandelsgesetz) als auch unternehmensweite Gültigkeit haben (z.B. ethisches Verhalten der Mitarbeiter).

- Welcher Änderungsbedarf für Geschäftsprozesse besteht bzw. müssen neue Prozesse eingeführt werden? Welchen neuen funktionalen Anforderungen werden an die Applikation gestellt?

- Ist nur eine Fachapplikation von den Compliance-Vorgaben betroffen oder mehrere Applikationen, bzw. ein End-to-Endprozess?

- Ist für das Einsatzgebiet eine globale IT-Lösung für alle Länder im Einsatz oder gibt es länderspezifische Lösungen?

- Welche Anforderungen bezüglich Projektvorgehen, Validierung der IT-Systeme und Standards für Dokumentenaufbewahrung werden gestellt, z.B. Branchenstandards oder Standards für Aktenmanagement in der öffentlichen Verwaltung.

- Sind cloudbasierte Fachanwendungen betroffen oder werden cloudbasierte Lösungen für die Erfüllung der Compliance-Vorgaben eingesetzt? In der Pharmabranche konnten sich in den letzten Jahren cloudbasierte Dienste erfolgreich etablieren. Obwohl in der Pharmaindustrie hohe Anforderungen bestehen konnten sich spezialisierte Anbieter erfolgreich am Markt etablieren. Ein Cloudanbieter gibt an, für den Reviewprozess von Dokumenten mehr als 40000 Nutzer aus 120 Pharmafirmen in 170 Länder erreicht zur haben [Zi14].

- Sind von den Compliance-Vorgaben Informationen betroffen die auf Mobilgeräten bzw. Tablets (z.B. für Vertriebsmitarbeiter) verarbeitet werden?

Müssen in diesem Bereich besondere Maßnahmen bzgl. Sicherheut und Datenschutz getroffen werden?

Je nach Unternehmen und Compliance-Vorgaben gibt es unterschiedlichen Szenarien. In einem Szenario kann es zu Genehmigungsprozessen in unterschiedlichen Prozessen und IT-Systemen führen, welches die Anpassung der jeweiligen Applikation erfordert. In anderen Szenarien kann eine zentrale BPM Plattform unterstützen, falls die Prozesse neu eingeführt werden müssen und durch eine Applikation unterstützt werden können.

5.3 Dynamik und Komplexität der Organisation und nationale Gesetzgebung

Werden in einer globalen Organisation Complianceprozesse eingeführt, stellt sich die Frage, wie das System die Anforderungen der Ländergesellschaften abdecken kann.

- Kann in allen Ländereinheiten der identische Prozess eingeführt werden oder muss pro Land eine Anpassung der Prozesse erfolgen bzw. müssen die Prozesse nach Herkunftsland des Kunden differenziert werden? Falls eine Anpassung notwendig ist, sind es Änderungen in der Abfolge der Approvals oder spezielle Prozesse für Ausnahmeregeln?

- Müssen pro Land spezielle Hinweise den Mitarbeitern gegeben werden (z.B. Verweise auf lokale Gesetze bzw. Hinweis in der Landessprache erfolgen)?

- Sind in den Ländern spezielle Trainings für Mitarbeiter vorgeschrieben, welche vor der Durchführung des Prozesses erfolgen müssen? Wie und wo wird der Nachweis über das Training der Mitarbeiter dokumentiert?

- Sind alle Ländergesellschaften identisch strukturiert (Aufbau und Rollen) oder sind in ausgewählten Ländern bestimmte Rollen nicht vorhanden?

- Sind ausgewählte Ländern zu Länderclustern zusammengefasst, so dass Anpassungen der Prozesse und Rollen notwendig sind? Die Prozesse müssen sowohl für Länder mit hoher Anzahl von Mitarbeitern als auch für Ländern mit geringen Mitarbeiteranzahl umgesetzt werden. So wird beispielsweise die Komplexität erhöht, falls kleinere Länder in aufstrebenden Märkten zu Clustern zusammengefasst sind.

- Sind ausgewählten Ländern spezielle Organisationsstrukturen, z.B. Joint-Ventures vorhanden?

In einem statischen Szenario kann ein einheitlicher Genehmigungsprozess top-down in alle Länder eingeführt werden. In einem dynamischen Szenario muss für jedes Land die Reihenfolge der Rollen für die Genehmiger festgelegt werden, um auf die rechtlichen Rahmenbedingungen und die Ressourcensituation des Landes einzugehen.

6. Fazit

Der Einsatz von BPM zur Abbildung von Geschäftsregeln, welche Compliance Anforderungen umsetzen, ist ein verbreitetes Einsatzgebiet. Meist sind ausgewählte Prozesse und Compliance-Vorgaben betroffen. Insbesondere große Unternehmen sind aufgefordert unternehmensweite Compliance Management Systeme einzuführen und den Nachweis über die Einhaltung der Compliance zu erbringen. Hierdurch entsteht der Bedarf Entscheidungen und Genehmigungsprozesse im Unternehmen zu strukturieren und zu dokumentieren. BPM Applikationen können hier eine wirkungsvolle Unterstützung leisten. Die angeführten Beispiele zeigen, dass besonders für global agierende Unternehmen die Komplexität der Umsetzung durch die Dynamik der Complianceprozesse beeinflusst wird. Für den zukünftigen Einsatz von BPM im Anwendungsgebiet Compliance stellt sich die Frage, wie Unternehmen ein organisatorisches Compliance Management System einführen können und welche Unterstützung das Prozessmanagement hierbei spielen kann.

Literaturverzeichnis

[AMA13] American Medical Association: Physician Financial Transparency Reports (Sunshine Act). http://www.ama-assn.org/ama/pub/advocacy/topics/sunshine-act-and-physician-financial-transparency-reports.page.

[Bu12] Burgwinkel, D.: Success factors for global rollout of BPM applications, Zürich, 2012.

[EA09] EABPM: Business Process Management Common Body of Knowledge – bpm cbok - Leitfaden für das Prozessmanagement. Verlag Dr. Götz Schmidt, Giessen, 2009.

[KF13] Koetter, F. et al.:Unifying Compliance Management in Adaptive Environments through Variability Descriptors, 2013.

[Ko11] Komus, A.: BPM Best Practice. Springer Berlin Heidelberg, 2011.

[Si09] Siemens: Siemens Business Conduct Guidelines – Prinzipien und Regeln für das Verhalten aller Mitarbeiter. http://www.siemens.com/sustainability/pool/cr-framework/business_conduct_guidelines_d.pdf.

[SN11] Slama, D.; Nelius, R.: Enterprise BPM: Erfolgsrezepte für unternehmensweites Prozessmanagement. dpunkt.verlag, 2011.

[SW08] Strasser, A.; Wittek, M.: IT-Compliance. https://www.gi.de/service/informatiklexikon/detailansicht/article/it-compliance.html.

[TÜV11] TÜV Rheinland TÜV: TR CMS 101 2011 - Standard für Compliance Management Systeme.

[Zi14] Zincahead: Zinc Homepage. http://www.zinc-ahead.com, 3.05.2014.

Compliance in BPM today - an insight into experts' views and industry challenges

Monika Kochanowski[1], Christoph Fehling[2], Falko Koetter[3], Frank Leymann[2], Anette Weisbecker[1]

[1] Fraunhofer Institute for Industrial Engineering
70569 Stuttgart, Germany
firstname.lastname@iao.fraunhofer.de
[2] Institute of Architecture of Application Systems
University of Stuttgart, 70569 Stuttgart, Germany
lastname@iaas.uni-stuttgart.de
[3] Institute for Human Factors and Technology Management
University of Stuttgart, 70569 Stuttgart, Germany
falko.koetter@iat.uni-stuttgart.de

Abstract: Compliance has evolved into a mature research topic in business process management (BPM). However, most companies and BPM tool providers do not incorporate advanced methods from scientific research into their everyday life. This work gives an insight on the state in the industry based on expert interviews in companies and BPM tool provider companies. Based on this, the findings are condensed in challenges and research questions, showing the need for additional studies.

1 Introduction

Compliance is a current research topic in business process management (BPM). Examples include frameworks for overall compliance management in the complete lifecycle [TZH+12], methods for solving single issues with certain design time checks [SFG+11] [SLSW10], as well as monitoring for compliance exceptions [SLS10].

However, the state of the practice shows a different picture. Companies often have a heterogeneous environment and deal with compliance issues on distributed levels. First of all, IT compliance is managed by the IT department. Secondly, compliance to certain laws is ensured by lawyers, situated in the departments law, revision, or controlling. Thirdly, compliance in business processes is focused on risk management and internal control systems, which is required from recent legislation and a feature that BPM tools offer.

In earlier work [KKR+13], the details of the unification of compliance in various departments have been tackled from a technical view point. The compliance descriptor concept was introduced, which interlinks necessary artifacts on several levels - IT, business process, and law. However, this has not included details on experts' opinions and methodological derivations.

In this work, related work on the topic compliance and BPM is reviewed with a focus on industry studies in Germany. The main contributions are expert interviews in the industry in different departments and telephone interviews with BPM tool providers. Finally, a conclusion and an outlook are given.

2 Related work

The related work is structured it in (1) overall compliance management BPM concepts, (2) compliance in certain application areas like IT or certain types of processes, and (3) empirical research on the needs of compliance based on the finance industry in Germany.

2.1 Compliance management in BPM

In a model of a business process one verified aspect often considered for compliance is termination of the process and reachability of modeled activities [GB13]. More specific, compliance may be described as requirements that either have to be checked or state how activities have to be done [SAL+10]. Requirements may already be expressed as verifiable rules to ensure a well-structured and compliant data-flow and control-flow [SGL+11]. One approach to guide the implementation of reoccurring compliance requirements in business processes is the use of process fragments [SLM+10]. In order to maintain a manageable overview of such modeled processes and especially the included compliance aspects, user-specific views on processes may be defined [SLS10]. During runtime, the monitoring of such business process models may be ensured using the a priori defined rules [LRMKD11]. For the verification of compliance in business processes, many approaches rely on formal models to which the business process and described compliance rules are translated for a formal evaluation. In [AKM09], Reo - a channel-based coordination language - is used for compliance verification. [GV06] uses temporal deontic logic focusing on permission associated with activities of a business process, [Gov13] provides an abstract framework to model deontic notions. Temporal logic and model checking are used for compliance verification in [MMJ98], [ADW08], and [SFG+11]. The work [AJKL06] describes how the usage of workflow management systems and middleware can support compliance enforcement. Specific compliance functionalities across the entire BPM lifecycle are described in [CRRC11], whereas the problem how different interpretations of rules can be dealt with has been partly tackled in an technical approach [CRRC12].

2.2 Compliance in specific application areas

Aside from the generic compliance management in BPM, further approaches have been developed for specific application domains. If compliance shall be ensured for business processes modeled in BPMN, [GK07] and [SLSW10] propose suitable frameworks

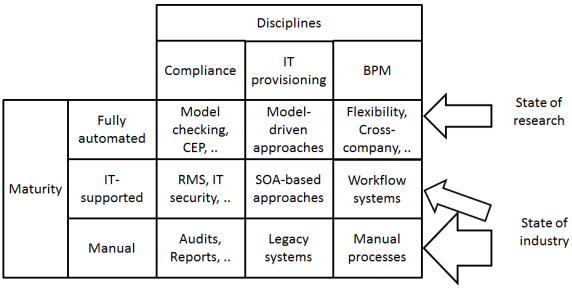


Figure 1: Compliance in research and industry

and BPMN modeling language extensions focusing on compliance. [LMX07] elaborates model checking techniques used for compliance verification in business processes expressed in BPEL. Regarding specific application domains, approaches can be found for handling compliance in service-oriented architectures (SOA) [DCD+09] and Cloud Computing [SLSW10] [BDA+10]. An overall model-driven approach for SOA is presented in [TZH+12].

2.3 Empirical studies on the state of compliance in industry

In the industry several companies, especially consulting companies, already conducted surveys on compliance. One is [Bea12], which discusses compliance management as an increasing challenge for insurance companies. Main findings include that process design will be part of the compliance task - products, services, processes, etc. have to be coordinated between compliance departments and operating departments. Main areas with focus on compliance are purchasing, law, and IT. Very interesting is that 93% of the participants mentioned that their compliance activities are not or only partly supported by IT. The authors follow that one of the main challenges in the future is to handle the multiple different compliance sources (international and national laws, code of conducts, etc.) and provide IT support thereof. It is seen as a challenge to monitor compliance in the process aspect as well. The authors state that control activities have to be automated in the future.

Another study discusses the compliance organization in insurance companies [BDO10]. One interesting aspect is that most reporting is done ad-hoc with 31%. One surprising aspect is that 53% state that the compliance functionality is not involved in any other company processes, which shows a lack of integration and communication between company departments. Only 11% state that other processes are involved, like risk management, and only 3% mention IT processes. The study states that compliance needs to be better integrated into business processes. The new edition of the survey in 2012 [BDO13] shows a very small improvement in the integration of compliance in the business processes. Additional potentials lie in a higher process efficiency and consolidation of compliance reports.

The study [KPM13] shows that compliance management is seen as very important, not

Figure 2: Layers of compliance (similar to [Bea12])

only by large companies, but also by medium sized companies. The integration of business partners is also seen as a main topic.

Figure 1 gives a view on the maturity of compliance management and the underlying layers - IT and BPM - in industry and in state of research. Whereas companies - especially in insurance industry - still perform compliance management mostly manually and do not have overall workflow systems in place, the state of research is much more sophisticated.

3 Expert interviews: compliance in companies

The three conducted expert interviews were targeted at three different peer groups in the industry: compliance departments, IT departments, and BPM departments. This matches the compliance layers (see Figure 2). Compliance is a cross-department issue, and depending on the company organization, the goals of the different department do not have to match completely. In preparation of the three interviews, an interview guide has been created which has been applied in all three interviews in order structure the answers, containing 10 questions on compliance and 2 specifically targeted on unification and adaptation. Additionally, the event *compliance compact* in the finance industry has been visited to grasp the topics discussed there. The findings are then summarized at the end of this section.

3.1 Expert interview on compliance within an IT department

The interview took place in an IT department in a company (more than 1000 employees). The company acts as a service provider for other branches in the Finance industry in Germany. The software architecture decision is to provide a private cloud - departments can host services in the company's IT department. Because of the importance of the IT applications and the IT positioning of the company in the finance industry, several audits take place every year, checking for IT compliance. To ensure the success of these audits, the governance framework COBIT 4 [IT 07]. The employed method is pragmatic: based on a main auditor, COBIT is applied and checked. When other auditors are involved, the result of the compliance check based on this framework is presented. If different requirements persist, they are discussed with the main auditor and incorporated into the application of the governance framework.

Considering compliance in business processes, the challenge is to begin with the documentation of processes across the departments. Currently the documentation is done using

standard office software. One person in each department is responsible for compliance. In the future, a centralized BPM is the goal of the company. As long as no input is given from the operating departments, the IT department cannot support BPM compliance within the company. This is currently perceived as a bottom-up approach from the IT department for ensuring compliance in BPM.

The current main issues are the identification of the matching and valid compliance regulations. To change, search, and find such regulations is a challenge. Compliance is not strictly defined - the most important issue is to extract measurable metrics from laws and rules. Documentation of compliance is not clearly defined. However, the interview partner's opinion is that automated reporting functionality is necessary in the future. Compliance of partners should be monitored by the operating departments, which is difficult to fulfill when considering the complexity of ensuring compliance requirements from business processes to IT.

3.2 Expert interview on compliance within a BPM department

The interview took place in a BPM department in one of the largest German insurance companies. The company also operates in parts of Europe and has over 10.000 employees. The company employs the six sigma methodology [BMK00] for BPM. Currently, process controlling is done by sample analysis of single files which document the process outcome. The process management is decentralized. One important topic within process management is the risk management and internal control system within financial reporting. Because of the size of the company a risk landscape has evolved, which is adapted according to the size of the company. Basically, every decision in the process is perceived as a risk.

The goal in BPM is to have complete process documentation in near future. The compliance department is responsible for defining how certain processes have to be done, the revision department is involved in the coordination. A BPM tool is employed for modelling business processes and risks. However, adaptations according to law changes are done manually. In the future this is perceived as a challenge, as changes are becoming more frequent and compliance requirements are manifold. Automation is seen as a possible solution.

Main issues for BPM and compliance are the acceptance of the users and resources within the company. Currently it is difficult to prove the main benefits of how controls actually support the company business processes.

3.3 Expert interview on compliance within a compliance and law department

The interview took place in a compliance and law department in a large German insurance company with above 10.000 employees. The software architecture is legacy software. Compliance management concepts are subdivided into three levels: (1) conceptual, (2)

matching the requirements, and (3) checking the results of compliance management. Currently the company is stated on level (2). One open question is where compliance actually starts and ends - beginning from risk management, as well as ensuring compliance not only in main processes but also in support processes like the food processing within the insurance company.

The organization of compliance management is done within a compliance management circle in a taskforce across several insurance companies. Law monitoring is also done in this way - the insurance circle informs the participants about relevant future legislation. In compliance management, all business processes are included - especially bonds processes, but also credit and real estate processes. It is interesting that the regulation Solvency II [Ric09] has been postponed to take effect twice in the last year, showing how difficult ensuring compliance is and how the industry struggles with regulation.

The main issue from a compliance perspective is that many different regulations have to be considered. Additionally, it is difficult to check the compliance of several processes, especially in marketing and sales departments, when performed on an operating level. Many changes in laws also pose additional challenges, because the compliance documentation has to be kept up to date and adapted manually. New products cause many process changes, with an increase in the last years. It is difficult to extract measurable metrics for compliance. Automation will be needed - by the increase in laws and changes a manual handling is not possible anymore in the future.

3.4 Findings in talks at a compliance conference in the finance industry

Additionally an event called *compliance compact* was visited, where insurance, banks, and the finance industry in general meet to discuss current compliance issues, with about 40 participants. Several talks were held by stakeholders from the industries and their service providers. Topics relevant to BPM and compliance are listed below.

Compliance and sustainability: The topic of corporate sustainability and social responsibility was discussed. In Germany, sustainability is seen as part of process management and governance. The German sustainability codex [Rat14] offers regulations and metrics for measuring that. It is interesting that this is a mixture of compliance requirements - a traditional documentation issue - as well as monitoring requirements of ecological key performance indicators.

FATCA: The foreign account tax compliance act (FATCA) [Int] is an American law which basically forces international banks to check if their customers have to pay taxes in the United States. This poses a challenge for institutes all over the world. A special system for mining and classifying customer data can be employed to solve this task. Here it is shown how data mining and BPM compliance requirements interleave.

Sales codex for insurances [GDV]: Insurance companies work together with agents, which distribute insurances of different insurers. Checking compliance is difficult in both ways: On the one side, each insurance company would have to check all their agents for compliance to the sales codex. On the other hand, all agents have to be compliant to all applicable

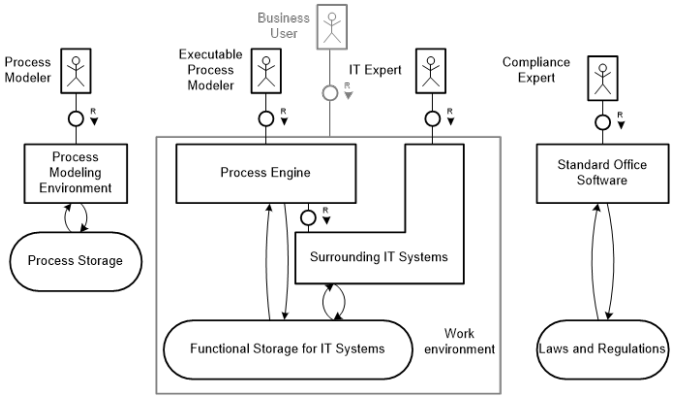


Figure 3: Current state of the architecture from experts' interviews

national and international laws and regulations as well as to the individual interpretations thereof.

Compliance software: Special compliance software can be used for managing regulations, document workflows, the quality management handbook, and process documentation in a revision safe way. The risk relevant decisions are documented. Altogether, compliance is basically brought together with document management processes.

Summarizing the above points, it shows that compliance in special application areas is discussed most, followed by the integration in IT and integrating partners. In individual discussions it showed that especially current law changes are most urgent. Depending on the stakeholder, the view is limited on their operating departments and their current problem. It is difficult to maintain and report compliance as well as to generate a documentation, as it has to be redone for every law / regulation. As many departments work isolated on some topics, coordination is difficult.

3.5 Summary of expert interviews in companies

According to the findings in the sections above, the current state of compliance management is segregated for each of the participating roles shown in Figure 3. The process modelers work mostly on their own, within a process modeling environment and a process storage depicted on the left of Figure 3. The compliance experts also work with their separate tools, mostly standard office software, generating compliance reports based on internal audits and single case analysis as seen on the right of Figure 3. Finally, the business user works within his work environment depending on functional IT systems, which can be enhanced by a process engine in some cases. The challenges from the interviews are summarized in Table 1.

A *unification concept* for bridging the gaps between different functions and IT systems within the company is not given. Further on, apart from round-trip BPM a *round-trip com-*

Table 1: Insights: Challenges in expert interviews, legend: ✓mentioned

Expert interview	IT department	BPM department	Compliance department	Compliance conference
Communication and coordination across departments	✓	✓		✓and partners
Increasing number of requirements and manual effort	✓		✓	✓
Identification of relevant laws and definition of measurable metrics	✓		✓	
Increasing number of changes and needed automation		✓	✓	✓
Other	Increasing number of audits	Acceptance of BPM in general	Compliance scope definition	Integration across systems

pliance management is needed, which interconnects the requirements and the measures taken. However, to generalize the findings, a much larger data set is needed in the future.

4 Expert interviews: compliance from a BPM tool provider point of view

4.1 Interviews with BPM tool providers on compliance

Compliance is a topic incorporated in many BPM tools. Currently, an internet research has shown that especially the topics of risk management and internal control system play an important role in the branch. Based on these findings, six BPM tool providers have been informally interviewed with an interview guideline containing seven questions in short telephone conferences in order to evaluate the importance of the topic and the willingness to give in-depth information in a further study.

Interview 1 The tool provider is very interested and sees an important topic in compliance in BPM. It is possible to design risk models and add risks to process tasks in the tools. In process execution, check tasks are triggered. However, no monitoring is included, but the tool provider counts on that development in the future. The most important industry branches for compliance are the finance and life science industry.

Interview 2 Another tool provider is focused on process execution. Compliance is ensured

by providing a business rule management framework, which works on data extracted from processes. The current process execution is not interlinked with compliance tools. The hosting of the processes is decided on a case by case basis.

Interview 3 The third tool provider claims to have a full compliance support. Included are reporting features as well as conformity to several governance frameworks and guidelines. Process lifecycle support is also given. However, the details of this statements have to be evaluated in the following.

Interview 4 The fourth tool provider has a model-driven compliance methodology. Risks and controls are mapped to process steps and can be analyzed in a list export. The documentation is based on quality criteria. The method consists of creating documentation, extracting it to workflow layer, add controls, and check these controls within control goals. Periodic assessments are done within the internal control system. Measures are extracted if the controls are not adequate or not effective. The company sees compliance and quality management as interleaved. Additionally versioning and tracking of changes are part of their solution. Compliance is seen as a lever for BPM in companies.

Interview 5 The next provider also offers a process platform, consisting of 12 single products like BPM tools, document management systems, etc. Additionally risk management and revision management is done. Internal control systems can be introduced together with the BPM tool. Risk management is focused on risk reporting for centralized management. Additionally question catalogues can be created to add risks to a process in a structured way. They are evaluated qualitatively and quantitatively. The analysis and documentation as well as measures are documented. Life science is seen as main industry, but finance industry is developing fast, especially in the field of risk management.

Interview 6 The last provider uses a workflow engine to support compliance and process management activities, but not for execution. However, internal control system activities will be in the product in the future. Risk management is included, as well as version management and revision management. Reporting is supported based on the results mentioned above. Main branches are finance and government. Quality management will be integrated. The process execution environment from a partner provider can be integrated, providing additional monitoring technology.

4.2 Summary and arising research questions for future BPM study

Preliminary insights are summarized in Table 2, but a structured BPM tool study is needed in the future. The limitations of the telephone interviews include the small sample size as well as the short interview duration. The research questions for a future study are:

What is the current understanding of *compliance* within BPM tools? Which are the current main features of compliance management supported by the tools ? How does compliance management in BPM tools integrate with document management, revision management, quality management, and IT governance? Which features will be needed by the customers in the future? How is the future development towards compliance monitoring seen by the BPM tool providers within the context of complete process lifecycle support?

Table 2: Insights: BPM tools and compliance features and industries, legend: ✓explicitly mentioned as included, - explicity not included, else: no information

Tool Provider	1	2	3	4	5	6
Risk management	✓	-		✓	✓	✓
Quality management				✓		
Revision management			✓	✓	✓	✓
Reporting			✓	✓	✓	✓
Internal control system	✓	-		✓	✓	✓
Finance industry focus	✓	-		✓	✓	✓
Life science industry focus	✓	-		✓	✓	-

5 Conclusion and outlook

In this paper the topic of compliance and BPM has been tackled from three different sides: (1) Literature research in related work, (2) the view on compliance management has been accessed in interviews with company employees in different departments, considering IT, business processes, and law, and (3) the view on compliance management by BPM tool providers, which have been contacted in telephone interviews. It has been shown that a gap between industry and current research persists. An important step would be to generalize the answers based on a larger data set. The small sample size of the interviews and the broad answers show that it is necessary to perform a larger survey with a structured questionnaire. In future work, a study of BPM tool providers will be conducted. It will also give an insight into the compliance needs of BPM customers in the future. The findings will serve as a basis for further developments based on recent research of the authors.

Acknowledgements

The work published in this article was funded by the Co.M.B. project of the Deutsche Forschungsgemeinschaft (DFG) under the promotional reference SP 448/27-1. The authors would like to thank all interview partners.

References

[ADW08] Ahmed Awad, Gero Decker, and Mathias Weske. Efficient compliance checking using bpmn-q and temporal logic. In *Business Process Management*, pages 326–341. Springer, 2008.

[AJKL06] R. Agrawal, C. Johnson, J. Kiernan, and F. Leymann. Taming Compliance with Sarbanes-Oxley Internal Controls Using Database Technology. In *22nd International*

Conference on Data Engineering (ICDE'06), pages 92–92, Atlanta, 2006. IEEE.

[AKM09] Farhad Arbab, Natallia Kokash, and Sun Meng. Towards using reo for compliance-aware business process modeling. In *Leveraging Applications of Formal Methods, Verification and Validation*, pages 108–123. Springer, 2009.

[BDA+10] Ivona Brandic, Schahram Dustdar, Tobias Anstett, David Schumm, Frank Leymann, and Ralf Konrad. Compliant cloud computing (c3): Architecture and language support for user-driven compliance management in clouds. In *Cloud Computing (CLOUD), 2010 IEEE 3rd International Conference on*, pages 244–251. IEEE, 2010.

[BDO10] BDO AG Wirtschaftsprüfungsgesellschaft. Compliance Survey bei Versicherungen, 2010.

[BDO13] BDO AG Wirtschaftsprüfungsgesellschaft. Compliance Studie bei Versicherungen, 2013.

[Bea12] BearingPoint GmbH. Agenda 2015 : Compliance Management als stetig wachsende Herausforderung für Versicherungen, 2012.

[BMK00] Bo Bergman, Kjell Magnusson, and Dag Kroslid. *Six Sigma: The Pragmatic Approach*. Studentlitteratur, Lund (Schweden), 2000.

[CRRC11] Cristina Cabanillas, Manuel Resinas, and Antonio Ruiz-Corts. Exploring Features of a Full-Coverage Integrated Solution for Business Process Compliance. In *Advanced Information Systems Engineering Workshops*, volume 83 of *Lecture Notes in Business Information Processing*, pages 218–227. Springer Berlin Heidelberg, 2011.

[CRRC12] Cristina Cabanillas, Manuel Resinas, and Antonio Ruiz-Corts. Introducing a Mashup-Based Approach for Design-Time Compliance Checking in Business Processes. In *Advanced Information Systems Engineering Workshops*, volume 112 of *Lecture Notes in Business Information Processing*, pages 337–350. Springer Berlin Heidelberg, 2012.

[DCD+09] Florian Daniel, Fabio Casati, Vincenzo D'Andrea, Emmanuel Mulo, Uwe Zdun, Schahram Dustdar, Steve Strauch, David Schumm, Frank Leymann, Samir Sebahi, et al. Business compliance governance in service-oriented architectures. In *Advanced Information Networking and Applications, 2009. AINA'09. International Conference on*, pages 113–120. IEEE, 2009.

[GB13] Heerko Groefsema and Doina Bucur. A survey of formal business process verification: from soundness to variability. In *Proceedings of International Symposium on Business Modeling and Software Design (BMSD)*, 2013.

[GDV] GDV e.V. Verhaltenskodex für den Vertrieb.

[GK07] Aditya Ghose and George Koliadis. Auditing business process compliance. In *Service-Oriented Computing–ICSOC 2007*, pages 169–180. Springer, 2007.

[Gov13] Guido Governatori. Business Process Compliance: An Abstract Normative Framework. *it - Information Technology*, 55(6):231–238, 2013.

[GV06] Stijn Goedertier and Jan Vanthienen. Designing compliant business processes with obligations and permissions. In *Business Process Management Workshops*, pages 5–14. Springer, 2006.

[Int] Internal Revenue Service. Foreign Account Tax Compliance Act.

[IT 07] IT Governance Institute. *COBIT 4.1: Framework, Control Objectives, Management Guidelines, Maturity Models*. IT Governance Insitute, Rolling Meadows (USA), 2007.

[KKR+13] Falko Koetter, Monika Kochanowski, Thomas Renner, Christoph Fehling, and Frank Leymann. Unifying Compliance Management in Adaptive Environments through Variability Descriptors (Short Paper). In *SOCA 2013, IEEE 6th International Conference on*, pages 214–219. IEEE, December 2013.

[KPM13] KPMG AG Wirtschaftsprüfungsgesellschaft. Analyse des aktuellen Stands der Ausgestaltung von Compliance Management- Systemen in deutschen Unternehmen, 2013.

[LMX07] Ying Liu, Samuel Muller, and Ke Xu. A static compliance-checking framework for business process models. *IBM Systems Journal*, 46(2):335–361, 2007.

[LRMKD11] Linh Thao Ly, Stefanie Rinderle-Ma, David Knuplesch, and Peter Dadam. Monitoring business process compliance using compliance rule graphs. In *On the Move to Meaningful Internet Systems: OTM 2011*, pages 82–99. Springer, 2011.

[MMJ98] S Mauw, R Mateescu, and W Janssen. Verifying business processes using SPIN. In *Proceedings of the International SPIN Workshop*, pages 21–36, 1998.

[Rat14] Rat für Nachhaltige Entwicklung. Der Deutsche Nachhaltigkeitskodex, 2014.

[Ric09] Richtlinie Solvabilität II. Richtlinie 2009/138/EG des europäischen Parlaments und des Rates vom 25. November 2009 betreffend die Aufnahme und Ausübung der Versicherungs- und der Rückversicherungstätigkeit, 2009.

[SAL+10] David Schumm, Tobias Anstett, Frank Leymann, Daniel Schleicher, and Steve Strauch. Essential Aspects of Compliance Management with Focus on Business Process Automation. In *ISSS/BPSC*, pages 127–138, 2010.

[SFG+11] Daniel Schleicher, Christoph Fehling, Stefan Grohe, Frank Leymann, Alexander Nowak, Patrick Schneider, and David Schumm. Compliance domains: A means to model data-restrictions in cloud environments. In *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, pages 257–266. IEEE, 2011.

[SGL+11] Daniel Schleicher, Stefan Grohe, Frank Leymann, Patrick Schneider, David Schumm, and Tamara Wolf. An approach to combine data-related and control-flow-related compliance rules. In *SOCA 2011, IEEE International Conference on*, pages 1–8. IEEE, 2011.

[SLM+10] David Schumm, Frank Leymann, Zhilei Ma, Thorsten Scheibler, and Steve Strauch. Integrating compliance into business processes. *Multikonferenz Wirtschaftsinformatik 2010*, page 421, 2010.

[SLS10] David Schumm, Frank Leymann, and Alexander Streule. Process views to support compliance management in business processes. In *E-Commerce and Web Technologies*, pages 131–142. Springer, 2010.

[SLSW10] Daniel Schleicher, Frank Leymann, David Schumm, and Monika Weidmann. Compliance scopes: Extending the BPMN 2.0 meta model to specify compliance requirements. In *SOCA 2010, IEEE International Conference on*, pages 1–8. IEEE, 2010.

[TZH+12] Huy Tran, Uwe Zdun, Taâid Holmes, Ernst Oberortner, Emmanuel Mulo, and Schahram Dustdar. Compliance in service-oriented architectures: A model-driven and view-based approach. *Information and Software Technology*, 54(6):531–552, 2012.

Soziale Software und Dynamik im Geschäftsprozessmanagement

Jens Drawehn[1], Dominic Walter[2]

[1]CC Softwaremanagement
Fraunhofer IAO
Nobelstraße 12
70569 Stuttgart
jens.drawehn@iao.fraunhofer.de

[2]Hochschule für Technik Stuttgart
Schellingstraße 24
70174 Stuttgart
dominicwalter@web.de

Abstract: Geschäftsprozessmanagement wird seit mehreren Jahrzehnten in vielen Unternehmen und Organisationen für unterschiedliche Aufgaben von der Prozessmodellierung über die Prozessanalyse bis hin zur Prozessausführung mittels IT-Systemen eingesetzt. Parallel dazu sind in den letzten Jahren Werkzeuge zur Unterstützung des Personal Networking, des Informationsaustauschs und der Zusammenarbeit entstanden, die zusammenfassend als *soziale Software* bezeichnet werden und zunehmend auch in Unternehmen zum Einsatz kommen. Im Beitrag werden die möglichen Berührungspunkte zwischen sozialer Software und dem Geschäftsprozessmanagement aufgezeigt und anhand aktueller Studien die bestehenden Einsatzmöglichkeiten sozialer Software für das Geschäftsprozessmanagement dargestellt. Dabei wird der Schwerpunkt auf eine verbesserte Dynamik bei der Durchführung von Änderungen an Prozessmodellen gelegt, es werden aber auch weitere Einsatzmöglichkeiten betrachtet.

1 Ausgangssituation und Motivation

Im Geschäftsprozessmanagement (GPM) werden u.a. die mangelnde Kommunikation zwischen verschiedenen beteiligten Gruppen und eine ungenügende Unterstützung des Wissensmanagements als konkrete Defizite [Wa14, S. 13f] benannt. Diese erschweren bzw. verzögern die Erstellung und Pflege von Prozessmodellen, was dazu führen kann, dass die bestehende Prozessdokumentation nicht den aktuellen Stand der Prozesse wiedergibt oder nicht alle Aspekte beinhaltet, die für den konkreten GPM-Anwendungsfall relevant sind. Weiterhin gibt es bisher nur wenige verfügbare GPM-Werkzeuge, die eine gute Unterstützung für die verschiedenen Arten von Prozessen bieten [RM13, S. 8]. So verfügen beispielsweise viele Workflow-Management-Systeme, die für die Unterstützung strukturierter Prozesse entwickelt wurden, nicht über passende Mechanismen für wissensintensive, kreative Prozesse, wie sie unter dem Schlagwort „Dynamic Case Management“ (DCM) oder auch „Adaptive Case Management“ (ACM)

zusammengefasst werden. Dadurch ist eine Unterstützung kreativer Prozesse mit GPM-Werkzeugen nur eingeschränkt möglich, so dass die Ausführung dieser Prozesse entweder nicht ausreichend unterstützt wird oder ihre Implementierung losgelöst vom übrigen Prozessmanagement erfolgt.

Soziale Netzwerke verwenden die Techniken des Web 2.0, um Nutzergemeinschaften beim Austausch von Informationen zu unterstützen [Pä13, S. 22]. Dabei werden soziale Netzwerke (oder allgemeiner: Soziale Software) nicht mehr nur im privaten Bereich eingesetzt, sondern zunehmend auch in Unternehmen und Organisationen. Die Einsatzbereiche sind dabei sehr unterschiedlich und gehen von der Nutzung öffentlicher sozialer Netzwerke für Marketingzwecke bis zum Einsatz von Enterprise Social Networks (ESN) für unternehmensinterne Aufgaben.

Soziale Software scheint grundsätzlich geeignet, einen wesentlichen Teil der oben angesprochenen Defizite im Bereich GPM zu beheben. Die hohe Bedeutung des GPM in den Unternehmen und die zunehmende Verfügbarkeit von sozialer Software haben in den letzten Jahren bereits dazu geführt, dass Funktionen aus dem Bereich der sozialen Software auch für das GPM genutzt werden. Dabei kann die erfolgreiche Einbindung größerer Benutzergruppen in das Prozessmanagement zu einer größeren Dynamik führen, weil Prozessänderungen schneller abgestimmt und umgesetzt werden können.

Im folgenden Abschnitt werden die verschiedenen Aufgabestellungen des GPM beschrieben und eine Definition und Abgrenzung des Begriffs „soziale Software“ anhand der unterschiedlichen Funktionen vorgenommen. Anschließend wird der aktuelle Stand des Einsatzes von sozialer Software für das GPM betrachtet. Dazu werden verschiedene vorliegende Untersuchungen herangezogen sowie die Vorabergebnisse einer Marktstudie, die derzeit durch das Fraunhofer IAO erstellt wird. Dabei wird gezeigt, welchen Beitrag das Social BPM zur Dynamik im GPM leisten kann. Den Abschluss bilden eine Zusammenfassung der wesentlichen Aussagen und ein Ausblick auf die zu erwartenden Entwicklungen in der nahen Zukunft.

2 Das Themenfeld „Social BPM“

Die Aufgaben des GPM lassen sich am besten anhand des Lebenszyklus von Geschäftsprozessen gliedern, wie er z.B. (basierend auf [DLMR13, S. 21]) mit den Phasen Prozessidentifikation, Ist-Modellierung, Prozessanalyse, Soll-Modellierung, Implementierung, Prozessausführung und Überwachung/Controlling in [Wa14, S. 5] beschrieben und in der folgenden Abbildung dargestellt ist.

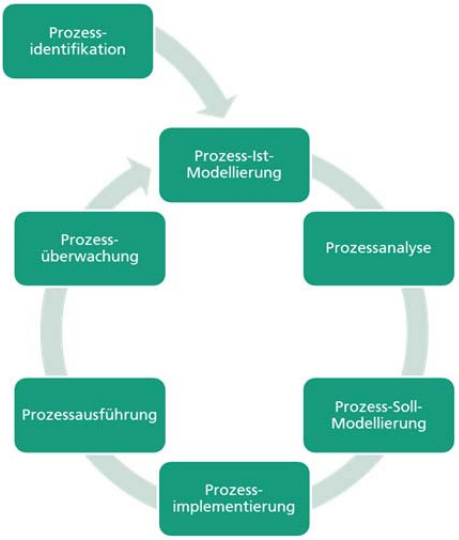


Abbildung 1: Der Lebenszyklus von Geschäftsprozessen

Ebenfalls in [Wa14, S. 8] findet sich eine Liste der Kategorien sozialer Software, gegliedert in die drei Bereiche Personal Networking, Information Sharing und Collaboration (s. Abbildung 2).

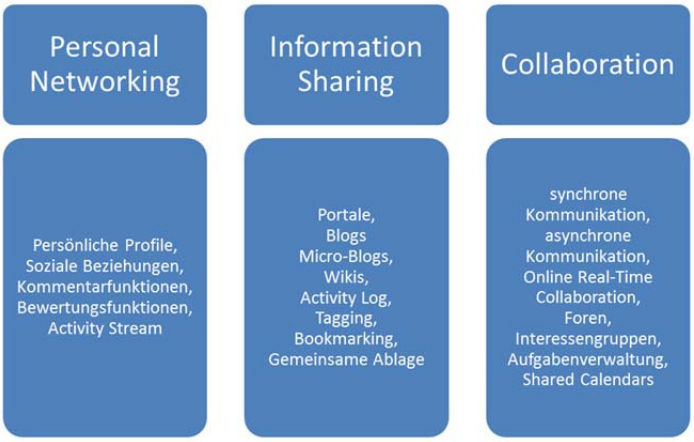


Abbildung 2: Kategorien sozialer Software nach [Walter 2013]

Interpretiert man den Begriff „Social BPM“ als den Einsatz von Funktionen aus dem Bereich der sozialen Software für das GPM, dann lassen sich die Kategorien sozialer Software den Phasen des GPM-Lebenszyklus gegenüberstellen, wobei verschiedene

Phasen des GPM-Lebemszyklus durch verschiedene Kategorien von sozialer Software auf unterschiedliche Art unterstützt werden können. Eine solche Gegenüberstellung ist in [Wa14, S. 16] vorgenommen worden, sie ist im Überblick in Abbildung 3 dargestellt.

		Prozesslebenszyklus	Identifikation	Ist-Modellierung	Analyse	Soll-Modellierung	Implementierung	Ausführung	Überwachung/Controlling	eigenständige Software	unterstützende Funktion
Personal Networking	Soziale Netzwerke	Persönliche Profile		+				+			X
		Soziale Beziehungen									X
		Kommentarfunktionen		+		+			+		X
		Bewertungsfunktionen	+		+	+			+		X
	Activity Stream	Live Status Update	-	-	-	-	+	-	+		
		News/RSS Feed	-	-	+	+	+	+	+		
Information Sharing		Portal/Prozessportal								X	X
		Blogs	+							X	X
		Micro-Blogging			+					X	X
		Wikis		+		+				X	X
		Activity Log	-	+	+	+	+	+	-		
		Tagging-Funktionen									X
		Bookmarking-Funktionen									X
		Gemeinsame Dokumentenablage						+			X
Collaboration	synchrone Kommunikationsfunktionen	Voice/Video								X	X
		Instant Messaging								X	X
	asynchrone Kommunikationsfunktionen	E-Mail								X	X
	Online Real-Time Collaboration	Online Real-Time Collaboration				+		+		X	X
		Real-Time Editing		+		+		+			X
		Group Decision Support	+	-	-	+	-	+	+	X	
		Foren	+	-	-	-	+	-	+	X	
		Interessengruppen	-	+	+	+	+	-	+		
		Aufgabenverwaltung									X
		Shared Calendars									X

Abbildung 3: Zuordnung von Funktionen sozialer Software zu den Phasen des GPM

In Abbildung 3 wird unterschieden zwischen der gezielten Unterstützung sozialer Software für einzelne Phasen, die in den Spalten „Identifikation“ bis „Überwachung / Controlling“ mit + oder – gekennzeichnet ist. Kategorien sozialer Software, die eine

unterstützende Funktion darstellen und grundsätzlich in allen Phasen des GPM eingesetzt werden können, sind in der Spalte „unterstützende Funktion“ mit „X“ gekennzeichnet. Unterstützende Funktionen, die darüber hinaus in einzelnen Phasen des GPM gezielt für konkrete Aufgabenstellungen eingesetzt werden können, sind in den entsprechenden Spalten mit einem + markiert. Das ist zum Beispiel für Wikis in den Phasen der Ist- und Soll-Modellierung der Fall, Wikis können hier gezielt dafür verwendet werden, Modellelemente mit textuellen Beschreibungen zu hinterlegen. In der Spalte „eigenständige Software“ sind diejenigen Kategorien mit einem „X“ markiert, die als eigenständige Software auftreten können (nicht: müssen), die hier nicht markierten Kategorien treten immer nur als Bestandteil umfassenderer Lösungen auf.

3 Aktueller Stand

Die Ergebnisse der in [Wa14, S. 37] durchgeführten Untersuchung zeigen, dass es beim Grad der Umsetzung von Funktionen aus dem Bereich der sozialen Software in den bestehenden GPM-Werkzeugen erhebliche Unterschiede gibt. So sind Portale, Kommentar- und Bewertungsfunktionen, asynchrone Kommunikationsfunktionen (Nachrichtenversand, meist in Form von E-Mails) und die Aufgabenverwaltung (u.a. im Zusammenhang mit Freigabeworkflows) in den meisten aktuellen GPM-Werkzeugen bereits vorhanden. Bei anderen Funktionen aus dem Bereich der sozialen Software ist der aktuelle Verbreitungsgrad deutlich geringer, so werden z.B. Wikis, Blogs und Foren nur von wenigen GPM-Werkzeugen unterstützt.

Eine aktuelle Untersuchung [RM13, S. 7] zeigt, dass GPM-Werkzeuge zunehmend verschiedene Prozessarten unterstützen. Als Kategorien von GPM-Werkzeugen werden hier BPM Suites (mit dem Schwerpunkt auf der Unterstützung von Workflow mit Benutzerbeteiligung), Integration Suites (IT-Integrationsplattformen) und DCM Suites (Werkzeuge zur Unterstützung des Case Management) angeführt. Die Untersuchung zeigt, dass bisher nur wenige Werkzeuge alle drei Kategorien umfassend unterstützen.

Eine weitere Untersuchung [KSC11, S. 18] zeigt, dass im GPM die Zielstellungen, die mit einer verstärkten Einbindung der Mitarbeiter in die Aktivitäten des GPM in Verbindung stehen (positive Veränderung der Unternehmenskultur, höhere Mitarbeitermotivation), nur selten (jeweils bei ca. 25% der Umfrageteilnehmer) erreicht werden. Da die einfache Einbindung von Personen bzw. die Vernetzung von Gruppen gerade ein wesentliches Merkmal sozialer Software ist, kann der Einsatz sozialer Software für das GPM dazu beitragen, diese Ziele in größerem Umfang zu erreichen.

Die bereits erwähnte Untersuchung [Wa14, S. 39f] zeigt, das der Einsatz bestimmter Funktionen aus dem Bereich der sozialen Software in allen oder zumindest in einigen Phasen des GPM-Lebenszyklus als sinnvoll angesehen wird. Dazu gehören die Funktionen, die wie oben beschrieben bereits in vielen GPM-Werkzeugen vorhanden sind, also Portale, Kommentar- und Bewertungsfunktionen, asynchrone Kommunikationsfunktionen und Aufgabenverwaltung. Weiterhin werden Persönliche Profile/Soziale Beziehungen, Wikis, Activity Stream / Activity Log sowie Tagging / Bookmarking als sinnvoll angesehen. Der Einsatz von Blogs oder Foren für das GPM wird dagegen eher

als nicht sinnvoll angesehen. Beim Vergleich zwischen den von den Anbietern unterstützten Funktionen und den von den Anwendern als sinnvoll angesehenen Einsatzmöglichkeiten fällt auf, dass Wikis zwar nur von wenigen Werkzeugen unterstützt werden, aber auf Seiten der Anwender durchaus Interesse am Einsatz von Wikis für das GPM, insbesondere für die Prozessmodellierung, vorhanden ist.

Weiterhin zeigt die Untersuchung [Wa14, S. 42f], dass ein großer Teil der Unternehmen bereits über Erfahrungen mit dem Einsatz von sozialer Software verfügt und daran interessiert ist, diese auch für das Social BPM einsetzen. Die Untersuchung zeigt aber auch [Wa14, S. 43], dass ein Teil der Unternehmen in der Vergangenheit negative Erfahrungen mit sozialer Software gemacht hat und daher zögert, soziale Software für das GPM einzusetzen.

Dass der Einsatz sozialer Software an sich noch nicht unbedingt zur Beseitigung aller wesentlichen Probleme im GPM führt, zeigt [HW14, S. 17]. In dieser Studie wurde gefragt, in welchem Umfang die verfügbaren IT-Systeme die von der Organisation definierten Prozesse unterstützen. Dabei antworten nur 14% der Befragten mit „immer“ oder „meistens“, 24% mit „häufig“ und 63% mit „gelegentlich“ oder „nie“, wobei sich diese Anteile im Zeitraum von 2005 bis 2013 nicht wesentlich verändert haben. Das bedeutet, dass auch bei der Umsetzung der definierten Sollprozesse im Unternehmen Defizite auf Seiten der verfügbaren Unternehmensanwendungen bestehen.

Insgesamt zeigen die vorliegenden Studien, dass ein Bedarf an der Nutzung von sozialer Software für das GPM vorliegt und dass die Anbieter von GPM-Werkzeugen bereits einige soziale Funktionen mit ihren Werkzeugen unterstützen.

4 Die Nutzung sozialer Software für das GPM

Das Fraunhofer IAO erstellt derzeit eine Marktstudie zu GPM-Werkzeugen im deutschsprachigen Raum mit dem Studienschwerpunkt Social BPM[1], an der zehn Anbieter von GPM-Lösungen beteiligt sind. Die Studie wird voraussichtlich im Sommer 2014 erscheinen, so dass zum Zeitpunkt des Workshops die vollständigen Ergebnisse vorliegen werden und zur Diskussion gestellt werden können. Derzeit wurde die Datenerhebung, die in Form von Interviews erfolgte, abgeschlossen, so dass in diesem Beitrag eine vorläufige Zusammenstellung der Ergebnisse vorgenommen werden kann.

4.1 Historie von Social BPM

Viele der befragten Anbieter nennen als Zeitpunkt für den Beginn einer intensiveren Beschäftigung mit dem Thema Social BPM in etwa das Jahr 2010. Als Auslöser für diese intensivere Beschäftigung werden genannt:

[1] http://www.swm.iao.fraunhofer.de/bpmt2014

- Kundennachfragen
- Gespräche mit Analysten
- die Entwicklung der sozialen Netzwerke, einschließlich deren Fähigkeiten zum Einsatz im Unternehmen
- die Verfügbarkeit von sozialen Funktionen in verbreiteten Plattformen anderer Anbieter, zu denen eine Integration der GPM-Werkzeuge besteht (Hier wurde Microsoft SharePoint mehrfach benannt.)
- der technische Fortschritt allgemein, mit einem Schwerpunkt auf der Erstellung von Web 2.0-Anwendungen

Alle Anbieter betonen aber auch, dass die Beschäftigung mit den Themen aus dem Bereich Social BPM für sie nicht neu ist, sondern eine Weiterentwicklung bereits länger bestehender Aktivitäten darstellt. Solche „Vorläuferaktivitäten" reichen teilweise bis in die Mitte der 90er Jahre zurück. Sie finden sich in mehreren Bereichen. Grundlegend ist dabei das Zusammenspiel aus einem Repository (der zentralen Komponente zur Speicherung der Prozessmodelle und sonstigen Daten), einem (Prozess-)Portal für den Zugriff der Benutzer auf die Modelle und den Publikationsfunktionen zur Verteilung der Modellinhalte an einen größeren Benutzerkreis. Diese Funktionen sind zwingend erforderlich, um die Zusammenarbeit größerer Benutzergruppen im GPM zu unterstützen. Früher bestand die Hauptaufgabe dieser Lösungen in der Verteilung von Informationen, die von einem kleinen Kreis der aktiven Prozessmodellierer erstellt und von den meisten Benutzern lediglich passiv zur Kenntnis genommen wurden. Der aktuelle Stand ist im Abschnitt 4.2 dargestellt.

Als weitere „Vorläuferaktivitäten" für das Social BPM werden Mitarbeiterprofile benannt, die sich aus den üblichen Benutzerprofilen in den GPM-Lösungen entwickelt haben sowie das Semantic Web, das genutzt werden kann um die Zusammenhänge zwischen Autoren und Informationsobjekten abzubilden.

4.2 Aktueller Stand der Umsetzung von Social BPM

Die Anbieter beschreiten verschiedene Wege, um soziale Funktionen in den GPM-Lösungen umzusetzen:

- Die Umsetzung der sozialen Funktionen im eigenen GPM-Werkzeug erfolgt durch sechs der zehn beteiligten Anbieter.
- Die Anbindung der GPM-Werkzeuge an andere Lösungen, die soziale Funktionen bieten, erfolgt durch vier Anbieter. Meist wird hier MS SharePoint genannt, aber auch soziale Netzwerke wie Yammer.
- Die Umsetzung sozialer Funktionen in eigenen, eigenständigen Lösungen (d.h. die nicht Bestandteil des GPM-Tools sind), erfolgt durch zwei Anbieter.

Einige soziale Funktionen sind bereits von mehreren an der Studie beteiligten Anbietern umgesetzt worden. Diese Funktionen werden im Folgenden kurz beschrieben.

Ein Prozessportal ist bei den meisten Anbietern bereits seit vielen Jahren verfügbar und wird heute von allen GPM-Lösungen unterstützt. Die Portale bieten heute auch Funktionen zur (nachvollziehbaren) Kommentierung von Modellen, unterstützen den Nachrichtenversand (meist in Form von Newslettern und mit der Möglichkeit für Benutzer, gezielt Inhalte zu abonnieren), sie bieten die Möglichkeit von Freigabe-workflows und unterstützen die Erfassung und Verwaltung von Verbesserungsvor-schlägen und Maßnahmenkatalogen. Insgesamt ist festzustellen, dass die Portale – dem Web 2.0-Ansatz folgend – eine direkte und aktive Beteiligung aller Benutzer erlauben, so dass der Informationsfluss in unterschiedlichen Richtungen erfolgen kann und nachvollziehbar gemacht wird.

Der Einsatz von Wikis für das Verfassen von textuellen Beschreibungen zu einzelnen Modellierungselementen (Akteure, Ressourcen, …) wird von fünf Anbietern unterstützt.

Praktisch alle Anbieter denken darüber nach, wie eine aktive Beteiligung größerer Benutzergruppen an der Modellierung erreicht werden kann. Die Erstellung hochwertiger Modelle (die von allen Anbietern nach wie vor als nötig angesehen wird) muss aber nach wie vor durch Modellierungsspezialisten erfolgen. Deshalb bieten fünf Anbieter vereinfachte Modellierungsumgebungen an, in denen Modelle meist in Listenform erfasst werden können, die später in „richtige“ Modelle umgewandelt und ausmodelliert werden. Ein Anbieter nimmt gerade an einem Forschungsprojekt teil, um die vereinfachte Erstellung von Prozessmodellen in Form von Wikis zu untersuchen. Eine wesentliche Herausforderung dabei besteht darin, aus den semi-strukturierten Inhalten eines Wikis formale Modelle zu erzeugen.

Vier Anbieter ermöglichen die Erfassung von Informationen zu bestimmten Themen (typischerweise zu einem Prozess, aber auch andere Kontexte sind denkbar) in Form von Blogs. Drei Anbieter erfassen Aktionen und Änderungen von Modellinhalten in Form von Activity Logs.

Weitere Funktionen sind bisher nur bei einzelnen Anbietern verfügbar bzw. in Planung. Dazu gehören

- die direkte Unterstützung der großen mobilen Plattformen (iOS, Android, Windows Phone) durch die GPM-Clients,
- die Unterstützung des Adaptive Case Managements (hier zögern einige Anbieter, weil sie den erst kürzlich veröffentlichten OMG-Standard CMMN[2] berücksichtigen wollen)
- die Unterstützung von Abstimmungen im GPM-Werkzeug (z.B. über Änderungsvorschläge an Prozessmodellen)
- die Übernahme von aus Spielen bekannten Konzepten („Gamification“) für das GPM (z.B. Ranglisten, Status von Personen).

[2] http://www.omg.org/spec/CMMN/

4.3 Unterstützung für das organisationsübergreifende GPM

Eine andere Art der Netzwerkbildung besteht in der Bildung von Plattformen für den organisationsübergreifenden Austausch von Modellinhalten. Beispielhaft genannt sei hier die Nationale Prozessbibliothek[3] (NPB), die den Austausch von Modellinhalten für Organisationen aus dem öffentlichen Bereich ermöglicht. Mehrere Anbieter von GPM-Werkzeugen haben bereits Adapter für die NPB entwickelt, mit denen Modellinhalte zwischen ihren Werkzeugen und der NPB ausgetauscht werden können. Da die NPB noch recht neu ist, liegen bisher keine Erfahrungen über das Nutzungsverhalten vor. Es ist aber zu vermuten, dass die oben beschriebenen sozialen Funktionen auf einer solchen Plattform einen erheblichen Beitrag zur Unterstützung des GPM leisten können.

Mit der NPB vergleichbare offene Plattformen zum unternehmensübergreifenden Geschäftsprozessmanagement gibt es in der freien Wirtschaft nach Aussage der Studienteilnehmer nicht und das ist auch nicht zu erwarten, da die Anwenderunternehmen ihre Prozessmodelle, die eine detaillierte Einsicht in die internen Belange des Unternehmens bieten, nicht der Öffentlichkeit zur Verfügung stellen wollen. Unternehmen, die einzelnen Partnern gezielten Zugriff auf einzelne Modelle oder Modellierungsbereiche geben wollen, können dafür die bei allen GPM-Werkzeugen verfügbaren Prozessportale nutzen und dort für die Partner Accounts anlegen.

4.4 Bewertung von Social BPM

Dass die an der Studie beteiligten Anbieter dem Social BPM durchweg eine hohe Bedeutung bescheinigen ist wenig überraschend, schließlich dürfte das einer der wesentlichen Gründe dafür sein dass sie sich an der Studie beteiligen.

Die Anbieter sehen in Bezug auf Social BPM sowohl Chancen als auch Risiken. Die Chancen lassen sich am besten unter dem Motto zusammenfassen, dass „alle Betroffenen zu Beteiligten“ gemacht werden sollen – ein Ziel, an dem die GPM-Anbieter zum Teil schon seit 20 Jahren arbeiten und dem sie mit Social BPM ein Stück näher kommen.

Die Risiken werden, wie beim Einsatz von sozialen Lösungen in anderen Bereichen auch, allgemein darin gesehen, dass durch die sozialen Funktionen ein Ablenkungseffekt entsteht und die Mitarbeiter von ihren „eigentlichen“ Aufgaben abgehalten werden. Alle GPM-Anbieter weisen darauf hin, dass der Umgang mit diesen Risiken entscheidend für die Akzeptanz der angebotenen Lösungen bei den Kunden ist. Als Konsequenz daraus werden soziale Funktionen von den Anbietern nur dann in den GPM-Lösungen implementiert, wenn sie einen klar erkennbaren Nutzen für das GPM haben.

4.5 Social BPM, Dynamik und Compliance

Viele der im Abschnitt 4.2 aufgeführten Funktionen des Social BPM zielen darauf ab, mehr Benutzer direkt in die Aktivitäten zum Geschäftsprozessmanagement einzubezie-

[3] http://www.prozessbibliothek.de/index.php/portal.html

hen. Die Prozessportale sorgen dafür, dass die aktuellen Inhalte jederzeit und überall verfügbar sind. Eine gezielte Nachrichtensteuerung sorgt dafür, dass Änderungen oder andere wichtige Ereignisse gezielt an die Benutzer verteilt werden, für die sie relevant sind. Die Oberflächen der GPM-Werkzeuge sind durchweg webbasiert und so gestaltet, dass die Einstiegshürden für neue Benutzer und auch für Gelegenheitsbenutzer so weit wie möglich abgesenkt werden. Ein Beispiel dafür sind die verschiedenen, meist listenbasierten Ansätze zur vereinfachten Modellierung. Weiterhin können alle Benutzer direkt im Werkzeug aktiv werden und Kommentare, Bewertungen oder anderes Feedback erfassen, das für die anderen Benutzer sichtbar ist. Auch mobile Plattformen werden zunehmend unterstützt.

In der Summe ergibt sich durch die Nutzung von Social BPM die Möglichkeit, die gesamten Aktivitäten der Modellerstellung, -analyse, -optimierung und -implementierung für alle Beteiligten in einer einheitlichen Prozessplattform zu unterstützen, den Informationsfluss und die erforderlichen Abstimmungsprozesse erheblich zu beschleunigen und damit einen Beitrag zu einer größeren Dynamik im GPM-Lebenszyklus zu leisten.

Social BPM kann auch einen Beitrag zur Verbesserung der Compliance im GPM leisten. Bisher fand ein Teil der für das GPM erforderlichen Aktivitäten außerhalb der GPM-Werkzeuge statt. Diese Aktivitäten konnten durch die GPM-Werkzeuge weder unterstützt noch gesteuert werden und die Nachvollziehbarkeit der Interaktionen ist ebenfalls nicht gegeben. Durch die Einbeziehung von sozialen Funktionen wie Kommentar- und Bewertungsfunktionen können diese Aktivitäten innerhalb der GPM-Werkzeuge erfolgen und in den Activity Logs aufgezeichnet werden. Damit können die genannten Defizite beseitigt werden.

Die direkte Einbindung weiterer Benutzergruppen ist auch für die Fachkräfte mit juristischem Spezialwissen möglich, die Compliance-Anforderungen formulieren und bewerten können, ob sie erfüllt sind. Auch diese Benutzergruppe kann bei der Modellierung mitwirken, bei Änderungen informiert werden etc., was ebenfalls zu einer Verbesserung der Compliance im Prozessmanagement führen kann. Dazu sind aber einige weitere Voraussetzungen zu erfüllen, unter anderem ist zu klären wie die Compliance-Aspekte in den Modellen abgebildet und wie spezifische Compliance-Aktivitäten unterstützt werden.

5 Zusammenfassung

Aufgrund der vorliegenden Untersuchungen und der (vorläufigen) Ergebnisse der Anbieterbefragungen im Rahmen der Marktstudie lässt sich sagen, dass der Einsatz von sozialer Software für das GPM bereits erfolgt, dass bei den GPM-Anwendern Interesse an einem weiteren Ausbau der sozialen Funktionen besteht und dass auch die GPM-Anbieter mehr und mehr soziale Funktionen in ihre Lösungen integrieren bzw. die Nutzung solcher Funktionen über Schnittstellen zu anderen Plattformen unterstützen.

Die Aussagen der Anbieter zum Grad der Umsetzung der verschiedenen sozialen Funktionen in den GPM-Lösungen decken sich weitgehend mit [Wa14, S. 37]: Prozessportale, Funktionen für Kommentare und Bewertungen sowie News und Abonnements sind praktisch zum Standard geworden. Ansätze zum Absenken der Einstiegshürden für neue und Gelegenheitsbenutzer werden entwickelt und sind zum Teil bereits verfügbar. Wikis und Blogs können für verschiedene Aufgabenstellungen im GPM eingesetzt werden.

Für das Social BPM ergeben sich im GPM zwei grundsätzlich verschiedene Einsatzbereiche. Der erste Bereich besteht darin, die Kommunikation über Modelle und die Bearbeitung der Modelle durch soziale Funktion so zu gestalten, dass ein größerer Teilnehmerkreis als bisher direkt mit den Modellen in Kontakt kommt. In diesem Bereich gibt es vielfältige Aktivitäten der Anbieter im deutschsprachigen Markt, es findet eine intensive Weiterentwicklung der GPM-Werkzeuge statt und die bestehenden Plattformen und Portallösungen werden um weitere soziale Funktionen ergänzt. Im Bereich der Modellierung tragen die geringeren Einstiegshürden und eine bessere Vernetzung dazu bei, Prozessmodelle schneller zu aktualisieren und gleichzeitig die Qualität der Modelle zu erhöhen. In diesem Bereich leistet Social BPM bereits einen wichtigen Beitrag zu einer größeren Dynamik im GPM.

Der zweite Bereich besteht in einer Unterstützung des Adaptive Case Management durch die GPM-Werkzeuge, wodurch eine bessere Einbindung wissensintensiver Prozesse in den GPM-Lebenszyklus erreicht werden soll. In diesem Bereich ist von Seiten der Anbieter ebenfalls Interesse vorhanden, es finden aber bisher kaum Aktivitäten zur Produktplanung oder gar Implementierung statt. Ein Grund dafür besteht in der Unsicherheit darüber, wie ein zukunftsfähiger Modellierungsansatz für wissensintensive Prozesse aussieht.

6 Ausblick

Die in Abschnitt 4 erwähnte Studie des Fraunhofer IAO mit dem Themenschwerpunkt Social BPM wird voraussichtlich im Sommer 2014 erscheinen, so dass die vollständigen Ergebnisse im Workshop vorgestellt und diskutiert werden können. Es ist zu erwarten, dass noch weitere relevante Schnittpunkte zwischen den Themengebieten Soziale Software und GPM festgestellt werden können.

Im Bereich der Prozessausführung würde ein erfolgreicher Ansatz für das Adaptive Case Management dazu führen, das wissensintensive Prozesse nicht im Detail ausmodelliert werden müssen, dennoch den bestehenden Anforderungen entsprechen und vollständig in den GPM-Lebenszyklus eingebunden werden können. Die Dynamik bei der Prozessausführung, d.h. die Möglichkeit der flexiblen Ausführung von Prozessen oder Prozessschritten, würde unter anderem auch den Änderungsbedarf bzw. die Änderungshäufigkeit wissensintensiver Prozesse senken. Interessant erscheint in diesem Zusammenhang der kürzlich veröffentlichte OMG-Standard Case Management Model and Notation (CMMN).

Literaturverzeichnis

[DLMR13] Dumas, M.; LaRosa, M.; Mendling, J.; Reijers, H.: Fundamentals of Business Process Management. Springer-Verlag Berlin Heidelberg 2013.

[HW14] Harmon, P.; Wolf, C.: The State of Business Process Management 2014 – A BPTrends Report.

[KSC11] Knuppertz, T.; Schnägelberger, S.; Clauberg, K.: Status Quo Prozessmanagement 2010/2011. BPM&O Akademie, Köln, 2011.

[Wa14] Walter, D.: Social BPM - Konzeption und Durchführung einer Marktanalyse. Bachelorarbeit an der Hochschule für Technik Stuttgart, 2014.

[Pä13] Pätz, M.: Business Process Management 2.0 – Impact, Opportunities, and Application Scenarios of Web 2.0 regarding Business Process Management Systems. Master Thesis an der Hochschule für Technik Stuttgart, 2013.

[RM13] Richardson, C.; Miers, D.: The Forrester Wave™: BPM Suites, Q1 2013.

Anbieterübergreifende Konfiguration komplexer Dienstleistungsportfolios: Ein Anwendungsfall im Bereich erneuerbarer Energien[1]

Michael Becker, Stephan Klingner, Michael Sonnenberg

Abteilung für Betriebliche Informationssysteme
Universität Leipzig
Hainstraße 11
04109 Leipzig
[mbecker,klingner,sonnenberg]@informatik.uni-leipzig.de

Abstract: Im Bereich erneuerbarer Energien existiert eine große Anzahl von Dienstleistungsanbietern. Dieser Fakt trägt zusammen mit der zunehmenden Komplexität von Dienstleistungen dazu bei, dass es für Kunden schwieriger wird, passende Angebote zu wählen. In diesem Beitrag wird ein Ansatz vorgestellt, mit dem Dienstleister ihr Portfolio anhand eines Standardportfolios der Domäne präsentieren können. Der damit einhergehende Anstieg der Transparenz ist sowohl für Kunden als auch für Anbieter von Vorteil.

1 Einleitung

Am Dienstleistungsmarkt stellen verschiedene Anbieter komplexe Portfolios bereit, mit denen sie Kunden ihre Dienstleistungen präsentieren. Dabei sind in vielen Branchen die Zusammensetzung und Beschreibung ähnlicher bzw. gleicher Dienstleistungen zwischen den Anbietern sehr heterogen. Abweichungen sind insbesondere in der Strukturierung und Terminologie zu finden. Durch die dadurch entstehende mangelnde Vergleichbarkeit verschiedener Angebote steigt einerseits das Kaufrisiko der Kunden. Andererseits bedeutet dies für den Anbieter unter Umständen eine geringere Kundenzufriedenheit, da eine Lücke im Verständnis zwischen dem, was der Kunde von der Dienstleistung erwartet und dem, was der Anbieter leistet, entstehen kann [ZBP93]. Eine vereinheitlichte Methode zur Gestaltung von Portfolios zum optimierten Suchen und Finden von Dienstleistungen kann entsprechend zu einer Erhöhung der Markttransparenz beitragen und für Kunden und Anbieter erhebliche Mehrwerte generieren [Moe07].

Zielsetzung des Forschungsprojektes EUMONIS ist die Entwicklung einer Software- und Systemplattform für Energie- und Umweltmonitoringsysteme zur Erbringung konkreter industrieller Dienstleistungen im Bereich erneuerbare Energien (EE). Diese Dienstleistungen werden durch Softwareservices repräsentiert. Mittels eines Standard-

[1] Dieser Beitrag wurde ermöglicht durch die Förderung des Projekts „EUMONIS“ (www.eumonis.org) mit Mitteln des Bundesministeriums für Bildung und Forschung (BMBF). Das Projekt (FKZ 01IS10033D und 01IS10033K) wird betreut vom Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (PT-DLR).

portfolios werden Dienstleistungen abgebildet, die über die Plattform angeboten und erbracht werden sollen. Anbieterportfolios sind individuelle Angebote verschiedener Anbieter, die ihre Dienstleistungen über die Plattform für Betreiber von EE-Anlagen bereitstellen wollen. Dazu wurden im Rahmen eines konstruktionsorientierten Forschungsprogramms verschiedene Software-Werkzeuge anhand von Anforderungen aus der Praxis entwickelt und evaluiert[2].

Für die EE-Domäne wurde in den letzten Jahren eine Vielzahl unterschiedlicher Standards entwickelt. Beispielsweise zählt alleine [IRE13] 573 Standards in den verschiedenen Technologiebereichen (Wind, Solar etc.) und Kategorien (Installation, Datenverarbeitung etc.) auf. Im Gegensatz zu Produktstandards gibt es für EE-Dienstleistungen bisher nur wenige Vorarbeiten. Dabei lassen sich Klassifikationsansätze vor allem dahingehend unterscheiden, ob sie Bestandteil eines größeren Klassifikationsrahmens sind (z.B. als Teil der GATS-Klassifikation [US13]) oder die EE-Spezifika adressieren, z.B. [HW09].

Es gibt verschiedene Branchen, in denen die existierenden Konzepte zur strukturierten Entwicklung, Erbringung und Vermarktung von Dienstleistungen nur unzureichend genutzt werden [SAB13]. Grundlage des vorliegenden Beitrags ist die Bereitstellung einer Methodik zur strukturierten Beschreibung von Portfolios verschiedener Anbieter, welche jedoch auch die Modellierung und Konfiguration komplexer Dienstleistungen unterstützt (Abschnitt 2). Es wird die Notwendigkeit eines Standardportfolios propagiert, anhand dessen die verschiedenen Anbieter Portfolios einheitlich und strukturiert modellieren und zur kundenindividuellen Konfiguration bereitstellen können (Abschnitt 3).

Die Vereinheitlichung erfordert ein schrittweises Vorgehen. In einem ersten Schritt werden Referenzkomponenten für einen bestimmten Dienstleistungsbereich in Form eines Standardportfolios modelliert. Hierfür bedarf es einheitlicher Vorlagen, die dem allgemeinen Branchenverständnis entsprechen, wie beispielsweise die DIN-Klassifikation von EE-Dienstleistungen [DIN14]. Im zweiten Schritt werden die Portfolios der Anbieter entsprechend ihrer internen Beschreibung modelliert. Ein Mapping zwischen Standard- und Anbieterportfolios erfolgt in einem letzten Schritt und führt zu einer vereinheitlichten Dienstleistungsbeschreibung (Abschnitt 4). Das Standardportfolio wird anschließend dem Kunden zur Konfiguration seiner gewünschten Dienstleistung bereitgestellt und ermöglicht so eine über die Grenzen eines Anbieters hinweg transparente Angebotserstellung (Abschnitt 5). Der Beitrag wird mit einer Diskussion weiterführender Herausforderungen beim Mapping und der Konfiguration abgeschlossen (Abschnitt 6).

2 Modellierung von Dienstleistungskomponenten

Grundlage der Modellierung von Anbieter- und Standardportfolio ist der komponentenbasierte Ansatz zur Modellierung von Dienstleistungen von [BBK11]. Ziel dieses Ansat-

[2] Tool zum Mapping von Anbieter- zu Standardportfolio: http://pcai042.informatik.uni-leipzig.de/~swp14-smd/ Werkzeug zur Konfiguration des Standardportfolios: http://pcai042.informatik.uni-leipzig.de/~swp14-kde/

zes ist es, eine leistungsfähige Methode zur Modellierung und kundenindividuellen Konfiguration komplexer Dienstleistungen bereitzustellen.

Der Ansatz ist formal durch ein Metamodell beschrieben. Dem Metamodell entsprechend werden monolithische Dienstleistungen in Komponenten, die abgegrenzte Funktionalitäten repräsentieren, unterteilt. Vorteile der Unterteilung in logisch abgegrenzte Dienstleistungskomponenten sind u.a. eine Reduktion der Komplexität oder eine einfachere Wiederverwendung sowie die Erstellung kundenindividueller Dienstleistungen auf Basis standardisierter Komponenten.

Die Dienstleistungskomponenten werden in einem hierarchischen Baum angeordnet und sind über Konnektoren verbunden. Konnektoren definieren dabei die Kardinalität, welche die Zusammensetzung von Komponenten in Konfigurationen quantitativ beschreiben. Auf diese Weise kann beispielsweise dargestellt werden, dass alle, keine oder genau eine Kindkomponente bei Auswahl der Elternkomponente zu wählen ist. Eine beispielhafte graphische Darstellung eines hierarchisch aufgebauten Portfolios sowie eine Übersicht verschiedener Kardinalitäten findet sich in Abbildung 1.

Neben der Darstellung solcher hierarchischer Zusammenhänge können mittels logischer Abhängigkeiten auch hierarchieübergreifende Zusammenhänge abgebildet werden (wenn Komponente A gewählt ist, dann muss auch Komponente B gewählt werden). Darüber hinaus ist zur Ableitung von Erbringungsprozessen auch die Definition zeitlicher Abhängigkeiten zwischen Komponenten möglich (z.B. Komponente A muss vor Komponente B erbracht werden).

Weitere auf Ebene der Komponenten zu definierende Elemente sind Key Performance Indicators zur Abbildung von Produktivitätseigenschaften von Komponenten, Attribute zur Darstellung nichtfunktionaler Eigenschaften sowie Variablen zur Berücksichtigung von Umgebungsfaktoren bei der Erbringung der Dienstleistungen. Auf Basis der beschriebenen Elemente ist es möglich, Dienstleistungsmodelle zu erstellen, zu konfigurieren und diese Konfigurationen hinsichtlich Validität und Produktivität zu bewerten.

3 Anwendungsfall

Der im Folgenden beschriebene Anwendungsfall steht exemplarisch für die Verwendung in einem mittelständischen Unternehmen und zeigt eine Möglichkeit, sich der Problematik des Mappings beider Portfolios zu nähern. Das hier verwendete Standardportfolio ist aus Gründen der Übersichtlichkeit fiktiv; das Portfolio der Plattform lehnt sich an existierende Klassifikationen wie z.B. [DIN14] und [SSK12] an.

Das Unternehmen bietet verschiedene Dienstleistungen in den Domänen Wind und Solar an und beschreibt diese mit Hilfe strukturierter Prozesse. Aus der Sammlung von Prozessen wurde das Portfolio in Abbildung 1 für den Bereich Wind erstellt. Dieses ist zunächst in die grundlegenden Komponenten Ersatzteilmanagement, Betriebsführung, Fernüberwachung, Anlagenmanagement und Störungsmanagement unterteilt. Aus

Gründen der Übersichtlichkeit wird an dieser Stelle nur der Bereich Anlagenmanagement genauer diskutiert und in weitere Unterkomponenten unterteilt.

Das Portfolio in Abbildung 1 ist hinsichtlich einer kundenspezifischen Konfiguration wie folgt aufgebaut. Wird das Anlagenmanagement gewählt, muss die Kritikalitätsprüfung durchgeführt werden. Darüber hinaus müssen die Kunden sich entscheiden, ob sie ihre Anlage selber mit Ersatzteilen bevorraten. In diesem Fall ist die Konzeptualisierung mit den entsprechenden Unterkomponenten zu wählen. Soll die Bevorratung durch den Dienstleister erfolgen, sind die entsprechenden Vertragsdetails auszuhandeln.

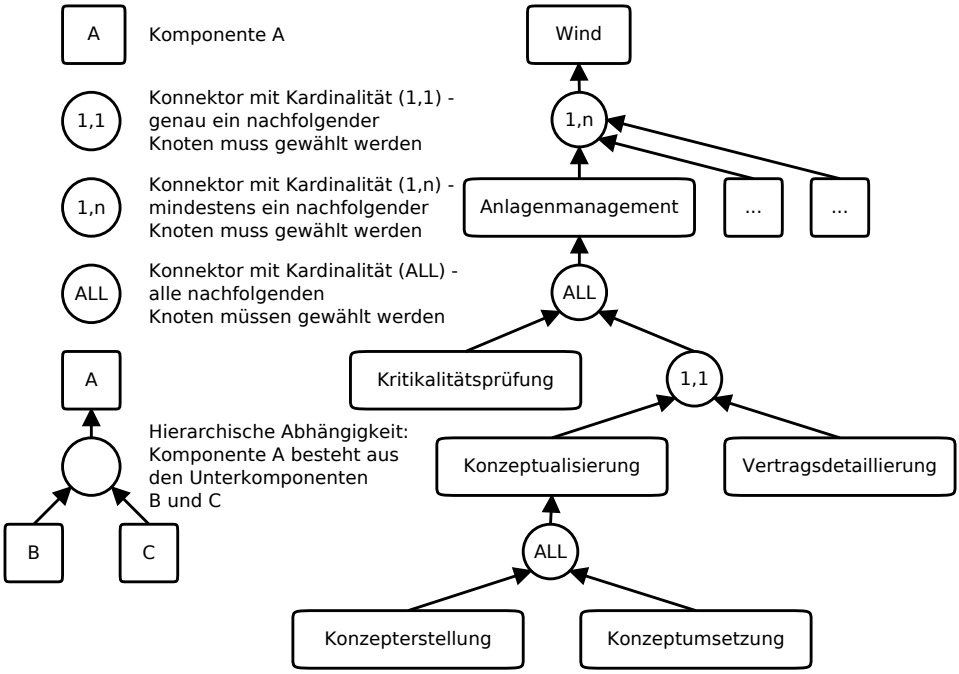


Abbildung 1: Anbieterspezifisches Portfolio

Anhand dieses Ausschnitts des Unternehmensportfolios lassen sich die oben angesprochenen Herausforderungen bei der Konfiguration von Dienstleistungen relativ einfach ablesen. Es ist nicht zu erwarten, dass andere Unternehmen ihr Portfolio in gleicher Art und Weise strukturiert haben. Darüber hinaus wird auch die Benennung der Komponenten nicht identisch sein. Aus diesem Grund ist es für Kunden nur mit relativ hohem Aufwand möglich, die Angebote verschiedener Anbieter miteinander zu vergleichen.

Diesen Herausforderungen soll mit dem in Abbildung 2 dargestellten Standardportfolio entgegen getreten werden. Ein Standardportfolio stellt eine allgemeingültige Klassifikation von Dienstleistungen für einen dedizierten Bereich bereit. Auch hier ist wiederum nur ein Ausschnitt skizziert. Im Standardportfolio ist das Anlagenmanagement in die beiden Bereiche Errichtung und Nutzung unterteilt. Die Errichtung, welche hier betrachtet werden soll, gliedert sich darüber hinaus u.a. in die Komponenten Planung Probe,

Koordination Beteiligte, Prüfung Gutachten, Planung Abnahme und Vorbereitung Garantie.

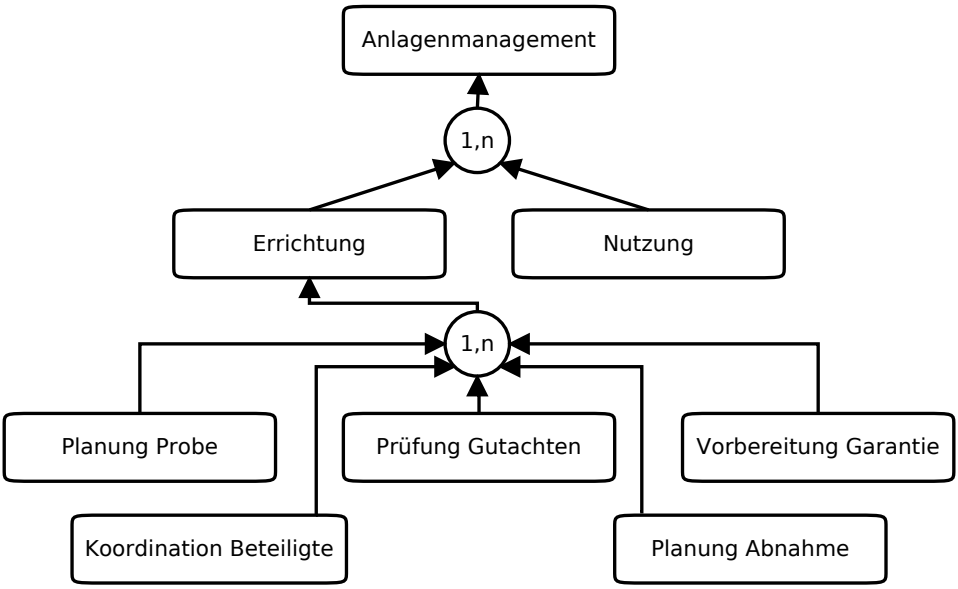


Abbildung 2: Standardportfolio

4 Mapping zwischen Anbieter- und Standardportfolio

Die beiden Portfolios in Abbildung 1 und Abbildung 2 zeigen sowohl strukturelle als auch terminologische Unterschiede. Die strukturellen Unterschiede liegen hierbei in zwei Ausprägungen vor: Einerseits ist in beiden Portfolios eine unterschiedliche Anzahl an Komponenten vorhanden. Dies lässt auf einen unterschiedlichen Abstrahierungsgrad bei der Modellierung der Portfolios schließen. Darüber hinaus sind die Komponenten auf verschiedene Art und Weise unterteilt. Der Gedanke hinter dem Standardportfolio ist, keine allzu detaillierten Vorgaben an die Konfiguration zu stellen, um möglichst generisch einsetzbar zu sein. Im Gegensatz dazu ist das Anbieterportfolio auf die spezifischen Gegebenheiten des Anbieters angepasst und dementsprechend strukturiert.

Um Kunden die Möglichkeit zu geben, Dienstleistungen aus einer größeren Anzahl von Anbietern auszuwählen, soll in diesem Abschnitt vorgestellt werden, mit welchen Methoden Komponenten des anbieterspezifischen Portfolios auf das Standardportfolio abgebildet werden können. Basierend auf dem damit erzeugten Mapping können Kunden das Standardportfolio wie im nächsten Abschnitt beschrieben anhand ihrer individuellen Anforderungen konfigurieren.

4.1 Zuordnungstypen

In der Praxis hat sich gezeigt, dass in der Regel alle oben angesprochenen terminologischen und strukturellen Abweichungen auftreten und durch das Mapping aufzulösen sind. Terminologische Unterschiede lassen sich relativ schnell beseitigen, da das Standardportfolio die Terminologie vorgibt. Hier ist allerdings zu beachten, dass ggf. nicht alle Begrifflichkeiten der Domäne in ihrer Semantik exakt definiert sind. Beispielsweise kann das Verständnis über die Dienstleistung Wartung je nach Anbieter sehr unterschiedlich ausfallen. Aus diesem Grund ist neben der reinen Benennung einzelner Komponenten weiterhin eine detaillierte Angabe dahingehend notwendig, welche Funktionalität die jeweilige Komponente darstellt. Dies ist als zusätzliche Information zu verstehen, um den Anbietern die korrekte Zuordnung zu ermöglichen. Darüber hinaus können die Beschreibungen der Komponenten den Kunden bei der Konfiguration präsentiert werden, womit auch für diese die Transparenz über den Inhalt der Dienstleistung steigt.

Abweichungen, die aus der unterschiedlichen Anzahl und Struktur der Komponenten resultieren, werden mit Hilfe der im Folgenden beschriebenen Zuordnungsarten aufgelöst. Jede dieser Zuordnungsarten verbindet eine oder mehrere Komponenten des anbieterspezifischen Portfolios mit einer oder mehreren Komponenten des Standardportfolios. Die verschiedenen Arten der Zuordnung lassen sich in zwei Dimensionen unterteilen: der Anzahl der an der Zuordnung jeweils beteiligten Komponenten sowie dem Zuordnungsziel in der Baumstruktur des Portfolios. Basierend auf der jeweiligen Art der Zuordnung ergeben sich verschiedene Besonderheiten hinsichtlich der Konfiguration.

Die verschiedenen Zuordnungsarten sind in Tabelle 1 anhand eines Beispiels dargestellt und werden im Folgenden genauer vorgestellt. Dabei steht die Bezeichnung A_n für eine Komponente des anbieterspezifischen Portfolios; die Bezeichnung S_n bezeichnet eine Komponente des Standardportfolios. Ein gerichteter Pfeil von links nach rechts stellt eine Zuordnung einer Anbieter- zu einer Standardkomponente dar.

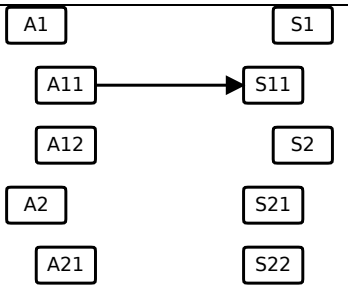
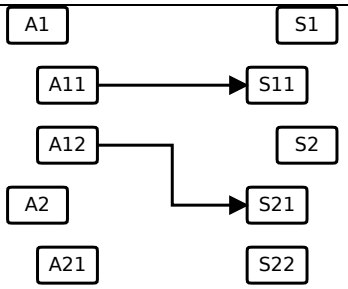
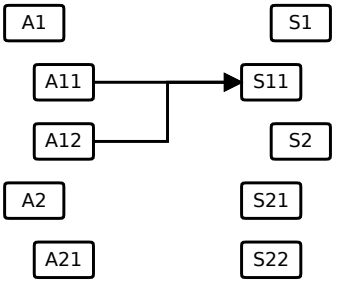
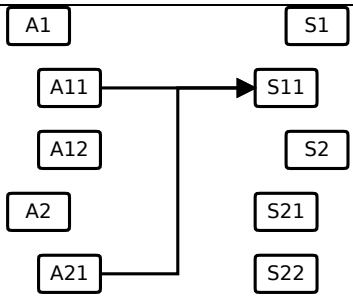
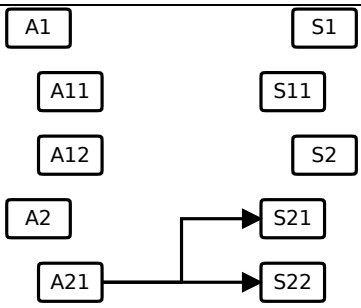
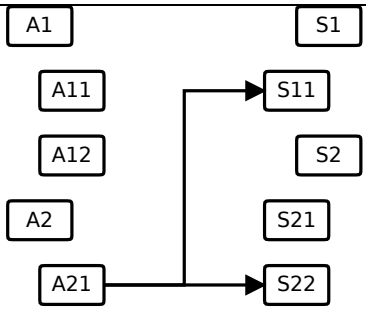
Im einfachsten Fall kann eine Eins-zu-Eins-Zuordnung vorgenommen werden, d.h. eine anbieterspezifische Komponente wird auf genau eine Komponente des Standardportfolios abgebildet. Diese Zuordnung kann dann durchgeführt werden, wenn die Standardkomponente die gleiche Funktionalität wie eine entsprechende Komponente des Anbieters bereitstellt.

Bei der Zusammenfassung werden mehrere Komponenten des Anbieters auf genau eine Komponente des Standardportfolios abgebildet. Dieser Fall kann eintreten, wenn der Anbieter intern eine Komponente des Standardportfolios noch weiter unterteilt hat. Das heißt, die jeweilige Standardkomponente kapselt die Funktionalität verschiedener Komponenten des Anbieters. Im umgekehrten Fall, d.h. wenn genau eine anbieterspezifische Komponente auf mehrere Komponenten des Standardportfolios abgebildet wird, liegt eine Erweiterung vor. Hierbei ist also das Portfolio des Anbieters weniger detailliert.

Die drei Typen von Zuordnungen lassen sich weiterhin dahingehend unterscheiden, ob sie eine hierarchieinterne oder hierarchieexterne Zuordnung darstellen. Bei der internen Zuordnung liegen die Zuordnungsziele jeweils nur innerhalb einer Hierarchieebene. Im

Gegensatz dazu werden die anbieterspezifischen Komponenten bei der externen Zuordnung in unterschiedlichen Hierarchieebenen neu strukturiert.

Tabelle 1: Zuordnungstypen nach Anzahl und Zuordnungsziel

Anzahl\ Zuordnungs-ziel	Hierarchieintern	Hierarchieextern
Eins-zu-Eins-Zuordnung		
Zusammen-fassung		
Erweiterung		

4.2 Mapping des Anwendungsfalls

Mit Hilfe der verschiedenen Zuordnungstypen können die Komponenten des Anwendungsfalls wie in Abbildung 3 dargestellt abgebildet werden. Das Portfolio des Anbieters findet sich auf der linken und das Standardportfolio auf der rechten Seite der Abbildung. Da die hierarchischen Beziehungen beim Mapping keine Rolle spielen, sind in der Ab-

bildung nur die entsprechenden Komponenten vorhanden und keine Konnektoren. Es zeigt sich, dass die Zuordnung zunächst zu einer Auflösung terminologischer Unklarheiten beitragen kann (z.B. die Abbildung der Komponente Vertragsdetaillierung auf die Komponente Vorbereitung Garantie). Darüber hinaus konnte ein höherer Detaillierungsgrad erreicht werden, da die Komponente Konzepterstellung auf die Komponenten Planung Probe, Prüfung Gutachten und Planung Abnahme abgebildet wurde.

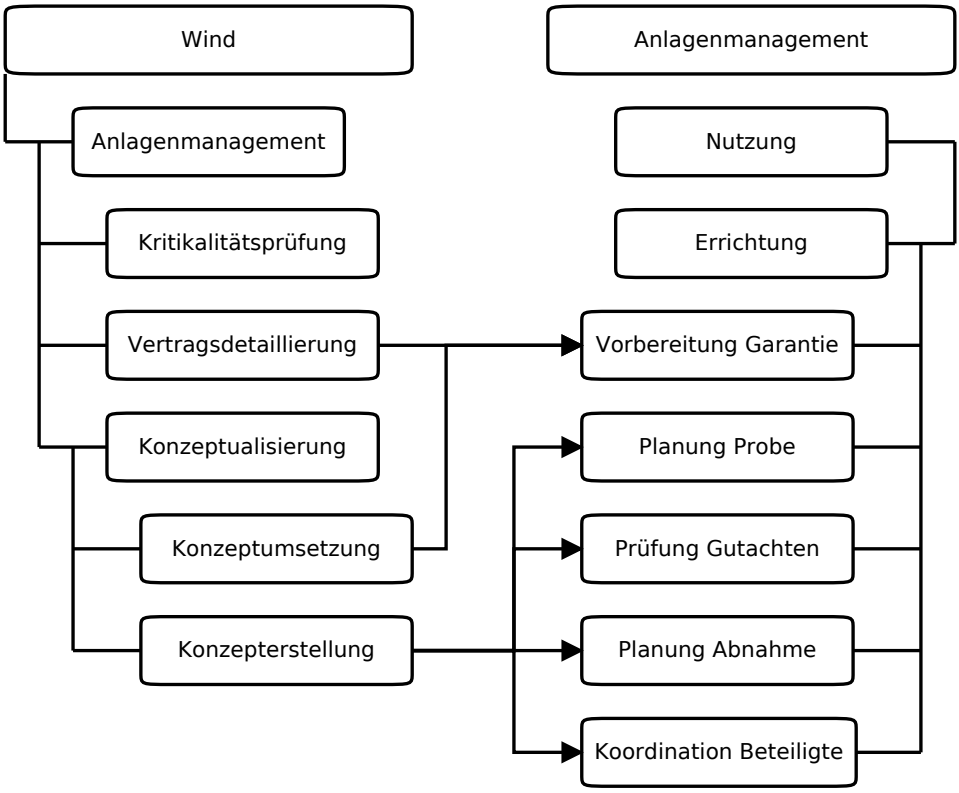


Abbildung 3: Mapping des anbieterspezifischen Portfolios auf das Standardportfolio

In der Abbildung zeigt sich allerdings noch ein weiterer Effekt, der zu beachten ist, wenn das Standardportfolio genutzt wird, um Dienstleistungen zu verkaufen. Die Komponente Kritikalitätsprüfung des Anbieters hat keine Entsprechung im Standardportfolio. Darüber hinaus sind auch im Standardportfolio Komponenten vorhanden, die nicht mit einer Komponente des anbieterspezifischen Portfolios verbunden sind. Dies liegt darin begründet, dass der Anbieter die entsprechende Funktionalität nicht bereitstellt.

Die beiden Zuordnungstypen, die im Mapping vorhanden sind, sind einerseits eine hierarchieinterne Erweiterung: die Konzepterstellung wird auf die Komponenten Planung Probe, Prüfung Gutachten, Planung Abnahme und Koordination Beteiligte abgebildet. Darüber hinaus ist mit der Abbildung der anbieterspezifischen Komponenten Vertrags-

detaillierung und Konzeptumsetzung auf die Standardkomponente Vorbereitung Garantie eine hierarchieexterne Zusammenfassung vorhanden.

5 Konfiguration eines Standardportfolios und Anbieterauswahl

Basierend auf dem im vorherigen Abschnitt entwickelten Mapping zwischen Komponenten eines anbieterspezifischen Portfolios sowie den Standardkomponenten, können Kunden nun Dienstleistungen hinsichtlich ihrer individuellen Anforderungen konfigurieren. Unter der Voraussetzung, dass verschiedene Anbieter ihre Portfolios dem Standardportfolio zugeordnet haben, können Kunden aus einer Reihe von Anbietern wählen. Bei der Konfiguration sind einige Besonderheiten zu beachten, die exemplarisch an dem Mapping des Anwendungsfalls vorgestellt werden. Grundsätzlich ist zu beachten, dass zur Konfiguration eine Reihe verschiedener Mappings vorliegen, so dass das Standardportfolio auf verschiedene Anbieter abgebildet wird. Anhand der Konfiguration durch den Kunden werden Anbieter präsentiert, welche die geforderten Funktionalitäten bereitstellen. Mit Hilfe von Attributen wie z.B. Preis oder Ausführungszeiten lassen sich verschiedene Angebote hinsichtlich nichtfunktionaler Kriterien miteinander vergleichen.

Die Konfiguration besteht dementsprechend aus drei Schritten. Zunächst wählen Kunden die Komponenten des Standardportfolios aus, mit denen sie ihre Anforderungen umsetzen können. Anhand dieser Auswahl und basierend auf den Mappings werden Anbieter ermittelt, welche die entsprechende Funktionalität bereitstellen. Diese Anbieter werden den Kunden in einer vergleichenden Übersicht dargestellt. Aus dieser Übersicht wählen die Kunden dann den oder die Anbieter, die das für sie beste Angebot bereitstellen; die weitere Kommunikation und eventuell notwendige Anpassung von Dienstleistungen ist nicht mehr im Fokus dieser Arbeit.

5.1 Auswahl von Komponenten des Standardportfolios

Aufgrund der hierarchischen Baumstruktur zur Darstellung von Dienstleistungen können Kunden eine komplexe Dienstleistung entweder top-down oder bottom-up konfigurieren. Bei der top-down-Konfiguration werden zunächst grobgranulare Komponenten ausgewählt und sukzessive verfeinert. Die top-down-Konfiguration kann genutzt werden, um eine Konfiguration ohne Kenntnis aller Details des Standardportfolios zu erstellen. Dieses enthält es keine Regeln bezüglich der minimalen oder maximalen Anzahl auszuwählender Unterkomponenten einer Komponente, weshalb keine nachfolgenden Komponenten automatisch selektiert werden.

Bei der bottom-up-Konfiguration suchen Kunden gezielt nach spezifischen Funktionalitäten innerhalb des Standardportfolios. Bei der Auswahl einer Komponente werden die entsprechenden Oberkomponenten automatisch mit ausgewählt, so dass stets eine vollständige Konfiguration sichergestellt ist. Dadurch können Kunden mit detaillierten Kenntnissen über die vorhandenen Komponenten des Standardportfolios relativ schnell und effizient die von ihnen benötigten Funktionalitäten konfigurieren.

5.2 Ermittlung und Vergleich von Anbietern

Basierend auf der im vorhergehenden Schritt erstellten Konfiguration werden nachfolgend Anbieter identifiziert, welche die gewünschte Funktionalität der Kunden erbringen können. Dazu werden die vorhandenen Mappings der verschiedenen Anbieter verarbeitet und mit den gewählten Komponenten des Standardportfolios verglichen. Grundsätzlich können dabei drei Szenarien eintreffen: ein Anbieter kann die Kundenanforderungen entweder vollständig, teilweise oder gar nicht erfüllen. Im Folgenden wird nur der Fall betrachtet, dass ein Anbieter die Anforderungen vollständig erfüllen kann. Die teilweise Erfüllung ist Gegenstand der Diskussion im nachfolgenden Abschnitt.

Anbieter, welche die Funktionalität vollständig erbringen können, haben alle gewählten Komponenten des Standardportfolios auf Komponenten ihres eigenen Portfolios abgebildet. Um einen detaillierten Vergleich zwischen diesen Anbietern zu ermöglichen, müssen deren Portfolios noch auf implizite logische oder hierarchische Abhängigkeiten hin untersucht werden, die dazu führen, dass weitere Komponenten ausgewählt werden müssen. Dieser Schritt ist in Abbildung 4 dargestellt.

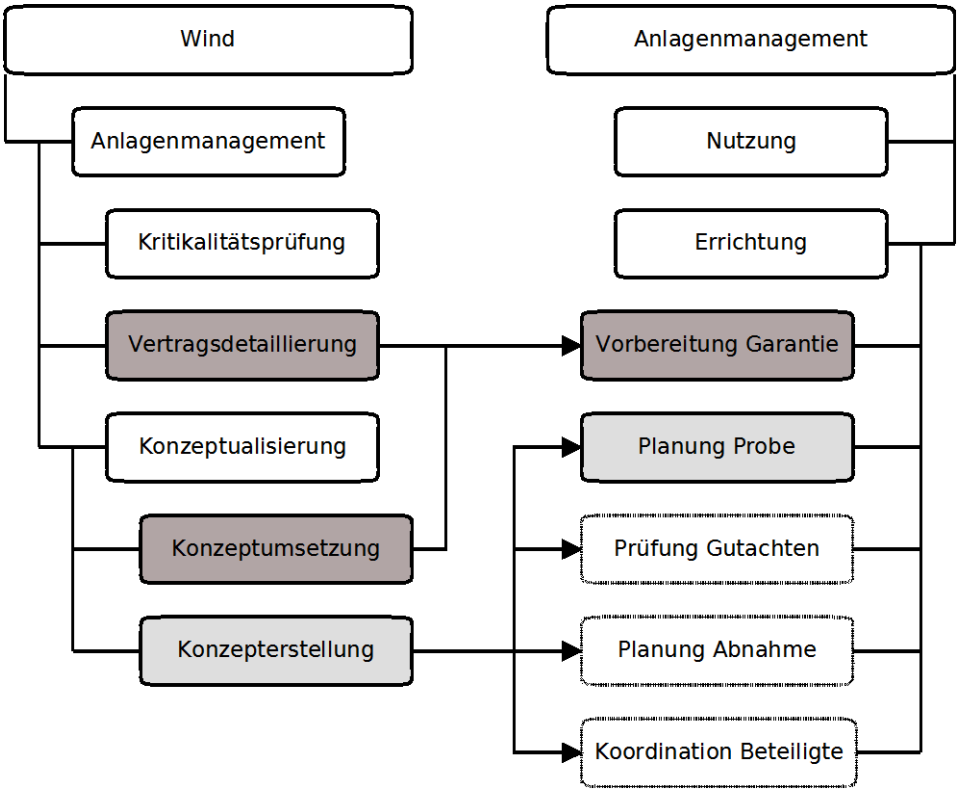


Abbildung 4: Konfiguration und implizite Abhängigkeiten durch das Mapping

In der Konfiguration aus Abbildung 4 hat der Kunde die beiden markierten Komponenten Vorbereitung Garantie und Planung Probe gewählt. Beide sind im Mapping des Anbieters enthalten, so dass dieser die Anforderungen vollständig erfüllen kann. Die Komponenten, mit denen die geforderten Funktionalitäten umgesetzt werden, sind die Vertragsdetaillierung und Konzeptumsetzung für die Komponente Vorbereitung Garantie sowie die Komponente Konzepterstellung für Planung Probe.

Die anbieterspezifische Komponente Konzepterstellung ist auch auf die in der Abbildung gestrichelt dargestellten Komponenten Prüfung Gutachten, Planung Abnahme und Koordination Beteiligte des Standardportfolios abgebildet. Diese sind dementsprechend in der Konfiguration implizit mit enthalten, wenn der Referenzanbieter als ausführender Dienstleister gewählt wird.

Nach Beendigung der Konfiguration werden dem Kunden die verschiedenen Anbieter präsentiert, welche die geforderte Funktionalität umsetzen können. Dabei kann basierend auf den vorhandenen Attributen der Portfolios ein Vergleich beispielsweise hinsichtlich des Preises oder der Verfügbarkeit erfolgen. Insbesondere durch die beiden Zuordnungstypen Zusammenfassung und Erweiterung können Unterschiede zwischen den Anbietern auftreten, die sich aus verschieden strukturierten Portfolios ergeben.

Im Beispiel aus Abbildung 4 ist ersichtlich, dass Kunden bei der Auswahl der Komponente Vorbereitung Garantie zwei Komponenten des Anbieters in ihrer Auswahl haben. Darüber hinaus stellt dieser Anbieter die Komponente Planung Probe nur gemeinsam mit den anderen Komponenten Prüfung Gutachten, Planung Abnahme und Koordination Beteiligte an. Dies kann bei anderen Anbietern, die ihr Portfolio anders strukturiert haben, abweichend gelöst sein.

6 Fazit

Kern dieser Arbeit ist ein Ansatz zur Konfiguration von Dienstleistungen basierend auf einem vordefinierten Standardportfolio. Anhand eines Anwendungsfalls aus dem Bereich erneuerbarer Energien wurde gezeigt, wie anbieterspezifische Portfolios auf das Standardportfolio abgebildet werden können. Mit Hilfe dieser Abbildung lassen sich basierend auf der Konfiguration des Standardportfolios die Angebote verschiedener Anbieter miteinander vergleichen.

Die Ergebnisse dieser Arbeit sind sowohl für Kunden als auch für Anbieter von Dienstleistungen nutzbar. Kunden werden in die Lage versetzt, den für sie passenden Dienstleister auszuwählen bzw. verschiedene Dienstleister zu kontaktieren, die in Kombination die Anforderungen erfüllen können. Anbieter haben darüber hinaus eine einfache Möglichkeit, ihre Dienstleistung in standardisierter Form zu präsentieren. Dadurch vergrößert sich der Kreis potentieller Kunden.

Anbieter haben grundsätzlich zwei Möglichkeiten, wie sie auf Abweichungen ihres Portfolios vom Standardportfolio reagieren können. Einerseits ist es möglich, dass sie ihr Portfolio entsprechend anpassen. Dies ist im Allgemeinen für Unternehmen relevant, die

kein stark strukturiertes Portfolio mit diversen Abhängigkeiten zwischen Komponenten besitzen. Durch die Anpassung des Portfolios an das Standardportfolio lassen sich eine Vielzahl von Abbildungen durch Eins-zu-Eins-Zuordnungen vornehmen.

Im Gegensatz dazu haben Unternehmen mit einem umfangreichen Portfolio, welches in der Regel komplexe Abhängigkeiten der Komponenten miteinander enthält, nicht die Möglichkeit dieses Portfolio anzupassen. Diese können dann die Zuordnungstypen Zusammenfassung und Erweiterung nutzen. Es wurde bereits angesprochen, dass nicht alle Anbieter auf alle Komponenten des Standardportfolios abbilden können, da sie unterschiedliche Funktionalitäten erbringen. Dadurch gibt es die Möglichkeit, dass ein Anbieter die gewählten Komponenten einer Konfiguration nur teilweise oder überhaupt nicht bereitstellt. Insbesondere die teilweise Erbringung ist für die weitere Analyse von Interesse. Hier ist ein Kompromiss zu finden zwischen dem Ausschluss von Anbietern und der passgenauen Präsentation für Kunden.

Zu beachten hierbei ist jedoch, dass durch die Plattform die Erbringung industrieller Dienstleistungen erfolgt, die durch das Standardportfolio beschrieben werden. Entsprechend gelten für Anbieter diese Vorgaben, an die sie beim Mapping gebunden sind. Gerade bei automatisiert erbrachten Dienstleistungen ist die strikte Definition konkreter Funktionalitäten von Bedeutung, nicht nur aus Sicht der Präsentation für den Kunden sondern vor allem aus Gründen einer technischen Realisierbarkeit. So müssen Anbieter ggf. einzelne Komponenten zu neuen Dienstleistungen zusammenfassen, um die angebotenen Funktionalitäten zu erfüllen.

Literaturverzeichnis

[BBK11] Böttcher, M., Becker, M. & Klingner, S., 2011. Dienstleistungsmodularisierung zur kundenindividuellen Konfiguration. Berlin, Köllen Verlag.

[DIN14] DIN, 2014. Klassifikation von Dienstleistungen für die technische Betriebsführung von Erneuerbare-Energie-Anlagen, Berlin: Beuth Verlag .

[HW09] Hirschl, B. & Weiß, J. Hrsg., 2009. Dienstleistungen im Bereich erneuerbare Energien. Wirtschaftliche Bedeutung, Exportpotenziale und Internationalisierungsstrategien. München: Oekom Verlag.

[IRE13] IRENA, 2013. International Standardisation in the Field of Renewable Energy, Abu Dhabi: IRENA.

[Moe07] Mörschel, I., Behrens, H., Fähnrich, K.-P. & Elze, R., 2007. Standardisation in the Service Sector for Global Markets. In: D. Spath & K. Fähnrich, Hrsg. Advances in Services Innovations. Berlin Heidelberg: Springer Berlin Heidelberg, pp. 257-277.

[SAB13] Sonnenberg, M., Ansorge, B. & Becker, M., 2013. Potential of Service Engineering in the Field of Renewable Energies. Leipzig, s.n.

[SSK12] Sonnenberg, M., Schmidt, J. & Kühne, S., 2012. Klassifikation von Dienstleistungen der technischen Betriebsführung regenerativer Energieanlagen. Leipzig, LIV/InfAI, pp. 151-161.

[US13] USITC, 2013. Renewable Energy and Related Services: Recent Developments, Washington, DC: United States International Trade Commission.

[ZBP93] Zeithaml, V., Berry, L. & Parasuraman, A., 1993. The nature and determinants of customer expectations of service. Journal of the Academy of Marketing Science, 21(1), pp. 1-12.

Kaskadierender Widerruf von Delegationen in prozessbasierten Informationssystemen

David Hopfmüller, Sigrid Schefer-Wenzl und Mark Strembeck
Institut für Wirtschaftsinformatik und Neue Medien
Wirtschaftsuniversität Wien
{vorname.nachname}@wu.ac.at

Abstract: Das *BusinessActivities Framework* stellt einen integrierten, modellgetriebenen Ansatz zur Modellierung von Sicherheitseigenschaften für prozessbasierte Informationssysteme zur Verfügung. Ein wichtiger Teil des BusinessActivity Frameworks ermöglicht die Modellierung von rollenbasierten Zugriffskontrollmodellen. In diesem Kontext ist auch die Möglichkeit zur Delegation von Rollen und Rechten vorgesehen. Der vorliegende Beitrag erweitert den Ansatz um einen Algorithmus für den kaskadierenden Widerruf von Delegationen. Der Ansatz umfasst eine prototypische Implementierung, die als Erweiterung der BusinessActivity LRE (von engl.: Library and Runtime Engine) fungiert.

1 Einleitung

Ein Geschäftsprozess besteht aus einer Menge an Aufgaben, die in einer Organisation durchgeführt werden müssen, um bestimmte Ziele zu erreichen. Zur Unterstützung der sicheren Ausführung von Geschäftsprozessen müssen alle Subjekte, die in einem bestimmten Prozess mitarbeiten, die benötigten Rechte besitzen (siehe z.B. [GMPT01, OP03, SM11]). In den vergangenen Jahren hat sich die *rollenbasierte Zugriffskontrolle* (engl.: role-based access control, RBAC, [FKC07, SCFY96]) sowohl in der Industrie als auch im wissenschaftlichen Bereich zu einem De-facto-Standard entwickelt. In RBAC bilden Rollen verschiedene Arbeitsbereiche und Verantwortlichkeiten innerhalb einer Organisation bzw. in einem Informationssystem ab. Einer Rolle werden einerseits Zugriffsberechtigungen gemäß ihres Aufgabenbereichs zugewiesen. In weiterer Folge werden diese Rollen menschlichen Benutzern entsprechend ihrer Arbeitsprofile zugewiesen (siehe [Str10]).

Delegation ist ein Mechanismus zur Flexibilisierung des organisationalen Rechtemanagements, der auch als wichtiges Konzept für prozessbasierte Informationssysteme identifiziert wurde (siehe z.B. [CK08]). Subjekte haben hierbei die Möglichkeit, zugewiesene Aufgaben, Pflichten oder komplette Rollen an andere Subjekte weiterzuleiten (siehe z.B. [SM02, SWSB12]). In weiterer Folge agiert der Empfänger der Delegation (engl.: delegatee) anstelle (bzw. im Auftrag) des delegierenden Subjekts (engl.: delegator). Eine *mehrstufige Delegation* liegt dann vor, wenn ein Subjekt Aufgaben, Pflichten oder Rollen, die es selbst durch Delegation erhalten hat, an ein drittes Subjekt weiter delegiert.

Der *Widerruf* (engl.: revocation) einer Delegation ist jener Vorgang, bei dem ein Sub-

jekt anderen Subjekten zuvor delegierte Berechtigungen wiederum entzieht (siehe z.B. [HJPW01]). Grob kann zwischen einfachem und kaskadierendem Widerruf unterschieden werden. Beim *einfachen Widerruf* wird ein bestimmtes, delegiertes Recht bzw. eine bestimmte, delegierte Rolle ohne Berücksichtigung etwaiger Weiterdelegationen wieder zurückgenommen. *Kaskadierender Widerruf* beinhaltet die Aufhebung der delegierten Rechte, Pflichten oder Rollen auch aller Weiterdelegationen. Sobald dem Subjekt, an das zuerst delegiert wurde, die delegierten Berechtigungen entzogen werden, verlieren auch alle Subjekte, an die weiter delegiert wurde, diese Berechtigungen.

In diesem Beitrag präsentieren wir einen Algorithmus zum systematischen Widerruf von Delegationen in prozessbasierten Informationssystemen. Der Fokus liegt hierbei insbesondere auf der Diskussion spezieller Herausforderungen im Zuge des kaskadierenden Widerrufs. Die Evaluierung unseres Algorithmus erfolgte im Rahmen einer prototypischen Implementierung als Erweiterung einer bestehenden Softwareplattform, dem *BusinessActivities Framework*. Dieses Framework ermöglicht die integrierte Modellierung von Sicherheitseigenschaften für prozessbasierte Informationssysteme.

Abschnitt 2 liefert einen Überblick zum Thema Delegation und Widerruf in prozessbasierten Informationssystemen und geht insbesondere auf problematische Aspekte beim kaskadierenden Widerruf ein. Anschließend stellen wir den entwickelten Algorithmus und seine Integration in das BusinessActivities Framework vor. Abschnitt 3 schließt mit Informationen zur praktischen Implementierung und den parallel dazu entwickelten Tests an. In Abschnitt 4 wird ein Überblick über verwandte Arbeiten in diesem Bereich vermittelt. Abschnitt 5 fasst schließlich zusammen und schlägt einige Ideen und Anregungen zur Weiterentwicklung des Ansatzes vor[1].

2 Delegation und Widerruf in prozessbasierten Informationssystemen

2.1 Grundlagen der Delegation

Jede Aufgabe in einem IT-unterstützten Geschäftsprozess (z.B. eine Vertragsverhandlung) ist üblicherweise mit bestimmten Zugriffsrechten verknüpft (z.B. zur Unterzeichnung eines Vertrags). Daher muss jedes Subjekt, das Aufgaben in einem Prozess auszuführen hat, zur Durchführung dieser Aufgaben berechtigt sein (siehe z.B. [GMPT01, OP03, SM11]). In einem organisationalen Kontext kann eine Aufgabe auch mit bestimmten Pflichten verbunden sein. Eine Pflicht (engl.: duty) definiert eine Aktion, die aufgrund gesetzlicher oder unternehmensinterner Regeln durchgeführt werden muss (siehe z.B. [SM02, SS11a]). In dem Modell, das dieser Arbeit zugrunde liegt, sind sowohl Rollen, als auch Aufgaben und Pflichten delegierbar (siehe [SWSB12]). Eine Pflicht, die einem Subjekt (oder einer Rolle) zugewiesen wurde, wird auch als Verpflichtung (engl.: obligation) bezeichnet.

Delegation beschreibt im Rechte-Management einen Mechanismus, der es Subjekten er-

[1] Um den LNI-Vorgaben zu genügen, musste der vorliegende Artikel gekürzt werden. Die vollständige Fassung ist online auf der Homepage zum BusinessActivities Framework verfügbar: `http://wi.wu.ac.at/home/mark/BusinessActivities/library.html`

möglicht, Teile von oder alle zugewiesenen Aufgaben und Pflichten an andere Subjekte zu transferieren (siehe z.B. [ZOS03]). In Zusammenhang mit RBAC existiert eine Reihe von Modellen, die das Konzept so genannter Delegationsrollen einsetzen (siehe z.B. [JB06, SW08, ZOS03]). Bei dem in diesem Beitrag verwendeten Delegationsmodell wird eine *Delegationsrolle* von einem Subjekt erzeugt und enthält eine Menge delegierbarer Aufgaben und Pflichten (siehe [SWSB12]). Delegationsrollen werden in weiterer Folge an andere Subjekte delegiert, die dann anstelle des delegierenden Subjekts (z.B. in Stellvertretung) handeln können. Im Unterschied zu Delegationsrollen werden alle anderen Rollen daher nun als *reguläre Rollen* bezeichnet.

Wir unterscheiden prinzipiell drei Arten von Delegationsvorgängen:

- Delegation einzelner Aufgaben oder Pflichten an Delegationsrollen,
- Delegation vollständiger Rollen an Delegationsrollen,
- Delegation einer Delegationsrolle an ein Subjekt (genannt *delegatee*).

Um die Menge der delegierbaren Elemente sprachlich besser fassen zu können, soll an dieser Stelle als Oberbegriff aller delegierbaren Typen die Bezeichnung *Delegable* eingeführt werden. Ein *Delegable* kann somit vom Typ Aufgabe bzw. Recht, Pflicht oder Rolle sein, und kann zu Delegationszwecken einer Delegationsrolle hinzugefügt werden.

Zu jedem der oben genannten Delegationsvorgänge muss es in weiterer Folge auch einen entsprechenden Umkehrvorgang zum Widerruf bzw. zur Zurücknahme der Delegation geben. Die folgenden Rahmenbedingungen sind für einen wohlgeordneten Delegationsprozess und den anschließenden Widerruf von Delegationen von zentraler Bedeutung (siehe z.B. [BS00, SWSB12]):

1. Subjekte können eine beliebige Anzahl an Delegationsrollen erstellen. Das erstellende Subjekt (engl.: *creator*) wird automatisch zum *delegator* der Delegationsrolle.
2. Aufgaben (bzw. Rechte) können einer Delegationsrolle zugeordnet werden. Hierbei gelten folgende Regeln:
   (a) Die zu delegierende Aufgabe muss dem Ersteller der Delegationsrolle aktuell zugeordnet sein (durch eine reguläre Rolle oder, im Fall einer mehrstufigen Delegation, durch eine andere Delegationsrolle).
   (b) Die Aufgabe muss als delegierbar (engl.: *delegable*) definiert sein.
   (c) Soweit zugehörige Pflichten vorhanden sind, müssen diese ebenfalls als *delegable* definiert sein.
3. Delegationsrollen können an ein oder mehrere Subjekte *(delegatees)* delegiert werden. Diese erhalten dadurch das Recht (oder die Pflicht), die enthaltenen Aufgaben durchzuführen.
4. Subjekte können nur jene Aufgaben, Rollen und Delegationsrollen (weiter) delegieren, die ihnen selbst zugeordnet sind. Ist nur die einstufige Delegation (engl.:

single-step delegation) vorgesehen, beschränkt sich die Menge der delegierbaren Elemente eines Subjekts auf jene, die ihm über reguläre Rollen zugeordnet sind. Ist die mehrstufige Delegation (engl.: *multi-step delegation*) von Rollen und Rechten erlaubt, können auch per Delegationsrollen empfangene Elemente weiter delegiert werden.

5. Delegationsrollen können in einer *Rollenhierarchie* angeordnet werden (siehe z.B. [SWSB12, ZOS03]). Eine Rollenhierachie ist im Wesentlichen ein gerichteter, azyklischer Graph, der Vererbungsbeziehungen zwischen Rollen definiert. Um unerlaubte Rechteweitergabe zu vermeiden, gilt, dass reguläre Rollen nur als Junior-, nicht aber als Seniorrolle von Delegationsrollen zugelassen sind (siehe hierzu z.B. auch [ZOS03]).

6. Delegationsrollen können als *permanent* oder *temporär* definiert werden. Die Ausübung temporärer Delegationsrollen ist auf eine bestimmte Prozessinstanz beschränkt. Im Gegensatz dazu können permanente Delegationsrollen in jeder Instanz des zugehörigen Prozesstyps verwendet werden.

In diesem Beitrag werden weiterhin vier Arten von Einschränkungen (engl.: *constraints*) betrachtet, die die Kombination von Aufgaben einschränken (siehe z.B. [SM11] und Abbildung 1):

1. *Subject-binding* (SB): Auf diese Art verbundene Aufgaben müssen vom selben Subjekt ausgeführt werden, beispielsweise eine medizinische Diagnose und das Verfassen des zugehörigen Befundes durch den behandelden Arzt in einem Krankenhaus.

2. *Role-binding* (RB): Derart verbundene Aufgaben müssen von Mitgliedern derselben Rolle ausgeführt werden, aber nicht notwendigerweise vom selben Individuum – beispielsweise Zeichnungsberechtigungen in einer Firma.

3. *Static mutual exclusion* (SME): Aufgaben, die einander gegenseitig statisch ausschließen, dürfen nie derselben Rolle oder demselben Subjekt *zugeordnet* sein. Um Betrug und Missbrauch vorzubeugen, darf zum Beispiel kein Subjekt gleichzeitig Einkäufer und Controller der selben Unternehmung sein.

4. *Dynamic mutual exclusion* (DME): Aufgaben, die einander dynamisch ausschließen, dürfen zwar derselben Rolle (bzw. dem selben Subjekt) zugeordnet werden, diese Aufgaben müssen jedoch in jeder Prozessinstanz von verschiedenen Subjekten ausgeführt werden (“Mehraugenprinzip”).

Die oben erwähnten Einschränkungen spielen bei der Delegation eine wichtige Rolle und erfordern einigen Aufwand, um ihre Einhaltung zu garantieren. *Subject-binding* und *role-binding* sind zudem auch für den Widerruf einer Delegation relevant, um einen konsistenten Zustand zu erhalten.

Sobald Aufgaben und Rollen delegiert werden können, entsteht die damit verbundene Anforderung, diese Delegationen auch widerrufen zu können. Dies trifft insbesondere dann

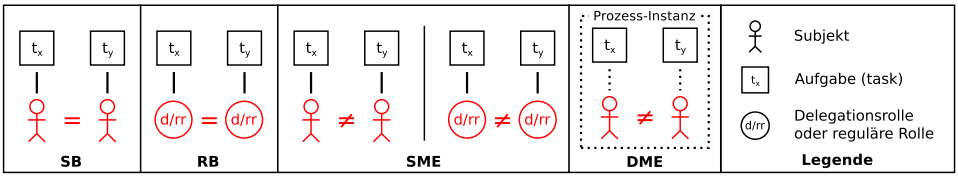


Abbildung 1: Mögliche Einschränkungen von Delegationen

zu, wenn die Delegation temporär erfolgen soll, wie es beispielsweise bei der Urlaubsvertretung in einer Firma der Fall sein kann. Es gilt der Grundsatz, dass grundsätzlich alles widerrufen werden kann, was delegiert werden kann. In Frage kommen somit:

- Widerruf einzelner *Aufgaben* (engl.: tasks) von einer Delegationsrolle,
- Widerruf von *regulären Rollen*, welche an eine Delegationsrolle delegiert wurden,
- Widerruf von *Delegationsrollen*, die an eine andere Delegationsrolle delegiert wurden,
- Entfernung einzelner *Empfänger* (engl.: delegatees) aus einer Delegationsrolle.

Für jeden dieser Widerrufsvorgänge unterscheiden wir grundsätzlich zwei Fälle:

Einfacher Widerruf: Beim einfachen Widerruf wird eine einzelne Delegation widerrufen. Bei dieser Delegation kann es sich um eine Aufgabe, Delegationsrolle oder reguläre Rolle handeln, die aus einer Delegationsrolle entfernt werden soll. Alternativ kann auch ein Empfänger einer solchen Delegationsrolle entfernt werden.

Kaskadierender Widerruf: Im Fall der mehrstufigen Delegation kann ein Empfänger die delegierten Elemente seinerseits weiterdelegieren. Beim kaskadierenden Widerruf sollen auch alle Weiterdelegationen rückgängig gemacht werden.

2.2 Problemstellungen des kaskadierenden Widerrufs

(1) Mehrfache Zuweisung

Das Grundproblem des kaskadierenden Widerrufs von Delegationen besteht darin, dass ein Empfänger eines delegierten Elements dasselbe Element mehrfach erhalten haben kann, nämlich durch eine (andere) Delegationsrolle, oder durch Zuweisung über regulärer Rollen.

Delegiert nun der Empfänger *(delegatee)* das Element seinerseits weiter, ist prinzipiell unklar, welche Instanz des Delegables weiter delegiert wurde. Zwar wäre eine Implementierung denkbar, die die einzelnen Instanzen unterscheidbar hält, diese Unterscheidung ist jedoch aus Sicht der Geschäftslogik nicht relevant und wird daher hier nicht weiter

betrachtet. Allerdings bedeutet dies auch, dass die weitere Delegation nicht einfach zurückgezogen werden darf, ohne zu prüfen, ob eine zweite Quelle (engl.: *second source*) des fraglichen Elements existiert, die als Ursprung der Weitergabe in Frage kommt: Das kann entweder eine andere Delegationsrolle oder eine reguläre Rolle sein.

(2) Bestimmung der Herkunft mehrfacher Zuweisungen

Werden für ein Delegable ausschließlich Delegationsrollen als zweite Quelle gefunden, besteht die Möglichkeit, dass deren Herkunft letztendlich ausschließlich auf dieselbe (nämlich die ursprünglich zu widerrufende) Delegationsrolle zurückführt. Ist das der Fall, handelt es sich *nicht* um eine (qualifizierte) zweite Quelle.

(3) Zyklen bei mehrfacher Zuweisung

Im Zuge der Auflösung der Herkunft von Delegationsrollen, die als zweite Quelle in Frage kommen, stellt neben deren Herkunft auch die mögliche Zyklenbildung ein Problem dar. Zwar ist die direkte Bildung eines Zyklus zwischen zwei Delegationsrollen nicht zulässig (vgl. Abschnitt 2.1), allerdings könnte ein Subjekt S1 ein Element an Subjekt S2 delegieren, welches dasselbe Delegable (mittels einer eigenen Delegationsrolle) wieder zurück an S1 delegiert. Das Erkennen solcher zyklischer Graphen ist Voraussetzung für eine korrekte Implementierung des zugehörigen Modells.

(4) Weiter-Delegation von Teilmengen

Hat die ursprüngliche Delegationsrolle reguläre Rollen oder Delegationsrollen enthalten, können diese Container ihrerseits prinzipiell beliebig viele und beliebig tief verschachtelte Elemente enthalten. Der Empfänger einer solchen Delegationsrolle hat (bei aktivierter Mehrfachdelegation) nun die Möglichkeit, beliebige Teilmengen der ursprünglichen Delegationsrolle weiter zu delegieren.

Beim Auflösen derartiger Delegationsrollen sind daher nicht nur die direkten Juniorrollen (Kind-Elemente) einer Delegationsrolle zu beachten und zu behandeln, sondern gleichermaßen alle in Rollen enthaltenen Delegables.

2.3 Algorithmus für den kaskadierenden Widerruf von Delegationen

Der aus den oben genannten Anforderungen resultierende Algorithmus wurde in zwei größere Funktionen sowie einige Hilfsfunktionen aufgeteilt. Erstere sind dabei von besonderem Interesse, da sie der Verfolgung der potentiell komplexen Delegationsbeziehungen dienen. Die Prozedur *cascadingRevocation* sowie der Algorithmus *findSecondSource* sind nachfolgend in Pseudocode wiedergegeben und zudem als Erweiterung der BusinessActivity LRE (von engl.: Library and Runtime Engine) implementiert (siehe Abschnitt 3)[2]. Der Pseudocode der dort verwendeten Nebenfunktionen *taskRoleDelegationRevoke* und *roleRoleDelegationRevoke* ist aus Platzgründen nur in der erweiterten Version dieses Artikels enthalten.

[2]Der Quellcode der BusinessActivity LRE ist zum freien Download verfügbar: `http://wi.wu.ac.at/home/mark/BusinessActivities/library.html`

Wir benötigen die Menge aller Subjekte S, die Menge aller Rollen R (die alle Delegationsrollen DR miteinschließt), sowie die Menge der Aufgabentypen T_T samt seiner Teilmenge der delegierbaren Aufgabentypen DT_T. Auf Seiten der Funktionen kommt $drh^*(dr) = R_{j^*}$ zum Einsatz, welche die Menge aller direkten und transitiven Junior-Rollen einer Delegationsrolle liefert. Selbiges ist durch $rrh^*(r) = R_{j^*}$ für die reguläre Rollen-Hierarchie definiert. Das Mapping $town(dr) = T_{T_{dr}}$ liefert alle (auch über Junior-Rollen) delegierten Aufgaben einer Delegationsrolle, $rown_{rrh}(s) = R_s$ hingegen alle einem Subjekt über die reguläre Rollenhierarchie direkt und transitiv zugeordneten Rollen. Mit $rsdel(s) = DR_s$ können alle Delegationsrollen ermittelt werden, die momentan einem Subjekt zugeordnet sind. Umgekehrt kann mittels $rsdel^{-1}(dr) = S_{dr}$ auf alle Delegatees einer Delegationsrolle zugegriffen werden. $creator(dr) = s_{dr}$ wird verwendet, um den Ersteller einer Delegationsrolle zu bestimmen. Die Umkehrung $creator^{-1}(s) = DR_s$ ist in [SWSB12] nicht ausdrücklich definiert, hierfür verwenden wir folgende Definition: *Das Mapping* $creator^{-1} : S \rightarrow \mathcal{P}(DR)$ *wird* ***created delegation roles*** *genannt. Für ein gegebenes Subjekt* $s \in S$ *retourniert* $creator^{-1}(s) = DR_s$ *alle von diesem erstellte Delegationsrollen* $DR_s \in DR$.

Prozedur 1. *Finden und Widerrufen aller weiter delegierten Instanzen eines Delegables.*

Name: *cascadingRevocation*
Input: $dr \in DR$, $revoke \in S \cup R \cup DT_T$, $handledSubjects \in S$

```
1:  create_empty_set delegatees
2:  create_empty_set revokes
3:  create_empty_set secondSources
4:  if revoke ∈ S then
5:      set delegatees = revoke
6:      add drh*(dr) ∪ town(dr) to revokes
7:  else
8:      add rsdel⁻¹(dr) to delegatees
9:      add revoke to revokes
10:     if revoke ∈ R then
11:         add rrh*(dr) ∪ town(dr) to revokes
12:     endif
13:     if revoke ∈ DR then
14:         add drh*(dr) ∪ town(dr) to revokes
15:     endif
16: endif
17: for each delegatee ∈ delegatees
18:     add delegatee to handledSubjects
19:     for each createdDg ∈ drh*(creator⁻¹(delegatee)) \
        ∪ town(creator⁻¹(delegatee))
20:         if createdDg ∈ revokes ∧ createdDg ∉ secondSources then
21:             if createdDg ∈ findSecondSource(createdDg, delegatee) then
22:                 add createdDg to secondSources
23:             else
24:                 if createdDg ∈ DT_T then
```

```
25:                 for each createdDr ∈ creator^-1(delegatee) \
                    | createdDg ∈ town(createdDr)
26:                     taskRoleDelegationRevoke \
                        (createdDg, createdDr, "cascation", handledSubjects)
27:             else if createdDg ∈ R then
28:                 for each createdDr ∈ creator^-1(delegatee) \
                    | createdDg ∈ drh*(createdDr)
29:                     roleRoleDelegationRevoke \
                        (createdDg, createdDr, "cascation", handledSubjects)
30:             endif; endif; endif; endfor; endfor
```

Algorithmus. *Bestimmen, ob es weitere Quellen als Grundlage für eine Delegation gibt.*

```
Name: findSecondSource
Input: dg ∈ R ∪ DT_T, subject ∈ S, creator ∈ S
 1: create_empty_set checkedSubjects
 2: if dg ∈ DR ∧ creator(dg) == subject then
 3:     return true
 4: elseif dg ∈ rown_rrh(subject) ∪ town(rown_rrh(subject))
 5:     return true
 6: else
 7:     add subject to checkedSubjects
 8:     for each dr ∈ rsdel(subject)
 9:         if (dg ∈ R ∧ dg ∈ drh*(dr))
10:             ∨ (dg ∈ DR ∧ dg ∈ drh*(dr))
11:             ∨ (dg ∈ DT_T ∧ dg ∈ town(dr)) then
12:             if creator(dr) ∈ checkedSubjects then
13:                 return false
14:             else
15:                 return findSecondSource(dg, creator(dr), checkedSubjects)
16:             endif; endif; endfor
17:     return false
18: endif
```

2.3.1 Die Hauptfunktion `cascadingRevocation`

Ist kaskadierender Widerruf gewünscht, wird die zentrale Methode `cascadingRevocation` von einer der Methoden zum einfachen Widerruf eingebunden. Diese Methode enthält grundsätzlich all jene Funktionalität, die zur Behandlung eines Delegationsschritts erforderlich ist. Sie erwartet zumindest folgende zwei Parameter:

(1) Das zu widerrufende Delegable: Bezüglich der Definition der in Frage kommenden Typen von Delegables sei auf Abschnitt 2.1 verwiesen; (2) die Delegationsrolle, aus der das Delegable widerrufen werden soll, beziehungsweise aus der ein Delegationsempfänger entfernt werden soll.

In einem ersten Schritt werden drei Listen angelegt: Die Liste `revokes` enthält die zu berücksichtigenden Delegables (Rollen können ihrerseits weitere Elemente enthalten), die Liste `delegatees` enthält alle Delegationsempfänger, die auf Weiterdelegation überprüft werden müssen. Anschließend wird überprüft, ob einer dieser Delegationsempfänger eines der Delegables weiterdelegiert hat. Die Listen `handledSubjects` und `secondSources` dienen der Zwischenspeicherung bereits gefundener Subjekte bzw. Sekundärquellen.

Ein kritischer Punkt ist jedoch, dass ein Delegable nur widerrufen werden darf, wenn dafür keine valide Sekundärquelle existiert. Diese Suche übernimmt die Methode `findSecondSource` (siehe Algorithmus *findSecondSource*), die für das Tupel (Delegable, Subjekt) sowohl die regulär zugewiesenen, als auch die per Delegation empfangenen Objekte eines Subjekts findet. Existiert eine solche Sekundärquelle, wird das Tupel nicht weiter behandelt. Wird allerdings keine solche Sekundärquelle ausfindig gemacht, muss die Delegation rückgängig gemacht werden. Dies geschieht, indem die entsprechende Methode (`roleRoleDelegationRevoke` beziehungsweise `taskRoleDelegationRevoke`) mit den neuen Parametern des Delegables und der Delegationsrolle aufgerufen wird. Um unnötige Redundanzen zu vermeiden, werden entlang der Ausführung gewonnene Erkenntnisse über Schleifeniterationen hinweg zwischengespeichert. Gerade im Fall wechselseitig rekursiver Methoden kann die Laufzeit dadurch deutlich reduziert werden. Zudem wird die Liste der bereits untersuchten Subjekte gespeichert, um eine mehrfache Untersuchung zu verhindern.

2.3.2 Suche nach Sekundärquellen: `findSecondSource`

Die Methode `findSecondSource` sucht, wie bereits im vorangegangenen Abschnitt erwähnt, für ein konkretes Subjekt die Sekundärquellen eines bestimmten Delegables. Folgende Sekundärquellen kommen in Frage:

- Die mittels regulärer Rollen zugewiesenen Elemente. Diese zu finden ist vergleichsweise einfach und in die Methode `getRegularPermissions` ausgelagert.
- Ist das Delegable eine Delegationsrolle und das zu überprüfende Subjekt deren Delegator, handelt es sich um eine gültige Sekundärquelle.
- Die Ermittlung der ebenfalls als Sekundärquelle in Betracht kommenden empfangenen Delegables ist hingegen interessanter: Zuerst werden die empfangenen Delegationsrollen des Subjekts gesammelt und mittels der Funktionen `hasJuniorRole` beziehungsweise `ownsTask` daraufhin überprüft, ob das gesuchte Delegable enthalten ist. Ist dies bei keiner der Delegationsrollen der Fall, war die Suche erfolglos. Ist das Ergebnis jedoch positiv, gilt es auszuschließen, dass die in der Delegationsrolle enthaltenen Delegables auf eine andere Delegationsrolle zurückgehen, die bereits widerrufen wurde oder noch zu widerrufen ist – Delegationen also letztendlich “aufgrund ihrer selbst” legitimiert werden.

Die Rückgabe der Methode ist ein einfacher Wahrheitswert (engl.: *boolean*), der angibt, ob eine valide Sekundärquelle für das angegebene Delegable existiert oder nicht.

3 Integration in BusinessActivities

Das BusinessActivities Framework ist im Rahmen der *BusinessActivity LRE* (von engl.: Library and Runtime Engine) implementiert (für weitere Details siehe auch [SM11]). Der Aufbau ist modular gehalten, Erweiterungen wie Pflichten (engl.: *duties*) oder Delegationen können dadurch unkompliziert bei Bedarf eingebunden werden (siehe z.B.: [SS11a, SS11b, SWS13, SWSB12]). Auch die Implementierung zum Widerruf von Delegationen wurde in einer *Revocation* genannten Erweiterung gekapselt und umfasst neben den (bereits bestehenden) Methoden zum Widerruf einfacher Delegationen auch jene für den kaskadierenden Widerruf.

Eine der im Zuge der Implementierung zu lösenden Herausforderungen war die Suche nach geeigneten Parametern für die Methode `cascadingRevocation`. Dieses Problem trat erst beim kaskadierenden Widerruf von Delegationen auf, da in diesem Fall mehrere Datentypen für einen Parameter in Frage kommen. Statt der Kurznamen werden daher die vollqualifizierten Objektnamen verwendet. Über den Zugriff auf das entsprechende Objekt kann auf weitere seiner Eigenschaften zugegriffen werden, unter anderem auf den entsprechenden Objekttyp.

Zur Verifikation der Erweiterung wurde eine Reihe an Tests definiert, die jeweils eine zu erwartende Ausgangssituation modellieren. Nach Ausführung einer oder mehrerer mit der Delegation bzw. dem Widerruf in Verbindung stehender Operationen wird abschließend der erwartete Endzustand überprüft.

4 Verwandte Arbeiten

Die Überlegungen zum Widerruf von Delegationen bauen in erster Linie auf Arbeiten zur Delegation selbst auf. Es existiert mittlerweile eine Vielzahl unterschiedlicher Delegationsmodelle. Wir beschränken uns hier auf die Auswahl einiger Modelle, die unserem Delegations- und Widerrufsmodell am ähnlichsten sind.

Barka und Sandhu entwerfen in [BS00] das früheste vollständig definierte rollenbasierte Delegationsmodell RBDM (von engl.: Role-based Delegation Model) mit Widerrufsmöglichkeiten. In RBDM kann ein Benutzer komplette Rollen an andere Benutzer delegieren. Ein darauf aufbauender Ansatz ist das *PBDM* (von engl.: Permission-based Delegation Model), das von Zhang et al. in [ZOS03] beschrieben wird. In [HMM10] wurde das Capability-based Delegation Model (CRBAC) präsentiert, das viele Funktionalitäten des PBDM-Modells übernimmt. Der spezielle Fokus des CRBAC-Ansatzes liegt auf domänenübergreifender Delegation von Rollen und Rechten im Rahmen von Capability-Transfers.

Ein Beitrag in Form eines regelbasierten Frameworks für Delegationen und deren Widerruf ist in [ZAC03] zu finden, wo Zhang et al. insbesondere auf den Widerruf von Delegationen eingehen.

Der vorliegende Beitrag unterscheidet sich somit insofern grundlegend von oben genann-

ten Arbeiten, als er umfassend konkrete Problemstellungen erörtert und löst, die sich beim kaskadierenden Widerruf in prozessbasierten Informationssystemen auf Basis von RBAC ergeben können. Die Arbeit trägt mit dem kaskadierenden Widerruf einen wesentlichen Teil zur Komplettierung des bereits vorgestellten Konzepts der Delegation im Rahmen des BusinessActivities-Frameworks bei, da Delegationen ohne ihren Gegenpart wenig Praxisrelevanz haben.

5 Zusammenfassung und Ausblick

In diesem Beitrag diskutierten wir die Problemstellungen, die sich beim kaskadierenden Widerruf von Delegationen in prozessbasierten Informationssystemen ergeben. Der kaskadierende Widerruf ermöglicht die Aufhebung von Delegationen über mehrere Stufen hinweg und erfordert die Durchführung einer Reihe von Schritten, um die Konsistenz des gesamten Zugriffskontrollmodells nicht zu gefährden. Wir stellten zunächst einen generischen Algorithmus zum systematischen Widderruf von Delegationen vor, der den kaskadierenden Widerruf in beliebigen prozessbasierten Informationssystemen adressiert. Weiterhin präsentierten wir eine konkrete prototypische Implementierung unseres Algorithmus, die im Rahmen der Erweiterung einer bestehenden Softwareplattform erfolgte.

Aufbauend auf unserem Ansatz ergeben sich einige konkrete Ansätze für weiterführende Arbeiten: Momentan ist nur der Widerruf einzelner Delegables vorgesehen. Gerade wenn eine Delegationsrolle als Sammlung von Elementen verwendet wird, die in Zusammenhang mit einem bestimmten Anwendungsfall stehen, ist es naheliegend, diese auch als Gesamtes widerrufen zu können. Zudem wäre interessant, gezielt einzelnen Subjekten Delegables wieder entziehen zu können, die diese per Mehrfach-Delegation erhalten haben, wenn sich beispielsweise Zuständigkeiten innerhalb eines Unternehmens ändern. Diese Option ist momentan nur im Rahmen des einfachen Widerrufs, also jeweils durch den unmittelbar Delegierenden, vorgesehen.

Literatur

[BS00] Ezedin Barka und Ravi Sandhu. Framework for Role-Based Delegation Models. In *Proceedings of the 16th Annual Computer Security Applications Conference (ACSAC)*, December 2000.

[CK08] Jason Crampton und Hemanth Khambhammettu. On Delegation and Workflow Execution Models. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC)*, March 2008.

[FKC07] David F. Ferraiolo, D. Richard Kuhn und Ramaswamy Chandramouli. *Role-Based Access Control*. Artech House, second. Auflage, 2007.

[GMPT01] Christos K. Georgiadis, Ioannis Mavridis, George Pangalos und Roshan K. Thomas. Flexible Team-Based Access Control Using Contexts. In *Proceedings of the 6th ACM Symposium on Access Control Models and Technologies (SACMAT)*, May 2001.

[HJPW01] A. Hagstrom, S. Jajodia, F.P. Presicce und D. Wijesekera. Revocations - a Classification. In *Proc. of the 14th IEEE Computer Security Foundations Workshop*, 2001.

[HMM10] Koji Hasebe, Mitsuhiro Mabuchi und Akira Matsushita. Capability-Based Delegation Model in RBAC. In *Proceedings of the 15th ACM Symposium on Access Control Models and Technologies (SACMAT)*, June 2010.

[JB06] James B. D. Joshi und Elisa Bertino. Fine-grained Role-based Delegation in Presence of the Hybrid Role Hierarchy. In *Proceedings of the 11th ACM Symposium on Access Control Models and Technologies (SACMAT)*, June 2006.

[OP03] Sejong Oh und Seog Park. Task-role-based access control model. *Information Systems*, 28(6), 2003.

[SCFY96] Ravi Sandhu, Edward Coyne, Hal Feinstein und Charles Youman. Role-Based Access Control Models. *IEEE Computer*, 29(2), 1996.

[SM02] Andreas Schaad und Jonathan D. Moffett. Delegation of Obligations. In *Proceedings of the 3rd International Workshop on Policies for Distributed Systems and Networks (POLICY)*, June 2002.

[SM11] Mark Strembeck und Jan Mendling. Modeling Process-related RBAC Models with Extended UML Activity Models. *Information and Software Technology*, 53(5), 2011.

[SS11a] Sigrid Schefer und Mark Strembeck. Modeling Process-Related Duties with Extended UML Activity and Interaction Diagrams. In *Electronic Communications of the EASST, Vol. 37*, March 2011.

[SS11b] Sigrid Schefer und Mark Strembeck. Modeling Support for Delegating Roles, Tasks, and Duties in a Process-Related RBAC Context. In *Proc. of the International Workshop on Information Systems Security Engineering (WISSE)*, 2011.

[Str10] Mark Strembeck. Scenario-Driven Role Engineering. *IEEE Security & Privacy*, 8(1), 2010.

[SW08] Qinghua Shang und Xingang Wang. Constraints for Permission-Based Delegations. In *Proceedings of the 8th IEEE International Conference on Computer and Information Technology Workshops (CITWORKSHOPS)*, 2008.

[SWS13] Sigrid Schefer-Wenzl und Mark Strembeck. Modeling Context-Aware RBAC Models for Mobile Business Processes. *International Journal of Wireless and Mobile Computing*, 6(5):448–462, 2013.

[SWSB12] Sigrid Schefer-Wenzl, Mark Strembeck und Anne Baumgrass. An Approach for Consistent Delegation in Process-Aware Information Systems. In *Proc. of the 15th International Conference on Business Information Systems (BIS)*, Jgg. 117, 5 2012.

[ZAC03] Longhua Zhang, Gail-Joon Ahn und Bei-Tseng Chu. A rule-based framework for role-based delegation and revocation. *ACM Trans. Inf. Syst. Secur.*, 6(3):404–441, August 2003.

[ZOS03] Xinwen Zhang, Sejong Oh und Ravi Sandhu. PBDM: a flexible delegation model in RBAC. In *Proceedings of the eighth ACM symposium on Access control models and technologies*, SACMAT '03, Seiten 149–157, New York, NY, USA, 2003. ACM.

Who may share with Whom?!
Mixed-Tenancy: Conditional Sharing of Cloud Resources

Stefan T. Ruehl[1], Urs Andelfinger[2], Andreas Rausch[1], and Stephan A.W. Verclas[3]
[1]Clausthal University of Technology
Albrecht-von-Groddeck-Str. 7, 38678 Clausthal-Zellerfeld, Germany
Email: {stefan.t.ruehl, andreas.rausch}@tu-clausthal.de
[2]University of Applied Sciences Darmstadt
Haardtring 100, 64295 Darmstadt, Germany
Email: urs.andelfinger@h-da.de
[3]T-Systems International GmbH
Heinrich-Hertz-Str. 1, 64295 Darmstadt, Germany
Email: stephan.verclas@t-systems.com

Abstract: Software-as-a-Service is a delivery model whose basic idea is to provide applications to the customer on demand over the Internet. SaaS thereby promotes multi-tenancy as a tool to exploit economies of scale. This means that a single application instance serves multiple customers. However, a major drawback of SaaS is the customers' hesitation of sharing infrastructure, application code, or data with other tenants. This is due to the fact that one of the major threats of multi-tenancy is information disclosure due to a system malfunction, system error, or aggressive actions. So far the only approach in research to counteract on this hesitation has been to enhance the isolation between tenants using the same instance. Our approach (presented in earlier work) tackles this hesitation differently. It allows customers to choose if or even with whom they want to share the application. The approach enables the customer to define their constraints for individual application components and the underlying infrastructure. Based on these constraints, the SaaS application is deployment in the way that all constraints are satisfied and only minimal resources are required.

This paper's contribution is a summary of all results that were developed during the creation of the mixed-tenancy approach. These results focus on the challenges of capturing customers' deployment constraints, computing a valid and optimal deployment, and an evaluation of the general applicability of mixed-tenancy in real-world.

1 Introduction – Problem Analysis

Software-as-a-Service (SaaS) is a delivery model whose basic idea is to provide applications to the customer on demand over the Internet. In contrast to similar but older approaches, SaaS promotes multi-tenancy as a tool to exploit economies of scale. Multi-tenancy means that a single application instance serves multiple customers at the same time. However, even though multiple customers use the same instance, each of them has the impression that the instance is designated only to themselves. This is achieved by isolating the tenants' data from each other [CC06].

In contrast to single-tenancy, multi-tenancy has the advantage that IT infrastructure may be used as efficiently as possible to host as many tenants as possible on the same instance. Thus, operational and maintenance cost of the application is decreased. However, one of the major threats of multi-tenancy applications is information disclosure due to data breach [Clo13]. This may occur through a system error, malfunction, or destructive actions. So far this problem has only been tackled by proposing new approaches to implement and improve the tenants' isolation on a single instance (e.g. [GSH+07], [BAT12], and [AGI12]).

Our approach presented in previous works ([RARV12], [RWV13], [LMRV14], [RRM+14], [Rue14]) is different, however, as it strives to solve the problem by finding a hybrid solution between multi-tenancy and single-tenancy. We refer to this approach as *mixed-tenancy* (MT). The approach tries to emphasize both, the customers' concerns or fear about sharing infrastructure as well as the operator's desire to utilize infrastructure as efficiently as possible in order to gain low cost.

The basic idea of the MT is to utilize component-based applications. These so-called composite SaaS-applications are composed of a number of application components (AC) that each offer atomic functionality. Based on these applications, the MT approach enables the customer to provide constraints that state if or even with whom they are willing to share. They may express their constraints not just for the ACs, but also for the underlying infrastructure stack an application component utilizes (e.g. application server, virtual machine). Based on these constraints, a deployment is computed that complies to all constraints (it is valid) and, thereby, only utilizes minimal resources (it is optimal). The computed deployment is afterwards used as a blueprint according to which the SaaS-application is deployed.

When developing the MT approach the following three research questions were identified to be important and, thus, were tackled first. ***Research Question 1 (RQ-1)*** – How can customers' constraints towards MT be described? ***Research Question 2 (RQ-2)*** – How can a valid and optimal deployment be found in polynomial time? ***Research Question 3 (RQ-3)*** – Is it possible to apply the MT approach to existing composite applications? In this paper we present a summary of the results related to these research questions. Thus, the rest of this paper structures accordingly. Section 2 discusses the work done with respect to research question RQ-1 and summarizes the results achieved. For research question RQ-2 the same is done by Section 3. And finally, the results produced for research question RQ-3 are discussed in Section 4. Section 5 will conclude this paper by summarizing its results and outlining opportunities to develop the MT approach further.

2 Capturing Customers Deployment Constraints

The first research question to be tackled in order to create the MT approach was RQ-1. It deals with capturing customers' deployment constraints towards MT. The special challenge involved is that it shall be possible for customers to express with which other tenants they do and do not want to share ACs (and the underlying infrastructure), even if the operator does not reveal its customer base to the customers. This section will provide the

description of a generic description model that allows that and, further, may be used for a wide variety of different applications.

2.1 Conceptual Analysis Based on Requirements

The first step towards the definition of the description model is to analyze what needs to be captured. The following is a summary.

Deployment Levels (DL) – It has previously been stated (Section 1) that customers shall not just be able to express their deployment constraints for the ACs but also for the underlying infrastructure (e.g. application server, virtual machine). Within this work we refer to those slices of the stack for which constraints shall be expressible as DLs. However, in order to have a slice serve as a DL, it is necessary that they fulfill certain characteristics, like that it is possible to instantiate multiple units of a single DL and that any unit of a DL may host multiple of a higher DL[1].

Deployment Models (DM) – It is the idea of this model to allow customers to express different types of deployment constraints. These types are called DMs and differ in the way how customers may state the tenants they may share with. The following five are those considered: *private* (no sharing shall be allowed), *public* (sharing is allowed with all other tenants), *white hybrid* (sharing shall only be possible with specific tenants that are included – e.g. only with companies from Europe), *black hybrid* (sharing shall be possible with all tenants but with those that are excluded – e.g. not with competitors), and *gray hybrid* (sharing shall be possible only with specific tenants but not with specific others – e.g. only with companies from Europe but not with competitors).

Groups – In order to be able to express constraints like the ones just mentioned, it is necessary to have means of categorizing tenants. Thus, a group is an entity to which all tenants are associated that share the commonality the group represents (e.g. geographic regions like Europe, USA).

Dimensions – Further, it shall be possible to create groups that belong to different topics, such as industries (containing different industries e.g. finance, IT, telecommunications), geographic (containing the countries of geographic regions tenants operate in), or other things like data privacy acts tenants may apply to. Within this work, these topics are referred to as dimensions. Each dimension is realized by multiple groups that belong to only this dimension. Further, groups may have a subset relationship between them in the shape of an acyclic graph. This allows tenants to more precisely express their constraints since they may, for example, choose to exclude banks but not insurances, or exclude all customers from the entire finance industry (thereby excluding both banks and insurances).

Virtual Tenants – In the introduction to this section, it has been stated that keeping customer base secret to the customers, is one of the major objectives of the description model. The concept of virtual tenants allows to define explicitly which tenants shall be included or excluded. They represent companies that exist in the market but are not customers of

[1] A full list of requirements may be found in [RWV13] and [Rue14].

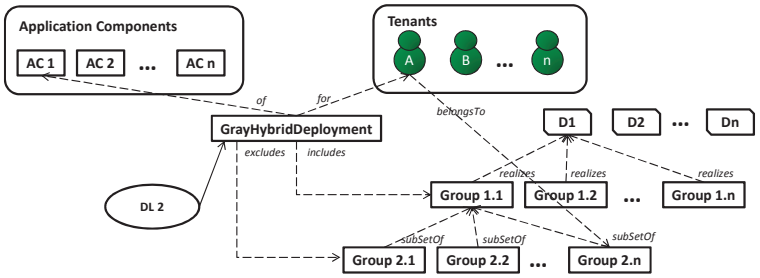


Figure 1: Description of a Deployment Constraint

the application yet. If one tenant demands to exclude another specific tenant, for example, but this tenant has not subscribed yet, it is possible to use virtual tenants as dummies to express constraints but they will not be considered when computing a deployment.

In order to verify that, the requirements just presented are sufficient to be used in reality. They have been discussed with several experts from both academia and industry.

2.2 Process of Deployment Constraint Definition

Due to the fact that the description model is intended to be as generic as possible, the definition of deployment constraints is performed not just by the tenants but also by the operator. The following introduces the necessary steps to capture customers' deployment constraints. *Step 1 – Customization* This step's purpose is to adapt the generic model to the specific application for which deployment constraints shall be captured. Thus, it covers the capturing of the specific ACs, DLs, dimensions, groups, and tenants. *Step 2 – Grouping* This step's purpose is to associate tenants to those groups they belong to (each tenant needs to be assigned to at least one group per dimension). The reason that this needs to be done is that otherwise it would not be certain that excluding a particular group may actually result in excluding the tenants that apply to this commonality. Thus, this step is crucial since if data is flawed, deployment constraints defined may not be properly realized. Again, these first two steps are executed by the operator. *Step 3 – Constraint Definition* Finally, in this step customers express their deployment constraints. Figure 1 illustrates an example for this. In the example tenant A chooses a *gray hybrid* DM to express that they want to share units of DL 2 for AC 1 only with tenants that belong to group 1.1 but not with tenants belonging to group 2.1. This means sharing will only be possible with tenants that belong to the groups 1.1, 2.2, ..., 2.n. *Step 4 – Deployment Information (DI) Extraction* The final step of the process is to extract the information about which tenants may share which instances of ACs and units of DLs, called DI. For a specific combination of AC and DL it is possible to represent this information as an undirected graph. In such a graph the vertices represent the tenants and an edge between two vertices states that those tenants are allowed to share. In order to comply with all constraints given by customers, there shall only be an edge between two tenants if none of the tenants has stated a constraint that prohibits

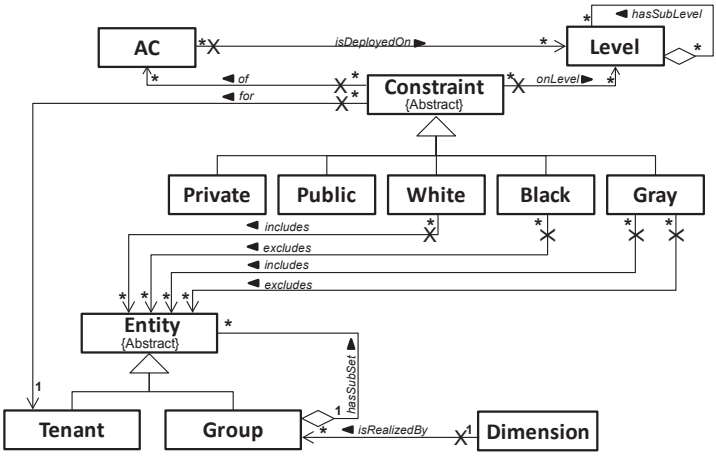


Figure 2: Description Model as UML Class Diagram (incomplete)

sharing. Thus, DI is a collection of graphs, one for every combination of an application component and a deployment level (where the application component is deployed on the deployment level). An example of it is illustrated on the left hand side of Figure 3.

2.3 Introduction of Description Model

Based on the previous two subsections, it is possible to introduce the description model. Figure 2 gives an overview of it using UML class diagram notation. It illustrates the entities that were discussed in previous sections as classes. Furthermore, there are relations between these classes that may be used to describe deployment constraints. In Figure 1 an example was given in which tenant A defined that they want to use AC-1 as a gray deployment on DL-2. Realizing this based on the UML Model would mean that there is an object of the type AC, called *AC-1*. In addition, there is an object *A* of type tenant and an object *DL-2* of type Level. These objects would have been created by the operator as part of the first step.

In order to create the aforementioned deployment constraint, they would create an object of type public. This object is then associated to *AC-1* by using the *of*-relationship, to *A* by using the *for*-relationship, to *DL-2* by using the *onLevel*-relationship, to group 1.1 by the *include*-relationship, and to group 2.1 by the *exclude*-relationship. Thus, the constraint is captured.

Please note that the structure between tenants, groups and dimensions is realized in the description model using the composite design pattern [GHJ94]. It allows that every group may have multiple subgroups as well as that tenants may be associated to any group. An inclusion or exclusion of a group shall also include all tenants or groups that are directly or transitively connected to the group through the *hasSubSet*-relationship.

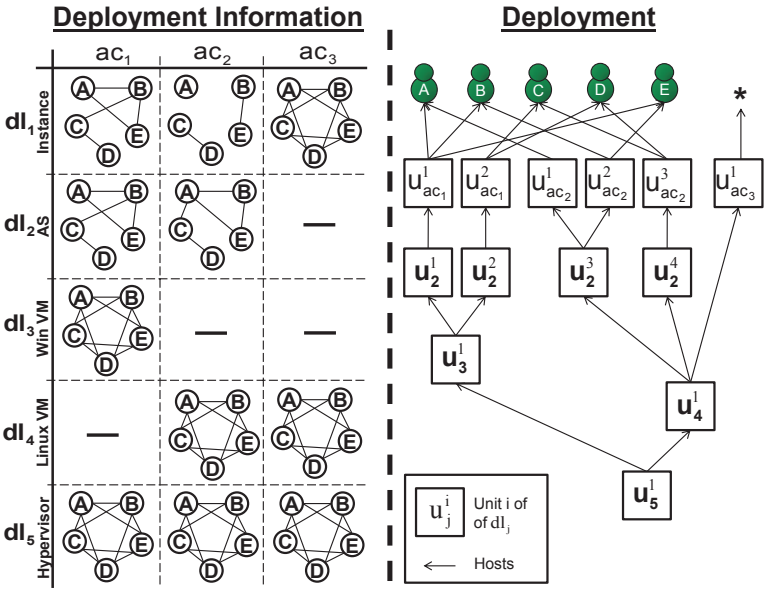


Figure 3: An Example of a Valid and Optimal Deployment

A realization of the model was created by utilizing the Web Ontology Language (OWL)[2]. Furthermore, it was evaluated if the model is actually able of providing the expressiveness required to model all requirements discussed. Based on an example that consisted of all cases, it was possible to show that the expressiveness of the model is sufficient. Details on both, the realization and the evaluation may be found in [RWV13] and [Rue14].

3 Computation of a Valid and Optimal Deployment

The second research question deals with the computation of a valid and optimal deployment. This question is addressed in this section.

3.1 Problem Description

The first step in this section is to introduce the deployment problem. This is done based on the example that is illustrated by Figure 3 (left hand side).

The input for the problem is all constraints given by customers in an aggregated from. It consists of multiple undirected graphs, one for every combination of AC and DL, where the AC shall be deployed on the DL. Each graph contains all tenants as vertices. An edge

[2] http://www.w3.org/TR/owl2-overview/

between two vertices states that these two tenants are allowed to share an AC on a given DL. It covers three ACs and five DLs. As visible, not all ACs are deployed on all DLs.

Based on the input, it is the goal to compute a deployment that is valid and optimal. A deployment shall be a collection of units for every DL. On the highest level units need to be created for every AC separately since each one needs to be instantiated (e.g. ac_1: $u^1_{ac_1}, u^2_{ac_1}$). On the other DLs only one type of unit needs to be created per DL (e.g. dl_2: $u^1_2, u^2_2, u^3_2, u^4_2$). Furthermore, units of the highest DL hosts tenants directly, the other units host units of higher DLs. A valid and optimal deployment for the example is illustrated on the right hand side of Figure 3. For each of the ACs, units were created. Each of them hosts tenants.

Due to the structure of a deployment, it is not possible to assign any tenants to the same units of the highest level but only those that are allowed to do so according to the related graphs. This is also considered in the example (e.g. $u^1_{ac_1}$ is shared by tenants A, B, and E which are all adjacent to each other in the corresponding graph). Units of the DLs dl_2 – dl_5 are used transitively through the host's relationship (e.g. all tenants use u^1_5). For these units there are two constraints that need to be considered in order to have two units share the same lower unit. The first one is that the DI must permit it. For this all graphs need to be considered that express the constraints for the evolved ACs (e.g. $u^1_{ac_4}$ is transitively used by all tenants and it deploys units from dl_2 and dl_3. Thus, sharing must be permitted between all tenants in the graphs of ac_2 and ac_3 on dl_4 - this is true since they are complete). The second requirement that needs to be fulfilled is that units of different ACs may only be hosted by the same lower unit if they share the entire lower DL stack. Based on the example, it is possible to describe why this is necessary. If this constraint was not considered, it would be possible to find a set of dl_2 units that host all dl_1 units (one for $u^1_{ac_1}, u^2_{ac_1}, u^2_{ac_2}$ and one for $u^1_{ac_2}, u^3_{ac_2}$). However, these units would not be deployable on the Levels dl_2 and dl_3 since they deploy instances of both ac_1 and ac_2. This would lead to an invalid deployment.

All this that has just been discussed needs to be fulfilled in order to have a valid deployment. However, in order to have a valid and optimal deployment the deployment needs to cause only minimal cost.

3.2 Analysis of the Deployment Problem

The first analysis that can easily be done is determining the complexity of the deployment problem. This can be done by reducing the clique cover problem to the problem at hand. The clique cover problem is an NP-hard graph theory problem whose goal is to split a given graph into a minimal number of disjoint cliques (complete subgraphs) [Kar72]. For the problem reduction, let's consider a problem instance of the deployment problem where there exists only one AC and one DL. For this, the DI would consist of only one graph. A valid and optimal deployment would require a minimal number of units that each hosts a set of tenants that are a clique in the graph (may share with each other). Therefore, the problem is equivalent to clique cover and, thus, is NP-hard.

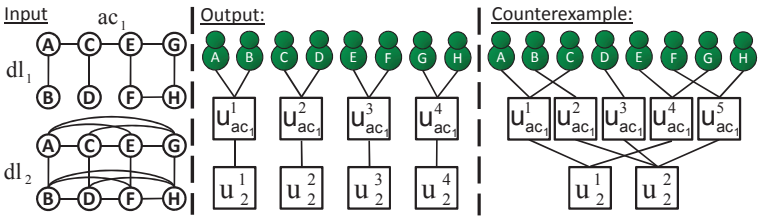


Figure 4: Example not utilizing a minimal Clique Cover on highest Level

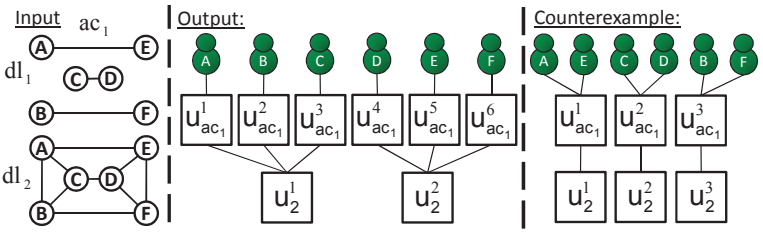


Figure 5: Example not utilizing a minimal Clique Cover on lowest Level

Based on the knowledge that clique cover is related to the deployment problem, an obvious assumption would be that finding a minimal clique cover on all DLs will allow computing a valid and optimal deployment. However, there are counter examples for this assumption.

Assuming that the cost of each instance is one, Figure 4 illustrates an example where a deployment utilizing a non-minimal clique cover on the highest DL causes less cost than one utilizing a minimal clique cover. Figure 5 illustrates the same for the lowest DL. For this example, choosing a minimal clique cover on the lowest DL will increase the cost. Furthermore, when combining the two examples, it is possible to create an example where a valid and optimal deployment will not utilize a minimal clique cover at all.

3.3 Introduction of Intuitive Algorithms

The conclusion gained from the previous subsection is that computing a minimal clique cover will not always allow computing a valid and optimal deployment. However, there are many algorithms available to approximate a valid and optimal deployment in polynomial time. Thus, we developed the following two intuitive algorithms: *Top-Down* – A valid deployment is approximating, by approximating a minimal clique cover for the highest DL. Based on this, the algorithm works its way down, always approximating a minimal clique cover by considering the higher clique cover. It was possible to prove that the relative performance guarantee of this approach is not better than 2. *Bottom-Up* – A valid deployment is approximating the other way around starting from bottom working its way up. For this approach it was possible to prove that its relative performance guarantee is not better than $3/2$. Details on both algorithms may be found in [LMRV14] and [Rue14].

In addition to the theoretical results, we also conducted an experimental evaluation. For this a total of 2.332.800 problem instances were randomly generated and both algorithms were used to compute a valid and optimal deployment. For all problem instances it was assumed that they follow the structure illustrated by Figure 3 (having five DLs and three ACs). However, each problem instance contained 50 tenants. Their deployment constraints were created randomly[3]. The experiments show that for about 67% of all problem instances the top-down approach outperforms the bottom-up approach (with an average improvement of 2.4%). In about 17% of the cases the results were equal. Thus, results suggest to prefer the top-down approach.

So far it has been assumed that all units cause cost of one. However, since this is probably not applicable in reality, we also created an algorithm that is capable of computing a valid deployment in any given order of DLs (for example from the most to the least expensive). Results for this approach are very promising and suggest that this approach may be beneficial. A detailed analysis of that is, however, still open.

4 Case Study: ERP-System as Mixed-Tenancy Cloud Service

The last research question introduced in the introduction of this paper was RQ-3. It deals with analyzing real-world applicability of the MT approach. In order to investigate this, a case study was conducted where MT has been introduced to an existing real-world application.

4.1 Deployment Platform

The first step towards deploying an existing application is to do the conceptual design of a deployment platform (DP). This platform has basically two primary jobs. The first challenge is deploying an application according a computed deployment. Realizing a platform that is capable of doing this for a specific application is straight forward using tools like Chef[4]. The second challenge is establishing communication between different instances of ACs. This is more difficult since it requires first of all that communication is decentralized and that platform logic is separated. *Decentralized Communication* – The goal of MT is to allow customers to express deployment constraints that restrict with whom they share resources. Following this idea it is necessary that the constraints are also enforced for the communication between AC instances. This means that tenants that do not wish to share resources shall also not be able share the same platform components that establish communication. This may be realized by creating a communication mechanism that is distributed without having a central system used by all tenants. *Separation of Platform Logic* – It is the aim of research question RQ-3 to evaluate if the MT approach is applicable to existing applications. Thus, the DP must have the ability to deploy existing applications according

[3] Details of the experiments may be found in [Rue14].

[4] `http://www.getchef.com/`

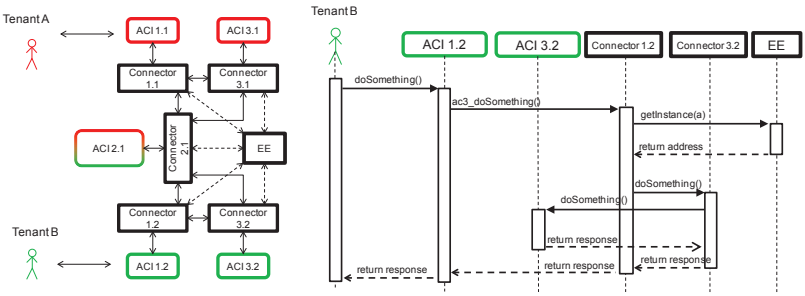


Figure 6: Overview of the Connector Pattern

to the MT paradigm. In order to have minimal effort needed to migrate an existing application onto the DP, it would be beneficial if the existing application's source code would not have to be altered. This goal can by achieved by separating the application logic from the platform logic.

In order to tackle this challenge, a communication pattern was created that is illustrated by Figure 6. Its basic idea is to realize the separation of platform logic by delegating all communication of an instance through an adapter. Depending of the tenant that is triggering communication, this adapter is capable of directing a request to the right instance.

4.2 Case Study: OpenERP as Mixed-Tenancy Application

In order to evaluate applicability of MT to real-world applications we analyzed applying it to the open source ERP-system OpenERP. The MT approach requires that ACs that compose an application may be deployed separately. This is due to the fact that there may be multiple instances of the same AC that are used by different tenants. Unfortunately, we had to find out that OpenERP's architecture does not allow deploying individual modules separately. Thus, it was necessary to split OpenERP into ACs that are separately deployable. This increased the resource demand of the application significantly. However, based on this version of OpenERP, it was possible to implement a DP that allows deploying and using OpenERP following the MT approach realizing the concepts introduced in the previous subsection.

Further, we evaluated that the deployment of OpenERP following MT behaves exactly like a regular deployed version by creating 151 test cases. Based on these, it can be concluded that it is possible to successfully introduce the MT approach to existing real-world applications.

As stated before, for OpenERP it was necessary to alter the architecture. This increased resource demand was so significant that the created MT deployment required more resources that a single-tenancy deployment of the original version of OpenERP. This implies that only those applications may beneficially be used for the MT approach where deploying ACs multiple times causes no or only minimal overhead. For applications where this is

not possible, it is still possible to describe the entire application as one single AC and have customers express constraints for sharing of the entire application. Thus, resource would be used more efficiently than in single-tenancy, any time there is sharing possible[5].

5 Conclusion

This paper summarized the current state of the development of the MT approach. Therefore, it addressed some of the major challenges involved.

It described a description model that allows customers to express their deployment constraints about if or with whom they are willing to share specific ACs on specific DLs. This model was designed to be very generic in order to be reusable for a wide variety of composite multi-tenancy applications. Furthermore, it allows operators to keep their customer base secret. Based on all deployment constraints that are captured, it is possible to extract the aggregation information about which tenants are allowed to share which ACs on which DL. This is the input for the algorithm that computes a valid and optimal deployment. A deployment is valid if it applies to all deployment constraints expressed by customers. Furthermore, it is valid and optimal if it only utilizes minimal cost. It was proven that the problem of computing such a deployment is NP-hard. Thus, it is not possible to do that in polynomial time. However, two intuitive approaches were introduced – top-down and bottom-up. They were analyzed both theoretically and experimentally. The theoretical analysis revealed that they only have a relative performance guarantee not better than 2 and $3/2$. The experimental analysis revealed that in most cases the top-down approach outperforms the bottom-up approach. However, due of the complexity of the problem, it was not possible to evaluate the overall quality of both approaches compared to an optimal solution. The final step undertaken so far was an evaluation of the applicability of the MT approach to existing real-world applications. This was done by performing a case study where the approach was introduced to OpenERP, an open source ERP-System. Based on this case study, it is possible to conclude that there are cases where the MT approach may successfully be applied in real-world.

Furthermore, even though the approach has only been applied to applications, it is possible to apply the approach also to systems that are used to deliver the other cloud computing service models Infrastructure-as-a-Service and Platform-as-a-Service (e.g. customers may choose with which other customers their virtual machines would be allowed to use the same hypervisor). Due to the lower complexity of these cases, both the description model and the deployment computation algorithm may be utilized without having to alter them.

Besides the results created so far, MT still offers many opportunities for future research. One significant opportunity is the creation of a sufficient approach to ensure that tenants are correctly associated to groups. This approach would be much needed because flawed data would results in constraints not being realized correctly (e.g. European company is associated to Asian group). Secondly, additional research can be done on finding a better algorithm to compute a valid and optimal deployment. So far only the two intuitive

[5] A full discussion of the case study may be found in [RRM+14] and [Rue14].

approaches were developed. Furthermore, another big point is that so far only the initial deployment of a MT application has been investigated. However, it is very likely that customers' deployment constraints will change over time and the application's deployment needs to be adapted.

References

[AGI12] M. Almorsy, J. Grundy, and A.S. Ibrahim. TOSSMA: A Tenant-Oriented SaaS Security Management Architecture. In *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, pages 981–988, 2012.

[BAT12] W.J. Brown, V. Anderson, and Qing Tan. Multitenancy - Security Risks and Countermeasures. In *2012 15th International Conference on Network-Based Information Systems (NBiS)*, pages 7–13, 2012.

[CC06] Frederick Chong and Gianpaolo Carraro. Architecture Strategies for Catching the Long Tail. *Microsoft MSDN*, April 2006.

[Clo13] Cloud Security Alliance. The Notorious Nine: Cloud Computing Top Threats in 2013. Technical report, February 2013.

[GHJ94] Erich Gamma, Richard Helm, and Ralph E. Johnson. *Design Patterns. Elements of Reusable Object-Oriented Software.* Addison-Wesley Longman, Amsterdam, 1st ed., reprint. edition, October 1994.

[GSH+07] Chang Jie Guo, Wei Sun, Ying Huang, Zhi Hu Wang, and Bo Gao. A Framework for Native Multi-Tenancy Application Development and Management. In *The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007)*, pages 551–558, Tokyo, Japan, July 2007.

[Kar72] Richard M. Karp. Reducibility among combinatorial problems. Technical report, Springer, 1972.

[LMRV14] Steffen Lange, Marian Margraf, Stefan T. Ruehl, and Stephan A. W. Verclas. On Valid and Optimal Deployments for Mixed-Tenancy Problems in SaaS-Applications. In *2014 IEEE 10th World Congress on Services (Services)*, July 2014.

[RARV12] Stefan T. Ruehl, Urs Andelfinger, Andreas Rausch, and Stephan A. W. Verclas. Toward Realization of Deployment Variability for Software-as-a-Service Applications. In *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, pages 622 –629, June 2012.

[RRM+14] Stefan T. Ruehl, Malte Rupprech, Bjorn Morr, Matthias Reinhardt, and Stephan A. W. Verclas. Mixed-Tenancy in the Wild - Applicability of Mixed-Tenancy for Real-World Enterprise SaaS-Applications. In *2014 IEEE 7th International Conference on Cloud Computing (CLOUD)*, July 2014.

[Rue14] Stefan T. Ruehl. *Mixed-Tenancy Systems - A hybrid Approach between Single and Multi-Tenancy*. PhD thesis, Clausthal University of Technology, 2014.

[RWV13] Stefan T. Ruehl, Holger Wache, and Stephan A. W. Verclas. Capturing Customers' Requirements towards Mixed-tenancy Deployments of SaaS-Applications. In *2013 IEEE 6th International Conference on Cloud Computing (CLOUD)*, pages 462–470, June 2013.

Management komplexer IT-Systeme und Anwendungen

(MITA)

Management komplexer IT-Systeme und Anwendungen (MITA 2014)

Wir begrüßen Sie zum Workshop „Management komplexer IT-Systeme und Anwendungen“ der GI/ITG-Fachgruppe „Kommunikation und Verteilte Systeme (KuVS)“ im Rahmen der 44. Jahrestagung der Gesellschaft für Informatik.

Das IT Management hat sich in den letzten Jahren von einem Management der rechentechnischen Infrastruktur, meistens dem Netzwerk, hin zu einem Service Management zur Unterstützung kritischer, verteilter Geschäftsprozesse in Unternehmen sowie der Verwaltung und produktionsnaher Prozesse entwickelt. IT-Umgebungen werden dabei durch Virtualisierung (Cloud Computing), durch Einbezug privater mobiler Endgeräte (Bring Your Own Device: BYOD), durch Verarbeitung großer Datenmengen in Echtzeit (Big Data), durch eine horizontale und vertikale Integration in produktionsnahen Umgebungen (Industrie 4.0, Internet-of-Things: IoT) wie auch durch eine zunehmende Hausautomation im privaten Bereich (Smart Home, Ambient Assisted Living) immer komplexer sowie in ihren Ursachen- und Wirkungszusammenhängen immer weniger durchschaubar. Dabei müssen unter anderem Dienstgütemerkmale bei möglichst geringem Ressourceneinsatz zugesichert werden, eine Automatisierung von Management-Funktionen zur Effizienzsteigerung und eine höhere Kosten- und Wertschöpfungstransparenz der IT-Dienste erreicht sowie Schutzziele eingehalten werden.

Der Workshop MITA 2014 bietet Interessenten aus Praxis und Forschung ein Forum zur Präsentation und Diskussion aktueller Probleme und Lösungsansätze, welche im Besonderen auf neue Herausforderungen des IT Management fokussieren.

Die Organisatoren danken allen Autoren für ihre Einreichungen, aus denen der Programmausschuss basierend auf jeweils drei Gutachten das vorliegende Programm zusammengestellt hat. Damit sind die Organisatoren den Gutachtern für ihre zeitlich eng bemessene Arbeit dankbar und sind sich sicher, dass der Workshop eine ausgezeichnete Gelegenheit darstellt, die Thematik in hinreichender Tiefe und Breite anzusprechen, zu diskutieren und neue Kontakte zu knüpfen.

Organisation: Reinhold Kröger (Hochschule RheinMain, Wiesbaden), Kurt Geihs (Universität Kassel), Burkhard Stiller (Universität Zürich)

Programmausschuss: Gabi Dreo Rodosek (Universität der Bundeswehr München), Kurt Geihs (Universität Kassel), Wolfgang Johannsen (it's okay, Bensheim), Odej Kao (Technische Universität Berlin), Samuel Kounev (Universität Würzburg), Reinhold Kröger (Hochschule RheinMain, Wiesbaden), Bernhard Neumair (Karlsruhe Institute of Technology), Helmut Reiser (Leibniz-Rechenzentrum München), Markus Schmid (T-Systems International GmbH, Darmstadt), Burkhard Stiller (Universität Zürich)

-

Konzepte der Information-Governance

- Der „Missing Link“ in der Governance of Enterprise IT -

Wolfgang Johannsen

It’s Okay - Governance in Practice
Wilhelm Busch Str. 23
64625 Bensheim
johannsen@its-okay.com , www.eaog.eu

Abstract: Mit der umfassenden Berücksichtigung der Informationen und der Daten einer Organisation als Gegenstand des strategischen, methodischen IT-Managements verändert sich die Perspektive auf die IT-Governance (auch als „Governance of Enterprise IT“ bezeichnet) von einer eher technologiezentrischen hin zu einer informationszentrischen Sicht. Nicht zuletzt die jüngere Auseinandersetzung mit den Wirkungen großangelegter Datenkonzentrationen (Big Data) und ihren Auswirkungen geben diesem Aspekt eine aktuelle Bedeutung.

Der Beitrag zeigt die Notwendigkeit für eine „Information-Governance“ als Komplement der „IT-Governance“ auf. Während IT-Governance den klassischen Fokus des IT-Managements als Gegenstand hat, erweitert Information-Governance den methodischen Rahmen um Daten und Informationen. Dabei finden unter einer informationszentrischen Perspektive syntaktische, semantische und sicherheitsrelevante Aspekte besondere Berücksichtigung. Am Beispiel einer neuen Erweiterung des COBIT 5 - Frameworks wird auf Anwendungsaspekte eingegangen.

1 Übersicht und Einordnung

Informationen und der Umgang mit ihnen stehen heute im Zentrum gesellschaftlicher Kritik. Dies haben u.a. die spektakulären Abhörskandale in jüngster Zeit deutlich gezeigt. Darüber hinaus wächst auch die gesellschaftliche Sensibilität im Umgang mit Informationen und Daten[1]. Dies gilt für soziale Netzwerke ebenso wie für die intensivierte Konsolidierung und Analyse von (Kunden-) Informationen durch

[1] Zur Unterscheidung: Informationen werden im Folgenden als Daten im Kontext verstanden (s.u.)

Wirtschaftsunternehmen. Letztere haben im Zuge der Konsolidierung großer Datenmengen („Big Data"), durch Vernetzung mit ihren Kunden und mit Hilfe hochinteraktiver Schnittstellen den Rohstoff für neue Analyseverfahren in die Hand bekommen. Dies ermöglicht ihnen beispielsweise, Produktdesign und Marketing auf immer kleinere Kundensegmente hin spezifisch auszurichten.

In der Konsequenz ergibt sich sowohl aus dem Bedarf nach Sicherung vor Datenmissbrauch als auch aus dem wirtschaftlichen Umgang mit bisher nicht gekannten Datenmengen die Notwendigkeit einer genaueren Betrachtung der Eigenschaften und Strukturen von Informationen. Diese Art der Betrachtung bildet die Voraussetzung für ein effektiveres und effizienteres Informationsmanagement. Sie bildet auch die zwingende Antwort auf eine zu erwartende, verschärfte Regulierung des Umgangs mit persönlichen Daten seitens staatlicher Instanzen. Frameworks zur Bewältigung dieser komplexen Situation sollen einen methodischen Rahmen für das Management der Informationen bieten. Dafür müssen sie sowohl als Instrument des Managements zur Leistungsoptimierung als auch zur Absicherung der Compliance im regulatorischen Umfeld[2] einsetzbar sein.

Im Folgenden wird zunächst die Notwendigkeit eines methodisch umfassenden Umgangs auch mit unstrukturierten Informationen verdeutlicht. Anschließend wird eine Einordnung der verwendeten Begriffe vorgenommen, die um eine Abgrenzung zwischen „Information-Governance" und „IT-Governance" ergänzt wird.

Im zweiten Teil wird gezeigt und diskutiert, wie Frameworks bei den Aufgaben des Informationsmanagements helfen können. Ausgewählte Ansätze hierfür werden vorgestellt. Eine vor kurzem veröffentlichte Erweiterung des COBIT 5 - Frameworks [IS12a] [GJ13] wird daran anschließend diskutiert.

2 Information-Governance - ein Perspektivenwechsel

Die in Wirtschaft und Verwaltung zu spürenden wachsenden Schwierigkeiten im Umgang mit Informationen und Daten sind nur zum Teil mit dem Phänomen „Big Data" in direkten Zusammenhang zu bringen. Sie werden sowohl aufgrund der immer intensiveren Nutzung vernetzter kommerzieller Systeme als auch aufgrund der Verwendung sozialer Netze weiter verstärkt. Beide wirken als Beschleuniger der schnell zunehmenden Datenkonsolidierung. Insbesondere die explosionsartige Zunahme von unstrukturierten und/oder sensitiven Informationen und Daten verlangt neue Antworten hinsichtlich der Effektivität des Informationsmanagements wie auch der Compliance. Dies bedeutet neue regulatorische Herausforderungen.

Das Wachstum heutiger (verteilter) Systeme erfolgt weitgehend organisch und unkontrolliert. Die Informationen, die sich in ihnen „aufhalten", sind oftmals

[2] Für die Fähigkeit, dies zu tun, wird seit kurzem auch der Begriff „Datability" verwendet.

unstrukturiert (z.B. Soziale Netze, Chats, Blogs etc.), häufig semi-strukturiert (z.B. Emails) und zu einem abnehmenden Anteil strukturiert (z.B. Datenbanken, Data Warehouses). Insgesamt ist eine Situation herangereift, in der zwar das Zusammenspiel der Technologiekomponenten möglich ist, nicht aber das strukturierte Management sensitiver Informationen - wie auch immer die Kriterien für diese festgelegt sein mögen.

Zwei alternative Ansätze, der Ausbau traditioneller Managementtechniken oder ein Perspektivenwechsel hin zu einer informationszentrischen Information-Governance, zeichnen sich als Entwicklungslinien künftiger Management-Disziplinen ab. Die erste Variante bleibt technologiezentriert und misst Informationen und Daten eher sekundäre Bedeutung bei. Mit der zweiten Variante hingegen werden Informationen nicht mehr als eine Art Seiteneffekt der Technologie, sondern als der zentrale Gegenstand der Governance selbst gesehen. Es wird also eine informationszentrische Perspektive [IS13] eingenommen. Darüber hinaus werden die Grundlagen für einen neuen Umgang mit Informationen im wirtschaftlichen Kontext gelegt.

Die gegenwärtige Situation lässt das Fehlen einer Information-Governance als Teil des Informationsmanagements deutlich zu Tage treten. Dem Wunsch insbesondere auf der Nutzer-/Verbraucherseite nach einem Umgang mit Informationen, der ihre Diffusion, ihre Replikation und ihren gesamten Life Cycle steuerbar macht, kann nicht entsprochen werden, solange es an allgemein akzeptierten Informationsmodellen im Informationsmanagement mangelt. Erst eine klare Strukturierung mit methodisch fundierten und holistischen Managementtechniken (Information-Governance und IT-Governance) wird es künftig ermöglichen, Daten und Informationen so zu handhaben, dass Verbraucherschutz und betriebswirtschaftliche Interessen wieder in eine Balance miteinander kommen.

2.1 Der strategische Bedarf an Information-Governance

Die Notwendigkeit einer Information-Governance ist heute durch die bestehenden dynamischen Anforderungen gegeben. Wie bereits in [Jo10] ausgeführt, kann zunächst festgehalten werden, dass Information-Governance sich mit den strategisch orientierten Fragestellungen im Zusammenhang mit dem Informationsmanagement zu befassen hat. Dies soll ganzheitlich unter Einbezug von Prozessen, Menschen, Technologien und Strategien erfolgen und sich auf große Datenmengen konzentrieren. Dabei stehen u.a. die folgenden Zielsetzungen im Mittelpunkt:

- Den strategischen Wert der Informationen erhöhen
- Informationen in semantische und syntaktische Aspekte kategorisieren
- Das Unternehmen/die Organisation auf erhöhte Compliance-Anforderungen vorbereiten
- Qualitätskriterien für Informationen festlegen, überwachen und steuern
- Vervielfältigungen und Migrationswege von Informationen verfolgen
- Schutzrechte (Intellectual Property Rights, IPR) sichern
- Integrität und Verfügbarkeit sichern
- Informationsfluss bei diversifizierter Mediennutzung steuern

- Priorisierung und Bewertung der Informationen vornehmen
- Life Cycle Management der Informationen durchführen

Selbstverständlich werden die bekannten Methoden der IT-Governance [JG10] und des IT-Managements sowie die dabei verwendeten Methoden, Prozesse und Referenzmodelle ebenso wenig überflüssig wie Datenbanken, Data Warehouses und ECM (Enterprise Content Management) - Systeme. Die erweiterten Anforderungen einer Information-Governance zwingen jedoch zu deren Überprüfung und ggfs. Erweiterung und Ergänzung.

2.1 Begriffsdefinition und Abgrenzung

Mit "Enabling Information" [IS13], einer Ergänzung zum COBIT 5 Framework, liegt erstmals ein in IT-Governance integriertes Framework für Information-Governance vor. In Anlehnung an das Framework DMBOK (Data Management Body of Knowledge) [DM09] wird darin „Information-Governance" als die Bewältigung folgender Aufgabenstellungen definiert:

- **Analyse und Bewertung der Anforderungen:** die Anforderungen, Bedingungen und Optionen der Stakeholder (Anteilseigner, Mitarbeiter, Geschäftspartner, …) werden ausgewertet, um den Einsatz, den Erwerb und das Management von Informationsressourcen im Sinne von ausgeglichenen und vereinbarten Unternehmenszielen sicherzustellen.

- **Ziele vorgeben:** die Richtung für die Entwicklung von Informations-Management-Fähigkeiten durch Priorisierung und methodische Entscheidungsfindung wird festgelegt.

- **Steuern:** Performance und Compliance anhand vereinbarter Ziele werden überwacht.

Im Unterschied dazu wird „Information-Management" als eher operative Tätigkeit entlang der Life-Cycle - Phasen „Planen", „Entwickeln und Integrieren", „Betreiben und Warten" sowie „Überwachen" verstanden.

Damit folgt die Definition der Information-Governance der Trennung von operativer und strategischer IT-Steuerung im COBIT 5 Framework und bezieht sich auch auf die gleichen Schritte im dort definierten Life-Cycle (vgl. Abbildung 1).

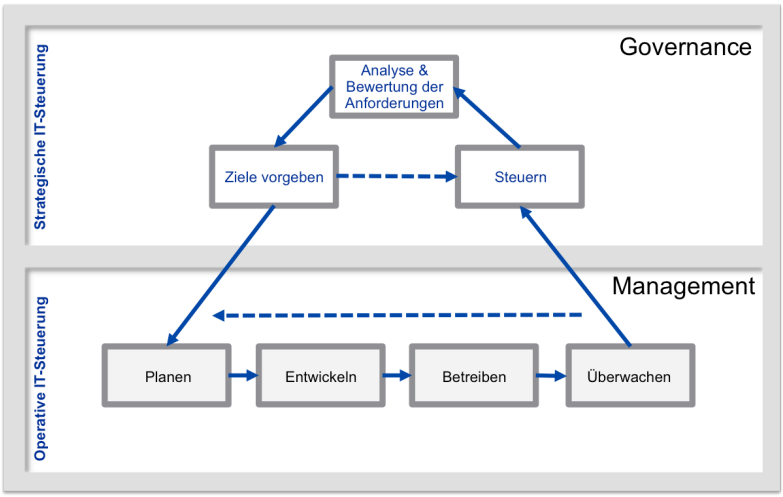


Abbildung 1: Unterscheidung Governance - Management (vgl. [IS12a], [AGM14])

Im Rahmen von COBIT 5 wird in „Enabling Information“ [IS13] auch eine Unterscheidung von Daten und Informationen vorgenommen.

- **Daten:** etwas was selbst ein Fakt ist oder einen Fakt repräsentiert. Daten können vielfältige Formen annehmen (Text, Zahlen, Grafik, Audio, Video, ...)

- **Information:** Informationen sind Daten im Kontext, d.h. im Zusammenhang mit einer Bedeutung. Dies umfasst sowohl das Format, in dem Daten präsentiert werden, als auch die Relevanz hinsichtlich eines bestimmten Nutzenkontextes.

Mit den Definitionen von IT-Governance, Information-Governance sowie von Daten und Informationen ist die Grundlage für die Modellierung von Daten und die zu ihrer Verarbeitung herangezogenen Verarbeitungs- und Governance-Prozesse gelegt.

Aus einem anderen Betrachtungswinkel heraus gesehen ergibt sich eine Teilmengenbeziehung der verschiedenen Managementdisziplinen wie in Abbildung 2 dargestellt. Hier bildet das klassische „IT-Management“ den operativen Kern, während „IT-Governance“ (auch „Governance of Enterprise IT“ [IS12]) durch seine ausschließlich strategisch ausgerichtete Aufgabenstellung davon abgegrenzt ist. Information-Governance umfasst nun die systematische und methodische Governance der Information. Die „Verteilte Information-Governance“ wird hier angeführt, um den ausgeprägten Verteilungsaspekten dezentraler Führungs- und Unternehmensstrukturen (ggfs. über mehrere Rechtsräume hinweg) und den darin komplexer werdenden Compliance-Anforderungen Rechnung zu tragen. Die Gesamtheit aller Governance-Aktivitäten in einem Unternehmen bzw. einer Organisation ist in eine „Corporate Governance Strategie“ [CO12] einzubetten.

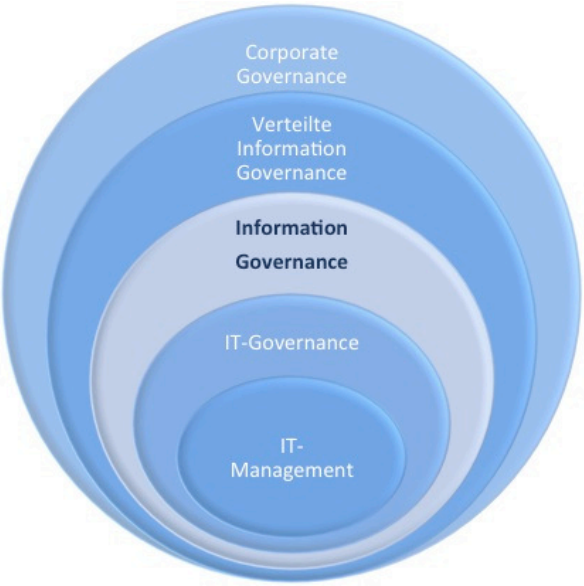


Abbildung 2: Einordnung von Information-Governance

Der Aspekt der Qualität der Informationen ist Gegenstand vielfältiger Untersuchungen und Systementwicklungen. Beispielsweise hat die Deutsche Gesellschaft für Informations- und Datenqualität e.V. (DGIQ) bereits 2007 ein Referenzmodell mit vier Kategorien (Nutzung, Inhalt, Darstellung, System), die in 15 Dimensionen aufgeteilt und beschrieben sind, vorgestellt[3]. Die darin vertretene Modellsicht ist jedoch weitgehend statisch und genügt daher nicht den Anforderungen an zukünftige Governance-Aufgaben.

Auch zur Information-Governance existieren weitere Definitionen:

- Gartner Group: [LO10], [LO13]: Information-Governance umfasst die Festlegung von Entscheidungsbefugnissen und die Festlegung eines „Frameworks“ zur Förderung von gewünschtem Verhalten bei der Bewertung, Erzeugung, Speicherung, Nutzung, Archivierung und Vernichtung von Informationen. Es umfasst Prozesse, Rollen, Metriken und Standards zur Sicherstellung von Effektivität und Effizienz bei der Nutzung von Informationen, die dazu dienen, Ziele zu erreichen, die sich eine Organisation gesetzt hat.

- IBM [IB07]: Data Governance ist die Qualitätskontrolldisziplin für die Bewertung, die Verwaltung, die Nutzung, die Verbesserung, die Überwachung, die Wartung und den Schutz der Informationen einer Organisation.

[3] Vergleiche: www.dgiq.de

- „The 451 Group“ [TG09]: Information-Governance umfasst die Praktiken und Technologien beim proaktiven Management der Einbehaltung, der Speicherung, der Entscheidung über Speicherdauer und bei der Vergabe von Zugriffsrechten und Sicherungsmaßnahmen.

Die Definition der Gartner Group nimmt, im Gegensatz zu den beiden anderen aufgeführten Definitionen, Bezug auf Prozesse, Rollen und Metriken. Im Folgenden werden wir die damit kompatible Definition der ISACA (s.o.) wegen ihrer gleichfalls dynamischen und prozessorientierten Zielrichtung zugrunde legen.

3 Information-Governance im COBIT 5 Framework

Während die Menge der Informationen, die in Unternehmen und Organisationen konsolidiert und verarbeitet werden, mit hoher Beschleunigung zunahm, wurden die Defizite hinsichtlich einer methodischen Governance, also des strategischen Umgangs mit diesen Informationen, deutlicher.

Obwohl die Fachdiskussion in den vergangenen Jahren des Öfteren auf den Bedarf für Information-Governance hinwies [Jo10], [IB07], [Lo10], blieben relevante Standards und Frameworks [IS01], [DD09] unterhalb einer kritischen Akzeptanzschwelle. Dies kann u.a. darauf zurückgeführt werden, dass ein integrierter Ansatz mit der eher technologiezentrierten IT-Governance ausblieb.

„Enabling Information“ [IS13] ist eine Ergänzung des IT-Governance Frameworks COBIT 5 [IS12a]. Es umfasst einen integrierten Ansatz beider oben genannter Managementdisziplinen „IT-Governance“ und „Information-Governance“. Entsprechend der oben eingeführten Definitionen wird darin zwischen „Information-Governance“ und „Information-Management“ unterschieden.

Information wird in „Enabling Information“ als Ressource aufgefasst, die optimal zur Erreichung der Unternehmensziele einzusetzen ist. Die Ziele ergeben sich aus den Ansprüchen diverser „Stakeholder“ (Anteilseigner, Mitarbeiter, Regulierung etc.). Stakeholder-Gruppen sind im COBIT 5 Framework Ausgangspunkt einer Zielhierarchie, die auf den folgenden Ebenen Unternehmensziele und dann daraus abgeleitete IT-Ziele umfasst.

In „Enabling Information“ werden die generischen Unternehmensziele in gleichfalls generische Qualitätskriterien für Informationen abgebildet. Diese Kriterien können dann fallweise herangezogen werden, um die konkreten Informationen eines Unternehmens in den Governance-Prozess zu integrieren. Das Modell umfasst prinzipiell alle Typen von Informationen in einem Unternehmen: interne Informationen der Geschäftsbereiche und der IT-Funktionen ebenso wie externe Markt-, Kunden- oder Lieferanteninformationen.

3.1 Struktur und Komponenten

Das Informationsmodell in „Enabling Information“ wird in [IS13] Top Down entwickelt. Aus einer Menge von Stakeholder-Anforderungen werden generische Zielsetzungen im Kontext von Nutzen/Gewinn-Realisierung, Risikooptimierung und Ressourcen-optimierung abgeleitet und in Form einer Balanced Scorecard als generische Unternehmensziele erfasst. Auf dieser Grundlage werden, unter Berücksichtigung eines Informationsfluss-Modells von den Stakeholdern über Governance- und Managementinstanzen bis hin zur Ausführungsebene, generische IT-Ziele bestimmt und gleichfalls als Balanced Scorecard dargestellt.

Dimensionen des Informationsmodells in „Enabling Information“ / COBIT 5:

Informations-Qualitätsziele: Die Qualitätskriterien für Informationen bilden einen strukturellen Kern im COBIT 5 Informationsmodell für Information-Governance. Es unterscheidet drei Qualitätskriterien: Intrinsische Kriterien, Kontextkriterien und Sicherheits-/Zugriffskriterien:

- Intrinsische Kriterien: Korrektheit, Objektivität, Glaubwürdigkeit, Reputation
- Kontextkriterien: Relevanz, Vollständigkeit, Aktualität, angemessene Menge, kompakte und konsistente Repräsentation, Interpretier- und Verstehbarkeit, Manipulierbarkeit
- Sicherheits-/Zugriffskriterien: Verfügbarkeit, Zugriffsbeschränkungen

Informations-Lebenszyklus: Damit der gesamte Lebenszyklus der Information im Informationsmodell repräsentiert werden kann, wird zwischen den einzelnen Phasen dieses Zyklus (planen, entwickeln, betreiben, erstellen, überwachen und entsorgen) und den IT-Prozessen des COBIT 5 IT-Governance Frameworks [IS12b] eine Beziehung hergestellt bzw. das Informationsmodell in COBIT 5 integriert.

Informationsattribute: Es existiert keine einheitliche, kontextunabhängige Definition für den Begriff Information. Um jedoch einen Anwendungskontext für die Praxis näher spezifizieren zu können, werden in „Enabling Information“ Informationsattribute definiert. Dies erfolgt in sechs Ebenen mit den folgend aufgelisteten Attributen:

Ebene 1: *Physische Welt Ebene*
- physische Träger wie Papier, elektrische Signale etc.

Ebene 2: *Empirische Ebene*
- *Zugriffskanal, Interface etc.*

Ebene 3: *Syntaktische Ebene*
- *Sprache, Regeln zur Kombination von Symbolen*

Ebene 4: *Semantische Ebene*
- *Informationstyp: Finanz vs. non Finanz, intern vs. extern etc.*
- *Informationsaktualität: Zeithorizont, Vergangenheit, Gegenwart*
- *Informationsebene: Detailliertheit (Verkauf p.a. vs. monatlich) etc.*

Ebene 5: *Pragmatische Ebene*
- *Aufbewahrungspflichten: Mindestdauer der Speicherung*
- *Status: operative vs. historische Information*
- *Neuigkeitswert: neues Wissens vs. bekanntes Wissen etc.*
- *Bedingtheit: Abhängigkeit von anderen Informationen*

Ebene 6: *Soziale Welt Ebene*
- *Kontexte, in der eine Information Sinn ergibt*

Metriken: Für jedes generisches IT - Ziel (bspw. Alignment von IT- und Geschäftsstrategie, Kostentransparenz, IT-Agilität) aus COBIT 5 werden indikativ Informationselemente (Unternehmensstrategie, Compliance-Anforderungen, Budgetdaten, Service-Portfolio etc.) angegeben und in Beziehung zu Informations-Qualitätskriterien gesetzt. Zu jeweils einem Kriterium oder zu mehreren, in einer Gruppe zusammengefassten Qualitätskriterien werden exemplarische Metriken aufgeführt. Die so entstehende Referenzliste erlaubt es, ein konkretes Informationsmodell eines Unternehmens mit messbaren Größen zu füllen.

3.2 Anwendung in unterschiedlichen Kontexten

Die Anwendung des skizierten Informationsmodells für Unternehmensdaten kann innerhalb des COBIT 5 Frameworks oder auch unabhängig davon erfolgen.

„Enabling Information" [IS13] ist als Ergänzung zum COBIT 5 - Framework konzipiert und hat als Strukturelement die Funktion eines „Governance-Ermöglichers". Dazu gehören Prinzipien/Leitlinien, Prozesse, Organisationsstrukturen, Kultur/Ethik, Services/Infrastruktur und Mitarbeiter/Fähigkeiten. Für das Framework als Ganzes ist die Integrationsfähigkeit der diversen Enabler ein entscheidender Erfolgsfaktor. Zur Reduzierung der Anwendungskomplexität kann jedoch - bspw. als ein erster Schritt - ein nicht integrierter Einsatz vorgezogen werden (vgl. Beispiel 1).

Beispiel 1:

Als Beispiel für die Bewältigung einer eingeschränkten, illustrativen Governance-Aufgabe wird im Folgenden die Beschreibung der Risiken, die mit der Handhabung von Unternehmensdaten entstehen, herangezogen. Die entstehenden Risiken können mit Hilfe des Informationsmodells entlang der folgenden Komponenten (s.o.) beschrieben werden:

Informations-Qualitätsziele: Für jedes Kriterium wird eine Zielsetzung für das Risikoprofil angegeben.

Informations-Lebenszyklus: Für jeden Schritt im Lebenszyklus werden die internen und externen Stakeholder am entstehenden Risikoprofil benannt und ihr Interesse an dem jeweiligen Schritt charakterisiert.

Informationsattribute: Zu jedem Attribut wird sein Wert im zu erstellenden Risikoprofil definiert. Bspw. kann das Attribut „Semantische Ebene -

Informationsaktualität“ verdeutlichen, dass das Risikoprofil historische, aktuelle und prognostische Daten enthält.

Es wird deutlich, dass Qualitätsziele, Attribute und Schritte im Lebenszyklus der Information Dimensionen eines Datenraumes bilden (vgl. Abbildung 3). Die „Raumpunkte“ wiederum können durch Inputs und Outputs aus anderen COBIT 5 Ressourcen, wie bspw. den IT Prozessen, in Verbindung gebracht werden.

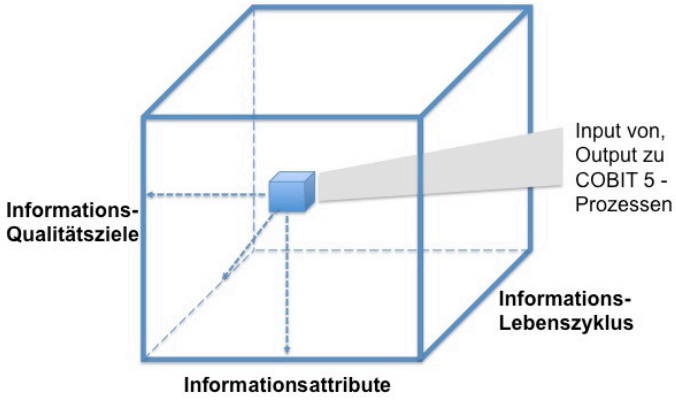


Abbildung 3: Datenraum bestehend aus Qualitätszielen, Attributen und Phasen im Lebenszyklus der Information

Für eine Reihe von typischen Management-Herausforderungen resp. Anwendungskontexten sind darüber hinaus Governance-Profile skizziert, die zwar generisch bleiben, jedoch bereits eine Fokussierung auf praktisch relevante Fragestellungen bedeuten. Die so behandelten Management-Herausforderungen sind: Nachfrage / Nutzung der Information, Big Data - Marketing, Big Data - Betrugsentdeckung, Big Data - Prognostische Analyse, Master Data Management, End User Computing, Disintermediation (suboptimale Informationsumschichtung), Regulation/Compliance, Datenschutz.

Für jede dieser Management-Herausforderungen wird eine strukturierte Handlungsleitlinie angeboten, die sich neben den hier behandelten Informationsaspekten der Governance auf das engere COBIT 5 Framework und insbesondere auf die darin enthaltenen IT-Prozesse beziehen. Die folgenden Punkte geben die Struktur dieser Leitlinie wieder:

- **Geschäftlicher Kontext:** Beschreibung der Herausforderung und Charakterisierung des geschäftlichen Kontextes, in der diese zu finden ist.

- **Betroffene Information:** Beschreibung der Informationskomponente, die zur Behandlung der Aufgabenstellung herangezogen wird (bspw. Marketing Report, Produktstrategie) und die Qualitätskriterien, die an diese angelegt werden sollen (bspw. Relevanz, Korrektheit, Verfügbarkeit, Aktualität,

Glaubwürdigkeit). Des Weiteren das Qualitätsziel, das mit diesen Kriterien verbunden wird (bspw.: die Entscheidungen des Unternehmens beruhen auf einer korrekten Marktanalyse).

- **Betroffene Unternehmensziele und IT-Ziele:** Unternehmens- und IT-Ziele, die in den dazugehörigen Balanced Scorecards des COBIT 5 Frameworks aufgeführt sind, werden zu den jeweils behandelten Management-Herausforderungen in Beziehung gesetzt.

- **Umsetzung:** Relevante „Enabler" (Ressourcen im COBIT 5 Framework), die eine Governance ermöglichen, sind als Hilfestellung zur Bewältigung der Management-Herausforderungen aufgeführt. Dies können bspw. Governance-Prozesse, organisatorische Strukturen oder kulturell/ethische Aspekte sein.

4 Schlussbemerkungen

Die informatorische Revolution, die durch das sehr schnelle Anwachsen von Informationsbeständen und dem hohen Grad an Vernetzung spürbar wird und weiter voranschreitet, hat zu einer Verschiebung des Fokus im strategischen IT-Management geführt. In ihr kommt die Einordnung von Information als zugleich zentraler Vermögenswert und kritischer Risikofaktor zum Ausdruck. Informationen bilden entlang der gesamten Wertschöpfungsketten heutiger Unternehmen den Schlüssel zur Produkt- und Serviceinnovation ebenso wie zur Erschließung neuer Märkte bzw. neuer Kundengruppen. Die zutage getretenen Sicherheitslücken der jüngeren Vergangenheit zeigten jedoch auch die Verletzbarkeit, die durch ihre wachsenden Informationsbestände den Unternehmen und Organisationen entsteht.

Über die rein technischen Sicherheitsvorkehrungen hinaus sind zunehmend informationsspezifische Maßnahmen gefragt. Diese umfassen gleichermaßen ein vertieftes Verständnis für die jeweils genutzten Informationen wie auch eine Methodik zur Modellierung dieser Informationen. Als drittes Element ist die Einbettung einer „Governance von Informationen" in eine „Governance von Technologie" vonnöten.

Ein Teil dieser Entwicklungen wurde in Fachkreisen seit einigen Jahren antizipiert. Die einsetzende Diskussion führte zu einer erweiterten, informationszentrischen Sicht auf die Aufgaben der IT-Governance, in der Information als Gegenstand der Governance zur Technologie Governance hinzugefügt wurde.

Mit „Enabling Information" [IS13] als Erweiterung des COBIT 5 Frameworks [IS12a] liegt nunmehr ein Best Practice - Ansatz vor, der den genannten Anforderungen gerecht wird. Dieser umfasst nicht nur ein Informationsmodell mit einer Strukturierung nach Qualitätskriterien und Attributen sowie eine Einordnung in ein Life-Cycle - Modell, sondern auch die Integration in das bestehende Framework.

Literaturverzeichnis

[AGM14] Andelfinger, U., Graf, A., Manneck, F-M.: Compliance und Performance verbinden - ein (strategie-)prozessorientierter Ansatz für die IT-Governance in Kreditinstituten. IT-Governance, dpunkt-Verlag, Heft 17, Heidelberg, 2014.

[CO13] Internal Control — Integrated Framework Executive Summary. COSO - Committee of Sponsoring Organizations of the Treadway Commission, Durham, North Carolina, USA, 2013.

[DM09] The DAMA Guide to Data Management Body of Knowledge (DAMA-DMBOK Guide), Ausgabe. Technics Publications, Bradley Beach, New Jersey, USA, 2009.

[GJ13] Goeken, M., Johannsen, W: COBIT 5 - Grundlagen, Anwendung und Toolimplementierung, 2013, in: Lang, M. (Hrsg.): CIO-Handbuch - Band 2, Symposion Publishing, Ettlingen, 2013.

[IB07] The IBM Data Governance Council Maturity Model: Building a roadmap for effective data governance, IBM Software Group, Somers, New York, USA, 2007.

[IS01] ISO 15489:2001 Information and Documentation - Records Management. Genf, Schweiz, 2001.

[IS12a] COBIT 5, Rahmenwerk für Governance und Management der Unternehmens-IT, ISACA, Rolling Meadows, USA, 2012.

[IS12b] COBIT 5, Enabling Processes, ISACA, Rolling Meadows, USA, 2012.

[IS13] COBIT 5, Enabling Information, ISACA, Rolling Meadows, USA, 2013.

[Jo10] Johannsen, W.: Information-Governance - Herausforderungen in verteilten Umgebungen. In Klaus-Peter Fähnrich & Bogdan Franczyk, ed., 'GI Jahrestagung (2)' , GI, S. 311-316, 2010.

[JG10] Johannsen, W; Goeken, M.: Referenzmodelle für die IT-Governance. dpunkt-Verlag, (2. Auflage), Heidelberg, 2010.

[Lo10] Logan, D.: What is Information Governance? And Why is it So Hard? Gartner Group, 2010.

[Lo 13] Logan, D.: Information Governance Best Practice: Adopt a Use Case Approach. Gartner Group, 2013.

[TG09] Reidy, K.: The Rise of Information Governance, the 451 group, London, UK, 2009.

Ein Ansatz für eine Ontologie-basierte Verbindung von IT Monitoring und IT Governance

Andreas Textor

Labor für Verteilte Systeme
Hochschule RheinMain
Unter den Eichen 5
65195 Wiesbaden
andreas.textor@hs-rm.de

Abstract: IT Governance hat zum Ziel, dass die Unternehmens-IT die Strategie und Geschäftsziele des Unternehmens möglichst effizient unterstützt. Um dies umzusetzen, existieren mehrere international anerkannte Verfahren und Standards wie z.B. COBIT, ITIL oder ISO 20000. Diese sind jedoch konzeptionell gehalten und machen meist keine Vorgaben zur technischen Umsetzung. Dieser Beitrag untersucht, wie mit Hilfe von Ontologien die Lücke zwischen solchen abstrakten Konzepten der Geschäftssicht und technisch erfassbaren Messwerten des IT-Monitorings geschlossen werden kann, konkret am Beispiel COBIT 5. Dies ermöglicht eine Top-Down-Sicht von Geschäftszielen über Services bis hin zu den Komponenten der IT, die zur Erbringung der Ziele beitragen. Hierdurch wird die Transparenz (u.a. über Kosten) erhöht. Außerdem ermöglicht eine solche Verbindung der Abstraktionsebenen eine Automatisierung der IT Governance sowie des IT Service Managements durch Regeln, die auf Konzepte der unterschiedlichen Ebenen Bezug nehmen können.

1 Einleitung

Es ist nicht nur essentiell für ein Unternehmen, eine funktionierende IT-Infrastruktur zu haben, um das Kerngeschäft zu unterstützten, sondern es ist ebenso wichtig für das Unternehmen zu wissen, wie diese IT-Infrastruktur funktioniert und wie sie genau zu den Geschäftszielen beiträgt. Die Erstellung und Verwaltung von Prozessen und Strukturen zur Sicherstellung der gewünschten Funktionsweise der Unternehmens-IT wird als IT Governance bezeichnet. IT Governance ist Teil des Unternehmensmanagements und betrachtet daher die IT auch aus dieser Perspektive. Sie umfasst nach [JG07] Business-IT-Alignment, Compliance, Erfolgsmessung, Ressourcenmanagement und Risikomanagement. Dabei wird der Fokus auf den Wertbeitrag der IT (Value Delivery) gelegt. Als Mittel zur Messung des Erfolgs der IT Governance werden häufig Metriken definiert, z.B. eine Obergrenze von Kosten der IT-Infrastruktur.

Für IT Governance und das verwandte IT Service Management existieren mehrere international anerkannte Verfahren und Standards. Die bekanntesten sind dabei das IT Governance Framework COBIT (Control Objectives for Information and Related Technology)

der ISACA (Information Systems Audit and Control Association) [Infb], das IT Service Management Framework ITIL (Information Technology Infrastructure Library) [AXE] sowie die international anerkannte und auf ITIL aufbauende IT Service Management Norm ISO/IEC 20000 [Int].

Auf der Seite des technischen Betriebs der IT werden die vom Management vorgegebenen Anforderungen umgesetzt. Für die Sicherstellung der gewünschten Funktionsweise werden hier technische, messbare Metriken im Rahmen des IT-Monitorings überprüft (Monitoring von Kenngrößen der überwachten Systeme, Netzwerke und Anwendungen). Für IT Monitoring existiert eine Vielzahl von Standards, Interfaces und Werkzeugen, die im Wesentlichen zur Laufzeit aktiv oder passiv Informationen von den überwachten Systemen abrufen, aggregieren, visualisieren und bei Ausfällen oder Unterschreitungen z.B. der Antwortzeiten einen Verantwortlichen benachrichtigen. Die beim IT Monitoring überwachten Kenngrößen beinhalten in der Regel keine Informationen darüber, wie der Zusammenhang des überwachten Systems zu den übergeordneten Organisationsstrukturen ist, also z.B. welche Services der jeweilige Server bereitstellt, in welchen IT-Prozessen diese Services verwendet werden, welche Geschäftsprozesse durch diese IT-Prozesse unterstützt werden und letztlich welche Geschäftsziele durch diese Geschäftsprozesse erreicht werden sollen.

Das IT Governance Framework COBIT stellt die Verbindung zwischen diesen Abstraktionsebenen her und beginnt Top-Down auf der Managementebene. Die Zusammenhänge zwischen den Abstraktionsebenen werden dabei aber nur bis hin zu IT-Prozessen und zu deren Beurteilung festgelegten Metriken definiert. Die Verknüpfung zwischen IT-Prozessen und messbaren IT-Monitoring-Daten ist kein Teil des Frameworks, da dies den Rahmen der Ziele der IT Governance aus Management-Sicht sprengt. Diese Verknüpfung ist jedoch notwendig, um drei Ziele zu erreichen:

- Erstens, um die IT in Richtung des Managements nicht weiter als Black Box erscheinen zu lassen. Wenn die Verbindungen zwischen den technischen Systemen, Services und darauf aufbauenden IT- und Geschäftsprozessen auch formal bzw. maschinenlesbar definiert sind, sind Analyse und Reporting möglich, die Fragen über diese Zusammenhänge direkt beantworten können, ohne den Umweg über manuelle Aufbereitung. Beispielsweise kann exakt und automatisch beantwortet werden, welche finanziellen Resourcen nötig sind, um einen bestimmten Geschäftsprozess durch die IT aufrecht zu erhalten.

- Zweitens, um IT Governance-Kennzahlen, die sich direkt aus technischen Messwerten herleiten lassen (z.B. Ausfallrate von Services, die einen bestimmten Geschäftsprozess unterstützen – ggf. mit Grundursachen) auszurechnen und dem Management in geeigneter Form in Echtzeit zur Verfügung zu stellen.

- Drittens, um Ansätze zur Automatisierung, die auf technischer Ebene bereits zum State of the Art gehören (Self-Management bzw. Autonomic Computing) auch auf den höheren Abstraktionsebenen des Unternehmens anwenden zu können. Wird die Brücke zwischen IT Governance und IT Monitoring durch einen entsprechenden Formalismus (Modell und Implementierung) geschlossen, ist es möglich, Regeln für Regelkreise nicht mehr nur Bezug auf technische Größen nehmen zu lassen,

sondern auch auf nicht-technische Entitäten. Beispielsweise könnte nicht mehr nur eine Regel definiert werden, die Services oder Virtuelle Maschinen auf ein anderes System migriert, wenn Prozessor- oder Arbeitsspeicher nicht mehr ausreichen und damit vereinbarte SLAs nicht mehr erfüllt werden können, sondern z.B. auch wenn Kosten gespart werden können, indem Services migriert oder abgeschaltet werden, weil durch Geschäftsprozesse bekannt ist, dass sie nicht in vollem Umfang benötigt werden.

Um eine solche Verknüpfung zwischen den Abstraktionsebenen umsetzen zu können, die diese Ziele erfüllen kann, ist ein formales, d.h. maschinenverarbeitbares Modell nötig, das flexibel genug ist, um die Belange der unterschiedlichen Ebenen ausdrücken zu können. Es muss möglich sein, unterschiedliche Domänen separat zu modellieren und dann Konzepte der Teilmodelle miteinander zu verbinden. Ebenso muss es möglich sein, in Teilmodellen Bezug auf gemeinsam genutzte Terminologien zu nehmen. Diese Anforderungen werden erfüllt von einem Ontologie-basierten Modell. Eine Ontologie ist ein formales Modell, das Entitäten und Relationen zwischen ihnen beschreibt. Durch die Verwendung der Ontologie-Beschreibungssprache OWL (Web Ontology Language), der Abfragesprache SPARQL (SPARQL Protocol and RDF Query Language), bereits verfügbaren Ontologien zu generischen Konzepten wie Einheiten sowie Werkzeugen und Libraries kann ein solches Ontologie-basiertes Modell mittels ausgereifter Technik umgesetzt werden.

Dieser Beitrag stellt einen Ansatz vor, mit dem Ontologie-basiert eine solche Verknüpfung von technischen Informationen (Messwerte des IT-Monitoring sowie technische Informationsmodelle) und nicht-technischen Strukturen (IT Governance, IT Service Management) erzielt werden kann. In Abschnitt 2 werden zunächst die Grundlagen des IT Governance Frameworks COBIT und des im weiteren Verlauf verwendeten Ontologie-Frameworks QUDT erläutert. Verwandte Arbeiten werden in Abschnitt 3 vorgestellt. Abschnitt 4 beschreibt den Kern des Ansatzes zur Verknüpfung zwischen den Abstraktionsebenen. Der Beitrag endet mit einer Zusammenfassung und einem Ausblick in Abschnitt 5.

2 Grundlagen

2.1 COBIT

COBIT (Control Objectives for Information and Related Technology) [Infb] ist ein umfassendes Framework, das Unternehmen dabei unterstützen soll, ihre Ziele zu Governance und Management der Unternehmens-IT zu erreichen. Dies soll dadurch erreicht werden, dass eine Balance zwischen der Umsetzung von Funktionalität der IT (und damit gesteigertem Nutzen) und der Optimierung von Risiken und Ressourcenverbrauch angestrebt wird. COBIT wird entwickelt und gepflegt von der ISACA (Information Systems Audit and Control Association), einer unabhängigen, gemeinnützigen und internationalen Organisation, die heute mehr als 115000 Mitglieder zählt [Infa].

In der aktuellen Version 5 von COBIT werden dazu alle nötigen Konzepte der IT Gover-

nance sowie ihre Abhängigkeiten definiert: Getrieben von den Bedürfnissen der Stakeholder (Nutzer, Entscheider usw.) werden Enterprise Goals (EG) definiert, die jeweils von IT Goals (ITG) unterstützt werden, die wiederum von IT Prozessen umgesetzt werden. Abbildung 1 zeigt vereinfacht den konzeptuellen Aufbau des Modells. Das Framework geht hierbei jedoch weiter ins Detail und unterscheidet beispielsweise auch zwischen primärer und sekundärer Unterstützung von EGs durch ITGs.

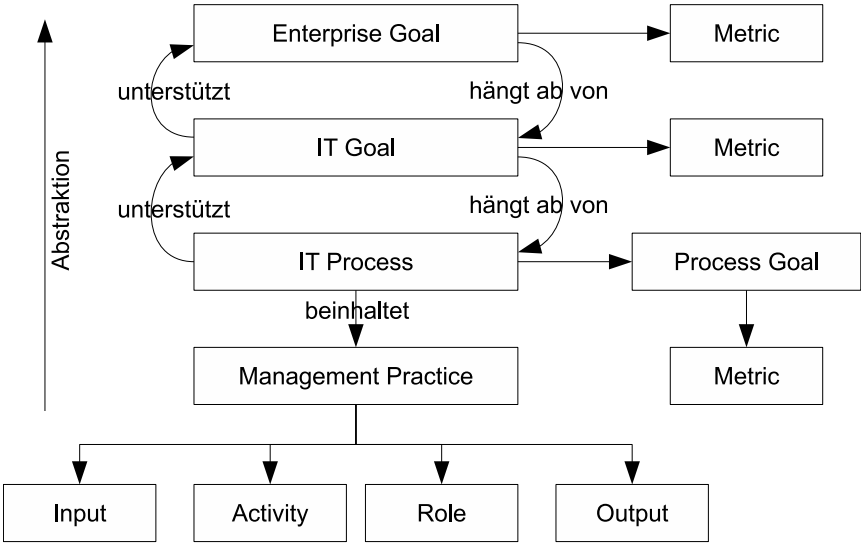


Abbildung 1: Konzeptueller Aufbau der COBIT Abstraktionsschichten

Jeder IT Prozess setzt sich zusammen aus einer Menge von Management Practices (MP), die vergleichbar sind mit detaillierteren Unterprozessen. Jede MP hat eine Reihe von nötigen Eingangs- und erzeugten Ausgangsdokumenten (z.B. Policies, Evaluationen, Vorgehensweisen usw.), eine Menge von Aktivitäten (Handlungsanweisungen) sowie eine Verantwortlichkeitstabelle (RACI-Matrix), die beschreibt, welche Rollen im Unternehmen in welcher Form mit den MPs interagieren. Zusätzlich können alle Goals (EG, ITG, Process Goals) eine Menge von definierten Metriken haben, die den Erreichungsgrad des Ziels quantifizieren sollen. COBIT 5 definiert neben dem Modell eine große Menge von Instanzen der Konzepte: Es sind 37 Prozesse mit mehr als 200 zugehörigen MPs definiert. Außerdem sind ca. 160 Goals, 440 Eingabe- und Ausgabedokumente, mehr als 1100 Aktivitäten und 380 Metriken definiert.

Das COBIT Framework ist ausschließlich in Form von Dokumenten spezifiziert (ca. 400 Seiten Fließtext, Tabellen und Diagramme) – von Seiten der ISACA ist keine maschinenverarbeitbare Form verfügbar. Alle Entitäten sind ausschließlich in menschenlesbarer Form beschrieben, dies beinhaltet auch Metriken. So hat beispielsweise das IT Goal 10 „Security of information, processing infrastructure and applications“ die Metrik „Number of security incidents causing financial loss, business disruption or public embarrassment“.

2.2 QUDT

QUDT (Quantities, Units, Dimensions and Data Types Ontologies) [Hod] ist ein Ontologie-Framework zur Beschreibung von Einheiten, das von der Firma TopQuadrant im Rahmen eines Projektes mit der NASA entwickelt wurde. Es umfasst ein Schema, das Terminologien und Definitionen zu Einheiten, Dimensionen und Typen beschreibt, mehrere Vokabulare von konkreten Einheiten, Dimensionen und Quantities sowie Faktor- und Offset-basierte Transformationsvorschriften für ausgewählte Einheiten. Abbildung 2 zeigt den Aufbau des Frameworks aus den unterschiedlichen Modulen. Grundlegende Konzepte in QUDT sind (informell):

- *Quantity Kind* oder *Kind*: Die Eigenschaft, die durch eine Einheit beschrieben wird, bekannte Beispiele sind Länge, Masse, Zeit, Energie, aber auch Währung, Verzinsung oder Kanalkapazität sind Quantity Kinds. Kinds, die den SI-Basiseinheiten entsprechen, werden als Base Kinds bezeichnet, alle anderen als Derived Kinds.
- *Unit* oder *Unit of Measure*: Eine bestimmte Größe einer Quantity Kind, die als Skala für andere Größen der gleichen Quantity Kind dient.
- *Quantity* oder *Quantity Value*: Das Produkt eines skalaren Wertes und einer Unit.
- *Quantity Symbol* bzw. *Dimension Symbol*: So wie eine Unit ein Symbol hat, so hat auch eine Quantity ein Symbol, beispielsweise ist das Symbol für Masse üblicherweise *m*.
- *Dimension*: Die mathematische Dimension der Teil-Units bei Units, die sich auf Derived Kinds beziehen (z.B. bei der Einheit m/s^2).

In einem Beispiel wäre „Length“ eine Quantity Kind (Base Kind) mit Unit Symbol *L*, die zugehörige Unit „Meter“ mit Symbol *m*; „10 m“ ist dann eine Quantity. Die Derived Kind „Area“ hat das Dimension Symbol L^2 und die Unit „Square Meter“ mit Symbol m^2. Das Vokabular von Units umfasst ca. 780 Einheiten. Um nicht im Vokabular mitgelieferte Einheiten zu definieren, können entsprechende eigene Quantity Kinds, Units und darauf aufbauend Quantities definiert werden.

3 Verwandte Arbeiten

Die Idee, Ontologien für die Integration verschiedener Datenquellen in einem Unternehmen einzusetzen, ist nicht neu, da sich Ontologien als domänenunabhängige Modellierungsform dafür gut eignen. Wache et al. [WVV+01] fassten bereits 2001 in einem Survey zahlreiche Ansätze zusammen. Cruz und Xiao geben in [CX05] eine Übersicht zu unterschiedlichen Aspekten: Zentrale oder P2P-Architektur, Repräsentation von Metadaten, Querying und Mapping. Unter dem Namen *Ontology-Based Data Management* präsentieren Semeraro et al. [SBB+13, Mau13] einen Ansatz zur Datenintegration in Unternehmen, bei dem eine Ontologie (die *Domain Knowledge Base*) den Kern eines Informationssystems bildet, die durch einen Mapping Layer auf bestehende Ressourcen wie

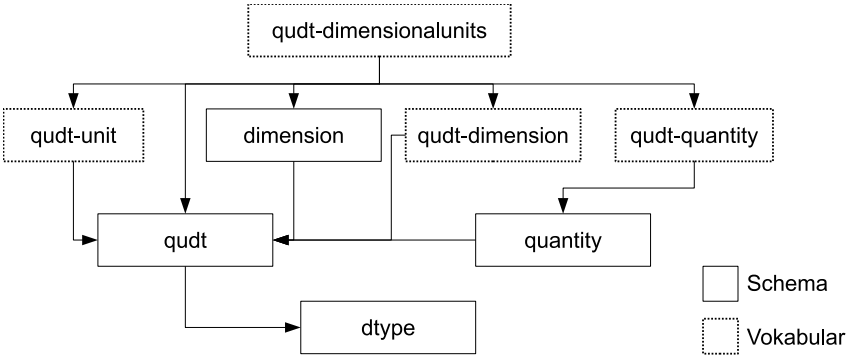


Abbildung 2: Aufbau der Module von QUDT

RDBMS zugreift. Es wird hierbei aber nur marginal darauf eingegangen, wie die Ontologien zu gemeinsam genutzten Terminologien entstehen sollen – sie müssen von Domain Experts modelliert werden. Auf viele Details zur Umsetzung des vorgestellten Ansatzes wird nicht eingegangen. Der Einsatz von Ontologien beschränkt sich zumeist auf die technische Verbindung von Datenquellen, nicht jedoch auf die weitergehende Nutzung der entstehenden Informationen für das bessere Verständnis der Vorgänge im Unternehmen.

Ein Ansatz zur Modellierung von COBIT in Form einer Ontologie wird in [NHSS11] von Niemann et al. vorgestellt. Dieser Ansatz bezieht sich auf die inzwischen veraltete Version 4.1 von COBIT. Die Autoren untersuchen zunächst Methoden zur strukturierten Modellierung einer Ontologie und extrahieren aus der COBIT-Spezifikation grundlegende Begriffe. Es wird jedoch nicht ersichtlich, welche Konzepte und Beziehungen die eigentliche Ontologie umfasst. Übersichten von Entitäten, Konzepten oder Beziehungen sowie eine resultierende Ontologie werden nicht angegeben.

Auf Seite des IT Service Managements gibt es die ITSMO (IT Service Management Ontology) [Fag], die Konzepte aus dem Umfeld des IT Service Managements kompatibel zur Struktur von ITIL in einer OWL-Ontologie formuliert. Diese ist laut eigener Angabe jedoch noch „Work in progress“. Mit dem Ziel, Modell-basierte Software-Entwicklung direkt mit dem IT Service Management zu verbinden, entwickeln Valiente et al. in [Mar11, VGBS12] u.a. eine umfassende ITIL-Ontologie, die ebenfalls für die Integration der Sichten im Unternehmen eingesetzt werden könnte.

Nicht nur Unternehmensdaten, sondern auch Geschäftsprozesse werden in Form von Ontologien modelliert, siehe dazu z.B. Semeraro et al. [SBB+13] oder Hepp und Roman [HR07].

4 Überbrückung von Abstraktionsebenen

Um die in Abschnitt 1 formulierten Ziele zu erreichen, sind mehrere Schritte nötig. Der erste Schritt besteht aus der Erstellung eines formalen Modells des COBIT-Modells. Im

Gegensatz zu anderen Spezifikationen aus dem Umfeld des IT Management, wie z.B. des Business Process Frameworks eTOM [For], das zumindest in Form einer Excel-Tabelle der Prozessbeschreibungen verfügbar ist, stellt COBIT keine vergleichbaren, automatisch konvertierbaren Formate zur Verfügung. Daher müssen die in der Spezifikation beschriebenen Konzepte analysiert und händisch modelliert werden. COBIT lässt sich grob in zwei Teile gliedern: Die Beschreibungen der gemeinsam genutzten Terminologien (z.B. Prozesse, Activities, usw.) und ihre Relationen sowie die Vorgehensbeschreibungen zur Einführung und Umsetzung von COBIT im Unternehmen. Für die hier angestrebte Nachvollziehbarkeit und Automatisierung sind letztere nicht relevant.

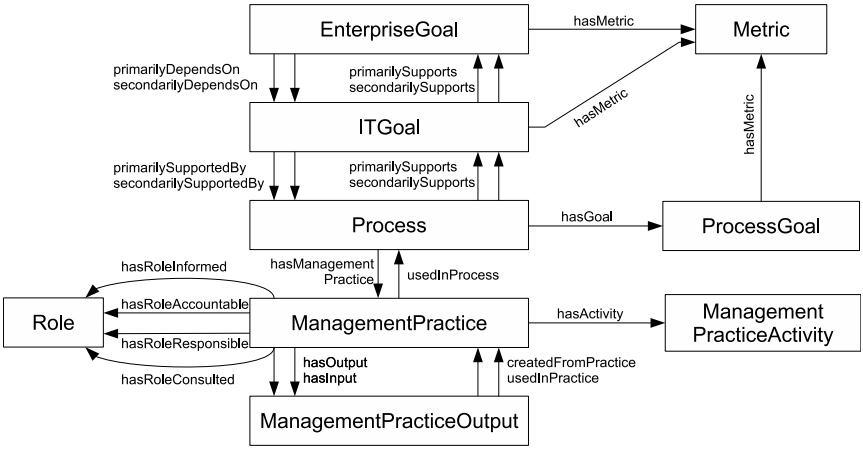


Abbildung 3: Ausschnitt: Kern der COBIT 5 Ontologie

Die COBIT-Konzepte, die in Abschnitt 2.1 beschrieben wurden, wurden in Form einer OWL-Ontologie implementiert. Abbildung 3 zeigt den Kern der resultierenden COBIT-Ontologie. Dazu wurden Klassen für die Kategorien von Konzepten erstellt, die wichtigsten sind dabei die Ziele (EnterpriseGoal, ITGoal, ProcessGoal), Process und ManagementPractice. Die in der Spezifikation tabellarisch angegebenen primären und sekundären Unterstützungsrelationen zwischen EnterpriseGoal und ITGoal sowie zwischen ITGoal und Process wurden als entsprechende ObjectProperties modelliert. Die Verantwortlichkeiten von Rollen der ManagementPractices, die in der Spezifikation als RACI-Matrix (Responsible, Accountable, Consulted, Informed) für jede ManagementPractice angegeben sind, wurden ebenfalls als ObjectProperties ausgedrückt.

Die konkreten Instanzen der Konzepte, die in COBIT spezifiziert sind (also Goals, Prozesse, ManagementPractices usw.) wurden als entsprechende OWL Instanzen in der Ontologie modelliert, die in Tabellen und Grafiken angegebenen Relationen zwischen Instanzen als Object Property Assertions bzw. Data Property Assertions, Bezeichner und Beschreibungen als RDF Labels und Description. Da in einer OWL-Ontologie jede Entität einen eindeutigen URI-Bezeichner hat, musste für zahlreiche Entitäten, die in der COBIT-Spezifikation selbst keine Namen oder Bezeichner besitzen (z.B. Metrics oder ManagementPracticeActivities) eine Nomenklatur entwickelt werden. Auf Details, wie diese

Nomenklatur, sowie auf Klassen, die weitere Details der Spezifikation umsetzen, wird hier aus Platzgründen nicht eingegangen. Die entwickelte COBIT-Ontologie umfasst 30 Klassen, 40 Object Properties, 2578 Individuals und hat mehr als 21,300 Axiome.

Um den Erfüllungsgrad von Zielen zu messen, enthält COBIT das Konzept der Metrik. Die Metriken sind als Fließtext formuliert und variieren stark in ihrem Abstraktionsgrad. Während Metriken von hohem Abstraktionsgrad nur schwer quantifizierbar sind (z.B.: IT Goal 17 „Knowledge, expertise and initiatives for business innovation“, Metrik 1 „Level of business executive awareness and understanding of IT innovation possibilities“), gibt es Metriken, die zwar quantifizierbar, aber nicht einfach automatisch aus Daten des IT Monitorings hergeleitet werden können, aber u.U. aus anderen Informationsquellen des Unternehmens (z.B.: Process 3 „Monitor, Evaluate and Assess Compliance with External Requirements“, Goal 1 „All external compliance requirements are identified“, Metrik 2 „Frequency of compliance reviews“). Metriken mit niedrigem Abstraktionsgrad können direkt aus technischen Messwerten übernommen werden (z.B.: Enterprise Goal 6 „Customer-oriented service culture“, Metrik 1 „Number of customer service disruptions due to IT service-related incidents (reliability)“).

Der nächste Schritt ist die weitere Formalisierung nötiger Entitäten. Konkret bedeutet dies: Selbst mit der Ontologie, die die in der Spezifikation des Frameworks natürlichsprachlich definierten Strukturen vollständig umfasst, bleiben viele Beschreibungen nach wie vor nur menschenlesbar. Um die Metriken automatisch zu verarbeiten, muss modelliert werden, auf welche Größen sie Bezug nehmen. Dabei kann es sein, dass diese Größen selbst zunächst im Ontologie-Schema definiert werden müssen, weil sie in der Spezifikation nur implizit erwähnt werden. Ein Beispiel hierfür sind IT Services, die durch Import einer ITIL- oder ITSM-Ontologie (siehe Abschnitt 3) in die Ontologie mit aufgenommen werden können. Darüber hinaus muss eine Möglichkeit geschaffen werden, die berechneten Größen formal beschreiben zu können, d.h., die Größen mit einer entsprechenden Einheit zu annotieren.

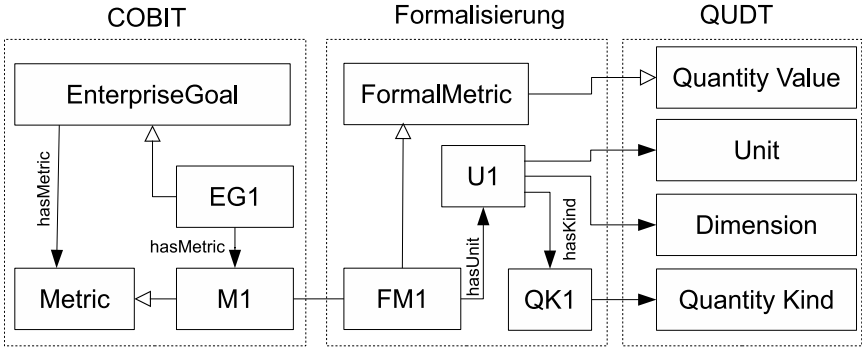


Abbildung 4: Formalisierung von COBIT-Metriken mittels QUDT

Für die Beschreibung von Einheiten in Ontologien wird hier die in Abschnitt 2.2 vorgestellte QUDT-Ontologie verwendet. Eine separate Ontologie, hier als Formalisierungson-

tologie bezeichnet, dient als Bindeglied zwischen den COBIT- und QUDT-Ontologien. Die Konzepte für Einheiten, Dimensionen und Quantity Kinds werden aus QUDT importiert. Abbildung 4 zeigt die Verbindung. In der Abbildung stellen *EG1*, *M1*, *FM1*, *U1* und *QK1* Instanzen dar, während die restlichen Objekte Ontologieklassen darstellen. Die konkrete COBIT-Metrik *M1* soll mit einer beschreibenden Einheit der entsprechenden Quantity Kind und Dimension verbunden werden. Dazu werden in der Formalisierungsontologie zunächst Instanzen für eine solche Unit *U1* und eine entsprechende Quantity Kind *QK1* angelegt. Dieser Schritt ist nur nötig, falls keine entsprechende Einheit oder Kind im QUDT-Vokabular vorhanden ist – der Katalog an Basisgrößen und physikalischen Einheiten ist zwar umfangreich, aber z.B. eine Einheit „Incidents pro Monat“, wie sie für eine COBIT-Metrik gebraucht werden könnte, ist ein QUDT nicht definiert. Eine Instanz *FM1* der Klasse FormalMetric wird als Container für den entsprechenden Quantity Value erzeugt und dann per Annotation oder Object Property mit der ursprünglichen Metrik *M1* verbunden. Werden diese Schritte für alle Instanzen von quantifizierbaren Metriken durchgeführt, kann in SPARQL-Queries, Reports, Visualisierungen und Automatisierungsregeln auf die tatsächlichen Einheiten Bezug genommen werden. Darüber hinaus können die Einheiten im Laufzeitsystem zur Validierung der Gültigkeit von Werten genutzt werden. Benutzeroberflächen, die Unterstützung bieten beim Editieren von Metriken oder bei der Verbindung mit Informationssystemen, aus denen Werte von Metriken berechnet werden, können die Einheiten ebenfalls zur Validierung nutzen.

Der letzte Schritt ist die Verbindung der Metriken mit konkreten Messwerten, die aus Monitoring-Werkzeugen, aber auch anderen Datenquellen, wie z.B. einem Incident Management System, stammen können. Außerdem müssen Fälle unterstützt werden, in denen Metriken hoher Abstraktionsebenen von Werten von Metriken niedriger Abstraktionsebenen abhängen, d.h., erstere können erst berechnet werden, wenn letztere bereits vorher berechnet wurden. Diese Abhängigkeiten von Metriken sind in COBIT nicht explizit ausgedrückt.

Abbildung 5 zeigt, wie diese beiden Fälle behandelt werden. Es wird eine weitere Ontologie eingeführt, die Runtime-Ontologie. Diese wird vom Laufzeitsystem dynamisch aktualisiert und beinhaltet zum einen den Berechnungsstatus der unterschiedlichen konkreten Metriken, zum anderen die Berechnungs- bzw. Transformationsvorschriften. Der Status wird, analog zur Verbindung zwischen COBIT und Formalisierung, mittels einer Data Property Assertion mit der entsprechenden Instanz der FormalMetric verbunden. Für jeden Aktualisierungszyklus wird das Attribut aktualisiert, sobald der neue Metrikwert berechnet wurde. Die in Abbildung 5 vereinfacht dargestellten Berechnungsvorschriften beinhalten die entsprechenden Ausdrücke bzw. Queries sowie Relationen zu den für die Berechnung benötigten Entitäten – entweder Entitäten, die den Zugriff auf externe Datenquellen repräsentieren, oder andere FormalMetrics. In welcher Syntax die eigentliche Berechnungsvorschrift formuliert ist, ist abhängig von der Art der Datenquelle. Für Regeln zwischen Metriken kann SPARQL verwendet werden, für andere Datenquellen kann entweder ein Query-Rewriting eingesetzt oder eine native Abfrageform angewendet werden (z.B. SQL-Query, REST-URL, SNMP GET usw.). Sofern alle notwendigen Details über die Datenquelle im Modell enthalten sind (z.B. Host, Credentials, Protokoll-Version usw.), können diese Abfragen weitestgehend generiert werden. Dies fällt in den Bereich der in

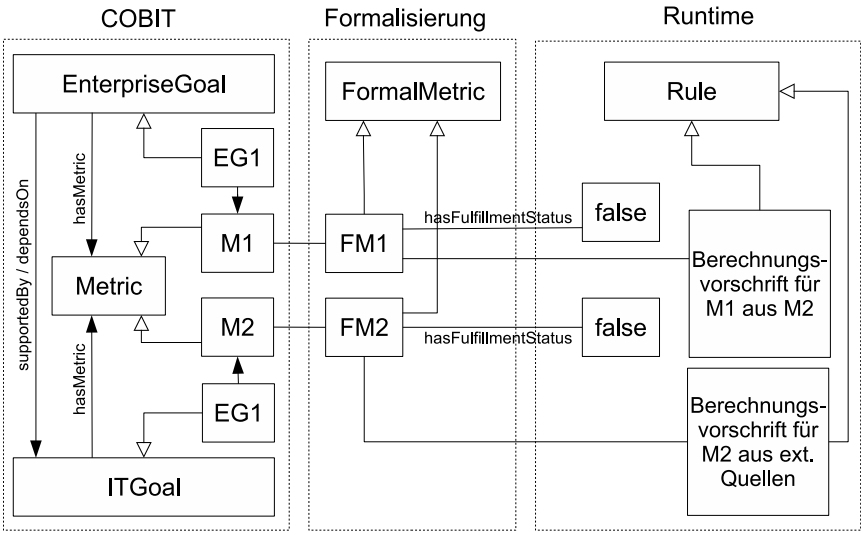


Abbildung 5: Berechnung von abhängigen Metriken

Abschnitt 3 vorgestellten Ontologie-basierten Datenintegration.

Wird im Laufzeitsystem ein regelbasierter Reasoner eingesetzt, können die Abhängigkeiten der Berechnungsstatus als Eingangsvoraussetzungen der entsprechenden Regeln formuliert werden. Der Pattern Matching Algorithmus des Reasoners sorgt dann für die Auswertung der Berechnungsvorschriften in der korrekten Reihenfolge.

5 Zusammenfassung und Ausblick

Dieser Beitrag stellt einen Ansatz vor, mit dem Ontologie-basiert abstrakte Konzepte aus Frameworks für IT Governance oder IT Service Management mit konkreten technischen Daten, wie IT Monitoring Daten, verknüpft werden können. Eine solche Verknüpfung verfolgt mehrere Ziele: Die IT soll nicht mehr als „Black Box“ auftreten, sondern es soll transparent ersichtlich sein, wo sie wie genutzt wird, um die abstrakteren Unternehmensziele zu erreichen, oder wie hoch die anteiligen Kosten der IT für die Unterstützung eines bestimmten Geschäftsprozesses sind. Darüber hinaus sollen für die Entscheidungsfindung nötige Kennzahlen direkt berechenbar gemacht werden, soweit dies anhand bestehender, aber nicht verknüpfter technischer Messdaten möglich ist. Schließlich soll eine Automatisierung ermöglicht werden, bei der Regeln Bezug nehmen können auf Entitäten aller Abstraktionsebenen im Unternehmen, von technischen Systemen über Prozesse bis hin zu Geschäftszielen.

Am Beispiel des IT-Governance Frameworks COBIT 5 wurde untersucht, welche Schritte nötig sind, um eine solche Verbindung zu ermöglichen. COBIT liegt vor in Form ei-

ner menschenlesbaren Spezifikation und wurde daher zunächst in einer OWL-Ontologie formuliert. Die hier in Ausschnitten vorgestellte COBIT-Ontologie umfasst 30 Klassen, 40 Object Properties und 2578 Individuals und beinhaltet die Kernelemente des COBIT-Datenmodells – Geschäftsziele, IT Ziele, Prozesse, Management Practices, Rollen, Metriken usw. – die Teile des Frameworks, die die Einführung von COBIT im Unternehmen beschreiben, wurden nicht berücksichtigt.

Es wurde weiterhin untersucht, wie die in COBIT nur als Fließtext formulierten Metriken, die mit Geschäftszielen, IT Zielen und Prozesszielen verbunden sind und zur Messung der Zielerreichung dienen, ebenfalls formal in einer Ontologie ausgedrückt werden können. Dazu wird die QUDT-Ontologie verwendet, ein Framework zur Definition von Einheiten in Ontologien. In einer separaten Formalisierungs-Ontologie werden Instanzen von COBIT-Metriken mit QUDT-Instanzen verbunden. Auf diese Weise wird formal erfassbar gemacht, was die errechneten Metriken genau ausdrücken.

In einer weiteren Teilontologie, der Runtime-Ontologie, die vom Laufzeitsystem dynamisch aktualisiert wird, werden Berechnungsstatus und Berechnungsvorschriften für Metrik-Werte gespeichert. Berechnungsvorschriften enthalten dabei auch die Referenzen auf für die Berechnung benötigte andere Metrik-Instanzen oder auf Instanzen, die externe Datenquellen repräsentieren. Durch diese Modellierung wird in Verbindung mit einem regelbasierten Reasoner eine automatische Auswertung von Metriken und damit eine Kaskadierung von berechneten Werten von niedrigen auf hohe Abstraktionsebenen im COBIT-Modell möglich.

Ein Laufzeitsystem, das sich augenblicklich in der Entwicklung befindet und hier nicht im Detail vorgestellt wurde, verwaltet die unterschiedlichen Teilontologien, Zugriff auf die externen Datenquellen, einen regelbasierten Reasoner sowie Schnittstellen für Reporting und Visualisierung in Echtzeit. Das System nutzt das Apache Jena Framework [Apa]. Der Reasoner nutzt die Jena-Regelengine und implementiert RDF- und RDFS-Semantik sowie das OWL 2 RL Profile [MGH+]. Basierend auf dem in diesem Beitrag vorgestellten Ansatz kann dieses System zukünftig in der Lage sein, die in Abschnitt 1 gestellten Ziele erfüllen können.

Literatur

[Apa] Apache. Jena. `https://jena.apache.org/`, abgerufen am 24.04.2014.

[AXE] AXELOS. Information Technology Infrastructure Library. `http://www.itil-officialsite.com/`, abgerufen am 24.04.2014.

[CX05] Isabel F. Cruz und Huiyong Xiao. The Role of Ontologies in Data Integration. *Jounal of Engineering Intelligent Systems*, 13(4), 2005.

[Fag] Enrico Fagnoni. ITSMO IT Service Management Ontology. `http://ontology.it/itsmo/v1/itsmo.html`, abgerufen am 24.04.2014.

[For] TM Forum. Business Process Framework (eTOM). `http://www.tmforum.org/businessprocessframework/1647/home.html`, abgerufen am 24.04.2014.

[Hod] Ralph Hodgson. QUDT - Quantities, Units, Dimensions and Data Types Ontologies. http://www.qudt.org/, abgerufen am 24.04.2014.

[HR07] Martin Hepp und Dumitru Roman. An Ontology Framework for Semantic Business Process Management. In A Oberweis, C Weinhardt, H Gimpel, A Koschmider, V Pankratius und B Schmizler, Hrsg., *Proceedings of the 8th International Conference Wirtschaftsinformatik*, Seiten 423–440, Karlsruhe, 2007. Universitätsverlag Karlsruhe.

[Infa] Information Systems Audit and Control Association (ISACA). About ISACA. http://www.isaca.org/about-isaca/Pages/default.aspx, abgerufen am 24.04.2014.

[Infb] Information Systems Audit and Control Association (ISACA). Control Objectives and Related Technology (COBIT). http://www.isaca.org/COBIT/, abgerufen am 24.04.2014.

[Int] International Organization for Standardization. ISO/IEC 20000-1:2011: Information Technology – Service Management. http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=51986, abgerufen am 24.04.2014.

[JG07] Wolfgang Johannsen und Matthias Goeken. *Referenzmodelle für IT-Governance. Strategische Effektivität und Effizienz mit COBIT, ITIL & Co*. dpunkt Verlag, 2007.

[Mar11] María-Cruz Valiente-Blásquez. *Improving IT Service Management using an Ontology-Based and Model-Driven Approach.* Dissertation, Universidad de Alcalá, 2011.

[Mau13] Maurizio Lenzerini. Ontology-based Data Management, Mai 2013. http://wp.sigmod.org/?p=871, abgerufen am 24.04.2014.

[MGH+] Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu, Achille Fokoue und Carsten Lutz. OWL 2 Web Ontology Language Profiles (Second Edition). http://www.w3.org/TR/owl2-profiles/, abgerufen am 24.04.2014.

[NHSS11] Michael Niemann, Sascha Hombach, Stefan Schulte und Ralf Steinmetz. Das IT-Governance-Framework CObIT als Wissensdatenbank - Entwurf, Umsetzung und Evaluation einer Ontologie. Bericht, 2011.

[SBB+13] Giovanni Semeraro, Pierpaolo Basile, Roberto Basili, Marco De Gemmis, Chiara Ghidini, Maurizio Lenzerini, Pasquale Lops, Alessandro Moschitti, Cataldo Musto, Fedelucio Narducci, Arianna Pipitone, Roberto Pirrone, Piero Poccianti und Luciano Serafini. Semantic technologies for industry: From knowledge modeling and integration to intelligent applications. *Intelligenza Artificiale*, 7:125–137, 2013.

[VGBS12] María-Cruz Valiente, Elena Garcia-Barriocanal und Miguel-Ángel Sicilia. Applying Ontology-Based Models for Supporting Integrated Software Development and IT Service Management Processes. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(1):61–74, Januar 2012.

[WVV+01] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann und S. Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *Proceedings of IJCAI'01 Workshop on Ontologies and Information Sharing*, Seiten 108–117, 2001.

Disaster-Readiness als Beispiel für Ontologie-basiertes IT-Management im Geschäftskontext

- Summary Paper -

Peter Klotz

Leiter R&D, Chef-Architekt
Blue Elephant Systems GmbH
Zettachring 2
70567 Stuttgart
peter.klotz@blue-elephant-systems.com

Abstract:

Am Beispiel eines Anwendungsbeispiels mit realem Hintergrund möchte dieser Beitrag die Notwendigkeit und Eleganz von Ontologie-basierter Integration aufzeigen. Viele in IT Management Systemen verborgenen Informationen können kombiniert eine immer aktuelle und vor allem der Realität entsprechende Abbildung liefern um sehr drängende und wichtige Entscheidungen und Bewertungen für Probleme in der Geschäftsebene zu liefern. Die dafür notwendige Integration und Anreicherung von technischen Informationen über System-Grenzen hinweg lässt sich sehr elegant mit Hilfe von Ontologien und Semantic Web Technologie erreichen.

Anhand des Risk Management Beispiels "Disaster-Readiness" werden die notwendigen Mechanismen und konkreten technischen Konzepte vorgestellt.

1. Use-Case

Der in diesem Artikel vorgestellte Use-Case entspringt realen Kunden-Szenarien aus dem IT-Management. Als Teil ihrer Riskomanagement Strategie müssen Firmen sich gegen katastrophale Ausfälle die einen großen Teil der IT-Infrastruktur eines Standortes bedrohen wappnen. Dies könnten Naturkatastrophen, Explosion von Weltkriegsbomben oder Terroranschläge sein. In der Regel wird man sich auf solche potentiellen großflächigen Ausfälle durch Aufbau eines zweiten Rechenzentrums vorbereiten und alle Services in beiden redundanten Rechenzentren vorhalten. Dazu wird jedoch viel Infrastruktur, getrennte Netze, Cluster, die sich über beide Zentren spannen, geteilte SAN, Hochgeschwindigkeits-Netzwerk-Kopplung etc. benötigt. Beide Umgebungen müssen im Monitoring aufgesetzt und aktiv überwacht werden.

Basierend auf diesen technischen Gegebenheiten gibt es nun einige geschäftskritische Fragen, wie z.B.:

- Welche der Firmen-IT Services sind vorbereitet für ein Desaster?
- In welcher Reihenfolge können die Services im Desaster-Fall im anderen Rechenzentrum gestartet werden und wie lange würde das dauern?
- Was kostet diese Ausfallsicherheit?

Um diese Fragen zu beantworten werden folgende Informationen benötigt:

- Information über Abhängigkeiten der IT Services und Infrastruktur bis auf die technische Ebene.
- Informationen sind über viele verschiedene Tools verteilt, wie z.B. im Netzwerk-Management, im Service-Management, im System und Application Management oder auch CMDBs und Inventories. Jedes Tool liefert jedoch nur einen Teilaspekte der Fragen und jedes Tool hat dabei eine eigene Repräsentation der Daten, sowie eigene inkompatible Konzepte und Modelle.
- Kein einziges Tool beantwortet die Frage direkt, da die Fragestellungen sowohl technische als auch Geschäfts-Aspekte umfassen

2. Lösung mit Hilfe von Ontologien

Eines der Kernprobleme ist die Integration von Daten aus verschiedenen Systemen und die Integration von technischen mit nicht-technischen Daten, nur in der Kombination können die oben vorgestellten Fragen beantwortet werden. Weiterhin müssen die Fragen permanent beantwortet werden, d.h. eine einmalige manuelle Analyse ist nicht zielführend.

Ontologien, insbesondere in Form von Semantic Web Technologien und OWL eignen sich hervorragend um IT Systeme und Services zu modellieren. So können Tool-spezifische hart-codierte Datenmodelle ebenso per Ontologie beschrieben werden, wie objekt-orientiert modellierte CMDB-Modelle oder kundenspezifische Datenmodelle. OWL bietet von Hause aus gute Möglichkeiten zur Datenintegration auf Modellseite.

Die Fragestellungen sind aber kein Fall für klassisches Model-Reasoning, vielmehr müssen Daten auf Instanz-Ebene kombiniert werden. Dafür müssen die benötigen Daten aus den verschiedenen Quellen, d.h. IT Management und Business Applikationen gesammelt und entsprechend der spezifischer Ontologien in ein Model eingefügt werden. Dabei kann die nützliche Eigenschaft von OWL ausgenutzt werden, dass ein Individual mehrere Klassen „asserted“ haben kann. Dies ist eine sehr einfach Möglichkeit zu beschreiben, dass ein Objekt Mitglied mehrerer Tools ist, jeweils assoziiert mit Tool-spezifischen anderen Objekten über Properties.

OWL erlaubt es zusätzliche Assertions zu machen über die Mitgliedschaft in Klassen von höherwertigen Business-Ontologien, wie ITSM oder COBIT, die zusätzlich zu den technischen Assertions gemacht werden. Außerdem können dynamisch und iterativ neue

Properties aus spezifischen Ontologien an existierende technische Individuals angehängt werden um diesen einen Business-Kontext zu geben. Eine Property für „Disaster-Readiness“ ist ein solches Beispiel. Es kann sich aber auch um Asset-Informationen, Kosten oder Risikofaktoren aus entsprechenden Ontologien handeln. Damit kann man plötzlich ursprünglich technische Objekte aus einer Business-Perspektive betrachten, z.B. mittels einer SPARQL-Query unabhängig von der technischen Quelle ohne die originale technische Information zu verlieren.

Im besprochenen Beispiel würde per komplexer Regel z.B. einer „constructing SPARQL-Query“ eine „Disaster-ready“ Property für einen Service berechnet auf Basis von z.B.

- der Service läuft auf einem Cluster
- der Cluster besteht aus mindestens zwei physikalischen Systemen (egal in welchem IT Inventory oder CMDB diese gefunden wurden)
- die physikalischen Systeme müssen in zwei verschiedenen Rechenzentren stehen
- die Systeme müssen in unabhängigen aber verbunden Netzen stehen
- die Systeme müssen in einem IT-Monitoring Tool (könnten evtl. unterschiedliche sein) als überwachte Geräte aufgeführt sein
- die Überwachung muss aktiviert und verteilt sein
- es muss Backup für die Systeme konfiguriert sein

Es können nun noch weitere Kriterien für „Disaster-ready“ definieren, je nachdem wie es die Geschäftsanforderungen gebieten. Entscheidend ist, daß die Kriterien auf einem abstrahierten und integrierten Model definiert werden können, das die Unterschiede der darunterliegenden Tools abstrahiert und versteckt. Sobald man die Informationen in die Welt der Ontologien gebracht hat, hat man alle benötigte Flexibilität und Mechanismen um sich auf den Geschäftswert und eine wirklich exakte und aussagekräftige Definition zu konzentrieren.

Ontologien revolutionieren das IT-Management!

Ontologie-basiertes Monitoring von IT-Systemen

Fabian Meyer

Labor für Verteilte Systeme
Hochschule RheinMain
Campus Unter den Eichen 5
D-65195 Wiesbaden
fabian.meyer@hs-rm.de

Abstract: In diese Arbeit wird ein Ontologie-basierter Ansatz zum Monitoring von IT-Systemen vorgestellt. Es wird untersucht, wie die Skalierbarkeit des Reasonings für die IT-Management-Domäne verbessert und zeitbehaftete Daten dargestellt werden können. Zudem wird ein Konzept vorgestellt, mit dem zusätzliche, nicht in der Ontologie darstellbare Semantik als Regelstruktur in der Domänenontologie modelliert und zur Laufzeit von einem speziellen Reasoner interpretiert werden kann, wodurch eine Fragmentierung der Wissensbasis vermieden wird. Abschließend werden die entwickelten Konzepte auf die MAPE-K-Loop abgebildet und eine prototypische Umsetzung vorgestellt.

1 Motivation

Moderne IT-Systeme sind komplexe Kompositionen unterschiedlicher Hard- und Softwarekomponenten, deren korrektes Zusammenspiel von einer Vielzahl von Faktoren abhängt. Um einen reibungslosen Ablauf der operationellen Prozesse zu gewährleisten, ist eine ständige Regelung des Gesamtsystems notwendig, das sogenannte IT-Management. Die Aufgaben der Systemadministratoren sind dabei vielfältig: von der reinen Überwachung der Quality of Service (QoS) basierend auf definierten Key Performance Indicators (KPIs), über die Einhaltung definierter Service Level Agreements (SLAs), das Deployment von neuen Komponenten bis hin zur Rekonfiguration der gesamten Systemlandschaft.

Um diese zeit- und kostenaufwendigen Prozesse zu unterstützen und zu teilautomatisieren, ist eine Vielzahl von zunehmend spezialisierten Management-Werkzeugen entstanden, ein ganzheitliches Management von heterogenen Systemen ist jedoch kaum möglich. Auch die flexible Anpassung und Erweiterung gestalten sich als schwierig, da das unterlagerte Modell und die Management-Regeln meist fest kodiert sind. Dies mündet in einer Pluralität von IT-Management-Lösungen, die Administratoren nutzen müssen, um das Gesamtsystem zu verwalten. Daher besteht ein hoher Bedarf an domänenübergreifenden, leicht erweiterbaren und leicht anpassbaren Lösungen.

Durch die Entwicklungen im Bereich des Semantic Webs in den letzten Jahren haben Ontologien aus der Künstlichen Intelligenz einen neuen Aufschwung erlebt. Diese domänenübergreifenden Wissensmodelle bieten die Möglichkeit, Zusammenhänge von Entitäten

nicht nur rein syntaktisch, sondern auch semantisch zu beschreiben. Eine Eigenschaft, die sie besonders für den Einsatz zur Modellierung von komplexen, heterogenen IT-Systemen geeignet macht.

Neu entstandene Standards und Technologien wie die Web Ontology Language (OWL) und die Semantic Web Rule Language (SWRL) wurden bereits von Forschergruppen für den Einsatz im IT-Management adaptiert (vgl. Abschnitt 4). Dabei zeigte sich, dass Ontologien für die Modellierung der komplexen IT-Systeme und die Abbildung existierender IT-Management-Modelle (z.B. CIM) geeignet sind, jedoch die schlechte Skalierbarkeit des Reasonings, die fehlende zeitliche Betrachtung und die fehlende Möglichkeit zur Wertaggregation Probleme bereiten. Es existieren zwar Ansätze die versuchen diese Aspekte unter Nutzung der SPARQL Protocol And RDF Query Language (SPARQL) anzugehen [Sch14], dabei entsteht jedoch ein hybrides Auswertungssystem, bei dem die anwendungsspezifische Semantik extern hinterlegt ist und zyklisch auf die Wissensbasis angewandt wird. Dies führt zu einer Wissensfragmentierung und zusätzlichen Hürden bei der Synchronisation.

Daher wird in dieser Arbeit ein alternativer Ansatz vorgestellt, bei dem die zusätzliche Semantik nicht extern gehalten wird, sondern als Konzepte, Instanzen und Relationen Teil der Domänenontologie ist, sodass die komplexen Zusammenhänge bei der Betrachtung des Modells erschlossen werden können. Zur Laufzeit wird das so modellierte Wissen von einem speziellen Reasoner interpretiert und auf Regeln abgebildet, die zusätzliche Schlussfolgerungen ermöglichen. Dieser Reasoner ist wiederum Teil einer Management-Architektur, die auf der Monitor-Analyze-Plan-Execute-Knowledge (MAPE-K) Regelschleife basiert und sich auf beliebige Domänen adaptieren lässt.

2 Ansatz

Der in diesem Kapitel vorgestellte Ansatz beschreibt ein neues Verfahren, bei dem Ontologien als semantisches Modell für das IT-Management eingesetzt werden. Die Hauptaspekte sind die Verbesserung der Skalierbarkeit des Reasonings, die Darstellung von zeitlichen Aspekten, die Modellierung der zusätzlichen Semantik und deren gemeinsame Einbettung in das Management-System.

2.1 Skalierbarkeit des Reasonings

In der ersten Version von OWL (2005) wurde die Sprache anhand der Ausdrucksmächtigkeit und Komplexität in die drei Klassen OWL-Full, OWL-DL und OWL-Lite [BvHH+04] unterteilt. Existierende Ontologie-basierte IT-Management-Ansätze (vgl. Abschnitt 4) nutzen meist Tableau-basierte Reasoner, die zwar in der Lage sind, das gesamte OWL-DL-Spektrum abzudecken, dadurch jedoch sehr schlecht skalieren.

Für OWL 2 (2012) wurde das Konzept der Sprachprofile eingeführt, die jeweils für einen bestimmten Anwendungsbereich optimiert sind. [MGH+12] charakterisiert die drei Profile

wie folgt:

OWL 2 EL eignet sich besonders für die Anwendung bei Ontologien mit einer komplexen Terminologie. Es basiert auf der Familie der $\mathcal{EL}^{++}$-Beschreibungslogiken [BBL08], ist entscheidbar und lässt sich mit verschiedenen Inferenzalgorithmen in *ExpTime* lösen.

OWL 2 QL eignet sich besonders für die Anwendung bei Ontologien mit einer großen Anzahl Individuen, bei der das Beantworten von Queries im Vordergrund steht. Es kann durch Query-Rewriting in Relationalen Datenbanken als *NP-vollständiges* Problem gelöst werden, beschränkt die Ausdrucksmächtigkeit jedoch auf das Niveau von UML- oder ER-Diagrammen.

OWL 2 RL eignet sich besonders für die Anwendung in Umgebungen, bei denen ein skalierbares Reasoning benötigt wird, ohne die Ausdrucksmächtigkeit zu stark einzuschränken. Es kann mit regelbasierten Systemen als *NP-vollständiges* Problem gelöst werden.

Betrachtet man diese drei Profile, erscheinen das QL- und das RL-Profil am geeignetsten für den Einsatz im IT-Management, da dort in der Regel keine komplexe Terminologie, sondern viele Instanzen und Relationen betrachtet werden. Unsere Untersuchung der Ontologien von existierenden IT-Management-Ansätzen hat gezeigt, dass die Ausdrucksmächtigkeit des QL-Profils nicht ausreichend ist. Das RL-Profil hingegen ist mächtig genug, um die benötigten Strukturen darzustellen und eignet sich daher sehr gut für den Einsatz im IT-Management. Durch die Wahl des RL-Profils und dessen Umsetzung mit einem Regelsystem kann die Komplexitätsklasse von *NExpTime-vollständig* für OWL-DL auf *NP-vollständig* reduziert werden, wodurch eine bessere Skalierbarkeit erzielt wird.

2.2 Darstellung von Zeit

Für viele Aspekte des IT-Managements spielt Zeit eine wichtige Rolle. Häufig müssen Mittelwerte, Abweichungen, etc. über Zeiträume gebildet, Differenzen von Werten zweier Zeitpunkte bestimmt oder Verläufe betrachtet werden, um die Dienstgüte einen Systems zu ermitteln. OWL-Ontologien bieten jedoch auf Sprachebene kein Zeitkonzept.

Ein Ansatz zur Lösung diese Problems wurde in [WF06] vorgestellt und in [FMK10] und [MFK12] erweitert. Darin werden sogenannte 4DFluents eingeführt, bei denen ein Individuum eine Menge von zeitlichen Ausprägungen (Temporal Parts) hat, die jeweils während eines bestimmten Zeitraums gültig sind. Der Zustand eines Individuums zum Zeitpunkt t ist definiert als die Vereinigung aller zeitlichen Ausprägungen für die gilt, dass t in ihrem Gültigkeitsintervall liegt. Der Vorteil des Ansatzes ist, dass keine Spracherweiterung nötig sind, sondern mit den existierenden OWL-Mitteln zeitliche Relationen abgebildet werden können. Der Ansatz findet bereits im Kontext anderer Domänen Verbreitung [BP10, HAC13].

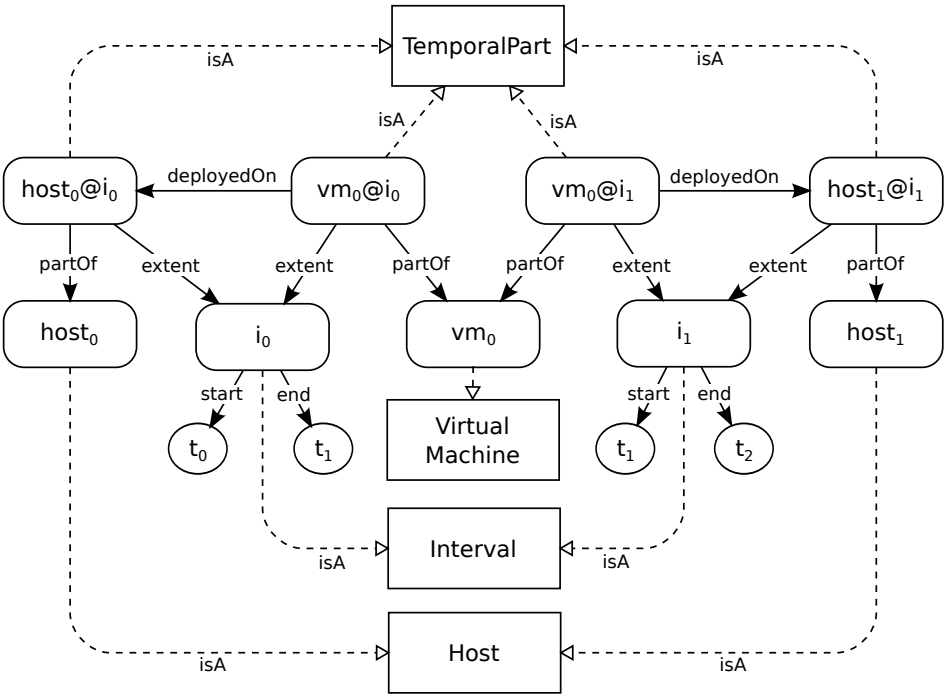


Abbildung 1: Beispielhafte Darstellung der Verschiebung einer Virtuellen Maschine zwischen zwei Hosts unter Nutzung von 4D Fluents. Eckige Kästen stellen Klassen, abgerundete Kästen Instanzen, Kreise Literale, durchgezogene Pfeile Objekt- und Datenrelationen und gestrichelte Pfeile Klassenrelationen dar.

Eine beispielhafte Adaption diese Ansatzes auf die IT-Management-Domäne ist in Abbildung 1 dargestellt. Es zeigt den Umzug einer Virtuellen Maschine vm_0 von Host $host_0$ nach $host_1$ zum Zeitpunkt t_1. Dabei werden die $deployedOn$-Relationen nicht zwischen den Individuen vm_0 und $host_0$ bzw. vm_0 und $host_1$, sondern zwischen den jeweiligen zeitlichen Ausprägungen $vm_0@i_0$ und $host_0@i_1$ bzw. $vm_0@i_1$, $host_1@i_1$ definiert, die während der Intervalle i_0 bzw. i_1 gültig sind.

2.3 Modellierung von zusätzlicher Semantik

Neben einer zeitlichen Betrachtung spielt im IT-Management bei der Bestimmung von Dienstgüte die Aggregation von Werten über Geflechte von Komponenten eine wichtige Rolle. Beispiele dafür sind die Summe des Speicherverbrauchs aller Virtueller Maschinen auf einem Host oder die durchschnittliche Antwortzeit einen Dienstes. Solche Zusammenhänge lassen sich jedoch mit den Standardmitteln von OWL nicht ausdrücken.

Daher haben wir die Ontology-based Management Language (OntML) entwickelt, die es ermöglicht, zusätzliche deklarative Closed-World-Regeln zu definieren, die keine Allgemeingültigkeit haben, sondern anwendungsdomänenspezifisch definiert werden. Um sie dennoch bei der Betrachtung der Zusammenhänge in einer Domänenontologie erfassen zu können, werden die Regeln als Teil der Ontologie modelliert. Ein ähnlicher Ansatz wurde in der Quantities, Units, Dimensions and Data Types (QUDT)-Ontologie der NASA [HKHS13] vorgestellt. Dort steht die umfassende Beschreibung der funktionalen Zusammenhänge von physikalischen Größen im Vordergrund.

Die Konzepte zur Modellierung der deklarativen Regeln (OWL-Klasse *DeclarativeRule*) in der Ontologie folgen dem Schema von Horn-Klauseln und werden als materiale Implikationen dargestellt. So setzt sich eine Regel aus konjugierten Bedingungen (*Condition*) und Konsequenzen (*Consequence*) zusammen. Bedingungen sind Variablenzuweisungen (*VariableDeclaration*) oder Einschränkungen (*Constraint*), Konsequenzen Relationserzeugungen (*AssertionAction*) bzw. Relationswiderrufe (*NeglectionAction*). Darin können wiederum unter Nutzung der folgenden Konzepte Ausdrücke modelliert werden:

Boolesche Ausdrücke als Konstanten (*True* und *False*) und logische Operatoren (*Not*, *And*, *Or*, *GreaterThan*, *LessThan*, *Equal*, *NotEqual*, *GreaterEqual* und *LessEqual*).

Datenausdrücke als Konstanten (*StringConstant*, *IntegerConstant*, *DoubleConstant*), arithmetische Ausdrücke (*Add*, *Subtract*, *Multiply* und *Divide*) und als Werte einer Datenrelation auf einem Objektausdruck (*DataPropertyExpression*).

Objektausdrücke als referenzierte Individuen der Domänenontologie (*IndividualExpression*), Elemente einer Klasse der Domänenontologie (*ClassExpression*), Werte einer Objektrelation auf einem Objektausdruck (*ObjectPropertyExpression*) und Filterausdrücke (*FilterExpression*).

Die Einführung von Konzepten für zusätzliche Aggregationsfunktionen ermöglicht es, nicht mehr nur Aussagen über einzelne Elemente, sondern über Mengen zu machen. Dazu werden die Aggregationsausdrücke zur Bildung einer Summe (*Sum*), des Mittelwerts (*Average*), Minimums (*Minimum*) und Maximums (*Maximum*) definiert.

Um das in Abschnitt 2.2 vorgestellt Zeitkonzept in Regeln nutzbar zu machen, ohne dass die gesamte Relationskette bis hin zu den zeitlichen Ausprägungen manuell definiert werden muss, wird eine zusätzliche *TemporalExpression* eingeführt. Sie erlaubt die Definition eines Zeitpunkts oder Zeitraums, für den ein bestimmter Ausdruck ausgewertet wird. Unterstützte Zeitausdrücke sind relative Zeitfenster (*WindowExpression*) und relative Zeitpunkte (*InstantExpression*), also z.B. alle Virtuellen Maschinen, die einem Host innerhalb der letzten fünf Stunden zugewiesen waren, oder alle Virtuellen Maschinen, die einem Host vor genau drei Stunden zugewiesen waren.

Auch die Abbildung von Monitoring-Informationen auf die Wissensbasis gestaltet sich bei existierenden Ansätzen schwierig, da dies i.d.R. in Adapterkomponenten geschieht. Dort werden eingehende Events interpretiert, benötigte Kontextinformation aus der Wissensbasis ausgelesen und anschließend programmatisch neue Instanzen und Relationen in die Wissensbasis einfügt, bzw. existierende Instanzen und Relationen entfernt. Dies führt

zu einer weiteren Fragmentierung der Logik. Deshalb stellt OntML die Möglichkeit bereit, domänenspezifische Events formal zu beschreiben und anschließend in den Regeln zu referenzieren. Ein Event wird beschrieben durch einen eindeutigen Namen und eine Menge von Feldern mit primitiven Datentypen. Die Nutzung in den Management-Regeln geschieht analog zu der Nutzung von Klassen und Relationen aus der Domänenontologie. Zudem werden spezielle Filterausdrücke eingeführt, die es erlauben, das erste bzw. letzte Event eines Stroms auszuwählen.

2.4 Interpretation der Regeln

Zur Laufzeit wird die mit Hilfe der OntML-Konzepte definierte zusätzliche Semantik der Ontologie interpretiert und als Wenn-Dann-Regeln auf einer Ebene oberhalb der OWL-RL-Regeln umgesetzt. Dazu werden alle Bedingungen und Konsequenzen jeder Regelinstanz betrachtet und die zugrundeliegenden Ausdrücke transformiert. Die logischen und arithmetischen Ausdrücke können direkt übernommen werden, Datenrelationen, Objektrelationen und Klassenreferenzen werden in entsprechende Prädikate und Aggregationen in Mengenoperationen überführt.

Im Kontext von temporalen Ausdrücken gilt eine Sonderbehandlung für die Interpretation von Relationen und Klassenreferenzen, bei der der Zeitparameter des temporalen Ausdrucks einbezogen und statt der eigentlichen Individuen deren zeitlichen Ausprägungen in den Prädikaten gebunden werden.

Binden alle Bedingungen in einer Regel, werden für jede *AssertionAction* neue zeitliche Ausprägungen der entsprechenden Individuen mit einem offenen Intervall, gültig ab dem aktuellen Zeitpunkt, erzeugt und diese über die definierten Relationen verknüpft. Für jede *NeglectionAction* werden die betroffenen zeitlichen Ausprägungen der entsprechenden Individuen gesucht und deren Intervalle mit dem aktuellen Zeitpunkt geschlossen.

2.5 Management-Architektur

Die vorgestellten Konzepte wurden auf die MAPE-K-Architektur [IBM06] adaptiert und ein Management Framework entwickelt. Im Folgenden werden die Knowledge-, Monitor- und Analyze-Komponenten beschrieben. Da der Fokus dieser Arbeit auf der Systemüberwachung liegt, werden die Plan- und Execute-Komponenten nicht betrachtet. Eine Übersicht der Gesamtarchitektur ist in Abbildung 2 dargestellt.

Die **Knowledge**-Komponente beinhaltet die für das Management benötigten Ontologiemodelle. Neben den systemspezifischen Domänenontologien des zu überwachenden Systems, die entweder aus einem existierenden Modell wie UML oder CIM transformiert oder manuell modelliert werden, ist dort die systemunabhängige OntML-Ontologie zu finden. Sie beinhaltet die in Abschnitt 2.3 vorgestellten Konzepte zur Modellierung von Monitoring Events und Management-Regeln. Die systemspezifische Management-Ontologie setzt sich aus der Domänenontologie und der OntML-Ontologie zusammen und beschreibt

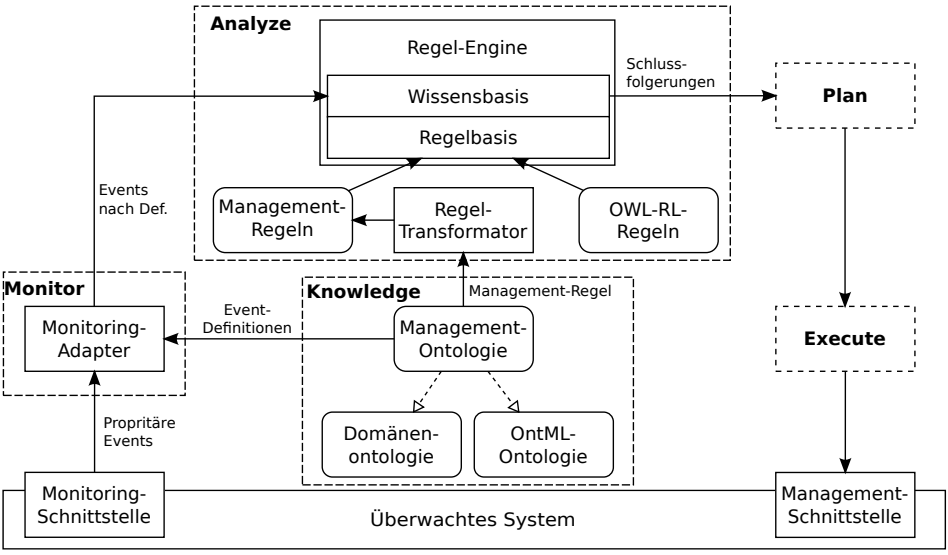


Abbildung 2: MAPE-K-basierte Management Architektur

durch die Instanziierung der OntML-Konzepte und Referezierung von Entitäten des Domänenmodells die Management-Regeln.

Die **Monitoring**-Komponente dient als Adapterschicht, die zwischen dem überwachten System und dem Management-System vermittelt. Dazu werden systemspezifische Monitoring-Informationen von der Monitoring-Schnittstelle des überwachten Systems gelesen (SNMP, CIM, proprietäre Schnittstelle, etc.) und in die von der Management-Ontologie definierten Events transformiert. Die transformierten Events werden dann zur Interpretation an die Analysekomponente weitergegeben.

Die **Analyze**-Komponente besteht aus einer Regel-Engine, in der die Management-Funktionalität umgesetzt wird. Die Regelbasis setzt sich aus den statischen, einmalig auf die genutzte Regel-Engine adaptierten OWL-RL-Regeln und den dynamischen, domänenspezifischen Management-Regeln zusammen. Die Monitoring-Events und Management-Regeln werden von einem Transformator beim Start der Komponenten aus der unterlagerten Management-Ontologie der Knowledge-Komponente extrahiert und in Enginespezifische Regeln übersetzt. Anschließend wird die Wissensbasis der Engine mit dem Domänenmodell gefüllt, bevor zur Laufzeit Events von der Monitoring-Komponente entgegengenommen, in die Wissensbasis eingefügt und dort interpretiert werden.

Das abgeleitete Wissen (QoS, SLA-Verletzungen, etc.) der Analysekomponente kann anschließend in der **Plan**-Komponente für Optimierung und in einer **Execute**-Komponente für die Rekonfiguration des überwachten Systems genutzt werden.

3 Umsetzung

Die in Abschnitt 2.5 vorgestellte Management-Architektur wurde in einem Prototyp umgesetzt. Dabei wurde als Programmiersprache Java, als Service Framework OSGi[1], als Schnittstelle zur Verarbeitung von Ontologien OWLAPI[2] und als Regel-Engine Drools[3] eingesetzt.

Es wurde ein Datenmodell für die Darstellung der Ontologien in der Regel-Engine definiert, und die in [MGH+12] definierten OWL-RL-Regeln wurden händisch in die Drools Rule Language (DRL) übersetzt. Anschließend wurde ein Regeltransformator entwickelt, der die geladenen Ontologien verarbeitet, die darin beschriebenen Management-Regeln extrahiert und entsprechend des in Abschnitt 2.4 definierten Schemas in DRL-Regeln umsetzt.

Zur Vermessung des Prototyps wurde ein synthetischer Versuchsaufbau entwickelt, bei dem das Monitoring von Virtuellen Maschinen simuliert wird. Jede Virtuelle Maschine ist einem physikalischen Host zugewiesen und erzeugt zyklisch Monitoring Events, die den genutzten CPU-Anteil der Maschine enthalten. Darauf basierend soll bewertet werden, ob sich ein Host in einem gefährdeten Zustand befindet.

Dazu wurde eine entsprechende Domänenontologie entwickelt, die die folgenden Entitäten definiert:

- Die Klassen *Host*, *ProneHost* und *VirtualMachine*,
- die Objektrelation *deployedOn*, die eine *VirtualMachine* mit einem *Host* verknüpft
- und die Datenrelationen *id*, *currentLoad* und *averageLoad* für eine *VirtualMachine* und *overallLoad* für einen *Host*.

Anschließend wurde eine Management-Ontologie entwickelt, die die Domänenontologie importiert und die folgenden Events und Management-Regeln unter Nutzung der Domänenentitäten definiert:

- Das Event *VMLoadEvent* mit den Attributen *vmId* vom Typ Integer zur Identifizierung der Virtuellen Maschine und *load* vom Typ Double zur Definition der aktuellen CPU-Load,
- und die Management-Regeln
  - *AssertAverageLoad*, die jede Minute den durchschnittlichen *load* aller *VMLoadEvents* einer *VirtualMaschine* über eine Minute bildet und eine entsprechende *averageLoad*-Relation einfügt,
  - *AssertOverallLoad*, die für einen *Host* die Summe des *averageLoad* aller ihm zugewiesenen *VirtualMaschines* bildet und eine entsprechende *overallLoad*-Relation einfügt

[1] http://www.osgi.org/
[2] http://owlapi.sourceforge.net/
[3] https://www.jboss.org/drools/

– und *AssertProneHost*, die einen *Host* der Klasse *ProneHost* hinzufügt, wenn dessen *overallLoad* größer als 80% ist.

Zur Laufzeit wurden die Ontologien in das prototypische System geladen, die darin enthaltenen Regeln transformiert und ein Monitoring-Adapter gestartet, der 100 virtuelle Maschinen erzeugt, die auf zehn Hosts verteilt sind. Anschließend generiert er für jede virtuelle Maschine alle zehn Sekunden ein *VMLoadEvent* mit einem zufälligen Load.

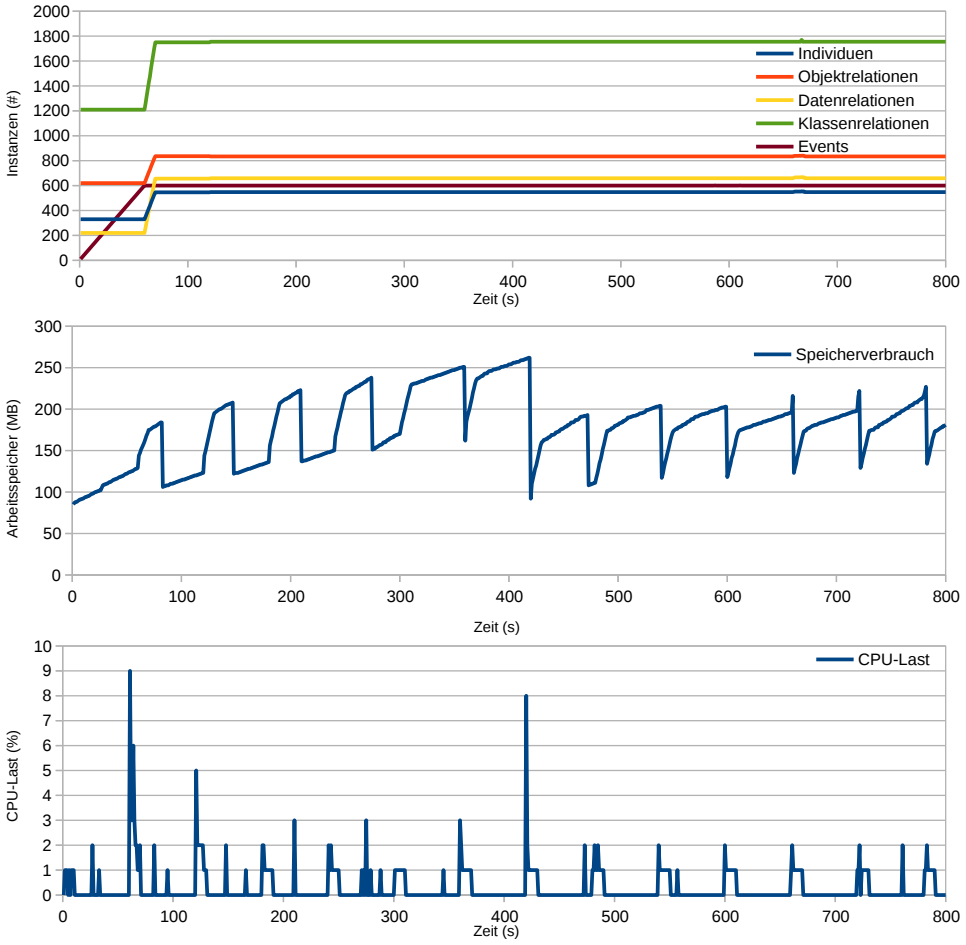


Abbildung 3: Performance-Messung zur Management-Architektur für einen Versuchsaufbau mit 100 Virtuellen Maschinen.

Die in Abbildung 3 dargestellte Messung zeigt, dass sich das System nach ca. 60 Sekunden eingeschwungen hat und anschließend mit einer konstanten Anzahl an Individuen, Objektrelationen, Datenrelationen, Klassenrelationen und Events arbeitet. Das unstetige Verhalten des Speicherverbrauchs lässt sich durch die Garbage Collection der Java Virtual

Machine begründen, im Mittel ist er jedoch auch nahezu konstant. Die CPU-Last steigt alle 60 Sekunden durch die Berechnung des durchschnittlichen und gesamten Loads leicht an. Die Messungen wurden auf einem System mit einem Intel® Core™ i7-3770 CPU @ 3.40GHz x 8 und 16GB Hauptspeicher durchgeführt. Dieses Ergebnis zeigt, dass trotz des sehr frühen Zustands des Prototyps bereits die Überwachung von mittelgroßen Systemen möglich ist.

4 Verwandte Arbeiten

Der Einsatz von Ontologien für das IT-Management hat bisher noch keine breite Anwendung gefunden, es existieren jedoch bereits einige Publikationen, die sich mit dem Thema beschäftigen.

In [VVB04] wird die Abbildung von existierenden Management-Modellen auf OWL-Ontologien betrachtet. Als Grundlage dient die Structure of Management Information (SMI), die Guidelines for the Definition of Managed Objects (GDMO), das Managed Object Format (MOF) und das Common Information Model (CIM). Die resultierende Ontologie kombiniert die angeführten Modelle zu einem gemeinsamen Management-Modell, das die Konzepte der unterschiedlichen Basismodelle miteinander verknüpft.

[DGVB09] beschreibt, wie Systeme in verschiedene Abstrationsschichten aufgeteilt und in einer hierarchischen Ontologienstruktur dargestellt werden können. Ähnlich geartete Systeme nutzen dabei auf den unteren Abstraktionsebenen dieselben Ontologien und Konzepte. Dadurch können beim automatisierten Management Aussagen über abstrakte Komponenten auf eine Vielzahl von konkreten Systemen adaptiert werden. Das Papier zeigt damit, dass OWL in der Lage ist, komplexe Systeme in verschiedenen Abstraktionsebenen darzustellen und unterschiedliche Systeme auf den unteren Ebenen zu vereinheitlichen.

Ein realer Anwendungsfall von Ontologie-basiertem IT-Management wird in [VV06] vorgestellt. Als überwachtes System wird eine Netzwerkinfrastruktur herangezogen, die basierend auf dem Policy-based Network Management (PBNM) [WSS+01] gemanaged wird. Dazu werden für das Domänenmodell SWRL-Regeln definiert, die beim Ausfall einer Verbindung eine Rekonfiguration der Netzwerkinfrastruktur durchführen. Eine Management-Komponente überwacht zur Laufzeit die Infrastruktur und bildet deren Zustand auf die Ontologie ab. Dort abgeleitete Rekonfigurationen werden interpretiert und auf das reale System abgebildet. Das Resultat der Arbeit war, dass sich die entwickelten Konzepte zum Ontologie-basierten Management anwenden und auf beliebige andere IT-Management-Probleme übertragen lassen.

In [TMK11] wurde eine MAPE-K-basierte Architektur für automatisiertes, wissensbasiertes IT-Management vorgestellt. Laufzeitinformationen eines überwachten Systems werden auf ein ontologisches Modell abgebildet und anschließend mit einem hybriden Ansatz aus semantischem Reasoner (exaktes Reasoning) und Bayes-Netz (probabilistisches Reasoner) Schlüsse gezogen. Darauf basierend werden Entscheidungen zur Rekonfiguration des Systems getroffen und diese in Form von Kommandos auf das reale System umgesetzt.

5 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein neuer Ansatz für ein Ontologie-basiertes Monitoring von IT-Systemen vorgestellt. Es wurden Lösungen für die Probleme der Skalierbarkeit des Reasonings, der Darstellung von zeitlichen Aspekten sowie der Aggregation über Geflechte von Komponenten vorgestellt. Dazu wurden die unterschiedlichen Sprachprofile von OWL 2 untersucht und das OWL-RL-Profil als am besten geeignet für den Einsatz im IT-Management ausgewählt. Für die Darstellung von Zeit wurde das in [WF06] vorgestellte 4D-Fluents-Konzept aufgegriffen und dessen Anwendung im IT-Management veranschaulicht. Um zusätzliche Semantik zu modellieren, wurde die Sprache OntML definiert, die es erlaubt, Regeln als Teil der Domänenontologie zu modellieren, die zur Laufzeit von einem speziellen Reasoner interpretiert und umgesetzt werden. Abschließend wurde die Integration des Ansatzes in die MAPE-K-Loop vorgestellt und dessen prototypische Umsetzung erläutert.

In zukünftigen Arbeiten soll eine umfangreiche Evaluation des Prototyps durchgeführt werden, um dessen Performance besser beurteilen zu können. Dazu wird als reales Testsystem das Air Traffic Management (ATC) System PHOENIX der Deutsche Flugsicherung GmbH herangezogen.

Zudem soll das Konzept auf die Plan- und Analyze-Komponente der MAPE-K-Loop ausgedehnt werden, um ein ganzheitliches System-Management zu ermöglichen. Dazu sollen in OntML neben den deklarativen Regeln auch Management-Aktionen beschrieben werden können, die in der Planungkomponente zur Optimierung des Modells genutzt werden. Die daraus resultierenden Optimierungsschritte sollen dann in der Execute-Phase auf ein reales System angewandt werden.

Literatur

[BBL08] Franz Baader, Sebastian Brandt und Carsten Lutz. Pushing the EL Envelope Further. In *In Proceedings of the OWLED 2008 DC Workshop on OWL: Experiences and Directions*, 2008.

[BP10] Sotiris Batsakis und Euripides G. M. Petrakis. SOWL: Spatio-temporal Representation, Reasoning and Querying over the Semantic Web. In *Proceedings of the 6th International Conference on Semantic Systems*, I-SEMANTICS '10, Seiten 15:1–15:9, New York, NY, USA, 2010. ACM.

[BvHH+04] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider und Lynn Andrea Stein. OWL Web Ontology Language Reference. `http://www.w3.org/TR/owl-ref/#Sublanguage-def`, February 2004.

[DGVB09] Jorge E. López De Vergara, Antonio Guerrero, Víctor A. Villagrá und Julio Berrocal. Ontology-Based Network Management: Study Cases and Lessons Learned. *Journal of Network and Systems Management*, 17(3):234–254, September 2009.

[FMK10] Flavius Frasincar, Viorel Milea und Uzay Kaymak. tOWL: Integrating time in OWL. In *Semantic Web Information Management*, Seiten 225–246. Springer, 2010.

[HAC13] Benjamin Harbelot, Helbert Arenas und Christophe Cruz. Continuum: A Spatiotemporal Data Model to Represent and Qualify Filiation Relationships. In *Proceedings of the 4th ACM SIGSPATIAL International Workshop on GeoStreaming*, IWGS '13, Seiten 76–85, New York, NY, USA, 2013. ACM.

[HKHS13] Ralph Hodgson, Paul J. Keller, Jack Hodges und Jack Spivak. QUDT - Quantities, Units, Dimensions and Data Types Ontologies. `http://www.qudt.org/`, August 2013.

[IBM06] IBM Corporation. An Architectural Blueprint for Autonomic Computing, Technical Whitepaper (Fourth Edition), June 2006.

[MFK12] Viorel Milea, Flavius Frasincar und Uzay Kaymak. tOWL: A temporal web ontology language. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(1):268–281, 2012.

[MGH+12] Boris Motik, Bernardo Cuenca Grau, Ian Horrocks, Zhe Wu Achille Fokoue und Carsten Lutz. OWL 2 Web Ontology Language Profiles (Second Edition). `http://www.w3.org/TR/owl2-profiles/`, 2012.

[Sch14] Jan Schäfer. A Semantic Self-Management Approach for Service Platforms. In *IEEE/IFIP Network Operations and Management Symposium (NOMS '14), Krakow, Poland, May 5-9, 2014 (accepted for publication as poster)*. IEEE, May 2014.

[TMK11] Andreas Textor, Fabian Meyer und Reinhold Kroeger. Semantic Processing in IT Management. In Pascal Lorenz und Eckhard Ammann, Hrsg., *Proceedings of the Fifth International Conference on Advances in Semantic Processing (SEMAPRO)*, Lisbon, Portugal, November 2011.

[VV06] Víctor A. Villagrá und Jorge E. López De Vergara. Ontology-based Policy Refinement Using SWRL Rules for Management Information Definitions. In *in OWL. In: Proc. 17th IFIP/IEEE International Workshop on Distributed Systems, Operations and Management (DSOM*, Seiten 227–232, 2006.

[VVB04] Jorge E. López De Vergara, Víctor A. Villagrá und Julio Berrocal. Applying the Web ontology language to management information definitions. *IEEE Communications Magazine*, 42:68–74, 2004.

[WF06] Chris Welty und Richard Fikes. A Reusable Ontology for Fluents in OWL. In *Proceedings of the 2006 Conference on Formal Ontology in Information Systems: Proceedings of the Fourth International Conference (FOIS 2006)*, Seiten 226–236, Amsterdam, The Netherlands, The Netherlands, 2006. IOS Press.

[WSS+01] A. Westerinen, J. Schnizlein, J. Strassner, M. Scherling, B. Quinn, S. Herzog, A. Huynh, M. Carlson, J. Perry und S. Waldbusser. Terminology for Policy-Based Management, 2001.

Optimizing resource topologies of workload in the cloud by minimizing consumption and maximizing utilization while still meeting service level agreements

Cataldo Mega,
ECM Cloud Services Development
IBM Germany Research & Development,
Schönaicherstraße 220,
71032 Boeblingen, Germany
(cataldo_mega@de.ibm.com)

Christoph Lange
ECM Cloud Services Development
IBM Germany Research & Development,
Schönaicherstraße 220,
71032 Böblingen, Germany
(chrisla@de.ibm.com).

Abstract: Reducing the total cost of ownership (TCO) and meeting the service level agreements (SLA) of a computing service or application in the Cloud remains a challenge. The traditional method of meeting SLA is to size production systems based on expected peak workloads, thus this leads to very low compute resource utilization. Our experience shows typical average utilization rates of 35 - 40%, topped only by IBM Mainframe systems with utilization rates of up to 80 - 85%. In the 'Cloud' with an inherent multi-tenant environment resource utilization rates of below 40% translates into a prohibitive cost factor and therefore into a non-competitive business model.

In this paper we discuss the dynamics of a policy based automated resource orchestration approach for servicing multi-tenant workloads in the Cloud. Our aim is to adapt system topologies dynamically based on changes in workload at runtime. The condition for being able to achieve these goals are: 1) a composable, self-tuned infrastructure (smart IaaS), 2) automatic coordination and provisioning of managed middleware (smart PaaS) and last but not least 3) elastic, self-driven orchestration of software defined workloads and system topologies (smart SaaS).

Introduction

Companies are starting to embrace and adopt the "cloud" delivery model for managed enterprise services i.e. in the Business-to-Business (B2B) market. Companies moving their on premise applications off-premise into the cloud are trying to reduce their total cost of ownership (TCO) while maintaining service level agreements (SLAs) the same or even improve. Cloud services like webmail and cloud storage originate from the consumer market and typically utilize a pricing model different from what is common in

B2B. Pay per use pricing models from business to consumer (B2C) scenarios are becoming a favorite model of every Chief Financial Officer (CFO) seeking to reduce capital expenditure (CAPEX). However, the new business model comes at the cost of integration rework and optimization of the solution for adjusting to the changed ecosystem when moving existing solutions into the cloud. In general, moving complex enterprise applications from an on premise to an off-premise production context is not a trivial task. The Cloud enforces a paradigm shift that affects the way software solutions are developed, delivered and operated. Providers of enterprise services must consequently cope with this new challenge as software services must be able to communicate with the underlying platform and infrastructure to dynamically request or release components as required, including the orchestration of resource topology changes[1] at varying workloads.

Characteristics of Software Defined Environments

Traditionally software is run on proprietary hardware using custom-built resource topologies i.e. a physical and logical layout of the components and their configuration. In software defined environments such as [8], [9] it is possible to use modeling languages such as the "Topology and Orchestration Specification for Cloud Applications" [10] allow the standards based definition of topologies and services as well as their dynamic management. This is in contrast to purpose built topology adaption methods based on the integration with proprietary Infrastructure-as-a-Service (IaaS) models. Enabling a dynamic automated system topology adaption requires all topology components and resources to be programmatically reconfigurable during runtime using new open standard cloud technology. This includes common infrastructure components such as network, compute, memory and storage (IaaS) elements, also platform (PaaS) components such as database schemas, application server, full text indexing and search engines, all the way up to software solution (SaaS) components like Content and Process Engine, Content Navigation Web-Clients, eDiscovery, Retention and Records Managers.

At the IaaS level scale is achieved by adding infrastructure elements but if the application component itself is saturated this does not yield the desired effect as adding new compute resources would not eliminate the bottleneck. With software defined environment application resources are considered and dynamically added or removed allowing a solution specific fine granular resource allocation. This approach lends itself more effectively to adapt the topology for a given SaaS workload.

2.1 Characteristics of ECM workloads

In order to validate our method we use a typical Enterprise Content Management (ECM) solution. ECM solution components include Business Support Services (BSS) and their respective administrative clients most of which are Java Enterprise Edition (Java EE) applications running on a Java EE application server complemented by a database and a filesystem among other software based topology components.

A workload is described by the actions of an end user population using one or more applications to perform a typical series of operations against a system. The set of operations and the sequence in which they are performed is called a workload pattern. The overall system workload consists of the aggregated individual workloads of all actors using the system. Service Level Agreements (SLA) specifies the Quality of Service (QoS) upon which the service provider and consumer [2] do contractually agree. SLA are then split into system component specific service-level objectives (SLO). SLO define system performance characteristics that must holistically be met by all components to sustain the workload at hand.

A typical ECM workload consists of a mix of CRUDS operations i.e. Create; Retrieve; Update; Delete; and Search presented in three different production contexts, interactive, batch and the administrative. Within a given context the assumption is that individual operations must process a request within a given time thus the KPI is 'response time' or 'throughput' or both. Examples of other type of KPIs are: document index time latency, archive capacity, storage size, system availability (HA), recovery time objectives (RTO), recovery point objective (RPO)… The cloud infrastructure is therefore forced to serve the different tenant-specific workload by translating each workload into a resource topology that defines the actual system layout. Under the above conditions and in order to save costs, as many resources as possible must be shared, which means that we have a resource optimization problem to solve.

Similar to a database query optimizer that is used to produce an optimized access plan in our case we need to develop a resource topology optimizer that produces an optimized topology change plan. The problem to solve can be defined as:

"What is the resource topology that best satisfies the load predicted and what are the topology changes required to get there."

The answer we seek consist of an execution plan specifying the necessary steps to chance the current resource topology considering the elasticity [1]of the cloud environment and that is executed by an orchestrator during runtime such for being able to define the system topology that best satisfies the given load, optimize utilization while minimizing consumption.

We examined three current approaches to solve this problem [3]: 1) utility based optimization techniques, 2) machine learning adaptive techniques and a 3) feedback loop based on control theory.

Utility-based optimization techniques are based on a performance model of a system. Creating such a model is a very complex and inflexible procedure. It implies longer times of the system to react to topology changes thus resulting in a reduced efficiency. As an example, a model based on queuing theory is used e.g. in [4].

The disadvantage of machine learning adaptive techniques is that they usually require several training hours. Furthermore, examples using this approach often assume a

[1] Note: In this paper, we define elasticity as the 'time to execute a topology change'.

maximum set of applications. Thus, an unlimited scalability of a service would not be possible. Examples for machine learning adaptive approaches are found in [5] and [6].

We have chosen to implement procedure that utilizes a feedback loop based on control theory employing a Monitor-Analyze-Plan-Execute (MAPE) concept and have adapted it to use heuristics derived from measured system performance base lines[7]. MAPE is a generic control-loop concept that uses 4 phases to control a target system. In the first monitor phase certain metrics of the system are measured and afterwards in the analyze phase compared to set baseline values. The plan phase then uses the result of this analysis to compile a plan of execution steps that adjust the system topology to meet the target values. Finally, this plan is executed in the final phase by an orchestrator that drives provisioning and de-provisioning workflows.

Gathering heuristics from performance measurements

As a first step a reference system and a normalized workload are defined. In our tests we used a typical enterprise content management solution i.e. compliance archive applications. The resource topology is defined in terms of domain specific resources (nodes), the relation among them (edges) and the hierarchy that describes a production system. The resource nodes cover infrastructure (IaaS), platform (PaaS) and solution entities (SaaS). In our case the common infrastructure components such as network, compute, memory and storage elements, then platform elements such as database schemas, application server, full text indexing and search engines, and finally software solution elements like Content Engine, Process Engine, Archive Web-Client, eDiscovery, Retention and Records Managers complemented by common Business Support Services (BSS). For that reason, we create an ECM specific vocabulary and a Domain Specific Language used as an abstraction suited for the domain considered. The ECM DSL was then complemented by a numerical normalization based on resource type and amount of capacity that the respective resource provides. This allows us to transfer measured performance data to other, even different systems and cloud environments.

Consider the following use case: A user logs on to the ECM application, navigates to her workplace and opens the workbasket. In the next step, she picks a work item e.g. an insurance claim, retrieves the document and starts reading the claim. A few minutes later, she opens the search user interface and performs a search against the ECM repository in order to find collateral documents relevant to the claim. The search yields a hit list with 20 documents, out of which the user selects to retrieve three, one after the other and annotates the 3rd, updating the document. After many iterations of the same mix of operations she eventually logs out of the system.

By using this and other representative usage scenarios a set of automated performance tests are developed that simulated end user behavior. During the performance test runs KPIs are measured and performance baselines created for all primitive operations. Then a final analysis is conducted to determine upper and lower threshold, saturation points and relationships among resource types utilizing best practices and experience. The outcome of the analysis is knowledge characterizing the system from which provisioning policies are derived that help the topology optimizer find a satisfactory resource

topology in reasonably short time. Figure 1 shows an overview of system performance graphs, the performance baselines measured and a few of the characteristic points used to derive heuristics and rules. It also outlines the topology optimization process.

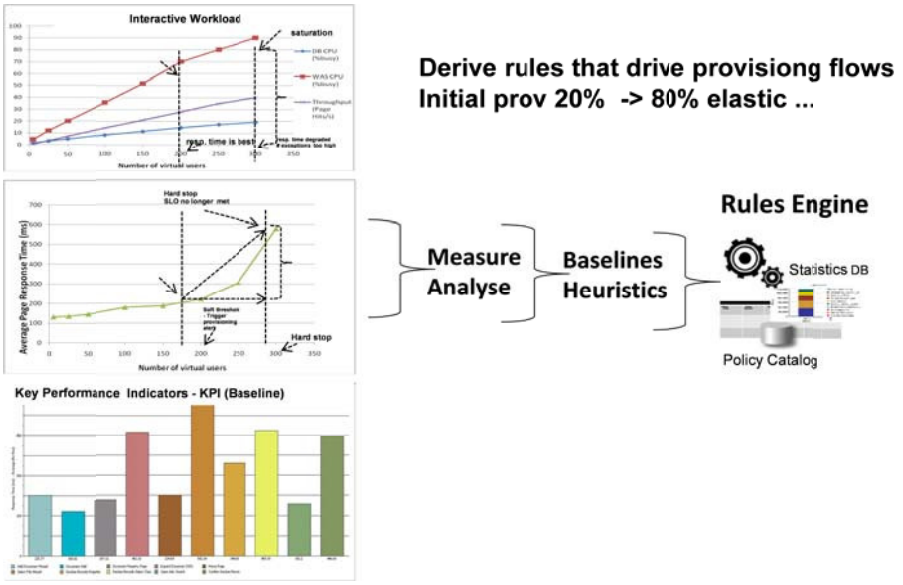


Figure 1 Heuristics are translated in provisioning policies.

Performing topology changes using heuristics

The optimizer uses the heuristics and rules to determine the optimal topology and to create a plan of the steps required to perform the changes. The orchestrator executes the plan and drives the creation of the new production system, as seen in Figure 2.

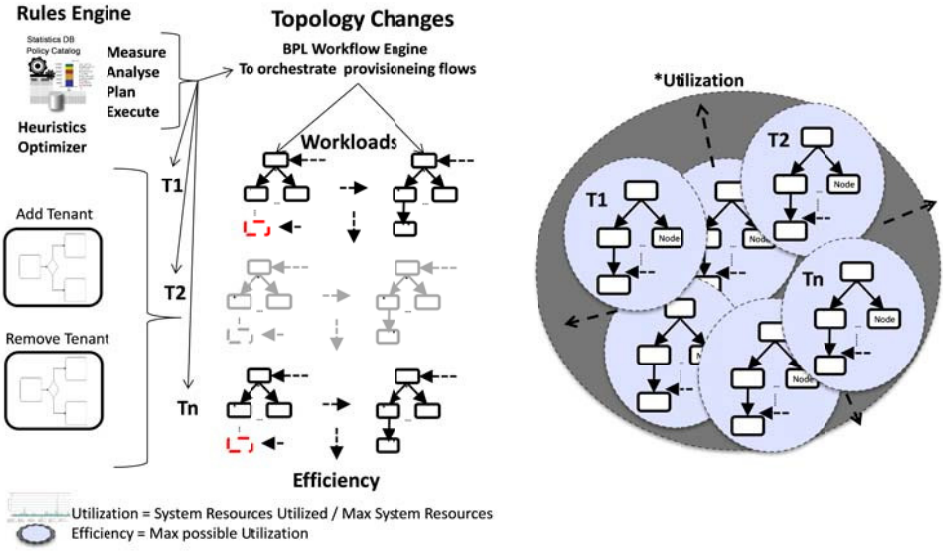


Figure 2 Reference System Topology and plans to apply elasticity for multiple tenants

In summary the MAPE-loop is continuously executed. The information collected in the monitor phase is used to predict how much time the service provider has to provision new resources since he is able to derive the slope of the workload from the last samplings until the most recent one.

Evaluation and Findings

In order to evaluate our approach we have set up a reference system running the Enterprise Content Management (ECM) cloud application and performed KPI measurements in order to derive performance baselines. Our ECM-based reference cloud-application provides a compliance archive as a Service for documents and emails in the cloud. Its three-tier Database, Application, Web architecture makes it well suited for our dynamic provisioning approach since the individual components can be scaled independently of each other.

The relevant components for our evaluation are a Web2.0 Client Application, the Batch Loader and an service administration client. The Web 2.0 Client Application serves as a graphical user interface for interactive access to the content in the cloud allowing customers to store and retrieve documents. The Batch Loader is used to ingest large amounts of documents according to batch-job definitions while the administrative client

application allows customers to configure and manage the whole system from a single user interface. The test system was loaded with several millions of documents of varying types and with sizes. The following performance baselines resulted from our initial interactive Web-Client tests.

The key baseline we measured by running our tests is shown in Figure 3.

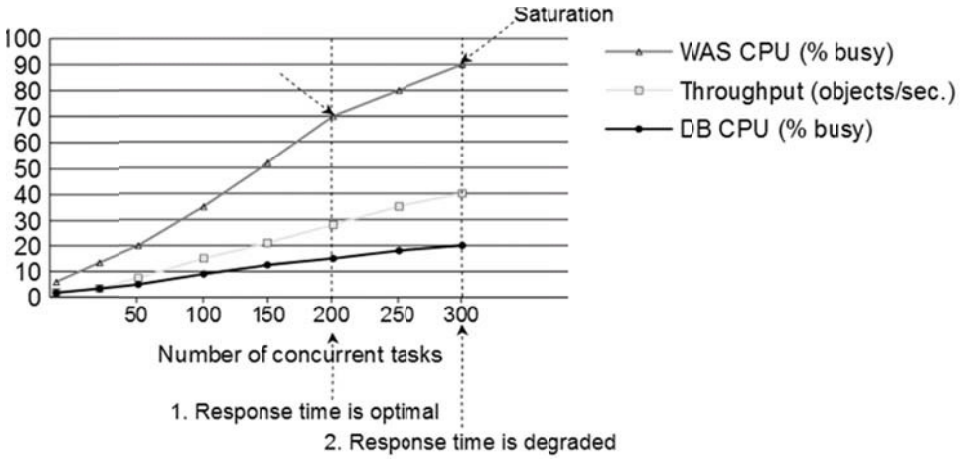


Figure 3. Batch workload performance. (WAS: WebSphere Application Server; DB: Database)

The term task on the x-axis represent a "virtual users". Every "virtual user" performs a certain amount of operations following a specific profile. The curve shows that at 300 simulated users, the WebSphere application server (WAS) on the reference system reaches saturation. Response times suffer and the service level objectives can no longer be met for this component. We therefore conclude that when the system hits 300 concurrent virtual users another application server needs to be deployed. From the picture above, one can also observe that at 200 concurrent virtual users the system load on the database system is below 20% CPU, therefore the database to application server CPU-ratio is ~ 1:3, leaving a CPU reserve to compensate load spikes.

From this we learn that a 2 CPU database node can serve up to 3 application sever nodes running the interactive workloads. Response time measurements also show that at 200 concurrent users the response time is optimal in relation to one application server, smaller load oscillations can be compensated and application server CPU reserves are sufficient for the Operating System (OS) to run smoothly.

Conclusion

The novelty of our approach is to: a) measure performance baselines on a reference test system using representative workloads, b) derive rules that describe the relationships (edges) among the resources (nodes) in the topology graph and c) use the heuristics learned to compute the optimized topology and the execution plan perform the topology changes. Finally an orchestrator is introduced for orchestrating resource topologies

changes and to programmatically create the system environment capable to sustain a given workload that fulfills service level agreements (SLA).

The benefit of our approach is to enable the sharing of resources between tenants on systems with different vertical layers of abstraction (virtual machine, application server, application component, etc.). Resources capacity is computed for tenant specific workloads at a fine granular level allowing SaaS providers to minimize the resources necessary to provide an SLA-compliant service offering and to reduce cost by eliminating unnecessary resources.

References

[MWLSB01] Mega, C. ; Waizenegger, T. ; Lebutsch, D. ; Schleipen, S. ; Barney, J.M. "Dynamic cloud service topology adaption for minimizing resources while meeting performance goals" [in IBM Journal of Research and Development Vol. 58 NO. 2/3 Paper 8 March/MAY 2014

1. F. Leymann, "Cloud Computing: The Next Revolution in IT",[in Proc. Photogramm. Week, 2009, vol. 52, pp. 3–12.
2. A. Paschke and E. Schnappinger-Gerull,"A categorization scheme for SLA metrics, [in Proc. Multi-Conf. Inf. Syst., 2006, vol. 80,pp. 25–40.
3. D. Ardagna, B. Panicucci, M. Trubian, and L. Zhang, "Energy-aware autonomic resource allocation in multi-tier virtualized environments",[IEEE Trans. Serv. Comput., vol. 5, no. 1, pp. 2–19, Jan.-Mar. 2012.
4. B. Urgaonkar, P. Shenoy, A. Chandra, and P. Goyal, "Agile,dynamic provisioning of multi-tier Internet applications",[ACM Trans. Auton. Adap. Syst. (TAAS), vol. 3, no. 1, pp. 1:1–1:39,2008.
5. S. Islam, J. Keung, K. Lee, and A. Liu, "An empirical study into adaptive resource provisioning in the cloud",[in Proc. IEEE Int. Conf. UCC, 2010, p. 8.
6. J. Parekh, G. Jung, G. Swint, C. Pu, and A. Sahai, "Comparison of performance analysis approaches for bottleneck detection in multi-tier enterprise applications",[in Proc. IEEE Int. Workshop Quality Service, 2006, pp. 293-1–293-9.
7. A. Boerner, "Orchestration and provisioning of dynamic system topologies", [M.S. thesis, University of Stuttgart, Stuttgart, Germany, 2011.
8. The Open Networking Foundation standard. [Online]. Available: http://www.opennetworking.org
9. The OpenStack Project. [Online]. Available: http://www.openstack.org
10. The OASIS-TOSCA standard. [Online]. Available: https://www.oasis-open.org/committees/tosca/
11. IBM Corporation, "An architectural blueprint for autonomic computing", [in Autonomic Computing White Paper,2006. [Online]. Available: http://www-03.ibm.com/autonomic/pdfs/ AC%20Blueprint%20White%20Paper%20V7.pdf
12. A. Arefin, V. Singh, G. Jiang, Y. Zhang, and C. Lumezanu, "Diagnosing data center behavior flow by flow,[in Proc. 33rd IEEE Int. Conf. Distrib. Comput. Syst., 2013, pp. 11–20.
13. C. Mega, F. Wagner, and B. Mitschang, " From content management to enterprise content management, [in Proc. Datenbanksysteme für Business, Technologie und Web (BTW), 2005, pp. 596–613. 14. C. Mega, K. Krebs, F. Wagner, N. Ritter, and B. Mitschang, B Content-anagement-systeme der na¨chsten generation,[in Proc Wissens- und Informationsmanagement; Strategien,Organisation und Prozesse, 2008, pp. 539–567.

15. T. Ritter, B. Mitschang, and C. Mega," Dynamic provisioning of system topologies in the cloud,[in Proc. I-ESA, 2012,pp. 391–401.
16. H. Zhang, G. Jiang, K. Yoshihira, H. Chen, and A. Saxena, "Intelligent workload factoring for a hybrid cloud computing model,[in Proc. Congr. Serv., 2009, pp. 701–708.
17. T. Waizenegger, O. Schiller, and C. Mega, "Datensicherheit in mandantenfaehigen Cloud Umgebungen",[in Proc.Datenbanksysteme fuer Business, Technologie und Web (BTW), 2013, pp. 477–489.
18. F. Wagner, K. Krebs, C. Mega, B. Mitschang, and N. Ritter, „Towards the design of a scalable email archiving and discovery solution",[in Proc. 12th East Eur. Conf. Adv. Databases Inf. Syst., 2008, pp. 305–320.
19. F. Wagner, K. Krebs, C. Mega, B. Mitschang, and N. Ritter, "Email archiving and discovery as a service",[in Proc. Intell. Distrib. Comput., Syst. Appl., 2008, pp. 197–206

Agile Migration

Michael Pisula

TNG Technology Consulting GmbH
Beta-Str. 13A
85774 Unterföhring
mihcael.pisula@tngtech.com

Abstract: Die Migration komplexer Systeme ist oft teuer, langwierig und birgt große Risiken. Das liegt häufig daran, dass ein System nicht einzeln migriert werden kann, weil starke Abhängigkeiten bestehen. Nicht selten führt dies dazu, dass die Migration als nicht durchführbar angesehen wird. Wir haben für unsere Kunden ein Verfahren entwickelt, um die Migration zu entkoppeln. Statt die Migration in einem Big Bang durchzuführen, werden die Systeme zuerst entkoppelt, um dann Schritt für Schritt migriert zu werden. Somit kann man die Risiken erheblich senken und die Komplexität verringern. Durch die Entkopplung der Migration wird es möglich, agil auf Verzögerungen im Projektplan und Änderungen der Anforderungen einzugehen.

1 Einleitung

Die Situation dürfte den meisten bekannt vorkommen. In der Systemlandschaft schlummert eine Zeitbombe, ein System, vor langer Zeit entstanden, kaum noch wartbar, fehleranfällig. Neue Features in das System zu bringen ist so gut wie unmöglich. Ganz klar, das System muss migriert werden. Doch damit fangen die Probleme oft erst an. In diesem Beitrag möchte ich die Probleme mit Migrationen von komplexen Systeme aufzeigen, und eine Lösung vorstellen, mit der man einige dieser Probleme umgehen kann.

2 Die Migration

Die Migration eines Systems ist immer dann problematisch, wenn dieses System in einer Systemlandschaft eingebettet ist. Denn so ein System kommuniziert mit anderen Systemen. Somit muss man sich darüber Gedanken machen wie diese Kommunikation nach der Migration aussehen soll. Alle Schnittstellen des alten Systems 1:1 zu übernehmen, ist normalerweise keine ideale Option. Zum einen soll das neue System mehr können, zum anderen sind die alten Schnittstellen auch oft nicht mehr im besten Zustand, die eingesetzte Technologie ist veraltet, es wurden Workarounds eingebaut, die man loswerden will etc. Bei einigen der anderen Systeme kann man die Schnittstellen anpassen, gerade dann wenn diese Systeme noch jung sind, und es einen aktiven Support

gibt. Das trifft aber nicht auf alle zu, meist sind einige der anderen Systeme auch schon veraltet, und eine Ablösung ist schon geplant oder zumindest angedacht.

Migriert man jedes dieser Systeme einzeln, hat man das Problem, dass die noch nicht migrierten Systeme und ihre Schnittstellen weiter unterstützt werden müssen. Genauso muss aber auch schon die Unterstützung der neuen Systeme vorhergesehen und eingebaut werden. Dies macht die Entwicklung der neuen Systeme deutlich teurer und langwieriger. Oftmals möchte man ein altes System auch in mehrere neue Systeme unterteilen, damit wird es deutlich komplexer die gleichen Schnittstellen zu unterstützen, die das alte System angeboten hat. Oder es gibt Workflows die in den neuen Systemen nicht umgesetzt werden, die aber trotzdem für die Übergangszeit eingebaut werden müssen. In anderen Fällen ist die neue Lösung ein Standardprodukt, welches die alten Systeme nicht unterstützt und erst kostspielig angepasst werden müsste.

Alles in allem ist eine Schritt-für-Schritt Migration nur selten sinnvoll, wenn zum Beispiel nur sehr wenige Systeme betroffen sind oder keine Änderungen an den Schnittstellen gewünschnotwendig sind.

3 Der Big Bang

Eine häufig verwendete Lösung der Big Bang Ansatz. Bei der Big Bang Migration werden alle Systeme zum gleichen Zeitpunkt migriert. Dieser Ansatz beseitigt die eben beschriebenen Probleme, es müssen in allen neuen Systemen nur die Schnittstellen der Systeme unterstützt werden, die nach der Migration übrig bleiben. Der Preis dafür ist aber eine hohe Komplexität bei der Planung und Umsetzung, sowie große Risiken, gerade in Bezug auf Verzögerungen im Projektverlauf. Da die Systeme miteinander kommunizieren müssen, ist eine sehr detaillierte Abstimmung zwischen den einzelnen Projekten unabkömmlich. Verzögerungen bei nur einem System führen zu einer Verzögerung der kompletten Migration. Auch die Umschaltung der Systeme selbst ist sehr riskant. Zum einen wegen der hohen Komplexität, da viele Systeme gleichzeitig umgeschaltet werden, zum anderen weil die einzige Möglichkeit auf schwere Probleme bei der Migration zu reagieren, ein komplettes Rollback aller Systeme ist. Die Risiken sind schwer zu kalkulieren und je mehr Systeme betroffen sind, desto aufwändiger wird die Migration.

Oft genug führen diese Gründe dazu, dass die Migration immer wieder verschoben wird, weil niemand das Risiko auf sich nehmen möchte. Das führt oft dazu, dass Systeme weit über ihr Haltbarkeitsdatum erweitert und gewartet werden, bis unweigerlich der Zeitpunkt kommt, an dem man um die Migration nicht mehr umhin kommt.

4 Agile Migration

4.1 Ansatz

Für unsere Kunden in solchen Situationen haben wir einen Ansatz entwickelt, der die meisten der Probleme löst und komplexe Migrationen deutlich erleichtert.

Während die Big Bang Migration dadurch kompliziert wird, dass sie die Migration aller Systeme synchronisiert, hat die Migration einzelner Systeme das Problem, dass jedes neue System deutlich komplexer wird, da es sowohl Schnittstellen zu den alten als auch zu den neuen Systemen unterstützen muss. Unser Verfahren kombiniert die Vorteile: die neuen Systeme müssen nur die Schnittstellen unterstützen die nach der kompletten Migration verwendet werden sollen, die Migration der einzelnen System ist aber trotzdem entkoppelt, jedes System kann unabhängig von den anderen migriert werden.

Bei der Big Bang Migration wird den einzelnen Migrationsprojekten ein starres Korsett aufgezwungen, alle müssen zum gleichen Zeitpunkt testbar sein und für die Migration fertig sein. Darin ähnelt die Big Bang Migration dem Wasserfallmodell, dass ebenfalls eine starre Abfolge von Phasen hat. Wie beim Wasserfall muss bei Problemen in eine frühere Phase zurückgefahren werden. Treten beim Testen der Systeme Probleme auf, wirkt sich das auf alle Systeme aus. Gibt es bei der Umschaltung eines Systems Probleme, müssen alle Systeme wieder in die Testphase zurück. Wie beim Wasserfallmodell auch ist die Flexibilität im Umgang mit Änderungen und Problemen auch bei der Big Bang Migration sehr eingeschränkt. Der hier vorgestellte Ansatz ist im Gegensatz dazu agil. Die starren Phasen werden aufgelöst, während ein System entwickelt wird, kann ein anderes schon getestet und ein drittes bereits in Produktion gebracht werden. Treten bei der Migration eines Systems Probleme auf, wirkt sich dies nicht auf die Projektplanung der anderen Systeme aus. Deshalb haben wir unseren Ansatz Agile Migration getauft.

Die Kernidee bei dem Verfahren ist es, die einzelnen Systeme voneinander zu entkoppeln. Sind die Systeme entkoppelt, können sie in einer beliebigen Reihenfolge migriert werden. Um die Entkopplung zu erreichen wird ein sogenannter Adapter zwischen die Systeme eingebaut. Dieser Adapter wächst mit der Migration und verbindet die bereits migrierten Systeme und die Altsysteme die sich noch in Produktion befinden. Er fungiert sozusagen als Übersetzer zwischen neuen und alten Systemen, die Daten kommen immer noch von den Systemen selbst, der Adapter enthält die Logik die notwendig ist, um für jede Anfrage die notwendigen Daten zu holen und ins richtige Format zu bringen. Wenn nötig kann der Adapter auch Daten verwalten, die von Altsystemen noch gebraucht werden, nach der kompletten Migration aber wegfallen. Indem man die Aufgabe zwischen neuen und alten Systemen zu vermittlen in eine externe Komponente verlagert, muss beim Entwurf der neuen Systeme nur der Endzustand betrachtet werden. Sobald die Migration abgeschlossen ist, kann der Adapter abgeschaltet werden. Es verbleibt somit kein Code in Produktion, der nur dazu dient, mit abgelösten Systemen zu kommunizieren. Durch den Adapter wird auch der Druck genommen die Migration in einem engen Zeitrahmen umzusetzen: jedes System kann

migriert werden, sobald es fertig implementiert wurde. Dadurch werden die Projektpläne der einzelenen Systeme voneinander entkoppelt und Verzögerungen wirken sich somit nicht mehr projektübergreifend aus. Es ist ebenfalls möglich neue und alte Systeme parallel laufen zu lassen, etwa weil die Daten nicht auf einen Schlag, sondern nach und nach migriert werden. Insofern kann die Agile Migration auch bei der Datenmigration den Big Bang Ansatz unnötig machen.

4.2 Umsetzung

Bei der Agilen Migration werden zuerst die Systeme identifiziert, die migriert werden sollen. Im Normalfall migriert man zuerst das größte System, welches oft das zentrale System mit den meisten Schnittstellen ist. Im weiteren Verlauf nennen wir dieses System Hauptsystem. Dieses System ist oft auch der eigentliche Fokus der Migration, es hat die meisten Abhängigkeiten und ist am schwierigsten zu migrieren. Die Systemlandschaft vor und nach der Migration wird festgelegt. Die Umsetzung des neuen Hauptsystems und des Adapters kann zeitgleich starten. Als ersten Migrationsschritt kann der Adapter in das intiale System integriert werden und dort bereits die zu migrierenden Systeme entkoppeln. Gerade wenn die Dokumentationslage der abzulösenden Systeme dürftig ist, gibt dieser Schritt die Möglichkeit die Daten zu analysieren die zwischen den Systemen fließen. Zu diesem Zweck können Methoden aus der Big Data Analyse im Adapter verwendet werden um die Kommunikation zwischen den Systemen zu kategorisieren. Damit kann sichergestellt werden, dass wichtige Use Cases nicht übersehen werden und dass die Anfragen, die der Adapter an die alten Systeme schickt auch von diesen verstanden werden. Diese Informationen sind mitunter für das neue Hauptsystem elementar. Oft ist die Dokumentation von alten Systemen in einem schlechten Zustand, so dass es keine gesicherten Erkenntnisse gibt, welche Daten benötigt werden oder welche Abhängigkeiten bestehen. In diesem Modus beobachtet der Adapter nur, und leitet die Daten 1:1 weiter.

Im nächsten Schritt erfolgt dann die Migration des Hauptsystems. Das neue Hauptsystem kommuniziert nun direkt mit den von der Migration nicht betroffenen Systemen und über den Adapter mit den noch zu migrierenden Systemen. Auch in diesem Schritt können die Daten, die durch den Adapter fließen, analysiert werden, um frühzeitig Probleme zu finden. Zu diesem Zeitpunkt sind keine weiteren Arbeiten am Adapter notwendig, wenn nicht weitere Systeme in den Fokus der Migration rücken.

Als nächstes Folgen nun Schritt für Schritt die weiteren Systeme. Migrierte Systeme werden direkt an das neue Hauptsystem angeschlossen. Nach der Migration des letzten Systems kann der Adapter abgeschaltet werden. Wenn die Migration eines Systems erst viel später erfolgen kann, kann der Adapter auch über längere Zeit in Produktion eingesetzt werden.

4.3 Umgang mit Verzögerungen

Der Big Bang Ansatz ist besonders empflindlich wenn bei der Implementierung eines der neuen Systeme Verzögerungen entstehen. Da die Migration viele Systeme umfasst, ist

eine Verzögerung teuer. Die alten Systeme müssen laufen, bis alle neuen Systeme fertig umgesetzt sind, eventuell vorhandene neue Features können nicht genutzt werden. Oft genug endet dies in vielen Eskalationsmeetings. Bei der Agile Migration ist die Lage etwas entspannter. Da jedes System einzeln migriert wird, wirken sich Verzögerungen nur auf dieses System aus. Sobald ein System fertig entwickelt wurde kann es produktiv gebracht werden. Auch hier können natürlich Probleme entstehen, gerade wenn zwischen den neuen Systemen Abhängigkeiten bestehen. Angenommen zwei Systeme A und B sollen noch migriert werden, dabei greift B auf Schnittstellen von A zu. Deshalb war vorgesehen, zuerst A, dann B zu migrieren. Durch eine Verzögerung bei der Implementierung wird aber System A nicht vor System B fertig. Statt mit der Migration von B zu warten, kann man die Schnittstellen von A im Adapter bereitstellen, und so bereits früher B in Produktion bringen. Dieser flexible Ansatz ist besonders dann sinnvoll, wenn die Verzögerungen groß oder die Schnittstellen einfach zu implementieren sind. Auch im Falle von Abhängigkeiten die spät entdeckt werden, kann der Adapter schnelle Abhilfe schaffen, bis eine Lösung im eigentlichen System bereitgestellt werden kann. Da der Adapter ein Wegwerf-Projekt ist, können hier auch Migrations-Workarounds, die man im fertigen System nicht haben möchte, untergebracht werden.

4.4 Der Adapter

Offensichtlich ist das beschriebene Verfahren sehr abhängig von der Qualität des Adapters. Viele der Unsicherheiten des Projekts lasten auf dieser Komponente, da er die Kommunikation mit den Legacy-Systemen leisten muss, für die es unter Umständen keine Dokumentation mehr gibt. Da alle Kommunikation über den Adapter läuft, muss er auch sehr performant sein, und darf die anderen Systeme nicht ausbremsen. Da er eine zentrale Rolle innerhalb der Infrastruktur einnimmt, ist es wichtig, dass diese Komponente besonders ausfallsicher ist. Ein weiterer Punkt den man bedenken sollte, ist dass der Adapter viele verschiedene Schnittstellen unterstützen können muss, selten sind Legacy-Systeme homogen in der Verwendung von Schnittstellen-Technologien. Der Adapter sollte auch sehr leicht zu erweitern und zu warten sein. Durch die Natur des Anwendungsfalls werden während der Migration sichere Annahmen über Workflows über den Haufen geworfen, und der Adapter muss schnell in der Lage sein, sich auf neue Gegebenheiten anzupassen. Die Agile Migration steht und fällt mit einer professionellen Umsetzung des Adapters.

5 Zusammenfassung

Wenn es um die Migration von komplexen Systemen geht, wird häufig der Ansatz der Big-Bang Migration eingesetzt. Dieser Ansatz bringt aber viele Probleme mit sich, und ist sehr unflexibel. In vielen Situationen kann hier die Agile Migration helfen, die viele Probleme des Big Bang Ansatzes löst und sehr flexibel mit Risiken und Problemen umgeht. Bei der Agilen Migration wird die Migration dadurch vereinfacht, dass Systeme mittels einer Adapter-Komponente voneinander entkoppelt werden, was eine flexible Migration erlaubt. Der Adapter verbindet bereits migrierte Systeme mit noch nicht

migrierten Systemen und fungiert als Schnittstellen-Übersetzer. Nach der fertigen Migration wird der Adapter abgeschaltet, so dass keinerlei Spuren der alten Schnittstellen in der Systemlandschaft verbleiben.

Geschäftsprozessflexibilität durch Service-orientierte Architekturen – grundlegende Zusammenhänge und weiterer Forschungsbedarf

Frank Termer, Volker Nissen, Nathalie Wessels

Fachgebiet Wirtschaftsinformatik für Dienstleistungen
Technische Universität Ilmenau, PF 10 05 65, 98684 Ilmenau
{frank.termer, volker.nissen}@tu-ilmenau.de

Abstract: Bei der Einführung einer Service-orientierten IT-Architektur (SOA) wird meist argumentiert, diese trage zu der heute geforderten hohen Unternehmensflexibilität bei. Dieser behauptete Zusammenhang wird im vorliegenden Beitrag näher untersucht. Ausgehend von der Überlegung, dass eine flexible IT-Architektur nicht direkt zur Unternehmensflexibilität beitragen kann, wird unterstellt, dass diese zu einer Flexibilisierung von Geschäftsprozessen führt und damit indirekt auf die Unternehmensflexibilität wirkt. Der Beitrag trägt daher Eigenschaften flexibler Prozesse und flexibler IT-Architekturen zusammen und stellt sie gegenüber. Daran anknüpfend wird untersucht, in wie fern eine SOA die für die Flexibilisierung von Prozessen nützlichen Eigenschaften flexibler IT-Architekturen erfüllt. Daraus wird weiterer Forschungsbedarf abgeleitet.

1 SOA als Mittel zur Erreichung von Unternehmensflexibilität

1.1 SOA-Begriff, Forschungsfragen und Zielstellung

Sich schnell an veränderte Rahmenbedingungen anpassen zu können, stellt für Unternehmen eine große aktuelle Herausforderung dar. Diese Fähigkeit hängt in zunehmendem Maße von der Veränderungsfähigkeit der Unternehmens-IT ab [NRT12]. Die Mehrheit der Unternehmen investiert daher in die Modernisierung ihrer IT-Architektur. Die Architektur beschreibt die grundlegende Organisation eines Systems, „welche sich in den Komponenten und deren Beziehungen zueinander sowie zum Systemumfeld manifestiert“ [HS11]. Generelles Ziel der Architekturgestaltung ist es, die Flexibilität in der IT zu steigern und damit die Unternehmensflexibilität durch die IT zu erhöhen. Als geeignetes Mittel werden häufig Service-orientierte Architekturen (SOA) angeführt [RHS05]. Eine SOA wird definiert als „offene, agile, erweiterbare, verbundene, komponierbare Architektur, die aus autonomen, […], interoperablen, auffindbaren und potenziell wiederverwendbaren Services besteht“ [Er05]. Ein Service ist ein Softwareelement mit einer klar definierten Leistung [KBS05]. Informationen, die zur Identifikation und Nutzung geeigneter Services benötigt werden, sind in einem Service-Repository abgelegt. Die Verbindung zwischen Services sowie zwischen Services und Application Frontends, d. h. den Schnittstellen nach außen, wird durch einen Service-Bus hergestellt [Re07]. Dass durch den Einsatz von SOA auch tatsächlich Unternehmensflexibilität erreicht wird, ist allerdings häufig nur eine Behauptung ohne Belege [Be11]. Eine

strukturierte Untersuchung anhand definierter Kriterien, ob und inwieweit SOA die Flexibilität in Unternehmen tatsächlich erhöhen kann, ist bislang nicht erfolgt. Da der Einsatz von SOA jedoch mit hohen Investitionen verbunden ist und meist mit der Steigerung der Flexibilität gerechtfertigt wird [Co07], bedarf gerade diese Frage einer genaueren Betrachtung. In Analogie zur Diskussion, welchen Wertbeitrag die IT zum Unternehmenserfolg leisten kann [WPR97] unterstellen wir, dass es keinen direkten Wirkungszusammenhang zwischen einer flexiblen Unternehmens-IT (z. B. in Form einer SOA) und der geschäftlichen Flexibilität eines Unternehmens gibt. Eine flexible Unternehmens-IT kann allerdings zur Flexibilisierung von Geschäftsprozessen beitragen, wodurch wiederum die Flexibilität des Unternehmens unterstützt wird [PM11]. Die grundsätzlich positive Wirkung einer SOA auf die Flexibilität eines Unternehmens wird damit nicht in Frage gestellt, allerdings soll die Art und Stärke des Zusammenhangs transparenter gemacht und auf eine argumentativ abgesicherte Basis gestellt werden. Wir fragen daher: In welchem Umfang kann SOA zur Flexibilisierung von Geschäftsprozessen beitragen? Zunächst gilt es zu untersuchen, welche Merkmale flexible Geschäftsprozesse (GP) charakterisieren. Weiterführend ist zu untersuchen, durch welche Merkmale flexible IT-Architekturen definiert sind und wie diese durch SOA erfüllt werden. Dem folgend kann dann untersucht werden, welche Eigenschaften flexibler Architekturen die definierten Merkmale flexibler GP direkt unterstützen. Damit kann abschließend die Forschungsfrage beantwortet werden, in welchem Maße SOA ein geeignetes Konzept ist, um die Flexibilität von GP positiv zu beeinflussen.

1.2 Methodisches Vorgehen

Als Grundlage der Untersuchung wurde eine Literaturanalyse nach FETTKE durchgeführt [Fe06]. Die Phase der Literatursuche wurde unter Berücksichtigung der Schlagworte Flexibilität, Service-orientierte Architektur, SOA, Geschäftsprozess und Prozess vorgenommen. Dabei fokussieren wir zum einen auf die Publikationsorgane der deutschsprachigen Wirtschaftsinformatik und legen die Zeitschriften Wirtschaftsinformatik sowie HMD – Praxis der Wirtschaftsinformatik zugrunde. Weiterhin werden die Tagungsbände der Internationalen Tagung Wirtschaftsinformatik sowie der Multikonferenz Wirtschaftsinformatik für die Untersuchung herangezogen. Zum anderen wurde die Schlagwortsuche auch in wissenschaftlichen Datenbanken vorgenommen, um Monographien, Sammelbände, Dissertationen und Habilitationen einschließen zu können. Als Recherchezeitraum wurden die Jahre 2000 bis 2012 betrachtet. Durch die Wiederholung einer vorwärts- und rückwärtsgerichteten Suche in relevanten Publikationen werden weitere Quellen erschlossen, die den aufgestellten Kriterien entsprechen.

2 Zur Flexibilität von Geschäftsprozessen

2.1 Begriffliche Grundlagen

WAGNER ET AL. definieren Flexibilität als „die Fähigkeit eines Systems, auf System- oder Umweltveränderungen unter Berücksichtigung gegebener Ziele durch Anpassung von Struktur und / oder Verhalten zu reagieren oder sie zu antizipieren“ [Wa11].

Strukturflexibilität ist gegeben, wenn Bestandteile eines Systems hinzugefügt oder ausgetauscht werden können, z. B. Vor- oder Nachereignisse, Sach- oder Formalziele oder auch Bestandteile des Lösungsverfahrens in einem System. Verhaltensflexibilität liegt dann vor, wenn Ereignisse zu einem anderen Zeitpunkt eintreten können, Sach- und Formalziele in der inhaltlichen Ausgestaltung variieren können oder Attribute bzw. Variablen bei der Verarbeitung veränderbar sind. Diese Definition ist als Ausgangspunkt geeignet, da sie sowohl für den Begriff der Architektur anwendbar ist (Strukturflexibilität) als auch für den Begriff des Prozesses (Struktur- und Verhaltensflexibilität).

Ein Prozess stellt „die inhaltlich abgeschlossene, zeitliche und sachlogische Abfolge der Funktionen dar, die zur Bearbeitung eines betriebswirtschaftlich relevanten Objekts ausgeführt werden“ [Ro96]. Bei einem GP handelt es sich in der Fortführung um einen speziellen Prozess, welcher der Erreichung der Unternehmensziele dient und das zentrale Geschäftsfeld eines Unternehmens beschreibt [BK08]. Dabei bestehen GP aus einer funktions- und organisationsüberschreitenden Verknüpfung wertschöpfender Aktivitäten, die auf die Erfüllung von Kundenanforderungen abzielen und die Prozessziele umsetzen, die aus der Geschäftsstrategie abgeleitet wurden [SS08].

2.2 Eigenschaften flexibler Geschäftsprozesse

In der Übertragung des Flexibilitätsbegriffs auf den Gegenstand des Prozesses [Ba11] stellt sich ein Prozess dann als inflexibel dar, wenn er ein singuläres Verhalten und eine exakt vorgegebene Struktur aufweist. Verhaltensflexibilität eines Prozesses zeigt sich hingegen z. B. in der Veränderung des Outputs bezüglich Art, Menge oder Zeit. Von Strukturflexibilität eines GP kann gesprochen werden, wenn er durch die Modifikation seiner Komponenten oder deren Beziehungen angepasst werden kann. Soweit nachfolgende Eigenschaften in der operativen Ausführung eines Prozesses vorliegen, ist es gerechtfertigt, diesen als flexibel zu bezeichnen [Al07]:

- Neue Aufgaben können zu Prozessen hinzugefügt werden.
- Nicht mehr gebrauchte Aktionen können gelöscht / entfernt werden.
- Die Reihenfolge von Aufgaben innerhalb eines Prozesses kann verändert werden.
- Es können neue Routen durch den Prozess definiert werden.
- Rollen innerhalb eines Prozesses können angepasst werden.
- Bedingungen zur Ablaufsteuerung können verändert werden.
- Die alternative Nutzung von Ressourcen innerhalb eines Prozesses ist möglich.

Damit Prozesse die aufgeführten Eigenschaften erfüllen, ist es notwendig, dass sie Anforderungen adressieren können, die sich aus Flexibilitätsbedarfen ergeben. Die Literaturanalyse hat hierzu nachfolgende Anforderungen an flexible Prozesse ergeben:

1. Flexible Prozesse können eine Überlappung von Prozessplanung und Prozessausführung berücksichtigen [FLW11]. Ein Prozess ist dann flexibel, wenn er nicht notwendigerweise vollständig im Vorfeld seiner Durchführung definiert ist, sondern sich die Prozessausgestaltung erst bei der tatsächlichen Ausführung ergibt, wobei die Planung von Prozessbestandteilen erst während der Prozessausführung erfolgt. Es ist

daher nicht notwendig (und oft nicht möglich), alle Prozessvarianten im Vorfeld zu definieren, sondern es sollten gezielt Stellen im Prozess vorgesehen werden, an denen der weitere Prozessablauf erst während der Durchführung festgelegt wird.

2. Flexible Prozesse können die Kontextsensitivität des Prozesses berücksichtigen [WF10]. Der Kontext eines Prozesses beschreibt die Umweltbedingungen, die den Prozessablauf beeinflussen. Durch die Betrachtung des Kontextes können mehrdeutige Situationen aufgelöst werden. Das erfordert eine verbesserte Abbildungstreue von Prozessmodellen, was die Komplexität der Modelle erhöht.

3. Flexible Prozesse können die beteiligten Mitarbeiter des Prozesses integrieren [Al98, BK08] bzw. erlauben Selbstorganisation der beteiligten Personen [Ku10]. Die kreative Beteiligung der Mitarbeiter motiviert diese durch zusätzliche Verantwortung, einen größeren Handlungsspielraum und zusätzliche Erfolgserlebnisse [BK08]. Durch diese Integration müssen Prozesse nicht so detailliert geplant werden, sondern es genügt Zielvorgaben zu setzen; der tatsächliche Ablauf kann dann schnell von den handelnden Personen verändert werden.

4. Flexible Prozesse erkennen Umweltveränderungen [Al98, Al09]. Veränderte Umweltbedingungen können Prozessanpassungen erfordern. Somit ist das Erkennen von Umweltveränderungen wesentlich dafür, wie flexibel ein Prozess tatsächlich ist. Mittels eines Umweltmonitorings werden Umfeldvariablen beobachtet, ausgewertet und Rückschlüsse zur Anpassung von Prozessen gezogen.

5. Flexible Prozesse besitzen Referenzwissen [Al98, Al09]. Um Prozessveränderungen umsetzen zu können, werden Kenntnisse über aktuell implementierte Prozesse benötigt. Dies beinhaltet u. a. Gründe für Prozessveränderungen, Annahmen über Ausprägungen von Umweltfaktoren, relevante Einflussfaktoren und Ziele. Durch das Vorhalten dieser Wissensbasis können Fehler bei der Prozessanpassung vermieden werden, die in der Vergangenheit evtl. bereits gemacht wurden.

6. Flexible Prozesse können durch die Verbindung von unternehmensinternen Prozessbausteinen erstellt werden (Prozessorchestrierung; [Al09, We12]). Bei der Orchestrierung werden die Aktivitäten von Prozessen, die Bedingungen ihrer Ausführung und deren Beziehungen untereinander betrachtet. Damit ergeben sich Flexibilitätspotenziale für Prozesse, da die tatsächliche Abfolge innerhalb eines Prozesses durch die Orchestrierung gestaltet werden kann.

7. Flexible Prozesse können mit anderen Prozessen gekoppelt und entkoppelt werden (Prozesschoreografie; [Al09, We12]). Das Prinzip entspricht der Orchestrierung, wird allerdings auf höherem Aggregationsniveau betrachtet. Eine Prozesschoreografie fokussiert Beziehungen zwischen verschiedenen Prozessorchestrierungen und ermöglicht die Festlegung der Zusammenarbeit von Prozessen untereinander.

8. Flexible Prozesse besitzen eine geringe Komplexität [SW05]. Wenn Prozesse wenig komplex sind, lassen sich Änderungen relativ schnell umsetzen. Die Komplexität bezieht sich insbesondere auf die Zahl der Prozessvarianten. Durch die Modularisierung von Prozessen, wie diese für die Prozessorchestrierung und Prozess-

choreografie notwendig ist, steigt allerdings die Komplexität von Prozessen an, so dass diese Eigenschaften in einem gegensätzlichen Verhältnis zueinander stehen.

Die vorgestellten Eigenschaften beeinflussen sich z. T. gegenseitig. Insbesondere kann die Umsetzung der Anforderungen eins bis sieben die Komplexität erhöhen, so dass bei Berücksichtigung aller Eigenschaften eine entsprechende Balance zwischen Flexibilität und der sich daraus ergebenden Komplexität gefunden werden muss.

3 Eigenschaften flexibler IT-Architekturen

Um zu untersuchen, in wie fern flexible Prozesse durch eine SOA unterstützt werden, ist zunächst zu klären, welche generellen Eigenschaften von IT-Architekturen die Flexibilität von Prozessen unterstützen. Anschließend kann das Vorhandensein dieser Architektureigenschaften bei SOA betrachtet werden. Unter der Architektur eines Systems wird „dessen Organisation […], welche sich in den Komponenten und deren Beziehungen zueinander sowie zum Systemumfeld manifestiert“ verstanden [HS11]. Die Informationssystemarchitektur umfasst Aufbau- und Ablauforganisation, Anwendungssysteme und Daten sowie technische Infrastruktur [HS11]. Diese Sicht ist für die weiteren Betrachtungen geeignet, da er sowohl Prozesse als auch technische Gegebenheiten beinhaltet. Das Architekturmanagement (AM) hat insbesondere die Weiterentwicklung einer bestehenden Architektur hin zu einer Soll-Architektur zum Ziel. Dabei kommt dem AM neben der Beschreibungs- und Kommunikationsfunktion auch eine Gestaltungsfunktion zu [HS11], bei der konkrete Handlungen zur Weiterentwicklung einer Architektur entworfen werden. Hierbei wird gefordert, dass eine Architektur flexibel gestaltet sein soll, um die schnelle Anpassung an veränderte Anforderungen zu ermöglichen. Aus der Literatur wurden folgende Eigenschaften flexibler Architekturen identifiziert:

- Die Grundsätze ordnungsmäßiger Modellierung (GoM) können auch für die Gestaltung von Architekturen herangezogen werden [HS11]. Um Flexibilität bereitzustellen, ist v. a. der Grundsatz des systematischen Aufbaus wesentlich, da sich hierdurch Komponenten mit klar definierten Schnittstellen ergeben.
- Die Modularität beschreibt das Zerlegen eines Systems in kleine, teilautonome Subsysteme [AGS05] und weist damit Ähnlichkeiten zur Forderung nach einem systematischen Aufbau auf. Die Funktionsweise eines Moduls wird dabei gekapselt und die Module werden über (möglichst wenige) Schnittstellen entkoppelt.
- Aus der Modularität ergeben sich die Prinzipien der losen Kopplung und der hohen Kohäsion [NRT12]. Kopplung beschreibt die Stärke der Verbindungen bzw. den Grad der Abhängigkeit der Komponenten voneinander. Lose Kopplung liegt vor, wenn der innere Zusammenhang einer Komponente hoch und die Verbindung zwischen Komponenten gering ist, wodurch leichtere Konfigurationsmöglichkeiten entstehen. Für eine flexible Architektur wird die Entwicklung kombinierbarer, wiederverwendbarer und leicht änderbarer Module gefordert [AS04].
- Die Interoperabilität [AGS05] bezeichnet die Fähigkeit von Anwendungen zusammen zu arbeiten, unabhängig von der technischen Realisierung. Inter-

operabilität verlangt, dass Komponenten einer Architektur ein hohes Maß an Kompatibilität zur Verfügung stellen.

- Bei Selbstähnlichkeit [AGS05, Gr00] einer Architektur bringt diese durch Zusammenlegen und Aufteilen von Subsystemen immer wieder gleiche Muster hervor. Eine solche Vorgehensweise reduziert die Komplexität des Gesamtsystems und trägt somit zu einer erhöhten Flexibilität bei.
- Selbstorganisation [AGS05, Gr00] bedeutet, dass ein System in der Lage ist, seine Struktur zu verändern, in dem Informationen aus der Umwelt aufgenommen, hieraus Schlussfolgerungen für die Systemstruktur gezogen und umgesetzt werden.
- Wissen über die Architektur [AGS05] trägt zur Flexibilität von Systemen bei, da prinzipielle Möglichkeiten der Verwendung von Komponenten, der Abbildung von Prozessen und der Nutzung von Teilsystemen bekannt sind. Damit lassen sich Änderungswünsche flexibel umsetzen.
- Eine geringe Komplexität und die Redundanzfreiheit [NRT12] führen zu einer leichteren Anpassbarkeit von Systemen, da es nur noch wenige Stellen innerhalb einer Architektur gibt, an denen Änderungen vorgenommen werden müssen.
- Parametrierbarkeit [NRT12] beschreibt die Möglichkeit, Änderungen an einem System ohne Programmierung (und damit zügig) vornehmen zu können.

4 Beitrag flexibler Architekturen zur Flexibilisierung von Prozessen

Da nun die Kriterien flexibler Prozesse und die Eigenschaften flexibler Architekturen herausgearbeitet wurden, können diese nachfolgend gegenübergestellt werden, um zu prüfen, welche Architektureigenschaften den Kriterien flexibler GP zugeordnet werden können (vgl. Abbildung 1). Die Beurteilung der Wirkung zwischen diesen beiden Dimensionen erfolgt dabei durch eine argumentativ-deduktive Diskussion, wobei damit kein Anspruch auf Abgeschlossenheit bzw. Vollständigkeit erhoben werden soll.

Der systematische Aufbau einer Architektur unterstützt die Prozessorchestrierung und die Prozesschoreografie positiv, da für diese die Existenz von wohldefinierten Schnittstellen eine Grundvoraussetzung ist. Ebenso wird die Integration der Mitarbeiter positiv beeinflusst, da durch einen systematischen Aufbau ein besseres Verständnis für die Strukturen bei den Mitarbeitern geschaffen werden kann. Wenn nun die Mitarbeiterintegration bei Prozessen zunimmt, so trägt ein systematischer Aufbau zur Beschleunigung von Veränderungen bei.

Modularität, lose Kopplung und hohe Kohäsion unterstützen die Überlappung der Planung und Ausführung von Prozessen, da besser mit Planungsunsicherheit umgegangen werden kann. Auch kann der Kontextsensitivität von Prozessen durch Modularität besser begegnet werden. Ebenso unterstützt Modularisierung die Prozessorchestrierung und Prozesschoreografie, da eine Komposition von klar abgegrenzten Modulen leichter möglich ist. Durch die Kombination von Modulen können einfach neue Prozesse zusammengestellt oder bereits bestehende Prozesse miteinander verbunden werden.

Kriterien flexibler GP / Eigenschaften flexibler Architekturen	Überlappung von Planung und Ausführung	Kontextsensitivität	Integration der Mitarbeiter	Umweltmonitoring	Referenzwissen	Prozessorchestrierung	Prozesschoreografie	Komplexitätsreduktion
Systematischer Aufbau			✓			✓	✓	
Modularität	✓	✓				✓	✓	✗
lose Kopplung	✓	✓				✓	✓	✗
hohe Kohäsion	✓	✓				✓	✓	✗
Interoperabilität						✓	✓	
Selbstähnlichkeit								✓
Selbstorganisation			✓	✓				
Wissen über Architektur			✓		✓			
geringe Komplexität								✓
Parametrierbarkeit								✓

Legende: ✓ Eigenschaft wirkt positiv ✗ Eigenschaft wirkt negativ; leer = kein Zusammenhang

Abbildung 1: Eigenschaften flexibler Architekturen und Kriterien flexibler Geschäftsprozesse

Die Modularisierung steht der Komplexitätsreduktion entgegen, da sich durch das Vorhandensein von Modulen zahlreiche Kombinationsmöglichkeiten ergeben und somit prinzipiell eine verstärkte Variantenbildung möglich ist. Durch Interoperabilität innerhalb einer Architektur können innerbetriebliche und überbetriebliche Prozesse gebildet werden, indem mehrere Anwendungssysteme miteinander verbunden werden. Die Selbstorganisation beinhaltet, dass Systeme Informationen über ihre Umwelt aufnehmen und verarbeiten können. Somit liegt eine Unterstützung der Eigenschaft des Umweltmonitorings vor. Gleichzeitig wirken auch Mitarbeiter auf Systeme ein, so dass Menschen aus Sicht von Architekturen ebenfalls der Umwelt zuzurechnen sind. Wenn nun die Gestaltung von Prozessen durch Selbstorganisation in der IT-Architektur mit mehr Freiheitsgraden versehen wird, nimmt die Integration der Mitarbeiter zu, da diese die Freiräume durch die Wahl der jeweils besten Möglichkeit entsprechend nutzen können. Folglich unterstützt Selbstorganisation auch die Mitarbeiterintegration. Durch Selbstähnlichkeit innerhalb einer Architektur ergeben sich gleiche und wiederkehrende Muster, die zu einer Reduzierung der Komplexität beitragen. Die Forderung, nach der Wissen über die Architektur vorliegen muss, unterstützt die Integration der Mitarbeiter, die als genereller Wissensträger auch in Prozesse eingebunden sind. Gleichzeitig trägt das Wissen über die Architektur einen Teil zum Referenzwissen von Prozessen bei und unterstützt folglich diese Eigenschaft flexibler Prozesse. Die geringe Komplexität und die Parametrierbarkeit innerhalb von Architekturen wirken positiv auf die Prozesseigenschaft der Komplexitätsreduktion. Insgesamt konnte allen Kriterien flexibler Prozesse mindestens eine Eigenschaft flexibler Architekturen mit einem positiven Zusammenhang zugeordnet werden. Damit kann angenommen werden, dass flexible Architekturen prinzipiell zur Flexibilisierung von GP beitragen.

5 Erfüllt SOA die Kriterien flexibler Architekturen?

Abschließend gilt es zu untersuchen, in wie weit eine SOA die (abstrakten) Eigenschaften flexibler Architekturen konkret umsetzt. Das zentrale Konstruktionsprinzip einer SOA sind lose gekoppelte Services, die Geschäftsfunktionen realisieren, welche über eine implementierungsunabhängige Schnittstelle gekapselt sind [TS07]. Generell können Systeme und Anwendungen zeitliche und örtliche Abhängigkeiten, aber auch Struktur-, Implementierungs- oder Datenabhängigkeiten aufweisen. Dies wird durch eine lose Kopplung der Services verhindert, da jeder Service lediglich eine Schnittstelle zur Verfügung stellt und seine Implementierung nach außen verbirgt. Die Funktionalität und der Datenaustausch eines Service werden allein über dessen Schnittstelle angeboten. So kann die Implementierung eines Services geändert oder ausgetauscht werden, ohne dass die Servicenutzer dies bemerken. In Abbildung 2 sind die Eigenschaften flexibler Architekturen den Eigenschaften einer SOA gegenübergestellt.

Eigenschaften flexibler Architekturen / **Eigenschaften einer SOA**	Systematischer Aufbau	Modularität	lose Kopplung	hohe Kohäsion	Interoperabilität	Selbstähnlichkeit	Selbstorganisation	Wissen über Architektur	geringe Komplexität	Parametrierbarkeit
lose gekoppelte Services		✓	✓		✓					
Entwurfsstandards	✓					✓			✓	
Modularität		✓								
jeder Service kann Teil einer SOA werden	✓	✓		✓	✓					
Service-Broker						✓	✓			
Service-Bus	✓				✓				✓	✓
Servicedokumentation	✓							✓	✓	
Servicekatalog	✓							✓		
Kapselung									✓	✓
Aufruf über definierte Schnittstellen	✓				✓					
Legende: ✓ trägt zu Umsetzung bei; leer = kein Zusammenhang										

Abbildung 2: Umsetzung der Eigenschaften flexibler Architekturen durch SOA

Die Eigenschaft des systematischen Aufbaus wird durch eine SOA erfüllt, denn es existieren Entwurfsstandards, nach denen eine SOA entwickelt und entworfen werden sollte. Dazu zählen u. a. der Aufruf einzelner Services durch einen einheitlichen Mechanismus sowie die Verwendung eines Service-Bus [RHS05]. Die Schnittstellen von Services sind gleich konstruiert und werden mittels standardisierter Beschreibungssprachen definiert.

Die Eigenschaft der Modularität wird ebenfalls durch eine SOA erfüllt. Die einzelnen Services innerhalb einer SOA bilden die geforderten Teilsysteme. Jeder Service erfüllt einen spezifischen Zweck. Durch die per Definition festgelegten Eigenschaften der losen Kopplung von Services, kann die Modularität als erfüllt angesehen werden. Weiterhin sollen Services so entworfen werden, dass jeder Service potenziell Teil einer SOA werden kann. Somit soll ein redundantes Angebot von Funktionalitäten verhindert werden und die Forderung nach Wiederverwendung und damit der hohen Kohäsion in Zusammenhang mit dem systematischen Aufbau führt zu einer positiven Interoperabilität von Services. Somit erfüllt eine SOA auch dieses Prinzip. Die Eigenschaften der Selbstähnlichkeit und der Selbstorganisation werden ebenso durch eine SOA abgebildet. Die Selbstähnlichkeit von Services ist durch die Verwendung von Konstruktionsstandards gegeben. Ebenso kann die Selbstorganisation durch die Verwendung eines Service-Brokers abgebildet werden [We12]. Dieser kann bei Bedarf aus mehreren Services einen passenden auswählen und in die SOA einbinden. Jeder Service wird mit einer umfangreichen, nach definierten Standards erstellten Dokumentation ausgestattet. Darin sind eine fachliche Beschreibung der Funktion und eine technische Beschreibung des Aufrufs enthalten. Mittels der Verwendung von Service-Katalogen kann zudem ein Überblick erhalten werden, welche Services in welcher Form in einer SOA eingebunden sind. Diese Eigenschaften einer SOA erfüllen die Forderung nach einem guten Wissen über die Architektur. Durch die Kapselung der technischen Realisierung eines Service, die Verwendung eines Service-Bus und die Berücksichtigung von Standards bei der Service-Beschreibung und Konstruktion wird eine geringe Komplexität der Architektur unterstützt. Abschließend wird auch die Forderung nach Parametrierbarkeit durch eine SOA erfüllt. Wie bereits beschrieben, verbergen Services die technische Implementierung ihrer Funktionalität und machen lediglich Angaben über deren Aufruf und deren Verwendung. Parameter, die für die Funktionalität eines Service notwendig sind, werden beim Aufruf mit übergeben. Sollten nun Änderungen notwendig werden, so können die Aufrufparameter entsprechend gesetzt werden, ohne die eigentliche Funktionalität des Service zwingend ändern zu müssen.

In Abbildung 3 werden nun die beiden Tabellen aus Abbildung 1 und Abbildung 2 zusammengeführt, indem ein Mapping zwischen den Eigenschaften einer SOA und den Kriterien flexibler GP über die Eigenschaften flexibler Architekturen erfolgt. Dazu wird in Abbildung 1 bei jeder Zelle, die eine positive oder negative Wirkung darstellt, in Abbildung 2 nachgesehen, durch welche Eigenschaften einer SOA diese Wirkung erzielt werden kann. Nachfolgend werden nur besonders erklärungsbedürftige Zusammenhänge diskutiert, da die positiv erkannten Zusammenhänge als plausibel nachvollziehbar angesehen werden. Es fällt auf, dass ein Umweltmonitoring lediglich durch den Service-Broker erreicht werden kann. Problematisch daran ist, dass bei einer SOA kein automatisierter Beobachtungsprozess der Umwelt stattfindet, der Service-Broker also nicht selbständig handelt. Der Anstoß zur Selbstorganisation muss durch einen Menschen gegeben werden. Somit wird die Forderung nach einem automatischen Umweltmonitoring von flexiblen GP nur bedingt durch eine SOA erfüllt. Ein zweiter kritischer Bereich ist dort zu erkennen, wo Eigenschaften einer SOA einen negativen Einfluss auf geforderte Kriterien flexibler GP ausüben. Dies ist bei der geforderten Komplexitätsreduktion der Fall. In der praktischen Umsetzung einer SOA muss darauf geachtet werden, dass Services nicht zu feingranular definiert werden. Dies würde zwar

die Wiederverwendbarkeit einzelner Services erhöhen, allerdings kann dadurch die Komplexität der Architektur steigen und damit insbesondere die Verhaltensflexibilität von Prozessen negativ beeinflusst werden.

Kriterien flexibler GP / **Eigenschaften einer SOA**	Überlappung von Planung und Ausführung	Kontextsensitivität	Integration der Mitarbeiter	Umweltmonitoring	Referenzwissen	Prozessorchestrierung	Prozesschoreografie	Komplexitätsreduktion
lose gekoppelte Services	✓	✓				✓	✓	✗
Entwurfsstandards			✓			✓	✓	✓
Modularität	✓	✓				✓	✓	✗
jeder Service kann Teil einer SOA werden	✓	✓	✓			✓	✓	✗
Service-Broker			✓	(✓)				✓
Service-Bus			✓			✓	✓	✓
Servicedokumentation			✓		✓	✓	✓	✓
Servicekatalog			✓		✓	✓	✓	
Kapselung								✓
Aufruf über definierte Schnittstellen			✓			✓	✓	
Legende: ✓ Eigenschaft wirkt positiv ✗ Eigenschaft wirkt negativ; leer = kein Zusammenhang								

Abbildung 3: Erfüllungsgrad der Kriterien flexibler GP durch Eigenschaften einer SOA

6 Abschließende Betrachtungen

Die Analysen in diesem Beitrag legen nachvollziehbar offen, dass SOA einen Beitrag zur Flexibilisierung von GP leisten kann und wie dies geschieht. Aufgrund der immanenten Beschränkungen der Forschungsmethode Literaturreview ist eine genaue Quantifizierung der Wirkungszusammenhänge allerdings nicht möglich ist. Daher wird angeregt, die in Abbildung 3 dokumentierten Beziehungen in weiteren Studien einer kritischen Prüfung zu unterziehen, wobei folgende Prioritäten naheliegen:

1. Am schwierigsten dürfte die Unterstützung des Umweltmonitorings durch eine SOA sein. Hier ist zu fragen, wie das vorhandene Konzept des Service-Brokers besser zur Erfüllung dieser Eigenschaft weiterentwickelt werden kann. Denkbar wäre ein Modul innerhalb einer SOA, das diesen Beobachtungsprozess implementiert. Ebenfalls ist zu prüfen, ob zukünftig technische Entwicklungen zur Unterstützung des Umweltmonitorings genutzt werden können, wie bspw. das organic computing [SR05].
2. Des Weiteren stehen einige Eigenschaften von SOA in einem negativen Verhältnis

zur gewünschten Komplexitätsreduktion flexibler GP. Da diese Eigenschaften, wie lose Kopplung und Modularität, aber wesentliche Merkmale einer SOA sind, müssen entsprechende Handlungsempfehlungen und Konstruktionsprinzipien angegeben werden, die eine Balance zwischen Serviceentwurf und Komplexität berücksichtigen. Der Erfolg einer SOA hängt damit wesentlich von der Qualität ihrer Implementierung ab. So sind z. B. die „richtige" Granularität und Wiederverwendung von Services kritische Themen, ebenso wie die sinnvolle Ableitung von IT-Services aus Business Services.

Abschließend ist auf die Grenzen der vorliegenden Untersuchung hinzuweisen. Auch wenn eine umfangreiche Literaturbasis diesem Beitrag zugrunde liegt, kann nicht ausgeschlossen werden, dass die Kriterien flexibler GP und flexibler Architekturen unvollständig sind. Für zukünftige Arbeiten scheint es daher lohnend noch genauer zu untersuchen, welche Eigenschaften die Flexibilität von GP und Architekturen hinreichend definieren. Diese Eigenschaften sollten eindeutig und überschneidungsfrei sein, um besser auf (negative) Wechselwirkungen eingehen zu können. Weiterführend müssen die vorgebrachten Zusammenhänge und Wirkbeziehungen der Eigenschaften flexibler Architekturen auf die Eigenschaften flexibler Prozesse weiter abgesichert und (möglichst empirisch) validiert werden. Durch den hier gewählten zweistufigen Ansatz und die argumentativ-deduktive Betrachtung kann eine unzulässige Verkürzung der Wirkungszusammenhänge nicht ausgeschlossen werden. Daher wäre anzuraten, Kennzahlen für die Messung der Flexibilitätseigenschaften zu erarbeiten, um Wirkungsbeziehungen direkt nachzuweisen. Ebenso ließe sich untersuchen, inwiefern SOA besser als andere Architekturkonzepte (wie bspw. Client-Server-Architekturen oder Peer-to-Peer-Architekturen) geeignet ist, die Flexibilität von GP zu unterstützen.

Literaturverzeichnis

[AS04] Aier, S.; Schönherr, M.: EAI als Enabler flexibler Unternehmensarchitekturen. In (Hasselbring, W.; Reichert, M., Hg.): Enterprise Application Integration. GITO, Berlin, 2004.

[Al07] Alexopoulou, N.; Nikolaidou, M.; Chamodrakas, Y.; Martakos, D.: Enabling On-the-fly Business Process Composition through an Event-based Approach. In (Sprague, R.H., Hrsg.): Proceedings of the 41st HICSS. IEEE Computer Society Press, Los Alamitos, California, 2007.

[Al98] Allweyer, T.: Adaptive Geschäftsprozesse. Gabler, Wiesbaden, 1998.

[Al09] Allweyer, T.: Geschäftsprozessmanagement. 3. Auflage. W3L-Verlag, Herdecke, 2009.

[AGS05] Andresen, K.; Gronau, N.; Schmid, S.: Ableitung von IT-Strategien durch Bestimmung der notwendigen Wandlungsfähigkeit von Informationssystemarchitekturen. In (Ferstl, O.K.; Sinz, E.J.; Eckert, S.; Isselhorst, T., Hrsg.): Wirtschaftsinformatik 2005. Physica-Verlag, Heidelberg, 2005.

[Ba11] Bartmann D.; Bodendorf F.; Ferstl O.K.; Sinz E.J.: Merkmale, Systemarchitekturen und Management hochflexibler Geschäftsprozesse. In (Sinz E.J.; Bartmann D.; Bodendorf F.; Ferstl O.K., Hrsg) Dienstorientierte IT-Systeme für hochflexible Geschäftsprozesse. University of Bamberg Press, Bamberg, 2011; S. 1-14.

[Be11] Becker, A.: Nutzenpotenziale und Herausforderungen service-orientierter Architekturen. Gabler, Wiesbaden, 2011.

[BK08] Becker, J.; Kahn, D.: Der Prozess im Fokus. In (Becker, J.; Kugeler, M.; Rosemann, M., Hrsg.): Prozessmanagement. Springer, Berlin [u.a.], 2008.

[Co07] Coldewey, J.: Lohnt sich SOA? In (Starke, G.; Tilkov, S., Hrsg.): SOA-Expertenwissen. dpunkt.Verlag, Heidelberg, 2007.
[Er05] Erl, T.: Service-oriented architecture. concepts, technology, and design. Prentice-Hall PTR, Upper Saddle River, New Jersey [u.a.], 2005.
[FLW11] Ferstl, O.K.; Leunig, B.; Wagner, D.: Analyse und Handhabung unvollständig planbarer Geschäftsprozesse. In (Sinz, E.J.; Bartmann, D.; Bodendorf, F.; Ferstl, O.K., Hrsg.): Dienstorientierte IT-Systeme für hochflexible Geschäftsprozesse. Schriften aus der Fak. WI und AI der Universität Bamberg. Band 9. Univ. Bamberg Press, Bamberg, 2011.
[Fe06] Fettke, P.: State-of-the-Art des State-of-the-Art. Wirtschaftsinformatik Bd.48(4), 2006; S. 257–266.
[Gr00] Gronau, N.: Modellierung von Flexibilität in Architekturen industrieller Informationssysteme. In (Schmidt, H., Hrsg.): Modellierung betrieblicher Informationssysteme. MobIS 2000, 2000.
[Ha10] Hanschke, I.: Strategisches Management der IT-Landschaft. 2. Auflage. Hanser, München, 2010.
[HS11] Heinrich, L.J.; Stelzer, D.: Informationsmanagement. Grundlagen, Aufgaben, Methoden. 10. Auflage. Oldenbourg, München, 2011.
[KBS05] Krafzig, D.; Banke, K.; Slama, D.: Enterprise SOA : service-oriented architecture best practices. Prentice Hall, Upper Saddle River, New Jersey, 2005.
[Ku10] Kurz, M.: BPM 2.0. Kollaborative Gestaltung von Geschäftsprozessen. In (Schumann, M.; Kolbe, L.M.; Breitner, M.H.; Frerichs, A., Hrsg.): Multikonferenz Wirtschaftsinformatik 2010. Universitätsverlag, Göttingen, 2010.
[NRT12] Nissen, V.; Rennenkampff, A. von; Termer, F.: Agile IT-Anwendungslandschaften als Strategische Unternehmensressource. HMD Praxis der WI (284), 2012; S. 24–33.
[PM11] Paschke, J.; Molla, A.: Definition and Measurement of the Adaptive IT Capabilities Construct. Proceedings of the 17th AMCIS, 2011.
[Re07] Reinheimer, S.; Lang, F.; Purucker, J.; Brügmann, H.: 10 Antworten zu SOA. HMD Praxis der Wirtschaftsinformatik. Band 44 (253), 2007; S. 7–17.
[RHS05] Richter, J.; Haller, H.; Schrey, P.: Serviceorientierte Architektur. Informatik Spektrum. Band 28 (5), 2005; S. 413–416.
[Ro96] Rosemann, M.: Komplexitätsmanagement in Prozessmodellen. Methodenspezifische Gestaltungsempfehlungen für die Informationsmodellierung. Gabler, Wiesbaden, 1996.
[SW05] Scheer, A.; Werth, D.: Geschäftsprozessmanagement und Geschäftsregeln. In (Institut für WI im DFKI, Hrsg.): Veröffentlichungen des Instituts für WI, Saarbrücken, 2005.
[SS08] Schmelzer, H.J.; Sesselmann, W.: Geschäftsprozessmanagement in der Praxis. Kunden zufrieden stellen-Produktivität steigern-Wert erhöhen. 6. Aufl. Hanser, München, 2008.
[SR05] Strohmaier, M.; Rollett, H.: Future Research Challenges in Business Agility. In (Wombacher, A.; Huemer, C.; Stolze, M., Hrsg.): Seventh IEEE Int. Conf. on E-Commerce Technology. IEEE Computer Society, Los Alamitos, California, 2005.
[TS07] Tilkov, S.; Starke, G.: Einmaleins der serviceorientierten Architekturen. In (Starke, G.; Tilkov, S., Hrsg.): SOA-Expertenwissen. dpunkt.Verlag, Heidelberg, 2007.
[WF10] Wagner, D.; Ferstl, O.K.: Erhöhte Abbildungstreue von Geschäftsprozessmodellen durch Kontextsensitivität. In (Engels, G.; Karagiannis, D.; Mayr, H.C., Hrsg.): Modellierung 2010. GI, Bonn, 2010.
[Wa11] Wagner, D.; Suchan, C.; Leunig, B.; Frank, J.: Methode zur Analyse der Flexibilität von IS. In (Sinz, E.J.; Bartmann, D.; Bodendorf, F.; Ferstl, O.K., Hrsg.): Dienstorientierte IT-Systeme für hochflexible Geschäftsprozesse. Schriften aus der Fakultät WI und AI der Universität Bamberg. Band 9. Univ. Bamberg Press, Bamberg, 2011.
[We12] Weske, M.: Business process management. concepts, languages, architectures. 2. Auflage. Springer, Berlin [u.a.], 2012.
[WPR97] Wigand, R.T.; Picot, A.; Reichwald, R.: Information, organization, and management. Expanding markets and corporate boundaries. Wiley, Chichester, 1997.

Continuous Auditing in Big Data Computing Environments:

Towards an Integrated Audit Approach by Using CAATTs

Andreas Kiesow, Novica Zarvic, Oliver Thomas

Information Management and Information Systems
University of Osnabrück
Katharinenstr. 3
49069 Osnabrück, Germany
andreas.kiesow@uni-osnabrueck.de
novica.zarvic@uni-osnabrueck.de
oliver.thomas@uni-osnabrueck.de

Abstract: Through the risen operating of autonomous and automated Accounting Information System (AIS), enterprises are increasingly faced with the growing size of accounting-relevant data. Therefore, the audit's purpose, which is namely to examine the true and fair view of financial statements, is heavily increasing in complexity. To cover the requirements to AIS in Big Data Computing Environments, we strive for the development of a Continuous Auditing (CA) approach. Since, the implementation of CA is a recognized challenge among researchers and practitioners, and traditional audit tools and techniques neglect the potential of Big Data Analytics, we strive for the development of appropriate computer-assisted audit tools and techniques (CAATTs). Therefore, we analyzed established CAATTs considering the dimensions of the Big Data paradigm in this paper. To do so, we analyzed relevant literature and viewpoints of occupational organizations. This led us to a first proposal of a conceptual architecture for the integrated audit approach. We conclude this paper with a critical discussion and emphasize implications for further work.

1 Introduction

The paradigm of Big Data has tremendous impacts on both IT departments and audit. On the one hand, since financial statements are produced in automated Accounting Information Systems (AIS), the auditor is faced with risen complexity and risks due to an increasing processing of ever-growing data [GDW11], [VCK12]. With respect to this development, prevention and detection of fraud are gaining in importance [FSP13]. Furthermore, through the permanent processing of unstructured data, increased demands on data privacy and information security have to be addressed. On the other hand, chances occur in using data analytics techniques for the auditor's purposes, such as fraud detection or risk-assessment [ISACA14]. Hence, the audit needs the support of the IT

department to evaluate the completeness and the accuracy of the financial statements (e.g. data sets, system access) according to national and international laws and standards (e.g. §317 HGB [BMJV14] or ISA 200 [IFAC13]). In general, the financial statements result from business events, such as purchases, payments, inventory changes, or shipments [SB13]. However, correct financial statements require the proper development and implementation of AIS as well as the integrity of programs, data files and computer operations. Therefore, all procedures, processes, and controls which refer to the Information Technology environment (IT General Controls) have to be evaluated by the auditor [SB13]. In brief, auditors need the support of the IT Management to use chances and address risks in Big Data Computing Environments (BDCE).

The development of an appropriate audit approach demands the consideration of the complexity and the prevalence of the Big Data paradigm from a holistic and theoretical perspective. Remarkable findings in the field were reached by MOFFITT and VASARHELYI, who stated that "*auditors can also take advantage of Big Data*" by implementing continuous monitoring and auditing tools and techniques. In addition, they stated that due to the digitalization, storage, retrieval and analytics of financial data could improve the complete audit process [MV13]. SUN stated that the development of innovative computer-assisted audit tools and techniques (CAATTs) contributes to the handling of increasing data complexity. Moreover, the quality of audit could benefit from the usage of intelligent data analysis implemented with CAATTs [Su12]. Hence, a reasonable support of the auditor, Continuous Auditing (CA) has to be realized with appropriate Information Systems (IS), which represents the scope of this paper.

To set a baseline for further investigations, we analyzed established types of CAATTs in terms to their support of the IT auditor to conduct a proper examination of financial statements in a BDCE. This analysis required a sound exploration of specific requirements of audit in BDCE and the CA approach. Therefore, our paper is structured as follows: At first, the analysis of the related literature and the theoretical background in the areas of audit and Big Data, CA, and CAATTs is presented. Then, the applied research methodology is described. In the following section, the types of CAATTs are analyzed considering the requirements of Big Data and CA. Additionally we present a first conceptual architecture of the envisioned audit approach. This leads into a summary, where a concept of an integrated Audit Module is introduced, and limitations are discussed. Finally, we conclude with the results for researchers and practitioners and propose an outlook to further work in this research area.

2 Related Work and Theoretical Background

2.1 Audit in a Big Data Computing Environment

As shown in the prior section, the prevalence of the Big Data paradigm has manifold impacts on the accounting-relevant processes. Therefore, internal audit departments as well as external audit companies have to face the upcoming consequences of these impacts. Particularly, the Information Systems Audit and Control Association (ISACA)

provides possible considerations in different white papers. The results of these papers are shown in this section.

Obviously, the occurrence of data in a BDCE is manifold. According to DOUG LANEY, data are defined as Big Data, if it is characterized through the dimensions *Volume*, *Velocity* and *Variety* [La01]. In literature, *Veracity* has become apparent as a 4th dimension [IBM12]. Before selecting appropriate tools and techniques, these four dimensions are analyzed from the perspective of accounting and audit. Additionally, we consider the usage of *Big Data Analytics* in terms of audit.

Computer-assisted accounting processes are defined by automatic and frequent generation and processing of data, which result in a higher and ever-growing size of data [VCK12] [MV13]. The *Volume* and *Velocity* of accounting-relevant data disable the manual evaluation (e.g. sample selection) completely [RES01]. ISACA suggests the use of CAATTs since it "*serves as an important tool for the IS auditor to evaluate the control environment in an efficient and effective manner*" [ISACA08].

Furthermore, manifold interfaces and data formats lead to increasing complexity and complicate the ex post traceability of data source (*Variety*). The source of accounting-relevant data has to be identified and defined. Generally, accounting-relevant data are processed within the Enterprise Resource Planning Systems [Bl12]. This *structured accounting-relevant data* are usually generated as "*a result of common business events, such as purchases, payments, inventory changes or shipments for instance*" [SB13]. Additionally, structured accounting-relevant data are also generated by recording events related to security, availability modifications and approvals of IT systems. *Unstructured accounting-relevant data* includes text, such as documents for organizational regulation, e-mails, test results, as well as data from mobile devices, audio and video files, server logs, web data for instance.

This dimension *Veracity* refers to the uncertainty of data that incorporates their confidence and quality i.e. accuracy of data [Fa13]. Confidence and quality of data are strongly related to their privacy and security. Hence, to prevent data breaches and establish data security along the accounting-relevant processes, robust data-privacy solutions are needed. Especially, the proactive monitoring of sensitive data and the processing IT environment as well as the fast reaction to data or privacy breaches are prevalent requirements [ISACA13b].

Moreover, ISACA defines *Big Data Analytics* (BDA) as "*the application of emerging statistical, processing and analytics techniques to big data for the purpose of advancing the business*" [ISACA14]. Additionally, the ISACA states that BDA, i.e. the analysis of data to find patterns and correlations, can be used to enforce business risk management and security operations [ISACA13a]. The chance to use BDA for the purposes of audit is increasingly recognized by researchers and auditors. SETTY and BAKHSHI propose the concept of a Big Data Refinery, which enables the analysis of logs generated by ITGC or the fraud analysis of large volumes of transactional data [SB13]. MOFFIT and VASARHELYI list several examples of using data analysis for audit purposes [MV13]: First, using Big Data for the continuity equations to link processes, to model lagged

information processes, and to perform automatic error correction according to KOGAN et al. 2011. Then, the cluster analysis for anomaly detection in accounting data according to THIPRUNGSRI and VASARHELYI (2011) and, finally, applying process mining on the audit process according to JANS et al. (2010). POULOVASSILIS sees the prevalence of large volumes of graph-structured data in the areas finance chains, fraud detection, risk analysis [Po13]. ABBASI et al. developed a design science approach to detect fraud in financial statements, which is based on a meta-learning Business Intelligence and Analytics (BI&A) framework [Ab12].

2.2 Reviewing Continuous Auditing

The paradigm of CA is well established in literature (e.g. [RES01], [Va12]). A substantial study in the field of CA was published by SUN in 2012. SUN stated, since CA is completely implemented as an automated, (i.e. computer-assisted) process, it contributes to efficiency whilst the audit of large volume of transaction data [Su12]. However, a CA approach is not limited to *handle* Big Data. Furthermore, CA can be used to *produce* Big Data which can be analyzed to improve audit efficiency and efficacy [MV13].

In 2012, the Americas Institute of Certified Public Accountants (AICPA) published a survey about the current state of CA from the perspective of external auditors [By12a]. A sustainable result of this survey is that CA approaches are rarely implemented in practice. AICPA carved out the following reasons for this deadlocked situation: First, practitioners stated that due to high complexity of the implementation of CA is recognized as a challenge and related to high cost. Furthermore, for external auditors, audit relationships are characterized as instable, which complicates the estimation of payback periods. Additionally, many businesses are described as protective of their data, which disables the ongoing access to systems, which is necessary for the implementation of CA. At last, external auditors see limitations in the current state of audit standards, which do not cover the requirements to CA or other future audit approaches. However, the survey carved out that practitioners are aware of the benefits of CA, which are namely reduced costs, time, and increased quality, if a CA solution is in place. Finally, the AICPA concludes that the full potential of the CA approach can be envisioned if the implementation of this approach can be effectively demonstrated, documented, and disseminated. In Germany, the current state of CA is mainly discussed among practitioners. In 2014, EULERICH and KALINICHENKO published an article about CA from a research perspective [EK14]. In 2011, HÖLZER and ARENDT stated, that the implementation of CA is the „*optimum of the auditors activities*“ [HA11]. However, according to AICPA, for the time being, we do not know any CA solutions in place.

2.3 Prevalence of CAATTs

Since accounting transactions are processed automatically by computer-assisted Information Systems, the need for tools and techniques, which support the audit of these transactions, has arisen [IFAC13b]. The definitions and application areas of computer-assisted audit tools and techniques (CAATTs) have changed through the decades,

depending on the state of technology. Regarding the aim of this paper, we define CAATTs according to HALL, who expresses CAATTs as the "*tools and techniques used to examine directly the internal logic of an application as well as the tools and techniques used to draw indirectly inferences upon an application's logic by examining the data processed by the application*" [Ha10]. In literature five different types of CAATTs can be distinguished, which are namely Test Data, Integrated Test Facility (ITF), Parallel Simulation, Embedded Audit Modules/ System Control and Audit Review Files (EAM/SCARF) and Generalized Audit Software (GAS) [BD03]. Additionally, the technique of tagging and tracing (snapshots) can be added to these types [RJG09]. These six types, their individual descriptions, and examples are shown in Table 1.

Types of CAATTs	**Description**
Test Data	Fictitious, auditor-prepared data, which will be processed by the audited systems. The evaluation bases on a comparison between the results of the test data and the auditor's expectations. The processing within the audited systems is a "black box".
Integrated Test Facility	Processing of Test Data in separated areas or modules within the audited system. The results of the internal system controls are visible for the auditor.
Parallel Simulation	Auditor-developed application, which is completely separated from the client's systems. The results of processing real data are compared with the results of the client's systems.
Embedded Audit Module, System Control and Audit Review Files (EAM/ SCARF)	Auditor-developed module which is implemented within a client's system. EAM evaluates real data by predefined criteria while it is processed. Results of EAM evaluations can be written into a SCARF, which is send to the auditors for further examination
Generalized Audit Software	Auditor-developed and self-contained applications, which evaluate extracted real data and analyze them, regarding predefined criteria.
Snapshot Method (tagging and tracing)	Selection and marking of accounting transactions and monitoring their processing within the AIS. After every step, a snapshot is created and analyzed.

Table 1: Typology of CAATTs [BD03] and [RJG09]

3 Research Methodology

The goal of our work is the development of an IT artifact, namely a computer-assisted audit solution for BDCE. The matter of this paper is the selection and analysis of appropriate CAATTs, which support the prevalence of CA of accounting-relevant processes. Generally, CAATTs are used to examine accounting-relevant information to support the audit to reach valid and sufficient evidences for the true and fair presentation of financial statements. According to MARCH and SMITH, the development of appropriate CAATTs and their usage to realize a CA approach support human (i.e. auditors, IT Management and, finally, stakeholders) purposes [MS95]. Furthermore, we strive for an audit approach in a BDCE which is an unsolved problem so far. Additionally, with the implementation of computer-assisted audit tools and techniques,

internal and external auditors should be empowered to gain assurance in effective and more efficient ways [He04]. For these reasons, we apply a Design Science approach, which is characterized by both building and evaluating IT artifacts.

ÖSTERLE et al. [Ös11] state that "*design-oriented IS research follows an iterative process*" and consists usually of four distinct phases. These phases are *analysis*, *design*, *evaluation* and *diffusion*. The work presented in this paper can mainly be situated in the first two phases, namely analysis and design (see Figure 1). Next to the identification of the practical problem, we represent an iteration considering the analysis of requirements from practitioner's perspective. To do so, we supplement our results from academic literature as well as the viewpoint of occupational organizations, such as ISACA and AICPA. This sets the baseline for the needs of audit in a BDCE. Our investigation of CAATTs is presented and discussed in terms of the needs that come up in the context of BDCE. Building on these results, we construct a first proposal of a conceivable IT architecture for an integrated audit approach. The evaluation of this instantiation according to HEVNER et al. [He04] will be performed in future work.

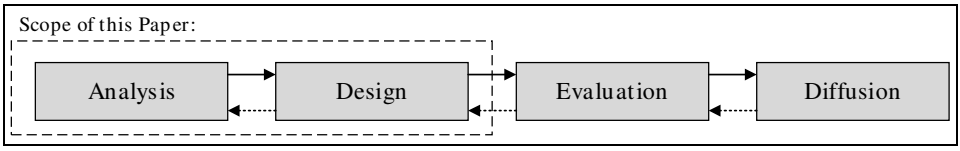


Figure 1: Design-oriented IS research according to Österle [Ös11]

4 A General Concept of an Integrated Audit Tool

4.1 Investigating Existing Audit Tools and Techniques

As mentioned above, CA has to be based upon computer-assisted tools and techniques. Therefore, the question arises which CAATTs contribute to an appropriate implementation of CA in a BDCE. With respect to this goal, the types of CAATTs have to be assessed considering the specifications of a Big Data. It has to be analyzed to what extent the types of CAATTs are able to support the audit of accounting-relevant processes within a BDCE. The results of the assessment are discussed in the following sub-sections.

As mentioned above, *Test Data* are data, which is created by the auditor and processed within the client's system. Considering the dimension *Volume*, Test Data can be used to test the load capacity of programs, i.e. to what extent is the program able to process large data sets. The dimension *Veracity* could be addressed through the examination of the completeness and accuracy of programs. Privacy aspects could be addressed by simulating personal data and examine their processing considering data leaks. *Velocity* is not covered by using Test Data, since the continuous input of data in the AIS cannot be tested by a single Test Data set. Testing *Variety* is generally possible, but not efficient, since the creation of Test Data with manifold data formats is complex and need a sound

exploration of the interfaces in use. Combining Test Data with BDA is at most useful to test the basic functions. Since the Test Data are created by the auditor, analyzing it with BDA does not result in additional value. In short, Test Data can support a CA approach through punctual analysis of large, various or personal related data sets. However, the creation of Test Data is related to both high costs and effort.

The *ITF* approach is strongly related to Test Data. Therefore, the same conclusions, which are mentioned for Test Data are also valid for ITF. However, since ITF demands the implementation of separated entities within the audited systems, and the simultaneous processing with live input, the realization of ITF covering Volume, Velocity, Variety and Veracity are from increasing complexity. A Continuous Approach realized by ITF requires the implementation thorough all accounting-relevant systems, which is unrealizable in practice. Furthermore, the results have to be gathered and analyzed, which is not proposed by ITF. Therefore, we conclude that ITF are inapplicable to implement CA in a BDCE.

The *PS* demands the complete rebuild of the audited AIS, considering all data processing steps and the integrated controls. Although this approach is certainly effective, it is quite inefficient and unrealizable in practice. Furthermore, it does not contribute to BDA, since Parallel Simulation replicates the results, which excludes further investigations. Therefore, we believe that Parallel Simulation is inapplicable to implement CA in a BDCE.

EAM are implemented as separated add-ons within accounting-relevant systems. Therefore, EAM enable the permanent evaluation of the data flow. The audit of large data sets (*Volume*) and the dimension of *Velocity* are covered by the continuous evaluation of data either before or after its generation or processing. To cover the *Variety* of accounting-relevant data, the implementation of relevant interfaces is required. *Veracity*, which is related to the completeness and accuracy of data, can also be addressed by the continuous evaluation, if the necessary checks are implemented within the EAM. The consideration of privacy is difficult, since the violation of privacy can be normally checked a-posteriori. The implementation of BDA functions within EAM is basically possible, but not efficient, since the analysis of Big Data sets (e.g. Data Mining) requires high processing capacity, which is normally not integrated in EAM. In short, a CA approach using EAM addresses *Volume* and *Velocity* of the BDCE. Furthermore, it can contribute to *Variety*, *Veracity* and BDA.

Generally, *GAS* is used to analyze extracted data sets *after* their processing in the system (e.g. annual Journal Entry Testing). Furthermore, the development of GAS-tools requires extensive test activities. Therefore, we state that GAS is inapplicable for the permanent audit of high volume of data, fast processing of data and various data formats. However, GAS offers several functions for fraud detection [La04] and, if it is combined with Data Mining techniques, it can be used for punctual analysis of large data sets. Therefore, we believe that GAS-Tools can contribute to audit *Veracity* and BDA.

The *tagging* of accounting transactions and the *tracing* throughout their processing within the BDCE is basically possible. However, it has to be considered that an adequate

large sample of transactions must be tagged, to address the dimension *Volume*. Furthermore, defining of an adequate sample is difficult, since data are generated quickly and in various formats (*Velocity*, *Variety*). The possibility to test *Veracity* is given, if completeness and accuracy of the tagged transactions are traceably documented after every step. The Snapshot Method does not contribute to BDA, since it does not enable further investigations of the source and structure of the transaction. Overall, we believe that the Snapshot Method is not applicable to realize an Audit Approach in a BDCE. However, it could be used to support the sought solution. The results of these considerations, i.e. the applicability of CAATTs to cover the requirements of a BDCE, are shown in the upcoming Table 2.

Types of CAATTs	**Dimensions of Big Data**					**Overall Applicability**
	Volume	**Velocity**	**Variety**	**Veracity**	**BDA**	
Test Data	◉	○	◉	◉	○	◉
ITF	○	○	○	◉	○	○
PS	○	○	○	○	○	○
EAM, SCARF	●	●	◉	◉	◉	◉
GAS	○	○	○	◉	●	◉
Snapshot Method	○	○	○	◉	○	○
Legend: ○ = low, ◉ = medium, ● = high						

Table 2: Applicability of CAATTs for the Dimensions of Big Data

As shown in Table 2, Embedded Audit Modules in relation with SCARF are most appropriate to cover the dimensions of Big Data. Test Data, GAS and Snapshots propose different supporting functions. Therefore, we propose a combination of different tools and techniques. The questions arises, *how* these techniques can be combined to an audit approach in a BDCE.

4.2 Conceptual IT Architecture for the Integrated Audit Approach

Through our investigations, we believe that a CA approach is the best way to address the complexity and requirements of a BDCE. This approach has to be realized with the computer-based techniques of EAM, which enable the audit of controls, Test Data to audit the processing within the system, and GAS to use BDA. To combine the results of these functions to a tangible and traceable solution, we revert to the pilot system, which was proposed by ALLES et al. in 2006 [Al06], and VASARHELYI, who achieved remarkable results by designing a CA approach for online systems in 1991 [VH91]. Basically, our approach consists of an audit database, which includes three major data tables, an Audit Cockpit as user front-end, and a computational core for analytics (Figure 2). In the data table "*Data Inventory*" all accounting-relevant data are included (Data Table 1). Furthermore, the origin of data (internal or external), sensitivity, and relevance

data should be labeled. Additionally, for every audited interface data producer and data consumer have to be defined. The second data table "*Controls*" includes all controls along the accounting-relevant processes, added with specific information (e.g. related risks, detective or preventive, frequency). The EAM enable the direct and permanent monitoring of the controls, and generate automated alerts in the case of control violation. The results of these monitoring activities will automatically be transferred in the third data table "*Results*". In the Audit Cockpit, the link between "*Data Inventory*", "*Controls*", and "*Results*" is visualized. Therewith, the auditor is able to understand, which data are controlled with which result. Overall, the audit data base and the computational core should be protected according to security and privacy standards. Finally, the CA infrastructure (i.e. EAM, Audit Cockpit, data base, hardware) has to be covered by disaster recovery arrangements. Additional functions should propose the visualization of the data flows and a list of current control violations.

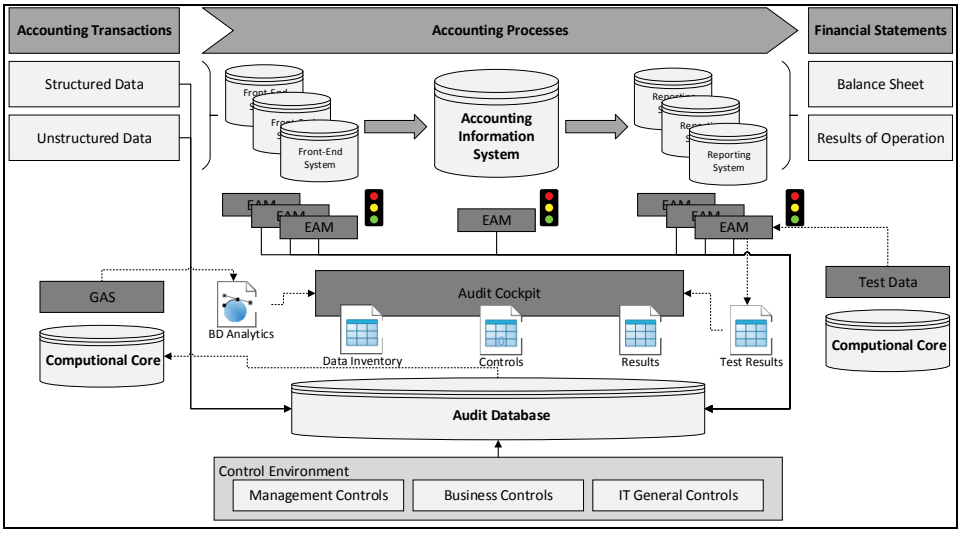


Figure 2: Integrated Audit Approach

5 Summary and Discussion

The discussion of our approach is based on an assessment according to the typology of audit methods, which is presented by MARTEN et al. [MQR07]. The results are subsequently summarized: The implementation of the approach can be risk-oriented in terms of the definition of high-risk areas, which are characterized by material impact on the financial statements. However, in the final stage, this approach covers all accounting-relevant areas, independent of the inherent risk. Furthermore, if the accounting-relevant areas are identified, the approach enables a direct examination of the subjects. The approach is designed system-oriented as well as goal-oriented, which means that internal controls can be examined as well as results from procedures. Incontestable, the approach enables the examination on transaction level, which empowers the auditor to investigate

further audit activities in a fast and precise manner. If implemented, the approach covers the population of all transactions, which leads into rigorous conclusions of the audited subject and reduces the probability of audit errors. Furthermore, if implemented, the approach set the baseline for progressive audits from the accounting transaction to the financial statement and, vice versa, for retrograde audits. Additionally, the approach empowers the auditor to conduct formal, and material audit checks, which means to assure *whether* and *to what extent* a transaction is accurate. However, we are aware that the implementation of this approach is a big challenge. The technical realization of this approach requires the implementation of manifold interfaces throughout the organization. Hence, this refers to the reorganization of the processes, which is related to tremendous costs and effort [Su12]. Controls and data flows have to be carved out and analyzed. Thus, the implementation is definitely a long-term solution and has to be considered through the strategic planning of the management. Furthermore, the approach is not able to cover manual controls, such as management approval by signature. The approach requires the implementation of automated controls within the systems and interfaces, possibly supported by a workflow management system. Finally, due to reduced system performance, code modifications, and the creation of excessive data, AICPA stated in 2012 that an approach using EAM "*exists as primarily an academic topic*" [By12b]. In summary, the usage of new technology in audit is a voluminous task for both IT management and audit. However, we believe that our findings imply new considerations in the field of computer-assisted audit which will benefit to researchers and practitioners. Since the proposed Audit Cockpit approach enables the permanent monitoring of the control environment, we set a further baseline for the alignment of IT and audit.

6 Conclusion

The design of appropriate audit solutions for BDCE is of increasing importance. On the one hand, opportunities could be exploited due to the usage of Big Data Analysis on transaction level, which could benefit to increased performance and quality of audit. On the other hand, the complexity in analyzing and auditing accounting-relevant data are recognized as a challenge by experts and researchers. Furthermore, the compliance with privacy and security guidelines in BDCE has to be audited. According to the relevant literature and occupational requirements, we suggest a CA approach, which has to be basically realized by computer-assisted audit tools and techniques (CAATTs). To do so, we analyzed existing CAATTs regarding the requirements of the Big Data paradigm (4 V's). Finally, we proposed the combination of different techniques, which are managed and monitored within an Audit Cockpit. Limitations were carved out in the technical reorganization of the accounting-relevant processes and the demand of automated controls along the processes. In future work, we strive for the cooperation with an organization, to analyze accounting-relevant processes and to develop a prototype of an Audit Cockpit. In this context, the requirements to Internal Control Systems as well as the application of common IT Frameworks such as COSO and COBIT have to be analyzed. Furthermore, we strive for a sound exploration of implied risks for auditors in BDCE.

7 References

[Ab12] Abbasi A.; Albrecht C.; Vance A. et al. (2012): MetaFraud: A Meta-Learning Framework for Detecting Financial Fraud. *MIS Quarterly*, 36 (4). pp. 1293-1327.

[Al06] Alles M.; Gerard Brennan; Kogan A. et al. (2006): Continuous monitoring of business process controls: A pilot mplementation of a continuous auditing system at Siemens. *International Journal of Accounting Information Systems*, 7 (2). pp. 137-161.

[BD03] Braun R. und Davis H. (2003): Computer-assisted audit tools and techniques: Analysis and perspectives. *Managerial Auditing Journal*, 18 (9). pp. 725-731.

[Bl12] Bloem J.; van Doorn M.; Duivestein pp. et al. (2012): *Creating clarity with Big Data*. (research report) Groningen, the Netherlands: The Sogeti Trend Lab vint.

[BMJV14] HGB §.3. (2014): *Handelsgesetzbuch*. Bundesministerium für Justiz und Verbraucherschutz. http://www.gesetze-im-internet.de/hgb/__316.html, zuletzt geprüft am 03.30.2014.

[By12a] Byrnes P.E.; Ames B.; Vasarhelyi M. et al. (2012): The Current State of Continuous Auditing and Continuous Monitoring. *AICPA White Paper*, 1-16.

[By12b] Byrnes P.E.; Al-Awadhi A.; Gullvist B. et al. (2012): Evolution of Auditing: From the Traditional Approach to the Future Audit. *AICPA White Paper*. pp. 1-9.

[EK14] Eulerich M. und Kalinichenko A. (2014): Die Continuous Auditing - Diskussion aus wissenschaftlicher Sicht. *Zeitschrift Interne Revision*, 49, 1, 34-45.

[Fa13] Fan W. (2013): Querying Big Social Data. In: Gottlob G.O.S. (Hg.): *29th British National Conference on Databases*. Oxford, UK: Springer. pp. 16.

[FSP13] Fochler K.; Schmidt A. und Paffrath R. (2013): IT-Revision 3.0 – Herausforderungen für die interne IT-Revision. *HMD – Praxis der Wirtschaftsinformatik* (289). pp. 20-30.

[GDW11] Gelinas U.; Dull R. und Wheeler P. (2011): *Accounting Information Systems*. Mason, OH: Cengage Learning.

[Ha10] Hall J. (2010): *Information Technology Auditing*. 3rd. Aufl. Mason, OH: South-Western Cengage Learning.

[HA11] Hölzer D. und Arendt pp. (2011): Massendatenanalysen. *Zeitschrift für Interne Revision*, 6, 2-6.

[He04] Hevner A.; March pp.; Park J. et al. (2004): Design science in information systems research. *MIS Quarterly*, 28 (1). pp. 75-105.

[IBM12] IBM Research (2012): *Global Technology Outlook 2012*. Warwick: IBM Research.

[IFAC13a] International Federation of Accountants (2013): ISA 200. In IFAC (Hg.), *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*. New York. pp. 74.

[IFAC13b] International Federation of Accountants (2013): ISA 315.19. In IFAC (Hg.), *Handbook of International Quality Control, Auditing, Review, Other Assurance, and Related Services Pronouncements*. New York. pp. 300.

[ISACA08] Information Systems Audit and Control Association (2008): *IS Auditing Guideline: G3 Use of Computer-Assisted Audit Techniques*. http://www.isaca.org/Knowledge-Center/Standards/Pages/IS-Auditing-Guideline-G3-Use-of-Computer-Assisted-Audit-Techniques1.aspx, zuletzt geprüft am 31.03.2014.

[ISACA13a] Information Systems Audit and Control Association (ISACA): Big Data Impacts & Benefits. *ISACA White Paper* (March 2013). pp. 1-14.

[ISACA13b] Information Systems Audit and Control Association (ISACA): Privacy Big Data. *ISACA White Paper* (August 2013). pp. 1-12.

[ISACA14] Information Systems Audit and Control Association (ISACA): Generating Value From Big Data Analiytics. *ISACA White Paper* (January 2014). pp. 1-12.

[La01] Laney D. (2001): *3D data management: Controlling data volume, velocity and variety.* META Group Research Note, 6.

[La04] Lanza R. (2004): Fraud data interrogation tools: COMPARING BEST SOFTWARE FOR FRAUD EXAMINATIONS. *FRAUD Magazine*, 18 (6). pp. 1-7.

[MQR07] Marten K.-U.; Quick R. und Ruhnke K. (2007): *Wirtschaftsprüfung*. 3rd. Aufl. Stuttgart, Germany: Schäffer-Poeschel Verlag.

[MS95] March pp.T. und Smith G.F. (1995): Design and natural science research on information technology. *Decision support systems*, 15 (4). pp. 251-266.

[MV13] Moffitt K. und Vasarhelyi M. (2013): AIS in an Age of Big Data. *Journal of Information Systems*, 27 (2). pp. 1-19.

[Ös11] Österle H.; Becker J.; Frank U. et al. (2011): Memorandum on design-oriented information systems research. *European Journal of Information Systems*, 20 (1). pp. 7-10.

[Po13] Poulovassilis A. (2013): Database Research Challenges and Opportunities of Big Graph Data. In: Gottlob G.O.S. (Hg.): *29th British National Conference on Databases*. Oxford, UK: Springer. pp. 29.

[RES01] Rezaee Z.; Elam R. und Sharbatoghlie A. (2001): Continuous auditing: the audit of the future. *Managerial Auditing Journal*, 16 (3). pp. 150-158.

[RJG09] Rittenberg L.; Johnstone K. und Gramling A. (2009): *Auditing: A Business Risk Approach*. 7th. Aufl. Mason, OH: Cengage Learning.

[SB13] Setty K. und Bakhshi R. (2013): What Is Big Data and What Does It Have to Do With IT Audit? *ISACA Journal*, Volume 3, 1-3.

[Su12] Sun C.-M. (2012): From CAATTs Adoption to Continuous Auditing Systems Implementation: An Analysis Based on Organizantional Routines Theories. *MIS REview*, 17 (2). pp. 59-85.

[Va12] Vasarhelyi M.A.; Alles M.; Kuenkaikaewa et al. (2012): The acceptance and adoption of continuous auditing by internal auditors: A micro analysis. *International Journal of Accounting Information Systems*, 132 (3). pp. 267-281.

[VCK12] Vasarhelyi M.; Chan D. und Krahel J. (2012): Consequences of XBRL standardization on financial statement data. *Journal of Information Systems*, 26, 1, 155-167.

[VH91] Vasarhelyi M.A. und Halper F.B. (1991): The continuous audit of online systems. *Auditing: A Journal of Practice & Theory*, 10 (1). pp. 110-125.

A Comprehensive Model for Revealing Anomaly in Network Data Flow

Maher Salem, Ulrich Buehler

Department of Applied Computer Science
University of Applied Sciences Fulda
Marquardstr 35
36043 Fulda
maher.salem@informatik.hs-fulda.de
u.buehler@informatik-hs-fulda.de

Abstract: Large computer and communication networks lead to the generation of massive data flows. The difficulty of analyzing and managing these data in network security degrades the online detection of intrusion and suspicious connections. To overcome this problem, we present a comprehensive model that handles the traffic of computer networks and uncovers intrusions in real time. The model consists of dataset generator and intrusion detector. The dataset generator captures, analyzes and manages the live traffic using a dynamic queuing concept. It continuously constructs connection vectors from the live traffic and exports them either as datasets or sequentially into a pipe for further processing. The intrusion detector is based on an enhanced growing hierarchical self organizing map which classifies exported vectors to normal, anomaly or unknown connections. The model has been evaluated using synthetic and realistic data sources. It is able to process data flows within significant time and classifies the connections in the online mode effectively.

1 Introduction

The online operation mode of Intrusion Detection System (IDS) has become a real challenge since the amount of generated heterogeneous and non-stationary data and the interconnection between communication networks are increasing rapidly [HB11],[LYW13],[GJP12]. Recent IDS models have been improved to classify network traffic in online operation mode, but the improvement could achieve limited targets such as detecting anomalous traffic in http request [LLZ10] or detecting only DoS attacks [XC07]. However, these models are not able to detect anomaly in networks which generate massive data flows.

Moreover, the detection models suffer from the inability to detect new attacks or attacks with updated signature in real time, because they adhere to the original offline training dataset and they are not superior in recognizing unknown data flow. Therefore, most recent IDSs still struggle to adapt with the current heterogeneous networks. They should

professionally be improved so that they can handle massive data flows and provide preciseness in detecting anomalous traffic. More specifically, they should be competent of converting massive data flows to connection vectors on 1 Gbps to 10 Gbps networks. These connection vectors are used for training and evaluating IDS models. Hence, IDS models need a state-of-the-art method to handle the previous challenges. Figure 1 depicts a general diagram of the major steps of an IDS model.

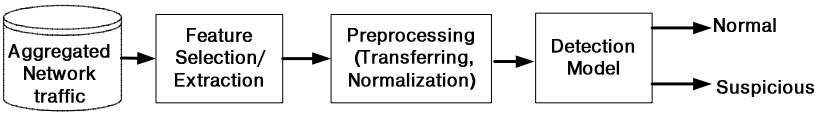


Figure 1: Major steps of an IDS model

Artificial Neural Networks are considered one of the superior methods in ensuring significance of network management. They have been using in IDS and then become the most efficient method in network security [ZCL11]. Growing Hierarchical Self Organizing Map (GHSOM) is considered as the most successful detection method based on neural network. It is a modified approach from the Kohonen Self Organizing Map (SOM) that can discover knowledge and abstract relations of high-dimensional data and map these data into a two-dimensional representation space [TK95]. GHSOM approach has improved the original SOM by structuring several SOMs in a hierarchical growing form [RMD02].

In this paper, we present a novel method that consists of two major parts: datasets construction using dynamic queuing concept and reinforcing the security over networks using an intelligent intrusion detection model. The first part continuously captures network packets and hosts' events in real time, processes and analyzes them in a queuing concept of dynamic window size, and then constructs sequential connection vectors based on selected features. These features are 22 valuable features, which are accurately selected based on our proposed feature selection method in [SBR11]. The second part is complementary to the first one, which is an enhanced GHSOM detection model (EGHSOM) that is able to uncover known and unknown anomaly from the constructed connection vectors [SB13]. The result of performance evaluation demonstrates the feasibility of the first part of the framework and shows a high performance of the second part.

The rest of this paper is organized as follows. Section 2 discusses previous approaches in the scope of our research. Our proposed method is described in section 3. While, section 4 presents the performance evaluation. Finally, section 5 concludes our work.

2 Related Work

Data flow in computer and communication networks is increasing rapidly which lead to the generation of voluminous amount of data. This issue makes network monitoring and security more complicated and demands more robust frameworks and methods to preserve the management. A framework that addresses this issue, by presenting a graph

differential anomaly visualization tool, is proposed in [LS12]. This kind of solution requires a persistent monitoring from the human side and reasonable reaction from the machine side as well. Therefore, it is very difficult to reach a satisfactory trade-off between these two sides.

Generally, IDS approaches often depend on available offline datasets for training and evaluating their models, and ignore the way of aggregating them. However, an acceptable framework (SSENet-2011) that synthetically generates datasets is addressed in [VHS11]. The framework has been installed in a similar simulation environment to that in DARPA intrusion detection evaluation program in 1999 [LYW13]. On the other hand, synthetic datasets cannot represent online traffic; hence, they are not appropriate for online IDS evaluation. In contrast, C. Hsu and S. Wang propose a packet capture module, called Flow Ring, to avoid packet loss and to enhance intrusion detection performance [HW11]. The proposed work consists of flow controller and buffer ring to ease packet capturing and to decrease losses as well. The experimental study demonstrates that the Flow Ring is a promising positive result especially when integrating to Snort IDS [MR99], [SN13] in computer networks.

In this regard, A. Biswas and P. Sinha outperform the Flow Ring method by proposing an efficient packet capturing method to improve the IDS performance [BS06]. The architecture performance is compared with Linux, NAPI, and PFRING against a comprehensive set of criteria. The result indicates that the proposed architecture DMA ring can be used with tcpdump or snort to improve their performance. The architecture can be helpful to dump packets without losses in high-speed networks. A similar and more scalable architecture has also been presented from C. Morariu and B. Stiller [MS08]. Most of the proposed works investigate only packet capturing and overlook other data sources such as hosts events. In this paper, dataset generation will be conducted for network traffic and hosts events in real time.

3 Proposed Method

The proposed method interconnects the above major steps of an IDS model and consists of two parts: The *Dataset Generator* and *Intrusion Detector* as illustrated in figure 2.

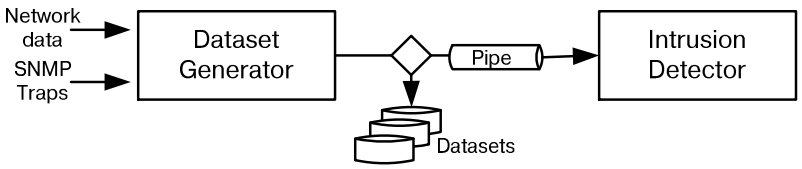


Figure 2: General overview of the comprehensive model

The constructed connection vectors from the *Dataset Generator* represent the link between these two parts. The output of the *Dataset Generator* is used as input for the *Intrusion Detector*. Finally, the classification of the constructed vectors delivers

information about the security state of the observed computer and communication network (i.e. normal, anomaly or unknown).

3.1. Dataset Generator

The model addresses network data flows and hosts' activities effectively by using the *Dataset Generator* that analyzes, processes, and correlates them based on a flow diagram as shown in figure 3.

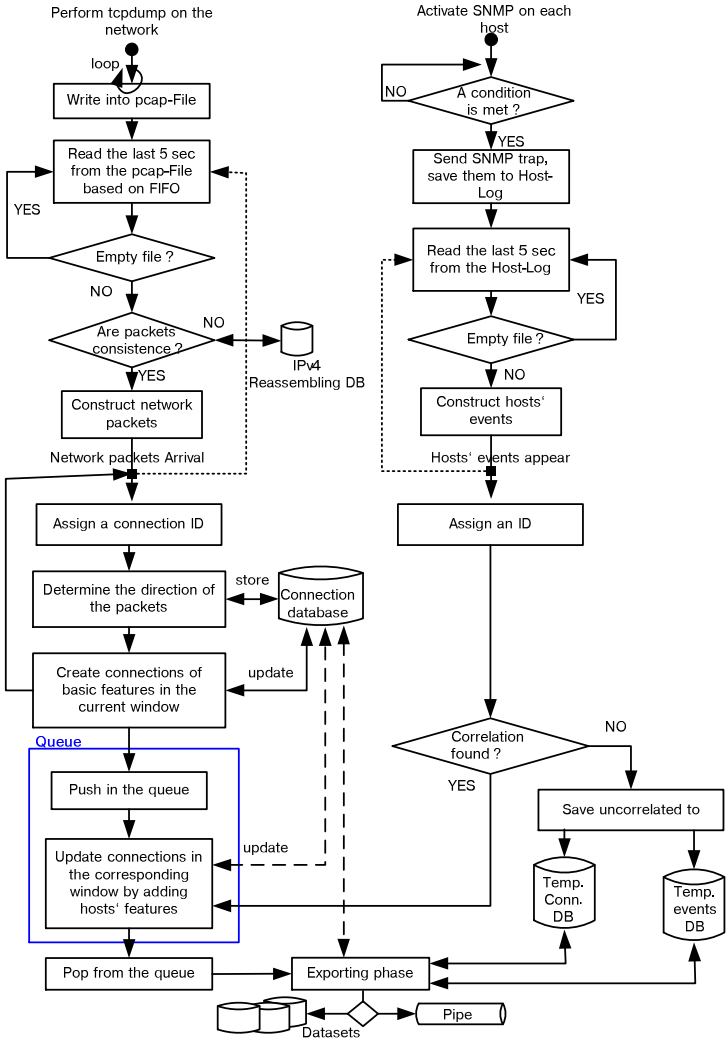


Figure 3: Flow diagram of the dataset generator

We use benefits of the well-known tool TCPDUMP to aggregate the online massive traffic by mirroring it on a specific port. Moreover, we exploit the SNMP mechanism to send traps from network hosts when a certain activity is matched, e.g. by a failure login of a user. To ease the illustration, we assume that the aggregator handles the dump data and traps into an appropriate time slot, e.g. every 5 seconds as figure 3 shows.

The generator analyzes and processes aggregated dump traffic and constructs corresponding network packets. Similarly, it constructs hosts' events from the received SNMP traps at the server side. It correlates the packets and traps upon ID match to construct connection vectors, based on certain features such as *protocol_type, service, count* [SBR11], inside the queue and export them as datasets.

Specifically, the generator performs the following main steps:

- Capture network traffic and dump it to a certain pcap file. Similarly, receiving SNMP traps from network hosts and saving them to a Host-log file.
- Read continuously from the pcap and Host-log files each certain time slot window (e.g. 5 seconds) based on First Input First Output (FIFO) concept.
- Check the packets consistency. If the packet flow is not consistence, it will be saved in an IPv4 reassembling database, else the generator constructs network packets for the current time slot window and assigns each packet with a unique-ID which is {*Timestamp, Protocol_Type, IP_Src, Port_Src, IP_Dst, Port_Dst*}.
- The generator performs the same procedure on Host-log file for the current time slot window, but without any reassembling because the SNMP traps send only three main users' activities which are logging status of each user, number of log in per user, and if the user uses root shell prompt. Likewise, each host event will be assigned by a unique ID as mention in the previous step.
- For the current time slot, the generator uses the network packets to determine the direction (i.e. from initiator to responder or vise versa) then constructs the corresponding connection from the following features {*timestamp, src_ip, dst_ip, src_port, dst_port, protocol_type, services, src_byte, flag, wrong_fragment*}. It then saves the constructed connections for the current time slot into a connection database and concurrently pushes them in the first window inside the dynamic queue.
- Meanwhile, the generator examines whether any host event and network packet have the same ID for the current time slot. In other word, when the generator pushes the constructed connections inside the queue, it checks the Host-log file if any host events is received. If there are any, the generator examines the ID for the host event and synchronizes it with the same constructed connections and updates the connection vectors of the current window inside the queue. If not, it saves uncorrelated events and packets to a temporary events and connections databases, respectively.
- While the window is still inside the queue, the generator keeps looking for any update in the connection database and adjusts all values of the connection vectors of the concerned window.

- At the end of the queue, the generator pops the window out of the queue and calculates some advanced features [SB14] for each connection vector in the window.
- Finally, the generator exports the connection vectors in a comma separated value (CSV) format either as a dataset or in sequentially in a pipe.

3.2. Intrusion Detector

Constructed connections by the generator can be exported via a pipe concept to the anomaly detector and so they can be classified as normal, suspicious, or unknown in real time. The anomaly detector is based on our enhanced growing hierarchical self organizing map (EGHSOM). Basically, the GHSOM construes high dimensional data on several layers with several maps to explore supplementary details [RMD02]. In the training process, the input vectors are presented to fixed number of neurons on the map using certain number of iterations. The final best matching units on the maps are called the GHSOM model. However, there are several shortcomings on the GHSOM model, which have been addressed on our work [SB13], major enhancements are:

- Meaningful map initialization
- Splitting threshold technique to boost the final topology
- Merging best matching units to robust the final GHSOM model
- Dynamic classification-threshold confidence to detect unknown attacks

Further detailed information about our EGHSOM is available in [SB13]. The main key aspect of the EGHSOM is using a dynamic classification-confidence distance threshold to classify the incoming connection vectors via the pipe as normal, suspicious or unknown. The EGHSOM is able to update its detection model in real time and hence it is able to uncover new attacks. Figure 4 illustrates the *Intrusion Detector* components.

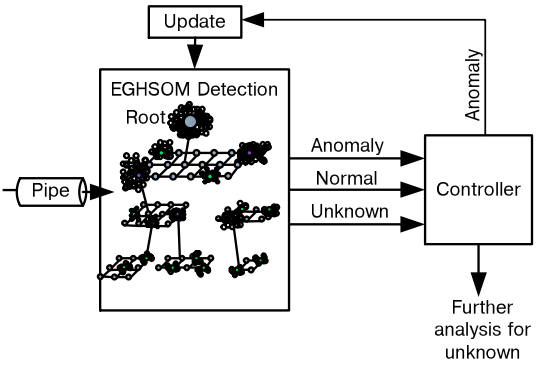


Figure 4: Internal components of the intrusion detector

The detection model classifies the connections and sends the detection result to the controller, which manages and distributes the connections as follows:

- It gathers the anomaly connections and sends them to the update model, which uses them to update the current EGHSOM detection model.
- It sends the unknown connections for further analysis by an expert.

The entire processes of EGHSOM perform the detection and update during the real time, which avoid degrading the network performance and emerging an overhead throughout the detection process.

Both parts of our model operate in the online operational mode in real time. The proposed model aggregates massive data flows of large scale network, analyzes and processes them to construct connection vectors, and sends them via a pipe for classifying the security state of the network system.

4 Performance Evaluation

We have evaluated the proposed model using synthetic and realistic data sources. In the university campus, we have installed a test network as seen in figure 5 and synthetically generated network traffic with a maximum bit rate of 200 Mbps. We have generated the traffic based on realistic network of our industrial partner. We capture the traffic at specific point (the bridge) and use it to evaluate the proposed model. There are several services running on the test network such as HTTP, SSH, FTP, SMTP, Rsync, Telnet.

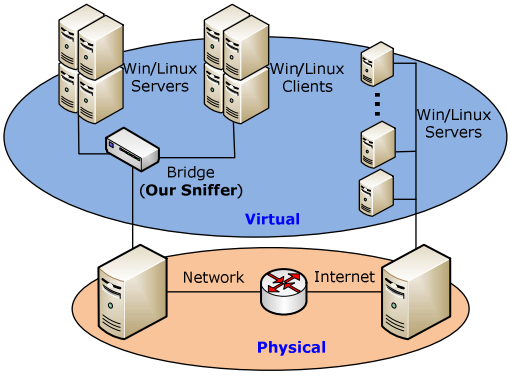


Figure 5: Architecture of the test network

Traffic generation in the test network has been performed using an automotive script. In addition, we have evaluated the model using live network traffic of an industrial firm. We have installed the model on network segmentation which has more than 10 servers and 300 clients connected. We have tried to evaluate the model while almost all users surf and browse the internet, sending emails, do other activities like sending or receiving requests. Results are summarized in the following sections.

4.1. Results of Dataset Generator

We have configured the window time slot in the dataset generator to be 5 seconds. For the gathered packets and events, we have monitored two relevant performance metrics, which are the number of processed packets inside each window and the required time to process these packets and export them as a dataset (connection vectors) based on the exporting phase.

Figure 6 illustrates the number of processed packets inside each window in the dynamic queue for the test network data sources.

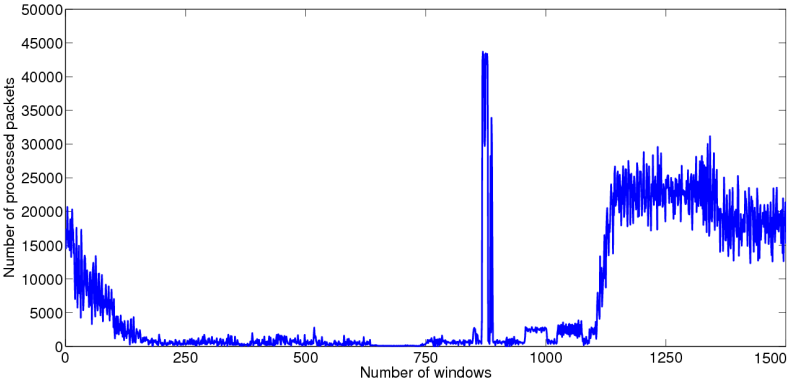


Figure 6: Number of processed packets per window for the test network

Our traffic generator generates network traffic randomly and so the number of processed packets is varying from window to another. Therefore, some windows could processed up to 45000 packets whereas others less than 1000 packets. Each window needs a specific time to process packets and events and then export them accordingly. This time should be not greater than the length of the window itself; otherwise the packet loss will increase rapidly. Figure 7 shows the processing time needed for each window.

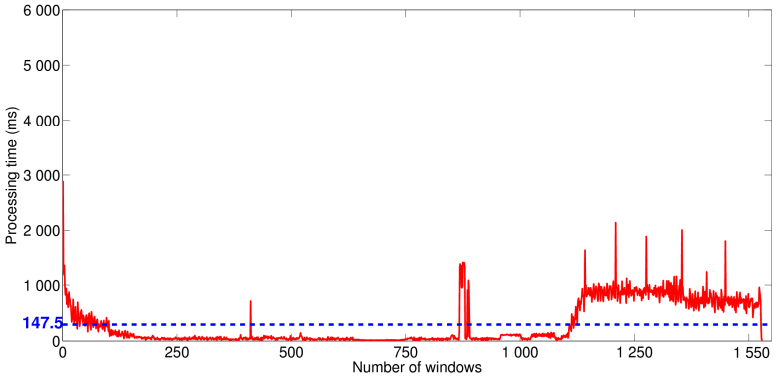


Figure 7: Processing time per window for the test network

The processing time was less than the window length and the packets could be processed without any loss. Similarly, the model could process the live traffic of a realistic network of industrial firm within the length of the window as shown in figures 8 and 9.

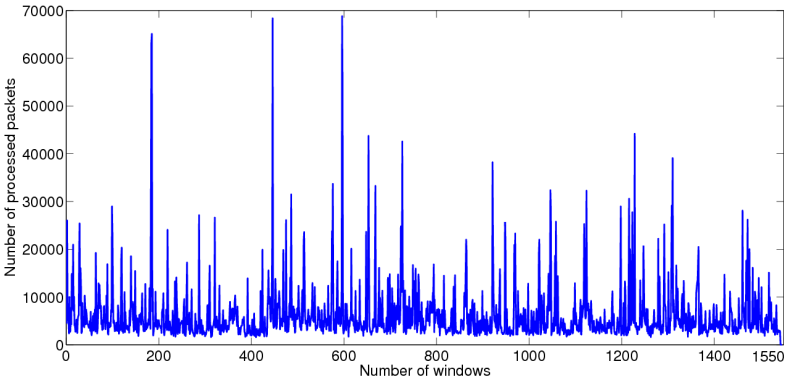


Figure 8: Number of processed packets per window of live traffic

The maximum number of packets has achieved 70000 packets per window within sufficient time as shown in figure 9.

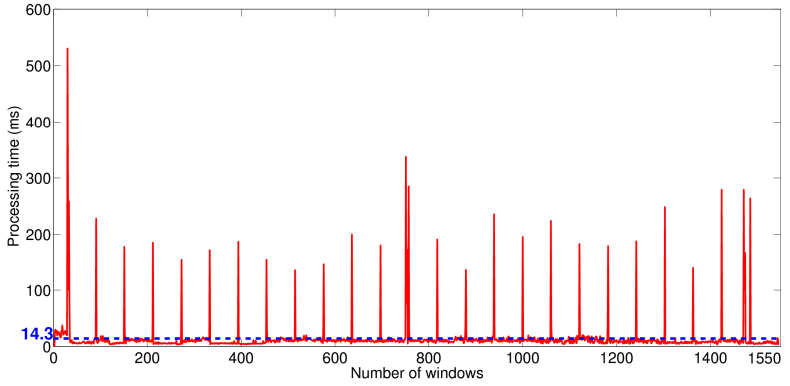


Figure 9: Processing time per window of live traffic

We notice that at the beginning the processing time is a bit more than the window length. This is related to the start phase of the model. If we consider the number of processed packets at the beginning we see that the number is less than 3000 packets; this number means that the model could process these packets surely in less than the window length but as we mentioned beforehand, at the start up of the model can cause some extra processing time.

4.2. Results of the Intrusion Detector

In this section we will briefly show that the detector can effectively classify the exported traffic from the dataset generator. The detector as shown in figure 4 is based on the proposed EGHSOM model [SB13] which is able to classify the constructed connections from the dataset generator to three labels, normal, anomaly and unknown.

We have evaluated the model in a realistic live network. The evaluation parameters can be summarized in table 1.

Characteristic	**Value**
Size of aggregated packets und events	4 GB
Total time of operation	3 hours
Number of constructed connections	303604

Table 1: Parameters of evaluation in real time

The final detection result throughout the evaluation in real time is shown in table 2.

Label of detected connections	**Number of connections**	**Percentage**
Normal	303544	99.9 %
Anomaly	57	0.0188 %
Unknown	3	0.0009 %

Table 2: Evaluation result of EGHSOM in real time

The detector could detect anomaly connections, which most likely related to advertisement servers, as our search confirmed. Other connections could be classified as unknown because they did not fall into the threshold classification margin (see [SB13]). These connections are considered suspicious and need to be further analyzed.

On both cases of synthetic and real time data flows, the model could effectively aggregate, process the traffic and reveal anomaly connections. However, the network speed of both cases was not large enough to proof the feasibility of the model. In the test network we have reached about 250 Mbps although we have installed more than 500 virtual clients on it. The network firm operates on 1 Gbps to 10 Gbps but our model is installed on network segmentation where about 300 users work on their daily activities. However, in the live test, the bit rate was almost 200 Mbps. Therefore, evaluating the model on a bit rate of 1Gbps is still not practically performed.

We have evaluated the proposed model on other synthetic and realistic environments and obtained almost the same result, which states that the generator can effectively aggregate network traffic and construct connection vectors continuously. The detector can precisely classify the constructed connections and update the EGHSOM model appropriately.

5 Conclusion

This paper presents a comprehensive model that consists of two main parts, the dataset generator and the intrusion detector. The former aggregates network packets and hosts' events, processes them and performs a correlation process to construct connection vectors using a dynamic queuing concept, and then it exports these connections in a dataset form or directly into a pipe for the detector. The detector is based on an intelligent EGHSOM model which is able to classify these connections to normal, anomaly or unknown based on a threshold margin. The model has been evaluated on a test network and on a real network firm. It could handle a large amount of packets for each window and effectively reveal anomaly connections. However, in the live test we could not reach a bit rate of 1 Gbps or even 10 Gbps to examine the plausibility, capability and scalability of the model in the high speed networks. This shortage refers to the limited number of users and small number of running services as well. In the future, the model will be enhanced to be able to classify the unknown connections. Moreover, it will be further evaluated on a large scale computer networks up to 10 Gbps.

References

[BS06] A. Biswas and P. Sinha, "On improving performance of Network Intrusion Detection Systems by efficient packet capturing," in *Proc. IEEE/IFIP NOMS*, pp.1-4, April 2006

[GJP12] V. Gulisano, R. Jimenez-Peris, M. Patiño-Martinez,C. Soriente, P. Valduriez, "StreamCloud: An Elastic and Scalable Data Streaming System," in *IEEE Trans. TPDS*, vol.23, no.12, pp.2351-2365, Dec. 2012.

[HB11] A. Hofmann, B. Sick, "Online Intrusion Alert Aggregation with Generative Data Stream Modeling," in *IEEE Trans. TDSC*, VOL. 8, NO. 2, pp. 282-294, MARCH-APRIL 2011.

[HW11] C. Hsu and S. Wang, "Embedded Network Intrusion Detection Systems with a Multi-core Aware Packet Capture Module," presented in *IEEE ICPP*, pp.207-213, Sept. 2011.

[LLZ10] Li Lin, C. Leckie, C. Zhou, "Comparative Analysis of HTTP Anomaly Detection Algorithms: DFA vs N-Grams," International Conference on Network and System Security (*NSS*), pp.113-119, Sept. 2010.

[LS12] Q. Liao and A Striegel, "Intelligent network management using graph differential anomaly visualization," in *Proc. IEEE/IFIP NOMS*, pp.1008-1014, 2012.

[LYW13] Y. Lee, Y. Yeh, Y. F. Wang, "Anomaly Detection via Online Oversampling Principal Component Analysis," in *IEEE Trans. TKDE*, pp. 1460-1470, VOL. 25, NO. 7, JULY 2013.

[MR99] M. Roesch, "SNORT-Lightweight Intrusion Detection For Networks," in *Proc. Of USENIX LISA* '99, pp. 229-238, 1999, Seattle, WA.

[MS08] C. Morariu and B. Stiller, "DiCAP: Distributed Packet Capturing architecture for high-speed network links," presented in IEEE Conf. on Local Computer Networks, pp.168-175, Oct. 2008.

[RMD02] A. Rauber, D. Merkl, M. Dittenbach, "The growing hierarchical self organizing map: exploratory analysis of high-dimensional data," in *IEEE Trans. TNNLS*, Vol. 13, Is. 6, pp. 1331-1341. Nov. 2002.

[SB13] M. Salem and U. Buehler, "An Enhanced GHSOM for IDS," in *Proc. IEEE SMC:Cybernetics*, Manchester, UK, 2013, pp. 1138-1143.

[SB14] M. Salem and U. Buehler, "Persistent Dataset Generation using Real time Operative Framework," in *IEEE Workshop on Computing, Networking and Communications*, Honolulu, Hawaii, USA, 2014, pp. 1023-1027.

[SBR11] M. Salem, U.Buehler, S. Reissmann, "Improved feature selection method using SBS-IG-Plus," in *Proc. of the ISSE Symposium*, Prague, Czech Republic, 2011, pp 352–36.

[SN13] Snort Network Intrusion Prevention and Detection System. Avaliable [Online] http://www.snort.org. Accessed 15 March 2014.

[TK95] T. Kohonen, Self-Organizing Maps. Berlin, Germany: Springer- Verlag, 1995.

[VHS11] A.R. Vasudevan, E. Harshini, and S. Selvakumar, "SSENet-2011: a Network Intrusion Detection System Dataset and its Comparison with KDD CUP 99 Dataset," presented in *AH-ICI*, Kathmandu, Nepal, 2011, pp 1-5.

[XC07] X. Chun, et al. "Research of DoS Intrusion Real time Detection Based on Danger Theory," Presented in the Symposium Data, Privacy, and E-Commerce, *ISDPE*, Nov. 2007, pp.209-211.

[ZCL11] J. Zhao, M. Chen, and Q. Luo, "Research of Intrusion Detection System Based on Neural Networks," in *IEEE-ICCSN*, Xi'an, China, May 2011, pp. 174-178.

Semantic Self-Management and Mediation for Home Service Platforms

Jan Schaefer,
RheinMain University of Applied Sciences,
Design Computer Science Media Department, Distributed Systems Lab,
Unter den Eichen 5, D-65195 Wiesbaden, Germany,
http://wwwvs.cs.hs-rm.de, jan.schaefer@hs-rm.de

Abstract:
Service-oriented computing has long outgrown the enterprise. Today, personal living environments feature an increasing number of convenience-, health- and security-related applications provided by distributed interacting devices and services (often called smart homes), which do not only support users but require management tasks such as application installation, configuration and continuous maintenance. These tasks are becoming complex and error-prone. One way to escape this situation is to enable service platforms to configure and manage themselves. The approach presented here extends services with semantic descriptions to enable platform-independent autonomous service level management and mediation using model driven architecture and autonomic computing concepts. It has been implemented as a semantic autonomic manager, whose concept, prototypical implementation and evaluation are presented.

1 Introduction

The *Service-Oriented Architecture* (SOA) principle has spread from global enterprise systems to personalized online services, mobile devices and increasingly smart homes. Services in these areas are usually implemented on top of a middleware called *service platform*. Platforms tailored to the requirements of personal living environments are typically called *home service platforms*.

Constant technological advances and a multitude of vendors with diverse intentions lead to an increasing fragmentation and heterogeneity between these platforms and services severely complicating interoperability and, thus, effective IT management. The situation is even worse in personal living environments as they usually lack a dedicated tech-savvy administrator, who is able or willing to constantly ensure correct operation of smart home services. As a consequence, service platforms should expose autonomic traits such as self-management according to *Service Level Agreements* (SLAs), which they are lacking at the moment.

Services offer public interfaces to enable interaction. These interfaces define a service's functionality syntactically, but they do not describe its semantics. A human service consumer bridges this gap by relying on additional information sources such as documentation or mutual understanding to learn what functionality a service is actually providing. How-

ever, with increasing numbers of services in personal living environments, the manual decision making process has become too complex and, thus, infeasible. Here, semantic web technologies are being discussed as means to provide machine-processable, semantic descriptions of service capabilities to enable automated service mediation. Platforms supporting semantic services are called *semantic service platforms*. Some platforms already use semantics to define context models.

This paper presents a semantic self-management approach for service-based distributed applications. Here, managed services provide semantic descriptions of their functional and management capabilities. *Model Driven Architecture* (MDA) [OMG03] concepts are used to define abstraction levels for platform-independent and platform-specific semantic service and service management models as well as accompanying platform-specific code. Using these abstractions and mappings, concrete instances of formally defined semantic services can be managed according to predefined QoS requirements. In addition, the service instance's QoS properties can be used for dynamic QoS-dependent service binding.

The approach relies on an autonomic manager called *Self-Manager* implementing the MAPE-K control loop — *M*onitor, *A*nalyze, *P*lan, *E*xecute and *K*nowledge — as defined by the *Autonomic Computing Initiative* (ACI) [KC03]. In addition, the self-manager uses semantic models, rules and queries in its knowledge base to manage services on an abstract level: To achieve this, runtime service monitoring data is mapped to high-level models, which are then processed. The derived (counter)actions are executed in managed services via their management agents or interfaces. This decouples the management knowledge and process from the underlying implementation technologies in the same way that semantic service descriptions are defined independently from their respective implementations. Also part of the approach but not presented here due to limited space is a concept for self-healing services, which enables fault-tolerant service execution on distributed computing nodes. As the self-manager itself is service-based, it can be executed redundantly as well, effectively preventing a single point of failure.

The remainder of this paper is structured as follows: Section 2 reviews work related to the approach presented here. Section 3 introduces the concept of the self-management approach for semantic applications, which is followed by a description and evaluation of its prototypical implementation in section 4. The paper ends with a conclusion and a discussion of future work in section 5.

2 Related Work

The *Semantic Markup for Web Services* (OWL-S) [W3C04] was originally intended to be used with Web services as implementation technology, but thanks to its modular design other technologies can be integrated. The SLM ontologies are based on the *UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms Specification* (QFTP) [Obj08], which defines QoS characteristics (properties), QoS constraints and fault tolerance definitions (among others), by the *Object Management Group* (OMG). Apart from the fault-tolerance concepts, the *QoS-MO Ontology for Semantic*

Modeling [TS08] already provides QFTP-derived concepts. As OWL-S and QoS-MO are described sufficiently by respective publications, however, they are not covered here in detail. This section only explains their basic purpose as well as extensions required for the work presented here.

The *OSGi Service Platform* (OSGi) [OSG12a] specification published by the OSGi Alliance provides a component model for the development of modular Java-based applications. Since its initial release, it has been tailored to the needs of embedded and enterprise applications alike. The increasing popularity of OSGi led to broadening tool support and integration into embedded environments such as car entertainment systems as well as enterprise applications.

Even before the arrival of semantic services in particular, several context modeling approaches for service platforms emerged. Gu, Pung, Zhang [G+04] propose using layered ontologies for context models in the home domain. Their work covers concepts that have become widely-acknowledged and, thus, also appear in more recent publications (e.g. activities, locations, persons, devices including custom service concepts). However, the interactions between formal service definitions and runtime systems (the actual service groundings) and how user preferences influence the platform's behavior are not discussed. Diaz Redondo et al. [R+08] propose the combination of OWL-S with an OSGi grounding called *OWL-OS*. Here, the authors use OWL-S service categories as a means to group services into aspects (e.g. lighting), which can be used by clients during service look-up.

With respect to home service platforms, especially *universAAL* [H+11] has to be considered. It combines the results of several European projects aiming to establish a platform in the market. It uses custom ontologies for modeling context and OWL-S for describing services.

Romero et al. [R+11] propose a platform based on the *Service Component Architecture* (SCA), which provides service and component abstractions similar to OSGi. The paper focuses on device integration over heterogeneous communication protocols into an event processing architecture, which is able to process data from a wide collection of devices (including sensor networks). Despite the similarities to the work presented in this paper, however, all these platforms lack SLM capabilities and support for autonomic behavior.

A-OSGi [F+10] implements the MAPE-K control loop on top of OSGi aiming at the management of the OSGi container. The logic of the analysis component, however, is programmed in Java and semantic modeling is not supported.

3 Semantic Self-Management

3.1 Abstraction Layers

Figure 1 presents the three abstraction layers (aligned horizontally) and four functional areas (aligned vertically) that form the foundation of the semantic self-management approach. The first functional area is the *Application Context*, with which semantic services interact by referencing context-provided entities. As the management model is defined

independently from context knowledge, however, the model is not covered in this paper.

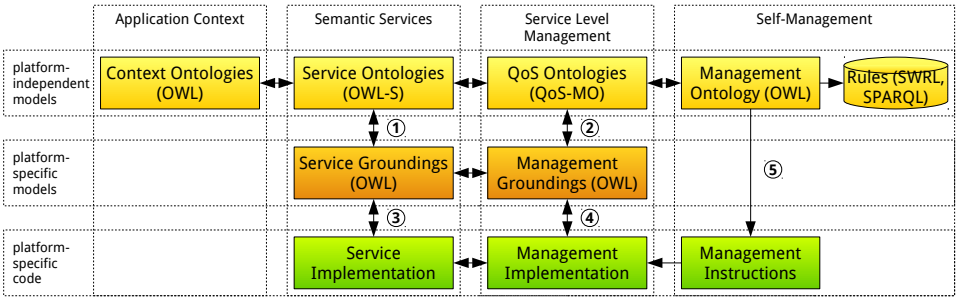


Figure 1: Self-Management Layers

Semantic Services are defined using the OWL-S ontologies. Providing the information required to control service execution, the *Service Level Management* must be integrated with the semantic services model delivering additional management information such as QoS properties. SLM-related properties are defined using the QoS-MO ontology. This modeling concept of this approach has been presented in [S+12] in detail already.

The last functional area, *Self-Management*, relies on a custom management model and accompanying management rules to evaluate formal service definitions against existing implementations at runtime to control service execution. Thus, the accompanying management rules can access both platform-independent and platform-specific models. On the platform-independent ontology layer, the *Quality of Service* (QoS) and *Service* ontologies are defined independently from actual *Service Grounding* and *Management Grounding* implementation technologies. This concept follows the *Separation of Mechanism and Policy* design principle as management queries can be adapted to specific requirements and can be replaced at runtime.

3.2 Grounding Mapping Ontologies

This approach features management logic that is independent from the actual technologies used to implement and manage services. Abstractly defined management decisions are only concretized before execution. Following the MDA paradigm, this separation requires mappings from platform-independent (abstract) to platform-specific (concrete) ontologies. As a result, each to be integrated service implementation technology only requires a new target grounding and accompanying mappings, which abstract from available service and management agent implementation details (e.g. methods and parameters).

To demonstrate the applicability of the selected approach, mapping ontologies from OSGi to OWL-S and *Java Management Extensions* (JMX) to QoS-MO have been developed. OSGi was selected as it is becoming the dominant service implementation technology in smart homes (cf. related work in section 2); JMX was selected as it is the OSGi standard management technology [OSG12b].

Figure 2 (left) shows how OSGi services are mapped to OWL-S entities (transition (1) in Figure 1). In OSGi, services feature an abstraction level themselves: They consist of a service interface (called *Service* in the mapping), whose implementation is provided in a bundle (*ServiceImplementation*). On an abstract level, an OWL-S service has a *Grounding* for each registered OSGi service interface and an *AtomicProcessGrounding* for each actual OSGi service implementation.

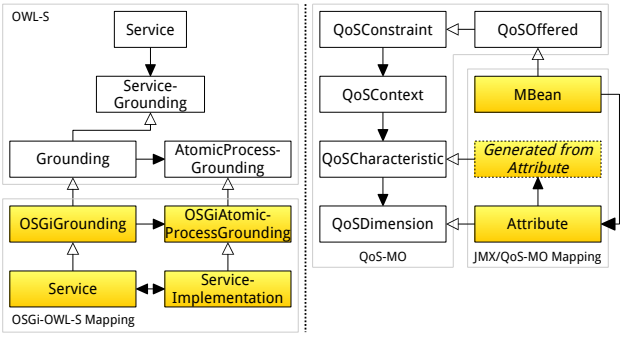


Figure 2: Ontology Groundings

Figure 2 (right) shows how JMX MBeans are mapped to QoS-MO entities (transition (2) in Figure 1). In JMX, managed entities are represented by MBeans, which are realized as Java classes implementing an associated MBean interface. For the approach presented here, services managed through JMX have to implement at least one OSGi service interface and MBean interface each. By convention, MBean attributes should be prefixed with "Current", so that the respective *QoSCharacteristic* sub-concept can be selected during runtime mapping (example: An *Attribute*/*QoSDimension* called *CurrentProcessingTime* results in the selection of the OWL concept *ProcessingTime*). MBeans and their attributes represent *QoSOffered* properties, which can be evaluated against QoS requirements.

3.3 Management Ontology and Rules

The management ontology defines a main concept called *Action*, which is used in the management cycle. Actions describe countermeasures for QoS violations. The ontology is accompanied by two types of management rules: First, there are generic management rules that constantly check predefined QoS requirements against QoS offerings of currently available service implementations. For example, QoS requirements define limits for QoS properties that service implementations have to honor. Depending on the qualification of the constraint (i.e. hard or soft limit), a violation can result in a warning or service fault. Second, service consumers can supply additional QoS requirements that mediated service implementations have to deliver.

3.4 Self-Manager Architecture

The *Self-Manager* (SM), whose architecture is depicted in Figure 3, is a semantic autonomous controller based on the MAPE-K control loop model. The SM consists of five modules, which communicate through events (publish/subscribe model): The *Monitor*, *Analyze*, *Plan* and *Execute* modules provide common functionality for their respective module type. These modules use a shared semantic knowledge base provided by the *Knowledge* module. The SM uses *SPARQL Protocol And RDF Query Language* (SPARQL) [W3C13] management queries representing *Event Condition Action* (ECA) rules to replace administration tasks usually executed by humans. It also extends service platforms that do not necessarily support semantics by providing a *Semantic Service Registry* (SSR), which is used to register services and their *Management Agents* (MA).

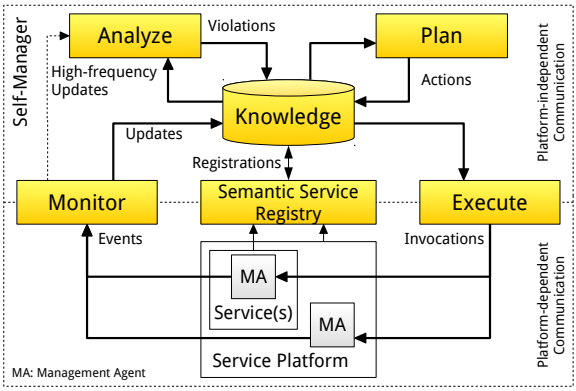


Figure 3: Self-Manager Architecture

Like the ontology layers presented in section 3.1, the SM also follows the MDA paradigm: The monitor and execute modules as well as the SSR map from implementation technologies to platform-specific and platform-independent ontology statements and vice versa (transition ③ in Figure 1). The other modules (analyze, plan and knowledge) only process platform-independent ontology statements resulting in a generic management logic. This allows reusing management rules, even if new service implementation technologies are introduced.

Monitor Modules retrieve platform-specific runtime management information and map it to the ontological representation of the service instance in the knowledge base as QoS offered properties using the corresponding ontology concepts described in section 3.2 (transition ④ in Figure 1). From this moment on, the properties can be used for the management of the service instance or as semantic filter properties by other services during the service selection and binding process (more details on this in the subsequent section 3.5).

Analyze Modules are notified, if QoS properties of monitored services have changed. Whenever an analyze module detects a QoS constraint violation (e.g. exceeded response time limit), it marks the respective QoS offering. Depending on the module implementation, analyze modules might also add facts derived from collected updates to the knowledge

base. The SPARQL query shown in Listing 1 serves as an example, which is used in the prototypical implementation. Here, the current execution time of a service implementation, i.e. of one of its atomic processes, is evaluated against the defined maximum execution time. In case the former is higher than the latter, a QoS violation has been detected.

```
CONSTRUCT {
  ?qosOffered itm:isInViolation true .
}
WHERE {
  ?atomicProcess a proc:AtomicProcess .
  ?qosRequired a qos:QoSRequired .
  ?qosRequired qom:hasTarget ?atomicProcess .
  ?reqExecTime a qos:ExecutionTime .
  ?qosRequired qos:hasContext ?reqExecTime .
  ?maxExecTime a qos:MaxExecutionTime .
  ?reqExecTime qos:hasDimension ?maxExecTime .
  ?maxExecTime qos:hasPrimitiveValue ?maxExecTimeValue .
  ?qosOffered a qos:QoSOffered .
  ?qosOffered qom:hasTarget ?atomicProcess .
  ?offExecTime a qos:ExecutionTime .
  ?qosOffered qos:hasContext ?offExecTime .
  ?offExecTime qos:hasDimension ?qosDimension .
  ?qosDimension a qos:CurrentExecutionTime .
  ?qosDimension qos:hasPrimitiveValue ?currExecTimeValue .
  FILTER( ?currExecTimeValue > ?maxExecTimeValue ) .
}
```

Listing 1: Analyze Query Example

Plan Modules are notified, if violations were added to the knowledge base. These modules are responsible to compute the response to a violation. As they are able to execute queries against the knowledge base, they select a management action based on service-related knowledge and management information provided by the targeted managed service. The response action is stored in the knowledge module and an update event is sent to available execute modules. The SPARQL query shown in Listing 2 is taken from the prototypical implementation: a service restart action is created for each service implementation that exceeds the predefined maximum execution time of its associated semantic service.

```
CONSTRUCT {
  ?qosOffered itm:requiresAction itm:RestartAction .
}
WHERE {
  ?qosOffered a qos:QoSOffered .
  ?qosOffered itm:isInViolation true .
  ?atomProcGrnd a grnd:AtomicProcessGrounding .
  ?qosOffered qom:hasTarget ?atomProcGrnd .
}
```

Listing 2: Plan Query Example

Execute Modules look for management actions they can execute on the managed service implementation respectively its platform. In contrast to analyze or plan modules, however, monitoring and execute modules are platform-specific, as they have to interact with the actual service implementations. The SPARQL query shown in Listing 3 serves as an example, which is used in the prototypical implementation (transition ⑤ in Figure 1). Here, the execute module queries the knowledge base for actions targeted at OSGi services executed in an OSGi framework manageable via JMX. For this, it needs the type of action

to be executed, the bundle ID of the targeted service and the URL of the MBean server. The action can be any action applicable to bundles deployed in an OSGi framework.

```
SELECT
  ?action ?bundleID ?url
WHERE {
  ?qosOffered a qos:QoSOffered .
  ?qosOffered itm:requiresAction ?action .
  ?target a osgi:ServiceImpl .
  ?qosOffered qom:hasTarget ?target .
  ?bundle a osgi:Bundle .
  ?target osgi:hostedByBundle ?bundle .
  ?bundle osgi:hasBundleID ?bundleID .
  ?mbean a jmx:MBean .
  ?mbean jom:manages ?target .
  ?mbs a jmx:MBeanServer .
  ?qosOffered jmx:hostedByServer ?mbs .
  ?mbs jmx:hasURL ?url .
}
```

Listing 3: OSGi/JMX Execute Query Example

If violations without counteractions are detected after the planning phase (i.e. when all plan modules are done processing), the self-manager is unable to handle the violation autonomously . In this case, the issue has to be escalated by specialized execute modules supporting escalation actions (e.g. by sending e-mails to human specialists).

Depending on the underlying service platform, the SSR can retrieve service registrations from existing (non-semantic) service registries to allow mixed usage of semantic and non-semantic services, which means that not all (existing) services have to be outfitted with semantic descriptions.

If a monitored source delivers high frequency updates or if raw data should be preprocessed first (e.g. high-volume sensor data), updates can be made available via a publish/subscribe message channel instead of being added to the knowledge module, which reduces ontology processing effort (i.e. model changes and reasoner updates). Analyze modules can subscribe to this topic allowing for modules working on either ontological knowledge or raw monitoring data (e.g. for *Complex Event Processing* (CEP) applications).

3.5 Manageable Semantic Service Development

The manageable semantic service development process consists of the phases depicted in Figure 4: During *Service Modeling*, the semantic service is defined in OWL-S; QoS properties and constraints, which are required from service implementations, are added during *QoS Modeling* (optional). During *Technology Integration*, the actual service and management agent implementations and the OWL-S and QoS-MO grounding definitions are created, which allows mapping implementations and semantic definitions. This step is required only once for each service/management implementation combination. At *Runtime*, the self-manager's monitor modules create facts reflecting the service's runtime state.

The development process for semantic applications presented here is a hybrid approach. An

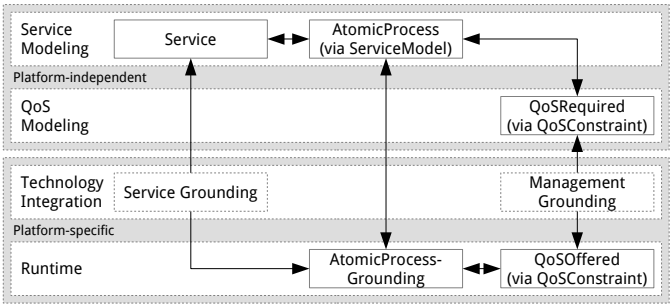


Figure 4: Manageable Semantic Service Development Phases

application can be implemented in a native programming language with platform-specific service interactions, if the application is unaware of semantics. However, to use semantic services (i.e., select services based on QoS properties), the client application has to define a semantic filter that is interpreted by the SSR and can feature requirements for the semantic service and actual instances (sample shown in Listing 4). To provide a semantic service, an application requires an ontological service description defined — if the formalized service or service version is unknown to the registry — and a service grounding describing how the syntactic service interface can be mapped to the abstract description. If both exist, the service can define additional QoS-related properties.

```
# semantic service requirement (predefined)
(MaxExecutionTime<=300)
# service instance requirement (monitored at runtime)
(CurrentExecutionTime<=150)
# Combined expression
(&(MaxExecutionTime<=300)(CurrentExecutionTime<=150))
```

Listing 4: Semantic service selection filter (sample)

Whenever a semantic service registers with the SSR, its provided ontologies are imported into the knowledge base and facts for the new service instance are created. If the service provides a management agent, the service is added to the list of monitored services. The monitoring module either subscribes to or polls for service attribute changes, which are used to update the available ontological facts. In contrast to traditional service property definitions, this approach supports dynamically changing service properties.

4 Prototypical Implementation

This section briefly describes the prototypical implementation of the self-management concept presented in section 3, which uses OSGi and JMX. For ontology processing, the *Apache Jena* framework is used in combination with the *Pellet OWL 2 Reasoner*.

OSGi event listeners are used to keep the models in the knowledge base and the appli-

cation's runtime state synchronized; they intercept bundle start notifications and import provided ontologies. OSGi service find hooks intercept registrations of semantic service implementations; They create OWL representations for them and map these to the associated abstract semantic services.

The JMX grounding is integrated with the QoS-MO ontology and allows representing runtime information provided by MBeans as QoS characteristics. Here, MBean server properties such as thread counts or memory consumption can be used as well as custom MBean attributes, constructors, operations and their respective parameters. Similar to the OSGi grounding, MBean notification listeners are used to synchronize the ontologies in the knowledge module with the MBean server's state. In addition, the prototypical implementation uses listeners to monitor MBean registrations and deregistrations as well as attribute changes and updates the knowledge module accordingly.

Both QoS dimensions defined for the associated semantic service (i.e. requirements) and QoS dimensions monitored at runtime (i.e. offerings) are added to the service properties of OSGi service instances, so that service consumers can use semantic filter expressions as shown in Listing 4 to express their requirements for mediated services. This allows defining service dependencies in a way that the standard OSGi service registry can bind and rebind services and consumers automatically depending on QoS properties and requirements.

A first evaluation of the prototype focused on the self-manager's module interaction and update processing capabilities. It featured a simple, arbitrary semantic test service, whose execution time was artificially increased with each processed request by 50ms starting at 100ms. The execution time was exposed via JMX and semantically represented as a QoS offering. A single consumer invoked the service continuously without additional delays between requests.

To demonstrate the self-manager's semantic management capabilities, the QoS requirement of a maximum execution time of 500ms was defined for service groundings implementing the associated semantic service. Once an exceeding execution time monitored by the JMX monitoring module was detected by the analyze module, the latter created a QoS violation alert, which resulted in the selection of a counteraction by the plan module (here: service grounding restart) and executed by the JMX execute module. After the restart, invocation times once again increased with each request. Table 1 presents the measurement results for 26853 service invocations recorded on a Core i5 (all times in milliseconds).

The difference between JMX monitor module updates and analyze module activity (26853 vs. 24803) was caused by management cycle activity: If an update arrives during cycle activity, the update is stored until the next cycle. However, only the latest update is buffered. If the self-manager received updates without QoS violations, the cycle ended after the analysis taking 30.3ms on average. If a violation was found and an action had to be taken, the cycle took 67.2ms on average (from MBean notification to finished OSGi service restart). Using the *Oracle VisualVM* diagnostics tool, it was verified that large parts of the management cycle execution time is spent on implicit ontology consistency checks that take place before each SPARQL query, which delays each query execution.

Service mediation, on the other hand, was not affected by ontology handling. There was a minimal delay updating OSGi service properties (<1 ms), whenever a service implementa-

Tag	Average	Min	Max	Std.Dev.	Count
JMX Monitor Module (Update)	0.2	0	56	0.9	26853
Inference Model Synchronization	5.1	4	85	1.2	24803
Analyze Module	23.7	21	199	3.0	24803
Plan Module	2.0	1	15	1.8	2502
JMX Execute Module (Query)	5.2	4	93	2.9	2502
JMX Execute Module (Action)	27.3	24	76	2.9	2502
Management Cycle (No Violation)	30.3	26	229	3.4	22301
Management Cycle (Execution Complete)	67.2	60	151	5.6	2502

Table 1: Execution Times

tion's QoS properties changed

5 Conclusion and Future Work

The semantic self-management concept and the accompanying autonomic manager prototype presented in this paper show that the integration of Semantic Web, Model Driven Architecture and Autonomic Computing concepts is viable. The separation of service modeling and implementation aspects allows integrating new service or management agent implementation technologies without modifications to the descriptions of the semantic services or existing management rules.

The initial evaluation of the prototype showed, however, that if real-world state changes are committed to the knowledge base immediately (without pre-processing using, for example, a CEP engine as described in section 3), model updates and synchronization delays caused by reasoner updates result in management cycle execution times, which might not be suitable for time-critical applications. For personal living environments with a limited amount of participating nodes, however, they might be acceptable.

For semantic service mediation, the prototype assumes that the services are implemented as OSGi services as methods, parameters and data types are not matched semantically during service look-up (i.e. the service name must be an OSGi interface name). However, semantic QoS properties are used as OSGi service properties (and continuously updated), which allows using them as service look-up filters and to control mediation between consumers and services through standard OSGi mechanisms (e.g. service rebinding in case of performance degradation). Since OSGi is becoming the dominating service platform technology in personal living environments, however, this does not restrict its applicability.

Although this semantic self-management approach focuses on home service platforms, it is not limited to this application area but can be used in the enterprise as well. In addition, the importance of autonomic properties and semantic web technologies is constantly increasing in especially but not limited to the areas of cyber-physical systems and machine-to-machine communications (e.g. [ETS12]), which opens up new possibilities for future applications.

The prototype has been evaluated in the WieDAS research project funded by the *German Federal Ministry for Education and Research* (BMBF) under contract number 17040B10. Here, the testing focus was on the general real-world applicability of the approach in personal living environments.

References

[ETS12] ETSI Technical Committee M2M Communications. Machine to Machine Communications (M2M); Study on Semantic support for M2M Data, December 2012.

[F+10] João Ferreira et al. A-OSGi: A Framework to Support the Construction of Autonomic OSGi-Based Applications. In Athanasius V. Vasilakos, Roberto Beraldi, Roy Friedman, et al., editors, *Autonomic Computing and Communications Systems*, volume 23, pages 1–16. Springer Berlin Heidelberg, 2010.

[G+04] T. Gu et al. Toward an OSGi-based Infrastructure for Context-Aware Applications. *Pervasive Computing, IEEE*, 3(4):66 – 74, 2004.

[H+11] Sten Hanke et al. universAAL – An Open and Consolidated AAL Platform. In Reiner Wichert and Birgid Eberhardt, editors, *Ambient Assisted Living*, pages 127–140. Springer Berlin Heidelberg, Berlin, Germany, 4. AAL-Kongress edition, 2011.

[KC03] J.O. Kephart and D.M. Chess. The Vision of Autonomic Computing. *Computer*, 36(1):41–50, January 2003.

[Obj08] Object Management Group. UML Profile for Modeling Quality of Service and Fault Tolerance Characteristics and Mechanisms Specification (Version 1.1), April 2008.

[OMG03] OMG Model Driven Architecture Working Group. MDA Guide Version 1.0.1, 2003.

[OSG12a] OSGi Alliance. OSGi Service Platform v5.0 Core Specification, March 2012.

[OSG12b] OSGi Alliance. OSGi Service Platform v5.0 Enterprise Specification, March 2012.

[R+08] R.P. Diaz Redondo et al. Enhancing Residential Gateways: A Semantic OSGi Platform. *IEEE Intelligent Systems*, pages 32 –40, Jan.-Feb. 2008.

[R+11] Daniel Romero et al. The DigiHome Service-Oriented Platform. *Software: Practice and Experience*, 2011.

[S+12] Jan Schaefer et al. QoS-based Testing and Selection of Semantic Services. In *ARCS 2012 Workshops (ASPRIT)*, Lecture Notes in Informatics (LNI), pages 15–26. Gesellschaft für Informatik e.V. (GI), Koellen Druck+Verlag GmbH, Bonn, February 2012.

[TS08] Gustavo Fortes Tondello and Frank Siqueira. The QoS-MO Ontology for Semantic QoS Modeling. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 2336–2340, New York, NY, USA, 2008. ACM.

[W3C04] W3C. OWL-S: Semantic Markup for Web Services (W3C Sub.), November 2004.

[W3C13] W3C. SPARQL 1.1 Query Language (W3C Recommendation), March 2013.

IT-Unterstützung in Emergency Management & Response

(IT-Rettung)

IT-Unterstützung in Emergency Management & Response (IT-Rettung)[1]

Abstract: Die Informationstechnik kann in Notfallsituationen sowie bei der Vorbeugung derartig kritischer Situationen wichtige Beiträge zum Schutz und der Rettung des menschlichen Lebens liefern. IT-Unterstützung für das Notfallmanagement kann in der Verbesserung der Vorbereitung, der Abwehr und der Bewältigung von Schadenslagen helfen. Anwendungsbeispiele zeigen, dass sie z. B. die Koordination zwischen Einsatzkräften verbessern und die Lagefeststellung beschleunigen kann. IT soll in diesen Situationen helfen, komplexe und kritische Situationen zu beherrschen. Allerdings zeigen Erfahrungen, dass der Einsatz von IT als operatives Einsatzmittel oder Entscheidungsunterstützungswerkzeug auch für eine Zunahme der gefühlten Komplexität einer Einsatzlage bei vielen Beteiligten sorgen kann. Im Fokus des Workshops stehen die besonderen Herausforderungen und technischen Konsequenzen, die sich für die IT in diesem Umfeld ergeben, sowie ein Austausch über aktuelle Erkenntnisse aus der Entwicklung und der Erforschung von Lösungen. Uns interessieren dabei sowohl Ansätze, deren Fokus auf der Unterstützung bei der Rettung und Versorgung einer einzelnen Person liegt, als auch Ansätze, die sich mit einer breiten IT-Unterstützung im Notfall- und Krisenmanagement beschäftigen. Beide Aspekte können z. B. in Großschadenslagen gleichzeitig eine hohe Relevanz haben, wenn z. B. ein Rettungseinsatz durch die Beteiligung unterschiedlicher Organisationen und Behörden schnell zu einer komplexen Situation für einzelne Akteure (z. B. Notärzte) und Führungsstellen (z. B. einen Führungsstab) wird. Ein schneller Lageüberblick und ein effizienter, bedarfsgerechter Informationsaustausch zwischen den Beteiligten sind Ziele, die durch geeignete technische Lösungen und (IT-)Equipment vor, während und nach Notfallsituationen erreicht werden sollen.

Themen

Im Workshop sollen verschiedene Aspekte, die mit der 'IT-Unterstützung in Emergency Management & Response' verbunden sind, diskutiert werden. Der Workshop soll zum einen die Präsentation von wissenschaftlichen Beiträgen umfassen, zum anderen aber auch Raum für intensiven Erfahrungsaustausch zwischen Wissenschaftlern, Forschungspartnern aus der Industrie und Anwendern bieten.

- Soziale Medien bieten eine große Menge an Daten, die jedoch selten explizit auf die Anforderungen und Anwendungsfälle der Gefahrenabwehr zugeschnitten sind. Bewertungsmechanismen und ihre Implementierung sind notwendig, um relevante, valide und aktuelle Inhalte zu identifizieren.

[1] http://www.cik.uni-paderborn.de/workshop-it-rettung/

- Die Vorbereitung auf Gefahrensituationen muss durch Planung und Training gestärkt werden. Dabei, aber auch in laufenden Einsätzen, können Simulationen unterstützen. Dabei ist abzuwägen, wie realitätsnah diese umgesetzt werden.

- Eine besondere Situation für Rettungskräfte ist bei einem Massenanfall von Verletzten (MANV) gegeben. Auf Grund ihrer Komplexität und ihres Ausmaßes, aber vor allem auch wegen ihrer Seltenheit ist für derartige Situationen besondere (IT-gestützte) Übungsunterstützung notwendig.

- Eine besondere Problemstellung in der Einsatzleitung einer Feuerwehr ist dadurch gekennzeichnet, dass Lageberichte „aus dem Feld“ empfangen werden, jedoch auf Grund der Kommunikationskanäle kaum eine direkte Rückfrage möglich ist. Eine mobile Reporting-Applikation soll diesbezüglich helfen.

- Das Land Rheinland-Pfalz setzt eine Portallösung für das Informationsmanagement im Brand- und Katastrophenschutz ein. Diese soll auch als Statusübermittler für den Digitalfunk in den Feuerwehreinsatzzentralen des Bundeslandes eingesetzt werden.

Im Rahmen des Workshops, der in diesem Jahr zum sechsten Mal in Rahmen der GI-Jahrestagung durchgeführt wird, werden fünf Beiträge vorgestellt. Diese Beiträge adressieren auf unterschiedliche Art die Herausforderungen und komplexen Situationen, in denen IT auf der Grundlage von Daten Unterstützung bieten soll. Neben Beiträgen zu diesen Themen werden verschiedene Kurzpräsentationen zur Diskussion im Workshop beitragen und Impulse zu weiteren Themengebieten geben.

Organisatoren: Christian Erfurth (Ernst-Abbe-Fachhochschule Jena), Jens Pottebaum (Universität Paderborn, C.I.K.)

Programmkomitee: Benedikt Birkhäuser (Bergische Universität Wuppertal), Patrick-Benjamin Bök (TU Dortmund), Tina Comes (University of Agder, Norwegen), Florian Daiber (DFKI), Wolf Engelbach (Fraunhofer IAO), Christian Erfurth (Ernst-Abbe-Hochschule Jena), Therese Friberg (Universität Paderborn), Rainer Koch (Universität Paderborn), Birgitta König-Ries (Friedrich-Schiller-Universität Jena), Andreas Mitschele-Thiel (TU Ilmenau), Simon Nestler (Hochschule Hamm-Lippstadt), Jens Pottebaum (Universität Paderborn), Christian Reuter (Universität Siegen), Thomas Rose (Fraunhofer FIT), Volkmar Schau (Friedrich-Schiller-Universität Jena), Christian Wietfeld (TU Dortmund)

Entwicklung einer mobilen Reporting-Applikation zur Artikulation entscheidungsrelevanter Informationsbedarfe im Katastrophenschutz

Thomas Ludwig & Christian Reuter

Institut für Wirtschaftsinformatik, Universität Siegen
Hölderlinstr. 3, 57076 Siegen
thomas.ludwig@uni-siegen.de; christian.reuter@uni-siegen.de

Abstract: Die im Katastrophenschutz beteiligten Behörden und Organisationen mit Sicherheitsaufgaben (BOS) benötigen für Entscheidungen stets möglichst aktuelle und genaue Informationen. Diese werden dabei je nach Lage von den Einsatzkräften, welche sich am Schadensort befinden, via Funk an die teilweise entfernt sitzenden Entscheidungsträger berichtet. Unsere Design-Fallstudie untersucht interaktive Unterstützungsmöglichkeiten dieser Berichts- bzw. Reporting-Prozesse: In einer qualitativen empirischen Studie deckten wir die Kommunikationspraktiken von BOS auf und entwickelten darauf aufbauend ein Konzept zur semi-strukturierten Artikulation von Informationsbedarfen, welches als mobile Android Applikation implementiert und mit potentiellen Nutzern evaluiert wurde. Unser Beitrag präsentiert, wie zielgerichtete Anfragen von Informationen unter Berücksichtigung angemessener Metadaten das Berichts- und Meldewesen örtlich verteilter Einsatzkräfte unterstützen können.

1. Einleitung und Grundlagen

In der Lagebeurteilung von Behörden und Organisationen mit Sicherheitsaufgaben (BOS) können Entscheidungen weitreichende Auswirkungen haben. Sie basieren dabei auf Informationen inkrementeller Meldungen der Einsatzkräfte, welche sich an dem jeweiligen Schadensort befinden. Einsatzkräfte müssen unter Bedingungen unvollständiger und ungenauerer Informationen, in einem Umfeld sich ändernder Gefahren und Ziele sowie unter enormen Zeitdruck Entscheidungen treffen [Pato03]. [BhLJ09] zeigen, dass Leiter lokaler Eingreiftruppen durchschnittlich mehr als 30 Minuten für das Sammeln der Informationen und die Koordination ihrer Teams benötigten. Verschiedene wissenschaftliche und technische Ansätze adressieren bereits die Unterstützung dieser Kooperation, wobei durch die Dynamik, welche insbesondere eine Großschadenslage mit sich bringen kann, mobile (ortsungebundene) Endgeräte ein große Rolle spielen [GOPC06].

Aktuelle wissenschaftliche Ansätze betrachten die Einsatzleitung (international teilweise äquivalent zur Leitstelle) meist sowohl als Informationsproduzenten als auch -konsumenten, wohingegen die Kräfte vor Ort primär als Informationsproduzenten aufgefasst werden, welche Informationen in Form der Meldungen an die Einsatzleitung herantragen [NiSt10]. In diesem Beitag umfasst die Einsatzleitung sowohl die jene vor Ort

(z.B. Feuerwehr) als auch jene in den Leitstellen (z.B. Polizei). Im Folgenden werden wir von Entscheidungträgern sprechen und meinen damit weisungsbefungte, hierarchisch höhere Einsatzkräfte. Unabhängig vom Standort münden Einsatzentscheidungen der Einsatzleiter in Maßnahmen, welche von den weiteren Einsatzkräften ausgeführt werden. Büscher & Morgensen [BüMo07] präsentieren Prototypen, die es Leitstellen ermöglichen, Live-Informationen über die Aktivitäten der Einsatzkräfte hinsichlich einer Lagebeurteilung zu erfassen, um eine umfassende Lageübersicht herstellen zu können. Durch den Einsatz ubiqitärer Technologien werden dabei die lokalen Einsatzkräfte nicht durch eine verbale Kommunikation gestört. Catarci et al. [CLMM10] beschreiben das System ‚Workpad', in dem jede örtliche Einheit einen von einem Prozessmanagement-System überwachten PDA verwendet, welcher die Einheiten steuert. Mithilfe der mobilen Geräte können sie Aufgaben empfangen und erfassten Bildern oder Videos Kommentare hinzufügen und sogar diese mit anderen teilen. Ein weiteres Konzept wurde von Singh & Ableiter [SiAb09] vorgestellt, deren Applikation ‚TwiddleNet' es erlaubt, multimediale Daten als Feed zu senden und empfangen, wobei die Smartphones eine Doppelrolle sowohl als Server als auch Client übernehmen. Solche Applikationen, die Echtzeitberichte ermöglichen und multimediale Daten mit lokalen Informationen verknüpfen, tragen zur Erhöhung der Awareness bei [BMPM05]. Auch Bergstrand & Landgren [BeLa11] analysierten die Wirkungen von Live-Videos der Einsatzstelle in die Leitstellen und stellten eine erhebliche Verbesserung der Lagebeurteilung im Lagezentrum fest. Aufgrund des Buttom-Up-Kommunikationsflusses wurde die Informationsbereitstellung der örtlichen Einsatzkräfte jedoch von der Eigenmotivation der Einsatzkräfte vor Ort angetrieben, was zu Problemen mit der Priorisierung führte: *„When you decide to use the camera you also decide not to do other things"* (Einsatzleiter in [BeLa11]). Wu et al. [WuYZ11] präsentierten im Unterschied hierzu ein Zwei-Weg-System mit Top-Down-Kommunikation, das auf ‚CIVIL' basiert, einer mobilen Applikation zum Up- und Download geo-referenzierter Daten sowohl für Einsatzkräfte als auch für Bürger. Probleme entstanden infolge der Datenmengen, da die Vielzahl der Bilder eine Überfrachtung der Karten verursachte. Eine vorgeschlagene Lösung bestand aus einem Bilder-Cluster, wobei die Frage offen blieb, wie ein repräsentatives Bild ausgesucht werden soll, das die gesamte Bildergruppe beschreibt [WuYZ11]. Ein gleichartiges Problem entstand bei der Verwendung der Applikation ‚diretto' [EKSS11], welche die Übertragung von Multimediadaten zu einer zuvor gestellten Anfrage erlaubt. Eine Applikation, die auf die Unterstützung von kollaborativem Situationsbewusstsein und Entscheidungsfindung im Spezialfall eines chemischen Industrieunfalls zielt, ist ‚DIADEM' [WMPG11]. Dort kann das Lagezentrum von im Vorfeld ausgewählten Akteuren Bilder anfordern, die automatisch geo-getagged und auf einer digitalen Karte im Lagezentrum angezeigt werden.

Der Großteil existierender Ansätze verfolgt eine Art Push-Mechanismus, bei dem die Informationen in Form von Berichten empfangen werden und der Empfänger keine Möglichkeit hat, seine Informationsbedarfe genauer zu artikulieren, was jedoch im tatsächlichen Einsatz im Rahmen einer Großschadenslage sinnvoll erscheint [RePi09]. Abgesehen von Problemen, die bei Sprachübertragungen auftreten, stellten [BeLa11] fest, dass trotz der erhöhten Awareness die bereits verfügbaren Bilder der örtlichen Einsatzkräfte unregelmäßig genutzt wurden und vereinzelt sogar gar keine Beachtung fanden. Daher sollten Entscheidungsträger die Möglichkeit haben, ihre individuellen Infor-

mationsbedarfe in improvisierter Weise zu benennen. Einige Prototypen [BüMo07] bieten zwar einen Pull-Mechanismus für die Informationsbeschaffung, adressieren jedoch nicht direkt Aktivitäten improvisierter Natur (d.h. spontane Anfragen), und versuchen, verbale Anfragen zu substituieren. Andere Systeme [CLMM10, WMPG11] erlauben Informationsanfragen, die allerdings häufig bloß aus Textnachrichten bestehen, und in denen die Entscheidungsträger keine Option haben, ihre Informationsbedarfe und -formate auf eine feingranulare und dennoch einfache Art und Weise zu spezifizieren.

Unter Berücksichtigung dessen lautet unsere Forschungsfrage: *Wie sollten die im Katastrophenschutz beteiligten Entscheidungsträger ihre Informationsbedarfe gegenüber den unterstellten Einsatzkräften artikulieren?* Und: *Wie kann diese Artikulationsarbeit durch mobile Applikationen unterstützt werden?* Die folgende Arbeit wird sowohl mobile Reporting-Praktiken von Einsatzkräften als auch Möglichkeiten zur Unterstützung dieser durch mobile Geräte und Applikationen untersuchen.

2. Empirische Studie: Reporting-Praktiken von BOS

Die dieser Arbeit zugrunde liegende empirische Studie der Informations- und Kommunikationspraktiken von BOS wurde in zwei geographisch und strukturell verschiedenen Kreisen Nordrhein-Westfalens durchgeführt. Zur Rekonstruktion der Praxis wurden verschiedene qualitative Methoden wie Dokumentenanalysen, 4 Beobachtungen, 26 teilstrukturierte Interviews und 4 Gruppendiskussionen ausgewertet. Ziel der *Dokumentenanalyse* ist die Untersuchung offizieller Arbeitspraktiken mithilfe von gesetzlichen Richtlinien, die von Organisation zur Organisation variierten. Die *Beobachtungen* dienten dazu, Erkenntnisse über die tatsächliche praktische Arbeit zu erhalten. Diese wurden in der Leitstelle während eines normalen Arbeitstages (9h), im Krisenstab und der Einsatzleitung einer Krisenübung (4h) und eines Großereignisses (6h) durchgeführt. Die semi-strukturierten *Interviews* (n=22 je 1–2h; I01-22) dienten der Reflexion der Arbeitspraxis, der Identifikation möglicher IT-Unterstützung und interorganisationaler Informationsbedarfe und wurden mit Einsatzkräften verschiedener Ebenen innerhalb verschiedener Organisationen (v.a. Polizei, Feuerwehr, Behörden) durchgeführt. Sie fokussierten im ersten Teil die Arbeitstätigkeiten in normalen Lagen, im zweiten Teil potentielle Aufgaben im vorgelegten Szenario und im dritten Teil Informations- und Kommunikationstechnologien. Ergänzende Interviews (n=5 je 1h; IM01-05) fokussierten mobile Kollaborationspraktiken. Die *Gruppendiskussionen* (n=4 je 4h) dienten der Erfassung kommunikativer Arbeitspraxen. Bei der Auswertung des vorliegenden Datenmaterials haben wir uns am Grounded Theory-Verfahren [Stra04] orientiert. Dazu kodierten wir die Transkripte offen und teilten die Aussagen in Textbausteine und spätere Kategorien auf.

2.1 Ergebnisse I: Informations- und Kommunikationspraktiken

Großschadenslagen erfordern das Treffen von Entscheidungen in-situ, d.h. in einer direkten Nutzungssituation, welche von der aktuellen Lage abhängt. Daher ist es notwendig, ständig über die Ereignisse auf dem Laufenden gehalten zu werden. Einige der Informationen, die für die Aufgabenbewältigung im Einsatzmanagement gebraucht wer-

den, werden von „offiziellen“ Informationssystemen bereitgestellt, wie beispielsweise Unwetterwarnungen. Gleichzeitig werden diese allerdings um informelle Informationsressourcen in verschiedenen Situationen ergänzt: *„Ich brauche für eine Entscheidung zu einer bestimmten Lage eine sichere Information“* (I05). Dabei muss eine Information *„so detailliert und so genau sein, dass sie wirklich eine tatsächliche Darstellung vor Ort bietet“* (I06). Dazu sind speziell visuelle Eindrücke des Einsatzortes enorm wichtig (I01), denn *„erkläre mal einem, dass da ein ICE mit 300 Verletzen ist. Da denken Sie und ich etwas anderes“* (I09). Aus diesem Grund ist *„die sicherste Information die, die ich selbst gesehen habe“* (I05), wobei die Schwierigkeit besteht, *„dass wir nicht durch das Telefon schauen können. Wir sehen nicht, wie es vor Ort aussieht.“* (I06).

Die Einsatzkräfte vor Ort wissen, wie wichtig es ist, visuelles Datenmaterial in die Leitstelle zu senden, da *„die [Leitstelle] das nur hören und nicht sehen“* und *„wenn man es mal selbst gesehen hat, kann man es besser beurteilen und ein besseres Empfinden der Kräfte vor Ort bekommen“* (I10). Derzeit besteht allerdings keine geeignete Übertragungsmöglichkeit, sondern lediglich die verbale Kommunikation via Funk. Die Menge dieser Funkmeldungen wird in der Leitstelle als zu groß wahrgenommen (I03) und es besteht dauerhaft die Gefahr *„mit Informationen überfrachtet zu werden, sodass man dann letztendlich gar nichts mehr versteht, weil man einfach so viel Input bekommt, dass man die Informationen nicht verarbeiten kann“* (I03).

Auf der anderen Seite klagen die Einsatzkräfte vor Ort über zu wenige Informationen (I07). Welche Informationen zur Entscheidungsfindung notwendig sind, ist sehr subjektiv und individuell und daher im Voraus nicht spezifizierbar (I03). Dennoch sollten Informationen immer bestimmte Attribute besitzen: Da die Entscheidungsträger unserer untersuchten Organisationen auf Basis von Lagekarten arbeiten, müssen Informationen immer (1) einen geographischen Bezug besitzen. Es ist wichtig, dass die Informationen bei der Erstellung *„direkt die entsprechenden GPS-Koordinaten erhalten, sodass man auch weiß, wo sie entstanden sind und nicht rätseln muss.* (IM05). Ein weiteres Bestreben ist es, Informationen mit (2) verschiedenen Datenformaten zu versenden und zu empfangen (IM05). Dabei werden sowohl Fotos als auch Videos als besonders relevant angesehen (IM04), kritisch wird sich aufgrund des Aufwands gegenüber längeren Texteingaben geäußert: *„Sie müssen sich vorstellen, es ist nachts um halb 4, Sie sind gerade wach geworden und schreiben jetzt einen längeren Text auf ihrem mobilen Gerät. Sie wissen, was dabei rauskommt, es ist nicht immer sachdienlich“* (IM03). Zu einer Information muss stets (3) der erstellende Akteur einsehbar sein (IM03). Weiterhin sollte eine Information (4) *„so gestrickt sein, dass ich einen Zeitstempel habe. Also ich muss wissen, wie war eine Situation“* (IM01).

2.2 Ergebnisse II: Artikulation von Informationsbedarfen

Zahlreiche Aktivitäten in der Katastrophenbewältigung zeigen, dass die Lagebeurteilung eine kollaborative Aufgabe ist. Um mit einer unsicheren Umgebung während Einsatzlagen umgehen zu können, ist in der Regel eine Vielzahl von Personen an der Datensammlung und -analyse sowie der Entscheidungsfindung beteiligt. Gewöhnlich (und in der Ausbildung bereits verankert) sind die örtlichen Einsatzkräfte für die Zustellung von

relevanten Informationen verantwortlich, wonach die Informationen als Berichte „gepusht“ werden. Obwohl sich diese Praxis über Jahre bewährt hat, bestehen teilweise Probleme: So wurde in einem unserer Interviews deutlich, dass die örtlichen Kräfte, insbesondere die freiwilligen Kräfte, oftmals nicht wissen, welche Informationen sie übermitteln sollen, oder ausgehende Berichte sehr niedrig priorisieren (IM01, IM02). Aus diesem Grund ist die aktuelle Praxis so gestaltet, *„dass man einem Abschnittsführer in einer Situation sagt, wo der erste Trouble mal weg ist: Jetzt schick uns mal ein Bild! Wirklich auf Anforderung“* (IM01). Durch das gezielte Anfordern kann man verhindern, *“dass jeder irgendwelche Bilder macht und die rückwärtig einspeist, ohne dass man genau weiß, worum es da geht“* (IM05). Während einer solchen dynamischen Informationsanforderung sollte das zurückgelieferte Format (Bild, etc.) spezifiziert werden können (IM04). Bei einer solchen Anforderung muss der Mitarbeiter *„in der Leitstelle sehen, dass eine Person an einer Stelle mit seinem Tablet herum rennt und ihm sagen: Mach mir doch mal ein Foto!“* (IM02). Derzeit funktioniert die Standortbestimmung nur verbal über Funk: *“Im besten Fall nehme ich mir ein Funkgerät und spreche ihn an: Wo bist du? Und dann sagt er: Ich bin hier oder dort.“* (IM05). Idealerweise könnte man zusätzlich zum Standort die aktuelle Verfügbarkeit sehen. (I03). Diese Artikulationen verbraucht zum jetzigen Stand sehr viel Zeit. Daher merkten die Teilnehmer an, dass es – auf dem Standort basierend – eine Option zur Fern-Navigation der Einheiten geben sollte: *“Wenn ich das* [Standorte der Einsatzkräfte] *auf der Karte sehe und hätte beispielsweise noch Hydrantenpläne, dann könnte ich sagen: Geh mal noch fünf Meter weiter, da müsste der nächste Hydrant sein“* (IM02). Aufgrund der hohen Dynamik und der Improvisation innerhalb der Einsatztruppen können den Personen spontan neue Rollen zugeteilt werden. Daher sollten Anfragen immer Rollen und niemals Personen zugeordnet werden (IM01, IM02).

3. Mobiles semi-strukturiertes Reporting-Konzept

Die Entscheidungsfindung bei der Krisenbewältigung hängt von den inkrementellen, schriftlichen oder verbalen Vor-Ort-Meldungen ab. Die empirische Analyse der gegenwärtigen Praxis der BOS zeigt eine Notwendigkeit für improvisierte Maßnahmen, um geeignete Berichte zu erhalten. Die „Informationsproduzenten“ vor Ort sind demnach nicht immer in der Lage, die Bedürfnisse ihrer Kollegen in den Leitstellen zu antizipieren, sodass die resultierende Informationsüberlastung oder der Mangel an Informationen Entscheidungen negativ beeinflussen kann. Der Kreislauf von semi-strukturierten Informationsanfragen und -berichten (Abbildung 2) visualisiert das auf der Empirie basierende Konzept.

Wenn die schriftlichen oder mündlichen Vor-Ort-Berichte nicht die Bedürfnisse der Leitstelle befriedigen, sollte diese die Möglichkeit besitzen, ihren Informationsbedarf aktiv zu artikulieren. Ein solches dynamisches Anfordern wird aktuell lediglich via Funk unterstützt, wobei die Einheiten vor Ort über einen zu großen Interpretationsspielraum klagen: *„Der Einsatzleiter möchte etwas erledigt haben und wir müssen dann verstehen, was er genau meint“* (I07). Daher kann ein Mechanismus, welcher semi-strukturierte Informationsanfragen und wenig Platz für Interpretationen erlaubt, die Kollaboration verbessern. Aufgrund der Tatsache, dass Informationsbedarfe von Individuum zu Indivi-

duum variieren, sollte eine Spezifikation von Berichten für jeden Benutzer individuell möglich sein, wobei allerdings drei Kontextinformationen immer automatisch erfasst werden müssen, um eine Information eindeutig zu identifizieren (IM03): Der Standort, die Quelle und die Zeit, welche alle einsehbar sein müssen. Beim Anfordern und Berichten von Informationen muss die vordefinierte Weisungs- und Meldehierarchie berücksichtigt werden. Beispielsweise darf ein Unterabschnittsführer keine Informationen direkt an die Leitstelle senden, da sonst der Abschnittsführer übersprungen würde. Der Adressat einer Informationsanfrage kann zum einen über seinen Standort, zum anderen über seine Rolle bestimmt werden: Eine standortbasierte Anfrage gibt einer Einheit einen Überblick über alle Standorte seiner direkt untergebenen Kräfte, eine rollenbasierte Anfrage ermöglicht die Auswahl anhand seiner Rolle (beispielsweise Abschnittsführer Bereich A). Ein solcher Mechanismus zum Anfragen und Berichten von Informationen sollte sowohl auf Smartphones als auch auf Tablets zur Anwendung kommen können, um einen umfassenden Einsatz zu ermöglichen (I06). Beide Geräteklassen finden bereits Anwendung zur Betrachtung satellitenbasierter Lagebilder oder zur Navigation (IM04).

Abbildung 2: Kreislauf semi-strukturierter Informationsanfragen und -berichte

Um der Organisationsstruktur inklusive deren Weisungs- und Meldewesen gerecht zu werden, müssen folgende Rechte gewährleistet sein:

- Das *Anfordern von Informationen* ermöglicht es einer Einheit, eine benötige Information fein zu spezifizieren und geeignet zu artikulieren. Dabei sollte das Übertragen eines Zielortes zur Remote-Navigation der Einsatzkraft vor Ort sowie das Priorisieren für eine angemessene Beurteilung der Aufgabendringlichkeit möglich sein.

- Das *unabhängige Berichten von Informationen* ermöglicht autorisierten Einheiten, Berichte direkt zu senden, ohne eine vorherige Anfrage (IM06).
- Das *Anfrage-basierte Senden von Informationen* schränkt eine Einheit in der Form ein, Informationen nur als Antwort auf eine frühere Anfrage zu senden. Diese Restriktion soll die Informationsüberlastung für Entscheidungsträger vermeiden.

4. Implementierung der mobilen Reporting-App „MoRep“

Um das Konzept zu überprüfen und dessen Wirkung zu erforschen, wurde auf Basis einer vorherigen Systemversion [LuRP13a] die mobile Anwendung ‚MoRep‘ als Teil einer mobilen Applikation zur Kollaboration zwischen Mitarbeitern vor Ort und der Leitstelle [ReLP14, ReRi13] implementiert. Durch die Verwendung von Android 4 werden sowohl Smartphones als auch Tablets adressiert. Das technische Konzept basiert auf einer REST-Architektur, in welcher die Services als HTTP-Servlets umgesetzt wurden. Durch den Einsatz von Google Cloud Messaging sind innovative Benachrichtigungsmechanismen implementiert, welche eine parallele Nutzung zum BOS-Funk bieten. Im Folgenden wird die Anwendung kurz vorgestellt:

Start up: Nach der Authentifizierung erhält der Nutzer aktuelle Informationen über seine Rolle, auf deren Basis der Main Screen aufgebaut wird. Ein Nutzer mit allen Rechten kann Berichte sehen, Berichte von untergeordnete Einheiten anfordern, Berichte an übergeordnete zum einen auf Anforderung, zum anderen aber eigenständig (Freier Bericht) senden (Abbildung 3)

Abbildung 3: Main Screen

Berichte anfragen: Die Bestimmung eines Adressaten von Berichtsanfragen ist sowohl standort- als auch rollenbasiert möglich. Bei erster Option kann der Nutzer die Standorte untergeordneter Einheiten abfragen und auf einer Karte anzeigen lassen (Abbildung 4). Nach der Auswahl einer Einheit öffnet sich ein Formular, auf dem der Nutzer spezielle Attribute einer Anfrage, wie die Beschreibung oder das gewünschte Format, angeben kann. Der Nutzer hat die Möglichkeit, einen Zielstandort für eine entfernte Navigation (zur Definition eines Wunschstandortes oder Ereignisstandortes und Navigation) zu übermitteln und die Anfrage abzusenden.

Anfragen sehen / Bericht erstellen: Bei Eingehen einer Anfrage macht sich die Anwendung akustisch und visuell bemerkbar. In einer Anfragenübersicht werden individuell alle offenen Anfragen sortiert dargestellt. Eine Anfrage kann direkt mit einem Bericht beantwortet werden, in welchem die Textfelder und gegebenenfalls ein Zielstandort bereits durch die anfragende Einheit vordefiniert sind. Durch das vordefinierte Format wird direkt die entsprechende Aktion (z.B. *Starten der Kamera bei Foto*) ausgeführt und es kann optional ein Text eingegeben bzw. eingesprochen werden (Abbildung 5). Sollte ein Bericht nicht als direkte Antwort auf eine Anfrage erstellt werden, müssen die Textfelder eigenständig eingegeben werden.

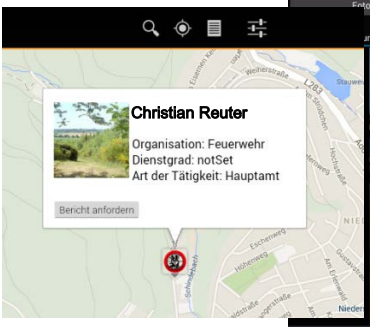


Abbildung 4: Standort-basierte Anfrage

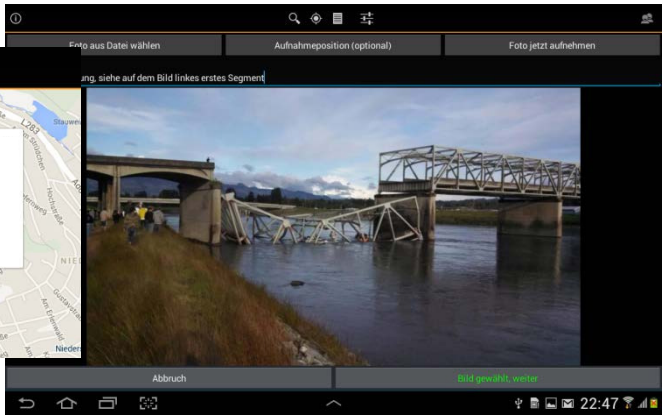


Abbildung 5: Bericht absenden

Berichte sehen: Das wesentliche Element ist eine Karte, auf welcher die zuvor erstellen Reports mit dem entsprechenden Icon angezeigt werden. Der Nutzer kann entweder alle oder nur die von ihm angeforderten Berichte sehen (Abbildung 6). In den Details sind die Einheit, welche den Bericht erstellt hat, sowie weitere Details, wie z.B. die Organisation oder die Erstellungszeit, einsehbar (Abbildung 7). Diese beiden Attribute charakterisieren mit dem Geostandort eine Information eindeutig (IM05). Es lässt sich direkt der Inhalt betrachten, welcher sofort an den eigenen Vorgesetzten übermittelt werden kann. Zusätzlich kann der Standort der Einheit sowie der des Ereignisses eingesehen werden, da beides nicht zwangsläufig übereinstimmen muss (Abbildung 8).

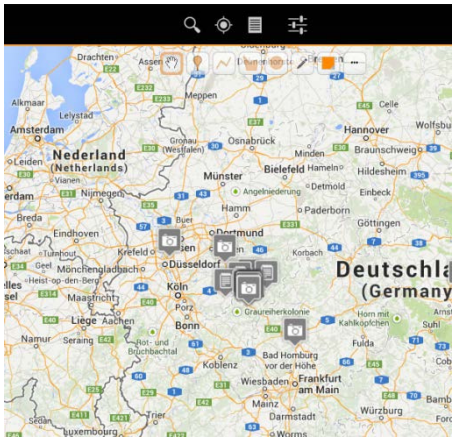


Abbildung 6: Berichte sehen

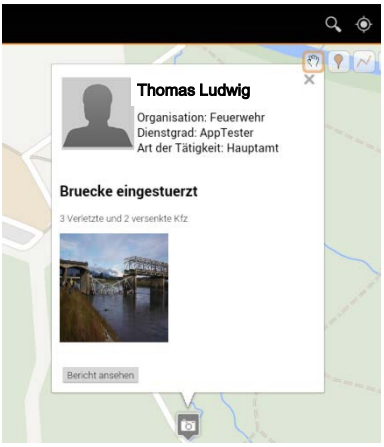


Abbildung 7: Berichtdetails anzeigen

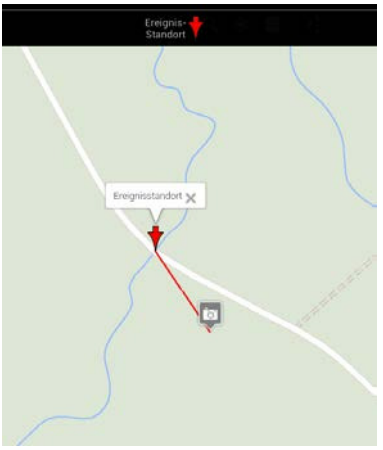


Abbildung 8: Ansichtskoordinaten

5. Evaluation von MoRep

Trotz vollständiger Implementierung unseres Systems hielten uns IT Sicherheitsregularien einzelner Organisationen von einer langfristigen tatsächlichen Nutzungsevaluation ab. Um dennoch eine realitätsnahe Überprüfung des Konzeptes sicherzustellen, evaluierten wir die Nutzung durch die Polizei und Feuerwehr in einem Szenario-basierten Walk-Through, gefolgt von Interviews. Neben professionellen Kräften wurden auch speziell freiwillige Einsatzkräfte einbezogen, da auch diese potenzielle Endbenutzer eines solches Systems wären (IM01, IM02). Die Evaluation umfasste insgesamt 11 Personen (IM06 bis IM16) und dauerte im Durchschnitt je 45 Minuten. Durch die Selektion der einzelnen Interviewpartner konnten Eindrücke von Einsatzkräften auf allen Ebenen der Hierarchie erhoben werden. Bei jeder Evaluation wurde MoRep funktional vorstellt und demonstriert, wie die in der Empirie genannten Einsatzsituationen geeignet unterstützt werden könnten. Die Sitzung war interaktiv gestaltet, sodass die Nutzer direkt die Anwendung ausprobieren konnten, wobei Anmerkungen durch *Thinking Aloud* [Niel93] festgehalten wurden. Nach der Demonstration folgten teilstrukturierte Interviews mit verschiedenen Leitfragen: *Welche Folgen hat die Nutzung von semi-strukturierten Anfragen im Einsatz? Unter welchen Bedingungen kann das Konzept die aktuelle Arbeitspraxis unterstützen? Was sind Nutzungslimitationen?* Durch die Evaluation waren wir in der Lage, die Auswirkungen mobiler semi-strukturierter Anfragen auf die Improvisationsarbeit von Entscheidungsträger und Vor-Ort-Einheiten abzuleiten.

Ergänzung der Artikulationsarbeit: Das Konzept semi-strukturierter mobiler Berichte als Konsequenz dynamischer, semi-strukturierter Informationsanfragen kann größere Einsatzlagen nicht über den gesamten Zeitrahmen abdecken, allerdings kann es sehr einfach für *„Basisinformationen und ein gemeinsames Verständnis der Situation“* (IM08) genutzt werden. Es bietet eine zusätzliche Möglichkeit der Artikulation, welche die Berichte bzw. Meldungen um visuelle Daten anreichert (IM09). Die Feinspezifizierung der Anforderung und automatisierte Absetzung wurde im Falle unzureichender Berichte als

sehr nützlich angesehen, da Entscheidungsträger die informelle Option besitzen, von routinierten Strukturen abzuweichen, wobei die Antwort seitens der Vor-Ort-Einheiten die Pflicht ist (IM04). Informationsanfragen auf Basis des Einheitenstandortes oder deren Rolle sind wichtige Faktoren, um Awareness zwischen den verteilten Einheiten und eine geeignete Situationsübersicht zu ermöglichen, was sich stressreduzierend auswirkt (IM07, IM08). Die schnelle Informationsweiterleitung an den Vorgesetzen innerhalb einer Lage ist unerlässlich und würde eine stark verbesserte Arbeitspraxis darstellen (IM06, IM08).

Erweiterte Situation Awareness: Der Umgang mit semi-strukturierten Informationsanfragen als eines der Kernelemente des Konzeptes hat zwei unterstützende Dimensionen: Zum einen werden lokale freiwillige Einsatzkräfte unterstützt, welche das Wissen über den Einsatzstandort besitzen, notwendigerweise aber nicht die Erfahrung über die Relevanz einer Information für die Entscheidungsträger zu entscheiden (IM06). Semi-strukturierte Informationsanfragen überbrücken dies und fördern Trainingseffekte. Zum anderen unterstützen semi-strukturierte Informationsanfragen professionelle Einheiten anderer Regionen, welche die Relevanz von Informationen besser einschätzen können, allerdings – speziell in größeren Lagen – oftmals kein Wissen über den Einsatzort besitzen. Der Leiter der Leitstelle erklärte, dass die Polizei in Kürze eine Vielzahl neuer Kräfte erhalten wird, die nicht mit der Region vertraut sind, weshalb das Übermitteln von Zielstandorten zur Schaffung der Situation Awareness unerlässlich sei (IM06).

Berücksichtigung organisatorischer Besonderheiten: Die verschiedenen Führungsstrukturen der Polizei ("von hinten" aus der Leitstelle) und Feuerwehr („von vorne" am Einsatzort) haben einen großen Einfluss auf eine Etablierung von MoRep in den Arbeitspraxen. Bei der Polizei liegt die Einsatzverantwortung bei der Leitstelle. Sie benutzt Software, welche alle Einsatzkräfte und deren dynamische Rollenzuweisungen verwaltet, wodurch sie eine ständige Übersicht über aktuelle Weisungs- und Meldewege besitzt (IM06). Dadurch kann eine automatisierte Verbindung zwischen MoRep und dem Leistellensystem über aktuelle Rollenzuweisungen und Meldewege hergestellt werden. Im Gegensatz dazu besitzt die Feuerwehr keine vergleichbare Software, denn die gesamte Verantwortung eines Einsatzes liegt beim Einsatzleiter vor Ort. Die Leitstelle wirkt unterstützend, wobei sie das Problem besitzt *„wenn ich ihm neue Koordinaten zuweise, dann ist das ein Prozess, wo ich lenkend einwirke, dann muss man sich überlegen, ob das sinnvoll ist oder auf welcher Stufe man das darf"* (IM11). Deshalb wäre der Hauptnutzer hier der Einsatzleiter vor Ort, um die Kommunikation zwischen ihm und den sonstigen, verteilten Einsatzkräften zu unterstützen, denn auch er befindet sich nicht immer direkt am Schadensort (IM11).

Anreicherung der Einsatznachbesprechungen: Nach einem Einsatz sollten die automatisch gespeicherten Anfragen und übertragenen Informationen zur Dokumentation genutzt werden (IM03). Diese Dokumentation kann als Basis für Nachbesprechungen vergangener Einsätze dienen: *„Aktuell haben wir nachher eine Nachbesprechung anhand von Internetvideos, Fotos von Fotografen und dem Pressemensch und das Ganze wird zusammengefügt"* (IM11), denn neben den geschrieben Berichten existiert keine vor allem multimediale Dokumentation zur Nachbereitung und Trainingszwecken.

Vordefinierte Kommunikationspfade vs. Improvisationsarbeit: Obwohl vordefinierte Kommunikationspfade als nützlich gelten, bestehen Zweifel, ob die Einsatzkräfte vor Ort sich daran halten (IM07). Durch die vordefinierte Weisungs- und Befehlshierarchie ist aktuell das Konzept zu statisch, um die gesamte Improvisationsarbeit während Einsätzen abzudecken. Dennoch waren sich die Teilnehmer einig, dass mobile Anwendungen wie MoRep eine erweiterte Kommunikationsmöglichkeit bilden und dass bei starker Improvisationsarbeit die Option eines Anrufes weiterhin besteht (IM09).

6. Fazit

Im Katastrophenschutz ist das Einholen notwendiger Informationen zur Schaffung einer Situationsübersicht und Entscheidungsfindung der Entscheidungsträger von hoher Relevanz. Das Zusammenspiel zwischen Leitstelle und Vor-Ort-Kräften ist dabei eine der wichtigsten Informationsketten und in Dienstvorschriften geregelt. In manchen Lagen kann das daran geknüpfte, routinierte Handeln nicht alle Informationsbedarfe decken, wodurch Einsatzkräfte zur Improvisation und hoher Artikulationsarbeit gezwungen sind.

Mit unserer Design-Fallstudie haben wir die Notwendigkeiten von Artikulationsarbeit in Bezug auf die Erfassung von Informationen untersucht und ein semi-strukturierte Multimedia-Berichte ermöglichendes Interaktionskonzept konzipiert und evaluiert. In unserer empirischen Studie der aktuellen Arbeitspraxen von BOS im Hinblick auf die Zusammenarbeit bei der Entscheidungsfindung haben wir festgestellt, dass die Spontanität und Divergenz der Informationsbedarfe große Herausforderungen an eine zielführende Kommunikation stellt. Dabei zeigt die existierende Praxis, dass um geeigneten Berichten und einer Zuverlässigkeit von Informationen gerecht zu werden, eine Reihe von Rahmenbedingungen adressiert werden müssen, um technische Unterstützungsmöglichkeiten zu schaffen:

(1) *Gezieltes Anfragen*: Die fehlenden Informationen in einer Entscheidungssituation sind oft sehr spezifisch für einen Ort, eine kritische Infrastruktur oder anderen situativen Aspekt. Diese Besonderheiten des Informationsbedarfs müssen artikuliert und verstanden werden.
(2) *Zuverlässige Berichte:* Entscheidungen der Leitstellen können große Auswirkungen und rechtliche Folgen haben. Deshalb benötigen Mitarbeiter qualitativ hochwertige Informationen, welche nicht nur im Bereich der technischen Informationsqualität betrachtet werden müssen, sondern auch im Hinblick auf Vertrauen, welches auf der fachlichen Kompetenz des explizit angefragten Informationserstellers basiert.
(3) *Dokumentiertes Vorgehen:* Sowohl die Anforderungen als auch Berichte von Informationen müssen dokumentiert werden, um zum einen die Nachbereitung zu ermöglichen und Trainingsmaterial zu erstellen.

Danksagung: Dieser Beitrag ist eine überarbeitete und erweiterte Version von [LuRP13b] und [LuRP13a]. Die Forschungsarbeiten wurden im Rahmen des Verbundprojektes InfoStrom durch das Bundesministerium für Bildung und Forschung (Fö.-Kz. 13N10712) unterstützt.

Literaturverzeichnis

[BeLa11] Bergstrand, F.; Landgren, J.: Visual reporting in time-critical work : Exploring video use in emergency response Chalmers University of Technology. In: *Proc. MobileHCI*. Stockholm, 2011, pp. 415–424

[BhLJ09] Bharosa, N.; Lee, J.; Janssen, M.: Challenges and obstacles in sharing and coordinating information during multi-agency disaster response: Propositions from field exercises. In: *Information Systems Frontiers* vol. 12 (2009), Nr. 1, pp. 49–65

[BMPM05] Betts, B. J.; Maah, R. W.; Papasin, R.; Del Mundo, R.; Mcintosh, D. M.; Jorgensen, C.: I*mproving situational awareness for first responders via mobile computing*. Moffett Field, California, 2005

[BüMo07] Büscher, M.; Mogensen, P.: Designing for material practices of coordinating emergency teamwork. In: *Proc. ISCRAM*. Delft, 2007

[CLMM10] Catarci, T.; Leoni, M.; Marrella, A.; Mecella, M.; Bortenschlager, M.; Steinmann, R.: The WORKPAD Project Experience: Improving the Disaster Response through Process Management and Geo Collaboration. In: *Proc. ISCRAM*. Seattle, USA, 2010, pp. 1–10

[EKSS11] Erb, B.; Kaufmann, S.; Schlecht, T.; Schaub, F.; Weber, M.: diretto: A Toolkit for Distributed Reporting and Collaboration. In: *Mensch & Computer 2011*. Chemnitz, Germany: Oldenbourg, 2011, pp. 151–160

[GOPC06] Guerrero, L.A.; Ochoa, S. F.; Pino, J. A.; Collazos, C. A.: Selecting Computing Devices to Support Mobile Collaboration. In: *Group Decision and Negotiation* vol. 15 (2006), Nr. 3, pp. 243–271

[LuRP13a] Ludwig, T.; Reuter, C.; Pipek, V.: What You See Is What I Need: Mobile Reporting Practices in Emergencies. In: *Proc. ECSCW*. Paphos, Cyrus : Springer, 2013, pp. 181–206

[LuRP13b] Ludwig, T.; Reuter, C.; Pipek, V: Mobiler Reporting-Mechanismus für örtlich verteilte Einsatzkräfte. In: Mensch & Computer 2013. Bremen: Oldenbourg, 2013, pp. 317–320

[Niel93] Nielsen, J.: *Usability Engineering*. San Francisco: Morgan Kaufmann, 1993

[NiSt10] Nilsson, E. G.; Stolen, K.: Ad Hoc Networks and Mobile Devices in Emergency Response – a Perfect Match? (Invited Paper). In: *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering* vol. 49. Victoria, BC, Canada (2010), pp. 17–33

[Pato03] Paton, D.: Stress in disaster response: a risk management approach. In: *Disaster Prevention and Management* vol. 12 (2003), Nr. 3, pp. 203–209

[ReLP14] Reuter, C.; Ludwig, T.; Pipek, V.: Ad Hoc Participation in Situation Assessment: Mobile Collaboration Practices in Emergencies. In: *TOCHI* (to appear), ACM (2014)

[ReRi13] Reuter, C.; Ritzkatis, M.: Unterstützung mobiler Geo-Kollaboration zur Lageeinschätzung von Feuerwehr und Polizei. In: *Proc. Wirtschaftsinformatik*. Leipzig, Germany, 2013, pp. 1877–1891

[RePi09] Reuter, C.; Pipek, V.: Krisenkommunikation trainieren: Ein webgestützter Ansatz. In: *Mensch & Computer 2009* Berlin: Oldenbourg-Verlag, 2009, pp. 93–102

[SiAb09] Singh, G.; Ableiter, D.: Twiddlenet: Smartphones as personal content servers for first responders. In: Löffler, J.; Klann, M. (eds.): *Mobile Response*. vol. 5424, 2009, pp. 130–137

[Stra04] Strauss, A.: Methodologische Grundlagen der Grounded Theory. In: Strübing, J.; Schnettler, B. (eds.): *Methodologie interpretativer Sozialforschung: Klassische Grundlagentexte*. Konstanz : UVK, 2004, pp. 427–453

[WMPG11] Winterboer, A.; Martnes, M. A.; Pavlin, G.; Groen, F. C. A.; Evers, V.: DIADEM: a system for collaborative environmental monitoring. In: *Proc. CSCW*. Hangzhou, China, 2011, pp. 589–590

[WuYZ11] Wu, A.; Yan, X.; Zhang, X.: Geo-tagged Mobile Photo Sharing in Collaborative Emergency. In: *Proc. VINCI*. Hong Kong, China, 2011, pp. 1–8

Framework zur Simulation von realistischen Einsatzszenarien

Silvia Krug, Maik Debes, Jochen Seitz

Fachgebiet Kommunikationsnetze
Technische Universität Ilmenau
Postfach 100565
98684 Ilmenau
silvia.krug@tu-ilmenau.de
maik.debes@tu-ilmenau.de
jochen.seitz@tu-ilmenau.de

Abstract:
Für die Analyse von Kommunikationssystemen für Rettungskräfte sind aussagekräftige Bewegungsszenarien erforderlich, die das gesamte Einsatzspektrum der Retter abdecken. Mithilfe dieser Bewegungsszenarien kann auf Basis von Simulationen evaluiert werden, welche Anforderungen an die Kommunikation bestehen. Diese Simulationen sind die Voraussetzung für die Entwicklung leistungsfähiger Kommunikationsprotokolle. In diesem Beitrag wird ein Framework zur Erzeugung und Visualisierung generischer Einsatzszenarien anhand von Bewegungsmustern und generischen Einsatzabläufen vorgestellt. Mit Hilfe des Frameworks können Traces für die Bewegung der Rettungskräfte in verschiedenen Szenarien erzeugt werden. Durch die Visualisierung der Traces in einer Simulationsumgebung erlaubt dies zusätzlich zur reinen Evaluation der Kommunikationsaspekte auch die Analyse auf Planungsebene.

1 Einleitung

Nach Naturkatastrophen aber auch bei anderen Rettungseinsätzen ist eine effiziente Koordinierung der Einsatzkräfte notwendig, um möglichst viele Menschen zu retten und den Schaden zu begrenzen. Die Kommunikationssysteme der beteiligten Rettungskräfte spielen für diese Aufgabe eine zentrale Rolle, da die Kommunikation zwischen den einzelnen Rettern die Basis für einen effektiven Einsatz darstellt. Wachsende Anforderungen an die zu übertragende Datenmenge zum Beispiel durch Videoübertragungen erfordern neue Kommunikationssysteme, die idealerweise unabhängig von einer eventuell beschädigten Infrastruktur arbeiten können.

Um solche Systeme entwerfen zu können, müssen Einsatzszenarien analysiert werden, die ein möglichst breites Einsatzspektrum realistisch abbilden können. Ziel der Analyse ist dann die Identifikation von Randbedingungen, wie zum Beispiel der erforderlichen Reichweite oder der verfügbaren Anzahl an Knoten, die Kommunikationsgeräte darstellen. Simulationen stellen ein geeignetes Werkzeug für diese Analyse dar. In Bezug auf

die Einsatzkommunikation ist insbesondere die Simulation der Bewegung von Einsatzkräften im Gelände interessant, weil sie Rückschlüsse auf zentrale Fragen zur Position und der Entfernung zwischen den Einsatzkräften ermöglicht. Die Position kann auch für die entsprechenden mobilen Kommunikationsgeräte angenommen werden. Damit die entsprechenden Szenarien realitätsnah sind, ist ein gewisses domänenspezifisches Wissen erforderlich. Dieses kann entweder durch Nachbildung realer Einsätze oder Übungen gewährleistet werden, wenn zum Beispiel GPS-Daten für Rettungskräfte erfasst wurden, oder indem mögliche Schadenslagen, wie Hochwasser oder Waldbrände, mit entsprechenden Experten evaluiert werden.

Für eine Analyse der Kommunikationswege ist in jedem Fall nur relevant, wie viele Retter mit welcher Kommunikationsausrüstung im Szenario unterwegs sind, wo sie sich zu bestimmten Zeitpunkten aufhalten und wie schnell sie sich bewegen. Je nach Detailierungsgrad ermöglichen diese Szenarien insbesondere bei entsprechender Visualisierung auch Rückschlüsse auf einsatztaktische Parameter, die den Rettungskräften selbst ermöglichen, bestimmte Situationen für Schulungszwecke und zur Gefahrenabschätzung zu verwenden.

In diesem Beitrag wird ein Framework zur Erzeugung und Simulation realitätsnaher Einsatzszenarien vorgestellt, das die flexible Konfiguration verschiedener Szenarien ermöglicht. Die so erzeugten Szenarien können anschließend in verschiedenen Simulationsumgebungen genutzt werden oder direkt mit der im Framework enthaltenen Simulationsumgebung evaluiert werden. Diese enthält auch eine Möglichkeit zur Visualisierung der entsprechenden Szenarien.

2 Simulation von Einsatzszenarien

2.1 Netzwerkevaluation

Für die Evaluation verschiedener Kommunikationsprotokolle in Einsatzszenarien werden in der Regel zwei Möglichkeiten verfolgt: Die Verwendung vorher aufgezeichneter GPS-Traces der Rettungskräfte oder die Generierung der Mobilität anhand von mathematischen Bewegungsmodellen.

Traces ermöglichen die realistischste Abbildung der Bewegung der entsprechenden Knoten. Dabei ist jedoch zu beachten, dass eine Aufzeichnung immer nur eine konkrete Situation zeigt. Deshalb sind für grundlegende Analysen sehr viele Daten erforderlich. Nachteilig ist ebenfalls, dass für komplette Szenarien alle Rettungskräfte mit geeigneten GPS-Geräten ausgestattet sein müssen, um entsprechende Traces aufzeichnen zu können.

Eine Alternative bilden verschiedene Bewegungsmodelle, mit deren Hilfe Traces für eine beliebige Anzahl Knoten erzeugt werden können. Diese Modelle werden zum Teil als separate Werkzeuge bereitgestellt oder sind in die normalen Netzwerksimulationsumgebungen integriert. Allerdings werden häufig Modelle verwendet, die rein zufällige Bewegungen der Knoten im angegebenen Gebiet erzeugen. Ein Beispiel dafür ist das Random-Waypoint Modell [CBD02]. Diese Modelle sind für Einsatzszenarien nicht geeignet, weil die Bewegungen der Rettungskräfte nicht zufällig erfolgen. Im Gegensatz dazu ergeben

sich diese Bewegungen aus taktischen Erfordernissen und den entsprechenden Dienstvorschriften.

Ein anderes Modell mit Bezug auf Katastrophenfälle beschreibt die Bewegung anhand von bestimmten Punkten, die auf Knoten anziehend oder abstoßend wirken können [NHIK07]. Die Knoten entfernen bzw. nähern sich diesen Punkten mit entsprechend höherer Wahrscheinlichkeit. Auch dieses Modell ist für die Modellierung der Bewegungen von Rettungskräften nicht ausreichend, ermöglicht aber eine Darstellung der Bewegungen von möglichen Opfern bzw. anderen betroffenen Personen.

Gezielt für Katastrophenfälle wurden Modelle entwickelt, die zum Beispiel ein spezielles Szenario mit sehr vielen Verletzen [AGPG+07] oder die Versorgung verschiedener Notunterkünfte nach einer Katastrophe [UNAK09] beschreiben. Diese sehr spezifischen Modelle können das jeweilige Szenario realistisch abdecken, ermöglichen jedoch keine Modellierung anderer Einsatzszenarien.

2.2 Simulation zur Analyse von Einsätzen

Eine andere Klasse von Simulationen wird für Planungszwecke im Katastrophenschutz bzw. Schulungszwecke in der Ausbildung von Rettungskräften benötigt. In diesem Bereich liegt der Fokus weniger auf der Bewegung aller Knoten und daraus resultierenden Kommunikationsoptionen. Vielmehr ist die realistische Nachbildung einzelner Situationen entscheidend, um Rettungskräfte gezielt auf Einsätze vorzubereiten bzw. eine Abschätzung der möglichen Gefahrensituation zu erreichen.

Zu Trainingszwecken existieren verschiedene Simulationswerkzeuge, die vordergründig die realistische Nachbildung verschiedener Einsatzsituationen aus Sicht einzelner Einsatzkräfte darstellen. Ein Beispiel dafür ist das von der Universität Kassel entwickelte KATIE-System [Fac], das virtuelle dreidimensionale Welten erstellt, in denen der Retter einen Avatar steuert, um einen Einsatz nachzuspielen und der dabei konkrete Aufgaben im jeweiligen Szenario erfüllen muss. Das Konzept wurde von Computerspielen übernommen. Vergleichbare Systeme wurden auch in anderen Ländern entwickelt [CST+13, MSZ+12]. Die entsprechenden Systeme ermöglichen in der Regel auch die Integration eigener Modelle, um die jeweils lokalen Bedingungen nachzustellen. Diese Ansätze sind sehr detailliert und erlauben eine direkte Sicht auf einzelne Akteure. Allerdings fehlt eine Möglichkeit, die strategischen Aspekte eines Einsatzes bzw. eine globale Sicht auf den Ablauf darzustellen, die insbesondere für die Planung und Koordination solcher Einsätze wichtig ist.

Neben diesen sehr spezifischen Modellen existieren verschiedene Ansätze, um die Ausbreitung von Feuern selbst zu simulieren. Diese Modelle werden teilweise in die oben genannten 3D-Simulationen eingebunden. Sie ermöglichen die Simulation von potentiellen Feuerereignissen und somit eine Abschätzung von der Ausbreitung eines Feuers. Dies ist insbesondere für die Evaluation von größeren Wald- bzw. Buschfeuern relevant [Fin04]. Ohne die entsprechende Berücksichtigung lokal verfügbarer Einsatzkräfte, ist eine geeignete Vorhersage verschiedener Optionen für die Brandbekämpfung jedoch nur schwer möglich.

3 Simulationsframework

Das in diesem Beitrag vorgestellte modulare Framework zur Erzeugung und Simulation von generischen Einsatzszenarien ermöglicht einen Überblick über alle Rettungskräfte und deren Zusammenarbeit in einem größeren Einsatz und eignet sich daher neben der Analyse von Kommunikationsanforderungen auch für die Evaluation ganzer Einsätze auf Planungsebene. Dazu können beliebige Einsatzarten modelliert werden.

3.1 Anforderungen

Da das vorgestellte Framework sowohl für die Analyse und Evaluation von Kommunikationsaspekten als auch für die Planung und Evaluation von Rettungseinsätzen verwendet werden kann, werden im Folgenden die Anforderungen dieser beiden Verwendungsmöglichkeiten an ein Framework vorgestellt. Wie sich zeigt überlappen sich einige Anforderungen, wodurch die Verwendung eines einzigen Frameworks sinnvoll ist.

3.1.1 Kommunikationsanalyse

Hinsichtlich einer Simulation zur Analyse vorhandener Kommunikationsmöglichkeiten beziehungsweise von Problemen, die bei der Einsatzkommunikation auftreten, werden Informationen zur verwendeten Kommunikationstechnik und die Position der Geräte benötigt. Werden mobile Geräte betrachtet, ist auch der zeitliche Ablauf der Bewegung interessant. Aus diesem Grund ist es wichtig die Bewegung der Konten zu erfassen und in Kombination mit Eigenschaften der gewählten Kommunikationstechnik, wie z.B. der Reichweite und der verfügbaren Bandbreite, zu evaluieren.

Mögliche Analysen sind dann:

- Verzögerungszeiten
- Netzabdeckung
- Mögliche Partitionierungen

Dazu sollte die Simulationsumgebung verschiedene Bewegungsszenarien erzeugen können oder eine Möglichkeit bieten, entsprechende externe Trace-Daten zu importieren. Neben der Erzeugung einzelner Szenarien, ist auch eine Möglichkeit einzelne Parameter zu variieren wichtig, um verschiedene Auswirkungen der Mobilität zu evaluieren. Dabei ist es auch wichtig nicht nur ein Szenario oder eine Gruppe ähnlicher Szenarien zu evaluieren, damit das zu entwickelnde System bzw. Protokoll nicht für dieses Szenario optimiert wird, in anderen Fällen aber nicht anwendbar ist. Gerade für die Kommunikation von Rettungskräften ist die wichtig, da Situationen mit wenigen Knoten in einem großen Gebiet (Hochwasser) und andererseits Situationen mit sehr vielen Knoten in einem kleinen Gebiet (Gebäudebrand) auftreten können und ein Kommunikationssystem in beiden Fällen funktionieren muss.

3.1.2 Einsatzevaluation

Sollen reale Einsätze oder entsprechende Übungen in einer Nachbesprechung analysiert werden, ist es interessant, die Positionen der beteiligten Rettungskräfte dynamisch nachzuvollziehen. Dazu können neben während des Einsatzes per GPS erfassten Bewegungsdaten in Form von Traces auch generisch erzeugte Traces verwendet werden, wenn das entsprechende Framework eine realistische Nachbildung der Einsatzbedingungen inklusive Kartenmaterial ermöglicht.

Die dynamische Darstellung des Ablaufes ist auch der wichtigste Vorteil eines solchen Frameworks im Vergleich zu einer rein text-basierten Analyse oder der Verwendung einzelner Momentaufnahmen, weil zeitliche Abläufe deutlich werden und die Auswirkungen einzelner Entscheidungen analysiert werden können. Durch die Integration weiterer zufällige Ereignisse ist es zudem möglich alternative Entscheidungen und ihre Auswirkungen auf den Einsatzverlauf zu evaluieren. Diese Zufallselemente sind daher ein wichtiger Bestandteil des zu entwickelnden Frameworks.

Neben der Analyse von Einsätzen ist die Vorausplanung oder Gefahrenabschätzung für bestimmte seltene Ereignisse ebenfalls ein Anwendungsfall für ein solches Framework. Beispiele für solche Evaluationen können Hochwasserereignisse oder auch Waldbrände sein. In diesem Fall kann die Simulation helfen, die benötigte Anzahl an Rettungskräften sowie kritische Punkte zu ermitteln. Auch in diesem Anwendungsfall ist die dynamische Darstellung des Ablaufes vorteilhaft, um Reaktionszeiten zu ermitteln.

3.2 Erstellung generischer Einsätze aus Bewegungsmustern

Die Bewegungen der Rettungskräfte richten sich nach den jeweiligen lokalen Gegebenheiten und der vorgefundenen Situation am Einsatzort. In jedem Fall werden die entsprechenden Dienstvorschriften berücksichtigt. Daraus lässt sich der in Abbildung 1 dargestellte, generische Ablauf eines Einsatzes ableiten.

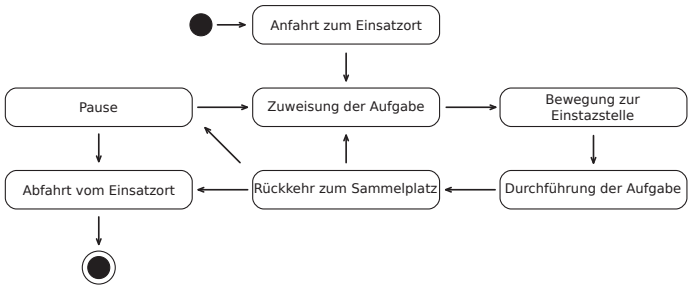


Abbildung 1: Generischer Einsatzablauf

Nach dem Ausrücken begeben sich die Einsatzkräfte zum eigentlichen Einsatzort. Dort

angekommen, werden den einzelnen Gruppen Aufgaben übertragen, die sie an der eigentlichen Einsatzstelle auszuführen haben. Anschließend begeben sich die Gruppen zur jeweiligen Einsatzstelle und beginnen mit der Durchführung der übertragenen Aufgabe. Jede Gruppe kommt unabhängig von den anderen Gruppen nach einer gewissen Zeit oder wenn die Aufgabe erfüllt wurde, zum zentralen Sammelplatz zurück. Dort kann sie ggf. nach einer Pause erneut eine Aufgabe übernehmen oder verlässt den Einsatzort, um zum Stützpunkt zurückzukehren. Diese generische Struktur ermöglicht die Modellierung verschiedener Einsätze vom Gebäudebrand bis zur Suche nach vermissten Personen und Hochwasserereignissen.

Ein solcher Einsatz modelliert ein Fahrzeug und die dazugehörige Mannschaft als eine Gruppe von Knoten. In jeder Phase wird den Gruppenmitgliedern ein entsprechendes Bewegungsmuster zugeordnet und anschließend konfiguriert. Zur Konfiguration gehören zum Beispiel Start- und Zielkoordinaten sowie die Geschwindigkeit mit der sich die Gruppe bewegt. Eine solche Gruppe kann am eigentlichen Einsatzort auch in mehrere Untergruppen, mit unterschiedlichen Aufgaben aufgeteilt werden. Der einzigen Unterschiede zwischen verschiedenen Einsätz sind die jeweiligen Bewegungsmuster der Rettungskräfte während der eigentlichen Durchführung der Aufgaben, die Anzahl an Rettungskräften insgesamt sowie deren Anzahl in den einzelnen Gruppen. Die übertragenen Aufgaben sind für jeden Einsatztyp spezifisch, kommen aber in vergleichbarer Form häufig vor. Beispiele für aufgabenspezifische Bewegungsmuster sind charakteristische Formationen bei der Suche nach vermissten Personen, das paarweise Vorgehen bei einem Löschangriff oder das Sichern von Deichen mit Hilfe von Sandsäcken. Alle Bewegungsmuster enthalten auch Mechanismen, um Hindernissen auszuweichen, vgl. [KSS+14].

Die Übergänge zwischen den einzelnen Phasen werden mit Wahrscheinlichkeiten beschrieben, die zum Teil zeitabhängig sind. Zum Beispiel steigt die Wahrscheinlichkeit, dass Einsatzkräfte nach Abschluss einer ersten Teilaufgabe pausieren mit der Gesamtdauer des Einsatzes an.

3.3 Konzept

Die generische Missionsstruktur mit den entsprechenden Bewegungsmustern bildet das Kernstück des hier vorgestellten Frameworks. Ziel ist es, generisch möglichst realitätsnahe Szenarien auch für große Schadenslagen zu erzeugen. Dafür ist entsprechendes Expertenwissen notwendig, um den Ablauf und die notwendigen Aktionen nachbilden zu können. Dieses Wissen fließt in die Beschreibung und Modellierung der Bewegungsmuster mit ein, wird aber auch für die Konfiguration des gesamten Einsatzes benötigt. Abbildung 2 zeigt die Architektur des Frameworks.

Momentan werden die Bewegungsmuster aus [KSS+14] modular nach einem Baukastenprinzip bereitgestellt. Weitere Muster können integriert werden, um weitere Einsatztypen zu unterstützen.

Neben den vordefinierten Bewegungsmustern, ist ein Konfigurationstool notwendig, das zunächst die oben beschriebene Erzeugung einzelner Einsätze für eine Gruppe von Ein-

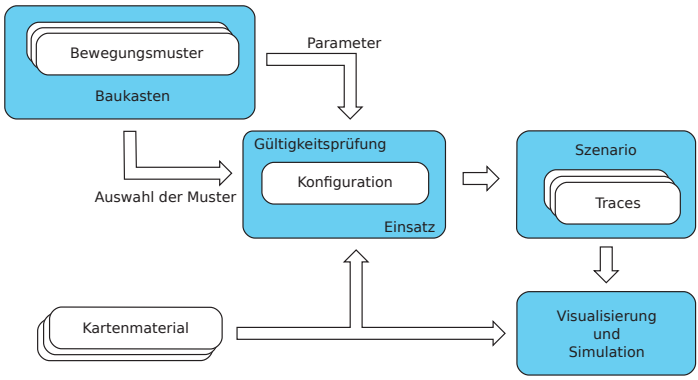


Abbildung 2: Architektur des Mobilitätsframeworks

satzkräften ermöglicht. Mehrere so konfigurierte Einsätze bilden dann ein Szenario, aus dem Traces erzeugt werden. Die Traces können anschließend in entsprechenden Simulationsumgebungen exportiert werden. Darüber hinaus können die Traces innerhalb der integrierten Simulationsumgebung visualisiert werden.

Um die Konfiguration zu erleichtern ist es möglich, verschiedene Einsätze als vordefinierte Konfigurationen abzuspeichern und später erneut zu verwenden. Außerdem werden Kartendaten im WellKnownText-Format (wkt) eingebunden, um die realistische Positionierung der Einsatzkräfte im entsprechenden Gebiet zu ermöglichen. Die Bereitstellung der Kartendaten erfolgt im wkt-Format, weil dies von ONE (Opportunistic Network Emulator) [KOK09], einer Simulationsumgebung für die Evaluation von verzögerungstoleranten Netzen, direkt unterstützt wird und somit eine Visualisierung der Traces und der Kartendaten in einer Umgebung gewährleistet ist. Die entsprechenden Karten können ergänzend zu Straßen oder Wegen auch Informationen zu Höhenlinien und Hindernissen enthalten. Diese Daten werden momentan statisch importiert, zukünftig ist eine Integration eines entsprechenden GIS Systems angedacht.

Damit die erzeugten Einsätze an die entsprechenden Anforderungen angepasst werden können, verfügt jedes Bewegungsmuster über mehrere Konfigurationsmöglichkeiten, die neben der Anzahl an Knoten und deren typischer Geschwindigkeit auch Zufallsaspekte und Ansätze zum Ausweichen von Hindernissen beinhalten. Details zur Realisierung der Bewegungsmuster sind in [KSS+14] beschrieben. Durch die Verwendung zufälliger Elemente wird auch deutlich, dass sich abhängig von den verwendeten Zufallsverteilungen unterschiedliche Situationen und damit auch unterschiedliche Varianten des gleichen Einsatzes erzeugen lassen.

4 Implementierung und Validierung

Das beschriebene Framework wurde in Java implementiert und so modular aufgebaut, dass weitere Bewegungsmuster einfach hinzugefügt werden können, wenn sie die entsprechenden Schnittstellen unterstützen und die Gültigkeitsparameter festgelegt werden.

Die Konfiguration erfolgt momentan rein textbasiert, weil diese sich an der von ONE vorgegeben Struktur orientiert und somit eine Integration mit der Simulationsumgebung vereinfacht wird. Perspektivisch ist aber die Entwicklung einer graphischen Oberfläche vorgesehen. Zur Konfiguration gehört auch eine Plausibilitätsprüfung, die verhindern soll, dass eine inkompatible Abfolge einzelner Muster erzeugt wird. Dies betrifft insbesondere die Übergänge zwischen den einzelnen Phasen, in denen aus einer großen Gruppe mehrere kleinere Untergruppen werden. Außerdem werden die Parameter der einzelnen Muster auf ihre Gültigkeit geprüft.

Die Visualisierung der Szenarien erfolgt in ONE. Die notwendige Konfigurationsdatei für ONE wird automatisch, anhand des gewählten Szenarios und den daraus resultierenden Traces, erzeugt. Um gleichzeitig Kartendaten und die erzeugten Traces, die im gleichen Koordinatensystem vorliegen, in einer Simulation zu verwenden, musste ONE zunächst um ein entsprechendes Modul erweitert werden. Dieses neue Modul, *ExternalMapBased-Movement*, erlaubt die erforderliche Transformation der Trace-Koordinaten, die in ONE für Kartendaten standardmäßig durchgeführt wird. Ohne dieses Modul werden die Traces zwar korrekt eingelesen, die Visualisierung erfolgt aber nicht an der ursprünglichen Position auf der Karte, weil ONE für kartenbasierte Mobilitätsmodelle den Koordinatenursprung an eine andere Position legt und Traces standardmäßig nicht entsprechend anpasst.

Das Framework wurde evaluiert, indem ein fiktiver Einsatz zur Vermisstensuche in einem Waldgebiet nachgebildet wurde. Das Szenario wurde in Zusammenarbeit mit einer Rettungshundestaffel entwickelt. Abbildung 3 zeigt die Aufteilung des gesamten Einsatzgebietes in die entsprechenden Suchabschnitte (rot), den zentralen Sammelplatz (gelb) und die Position der gesuchten Person (grüner Punkt in Suchabschnitt 22).

Nach der Generierung der Traces mit Hilfe des Frameworks wurde dieses Szenario in ONE visualisiert. Abbildung 4 zeigt einen Screenshot mit einem Ausschnitt des Beispielszenarios. Die einzelnen Knoten stellen die Rettungskräfte mit ihren Rettungshunden sowie die entsprechenden Einsatzfahrzeuge dar.

Die Kreise um die Knoten symbolisieren jeweils die Kommunikationsreichweite der Knoten, die in diesem Beispiel auf 150 m für Einsatzkräfte und 1000 m für Kommunikationssysteme in Fahrzeugen angenommen wurden. Nur Knoten, die miteinander verbunden (schwarze Linien) sind, können direkt kommunizieren. Dieser Screenshot zeigt, dass die Kommunikation zur Einsatzleitung am Sammelplatz irgendwann unterbrochen wird, wenn keine zusätzlichen Knoten mitentsprechender Reichweite eingesetzt werden. An dieser Stelle wird deutlich, wie das Framework zur Evaluierung verschiedener Kommunikationsaspekte verwendet werden kann. Auf Basis der vorliegenden Szenarien lassen sich anschließend verschiedene Protokolle sowie andere Lösungsansätze wie z.B. die Verwendung zusätzlicher mobiler Konten oder Geräten mit größerer Reichweite evaluieren.

Aber auch für die Bewertung bzw. Planung eines solchen Einsatzes sind diese Schluss-

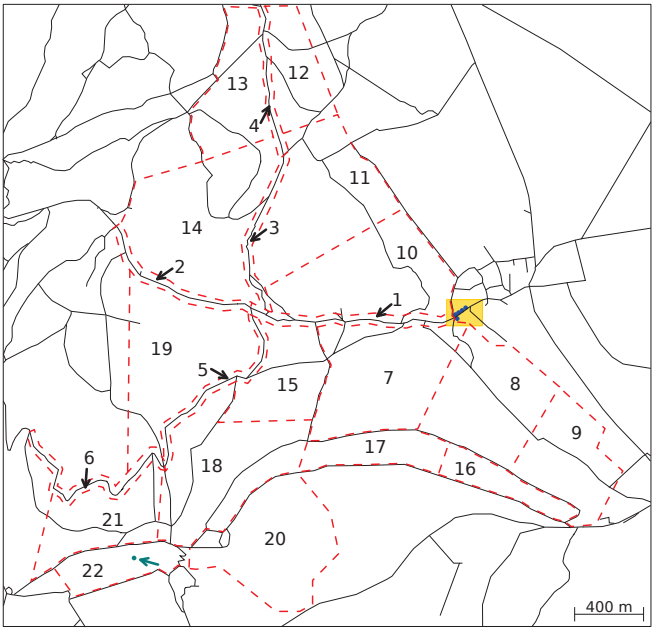


Abbildung 3: Beispielszenario Vermisstensuche in Waldgebiet [KSS+14]

folgerungen wichtig, denn die Simulation kann zweigen, warum die Kommunikation an einem bestimmten Punkt abgebrochen ist und Hinweise geben, wann ein zusätzlicher Knoten erforderlich ist, der von vorhandenen Rettungskräften entsprechend positioniert werden muss, um die Kommunikation aufrecht zu erhalten. Außerdem ermöglicht die Simulation eine Abschätzung der Einsatzdauer, wenn unterschiedlich viele Einsatzkräfte zur Verfügung stehen.

5 Zusammenfassung und Ausblick

In diesem Beitrag wurde ein Framework vorgestellt, das es ermöglicht, generische Traces für Einsatzszenarien zu erzeugen und an einem Beispielszenario gezeigt, wie diese Daten in einer Simulationsumgebung visualisiert werden können. Die Entwicklung des Frameworks erfolgte vor dem Hintergrund, entsprechende Daten für die Analyse und Entwicklung neuer Kommunikationsprotokolle und -systeme zu verwenden. Die so erzeugten Daten bieten jedoch interessante Ansätze, um mögliche Schadenslagen und benötigte bzw. verfügbare Einsatzkräfte zu evaluieren und haben deshalb über die Kommunikationsforschung hinaus eine Bedeutung.

Der aktuelle Stand des Frameworks ermöglicht es, bei entsprechendem Fachwissen realis-

Abbildung 4: Screenshot des Beispielszenarios in Simulationsumgebung ONE

tische Szenarien zu erzeugen und zu visualisieren. Aufgrund der noch fehlenden graphischen Oberfläche ist die Konfiguration der einzelnen Abläufe jedoch relativ kompliziert. Deshalb soll zukünftig eine geeignete Oberfläche entworfen werden.

Momentan können nur zweidimensionale Szenarien erzeugt werden. Durch die schon vorhandene Integration entsprechender Kartendaten, wäre eine Erweiterung um dreidimensionale Daten ebenfalls möglich. Die benötigten Informationen zum Geländeprofil könnten dann zusätzlich bei der Erzeugung der Bewegungen berücksichtigt werden und so die Realitätsnähe der Simulation weiter erhöhen. In diesem Zusammenhang ist auch die Anbindung an ein Geoinformationssystem denkbar. Insbesondere GRASS GIS [NBLM12] mit seinen Funktionen zur Simulation von Waldbränden und Hochwasserereignissen wäre dafür interessant.

Danksagung

Die Autoren danken Herrn Torsten Kiesewetter von der Freiwilligen Feuerwehr Wipfratal im Ortsteil Branchewinda sowie Herrn Hendrik Stolz von der Facheinheit Rettungshunde Ortungstechnik für die Unterstützung bei der Entwicklung der vorgestellten Szenarien und die kritische Betrachtung des vorgestellten Frameworks.

Literatur

[AGPG+07] N. Aschenbruck, E. Gerhards-Padilla, M. Gerharz, M. Frank und P. Martini. Modelling Mobility in Disaster Area Scenarios. In *10th ACM-IEEE International Symposium on*

Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWIM), Seiten 4–12, Oct. 2007.

[CBD02] T. Camp, J. Boleng und V. Davies. A Survey of Mobility Models for Ad Hoc Network Research. *Wireless Communications and Mobile Computing*, 2(5):483–502, Sep. 2002.

[CST+13] D. Cohen, N. Sevdalis, D. Taylor, K. Kerr, M. Heys, K. Willett, N. Batrick und A. Darzi. Emergency preparedness in the 21st century: Training and preparation modules in virtual environments. *Resuscitation*, 84(1):78–84, 2013.

[Fac] Fachgebiet Technische Informatik, Universität Kassel. KATIE - Katastrophen, Avatare, Technische Simulationen – in virtuellen Environments.

[Fin04] M. A. Finney. FARSITE, Fire Area Simulator–Model Development and Evaluation. Bericht, United States Department of Agriculture Forest Service Rocky Mountain Research Station, 2004.

[KOK09] A. Keränen, J. Ott und T. Kärkkäinen. The ONE Simulator for DTN Protocol Evaluation. In *2nd International Conference on Simulation Tools and Techniques (Simutools)*, Mar. 2009. Article 55.

[KSS+14] S. Krug, M. F. Siracusa, S. Schellenberg, P. Begerow, J. Seitz, T. Finke und J. Schröder. Movement Patterns for Mobile Networks in Disaster Scenarios. In *8th IEEE WoWMoM Workshop on Autonomic and Opportunistic Communications (AOC 2014)*, Sydney, Australia, June 2014.

[MSZ+12] A. Moreno, A. Segura, S. Zlatanova, J. Posada und A. García-Alonso. Introducing GIS-based simulation tools to support rapid response in wildland fire fighting. In C. Brebbia und G. Perona, Hrsg., *Modelling, Monitoring and Management of Forest Fires III*, Jgg. 158 of *Wit Transactions on Ecology and the Environment*, Seiten 163 –174. WIT Press, 2012.

[NBLM12] M. Neteler, M. H. Bowman, M. Landa und M. Metz. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software*, 31:124–130, 2012.

[NHIK07] S. C. Nelson, A. F. Harris III und R. Kravets. Event-driven, Role-based Mobility in Disaster Recovery Networks. In *2nd ACM Workshop on Challenged Networks (CHANTS)*, Seiten 27–34, Sep. 2007.

[UNAK09] M. Y. S Uddin, D. M. Nicol, T. F. Abdelzaher und R. Kravets. A Post-Disaster Mobility Model for Delay Tolerant Networking. In *Winter Simulation Conference (WSC)*, Seiten 2785–2796, Dec. 2009.

Übungsunterstützung für Einsatztrainings des Massenanfalls von Verletzten (MANV)

Johannes Sautter[1], Manuel Habermann[1], Sandra Frings[1], Friederike Schneider[2], Bernhard Schneider[3], Holger Bracker[3]

[1]Fraunhofer IAO
Nobelstr. 12
70569 Stuttgart

[2]DRK Generalsekretariat
Carstennstr. 58
12205 Berlin

[3]Airbus Defence and Space
Landshuter Str. 26
85716 Unterschleissheim

[1]{Vorname.Nachname}@iao.fraunhofer.de
[2]schneidf@drk.de
[3]{Vorname.Nachname}@cassidian.com

Abstract: Einsatzübungen sind ein wichtiges Instrument, um haupt- und ehrenamtliche Helfer und Führungskräfte im Rettungsdienst auf Großschadenslagen vorzubereiten. Die Auswertung solcher Übungen sollte so erfolgen, dass eine Vergleichbarkeit und Unabhängigkeit vom Einsatzkonzept gegeben ist. Außerdem sollten Zeitaspekte und die Qualität der Patientenversorgung eine Rolle spielen. Dieser Artikel beschreibt ein Trainingskonzept mit Bewertungsschema zur Auswertung sowie Anwendungsfälle und Benutzerschnittstellen für ein Übungsunterstützungssystem für die strukturierte Datenerfassung und Unterstützung der Auswertung.

1 Einleitung und Motivation

Ein Massenanfall von Verletzten (MANV) ist im Katastrophenschutz definiert als ein Notfall mit einer größeren Anzahl von Verletzten, Erkrankten oder Betroffenen, der mit vorhandenen Ressourcen aus dem Rettungsdienstbereich nicht bewältigt werden kann (DIN 13050:2009-02 (Rettungswesen), via BBK Website Glossar). MANV-Szenarien und damit verbundene Großschadenslagen sind selten. Die Koordination gestaltet sich insbesondere für die am Schadensort zuerst eintreffenden Einsatzkräfte sehr schwierig, da die organisatorischen und medizinischen Abläufe sich von denen bei kleineren Ereignissen signifikant unterscheiden (Max et al 2013). Regelmäßiges Training ist daher insbesondere auch bei ehrenamtlichen Helfern im ländlichen Raum bedeutend. In der Regel werden Einsätze bei MANV-Schadensszenarien auf Basis lokal-spezifischer Einsatzkonzepte für die Zusammenarbeit beteiligter Einsatzeinheiten und Organisationen ausgeführt. Beispiele für beteiligte Hilfsorganisationen sind das Deutsche Rote Kreuz (DRK), das Technische Hilfswerk (THW) sowie lokale Feuerwehren, und die Polizei.

Für eine adäquate Auswertung, sollte eine Evaluation einer Übung oder eines Trainings unter anderem Anforderungen an Vergleichbarkeit, Reproduzierbarkeit und Unabhängigkeit genügen. Wesentlich ist zudem eine Messung der Patientenversorgung. Dafür ist es erforderlich Messgrößen mittels eines Trainingskonzeptes zu erheben. Eine Unterstützung des organisatorischen Trainingskonzeptes durch ein IT-System ist nun naheliegend, da es erforderlich ist, Übungsdaten zu erfassen und innerhalb kurzer Zeit adäquat zu Analyse- und Vergleichszwecken zu aggregieren.

Um MANV-Übungen vergleichbar zu machen und ein einheitliches organisatorisches Trainingskonzept für die Praxis zu entwickeln, fand eine intensive Auseinandersetzung zu den Anforderungen der Endnutzer an ein IT-System statt – unter anderem durch die Teilnahme an Übungen, Workshops und Expertengesprächen. Bei der Entwicklung wurde an ein im Einsatz befindliches Trainingskonzept des Kreisverbands Garmisch-Partenkirchen des Bayrischen Roten Kreuzes angeknüpft (Geuther 2013). Zur Evaluation des Trainingskonzeptes sowie eines darauf basierenden Übungsunterstützungssystems sind zwei Übungen geplant. Im Folgenden wird zunächst auf den Stand der Technik bezüglich Übungsunterstützungssysteme eingegangen. Anschließend wird das organisatorische Trainingskonzept beschrieben gefolgt von einer Beschreibung der Anwenderperspektive angereichert mit grafischen Mockups als Prototypen der Benutzerschnittstellen des Übungsunterstützungssystems. Abschließend wird das konzeptuelle Datenmodell des Systems beschrieben und ein Ausblick auf noch offene Fragestellungen gegeben.

2 Stand der Technik zu Übungsunterstützungssystemen

Mit der umfangreichen Software Exonaut™ Training and Exercise Manager (TEM) (4C Strategies 2014) können Übungen und Trainings designt, implementiert und ausgewertet werden. Das System deckt die Planungs-, Ausführungs- und Nachbereitungsphase von Übungen ab. TEM ist über eine Webschnittstelle zugänglich und somit für die Übungsteilnehmer auch mobil verfügbar. Übungsdokumente können archiviert werden und somit kann TEM als Knowledge Management System eingesetzt werden. Als allgemeine IT-Lösung für Übungsunterstützung ist nicht erkennbar inwieweit spezifische Anforderungen des Einsatztrainings von notfallmedizinischen Arbeitsprozessen abgedeckt werden. Ein Forschungsprojekt des Bundesministeriums für Bildung und Forschung ist das Projekt RescueLab (Koch et al 2014). Das Ziel des Projektes ist die Steigerung der Qualität der Übungsabläufe mittels IT-gestützter Software. Unterschiedliche Rollen sind involviert zur Übungsbeobachtung und Nachbetrachtung. Die Datenerfassung erfolgt teilweise mittels Sensoren. RescueLab fokussiert sich auf Übungen mit technischen Rettungsorganisationen (Feuerwehr, THW) während im Folgenden die medizinische Übungsnachbetrachtung einen hohen Stellenwert erhält.

3 Organisatorisches Trainingskonzept

Um die Leistung von Einsatzeinheiten bei Trainings und Übungen adäquat auswerten zu können, sollte ein Trainingskonzept folgende Anforderungen erfüllen (Brauner et al 2014): Es sollte eine Vergleichbarkeit zwischen der erfolgten Abarbeitung eines Übungsszenarios (im Folgenden Übungslauf) gegeben sein. Weiterhin sollten Ergebnisse reproduzierbar sein, so dass aggregierte Kennzahlen immer auf konkret gemessene Daten zurückgeführt werden können. Den Standards der Deutschen Gesellschaft für Evaluation (DeGEval) folgend sollte auf die Nützlichkeit und Durchführbarkeit in der Praxis geachtet werden. Fairness und Genauigkeit sind weitere Standards, denen ein Evaluationskonzept im Allgemeinen folgen sollte. Da bei MANV-Szenarien und darauf abzielenden Einsatzkonzepten in Deutschland eine sehr große lokale Spezifität vorliegt, sollte ein Konzept zur Trainingsevaluation unabhängig von lokalen Einsatzkonzepten anwendbar sein. Als wesentlicher medizinischer Erfolgsfaktor sollte das Trainingskonzept zudem in der Lage sein, eine adäquate Patientenversorgung zu messen (Brauner et al 2014).

3.1 Indikatoren als Auswertungsschema für MANV-Übungen

Für den Erfolg eines MANV-Einsatzes ist es grundsätzlich am wichtigsten, die notfallmedizinische Individualversorgung schnellstmöglich wiederherzustellen. Taktische Kernpunkte sind hierbei die Festlegung von Prioritäten (Sichtung) und der Transport in umliegende Krankenhäuser. Diese Kernpunkte verstehen wir als Indikatoren und können mittels einer Berechnungsvorschrift aus Übungsdaten errechnet werden. Indikatoren umfassen konkrete Zeitspannen oder Verhältnisse. Zeitliche Indikatoren werden relativ erhoben und messen eine Zeitspanne ab Eintreffen des ersten Fahrzeugs an der Schadensstelle (Referenzzeitpunkt der Schadensabarbeitung):

- Zeit bis der letzte Patient abtransportiert wird
- Zeit bis alle Patienten der Sichtungskategorie rot die Schadensstelle verlassen haben
- Verhältnis von Einsatzkräften und Patienten insgesamt
- Verhältnis von Einsatzkräften und Patienten pro Intervall
- Zeit bis alle Patienten vorgesichtet wurden (nicht-ärztliche Vorsichtung)
- Zeit bis alle Patienten gesichtet wurden
- Anzahl der Rettungsmittel insgesamt
- Anzahl der durchgeführten Maßnahmen (Überprüfung der Atmung, Bewusstseinskontrolle, Blutstillung, Lagerung, Wärmeerhalt und Betreuung)
- Verhältnis zwischen Rettungsmittel und Patient

Unter Auswertungsschema verstehen wir die Werte der Indikatoren sowie die Ausgabe der Messgrößen in der zeitlichen Abfolge des Übungslaufs (z.B. in Diagrammen). Als Messgröße im Zeitverlauf ist zum Beispiel interessant wieviel Patienten sich zu jedem Zeitpunkt noch an der Schadensstelle befinden und demgegenüber wie viele Helfer bzw. Fahrzeuge.

Neben der Anzahl der durchgeführten Maßnahmen als Indikator für die Patientenversorgung vor Ort fließt auch eine Beurteilung dieser Maßnahmen durch freiwilligen Personen, die die Verletzten spielen (im Folgenden Mimen) in die Übungsevaluation mit ein. Wenngleich die Einschätzung der Mimen stets höchst subjektiv erfolgt, bietet dieses Wissen den Einsatzkräften einen wichtigen Anreiz, die Maßnahmen auch hinsichtlich qualitativer Aspekte durchzuführen und zu verbessern.

3.2 Trainingskonzept für Einsatztrainings

Das Trainingskonzept bildet den organisatorischen Rahmen für die Auswertung durch die Messung und Erhebung der Übungsdaten. Für übende Einsatzkräfte sollte, begleitet durch den Übungsleiter, ein Lerneffekt auf Basis des Auswertungsschemas erzielt werden. Das Trainingskonzept des BRK-Kreisverbands Garmisch-Partenkirchen (Geuther 2013) wird im Folgenden adaptiert, weiterentwickelt und setzt insbesondere die Erhebung benötigter Übungsdaten um. Dadurch ist ein effizientes Lernen der Einsatzeinheiten als Ganzes sowie der einzelnen Durchführenden möglich.

Vor der Durchführung einer Übung muss zunächst das genaue Schadensszenario definiert werden und Fragen zur Ausgangslage, zum Übungsort und zum Personal müssen beantwortet werden. Auf Grundlage des Schadenszenarios können die Auswahl der Verletzungsmuster und die Zuordnung der Mimen erfolgen. Die fiktiven Patientencharaktere umfassen neben Verletzungsmuster und Sichtungskategorie auch Informationen wie fiktiver Name, Übungsausweisnummer sowie fiktives Geburtsdatum. Bei Umsetzung des Trainingskonzeptes ohne Übungsunterstützungssystem können diese Informationen auch in Excel-Tabellen vorgehalten werden. Eine Bündelung und schnelle Auswertung der Übungsdaten, so dass eine qualitative Auswertung und Vergleichbarkeit verschiedener Übungsläufe erfolgen kann, ist jedoch mittels Tabellenkalkulation schwierig.

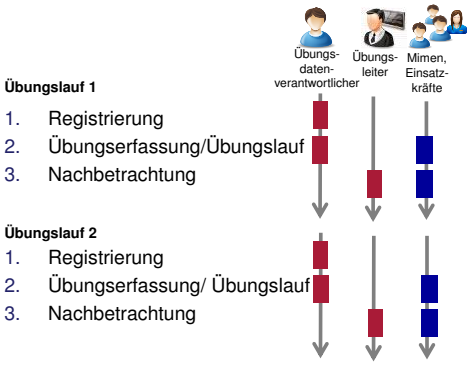


Abbildung 1: Typischer Ablauf eines Übungstermins

In Vorbereitung eines Übungsdurchlaufs werden die Mimen und Einsatzkräfte mit Übungsausweisen und Helfer-/Notarzt-Identifizierungskarten ausgestattet. Die Mimen erhalten zusätzlich eine Bewertungskarte, auf der die Zeitpunkte der Vorsichtung, der Sichtung, die jeweils erteilte (Vor-) Sichtungskategorie, die durchgeführten Versorgungsmaßnahmen sowie Bewertungen und der Zeitpunkt des Abtransports eingetragen werden. Deren Auswertung erfolgt im Rahmen des Trainingskonzeptes parallel bzw. nachgelagert zum Übungslauf oder händisch bzw. mit Hilfe der Tabellenkalkulation. Zur Koordination der Registrierung der Mimen und Einsatzkräfte sowie für die Datenerfassung bedarf es der Benennung eines

sogenannten Übungsdatenverantwortlichen, welcher vor und während der Übungsläufe ausschließlich mit dem Einpflegen und Verwalten der Daten betraut ist. Die Auswertung findet vor und während der Übungsnachbesprechung durch den Übungsleiter statt und wird entweder durch eine Benutzeroberfläche oder durch das Ergebnis zuvor erstellter Berechnungsvorschriften (Indikatoren) aus der Tabellenkalkulation unterstützt. Der Übungsleiter ist dadurch in der Lage, den Einsatzkräften gezieltes Feedback zu geben. Hierbei geht es nicht um die Nennung von Einzelpersonen und die Kritik an diesen, sondern um die Aufdeckung genereller Schwachstellen. Die Einblendung der zugrundeliegenden Detaildaten ist daher optional. Abbildung 1 illustriert die wichtigen Akteure und deren Rolle bei jedem Übungslauf. Ein Übungslauf umfasst in der Regel zwei Stunden, so dass mehrere Übungsläufe in Folge möglich sind. Durch das wiederholte Training derselben Schadenslage ergibt sich bei den Einsatzkräften ein erhöhter Trainingseffekt, der zu einem Zugewinn an Erfahrung und Selbstvertrauen beitragen kann (Geuther 2013, S.12). Es können so mehrere Übungsläufe an ein bis zwei Tagen erfolgen und wird in der Regel durch Einbindung ehrenamtlicher HelferInnen am Wochenende durchgeführt.

Insgesamt ist es wichtig, dass nicht nur Daten quantitativ erfasst werden sollen, sondern auch eine qualitative Interpretation der Daten ermöglicht wird. Es wäre beispielsweise wenig aussagekräftig nach der Übung zu wissen, wie schnell alle Patienten gesichtet wurden; wenn man aber Informationen über die Anzahl der zur Verfügung stehenden Ressourcen und Informationen über die Infrastruktur hinzufügt, kann man valide Aussagen darüber generieren, wie „erfolgreich“ der Übungslauf war, wo Schwachstellen liegen und vor allem wie man diese korrigieren kann. Bei den Übungsleitern handelt es sich um erfahrene Einsatz- und Führungskräfte, welche in der Lage sind, auf Basis der erfassten Daten und aggregierten Indikatoren selbstständig Rückschlüsse zu ziehen. Das Übungsunterstützungssystem oder eine Tabellenkalkulationssoftware kann und soll hierbei lediglich unterstützend wirken.

4 Benutzersicht und Anwendungsfälle

Die Benutzersicht des Übungsunterstützungssystems für Einsatztrainings kann in drei übergeordnete Rollen untergliedert werden. Administrator, Übungsdatenverantwortlicher und Übungsleiter (vgl. Abbildung 2).

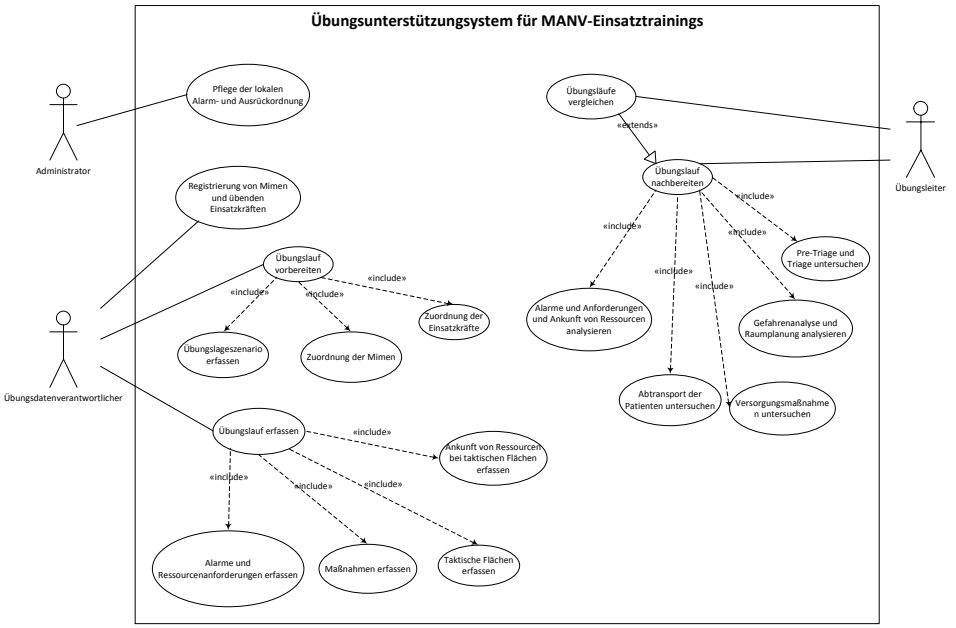


Abbildung 2: UML Use Case Diagramm des Übungsunterstützungssystems

Ein Administrator wird benötigt zur Wartung und Inbetriebnahme des Systems für einen spezifischen Orts- oder Kreisverband einer Hilfsorganisation. Der Übungsdatenverantwortliche ist mit der Registrierung von Personen, deren Zuordnung zu Übungsläufen und der strukturierten Erfassung der Übungsläufe betraut. Der Übungsleiter koordiniert das gesamte Einsatztraining und gibt den übenden Einsatzkräften mit Hilfe der Nachbetrachtungsansicht des Übungsunterstützungssystems fachliche Rückmeldungen.

4.1 Pflege der Alarm- und Ausrückeordnung

Eine Person, die regelmäßig mit der Ausstattung und der Ausrüstung einer Ortsgruppe oder eines Kreisverbandes einer Hilfsorganisation betraut ist, kann die Rolle des Administrators einnehmen. Diese ist verantwortlich für die Pflege der aktuellen Alarm- und Ausrückeordnung (AAO) im Übungsunterstützungssystem. Die AAO ist spezifisch festgelegt und an der regionalen Rettungsdienstinfrastruktur ausgerichtet. Für jedes Stichwort ist in der AAO eine Einsatzmittelkette definiert, die ausreichend Einheiten und Fahrzeuge dimensioniert und die ggf. mehrere Hilfsorganisationen umfasst. Um die korrekten Fahrzeug-/Truppkategorien bei einer Ressourcenanforderung im System hinterlegt zu haben, muss die AAO und das Personal des Land-/Stadtkreises, in dem das Übungsunterstützungssystem eingesetzt werden soll, im Programm eingepflegt werden. Das Stichwort MANV1 kann in einem spezifischen Landkreis beispielsweise mit 10 Rettungswagen (RTWs), 4 Krankentransportwagen (KTWs), 5 Notarzteinsatzfahrzeuge (NEF), einem Mannschaftstransportwagen (MTW) und einem Gerätewagen Sanitätsdienst (GW-San) verknüpft sein.

4.2 Registrierung, Vorbereitung und Übungslauferfassung

Die Rolle des Übungsdatenverantwortlichen sollte eine Person einnehmen, die zuvor bereits mit dem Übungsunterstützungssystem vertraut ist, um eine schnelle Eingabe sicherzustellen. Die drei wesentlichen Aufgaben werden im Folgenden beschrieben.

Registrierung von Mimen und Einsatzkräften: Die Übungsvorbereitung kann schon einige Stunden vor dem eigentlichen Übungsablauf durchgeführt werden, damit die Datenerfassung während der Übung nicht zu umfangreich wird. Dies betrifft insbesondere die Registrierung der Mimen und Einsatzkräfte. Die Zuordnung einer realen Person auf ihre Mimen-Identität hinterlegt der Übungsdatenverantwortliche im Übungsunterstützungssystem.

Vorbereitung des Übungslaufes: Bevor an einem Übungstermin ein konkreter Übungslauf gestartet werden soll, muss festgelegt werden, welche der zuvor registrierten Mimen und Einsatzkräfte an diesem Übungslauf teilnehmen werden.

Erfassen des Übungslaufes: Nach Start des Übungslaufes durch den Übungsleiter kann der Übungsdatenverantwortliche mit der Erfassung beginnen. Die Erfassung ist sowohl während des Übungslaufes als auch nach Abschluss des Übungslaufes möglich. Währenddessen können das Mithören des Sprechfunks, die Beobachtung des Übungsgeschehens (bspw. zur Definition taktischer Flächen) oder abgegebene Auswertebögen der Mimen als Informationsquelle dienen. Nach Ende des Übungslaufes können ebenso weitere Bewertungskarten von Mimen eingegeben werden, wie im Rahmen der Einsatzdokumentation erstellte Artefakte genutzt werden (bspw. Dokumentation der Führungskräfte von Bereitstellungsraum und Rettungsmittelhalteplatz über Ankunft und Weiterfahrt von Fahrzeugen).

Um beim späteren Betrachten der Übungsdaten im System die Randbedingungen und das zugrundeliegende Lageszenario des Übungslaufes bereit zu haben, erfasst der Übungsdatenverantwortliche zunächst das angenommene Lagebild nach den Vorgaben des Übungsleiters. In Abbildung 3 (a) wird die Benutzerschnittstelle zur Erfassung des Schadensszenarios gezeigt. Abbildung 3(b) zeigt die Maßnahmenerfassung.

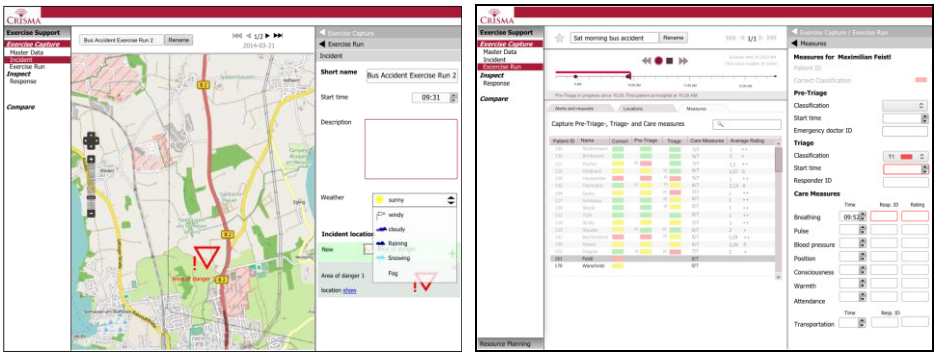


Abbildung 3: Benutzerschnittstelle zur Szenario-Erfassung (a) und Maßnahmenerfassung (b)

Bei der Gestaltung der Benutzerschnittstelle zur Datenerfassung wurde auf Fehlertoleranz und Plausibilisierung Wert gelegt. Fehlende Eingaben werden beispielsweise durch rot umrandete Markierungen hervorgehoben (vgl. Abbildung 3 (a) und (b)).

Alarme und Ressourcenanforderung: Die Ressourcenanforderung ist meist mit Fahrzeuganforderung gleichzusetzen, da das Equipment oft an die Fahrzeuge gebunden ist. Aufgabe des Übungsdatenverantwortlichen ist es, folgende Werte aufzufassen und zu dokumentieren: Alarmierungszeit, Anzahl/Art der Fahrzeuge und Bemerkungen. In Bemerkungen können Besonderheiten der einzelnen Alarme notiert werden, wie beispielsweise übermittelte Lageinformationen. Die Übersicht der bereits georderten Fahrzeuge wird zentral in einer Tabelle aufgelistet.

Maßnahmen erfassen: Verschiedene Maßnahmen, welche an einem Mimen während des Übungslaufs durchgeführt werden, werden von dem jeweiligen Mime mittels Bewertungsbogen dokumentiert und bewertet. Die Bewertungsbögen werden vom Übungsdatenverantwortlichen erfasst und in einer Tabelle ausgegeben. Die erfassten Daten der Mimen sind (Vor-)Sichtungskategorie, (Vor-)Sichtungszeit, sichtende/r Einsatzkraft/Notarzt, Bewertungen der einzelnen Maßnahmen.

Taktische Flächen erfassen: Die taktischen Flächen werden auf einer Übersichtskarte durch Drag and Drop oder mittels Geokoordinaten (long/lat) eingegeben. Die taktischen Flächen werden durch ihre speziellen Symbole auf der Karte deutlich erkennbar.

Ankunft von Ressourcen bei taktischen Flächen: Die genaue Erfassung der Ankunft von Ressourcen ist schwierig zu ermitteln, da die Fahrzeuge, welche die Ressourcen beinhalten, nur eine Statusmeldung über ihr Funkmeldesystem abgeben, wenn sie ihr Einsatzziel erreicht haben. Die Daten zur Erfassung können entweder von der Leitstelle bereitgestellt werden (sofern die Leitstelle in das Übungsgeschehen involviert ist) oder durch die Auswertung von Artefakten der Einsatzdokumentation.

4.3 Nachbetrachtung

Die Nachbetrachtung eines Übungslaufs ist für das beteiligte Personal eine wichtige und die oft einzige Beurteilung ihrer geleisteten Arbeit. Der Übungsleiter kann sich die Auswertung optional zeitnah zum Übungslauf vorab anschauen. Auf jeden Fall findet eine Analyse des Übungslaufes gemeinsam mit allen Beteiligten in einer Nachbesprechung statt.

In Abbildung 4 wird die Benutzerschnittstelle zur Nachbetrachtung gezeigt, in der die wichtigsten Informationen und Indikatoren des Übungslaufes dargestellt sind, um einen Überblick über den Verlauf des einzelnen Übungslaufes zu bekommen.

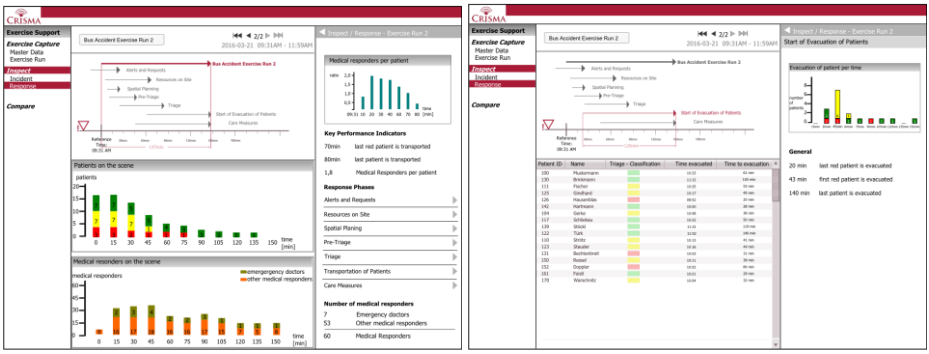


Abbildung 4: Benutzerschnittstelle zur Übungsnachbetrachtung als Übersicht (a) und zur Phase des Abtransports (b)

Anhand des Zeitstrahles ist die Navigation während der Nachbetrachtung möglich (Klick auf Einsatzphasen, vgl. Abbildung 4 (a)). Die Detailansichten der einzelnen Phasen (Vorsichtung, Sichtung, Abtransport etc.) werden so aufgerufen und können an passender Stelle der Übungsnachbesprechung gezeigt werden (vgl. Abbildung 4 (b)).

Alarme, Anforderungen und Ankunft von Ressourcen analysieren: Die Analyse der Ressourcenanforderungen bietet eine Übersicht über den Fluss der Rettungsmittel und Einsatzkräfte während des Einsatzgeschehenes. Beteiligte Fahrzeuge werden aufgeschlüsselt nach Fahrzeugtyp und Ankunftszeit dargestellt.

Start des Abtransportes der Patienten untersuchen: Wesentlicher Indikator für das erfolgreiche Absolvieren eines MANV-Einsatzes ist der Abtransport der Patienten. Die Benutzerschnittstelle erlaubt es dem Übungsleiter, die erfolgte Reihenfolge des Abtransports sowie die Entwicklung der in der Patientenablage anwesenden Patienten über die Zeit zu analysieren (vgl. Abbildung 4 (b)). .

Versorgungsmaßnahmen untersuchen: Die erfolgte Versorgung und Behandlung der Patienten und deren subjektive Bewertung kann dazu dienen, zu analysieren, welche Maßnahmen wiederholt schlecht bewertet bzw. vergessen wurden. Wichtig hierbei ist, dass in der Besprechung nicht auf Verfehlungen einzelner Einsatzkräfte eingegangen wird, sondern die Leistung der Einsatzeinheit als Ganzes im Vordergrund steht und ggf. mit gezielten Fortbildungsmaßnahmen verbessert werden kann. Die Übersicht der Patienten und die Beurteilung der katastrophenmedizinischen Behandlungen werden ähnlich wie in Abbildung 3 (b) in einer Tabelle zusammengefasst. Die Detailgenauigkeit ragt soweit, dass zu jeder Mime die genaue Uhrzeit und die behandelnde medizinische Einsatzkraft zugeordnet werden kann.

Gefahrenanalyse und Raumplanung analysieren: Zur schnellen Beurteilung der Gesamtsituation ist der zentrale Bestandteil der Analyseansicht für die Raumplanung eine Karte, auf der sich alle taktischen Flächen befinden. Eine chronologische Ordnung nach Erstellung der einzelnen taktischen Flächen wie auch die Entfernungen zwischen den einzelnen Flächen wird generiert.

Vorsichtung und Sichtung analysieren: Die Vorsichtung – ausgeführt durch medizinisches Fachpersonal – sowie die notärztliche Sichtung können ebenfalls im Zeitverlauf analysiert werden. In der Benutzerschnittstelle werden die Mimen in einer Tabelle mit ihrer korrekten Kategorisierung und der jeweils erfolgten Vorsichtung oder Sichtung aufgelistet. Wenn zwischen korrekter und erfolgter Klassifizierung ein Unterschied besteht, wird dieser hervorgehoben und auch im Zeitverlauf als Fehlsichtung angezeigt.

Übungsverläufe vergleichen: Übungsläufe können verglichen werden, wenn nahezu identische Randbedingungen bezüglich geographischer Lage und verfügbarer Ressourcen zugrunde liegen. Am interessantesten ist ein Vergleich von Übungsläufen für solche, die am selben Tag und mit den gleichen Übungsbeteiligten stattgefunden haben. Dadurch kann ersichtlich werden, welche Entscheidungen der Übenden einen starken Einfluss auf das Ergebnis bzw. die errechneten Indikatoren hatte.

5 Konzeptuelles Datenschema und Systemsicht

Zur Analyse der Zusammenhänge des Übungsgeschehens sowie zur späteren Umsetzung in eine Informationsarchitektur wurden, basierend auf der Nutzerperspektive sowie auf abstrahierten Arbeitsabläufen der an einer Übung beteiligten Rettungsinstanzen, Klassifizierungen und Zusammenhänge als UML-Klassendiagramme erhoben.

5.1 Mimen- und Helfer-Daten und Vorbereitung des Übungslaufes

Bevor ein Übungsdurchgang gestartet werden kann, weist der für die Datenerfassung zuständige Übungsdatenverantwortliche den Mimen ihre Patientenidentitäten zu. Weiterhin erhalten die an der Übung teilnehmenden Einsatzkräfte Einsatzkräfte-IDs bzw. Notarzt-IDs. Insbesondere für die Zuordnung der Mimen zu virtuellen Patientenidentitäten ist es denkbar, dass eine reale Person zeitlich aufeinanderfolgend in die Rolle von mehreren virtuellen Patienten schlüpft. Dies ist beispielsweise dann ein akkurates Mittel, wenn die Anzahl an für die Übung benötigten virtuellen Patienten die Anzahl der Mimen übersteigt. Ein Übungsdurchlauf wird dabei definiert durch ein definiertes Schadensszenario und eine Menge von Zuweisungen zwischen Rollen und realen Personen.

5.2 Daten des Übungslaufes

Sobald ein Übungsdurchlauf gestartet wurde, erfasst der Übungsdatenverantwortliche die Positionen der einzelnen taktischen Bereiche und den Zeitpunkt derer Etablierung sowie die Ankunftszeiten von Rettungsfahrzeugen am Unfallort (i.d. Regel aus Einsatzdokumentation). Sobald die Mimen am virtuellen Krankenhaus eintreffen, können Informationen von der Versorgungsbewertungskarte in das System eingegeben werden, wie Behandlungsmaßnahmen, Behandlungsdauer und Identität der behandelnden Helfer und Bewertung (vgl. Abbildung 5).

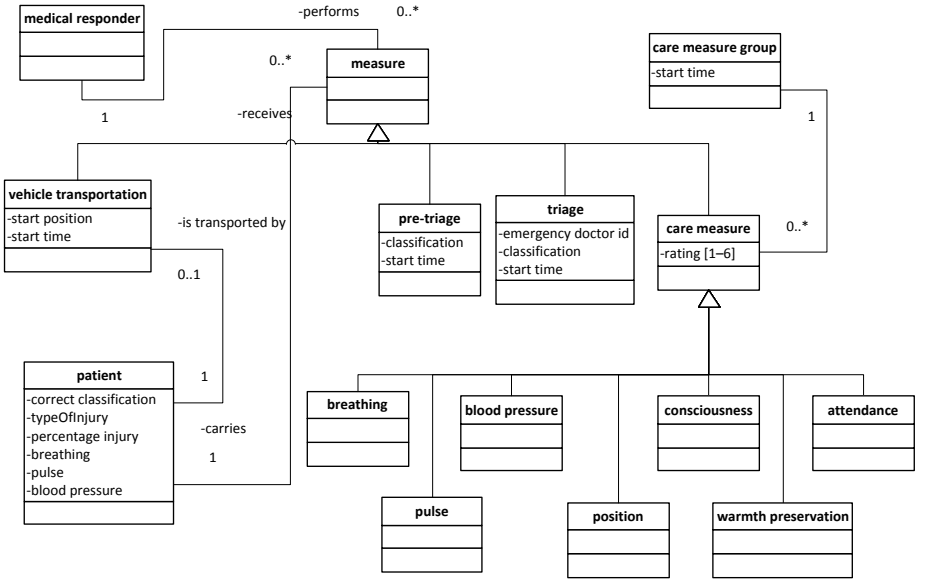


Abbildung 5: Datenschema für den Rettungsprozess

Der Patient wird ggf. an Ort und Stelle medizinisch versorgt und mit einem Rettungswagen oder einem Helikopter abtransportiert. Der Leitende Notarzt (LNA) sowie der organisatorische Leiter Rettungsdienst (OrgL) vor Ort sind verantwortlich für die Koordination von Aufgaben der medizinischen Helfer, angemessene medizinische Versorgungsmaßnahmen und Behandlungen sowie die Koordination der Rettungstransporte.

6 Zusammenfassung und Ausblick

Ein neues, für Praktiker einfach durchzuführendes Trainingskonzept mit Übungsunterstützungssystem wurde ausgearbeitet, welches die zuvor genannten Anforderungen erfüllt. Die Anforderung der Vergleichbarkeit für schnelle qualitative Rückschlüsse unter Experten wird durch die Indikatoren des Auswertungsschemas erreicht. Die Nutzung eines Bewertungsbogens durch die Mimen (organisatorisch) und die nachvollziehbare Speicherung der Ergebnisse (technisch) adressiert die Anforderungen Genauigkeit und Reproduzierbarkeit. Die bedarfsorientierte Einblendung von Detaildaten bei der Nachbetrachtung durch den Übungsleiter ermöglicht Fairness gegenüber den Einsatzkräften. Indem die Datenerhebung durch Mimen erfolgt, bleibt dabei das Trainingskonzept unabhängig von lokalen Einsatzkonzepten anwendbar. Zudem erhebt es Daten zur Patientenversorgung. Aktuell befinden sich Teile des webbasierten Übungsunterstützungssystems in einer Testphase, in der das System bei ausgewählten Übungen zum Einsatz kommt, wobei die Optimierung hinsichtlich der Anforderungen Nützlichkeit und Durchführbarkeit im Fokus steht.

Zusätzlich zu den im Projekt realisierten Funktionen, wurde im Gespräch mit Endnutzern der Wunsch nach einer simulationsgestützten Prognose laut, der den Patientenverbleib und Gesundheitszustand nach dem Abtransport bzw. bei dessen Ankunft in der Klinik bei der Nachbetrachtung ermöglicht. Als weiterer Ausblick könnte das System eine Berechnung von realistischen Anfahrtszeiten als individuelle Startzeit zur Einspielung an der Unfallstelle pro Einsatzfahrzeug vornehmen. Dies könnte auf Basis der Geoposition und Ausstattung aktueller Rettungsinfrastruktur eines Land-/Stadtkreises erfolgen. Realistischere Ankunftszeiten für Rettungsmittel während des Übungslaufes könnten mittels Simulationsmodellen erreicht werden. Zudem können die Mimen statt des Bewertungsbogens mit einem Sensor ausgestattet werden, der die Übermittlung der Zeiten (bspw. Zeit der Vorsichtung) per Knopfdruck erlaubt, somit würde die Fehlerquelle, die durch die manuelle Mitschrift der Mimen sowie die anschließende Erfassung durch den Übungsdatenverantwortlichen besteht, reduziert.

Literaturverzeichnis

Brauner, F./ Stiehl, M./ Lechleuthner, L./ Mudimu, O.A. (2014): Evaluation von Übungen des Massenanfalls von Verletzten (MANV), in: Notfall + Rettungsmedizin, 1/2014, Springer Verlag, Berlin/ Heidelberg.

Geuther, A. (2013): Von der Übung zum Training: Neues Trainingskonzept des BRK für Großschadensfälle, in: Im Einsatz, 20. Jahrgang, Februar 2013, Stumpf + Kossendey Verlagsgesellschaft, Edewecht.

4C Strategies (2014): 4C Strategies Sveavägen 13, Box 7637 103 94 Stockholm, https://www.4cstrategies.com/exonaut-products/training-and-exercise-manager#.U1YrD1dLerc (aufgerufen 2014).

Koch, R./ Marterer, R. (2014): IT-gestützte Möglichkeiten für die Auswertung/Visualisierung von Übungen, https://www.cik.uni-paderborn.de/fileadmin/mb/cik/Kurzbeschreibung_Projekte/093_Kurzbeschreibung_RescueLab_20130129_de.pdf

Max, M./ Sautter, J. (2013): Analysis of a German first responder exercise: Requirements for exercise-support and simulation, Comes, T. (Ed.), ISCRAM 2013, 10th International Conference on Information Systems for Crisis Response and Management. Proceedings : Baden-Baden, Germany, 12-15 May 2013, ISBN: 978-3-923704-80-4, S.923-924, www.iscramlive.org/ISCRAM2013/files/245.pdf.

Acknowledgements: Das Projekt CRISMA, welches durch das siebte Rahmenprogramm der Europäischen Union FP7/2007-2013 (grant agreement no. 284552) kofinanziert wird, bearbeitet das Modellieren und Simulieren von Krisenszenarien zur Unterstützung von Planung und Training im präventiven Krisenmanagement. Das CRISMA-Konsortium besteht aus siebzehn Partnern und wird geleite von dem Technischen Forschungszentrum Finnlands VTT (www.crismaproject.eu).

Entwicklung eines SOA-basierten und anpassbaren Bewertungsdienstes für Inhalte aus sozialen Medien

Christian Reuter, Michael Ritzkatis, Thomas Ludwig

Institut für Wirtschaftsinformatik, Universität Siegen
Hölderlinstr. 3, 57076 Siegen
christian.reuter@uni-siegen.de; michael.ritzkatis@student.uni-siegen.de;
thomas.ludwig@uni-siegen.de

Abstract: Dieser Beitrag soll aufzeigen, wie ein anpassbarer Bewertungsdienst die Nutzung bürgergenerierter Inhalte aus sozialen Medien unterstützen kann. Dabei soll insbesondere geklärt werden, wie dieser gestaltet werden kann und wie Nutzer die Qualitätskriterien angemessen artikulieren können. Nach einer Darstellung von Grundlagen und verwandten Arbeiten wird anhand einer empirischen Vorstudie der Umgang von Behörden und Organisationen mit Sicherheitsaufgaben (BOS) mit bürgergenerierten Informationen betrachtet. Basierend auf den dort gewonnen Erkenntnissen wurde ein service-orientierter Bewertungsdienst entwickelt und in eine Anwendung integriert, welche so den Zugang zu bürgergenerierten Informationen aus verschiedenen sozialen Medien inklusive einer anpassbaren Qualitätsbewertung ermöglicht. Eine abschließende Evaluation illustriert deren mögliche Anwendung in der Praxis.

1 Einleitung

Bürgergenerierte Inhalte spielen in Zeiten weiter Verbreitung interaktiver Technologien und sozialer Medien eine immer größere Rolle. Bürgergenerierte Inhalte sind dabei laut der Organisation für wirtschaftliche Zusammenarbeit und Entwicklung [Orga07] im Internet publizierte und eine gewisse kreative Eigenleistung erfordernde Inhalte, die außerhalb professioneller Routinen und Praktiken erstellt werden. In Schadenslagen, wie zuletzt dem Hurrikan Sandy 2012 oder auch dem mitteleuropäischen Hochwasser 2013, wurden soziale Medien von Betroffenen und freiwilligen Helfern aus der Bevölkerung genutzt, um sich untereinander auszutauschen und zu koordinieren [KaRe14]. Da das Engagement der Bürger weitgehend ungesteuert ist und Inhalte somit nicht zwingend strukturiert entstehen, muss die Vielzahl der entstehenden Daten analysiert werden. Zur Analyse sind geeignete Bewertungsmethoden unabdingbar, wobei eine konsistente Bewertung der Informationsqualität komplex sein kann [FrPK10]. Gerade wenn aus einer Vielzahl von Informationsquellen und -formaten eine Auswahl getroffen werden muss, ist es hilfreich, wenn verschiedene Qualitätskriterien die Bewertung erleichtern können. Dieser Beitrag adressiert die Herausforderungen der Integration bürgergenerierter Inhalte [ReMP12], insbesondere die Bewertung von Informationen aus sozialen Medien. Er

stellt die Ergebnisse einer Design-Fallstudie [Wulf09] dar, die basierend auf einer empirischen Vorstudie [ReRi14] (Kap. 2) und anhand der Entwicklung von IT (Kap. 3 und 4) sowie deren Evaluation (Kap. 5) untersucht, wie unterschiedliche Akteure durch die Verwendung eines anpassbaren Dienstes zur Bewertung bürgergenerierter Informationen aus sozialen Medien unterstützt werden können.

2 Verwandte Arbeiten: Informationsqualität

Es existieren verschiedene Definitionen für Informationsqualität (IQ), deren Gemeinsamkeit ist, dass eine Information dann eine hohe Qualität besitzt, wenn sie den subjektiven Ansprüchen des Adressaten entspricht. Dieses Maß lässt sich anhand verschiedener Kriterien messen. Während [Wolf99] beispielsweise Zeit, Ort, Menge, Form sowie Inhalt, Entscheidungsrelevanz und Wirtschaftlichkeit als Maßstab berücksichtigt, verwenden [ZeHe96] Übertragbarkeit, Gebrauchstauglichkeit, Wartbarkeit, Effizienz, Reliabilität und Funktionalität. Basierend auf 179 Attributen entwickelten [WaSG96] ein Framework mit 15 Dimensionen in vier Kategorien, die für Datennutzer von Relevanz sind.

1. **Intrinsische IQ:** Das Kriterium Präzision gibt an, ob die Information den beschriebenen Inhalt hinreichend genau wiedergibt [FrPK10]. Objektivität drückt aus, ob der Sachverhalt unvoreingenommen von persönlichen Eindrücken und Wertungen widerspiegelt wird; die Glaubwürdigkeit, ob die Informationen aus der Sicht des Empfängers plausibel und verwertbar sind. Dieses Maß entspricht nicht zwingend der objektiven Glaubwürdigkeit [FrPK10], die durch die Reputation ausgedrückt wird.
2. **Kontextabhängige IQ:** Neben den obigen eher allgemeinen Attributen werden solche beschrieben, die eine Information auf ihre Gültigkeit im jeweiligen Kontext überprüfen. Eine Information muss für den aktuellen Kontext einen Mehrwert liefern [ChRe13]. Dieser ist eng an die Aktualität gekoppelt; so ist ein veralteter Plan über Rettungswege nicht geeignet. Zusätzlich sollte sie ein hohes Maß an Relevanz für den Empfänger besitzen. Damit Informationen bewertet werden können, muss eine ausreichende Datenmenge vorhanden sein, die eine Einordnung in einen Kontext überhaupt erst ermöglicht [BZAZ09]. Gleichzeitig sollten Informationen vollständig sein, um die geschilderte Situation umfassend darzustellen [FrPK10].
3. **Begriffliche IQ:** Zusätzlich werden Informationen anhand des Informationsformats und ihrer Bedeutung klassifiziert. Am wichtigsten ist hierbei, dass diese in ihrer dargestellten Form für den Empfänger einfach zu verstehen und darüber hinaus interpretierbar sind. Somit kann der Empfänger die Information eigenständig in den entsprechenden Kontext einordnen. Für das Informationsformat gelten die Konsistenz des Formats und prägnante Darstellung, durch welche eine Eindeutigkeit von Abkürzungen und Symbolen sowie der Information selbst gewährleistet wird [WaSG96].
4. **Zugangs-IQ:** Die vierte Kategorie umfasst Eigenschaften der Sicherheit. Auf Informationen muss zugegriffen werden können, um sie verwenden zu können. Der Zugriff sollte aber nur unter Einhaltung der Sicherheitsanforderungen geschehen, sodass eine Person nur berechtigte Informationen benutzen kann [WaSG96].

Wie dargestellt, lässt sich Informationsqualität aus verschiedenen Perspektiven bestimmen. Im Folgenden wird sie als hoch betrachtet, wenn die Informationen den subjektiven

Ansprüchen des Empfängers genügen. Daraus motiviert sich, dass der Empfänger der Information selbst bestimmen muss, welche Dimensionen für ihn relevant sind.

Im Laufe der letzten Jahre beschäftigten sich verschiedene Vorhaben mit der Umsetzung obiger und weiterer Kriterien zur Bewertung und Integration bürgergenerierter Informationen in Schadenslagen. Eine Herangehensweise ist die Beschränkung der Informationsmenge. Dazu existieren Verfahren, die nicht auf der freien Informationseingabe der Benutzer basieren, sondern den Benutzern entweder eine spezielle Eingabesyntax vorschreiben [StPa11] oder aber konkrete Informationsanfragen an einzelne Benutzer stellen [DARO13]. Durch diese Einschränkung besteht allerdings die Gefahr, dass möglicherweise relevante Informationen verloren gehen. Andere Ansätze betrachten ebenfalls nur Teilmengen der Informationsbasis: [TVSP12] berücksichtigen lediglich die Daten eines sozialen Netzwerks. [ViBo12] setzen auf eine breitere Informationsbasis, ermöglichen jedoch nur BOS Zugang zu ihren Informationen. Gängige Verfahren zur Klassifizierung von Informationen reichen derzeit von maschinellem Lernen auf Basis von Wörterbüchern oder Sprachklassifizierungen [CACI13] über Text Mining [ZMTW13] und Bestimmung des Gefühlsausdrucks der Information durch den normalisierten „Sentiment Value“, sowie Einteilung der Nachrichten in verschiedene Klassen mit Bayesschen Netzen [NaSt12] und Bestimmung von Ähnlichkeiten anhand regulärer Ausdrücke [ZiBü12] bis hin zu Bestimmung des Informationsgehalts der Nachricht durch Entropie [VVCP11]. Einige Arbeiten beschränken sich auf die Ermittlung von Metainformationen, mit denen eine bessere Bewertung der Informationslieferanten möglich sein kann [StMP12]. Dennoch weist zum aktuellen Zeitpunkt jede Bewertungsmethode Fehlklassifizierungen auf. Weiterhin existieren zur Selektion und Nutzung bürgergenerierter Informationen einige Ansätze, die allerdings keine umfassende Qualitätsbewertung unterstützen [ReRi14].

3 Vorstudie: Nutzung bürgergenerierter Inhalte

Um ein tiefgehendes Verständnis über die Rolle bürgergenerierter Informationen in Schadenslagen zu bekommen, wurde das Datenmaterial einer empirischen Studie über Informations- und Kommunikationspraktiken von BOS (hier: Feuerwehr, Polizei) im Falle einer Großschadenslage in zwei geographisch und strukturell verschiedenen Kreisen Nordrhein-Westfalens verwendet. Die Ergebnisse der Vorstudie werden in [ReRi14] ausführlich behandelt und für diesen Beitrag nur zusammengefasst dargestellt.

Es wurden verschiedene qualitative Methoden wie Dokumentenanalysen, 4 Beobachtungen, 26 teilstrukturierte Interviews und 4 Gruppendiskussionen durchgeführt und ausgewertet. Ziel der Dokumentenanalyse war die Untersuchung offizieller Arbeitspraktiken mit Hilfe von Gesetzen und Richtlinien. Die Beobachtungen dienten dazu, Erkenntnisse über die tatsächliche praktische Arbeit zu erhalten. Diese wurden in der Leitstelle während eines normalen Arbeitstages (9h), im Krisenstab und der Einsatzleitung einer Krisenübung (4h) und eines Großereignisses (6h) durchgeführt. Die semi-strukturierten Interviews (n=22 je 1–2h; bezeichnet als I01-22) dienten der Reflexion der Arbeitspraxis, der Identifikation möglicher IT-Unterstützung und interorganisationaler Informationsbedarfe und wurden mit Einsatzkräften verschiedener Ebenen durchgeführt. Sie fo-

kussierten im ersten Teil die Arbeitstätigkeiten in normalen Lagen, im zweiten Teil potentielle Aufgaben im vorgelegten Szenario und im dritten Teil Informations- und Kommunikationstechnologien. Ergänzende Interviews (n=5 je 1h; bezeichnet als IM01-05) fokussierten mobile Kollaborationspraktiken. Die Gruppendiskussionen (n=4 je 4h, bezeichnet als W1-4) dienten der Erfassung kommunikativer Arbeitspraxen. Bei der Auswertung des vorliegenden Datenmaterials wurde sich an der Grounded Theory [Stra04] orientiert. Dazu wurden die Transskripte offen kodiert und die Aussagen in Textbausteine und spätere Kategorien aufgeteilt und Herausforderungen hergeleitet:

Plattformübergreifende Nutzung sozialer Medien: In Schadenslagen müssen sich Einsatzkräfte einen Überblick über die Situation verschaffen, um Tendenzen sowie Entwicklungen der Lage zu erkennen (I07, I18). Der Versuch, genügend Informationen vorzuhalten, scheitert dabei oftmals aufgrund der Ungewissheit der Situation (I01). In einigen Bereichen, wie der Beobachtung von Großevents oder Demonstrationen, werden externe Quellen mit dem Ziel der Erweiterung der Informationsbasis genutzt (W3, I02, I03). Hierzu gehören u.a. soziale Medien: „Diese elektronischen Medien, [...] werden immer mehr Einfluss gewinnen, auch auf die Art der Einsatzbearbeitung [...] und auf die Art der Informationserhebung“ (I02).

Automatisierte Verarbeitung: Gerade in nicht planbaren Einsätzen entstehen schnell Zeit- und Personalengpässe, welche die Notwendigkeit computerunterstützter Verarbeitung deutlich machen: „Wenn wir oben nur mit drei oder vier Mann sitzen und führen einen Einsatz, dann habe ich keine Zeit, mich noch mit Fremden zu beschäftigen“ (I2). Eine Vielzahl der Informationen soll idealerweise automatisiert und nicht manuell bearbeitet werden, um eine Informationsüberflutung der Einsatzkräfte, die diese paralysiert und weitere Handlungen behindert, zu vermeiden (I02, I03). Denn *„die Gefahr besteht*, mit Informationen überfrachtet zu werden und dass man dann letztendlich gar nichts *versteht, weil man so viel Input bekommt“* (I03).

Flexible und anpassbare Qualitätsbewertung: Insbesondere externe Quellen sind für Fehlinformationen anfällig (I14, I15) und müssen deshalb überprüft werden (I15): „Da zeigt sich natürlich immer wieder, dass wir da aufs Glatteis geführt werden sollen [...] und wir müssen die Informationen bewerten für uns“ (I02). Trotzdem scheitert der Einsatz von Bürgerinformationen aus sozialen Netzwerken besonders an der Bewertung der Informationen durch Einsatzkräfte, denn *„da ist einfach irgend so ein Flaschenhals, über den wir nicht drüber kommen“* (I02). Im Allgemeinen stellt die Bewertung von Informationen eine große Herausforderung dar, da sowohl die Herkunft der Informationen als auch die zur Bewertung notwendigen Metainformationen meist unbekannt sind (W3, W4). Diese Überprüfung benötigt Ressourcen, damit die Einsatzkraft nicht einer Informationsflut ausgesetzt wird: „Dann müssen sie noch 3-4 Filter-Funktionen davor schalten, damit Sie nur die wichtigen Informationen ankommen lassen“ (I01). Insbesondere muss sichergestellt sein, dass die gefilterten Informationen einen Mehrwert für die Einsatzkräfte darstellen (I02, I03). Erschwerend kommt hinzu, dass nicht im Vorhinein bestimmt werden kann, welche Hierarchieebene welche Informationen benötigt (IM01). Daher muss eine Flexibilität der Bewertungskriterien gegeben sein (I19). Letztlich muss die Einsatzkraft die Verantwortung über die Handlungen übernehmen und entscheiden, ob eine Information verwendet wird oder nicht (I15).

4 Konzept: Bewertung von Inhalten aus soziale Medien

Verwandte Arbeiten sowie die empirische Studie haben gezeigt, dass die Bewertung der Masseninformationen und die Extraktion relevanter Informationen eine große Herausforderung darstellen. Es scheint nachvollziehbar, dass verschiedene Situationen verschiedene Bewertungsmethoden erfordern und somit die Möglichkeit der Kombination derselben zur Verbesserung der Qualitätsbewertung beitragen könnten. Verschiedene Bewertungsmethoden wurden bereits in Kapitel 2 aufgezeigt. In Kapitel 3 wurde deutlich, dass unterschiedliche Bewertungsmethoden die subjektive Qualitätsbewertung in verschiedenen Situationen unterstützen können. Der Bewertungsdienst für soziale Medien ermöglicht eine Bewertung der Nachricht nach 15 Bewertungsmethoden (Tabelle 1), die sich an den oben dargestellten Kriterien orientieren und sich hinsichtlich der technischen Umsetzung in vier Kategorien unterteilen lassen:

A Die **Bewertung der Metadaten** beinhaltet fünf Bewertungsmethoden (Autorenhäufigkeit, zeitliche Nähe, örtliche Nähe, Anzahl Follower/Likes, Anzahl Metadaten), bei denen entweder die Abweichung zu den eingegebenen Suchkriterien oder aber das absolute Auftreten durch die Ermittlung der Differenz bestimmt wird.
B Die **Bewertung auf Basis des Inhalts** stellt zwei Bewertungsmethoden (Suchbegriffhäufigkeit, Stopwords) zur Verfügung, die das Vorkommen bestimmter Wörter anhand einer Wortliste bzw. einer Synonymliste ermitteln.
C Die **Bewertung auf Basis der Klassifikation der Nachricht** nutzt sechs Bewertungsmethoden (Sentiment Analyse, Fear Factor, Happiness Factor, Named Entity Recognition, Emoticon, Slang), die mittels Wortlisten das Auftreten von Wörtern untersuchen und die Information so in verschiedene Klassen einteilen.
D Die **Bewertung auf Basis wissenschaftlicher Methoden** verwendet zwei Bewertungsmethoden (Shannonscher Informationsgehalt, Term Frequency – Inverse Document Frequency).

Von diesen Methoden kann der Benutzer wählen, welche er in welcher Gewichtung verwenden möchte. Somit kann eine subjektive Informationsqualität bestimmt werden. Darüber hinaus ermöglichen diese Auswahl und die verschiedenen Kategorien einen Einsatz des Bewertungsdienstes in einer Vielzahl unterschiedlicher Szenarien. Generell gilt: Zuerst werden die einzelnen Nachrichten hinsichtlich der jeweiligen Methode absolut bewertet. Anschließend wird durch Suche des größten und kleinsten absoluten Wertes der relative Wert einer jeden Nachricht bestimmt. Die Nachricht mit dem größten absoluten Wert erhält den Score 1.0 (100%), diejenige mit dem kleinsten den Score 0.0 (0%). Im Anschluss werden Einzelscores gewichtet und ein Gesamt-Score gebildet.

Nr.	Methode / Kriterium	Beschreibung
A	Bewertung der Metadaten	
1	Autoren-häufigkeit (Reputation)	Anzahl von Nachrichten desselben Autors im Nachrichtenset. Je mehr Nachrichten ein Autor verfasst, desto mehr Wissen über die Situation wird unterstellt.
2	Zeitliche Nähe (Aktualität)	Zeitliche Nähe der Nachrichten mit dem Mittelpunkt des Suchzeitraums. Je näher die Nachricht am gesuchten Zeitpunkt ist, desto sicherer kann

		man sein, dass es sich um eine neue Information handelt.
3	Örtliche Nähe	Entfernung zwischen dem Erstellungsort einer Nachricht zu dem übergebenen Ort. Je geringer die Entfernung ist, desto höher ist die Wahrscheinlichkeit, dass die Nachricht die aktuelle Schadenslage thematisiert.
4	Follower/Likes (Glaubwürdigk.)	Mit steigender Anzahl an Likes bzw. Followern einer Nachricht / eines Autors wird ein steigendes Maß an Glaubwürdigkeit unterstellt.
5	Metadaten (Bilder/Links) (Ausreichende Datenmenge)	Gerade wenn eine bessere Betrachtung der Schadenslage fokussiert werden soll, ist es neben textuellen Informationen hilfreich, ein Bild oder andere mediale Informationen zu verwenden. Daher misst dieses Bewertungskriterium die Anzahl der Metadaten.
B	Bewertung auf Basis des Inhalts	
6	Suchbegriff-häufigkeit (Interpretierbarkeit)	Es wird sichergestellt, dass die Nachricht nicht nur zufällig den Suchbegriff enthält, sondern auch tatsächlich den Sachverhalt thematisiert. Dazu wird die Nachricht auch nach Synonymen durchsucht.
7	Stopwords	Die Anzahl an sogenannten Stopwords (bspw: „also“) erhöht nicht den Aussagegehalt der Nachricht, da diese Wörter keine Information liefern. Deshalb steigt mit sinkender Anzahl an Stopwords in der Nachricht ihr Nutzen.
C	Bewertung auf Basis der Klassifikation der Nachricht	
8	Sentiment Analyse (Objektivität)	Die Nachricht wird hinsichtlich ihrer emotionalen Eigenschaft bewertet. Emotionale Nachrichten können den Sinn verfremden, insbesondere wenn sie durch Angst motiviert sind.
9	Fear Factor	Bemisst den Grad des Ausdrucks von Angst in der Nachricht durch die Häufigkeit von Wörtern, die mit der Thematik Angst verbunden sind.
10	Happiness Factor	Bemisst hingegen den Ausdrucksgrad von Freude in der Nachricht durch die Häufigkeit von Wörtern, die mit der Thematik Freude verbunden sind.
11	Named Entity Recognition (NER)	Anzahl an Entitäten in der Nachricht. Eine Entität deutet dabei auf die Verknüpfung des Inhalts der Informationen zu einer anderen Informationsquelle hin. Daher steigt die Qualität der Information mit der Anzahl der Entitäten einer Nachricht.
12	Emoticon	Anzahl an Emoticons, wie Smileys, die ironisch formulierte Nachrichten kennzeichnen. Zum Verständnis werden die Emoticons durch ihre Bedeutung in der Nachricht ersetzt. Daraus motiviert sich, dass Informationen mit wenig Emoticons einen eindeutigeren Inhalt haben können.
13	Slang	Anzahl an Abkürzungen, dessen Einsatz durch Zeichenbeschränkungen ein beliebtes Instrument ist, um mehr Inhalt zu vermitteln. Wie Emoticons erschweren Abkürzungen eine automatisierte Verarbeitung.
D	Bewertung auf Basis wissenschaftlicher Methoden	
14	tf-Idf (term frequency – Inverse document frequency)	Das Auftreten der einzelnen Suchbegriffe (term frequency) mit der Auftrittshäufigkeit in der gesamten Nachrichtenanzahl (Inverse document frequency). Die Methode ist bei der Verwendung von mehreren Suchbegriffen hilfreich, da das Auftreten eines Teils der Suchphrase, welches in wenigen Dokumenten häufig auftritt, höher gewichtet wird, als wenn es in vielen Dokumenten absolut nicht so häufig vorkommt.
15	Informations-gehalt	Shannonscher Informationsbegriff. Je größer der Informationsgehalt einer Nachricht ist, desto mehr Informationen kann sie transportieren.

Tabelle 1: Bewertungsmethoden

5 Implementierung eines SOA-basierten Bewertungsdienstes

Der eigentliche Bewertungsdienst ist nach dem webbasierten service-orientierten Architekturmodell (SOA) realisiert. Durch dieses Architekturmodell ist es möglich, die Bewertung zentral zur Verfügung zu stellen und durch die Bereitstellung der Ergebnisse inklusive der Originaldaten im Format JSON (JavaScript Object Notation) eine Einbindung in verschiedene Anwendungen zu ermöglichen. Der Aufruf der Schnittstelle funktioniert per HTTP-GET (RESTful-Webservice). Der URL werden die Parameter (durch "&" getrennt) angehangen. Die serverseitige Bewertung der Informationen soll die Rechenlast des Clients verringern.

Für die Extraktion der Daten aus den sozialen Netzwerken werden APIs der jeweiligen Netzwerkbetreiber verwendet [ReSc14]. Insbesondere werden im Rahmen dieses Beitrags die sozialen Netzwerke Twitter und Facebook, begründet durch die Existenz einer notwendigen API, betrachtet. Die jeweiligen Betreiber-APIs bieten eine Fülle von Möglichkeiten, Daten sowohl in das soziale Netzwerk zu importieren als auch zu exportieren. Für die Erkennung von Begriffen aus dem Inhalt einer Nachricht wird ein NER (Nr. 11) verwendet. Der Stanford NER (http://nlp.stanford.edu/software/CRF-NER.shtml) steht kostenlos als Java-Bibliothek zur Verfügung. Für den NER wurde der Korpus „deWac generalized classifier“ eingesetzt, weil dieser vor allem auf deutsche Nachrichten aus sozialen Netzwerken trainiert wurde. Für die Erstellung eines Bayes Classifiers (Nr. 8), mit dem die Informationen in verschiedene Klassen eingeteilt werden können, wurde die Bibliothek Classifier4J (http://classifier4j.sourceforge.net/) verwendet, da dieser mit Wortlisten trainiert werden kann. Die Synonymliste (Nr. 6) wird mittels des Open Thesaurus Webservices (http://www.openthesaurus.de/) erzeugt.

Zur Visualisierung der Informationen wird ein geographischer Bezug benötigt. Da ein Großteil der Informationen keine geographischen Metadaten beinhalten, müssen diese geocodiert werden. Der Gisgraphy Geocoder (http://www.gisgraphy.com/) lässt sich über einen REST (Representational State Transfer) Webservice verwenden und geokodiert die Ortsinformationen für beliebiges Kartenmaterial. Aus Geschwindigkeitsgründen wird jedoch für jeden bereits geokodierten Ort eine Liste angelegt, aus der die Koordinaten auch ohne Geokodierung ermittelt werden können. GSON (https://code.google.com/p/google-gson/) bietet die automatische Erstellung eines JSON Objektes anhand eines Java-Objektmodells und wird deshalb zur Konvertierung verwendet.

Der Bewertungs-Service wurde in eine Facebook-Anwendung, die ungebundene Helfer in ihrer Arbeit der Schadensbewältigung durch die medienübergreifende Informationssuche und die -generierung unterstützt, integriert [KaRe14]. Innerhalb dieser Anwendung besteht die Möglichkeit, Informationen nach verschiedenen Parametern zu suchen und wahlweise zusätzlich eine Qualitätsbewertung durchzuführen (Abbildung 1). Dabei bestimmt der Benutzer welche Bewertungskriterien er verwenden möchte. Durch diese Art der Benutzerintegration werden die Anforderungen der Vorstudie nach einer flexiblen und kontrollierbaren Qualitätsbewertung realisiert. Dazu werden bereitgestellte Schnittstellen des Bewertungsdienstes verwendet. Die Methoden Nr. 8, 11 und 13 wurden zur Reduzierung der Bewertungskriterien hier nicht integriert. Die Ergebnisse dieser

Informationssuche und Bewertung werden sowohl tabellarisch mit verschiedenen Metainformationen, wie dem Erfüllungsgrad einzelner Methoden, die als MouseOver-Text angezeigt werden, als auch visuell auf einer Lagekarte zur direkten Bestimmung der Nähe der Information zu dem Suchstandort, dargestellt (Abbildung 2). Somit kann der Benutzer aus verschiedenen Verfahren der Ergebnisbetrachtung wählen und erhöht damit die Flexibilität der Anwendung. Dabei stellt diese Benutzeroberfläche nur eine von vielen Varianten zur Benutzung der Qualitätsbewertung dar.

Kompakt betrachtet vereint der Bewertungsdienst folgende Funktionalitäten: Die bürgergenerierten Informationen können anhand verschiedener Methoden bewertet werden. Diese Bewertung findet nicht nur auf Basis von Metadaten, sondern vor allem auf Basis des Inhalts statt. Der Benutzer bestimmt dabei jederzeit, welche Methoden er wie gewichten möchte. Die subjektive Qualität einer Nachricht bestimmt sich aus Kombination der Bewertung aller Methoden. Der Bewertungsdienst weist ein hohes Maß an Flexibilität durch die Möglichkeit der einfachen Erweiterung der Quellen und Bewertungsmethoden auf. Durch die SOA-basierte Implementierung wird die Einbindung und Verwendung dieser Informationen in weitere Applikationen ermöglicht.

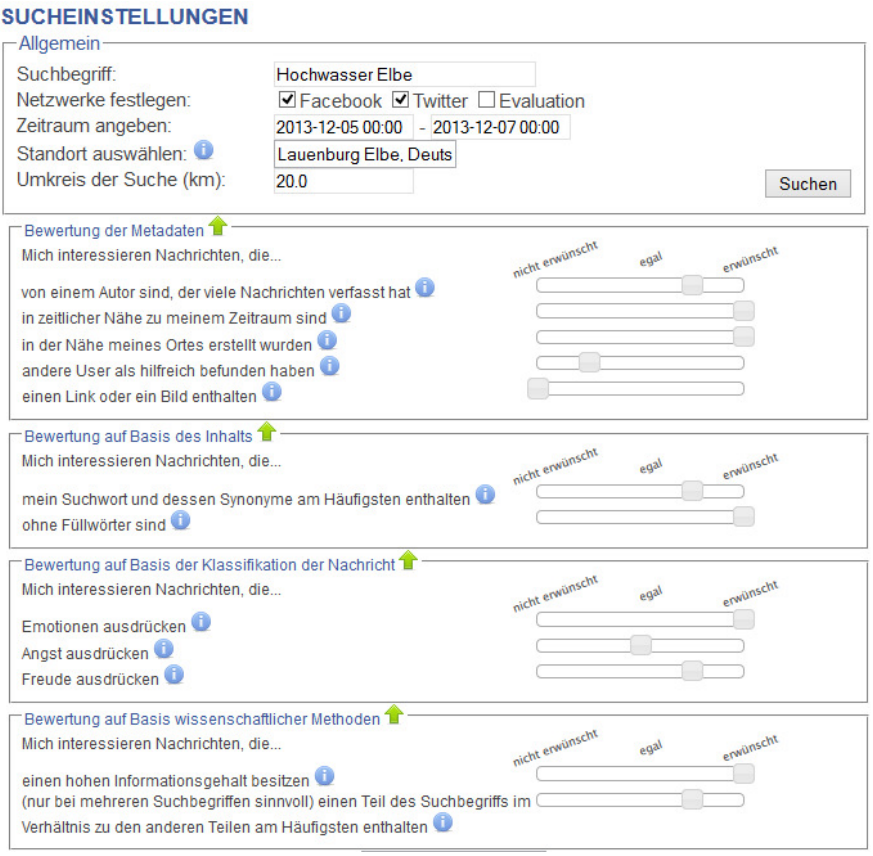


Abbildung 1: Informationssuche mithilfe von Schiebereglern

6 Evaluation: Anpassbare Bewertungsmechanismen

Um die Fragestellung zu beantworten, wie anpassbare Bewertungsmechanismen Benutzern geeignet zur Verfügung stellen können und wie Nutzer die Qualitätskriterien angemessen artikulieren können, wurde der Bewertungsdienst durch potenzielle Endnutzer evaluiert. Die Durchführung fand dabei anhand eines szenarienbasierten Walkthroughs mit anschließenden teilstrukturierten Interviews statt, um möglichst viele Erkenntnisse über das Potenzial des Dienstes und der Qualitätsbewertung bürgergenerierter Informationen zu erhalten. Die Teilnehmer sollten dabei gemäß Thinking Aloud [Niel93] ihre Gedanken während der Durchführung laut äußern, damit zugrundeliegende Gedankengänge und subjektive Eindrücke aufgenommen werden konnten. Die Evaluation dauerte jeweils ca. 45 Minuten und wurde mit insgesamt 19 Personen durchgeführt, wovon sechs Personen eine erste Version der Applikation bewertet haben. Sofern ein Teilnehmer kein ungebundener Helfer war, der sehr aktiv soziale Medien während des Hochwassers 2013 genutzt hat [KaRe14], wurde ihm zusätzlich eine Rollenbeschreibung vorgestellt, sodass er sich in die Lage eines Helfers versetzen konnte. Die Evaluation ergab, dass sich der Bewertungsdienst in vielen Bereichen als hilfreich erwies, jedoch noch weiterer Überarbeitungsbedarf besteht.

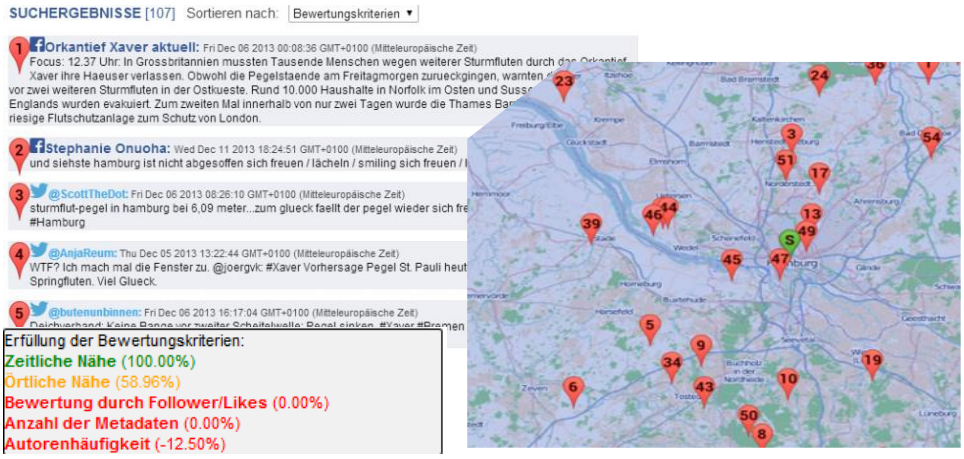


Abbildung 2: Suchergebnisse (links), Erfüllungsgrad (unten links) und Kartendarstellung (rechts)

Auswahl der Qualitätskriterien: Die Anzahl der Bewertungskriterien wirkte auf viele Nutzer im ersten Moment als zu hoch (E09, E07, E19). Dabei waren sich die Befragten darin einig, dass verschiedene Situationen verschiedene Bewertungskriterien erfordern (E12, E13, E08) und daher immer eine gewisse Anpassung der Kriterien an die Situation notwendig sei: *„Also es sind schon einige Kriterien, aber ich denke, dass ist auch wichtig, weil eben verschiedene Fragen verschiedene Suchbegriffe erfordern in dem Sinne“* (E13). Daher wurde auch der Vorschlag geäußert, dass die Bewertungskriterien *„ein* bisschen schon so eingestellt werden können, dass man vielleicht schneller zum Ziel *kommt“* (E12). Darüber hinaus wurden auch weitere Möglichkeiten, wie zum Beispiel die Schaffung einer Suche nach einem Unterbegriff (E19), als Ergänzung der Einstel-

lungsmöglichkeiten gewünscht. Die Evaluation hat gezeigt, dass die größte Herausforderung in der Identifikation passender Qualitätskriterien liegt. Während beispielsweise Aktualität von allen Testpersonen als wichtiges Kriterium verstanden wurde, konnten sich nur wenige Probanden unter der Koordinatenqualität etwas vorstellen. Jedoch konnte die Erklärung der Koordinatenqualität als Maß für die örtliche Nähe bereits Abhilfe schaffen. Ein Proband stellte zudem den Zusammenhang zwischen der Anzahl von Abonnenten eines Autors und seiner Reputation in Frage (E4).

Breite Informationsbasis und Informationsflut: Insbesondere die Möglichkeit, verschiedene Informationsquellen gleichzeitig mit einer Suche zu betrachten, wurde als Mehrwert dargestellt (E19, E15, E18): *„*Da wahnsinnig viel die öffentlichen Netzwerke genutzt werden, ist das also noch viel besser, die miteinander zu verknüpfen. Also es kann ja wahnsinnig gut helfen, den *Informationsbedarf zu decken"* (E17). Die Anzahl der Quellen sollte fortlaufend um weitere sinnvolle Quellen ergänzt werden, wobei nicht nur soziale Netzwerke, sondern beispielsweise auch Nachrichtenseiten betrachtet werden sollten. Des Weiteren zeigten die Befragten bei einer Informationssuche in den sozialen Netzwerken Angst vor einer Informationsflut (E16, E19). Diese konnte aber durch die Sortierung der Ergebnisse im Bewertungsdienst schnell beseitigt werden. Die meisten Benutzer wollten nicht die ganze Liste der Suchergebnisse durchgehen, sondern nur die ersten Einträge betrachten (E10, E11). Trotz allem sollten auch die weiteren Suchergebnisse für den Benutzer einsehbar bleiben, denn einige Szenarien erfordern ggf. die Sichtung weiterer Ergebnisse.

Automatische und manuelle Qualitätsbewertung notwendig: Die Qualitätsbewertung der Informationen findet automatisch statt. Dieser Automatismus wird durch die Benutzer dadurch akzeptiert, dass sie eine Kontrollmöglichkeit der Bewertung haben und zugleich nachvollziehen können, warum was wie bewertet wurde (E08; E11). Denn *„wie* immer, wenn Sachen bewertet werden, ne Bewertung von anderen ersetzt nicht die eige*ne Meinung"* (E10). Dabei wurde die generelle Möglichkeit der Kombination der Kriterien als Vorteil dargestellt: *„*Die Standardeinstellung ist egal. Das heißt, wenn ich nichts *mache, verändert das mein Suchergebnis nicht"* (E13). *„*Dadurch ist natürlich eine *vielfältige Kombination möglich, die ich auch für hilfreich erachte"* (E07). Um negativen Auswirkungen auf Handlungen entgegenzuwirken, sollte eine manuelle Nachbewertung umgesetzt werden, die die Korrektur bzw. Identifizierung fehlerhafter Einträge ermöglicht.

Verbesserung des Informationsflusses: Mit den aufgezeigten Möglichkeiten und Verbesserungsvorschlägen besteht das Potenzial, den Informationsfluss in Schadenslagen zu verbessern (E13, E19, E14). Dieses könnte insbesondere durch eine Klassifizierung der Schadenslagen und darauf aufbauend durch eine Voreinstellung der Gewichtungen realisiert werden. Wesentliche zeitliche und organisatorische Engpässe können vermieden werden, indem Informationen von Personen vor Ort (realer Helfer) oder im Netz (virtueller Helfer) [HeRe13] eingeholt werden (E07, E16): *„*Die Vorteile sind tatsächlich, dass ich Sachen schnell und zügig finden kann, [...] da man auch gezielt nach etwas suchen kann und es wirklich in den verschiedensten Plattformen ausgespuckt wird, so wie ich das möchte. Und ich kann anhand der Bewertungskriterien eben sehr gut ge*wichten" (E18; 00:19:23-*5).

7 Fazit

Schadenslagen stellen besondere Herausforderungen, wie situative Ungewissheit oder zeitliche sowie organisatorische Engpässe für Einsatzkräfte und ungebundene Helfer, dar. Um diese Herausforderungen bewältigen zu können, werden möglichst viele relevante Informationen, auch aus sozialen Medien, benötigt.

Dieser Beitrag hat gezeigt, wie die Qualitätsbewertung bürgergenerierter Informationen aus sozialen Medien unterstützt werden kann. Dabei liefert die Designfallstudie drei Beiträge zum Forschungsbereich: (1) eine Analyse des Umgangs mit bürgergenerierten Inhalten in Schadenslagen durch eine empirische Studie [ReRi14] (Kap. 3), (2) einen funktionsfähigen SOA-basierten und anpassbaren Bewertungsdienstes, der flexibel in verschiedene Anwendungen integriert werden kann (Kap. 4 und 5), sowie (3) die Ergebnisse dessen qualitativer Evaluation innerhalb einer Webanwendung (Kap. 6). In diesem Beitrag sollte insbesondere untersucht werden wie Nutzer die Qualitätskriterien angemessen artikulieren können; unser Ansatz hierzu wurde mithilfe der Evaluation überprüft. Zusammenfassend lässt sich sagen, dass die Flexibilität in Form der Anpassungsmöglichkeit von Sortierung, Quellenauswahl, Qualitätsbewertung hilfreich ist, da die Informationsbewertung immer subjektiv ist und somit auch diese Faktoren vom persönlichen Empfinden, der Erfahrung und der Lage abhängen.

Der durch diese Arbeit entwickelte Bewertungsdienst weist verschiedene Vorteile auf: Im Gegensatz zu [ViBo12] können so nicht nur Einsatzkräfte diese Informationen erhalten, sondern auch ungebundene Helfer können in ihrer Arbeit durch eine flexible Bewertung und freie Zugänglichkeit unterstützt werden. Eine sehr breite Informationsbasis wird durch die Verwendung verschiedener sozialer Netzwerke als Informationsquellen realisiert [TVSP12]. Im Vergleich zu [StPa11], die durch die Benutzung einer festen Syntax die Informationsmenge eingrenzen, werden hier aus den verschiedenen Quellen sämtliche Informationen betrachtet. Schließlich ermöglicht im Vergleich zu [VVCP11] erst die Verwendung von mehreren Bewertungskriterien, die miteinander kombiniert werden können, eine adäquate Qualitätsbewertung bürgergenerierter Informationen.

Gefördert im Rahmen des EU-FP7-Projekts „EmerGent“ (Fördernummer 608352).

Literaturverzeichnis

[BZAZ09] Bharosa, N., Zanten, B. Van, Appelman, J., & Zuurmond, A.: Identifying and confirming information and system quality requirements for multi-agency disaster management. In: Proc. ISCRAM. Gothenburg, Sweden, 2009

[CACI13] Chowdhury, S., Amer-Yahia, S., Castillo, C., Imran, M., & Asghar, M. R.: Tweet4act: Using Incident-Specific Profiles for Classifying Crisis-Related Messages. In: Proc. ISCRAM. Baden-Baden, Germany, 2013, pp. 834–839

[ChRe13] Christofzik, D. & Reuter, C. The Aggregation of Information Qualities in Collaborative Software. In: IJEV, vol. 5 (2013), Nr. 3, pp. 257–271

[DARO13] Díaz, P., Aedo, I., Romano, M., & Onorati, T. A framework to integrate large-scale participation in Disaster and Emergency Management. In: Workshop on Large-Scale Ideation and Deliberation Systems. Proc. C&T. München, 2013

[FrPK10] Friberg, T., Prödel, S., & Koch, R.: Analysis of information quality criteria in crisis situation as a characteristic of complex situations. In: Proc. IQ. Little Rock, 2010

[HeRe13] Heger, O. & Reuter, C.: IT-basierte Unterstützung virtueller und realer Selbsthilfegemeinschaften in Katastrophenlagen. In: Proc. WI. Leipzig, 2013, pp. 1861–1875

[KaRe14] Kaufhold, M.-A., & Reuter, C.: Vernetzte Selbsthilfe in Sozialen Medien beim Hochwasser 2013 in Deutschland. In: i-com vol. 13 (2014), Nr. 1, pp. 20–28

[NaSt12] Nagy, A. & Stamberger, J.: Crowd Sentiment Detection during Disasters and Crises. In: Proc. ISCRAM. Vancouver, Canada, Canada, 2012, pp. 1–9

[Niel93] Nielsen, J.: Usability Engineering. San Francisco: Morgan Kaufmann, 1993

[Orga07] Organisation for Economic Co-operation and Development: Participative Web: User-Created Content. URL http://www.oecd.org/internet/ieconomy/38393115.pdf

[ReMP12] Reuter, C.; Marx, A. & Pipek, V.: Crisis Management 2.0: Towards a Systematization of Social Software Use in Crisis Situations. In: IJISCRAM vol. 4 (2012), Nr. 1, pp. 1–16

[ReRi14] Reuter, C. & Ritzkatis, M.: Adaptierbare Bewertung bürgergenerierter Inhalte aus sozialen Medien. In: Mensch & Computer 2014. München: Oldenbourg, 2014

[ReSc14] Reuter, C. & Scholl, S.: Technical Limitations for Designing Applications for Social Media. In: Workshop-Proceedings Mensch & Computer 2014. München: Oldenbourg, 2014.

[StMP12] Starbird, K., Muzny, G., & Palen, L.: Learning from the Crowd: Collaborative Filtering Techniques for Identifying On-the-Ground Twitterers during Mass Disruptions. In: Proc. ISCRAM. Vancouver, Canada, 2012

[StPa11] Starbird, K., & Palen, L.: Voluntweeters: Self-Organizing by Digital Volunteers in Times of Crisis. In: Proc. CHI. Vancouver, Canada: ACM-Press, 2011

[Stra04] Strauss, A. L.: Methodologische Grundlagen der Grounded Theory. In: Strübing, J.; Schettler, B. (eds.): Methodologie interpretativer Sozialforschung: Klassische Grundlagentexte. Konstanz, Germany: UVK, 2004, pp. 427–453

[TVSP12] Terpstra, T., Vries, A. de, Stronkman, R., & Paradies, G. L.: Towards a realtime Twitter analysis during crises for operational crisis management. In: Proc. ISCRAM. Vancouver, 2012

[ViBo12] Vivacqua, A. S., & Borges, M. R. S.: Taking advantage of collective knowledge in emergency response systems. In: Journal of Network and Computer Applications vol. 35, Elsevier (2012), Nr. 1, pp. 189–198

[VVCP11] Verma, S., Vieweg, S., Corvey, W., & Palen, L.: Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency. In: Proc. ICWSM. Barcelona, 2011, pp. 385–392

[WaSG96] Wang, R., Strong, D., & Guarascio, L.: Beyond accuracy: What data quality means to data consumers. In: Journal of Management Information Systems (1996)

[Wolf99] Wolf, P.: Konzept eines TQM-basierten Regelkreismodells für ein "Information Quality Management" (IQM). Verlag Praxiswissen. Dortmund, 1999

[Wulf09] Wulf, V.: Theorien sozialer Praktiken zur Fundierung der Wirtschaftsinformatik: Eine forschungsprogrammatische Perspektive. In: Becker, J.; Krcmar, H.; Niehaves, B. (eds.): Wissenschaftstheorie und Gestaltungsorientierte Wirtschaftsinformatik. Berlin: Springer/Physika, 2009, pp. 211–224

[ZeHe96] Zeist, R. H. J., & Hendriks, P. R. H. Specifying software quality with the extended ISO model. In: Software Quality Journal vol. 5 (1996), Nr. 4, pp. 273–284

[ZiBü12] Zielinski, A., & Bügel, U.: Multilingual Analysis of Twitter News in Support of Mass Emergency Events. In: Proc. ISCRAM. Vancouver, Canada, 2012, pp. 1–5

[ZMTW13] Zielinski, A., Middleton, S., Tokarchuk, L., & Wang, X.: Social Media Text Mining and Network Analysis for Decision Support in Natural Crisis Management. In: Proc. ISCRAM. Baden-Baden, Germany, 2013, pp. 840–845

Brand- und Katastrophenschutzportal Rheinland-Pfalz als TETRA-Statusübermittler für die Feuerwehreinsatzzentralen des Bundeslandes

Sebastian Alsbach, Maria A. Wimmer

Institut für Wirtschafts- und Verwaltungsinformatik
Universität Koblenz-Landau
Universitätsstr. 1
56070 Koblenz
{salsbach|wimmer}@uni-koblenz.de

Abstract: Der weltweite digitale Sprech- und Datenfunkt befindet sich seit einigen Jahren im Aufbau. Für die reine Sprachkommunikation wird er in Deutschland nahezu flächendeckend eingesetzt. Die Datenkommunikation und speziell die Übermittlung von Statusmeldungen gestaltet sich hingegen aufwendig, da sie nicht wie im bisherigen Analogfunk über den Sprachkanal erfolgt. Die Bundesländer müssen eigene Konzepte und Lösungen zur zielgerichteten Übertragung entwerfen. Das Bundesland Rheinland-Pfalz setzt auf den Aufbau eines örtlichen Statusservers, der alle Statusmeldungen empfängt und an die Leitstellen verteilt. Die angestrebte Interimslösung bietet jedoch noch kein allumfassendes Konzept für alle beteiligten BOS-Stellen, wie z.B. die 260 im Land verteilten Feuerwehreinsatzzentralen (FEZen). Dieser Beitrag stellt daher eine mögliche Lösung vor, wie die FEZen trotz ihrer großen Heterogenität der IT-Landschaften und des dadurch fehlenden Zugriffs auf das Landesnetzwerk die für die taktische Einsatzleitung notwendigen Statusmeldungen erhalten. Das Konzept baut auf der Open-Source Web-Portal-Lösung Brand- und Katastrophenschutz Portal des Landes Rheinland-Pfalz auf und nutzt zur schnellen Übermittlung der Statusmeldungen das WebSocket-Protokoll.

1. Motivation

Die bundesweite Einführung des digitalen Sprech- und Datenfunks für die Behörden und Organisationen mit Sicherheitsaufgaben (BOS) ist ein Vorhaben, welches Herausforderungen an alle Beteiligte stellt[1]. Gerade der durch die föderalen Strukturen in Deutschland geprägte Brand- und Katastrophenschutz steht vor der Aufgabe, die zumeist ehrenamtlichen Einsatzkräfte mit allen erforderlichen Informationen zu versorgen. Allein in Rheinland-Pfalz gibt es 240 kommunale Aufgabenträger mit Sicherheitsaufgaben, die wiederum für die Einsatzbereitschaft von fast 2300 Feuerwehren[2] zuständig

[1] Zentralstelle für Polizeitechnik: https://www.digitalfunk-rlp.de/index.php?option=com_content&view=article&id=8&Itemid=15 (Abgerufen am: 14.03.2014)
[2] Feuerwehreinsatzstatistik RLP 2012: http://isim.rlp.de/no_cache/sicherheit/feuerwehr/infomaterial/feuerwehr-in-zahlen/?cid=31042&did=113957&sechash=72f42e45 (Abgerufen am: 14.03.2014)

sind. Koordiniert werden diese durch rund 260 im Land verteilte Feuerwehreinsatzzentralen (FEZen)[3].

Die Übermittlung des Status von Einsatzfahrzeugen mit Hilfe des Funkmeldesystems (FMS) der BOS im Analogfunk unterstützt und entlastet nicht nur die Leitstellen-Disponenten, sondern auch die Besetzung in der jeweiligen FEZ ([Ci08], S. 92). Mit der Umstellung auf den TETRA Digitalfunk ist die Übertragung von Statusmeldungen mittels Short-Data-Service (SDS) weiterhin möglich und belastet zudem nicht wie im Analogfunk den laufenden Sprechfunkverkehr[4] ([Li08], S.11).

Jedoch ist der gefilterte Empfang der Informationen nicht ohne zusätzliche Hard- und Software möglich. Die in Rheinland-Pfalz eingesetzte Lösung der Firma FREQUENTIS AG[5] stellt eine Schnittstelle zur Übertragung von Statusinformationen für die Einsatzleitsysteme bereit. Aus technischen Gründen ist diese jedoch auf maximal 12 Teilnehmer beschränkt [Bu13]. Die rund 260 Feuerwehreisatzzentralen können sich folglich nicht direkt mit der Schnittstelle verbinden.

Eine weitere Herausforderung für den Empfang von Statusmeldungen der Einsatzfahrzeuge ist die große Heterogenität der IT-Landschaften der einzelnen Feuerwehreinsatzzentralen. FEZen sind aus diesem Hintergrund heraus z.B. nicht an das Landesnetzwerk angebunden.

Das Brand- und Katastrophenschutz Portal des Landes Rheinland-Pfalz (in weiterer Folge BKS-Portal.rlp benannt)[6], welches in seinen Grundzügen in [TAW13] vorgestellt wurde, soll nun in einer Erweiterung die Bereitstellung der Statusinformationen für die FEZen ermöglichen. Die Open-Source Web-Portal-Lösung bietet gerade für die ehrenamtlichen Einsatzkräfte einen schnellen und dennoch sicheren Zugang zu den benötigten Informationen. Die Vision im Projekt ist, dass das BKS-Portal.rlp die ungefilterten Statusmeldungen aller BKS-Endgeräte über die oben genannte Schnittstelle aus dem Landesnetzwerk empfängt. Diese Statusmeldungen müssen gespeichert und den einzelnen Feuerwehreinsatzzentralen zugeordnet werden. Das BKS-Portal.rlp soll darüber hinaus eine visuelle Aufbereitung der Statusmeldungen, zugeschnitten auf die jeweilige FEZ, ermöglichen. Ziel dieser Ausarbeitung ist es, das Konzept der Statusübermittlung wissenschaftlich einzubetten und zu beschreiben.

Der Beitrag ist wie folgt strukturiert: Zunächst werden im Kapitel 2 die Problemstellung näher beschrieben und Grundlagen zur Thematik aufbereitet. In Kapitel 3 werden die Forschungsfragen sowie das Forschungsdesign vorgestellt. Darauf folgt eine Betrachtung des Stands der Technik mit einer Untersuchung relevanter Literatur und Lösungen. Kapitel 5 Stellt das zu entwickelte Konzept der Lösung vor. Danach wird die Lösung in Kapitel 6 diskutiert und es werden Wiederverwendbarkeit und Transferierbarkeit unter-

[3] Feuerwehrmagazin 25. November 2013: http://www.feuerwehrmagazin.de/nachrichten/news/35-000-digitale-melder-fuer-rp-39689 (Abgerufen am: 14.03.2014)
[4] BDBOS: http://www.bdbos.bund.de/DE/Digitalfunk_BOS/Digitalfunk_BOS_Vergleich/vergleich_node.html (Abgerufen am: 14.03.2014)
[5] FREQUENTIS AG: http://www.frequentis.com/de/ge/solutions-portfolio/public-safety/products-and-solutions/smart-solutions-for-regional-systems/#!
[6] BKS-Portal.rlp: https://bks-portal.rlp.de/

sucht. Kapitel 7 rundet den Beitrag mit einer Zusammenfassung und einem Ausblick auf weiteren Forschungsbedarf und weitere Arbeiten ab.

2. Problemstellung und Grundlagen zur Statusübermittlung im TETRA Netz

Mit der Einführung des TETRA Digitalfunks in Deutschland und speziell in Rheinland-Pfalz soll der in die Jahre gekommene Analokfunk abgelöst und ein einheitliches digitales Kommunikationsnetz für die BOS aufgebaut werden ([DL11], S.10). Im Bereich der Sprachkommunikation ist inzwischen eine nahezu flächendeckende Umstellung auf die digitale Technik geschehen. Die Übermittlung des Status von Einsatzfahrzeugen wird jedoch weiterhin mit Hilfe des Funkmeldesystems (FMS) über den Analogfunk abgewickelt, da hier noch keine praktikable Lösung im digitalen Netz realisiert ist. Wie die Statusübermittlung im TETRA Netz funktioniert und welche Problemstellungen sich für den BKS-Bereich ergeben, wird nachfolgend erläutert.

2.1 Statusübermittlung im TETRA Netz

Die Statusübermittlung im Analogfunk mittels FMS erfolgt durch das Senden einer Tonfolge über den analogen Sprachkanal. Diese Tonfolge ist für alle Teilnehmer der Funkgruppe hörbar und beinhaltet die Fahrzeugnummer sowie den Fahrzeugstatus. Dieser verlässliche, einheitliche Standard ist interoperabel zwischen diversen Herstellern. Nachteile sind jedoch, dass der Funkkanal während der Übertragung der Statusmeldungen belegt, die Übertragungsgeschwindigkeit gering und die Information nicht verschlüsselt ist ([He97], S.66).

Im TETRA Netz ist eine Statusübermittlung mit dem Short Data Service (SDS) weiterhin möglich. Vergleichbar mit dem SMS Versand im GSM Netz können über den SDS Dienst Statusmeldungen und weitere Kurznachrichten zwischen den Teilnehmern ausgetauscht werden ohne den Sprachkanal zu belegen ([Li08] S.11; [Ci08], S.92). Übermittelt werden hierbei die Teilnehmer ISSI[7] und der Fahrzeugstatus unabhängig von der Sprechfunkgruppe (Sprach-GSSI[8]), in der sich der Teilnehmer gerade befindet. Hieraus ergeben sich Probleme bei der taktischen Zuordnung der Einsatzmittel/Teilnehmer, wenn diese die Heimatleitstelle oder die übliche Sprach-GSSI verlassen. In welcher Sprechfunkgruppe sich ein Teilnehmer befindet kann nur durch das sogenannte „tracking“ ermittelt werden. Herzu sind jedoch Administrator Rechte erforderlich, die in der Regel nur die Heimatleitstelle besitzt.

Verlässt ein Einsatzmittel den Bereich der Heimatleitstelle, so wechselt es wie aus dem Analogfunk bekannt in die Sprach-GSSI der nun zuständigen Leitstelle. Anders als beim Analogfunk laufen die Statusmeldungen jedoch weiterhin bei der Heimatleitstelle auf, da diese unabhängig von der Sprach-GSSI an eine im Funkgerät hinterlegte Daten-GSSI

[7] Die ISSI (Individual Short Subscriber Identity) kennzeichnet ein Endgerät innerhalb eines Funknetzes eindeutig
[8] Die GSSI (Group Short Subscriber Identity) kennzeichnet eine Gesprächsgruppe innerhalb eines Funknetzes

versandt werden. Ein Sprechwunsch an die zuständige Leitstelle würde dort nicht ankommen. Diese kann das Einsatzmittel auch nicht mit Status führen. Die Heimatleitstelle erhält hingegen weiterhin die Statusmeldungen, hat jedoch keine Information über die aktuelle Sprach-GSSI des Einsatzmittels.

Zur Lösung des Problems gibt es verschiedene Ansatzmöglichkeiten, die in Abschnitt 4 näher vorgestellt werden. Rheinland-Pfalz hat sich für den Einsatz eines örtlichen Statusservers im Bundesland entschieden.

2.2 Problemstellung im Projektkontext

Die Firma FREQUENTIS AG wurde von der autorisierten Stelle in Rheinland-Pfalz damit beauftragt, einen zentralen Statusserver im Bundesland einzurichten. Dieser soll die Statusinformationen ungefiltert von den TETRA Endgeräten an die Leitstellen transportieren und zudem die Möglichkeit bieten, Anweisungen von der Leitstelle an einzelne TETRA Teilnehmer zu senden. Dabei ist es wichtig, dass sich alle Teilnehmer in einer zentralen Daten-GSSI befinden und jede Leitstelle zudem mindestens eine Sprach-GSSI verwendet. Die Leitstellen bekommen die Statusinformationen aller Teilnehmer der zentralen Daten-GSSI übermittelt und filtern selbstständig auf die für sie relevanten Einsatzmittel. Die eigenen Teilnehmer werden hierbei sofort über die ISSI identifiziert und im Einsatzleitsystem (ELS) angezeigt. Fremde Einsatzmittel werden erst angezeigt, wenn sie in einer der Sprach-GSSIs der Leitstelle eingebucht sind. Dies wird anhand der „Affiliation“ auf die eigenen Sprach-GSSIs ermittelt.

Die in Rheinland-Pfalz angestrebte Interimslösung bietet jedoch noch kein allumfassendes Konzept für alle beteiligten BOS-Stellen. So können sich die rund 260 Feuerwehreinsatzzentralen nicht direkt mit dem Statusserver verbinden. Dies hat mehrere Gründe. Zum einen ist die mögliche Nutzerzahl durch FREQUENTIS auf maximal 12 ELS begrenzt, da die zur Verfügung stehenden Statusserver die Anfragen anders nicht bewältigen können [Bu13]. Ein Ausbau der Statusserver könnte hier Abhilfe schaffen, erfordert jedoch einen großen finanziellen Aufwand. Zum anderen befinden sich die Statusserver in einem nach außen abgeschotteten Landesnetz. Im Vergleich zu den örtlichen Polizeidienststellen sind die FEZ nicht an dieses Landesnetz angebunden. Die zumeist ehrenamtlichen Kräfte der FEZ haben daher keine Möglichkeit, an die Statusinformation heranzukommen. Sie können nicht aktiv und in Echtzeit am Einsatzgeschehen teilhaben. Dies behindert eine transparente und schnelle Einsatzleitung.

Ziel ist es daher, die Statusinformationen auch den zumeist ehrenamtlich besetzten FEZ derart bereitstellen zu können, wie es im Analogfunk auch möglich war. Wie dies mittels des in Rheinland-Pfalz eingerichteten BKS-Portal.rlp funktionieren kann, wird in den nachfolgenden Kapiteln vorgestellt. Zunächst aber wird das Forschungsdesign beschrieben.

3. Forschungsfragen und Forschungsdesign

In dem vorliegenden Beitrag wird die Problemstellung anhand der folgenden Forschungsfragen eingefasst. Zentral ist die Frage, wie die Statusinformationen an eine größere Gruppe des Einsatzteams geteilt werden können. Mit rund 260 Feuerwehreinsatzzentralen liegt die Zahl der ELS deutlich über den durch den Statusserver in Rheinland-Pfalz vorgegebenen 12 möglichen [Bu13]. Der Bedarf ist daher gegeben, die Informationen des zentralen Statusservers weiter zu teilen. Die Kommunikation zwischen einem geschlossenen Landesnetz in das schwer zu schützende Internet ist hierbei eine weitere Forschungsfrage. Da die FEZ zumeist nur über einen Internetanschluss ohne Sicherheitskonzepte verfügen, müssen die Statusinformationen aus dem geschützten Landesnetz in das Internet überführt werden. Ein Schutz vor ungewolltem Abhören soll jedoch weiterhin gewahr sein.

Eine weitere Forschungsfrage ist, wie die richtige Filterung und gezielte Weitergabe der bereitgestellten Statusinformationen zielgruppengerecht und zweckorientiert aufbereitet werden soll. Der zentrale Statusserver gibt alle eingehenden Statusmeldungen an die angemeldeten ELS weiter. Eine einzelne FEZ benötigt jedoch nur Informationen über die für sie relevanten Einsatzmittel. Also die eigenen Teilnehmer und die Einsatzmittel, für die sie aktuell die Rolle der führenden Leitstelle übernimmt. Letztlich müssen die Statusinformationen für die FEZ visuell zweckmäßig aufbereitet werden. In den wenigsten FEZ gibt es ausgereifte ELS, die eine geeignete Schnittstelle zur Entgegennahme der Statusinformationen bereitstellen. Die Statusinformationen müssen daher visuell aufbereitet und letztlich auf einem gewöhnlichen PC darstellbar sein ([Ap11], S.41).

Die Frage der Zwei-Wege-Kommunikation, also das Senden von Statusmeldungen aus der FEZ an die Einsatzmittel wird in diesem Beitrag nicht behandelt, da hierzu noch umfangreiche Überlegungen und Anpassungen in der Interimslösung notwendig sind.

Erwartet wird von diesem Beitrag im ersten Schritt ein Aufzeigen der Problemsituation, insbesondere im Vergleich zu den örtlichen Polizeidienststellen, die sich in einer homogenen IT-Landschaft befinden. Zudem soll der Bedarf einer praxisnahen Lösung identifiziert werden. Wesentlich ist jedoch das Konzept zur Statusübermittlung an die FEZen, welches im diesem Beitrag vorgestellt wird. Es umfasst die Schnittstelle in das geschützte Landesnetz, die Filterung und Weitergabe sowie letztlich die effektive visuelle Darstellung der Informationen. Die nachfolgende Abbildung 3.1 stellt die Aufgabenstellung in einem Rich Picture dar.

Das Forschungsdesign bedient sich verschiedener Methoden der Systemanalyse, um das vorhandene soziotechnische System zu analysieren und ein Konzept zu erarbeiten, mit dem die erwarteten Ergebnisse erreicht werden ([Ru08], S.1; [Bi05], S.134). Zunächst wird eine Literaturrecherche (Desk Research) zum Stand der Technik durchgeführt. Hier werden Grundlagen zu der Thematik Digitalfunk und im Besonderen zur Statusübermittlung zusammengetragen. Eine Anforderungsanalyse gibt anschließend Aufschluss über die von allen Beteiligten erwarteten Anforderungen an die zu konzeptionierende Lösung [Ru05]. UML-Diagramme und Mock-Ups helfen, die Lösung

schon vor der Implementierung anschaulich zu machen und gemeinsam mit den künftigen Nutzern zu diskutieren.

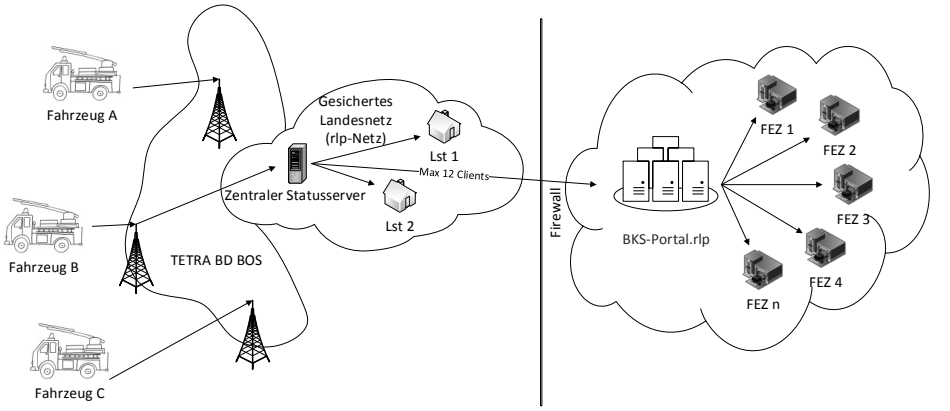


Abbildung 3.1: Rich Picture zur Statusübermittlung an die FEZ mittels BKS-Portal.rlp

4. Stand der Technik und Optionen der Statusübermittlung

Die Grundlage für die Einführung eines bundesweit einheitlichen digitalen Sprech- und Datenfunksystems bildet das „Schengen Abkommen“ 1985: Der „schrittweise Abbau der Kontrollen an den gemeinsamen Grenzen“ muss durch verbesserte Kommunikationsmöglichkeiten der Sicherheitsbehörden kompensiert werden.[9] Am 24. März 2004 schlossen Bund und Länder die "Vereinbarung zur Regelung der Zusammenarbeit beim Aufbau und Betrieb eines bundesweit einheitlichen digitalen Sprech- und Datenfunksystems für alle BOS in der Bundesrepublik Deutschland" und folgten damit einem weltweiten Trend ([Ci08], S. 351). Um die Planung, den Aufbau und den Betrieb des Digitalfunks BOS durchzuführen, wurde die Bundesanstalt für den Digitalfunk der Behörden und Organisationen mit sicherheitsaufgaben (BDBOS) eingerichtet und deren Aufgaben und Zuständigkeiten im BDBOS-Gesetz geregelt.[10] Ende 2006 wurde entschieden, den Aufbau des TETRA-Systems durch den Bund und die Länder in Eigenverantwortung durchzuführen.[11] Hier wirkte sich der Föderalismus in der BRD eher nachteilig aus, da er ein Hemmnis für eine von allen Beteiligten getragene Entscheidungsfindung darstellte [Be03]. Jedes der Bundesländer sowie der Bund verfügen demzufolge über eine Projektgruppe Digitalfunk. Diese arbeiten bei Planung, Aufbau und Betrieb des Digitalfunks eng mit der BDBOS zusammen und sind u.a. verantwortlich für die Bereitstellung, Ertüchtigung und Instandhaltung von

[9] Digitalfunk Rheinland-Pfalz: https://www.digitalfunk-rlp.de/index.php?option=com_content&view=article&id=9&Itemid=16 (Abgerufen am: 02.04.2014)
[10] BDBOS-Meilensteine: http://www.bdbos.bund.de/DE/Digitalfunk_BOS/Meilensteine/projekt_meilensteine_node.html (Abgerufen am 04.04.2014)
[11] Digitalfunk Rheinland-Pfalz: https://www.digitalfunk-rlp.de/index.php?option=com_content&view=article&id=9&Itemid=16 (Abgerufen am: 02.04.2014)

Basisstationen und Übertragungsstrecken.[12] Auch Rheinland-Pfalz bildet eine eigene „Projektgruppe Digitalfunk“, um das Projekt zu koordinieren und zu realisieren.

Eine Problemstellung, für die von Seiten der BDBOS bisher keine einheitliche Lösung angeboten wird, ist die Statusübermittlung. Hierzu gibt es verschiedene Lösungsansätze, die im Folgenden kurz diskutiert werden. Über die Nutzung sowie die Umsetzung des Dienstes, verbunden mit einer landesspezifischen Sonderlösung, entscheidet das jeweilige Bundesland.

Eine erste, dem Analogfunk ähnliche Lösungsmöglichkeit, ist die Anpassung des Statusziels (Daten-GSSI) in den Endgeräten in Abhängigkeit der aktuellen Sprechgruppe (Sprach-GSSI). Sobald der Nutzer sein Endgerät auf eine andere Sprach-GSSI umstellt, also die Sprechfunkgruppe wechselt, muss das Endgerät automatisch auch die Daten-GSSI entsprechend anpassen. So würden Status und Sprache immer an dieselben Leitstellen gesendet. Das Systemverhalten wäre dem Analogfunk gleich. Eine manuelle Anpassung der Daten-GSSI im Endgerät durch den Nutzer ist bei diesem Lösungsansatz ebenfalls denkbar. Jedoch gibt es für die Änderung der Daten-GSSI kein standardisiertes Leistungsmerkmal von Seiten der BDBOS. Es ist daher Gerätehersteller-abhängig, ob die Funktionalität zur Verfügung gestellt werden kann.

Eine andere denkbare Möglichkeit ist die Vernetzung der Leitstellen und eine damit mögliche Statusweiterleitung. Der Status eines Einsatzmittels läuft immer im ELS der Heimatleitstelle auf. Wechselt ein Endgerät in die Sprach-GSSI einer anderen Leitstelle, die nun die Rolle der führenden Leitstelle übernehmen soll, kann sie die benötigten Statusinformationen bei dem ELS der Heimatleitstelle anfragen. Konkret kann das ELS der führenden Leitstelle feststellen, welche Einsatzmittel in der eigenen Sprach-GSSI eingebucht sind. Für die ihm fremden Einsatzmittel kann es nun bei den vernetzten ELS nachfragen. Alternativ könnte das ELS der Heimatleitstelle z.B. über die GPS-Koordinaten oder Abfragen der eingewählten Sprach-GSSI ermitteln, in welchem Einsatzgebiet sich das Einsatzmittel gerade befindet und der zuständigen Leitstelle die Statusinformationen übertragen. Diese Lösungsmöglichkeit setzt jedoch eine hohe Interoperabilität zwischen den eingesetzten ELS verschiedener Hersteller voraus. Bei einheitlicher ELS Software ist diese Möglichkeit jedoch denkbar.

Eine weitere Lösungsmöglichkeit ist der Einsatz eines bundesweiten Statusservers. Alle Statusmeldungen werden z.B. über eine einheitliche Daten-GSSI an den Statusserver übermittelt. Dieser kann die Informationen dann jeder Leitstelle je nach Umsetzungsvariante bereits gefiltert oder ungefiltert zur Verfügung stellen. Der zentrale Statusserver könnte die empfangenen Statusmeldungen mit der Information über die eingewählte Sprach-GSSI nur den dafür zuständigen Leitstellen übergeben. Denkbar ist auch, dass der Statusserver alle Statusmeldungen allen ELS zur Verfügung stellt und diese selbstständig ermitteln, welche Informationen für sie relevant sind. Hiermit haben die Leitstellen einerseits eine hohe Datenverantwortung, da sie theoretisch über alle Statusmeldungen im Land verfügen. Andererseits haben sie eine hohe Flexibilität in der

[12] BDBOS-Projektgruppen der Bundesländer:
http://www.bdbos.bund.de/DE/Digitalfunk_BOS/Bundeslaender/bundeslaender_node.html (Abgerufen am: 04.04.2014)

Darstellung des Einsatzgeschehens, da sichergestellt ist, dass relevante Statusinformationen vorliegen und nicht durch den Server zurückgehalten werden. Dieser Lösungsansatz wird in einem Einzelprojekt von der BDBOS und der AG Leitstellen koordiniert. Es gibt jedoch noch keinen verbindlichen Termin oder Aussagen zur Funktionsweise.

Die letzte hier vorgestellte Lösungsmöglichkeit baut auf einen örtlichen Statusserver im Bundesland auf. Die Funktionsweise wird in diesem Ansatz gleich der zuvor beschriebenen Möglichkeit sein. Der entscheidende Unterschied ist, dass die Planung, Umsetzung und der Betrieb des Statusservers nur innerhalb eines Bundeslandes erfolgt. Die Funktion ist daher nur auf das jeweilige Bundesland beschränkt.

In Rheinland-Pfalz wird die zuletzt vorgestellte Lösungsmöglichkeit durch einen örtlichen Statusserver im Bundesland umgesetzt. Jedoch gibt es wie in Kapitel 2.2 beschrieben einige Problemstellungen, die mit dem im Folgenden vorgestellten Konzept zur Übermittlung der Statusinformationen in die FEZ gelöst werden.

5. Konzept zur Übermittlung der Statusinformationen an die FEZ

Das nun vorgestellte Konzept zur Statusübermittlung in die FEZen basiert auf zwei Annahmen, die zunächst erläutert werden. Zum einen soll die Lösung auf das von der Firma FREQUENTIS AG bereitgestellte Common Data Interface CDI des zentralen Statusservers zugreifen. Mit dieser SOAP-Schnittstelle ist es möglich, ungefilterte Statusinformationen von den TETRA Endgeräten zu empfangen und gezielte Anfragen an das TETRA-Netz zu senden. Die Schnittstelle hat jedoch kapazitive Einschränkungen. So können sich beispielsweise nur 12 Teilnehmer mit ihr verbinden [Bu13].

Zum anderen baut das Konzept auf der Infrastruktur des BKS-Portal.rlp auf. Das Portal richtet sich an alle Beteiligte im Brand- und Katastrophenschutz sowie Rettungsdienst in Rheinland-Pfalz. Sein Aufbau orientiert sich an den drei Kernbereichen Wissens- und Informationsmanagement, Organisation/Datenpflege und Zusammenarbeit. Über das Netzwerk des BKS-Portal.rlp lassen sich Informationen bis in die FEZen verteilen. Zudem bietet das feingranulare Berechtigungssystem des BKS-Portal.rlp zusammen mit einer zentralen Benutzerverwaltung den geforderten Sicherheitsrahmen.

Das Konzept sieht einen eigenen BKS-Statusdienst vor, der auf einem gesicherten Server, mit Anbindung an das Landesnetz, installiert wird. Dieser Statusdienst soll die XML basierten Statusmeldungen von der CDI-Schnittstelle entgegen nehmen, verarbeiten und in einer Datenbank zwischenspeichern. Parallel zur Ablage in der Datenbank sollen die Informationen an einen WebSocket[13] Server geleitet werden, der die Statusmeldungen unmittelbar an die angemeldeten FEZen weiterleiten kann, ohne das diese eine Aktualisierungsanfrage an den Server stellen müssen. Die Arbeitsweise

[13] WebSocket-Protokoll: TCP basierendes Netzwerkprotokoll für bidirektionale Verbindungen, zwischen Webanwendung und Server die aktiv vom Server verwaltet werden kann. (vergleiche V. Wang; F. Salim; P. Moskovits (2013): The Definitive Guide to HTML5 Web Socket. New York: Apress, S. 7)

des geplanten BKS-Statusdienstes ist mit dem UML-Diagramm in Abbildung 5.1 verdeutlicht.

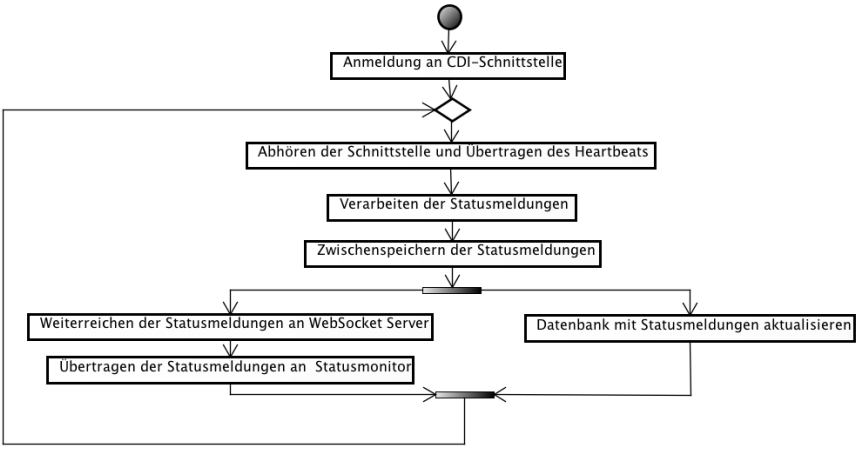


Abbildung 5.1: UML-Diagramm des BKS-Statusdienstes im BKS-Portal.rlp

Im BKS-Portal.rlp soll für jede FEZ ein eigener Bereich eingerichtet werden, auf den nur die Mitglieder der jeweiligen FEZ Zugriff erhalten sollen. Hier soll hinterlegt werden, für welche Einsatzmittel die jeweilige FEZ die Rolle der Heimatleitstelle übernimmt und welche Sprechgruppen (Sprach-GSSIs) ihr zugeteilt sind. Durch eine Umsetzungstabelle mit allen relevanten Informationen, wie ISSI, Geburts-OPTA, Alias-OPTA, Funkrufname und Heimatleitstelle, soll die Zuordnung der eigenen Einsatzmittel zu einer FEZ automatisiert erfolgen. Sind die eigenen Einsatzmittel und Sprechgruppen festgelegt, soll sich ein BKS-Statusmonitor starten lassen, der nur die Statusinformationen zu den relevanten Einsatzmitteln anzeigt. Die Logik zur Filterung der relevanten Statusmeldungen einer FEZ wird in Abbildung 5.2 gezeigt. Wie zu sehen sollen nur die Statusmeldungen der eigenen Einsatzmittel und die Meldungen der Einsatzmittel, die sich aktuell in einer der eigenen Sprach-GSSIs befinden dargestellt werden.

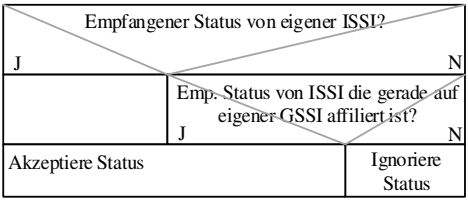


Abbildung 5.2: Logik zur Auswertung der relevanten Statusmeldungen (nach [Bu13])

Technisch soll die Aktualisierung des BKS-Statusmonitors mit Hilfe einer WebSocket Verbindung umgesetzt werden. Dies ist notwendig, da die Anforderung besteht, Statusänderungen möglichst schnell in den FEZen anzuzeigen. Eine Aktualisierung des Monitors durch z.B. eine Ajax-Anwendung[14] ist zu langsam und überlastet zudem die

[14] Asynchronous JavaScript and XML (Ajax): Asynchrone Datenübertragung zwischen Browser und Sever, um Seite zu verändern, komplett neu zu laden. (vergleiche B. McLaughlin (2006): Ajax von Kopf bis Fuß. O'Reilly Verlag)

Server des BKS-Portals.rlp. Bei Aufruf des BKS-Statusmonitors soll eine Verbindung zum WebSocket Server aufgebaut werden. Sobald diese besteht beginnt eine Synchronisierung des Monitors. Hierzu werden die aktuellen Meldungen aus der Datenbank gelesen und parallel alle eingehenden Statusmeldungen temporär zwischengespeichert. Sind alle erforderlichen Daten aus der Datenbank geladen, werden die Informationen mit den temporär zwischengespeicherten Statusmeldungen abgeglichen. Nach der Synchronisierung sollen die Statusmeldungen über die WebSocket Verbindung empfangen werden. Abbildung 5.3 zeigt die geplante Arbeitsweise anhand eines UML-Diagramms.

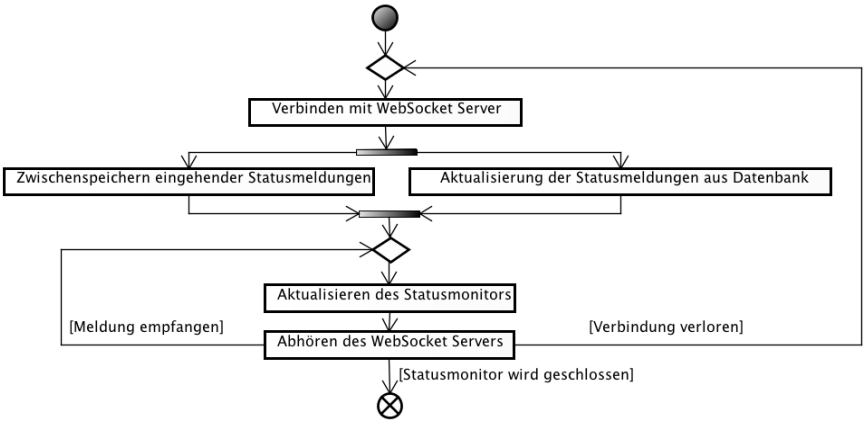


Abbildung 5.3: UML-Diagramm des BKS-Statusmonitors im BKS-Portal.rlp

Für die visuelle Darstellung des BKS-Statusmonitors ist ein eigener Inhaltstyp im BKS-Portal.rlp vorgesehen. So lässt sich die Seite, im Rahmen der Möglichkeiten des Drupal-CMS auf dem das BKS-Portal.rlp aufbaut, individuell an die Bedürfnisse der Nutzer anpassen. Abbildung 5.4 zeigt ein Mock-Up einer Darstellungsform im BKS-Portal.rlp.

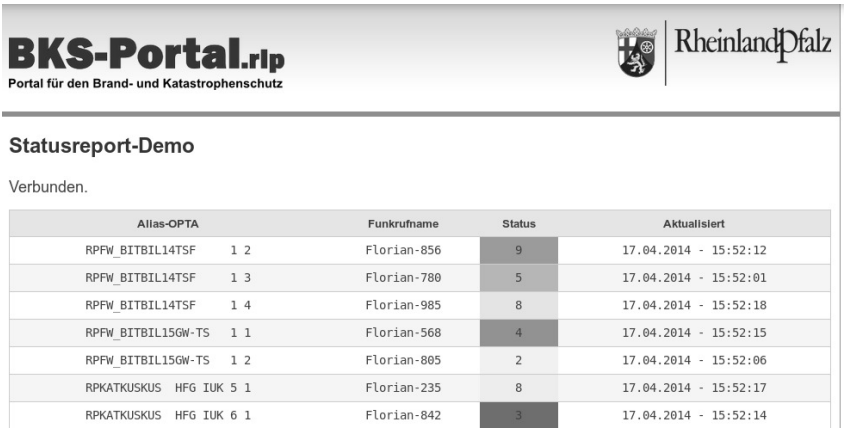


Abbildung 5.4: Mock-Up des BKS-Statusmonitors im BKS-Portal.rlp

6. Diskussion und Handlungsempfehlungen für eine Übermittlung der Statusinformationen

Mit diesem Konzept werden Statusmeldungen von einer zentralen Stelle entgegengenommen und mittels Webtechnologien an zahlreiche Teilnehmer multipliziert. Eine Besonderheit ist hierbei die Nutzung einer Web-Portal-Lösung. Diese erlaubt es Informationen auch in eine heterogene IT-Landschaft zu übertragen. Durch Wahl des WebSocket-Protokolls kann eine Übertragung und Darstellung der Statusmeldungen in Echtzeit erfolgen, die zudem ressourcenschonend über das Netz kommuniziert. Dieser Technologieansatz kann auch für andere Problemstellungen, wo die zeitkritische Übermittlung von einfachen Informationen im Fokus steht, Anwendung finden. So könnten beispielsweise zentrale Einsatztagebücher bei Großlagen ähnlich verteilt werden.

Kritisch muss erwähnt werden, dass die eigentliche Problematik, die mit der Einführung des Digitalfunks entstanden ist, mit dem Aufbau eines örtlichen Statusservers und dem daran anknüpfenden Konzept zur Statusübermittlung an die FEZen nicht unmittelbar zu lösen ist. Es bleibt zunächst die Frage offen, welche Einsatzmittel in welcher Sprechgruppe (Sprach-GSSI) eingebucht sind. Diese Information ist notwendig, um die Statusmeldungen eines Einsatzmittels, welches seine Heimatregion verlässt, auch in der dann führenden Leitstelle anzeigen zu können. Durch „Affiliation“ auf die eigenen Sprechgruppe ist es möglich, herauszufinden, welche fremden Einsatzmittel in der eigenen Gruppe eingebucht sind. Bei 925 Sprechfunkgruppen in Rheinland-Pfalz dauert es jedoch nach Schätzungen der Firma FREQUENTIS AG bis zu 6 Minuten, bis alle Informationen vorliegen [Bu13]. Zudem werden die Server des TETRA-Netz hierdurch stark belastet. Zum aktuellen Zeitpunkt liegen hierzu noch keine Lasttests vor.

7. Zusammenfassung und Ausblick

Das hier vorgestellt Konzept zur Statusübermittlung an die FEZen knüpft an einen örtlichen Statusserver im Bundesland an. Dieser kann die FEZen aus zwei Gründen nicht unmittelbar mit Statusinformationen versorgen. Zum einen befindet er sich in einem geschützten Landesnetz, zu dem die FEZen keinen Zugang haben und zum anderen ist der Statusserver auf maximal 12 Teilnehmer beschränkt. Das Konzept baut auf die Infrastruktur, gegeben durch das Open-Source Webportal BKS-Portal.rlp, auf und verteilt hierüber die Statusmeldungen. Zudem wird das WebSocket-Protokoll verwendet, um die Statusmeldungen, wie gefordert, möglichst zeitnah in den FEZen zur Verfügung zu stellen. Zur visuellen Aufbereitung werden die Möglichkeiten des Drupal- CMS auf dem das BKS-Portal.rlp aufbaut genutzt. Erste Tests zeigen hierbei, dass das vorgestellte Konzept umzusetzen ist.

Dennoch besteht weiterer Forschungsbedarf. Das hier vorgestellte Konzept sieht zunächst nur den Empfang von Statusmeldungen vor. Die Frage der Zwei-Wege-Kommunikation, also das Senden von Statusmeldungen aus der FEZ an die Einsatzmittel wird in diesem Beitrag nicht behandelt und erfordert weitere Forschung. Zudem muss

die in Abschnitt 6 angeschnittene Problematik mit der Frage nach der aktuell verwendeten Sprach-GSSI eines Einsatzmittels weiter untersucht werden. Hier sind Geschwindigkeit und erzeugte Last zu optimieren. Letztlich gibt es Organisatorisch noch offene Fragestellungen. So müssen Änderungen in der in Kapitel 5 erwähnten Umsetzungstabelle an alle Leitstellen und auch das BKS-Portal.rlp kommuniziert werden.

8. Literaturverzeichnis

[Ap11] Appel, B. (2011): Grundlagen der Leitstellenplanung. Berlin: Steinbeis-Hochschule

[Be03] Beckebanze, H.: Die BOS und der Digitalfunk – eine "unendliche Geschichte". Polizeitechnisches Institut der Polizei-Führungsakademie, Oktober 2003, Münster; S.12.

[Bi05] Binner, H.F. (2005): Handbuch der prozessorientierten Arbeitsorganisation - Methoden und Werkzeuge zur Umsetzung. 2. Auflage. München: Hanser.

[Bu13] Burgsteller M. (2013): TETRA Statusdienst Lösungsbeschreibung für Rheinland-Pfalz. FREQUENTIS AG (internes Dokument)

[Ci08] Cimolino, U.; Bayer, G.; Schneider, S.; Schweigger, A. (2008): Kommunikation im Einsatz: Planung, Organisation und Technik. 2. überarb. und erw. Aufl. Landsberg: ecomed Sicherheit

[He97] Henschel, J. (1997): Struktur und Management der nichtöffentlichen Funknetze. Diplomarbeit. Norderstedt: GRIN

[Li08] Linde, C. (2008): Aufbau und Technik des digitalen BOS-Funks. Poing: Franzis

[Ru05] Rupp, C. (2005): Requirements-Engineering und -Management. Professionelle, Iterative Anforderungsanalyse für die Praxis. München; Wien: Hanser.

[Ru08] Rupp, C. (2008): Systemanalyse kompakt; 2. Auflage, SOPHIST GROUP Heidelberg: Spektrum.

[DL11] Demel, J.T.; Linde C. (2011): Fachwissen Feuerwehr – Digitalfunk. Hamburg: ecomed Sicherheit.

[TAW13] Timm, E.; Alsbach S.; Wimmer, M.A. (2013): Wissens- und Kollaborationsplattform im Brand- und Katastrophenschutz: Open-Source als kostengünstige Option? In: Horbach, M. (Hrsg.): Informatik 2013. Informatik angepasst an Mensch, Organisation und Umwelt. 43. Jahrestagung der Gesellschaft für Informatik. Köllen Druck+Verlag GmbH Bonn. Nr. P-220. S. 1634-1648.

Datenmanagement und Management elektronischer Ressourcen in Bibliotheken

(DERM 2014)

Datenmanagement und Management elektronischer Ressourcen in Bibliotheken (DERM 2014)[1]

Abstract: The objectives of the international workshop DERM 2014, as part of the INFORMATIK 2014 conference in Stuttgart (Germany) is concerned with the transition of traditional information systems to a web-oriented and Linked Data oriented infrastructure in libraries. The initial focus within this research field is on evaluating the complexity of existing databases and accordingly providing correct data alignments. As a consequence thereof, another focus is on the development as well as the deployment of adaptive information systems. These systems can be used flexibly in new business processes within libraries, that are driven by tasks like managing and sharing heterogeneous data sets. Additionally, accumulating and augmenting data is made easier by providing tools that abstract from the technical details and the complexity of the Linked Data world as far as possible.

Within this context, the workshop addresses data engineers, knowledge engineers and software engineers, as well as users in the IT administration of libraries, information suppliers and other interested users.

Organization: Jens Mittelbach (Saxon State and University Library Dresden), Leander Seige (Leipzig University Library), Torsten Hartmann (Avantgarde Labs Dresden), Thomas Riechert (Leipzig University of Applied Sciences - HTWK), Natanael Arndt (Institute for Applied Informatics Leipzig), Andreas Nareike (Institute for Applied Informatics Leipzig)

[1] http://aksw.org/Events/2014/DERM2014.html

Design and Implementation of a Library Metadata Management Framework and its Application in Fuzzy Data Deduplication and Data Reconciliation with Authority Data

Martin Czygan
Leipzig University Library
Beethovenstraße 6
04107 Leipzig
martin.czygan@uni-leipzig.de

Abstract: We describe the application of a generic workflow management system to the problem of metadata processing in the library domain. The requirements for such a framework and acting real-world forces are examined. The design of the framework is layed out and illustrated by means of two example workflows: fuzzy data deduplication and data reconciliation with authority data. Fuzzy data deduplication is the process of finding similar items in record collections that can not be matched by identifiers. Data reconciliation with authority data takes a data source and enhances the metadata of its records – mainly authors and subjects – with available normed data. Finally, the advantages and tradeoffs of the presented approach are discussed.

1 Introduction

The system described in this paper is a result of a few product iterations on the problem of library metadata management. The problem emerged during project *finc*[1] – a project for the development of search engine based discovery interfaces for university libraries in Saxony – where a couple of heterogeneous data sources needed to be processed and merged in certain, sometimes convoluted, ways. The main contribution of this paper is to report on core requirements (section 3) of such a system, and how most of these requirements can be met by adapting a generic workflow management framework (sections 4 and 5). Subsequently, two example workflows from the library domain are described in short (section 6), followed by a discussion (section 7), where key tradeoffs are identified.

2 Related Work

There are a number of open source software products for workflow management, dependency graph handling and data processing. A prominent example of a tool that utilizes

[1] More information about the project can be found on its homepage at `http://finc.info` and in [Laz12].

a dependency graph is make[2], a build tool for software projects[3]. More data-oriented projects include Oozie[4], Askaban[5], joblib[6], Taverna[7] and luigi[8]. Features that made luigi a compelling choice[9] as the base for our system are a) a small code base[10], b) efficient command line integration and c) a single, lightweight notation to express tasks and their relationships.

3 Requirements

An easily formulated requirement of a metadata management application for libraries is the ability to work with different data sources and to combine, filter and transform the data, as it traverses the system. This requirement implies some kind of formalism of a unit, on which transformations can be applied. Typical notions for these units would be a file on the filesystem or on the web, a single record or a stream of records. This requirement is broad and leaves a lot of design decisions untouched. To add additional design constraints, we enumerate a few non-functional requirements.

Consistency Transformations on data should happen in a transactional manner or should in other ways be guaranteed to be atomic[11], that is a transformation should either be applied to a data element or not. Atomicity should be enforced at different data granularity levels, from a single field in a single record, that is rewritten, to sets of potentielly tens of millions of records, that need a transformation to be applied on all of its elements.

Composability It is desireable to design the data elements or the processing units in ways that support reuse. There are several approaches to this: A certain data package can be used to build various workflows atop. In that way, workflows do not need to recompute certain states or shapes of the data over and over again. Another way of reuse would be

[2] `http://www.gnu.org/software/make/`

[3] Other open source projects with similar scope are – among others – rake, ant, maven, gradle, scons, waf.

[4] From `https://oozie.apache.org/`: "Oozie is a workflow scheduler system to manage Apache Hadoop jobs. Oozie Workflow jobs are Directed Acyclical Graphs (DAGs) of actions."

[5] From `https://azkaban.github.io/`: "Azkaban is a batch workflow job scheduler created at LinkedIn to run Hadoop jobs. Azkaban resolves the ordering through job dependencies and provides an easy to use web user interface to maintain and track your workflows."

[6] From `https://pythonhosted.org/joblib/`: "Joblib is a set of tools to provide lightweight pipelining in Python."

[7] From `http://www.taverna.org.uk/`: "Taverna is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation."

[8] From `https://github.com/spotify/luigi`: "Luigi is a Python package that helps you build complex pipelines of batch jobs. It handles dependency resolution, workflow management, visualization, handling failures, command line integration, and much more."

[9] The system design is decribed in more detail in section 5.

[10] luigi 1.0.16 contains 5125 LOC, which included the task abstractions, a scheduler, support for hdfs, hadoop, hive, scalding, postgres, Amazon S3, date, configuration and compression handling and task histories.

[11] [HR83] describes atomicity as one of the four characteristics of a transaction: "Atomicity. It must be of the all-or-nothing type [...], and the user must, whatever happens, know which state he or she is in."

the application of a single task or processing unit on several data streams. For example, different data sources could be fetched in form of compressed files from an FTP server, so the task of fetching and extracting these files would present itself as a template task, because for several different data sources the processing would be similar up to a fixed number of parameters.

Efficiency Many data sources are updated in intervals and recomputations of data derived from that data source in turn should happen as frequently as the data source itself is updated. In general, a workflow, that encompasses a number of data sources should be executed as soon as any one of the constituent data sources changes in a significant way. Conceptually, one can view the combination of a number of data sources as a data source itself with the update times consisting of the update times of all subsumed data sources along with a total order.

Ease of deployment, agile development and change management The application should be easily relocatable[12] and should bootstrap its data environment as autonomously as possible. This proves to be useful for development since a simple setup enables developers to mirror a production environment on a development machine and therefore reduce the gap between both setups.

Development of new workflows and changes to existing workflows should not affect established setups, that might be part of a production system. It should be possible to distribute tasks sensibly among members of a team.

Testability, transparency and introspection Writing code that deals with millions of records and complicated workflows can be a challenge, as would be writing unit tests for those workflows. It would be desireable to be able to unit test components as well as to write integration tests for complete workflows with the help of mock services.

Most workflow management systems will break down a multi-stage workflow into several subtasks. One should be able to introspect the shape of the data along the way at the different steps to support transparency and debuggability.

4 Forces

There are a number of forces around the problem of library metadata management, of which a few should be introduced in this section.

[12] This is certainly interesting in a cloud setup, where machines may be spun up and down on a regular basis.

Data formats The format of the data to be processed varies widely. While there are standardized data formats for libraries such as MARC[13], MAB[14] or RDA[15], there is still enough room for variation. Other data sources, which might be of interest to the library community, have not yet been standardized or are not necessarily curated with the goal of standardization. Typical formats are XML (which in turn can come in many flavors to express similar things), JSON, HTML, N-Triples, CSV or plain text. All these formats can furthermore be compressed or span multiple files.

Data access As we described for the data formats the accessibility of the data varies widely, too. Typical examples are FTP servers, metadata harvesting protocols such as OAI-PMH [CLW08], Web-APIs, E-Mail attachments, database or index connections.

Update intervals and update signalling Typical for the use case, data sources are updated in intervals. The update intervals can range from a few seconds over days to many weeks, months or years. In other words they can range from real-time to a singular update. In these remarks of we do not consider real-time, but only cases, where the update granularity does not underrun 24 hours[16].

An interesting observation concerns the way, the availability of an update is signalled. This can happen in many ways as well, with two broad categories being explicit or implicit. An example for explicit signalling would be a file name containing a date or some other indicator of succession, such as a serial number. Implicit signalling means that the data source will not convey the availability of updates, unless it is preprocessed to reveal this information. An example for implicit signalling would be OAI-harvesting, where one would need to perform an actual request and a comparison with existing records to find out, whether updates are available[17].

Another aspect during update processing is the way in which an update to a data source is expressed. There are at least two ways in which this can be achieved in practice. A simple way to express an update to a data source is to provide a complete dump.

Using deltas is also common practice. Given an initial dump or a delta against an empty state, subsequent changes to the data source are expressed as changes, which need to be applied to the previous state of the data source. The upside of this approach is the limitation of the update to the affected records, the downside – which sometimes results in significant overhead in the amount of processing – is the need to keep all the metadata packages available if one wants to recompute some state of the data source after a number of deltas. The way in which the deltas are formulated can differ between data sources. One data source might provide an update or deletion list as an extra file, while other data sources ship updated or deleted records with a flag set at a predefined record field. The

[13] `http://www.loc.gov/marc/`

[14] `http://www.dnb.de/DE/Standardisierung/Formate/MAB/mab_node.html`

[15] `http://www.rda-jsc.org/rda.html`

[16] The workflow framework ships with a parameters class, that allows update intervals of 60 minutes. Below that, the inherently batch-oriented nature of the framework might not prove supportive.

[17] A *from* parameter in the request is required, which in turn can only be derived from some previous harvest.

variety of these notions does not assist in the definition of a consistent abstraction layer, when one tries to generalize the event of an update over a bigger number of data sources[18].

Processing speed It is desirable to utilize modern multicore computers by implementing systems, which take advantage of parallel processing whenever feasible. It should be noted that while certain parts of a workflow might not be suitable for parallelization, others might and the processing framework would ideally be able to identify these parts by itself and schedule the tasks accordingly.

5 Design

In this section, the overall system design is outlined along with the description of how the implementation addresses the requirements and forces.

Building blocks The system described in this report is based upon an open source software library called luigi[19], developed by Spotify[20] – a music streaming service. This library provides basic abstractions for workflow management and execution[21] and embodies many of the desired properties listed in section 3. The contribution of this report is the description of how this generic framework for implementing workflows can be adapted for the library domain.

The basic building block of the framework is the notion of a task, which is both simple and powerful. A task has a couple of essential properties: It has zero, one or more requirements, which are tasks themselves; furthermore a task has a body of code, which corresponds to the business logic that needs to be executed; finally every task has one or more targets[22]. Aside these three core notions each task has zero or more parameters. Every workflow is decomposed into a number of tasks. Tasks are implemented and wired together in the source code itself. Figure 1 gives an impression of an emerging depedency graph. Workflow execution is realized through command line integration, which supports interactive use as well as automation with standard tools such as cron[23].

A central scheduler, implemented as a deamon, handles task scheduling and visualisation. While the notion of a target supports many different implementations[24], the majority of the artefacts are files.

[18] This fact suggests, that the notion of an update to a data source alone might not be the abstraction level sweet spot.

[19] `https://github.com/spotify/luigi`

[20] "Luigi was built at Spotify, mainly by Erik Bernhardsson and Elias Freider, but many other people have contributed.", `https://github.com/spotify/luigi`

[21] The framework comes with more than just the basic abstractions. Most notable there are wrappers around HDFS [Bor08] and Hadoop [Whi09] to support cluster computing as well.

[22] While it is possible to create tasks with multiple targets, it is recommended, that each task has exactly one artefact as output.

[23] `http://unixhelp.ed.ac.uk/CGI/man-cgi?cron`

[24] The community implemented targets for Amazon Redshift, MySQL, Postgres, Elasticsearch.

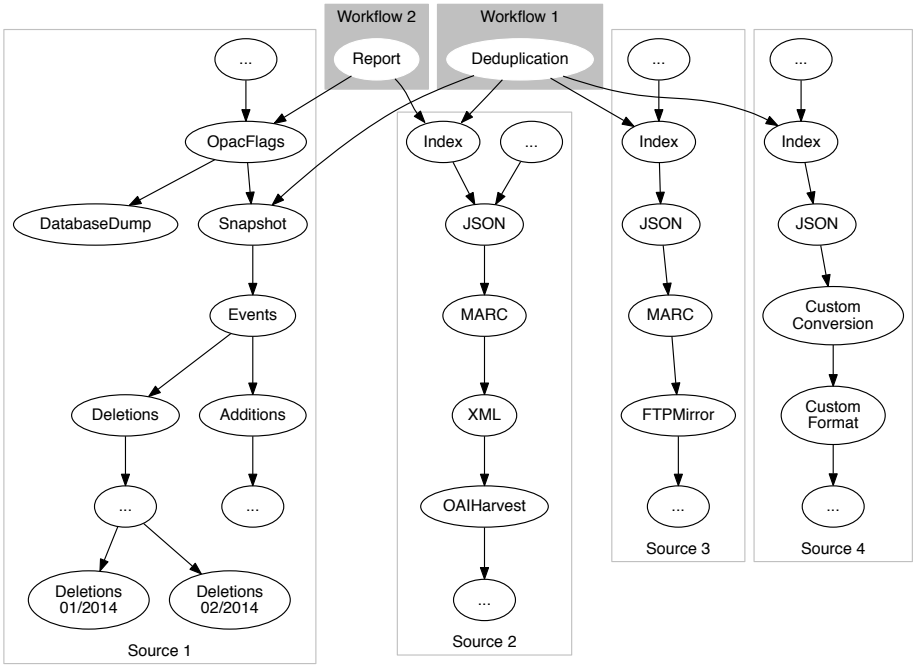


Figure 1: An impression of a dependency graph (with simplified task names and without task parameters). Different sources will differ in their internal complexity, depending on various factors, such as update intervals and signalling. Workflows can be built upon tasks from many sources.

Consistency The framework itself does not enforce consistency, but makes it easy to implement it. Following the rule, that each task will create an artefact only if the task has been run successfully, it is possible to carry over the consistency from tasks to workflows. A workflow (which is just a set of tasks) is successfully completed, only if every task executed successfully. When dealing with files, the atomicity can be enforced in the task by using the POSIX rename interface [IG13][25].

On the other side, it is possible to repeatedly execute workflows. Since each task has a defined target (e.g. a file), it will not execute the task again, if it has been run successfully in the past (e.g. if the output file exists) with the same set of parameters.

Idempotence and atomicity in the building blocks aid in reducing the number of states the system can be in, which benefits the resilience of the overall system and also helps to reason about what happened and what failed.

[25] "This rename() function is equivalent for regular files to that defined by the ISO C standard. Its inclusion here expands that definition to include actions on directories and specifies behavior when the new parameter names a file that already exists. That specification requires that the action of the function be atomic."

Composability The notion of a task underlines the provision of small units. While there is no limitation in what a task can do and how many lines of code it can contain, it is convenient to keep the amount of work and responsibility for a single task limited. In an evaluation of about 300 tasks, that implement several different workflows in our system, the average number of lines of code (LOC) inside the body of a task was 13.7. Figure 2 shows the distribution of the lines of code over the tasks. The number of tasks that constitute a workflow ranges between a few tasks up to several hundreds of tasks.

The nature of the framework supports composability also by letting the developer implement generic tasks, that can be reused by other tasks by adding these generic tasks to the workflow or derive special tasks from them. Examples for generic tasks, that have been implemented in our system are tasks for mirroring FTP files and directories, harvesting OAI-PMH interfaces or indexing documents.

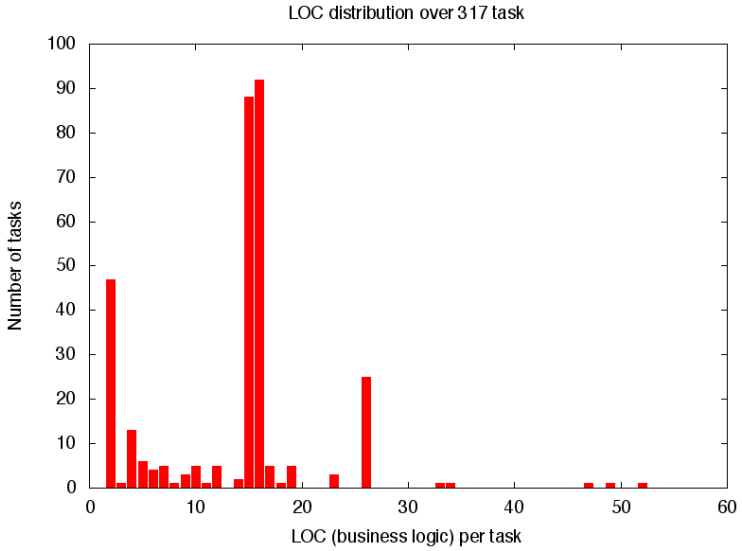


Figure 2: Distribution of lines of code (LOC) over 317 tasks. The average LOC per task (business logic only) is 13.7.

Efficiency Two important types of efficiency are efficiency concerning tasks and efficiency in the context of parallelism. Efficiency with regard to tasks means that tasks and workflows are only executed, if there has been an update to the data source in question. As described in section 4, the update intervals can vary widely between data sources. It can be sufficient to parameterize a task by date and to have it execute in regular intervals (e.g. in weekly, monthly or quarterly intervals). Other tasks might have some latest state, which can only be evaluated by looking at the data source itself. For these cases, a special parameter class has been implemented, that is able to map any given date to the closest date in the past, on which an update occured for a data source. In this way a task can

respond with an correct output artefact for arbitrary dates. The business logic of the task is only executed if absolutely necessary, therefore contributing to the efficiency of the overall setup.

The second type of efficiency is handled by the framework, which can use any number of workers. A worker corresponds to a separate OS process. Since the tasks in the system form a dependency graph, the framework is able to determine, which tasks are parallelizable and will spawn separate processes to execute them, given the user allows more the one worker to run.

Ease of deployment, agile development and change management Most of the system's artefacts are files, which are organized under a single directory. This makes it relatively easy to backup or replicate the metadata – both original and derived – to other machines by copying a single directory. The software consists of a single python package that can be installed with standard tools. One strength of the setup is the ability to almost completely bootstrap a metadata infrastructure, given a valid configuration has been provided. This, again, is possible because the tasks form a dependency graph. This feature supports agile development as well. A developer does not need to know, how a certain state or artefact of the metadata (e.g. a list of ISBNs from a few different data sources or a list of records from a data source, that have been updated in the last three weeks) is provided, as long as there is a task in the system that provides that specific artefact.

During development it has been our observation, that certain tasks are referenced more often than other. One could call these tasks *stable points*, since they underpin more workflows than other tasks. On the other hand, there are tasks, that merely serve as an intermediate step in some longer computation, making them less important in the overall setup. These kind of metrics[26] can also inform development. We saw that investing time in the refinement of the stable points in the architecture yielded performance benefits for many workflows, whereas it was often possible to contract less important intermediate steps into a single larger task for the sake of performance – without sacrificing too much conceptual clarity by agglutinating many steps within a single task.

Testability, transparency and introspection As described in section 5, the average number of LOC for a task is relatively low. Furthermore, a task has a defined set of inputs (often just a single file) and outputs (per recommendation just a single file). This helps writing unit tests, since a task resembles a function with a number of arguments and most of the time no or only a few references to any kind of state outside itself. Since the tasks themselves can be tested with moderate efforts, complicated workflows, which span hundreds of tasks, become manageable as well.

The mostly file-based setup of the system facilitates another aspect: transparency. Most tasks produce a single output and this output has a location[27] on the file system. In our

[26] Each task records its performance and reports it to the user. The benchmarks are implemented with python decorators, making it easy to extend this system into a metrics database that could monitors performance characteristics over time, should the need arise.

[27] We extended the built-in file system abstractions of the framework to be able to almost complety hide the

system, we wrote small wrapper scripts, that help to inspect tasks with a single command, abstracting from the real filename, referring to the task only by its name and parameters[28]. Another addition concerns documentation. By combining the documentation facilities[29] and the dynamic nature of the python programming language with the implicit dependency information encoded in the task, it is possible to autogenerate comprehensive command line documentation, similar to a UNIX man page[30].

Deployment setup The software is a single python package, splitted into two modules: One contains the data sources, another the higher level workflows. The data source modules take care of the data from its origin (e.g. FTP servers, web sites, databases), to some stable point, e.g. in our case this is mostly a single JSON file that represents the current state of the data source or a corresponding Elasticsearch index. The workflow modules work with those data sources, modeling a variety of things, such as data deduplication, reporting or data enrichment and reconciliation. The python package itself has a couple of external dependencies for database access, index access, ISBN handling, tabular data manipulation among other things. Common tasks and utilities have been factored out into a separate package.

6 Example workflows

Two workflows, that have been implemented atop various data sources are briefly described in this section. The important observation in both examples is the fact, that non-trivial tasks, that originate in the library data domain, can be made manageable, by builing upon a powerful data structure like a dependency graph.

6.1 Fuzzy data deduplication

As a first example we describe briefly the process of fuzzy data deduplication. We selected three data sources with about five million records as a first target. There are three workflows involved to get each data source into an Elasticsearch index. Two data sources work with regularly updated files on FTP-Servers, one data source uses a more involved scheme involving dumps, deltas and additional lists for deletions. Of the two data sources that use FTP, one conveys updates in form of complete dumps, whereas the other uses a set of deltas. As described in section 5, a stable point in our setup is an Elasticsearch index with a document structure, that is relatively homogenous across data sources. These indices are kept up-to-date as the data source commands it.

underlying directory structure from the developer or user.

[28] This feature is especially useful, when a new data source is added to the system and needs to be explored with the help of a few initial tasks.

[29] `http://legacy.python.org/dev/peps/pep-0257/`

[30] `http://manpages.bsd.lv/history.html`

For the fuzzy deduplication a dual approach of cosine and ngram similarity is used. First a list of ids is generated, for which similar items are collected with the help of the a heuristic query[31], based on record titles. The titles and authors of the candidates are then broken into trigrams and compared. Records with a similarity score above a threshold are then reported as fuzzy duplicates. Additional fields are evaluated in a postprocessing step. Records, for example, that do not share a similar edition statement are removed from the list of duplicates[32].

6.2 Data reconciliation with authority data

Another requirement concerns data enrichment and data reconciliation. Since our setup uses over 30 data sources, the quality of the metadata is not necessary alike. Especially there are data sources, that can benefit from the knowledge already encoded in form of library metadata in other data sources. One example is the use of high quality authority data produced and published by the German National Library in form of the Integrated Authority File (GND) [BNP11]. This authority file contains among other things naming variants for entities, for example authors or composers. On the other hand there are data sources in our metadata landscape, that carry a single naming variant of a person, thus making it impossible to find the record by using a valid name that is not part of the record. The workflow here takes the entries from the GND and successively compares the naming variants to names and entities found in the less rich metadata source, generating suggestions for metadata enrichment. In a quality control postprocessing step, the suggestions are checked for plausibility, by evaluating additional information about the subject[33].

7 Discussion

There are a couple of tradeoffs involved in a setup, as described in this report. The first and probably most significant is the space overhead of a file based approach. Similar to a map-reduce [DG08] style processing, where each intermediate step is persisted to disk, before another phase in the workflow picks it up, the disk space overhead is considerable. Depending on the concrete workflows, the increase in storage requirements in our setup can be up to tenfold – compared to the size of the raw original data source. It is possible to view the data generated by the tasks of the workflow as a cache, computing parts of the process once and reducing the number of computations required for subsequent steps. One example for this incremental computing[34] approach is the handling of data sources,

[31] This query uses Lucene's *MoreLikeThis* facility which uses a *bag-of-words* model to find similar documents.

[32] The definition of a duplicate will vary in different situations.

[33] We implemented this workflow for the University of Music and Theatre Leipzig. We could utilize information about the profession of a person, encoded in the GND taxonomy, to weigh suggestions. For example there are many people named Bach, of which not all are composers. In a second step, we used biographical data, also encoded in the GND, to further clean the list of suggestions.

[34] *Reactive Imperative Programming with Dataflow Constraints* [DFR11] describes incremental computing as follows: "When the input is subject to small changes, a program may incrementally fix only the portion of the

that ship updates with deltas. In one case, each delta package needs a few processing steps (e.g. extraction, conversion, filter). In another step all deltas and their derived artefacts are considered to create a snapshot of the data source. The amount of work that needs to be done when a new delta arrives is minimal, once the derived artefacts of the previous deltas are in place – making it possible to quickly create snapshots of that data source.

This extensive caching facilitates the solution of other problems in a straightforward manner. One example would be the computation of change reports for data sources. Since many stages of the metadata has been or can be recomputed, the comparison of the state of the metadata between two arbitrary time instances becomes surprisingly simple. One just calls the same task with two different date parameter values and extracts the difference[35].

A disadvantage of the presented approach is the discipline required to implement the tasks. Each tasks must be written in an manner, so that it supports atomicity and idempotency. Also not every task is suitable for this functional paradigm, although from empirical evidence we are able to conclude, that many different real world tasks are. Persisting intermediate steps can lead to inefficiencies, since performance could benefit from keeping data in memory for a certain amount of time or for a certain chain of tasks[36].

8 Summary

In summary, the advantages of a system like the one described in this report are the following: A workflow composed of a number of smaller tasks facilitates testable, reusable components. If it is easy to reason about a single task, complex workflows spanning multiple data sources can stay manageable. Supplying tasks with parameters for time can help with the isolation of update behaviours and can shield the developer from the idiosyncrasies of a data source by providing stable points in the architecture, which enable a multitude of workflows. Finally, trading extensive space requirements for a gain in simplicity and composability seems to be a worthwhile tradeoff in the library metadata domain.

References

[BNP11] Renate Behrens-Neumann and Barbara Pfeifer. Die Gemeinsame Normdatei – ein Kooperationsprojekt. *Dialog mit Bibliotheken*, (1):37–40, 2011.

[Bor08] Dhruba Borthakur. HDFS architecture guide. *HADOOP APACHE PROJECT http://hadoop. apache. org/common/docs/current/hdfs design. pdf*, 2008.

output affected by the update, without having to recompute the entire solution from scratch. “

[35] If, for example, the two tasks emit single column data, one can load the two artefacts into two sets and perform a set difference.

[36] This line of critique seems to apply to the map-reduce style framework Hadoop as well [Hea14]: “Spark is able to execute batch-processing jobs between 10 to 100 times faster than the MapReduce engine according to Cloudera, primarily by reducing the number of writes and reads to disc.“ Another

[CLW08] Michael Nelson Carl Lagoze, Herbert Van de Sompel and Simeon Warner. The Open Archives Initiative Protocol for Metadata Harvesting, 2008.

[DFR11] Camil Demetrescu, Irene Finocchi, and Andrea Ribichini. Reactive imperative programming with dataflow constraints. In *ACM SIGPLAN Notices*, volume 46, pages 407–426. ACM, 2011.

[DG08] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[Hea14] Nick Heath. Faster, more capable: What Apache Spark brings to Hadoop, 2014.

[HR83] Theo Haerder and Andreas Reuter. Principles of transaction-oriented database recovery. *ACM Computing Surveys (CSUR)*, 15(4):287–317, 1983.

[IG13] The IEEE and The Open Group. The Open Group Base Specifications Issue 7, IEEE Std 1003.1, 2013 Edition, 2013.

[Laz12] Jens Lazarus. *Projekt finc Ein Open Source Discovery System für sächsische Hochschulbibliotheken*, volume BIS - Das Magazin der Bibliotheken in Sachsen 5(2012)2 S. 72 - 76. Saechsische Landesbibliothek- Staats- und Universitaetsbibliothek Dresden Dresden, 2012. article.

[Whi09] Tom White. *Hadoop: The Definitive Guide: The Definitive Guide*. O'Reilly Media, 2009.

AMSL – Managing Electronic Resources for Libraries Based on Semantic Web

Andreas Nareike[1,2], Natanael Arndt[2], Norman Radtke[1,2],
Sebastian Nuck[2], Leander Seige[1] and Thomas Riechert[3]

[1] Leipzig University Library
Beethovenstr. 6, 04107 Leipzig, Germany
{lastname}@ub.uni-leipzig.de

[2] Institute for Applied Informatics (InfAI) e.V.
Hainstraße 11, 04109 Leipzig, Germany
{lastname}@infai.org

[3] Leipzig University of Applied Science (HTWK)
Gustav-Freytag-Str. 42A, 04251 Leipzig, Germany
thomas.riechert@htwk-leipzig.de

Abstract: In libraries a change from physical resources to electronic resources, with new licensing models and lending processes, has taken places. The existing managing infrastructure is not yet suitable for the upcoming requirements and does not provide support for flexible and extensible data models for being future-proof. In this paper we present a system that uses the generic RDF resource management system OntoWiki for managing library resources. OntoWiki is extended by components for adapting the generic system to the given domain, e.g. by using data templates. In addition the Linked Data capability of OntoWiki is used and extended to import various metadata to enrich the managing resource. Consequently using Linked Data further enables libraries to build up a Linked Data infrastructure in the library domain.

1 Introduction

Electronic resources are gaining more and more importance in modern libraries. While in the past every step in the classical workflow of acquiring, cataloguing and lending was related to a physical item, nowadays libraries have to deal with new categories of media and publication formats, such as e-journals, e-books or databases. Since those resources are not limited to physical items and can be copied without loss of information, new business models have to be developed by publishers. Current examples are pay-per-view, short term loan, big deal and Patron-Driven-Acquisition (PDA) but due to new markets, publication formats and user expectations, new models will keep evolving in the future. Since libraries have to manage these new complex models on their side, we introduce a new system for managing electronic resources in libraries with semantic web technology. This involves

various kinds of information to be stored and interconnected, such as contacts, contracts, packages, agreements and licenses.

In addition to the necessity of managing these resources, our endeavour contains a need for data integration: In the past years, many information sources have evolved on the vendor market which are necessary for the electronic resource management (ERM) or can help to improve it. All this information is available from different sources like Publishers, the German National Library[1], the Journal Database[2], authority files, Integrated Library Systems (ILS), Application Programming Interfaces (API) or DBpedia [KBAL09], while exported in heterogeneous formats e.g. MARC[3], ONIX-PL[4], COUNTER[5], knowledge bases, Microsoft Access databases, RDF and SPARQL.

Following conventional component based software engineering approaches, a system managing this complexity of information resources would need one component for each class of resources. With each new requirement or change in the data structure, the components have to be adjusted or new components have to be developed. In contrast to this conventional component based software engineering approach, we want to shift the complexity from the application layer to the data model. Our method of choice is to use the Resource Description Framework (RDF) [BGM14] and Linked Data [BL09] to specify this data model. Ideally, this means that the software engineering approach can be agnostic regarding the underlying data model. In practice, there will be some additional domain-specific components to further increase the usability in a library context. This design results in an extensible and adaptable software system which is able to manage all information within the life-cycle of an electronic resource.

2 Requirements for Electronic Resource Management in Libraries

Our electronic resource management solution for library data can be structured in two main aspects: (1) managing electronic resources and (2) enriching the electronic resources by aggregating and integrating bibliographic and business metadata from different sources. The aspect of managing electronic resources (1) is mainly done by creating, editing, linking and exporting electronic resources and their data, like license terms and contracts or the vendors and their contact data. Since we manage a lot of data, exporting and reporting is also a very important feature as providing a change history is one, too. Lastly, we want to provide the possibility to quickly address any resource in the system without the need to specify search fields, much like one would use a full text search. To achieve aspect (2), we further need to be able to import all that heterogeneous data into our data model to enrich existing electronic resources. External data will often differ in the file format as well as in the chosen model. Due to this, we also need tools to convert and transform data from different sources. Besides these two main requirements there are other non-functional

[1] Deutsche Nationalbibliothek (DNB): `http://portal.dnb.de`
[2] Zeitschriftendatenbank (ZDB): `http://www.zeitschriftendatenbank.de`
[3] MARC standards, The Library of Congress: `http://www.loc.gov/marc/`
[4] ONIX for Publications Licenses (ONIX-PL): `http://www.editeur.org/21/ONIX-PL/`
[5] National Information Standards Organization (NISO): `http://www.niso.org/schemas/sushi/`

requirements our solution should fulfil. The data model should be adaptable to offer the possibility to react to changes in the vendor market. And of course other libraries should be able to use the ERM or participate collaboratively in an existing instance, so they can interchange metadata and build up a Linked Open Data (LOD) infrastructure within the library domain.

3 State of the Art

Existing solutions which are currently used in libraries and are relevant for our endeavour can be distinguished in systems for managing print resources, electronic resources or more generic any kind of resources and tools for aggregation and integration of relevant data sources.

Traditionally, libraries have been primarily concerned with managing print resources. When it comes to electronic resources like e-books or e-journals, the situation is significantly changed. Not only are there different licensing models with accompanying complex access models, the contract terms and subscriptions terms often change with every new business year. Although Integrated Library Systems for managing print resources (e.g. LIBERO[6] used at the Leipzig University Library) are already available and in use in libraries, they are not prepared to be used for the management of electronic resources.

Recently *Next Generation Library Systems* (or *Unified Resource Management Systems*) which constitute a new approach to Electronic Resource Management have emerged.While those management systems imply a modern multitier architecture, there are also commercially driven cloud-based systems that come as Software-as-a-Service. Due to the commercial nature, these systems are not transparent and it is difficult to evaluate how well standards are implemented and which requirements have been incorporated [JAC+04, JAA+12]. Also these products come with their own commercial knowledge bases, which is an important part of the licensing model which makes it difficult to integrate external knowledge bases. Furthermore, these products have in common that the underlying data models are not sufficiently generic (as required in section 2) and can't be extended, as the structure needs to be scaled to fit new structural informational requirements.

A more general approach for managing electronic library resources is to use generic resource or document management systems e.g. specially adapted Wikis. OntoWiki [HDMA09, FMT+14] is a semantic Wiki system for managing resources of the Web of Data using RDF and OWL. OntoWiki is highly extensible and serves as an application development framework for knowledge intensive applications [FMT+14]. Besides the core functionality of organising RDF resources in classes, creating, querying and editing the RDF resources, it provides an application programming interface for third party extensions. The Erfurt-Framework is the abstraction layer, which is used to provide access to the underlying triple store for the high-level application. With the Linked Data Server and Linked Data Wrapper components OntoWiki can be used for publishing and consuming Data on the Linked Data Cloud resp. Web of Data.

[6] LIBERO – Library Management System: `http://www.libero.com.au/`

4 Using OntoWiki and RDF for Electronic Resource Management

One approach of building a highly agile system for managing electronic resources while being flexible to react to changing requirements to the data model is to implement a service oriented architecture (SOA) with an ecosystem of components which can be combined and extended as needed. In contrast to setups of SOAs in producing industries, the library electronic resource management doesn't consist of highly complex business processes with many combined tasks. Instead, we have a low complexity in the workflow management (e.g. simple processes like recording a new license) but a rather high complexity in the data model of interconnected resources. Implementing the system by developing individual customised components for each single step in the workflow is very expensive since each of them has to be adjusted, once new requirements arise for the data model. Our approach is to use a generic software framework for working on an highly expressive data model expressed in RDF and OWL. As a consequence, new requirements only require modification in the data model while the generic software can be used unchanged.

The implementation of AMSL[7] based on the OntoWiki Application Framework [HDMA09, FMT+14] with Erfurt Framework to gain access to our knowledge base and data model.[8] With its generic edit, query and visualisation components the data is accessible for collaboratively working on one instance and multiple instances via Linked Data. The edit and visualisation functions are adapted for domain experts by using a newly introduced flexible way of expressing data templates in RDF. External data is imported into the knowledge base using the generic Linked Data components and specialised adaptors. For faster access to the complex knowledge base, an adaptor Elasticsearch index is maintained.

4.1 The Data Model

A great amount of our data is imported from external knowledge bases. This is mostly bibliographic data about titles and publishers, scientific fields of research, etc. There is usually little or no choice regarding the vocabularies or the data model and we import this data 'as is'. To conveniently use this data, we will create some alignments for these vocabularies.

For business data about license types and conditions, information about contact persons and information about the budget, we have to create resources by our own. Our main focus is on using common vocabularies and vocabularies that have been developed by the library community.

For bibliographic data we use ontologies of the Dublin Core Metadata Initiative[9] and the Bibliographic Ontology Specification[10] (BIBO). For concepts we could not find in existing

[7] AMSL project: `http://amsl.technology/`

[8] AMSL project code repository: `https://github.com/AKSW/OntoWiki/tree/deployment/erm-hd`

[9] DCMI Metadata Terms: `http://purl.org/dc/terms/` and `http://purl.org/dc/elements/1.1/`

[10] Bibliographic Ontology Specification: `http://bibliontology.com/`

vocabularies mentioned before, we developed the vocabulary BIBRM[11] that is aligned to the ideas of the Electronic Resource Management Initiative (ERMI) [JAC+04]. For business data, we use the Academic Institution Internal Structure Ontology[12] (AIISO) and the Friend of a Friend vocabulary[13] (FOAF) to manage organisations, contacts and academic institutions.

4.2 Knowledge Base Concept for Collaborative Work in Library Consortia

OntoWiki is able to manage a large number of Knowledge Bases stored in the triple store. In addition, it provides user management that grants read and/or write access for each knowledge base. To meet the requirements for data aggregation and data enrichment described in section 2, we developed a knowledge base management concept.

We differentiate our knowledge bases into three kinds according to their role. Knowledge bases belonging to category (A) provide additional information for ERM and might contain complete dumps. Category (A) also includes knowledge bases providing data about faculties, organisations and contacts which was created mostly by hand. Knowledge bases of type (B) can be considered as stores for auxiliary data. They provide triples linking one resource with another, e.g. an ISSN with a ZDB resource. A knowledge base of this type could be obtained by exporting and transforming data from an Integrated Library System. Auxiliary knowledge bases are usually hidden even for administrators and are not edited within OntoWiki.

The types (A) and (B) both have in common that they will help to connect a high number of resources to their metadata counterparts right from the start, so resolving ISSNs and fetching resources from the ZDB linked data endpoint is only necessary if something is missing.

The final type (C) of knowledge bases are the ones used as working model and contain data that is mostly relevant to only one library. All the business data about contracts, licenses, prices etc. will be stored and edited there. To achieve multitenancy capability for the whole system there is one knowledge base of type (C) for each tenant (project partner).

To use the triples of knowledge bases belonging to (A) and (B) in the working models of (C), OntoWiki supports the OWL property `owl:imports`. As a result, triples of one knowledge base will be accessible in another knowledge base, as if they were imported. This enables the users to use the linking triples and access the metadata of e-journals with a few clicks. The advantage of using `owl:imports` is that the triples of a knowledge base can be used in many different knowledge bases without actually duplicating triples. Furthermore, having a separate knowledge base facilitates updating or even exchanging the knowledge base.

Data stored in (A) and (B) is shared knowledge of all libraries, therefore user privileges of OntoWiki will be set to allow read and write access for every contributor for these

[11] Vocabulary for Library ERM: `http://vocab.ub.uni-leipzig.de/bibrm/`
[12] Academic Institution Internal Structure Ontology: `http://vocab.org/aiiso/schema`
[13] FOAF Vocabulary Specification 0.99: `http://xmlns.com/foaf/0.1/`

knowledge bases. Since business data is more sensitive, the models of type (C) will only be read- and writable by users of that institution.

4.3 Templates for System Librarians

Initially, we were looking for a way to formally document in which way we combine different vocabularies to describe our data. Vocabularies are often not very precise in their constraints with the intention to be as widely usable as possible. The Dublin Core property `dc:creator` does not specify the kind of the creator (author, composer, painter, etc.) and also does not link the creator to other vocabularies (e.g. FOAF). To build a consistent database, it is not feasible to choose a semantic for each new date, i.e. to decide and write down for each new creator, that it also is a person as defined by the FOAF vocabulary. Especially for people with little or no experience in RDF, it is also easier to just fill out a form without having to worry too much about the vocabularies. On the most non-technical oriented level, we wanted to provide such forms that can be used on a daily basis. However, it is worth noting that forms, by 'shaping' the data, actually implement the underlying logic of an application or the application profile. On a more general level, when the application logic is sufficiently clear and machine-readable, forms are just one expression of this logic. An application profile should also

- filter imported data and possibly even apply transformation rules
- express the used ontology, without having to look at the data itself
- define/build an abstract API or SPARQL queries [TWSWG13] to access the stored data in a specific way

As we stated above, a requirement for an ERM system is the ability to quickly adapt to changes in the general framework without having to restructure the whole data store. Consequently, we store the template definitions in the triple store which is editable by the stakeholders. Our core idea is to define templates which bind a class and list properties that an instance of this class should use. The terms 'class' and 'property' here refer to the RDF/OWL constructs e.g. `rdfs:Class`, `owl:Class`, `rdf:Property`, `owl:DatatypeProperty` or `owl:ObjectProperty`. More specifically, listing 1 shows an example template.

```
@prefix bibrm: <http://vocab.ub.uni-leipzig.de/bibrm/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix bibo: <http://purl.org/ontology/bibo/> .
@prefix rdavocab: <http://rdvocab.info/> .

<http://vocab.ub.uni-leipzig/bibrm/Template/Periodicals> a bibrm:Template ;
    bibrm:bindsClass bibo:Periodical ;
    bibrm:providesProperty dc:title, bibo:issn, rdavocab:placeOfPublication .
```

Listing 1: An example template for a periodical

This template states how the class `bibo:Periodical` is used in the context of the application. Note that `bibo:Periodial` is a class of an external namespace, so we should not modify this class resource directly. However, a subclass of `bibo:Periodical` could be defined by using `rdfs:subClassOf`. This is a somewhat common approach, but requires additional reasoning when it comes to interlinking on the Web of Data. Occasionally, the use of subclasses might even be seen as ontology hijacking.

```
<http://ld.zdb-services.de/resource/2206346-8> a bibo:Periodical ;
    dc:title "Nature physics" ;
    bibo:issn "1745-2481" ;
    rdavocab:placeOfPublication "Basingstoke" .
```

Listing 2: A resource derived from the template in listing 1

The simple resource that is modelled after this template is shown in listing 2. It's left open, if the properties listed in the template are optional or mandatory, but we also provide properties that have a stricter semantic. Right now we use templates to describe our resources which can be seen as a formalised cookbook.

Another core usage are forms that simplify the creation of new resources. These forms are directly derived from a template. For instance, when creating a new resource of type `bibo:Periodical`, the template in listing 1 would generate a form that has input fields for `dc:title`, `bibo:issn` and `rdavocab:placeOfPublication`.

Where available, a suitable widget will be used. For instance, in the case of the publication place, an URI is often preferred to a simple string, e.g. instead of 'Heidelberg', `http://dbpedia.org/resource/Heidelberg` is more usable. Here we can provide a widget that uses auto-completion based on labels for URIs, so typing 'Heidelberg' will produce the a suitable resource. By using `rdfs:range`, we can further specify where OntoWiki will look for resources for auto-completion.

When importing external resource data, there might be more triples that are not really needed or wanted. For example, when resolving ZDB resource in listing 2, we will actually get more triples. By evaluating our template for `bibo:Periodical`, we can suppress unwanted triples when displaying a resource.

4.4 Periodical Import Process of License Data

Importing lists of subscribed journals is one of the main methods to solve requirement aspect (2) described in section 2. While the vendors will send their contract details and subscription lists in several formats and data structures to their library partners, we had to figure out which data is the most important for our data model. In dialogue with the librarians of Leipzig University Library, we learned that the import data is composed of both the electronic and the print version of International Standard Serial Number (ISSN), the title information and (optionally) the price of the e-journals. Librarians manually create CSV files which are structured, using the four columns, as mentioned above. This file than can be imported into OntoWiki which will create RDF resources accordingly. On the

import form, the user selects if the journals are part of a license contract or license package and writes a comment describing the contract/package. Each line of the import file will be accepted if there is at least one ISSN given or otherwise be dropped and not further processed. A line of the CSV file results in a new contract item resource (URI generation with hash of timestamp and the first found ISSN) and a link from the contract/package resource to that newly created resource. In listing 3 you can see an example for one CSV line and the resulting triples in RDF in listing 4 after the import process.

```
1234-5678;2345-6789;"An example";990,90
```

Listing 3: Example for one line of the CSV file

```
@prefix bibrm: <http://vocab.ub.uni-leipzig.de/bibrm/> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ex: <http://example.com/> .

ex:contract1 a bibrm:LicenseContract ;
    bibrm:hasItem ex:1 .

ex:1 a bibrm:LicenseContract ;
    dc:title "An example" ;
    bibrm:PISSN <urn:ISSN:1234-5678> ;
    bibrm:EISSN <urn:ISSN:2345-6789> ;
    bibrm:price "990,90" .
```

Listing 4: Resources created when importing the CSV from listing 3

4.5 Resolving ISSN as Identifier for Resources

For electronic journals, the ISSN is the main identifier for collecting and aggregating metadata. The ZDB provides a REST-API that will return RDF result sets for a given ISSN. In AMSL we developed an ISSN resolver that fetches the result set for an ISSN, extracts the important triples that link to the resources provided by the ZDB Linked Open Data service and creates an RDF (cf. listing 5) which is then imported into OntoWiki using the provided Linked Data Wrapper component. In our case, every resource with a URI starting with `urn:ISSN` followed by a valid ISSN will be resolved by the ISSN resolver to import the links to the ZDB resources into our data model. Linking these two kinds of resources helps to meet requirement aspect (2) (cf. section 2).

```
@prefix umbel: <http://umbel.org/umbel#> .

<urn:ISSN:1234-5678> umbel:isLike
      <http://ld.zdb-services.de/data/2754934-3> ,
      <http://ld.zdb-services.de/data/2662590-8> ,
      <http://ld.zdb-services.de/data/2089379-6> .
```

Listing 5: Example of linking a ISSN to resources of ZDB

4.6 Recording Changesets for each Resource

Every transaction done by OntoWiki is monitored by the built-in versioning mechanism. Hence, every addition and removal of statements is stored in a separate database. For reasons of comprehensibility, additions and removals relating to the same resource are consolidated into one single change statement event. As a result, every resource builds up a change log which provides an overview about the changes related to the resource. Since these changes are chronologically ordered, the built-in rollback feature allows to restore a resource to an earlier state. In addition to simple information about changes regarding a resource, the versioning also provides aggregated information which can be used to e.g. query recently updated resources by user or a list of users that edited one specific resource.

The versioning information is stored within a separate SQL table in the underlying store and thus is not accessible via SPARQL. The underlying data structure is related to the changeset vocabulary[14], as it already stores all of the required information which can be used to export the data as a knowledge base using the changeset vocabulary.

4.7 Using Search Engines for Faster Access to Resources

The full-text search extension replaces the SPARQL-based search function with a scalable, near real-time search powered by Elasticsearch[15]. To achieve this, the full-text search extension makes use of a class based index structure. Hence, all triples of one class are put into one index. The search is triggered via an autocomplete function that suggests results as the user types a search term into an permanently accessible search bar. The matched results are then displayed directly as result of the search. If no results have been found, the user will be redirected to a more detailed search result page which simultaneously triggers a fuzzy search. This makes the search function more robust against typing errors and provides the possibility of restricting the result set to the previously defined classes.

5 Conclusions and Prospect

Within this paper we presented the use case electronic resource management in libraries and how it can be handled by using RDF resources and managing them with the generic knowledge engineering system OntoWiki. The OntoWiki application framework is open source. For the AMSL project, OntoWiki was extended by components which make it easier to adapt the complete system to the library domain e.g. by using the presented data model (section 4.1) in combination with data templates as presented in section 4.3. Due to the presented knowledge base concept in section 4.2, the complete system is multitenancy capable and can be used collaboratively amongst a library consortium. Using RDF

[14] Changeset Vocabulary: `http://purl.org/vocab/changeset/schema#`
[15] Elasticsearch: `http://www.elasticsearch.org/`

and Linked Data for the resource management and metadata management in libraries can further lead to a Linked Open Data infrastructure for library metadata exchange.

6 Acknowledgments

We want to thank our colleagues Lydia Unterdörfel, Carsten Krahl and Björn Muschal from Leipzig University Library, Jens Mittelbach from Saxon State and University Library Dresden (SLUB) and our fellows from Agile Knowledge Engineering and Semantic Web (AKSW) research group for their help. This work was supported by the European Union and Free State of Saxony by a grant from the European Regional Development Fund (ERDF) for the project number 100151134 (SAB index).

References

[BGM14] Dan Brickley, R.V. Guha, and Brian McBride. RDF Schema 1.1. W3C Recommendation, W3C, February 2014. http://www.w3.org/TR/2014/REC-rdf-schema-20140225/.

[BL09] Tim Berners-Lee. Linked Data. Design issues, W3C, June 2009. http://www.w3.org/DesignIssues/LinkedData.html.

[FMT+14] Philipp Frischmuth, Michael Martin, Sebastian Tramp, Thomas Riechert, and Sören Auer. OntoWiki - An Authoring, Publication and Visualization Interface for the Data Web. *Semantic Web Journal*, 2014. to appear.

[HDMA09] Norman Heino, Sebastian Dietzold, Michael Martin, and Sören Auer. Developing Semantic Web Applications with the OntoWiki Framework. In Tassilo Pellegrini, Sören Auer, Klaus Tochtermann, and Sebastian Schaffert, editors, *Networked Knowledge - Networked Media*, volume 221 of *Studies in Computational Intelligence*, pages 61–77. Springer, Berlin / Heidelberg, 2009.

[JAA+12] Tim Jewell, Jeff Aipperspach, Ivy Anderson, Deberah England, Rafal Kasprowski, Bob McQuillan, Tim McGeary, and Angela Riggio. Making Good on the Promise of ERM: A Standards and Best Practices Discussion Paper. Technical report, NISO ERM Data Standards and Best Practives Review Steering Committee, One North Charles Street, Suite 1905, Baltimore, MD 21201, January 2012. http://www.niso.org/apps/group_public/document.php?document_id=7946&wg_abbrev=ermreview.

[JAC+04] Timothy D. Jewell, Ivy Anderson, Adam Chandler, Sharon E. Farb, Kimberly Parker, Angela Riggio, and Nathan D. M. Robertson. Electronic Resource Management – Report of the DLF ERM Initiative. Technical report, Digital Library Federation, Washington, D.C., 2004. http://old.diglib.org/pubs/dlf102/.

[KBAL09] Georgi Kobilarov, Christian Bizer, Sören Auer, and Jens Lehmann. DBpedia - A Linked Data Hub and Data Source for Web Applications and Enterprises. In *Proceedings of Developers Track of 18th International World Wide Web Conference (WWW 2009), April 20th-24th, Madrid, Spain*, April 2009.

[TWSWG13] The-W3C-SPARQL-Working-Group. SPARQL 1.1 Overview. W3C Recommendation, W3C, March 2013. http://www.w3.org/TR/sparql11-overview/.

Quantitative Analyses in Global and Area Studies using Graph-based Filtering of Heterogeneous Catalogue Data

Thomas Efer, Ninja Steinbach-Hüther

Computer Science Department | Graduate School "Global and Area Studies"
University of Leipzig
04109 Leipzig
efer@informatik.uni-leipzig.de
ninja.steinbach-huether@uni-leipzig.de

Abstract: In this contribution we describe the objective and technical considerations of an ongoing early-stage interdisciplinary collaboration at the University of Leipzig for the preliminary part of a quantitative analysis within a project of the Graduate School "Global and Area Studies".[1] This project is located at the meeting point between approaches stemming from the e-Humanities and approaches coming from the Global and Area Studies. Our intention is to provide first insights into new, data-driven methodology for bibliography-based analyses. We want to show the utility of recent technology for data modeling and management in the Global and Area Studies.

1 Research Questions

Africa seems to have become a focus of higher academic interest in various disciplines among which the Area Studies play a considerable role in this debate. Hence, if interest in a certain region grows significantly, we are most often confronted with the need for publications from and on this region. But questions on the presence and reception of African (academic) books and their place in academic debates are often answered by emphasizing their marginal role and their quasi non-existence on a global level [Bg06]. Nevertheless, despite critical statements about the global ignorance towards African academic literature, we seem to lack an all-encompassing overview over its quantitative status and how this has developed for the last five or six decades in certain regions and fields of research. Against this background, this contribution is placed in the broader scale of a project of the Graduate School "Global and Area Studies" that exemplarily analyses the presence of African academic literature in Germany and France in historical perspective.[2] The key question behind this project is whether we are dealing with a

[1] The methodological approach presented here is part of a larger PHD-project to be submitted by N. Steinbach-Hüther at the Faculty of Social Sciences and Philosophy, University of Leipzig, for which the methodology will be developed further. The results presented here are a first step for the larger study.

[2] This rather European angle and perspective can be explained by the fact that direct publishing in European or US American publishing houses as part of the global North has been and still is very common despite

growing attention to Africa and the knowledge that is produced by its academics or whether African knowledge is still situated at the margins of global knowledge production.[3]

The project theoretically follows the concept of cultural transfer as was established by Espagne, Lüsebrink and Werner in the French 'German studies' as part of the cultural studies and an alternative to the traditional comparatistic [Es06]; [KS03]; [Lü08]; [Mi01]; [WZ03]. The methodological work combines various methods to accomplish an all-embracing reconstruction of the whole transfer process in both a quantitative and a qualitative sense [Es12].

The scientific aim of this project is firstly to examine the quantitative presence of African academic literature from the Social Sciences with the Humanities also included, but the Life Sciences are excluded. In this work we concentrate on the French part of the analysis only.[4] Therefore, we will present the preliminary methodological approach to a rather complex dataset that was compiled by the Bibliothèque nationale de France (BnF) and that is based on its bibliographical catalogue[5]. By evaluating the entries we will accomplish the aim of getting bibliography-based quantitative results as the foundation for further quantitative and qualitative research in the Global and Area Studies.

Building on that primarily work, it can be analysed in a second step to what extent "academic Africa" has finally paved its way to the French book market and correspondingly may have become an issue of academic debates, discussions and discourses. While the first question is rather quantitatively oriented, the second is based on its results and approaches the results in a qualitative way. In a nutshell: based on the analysis, it will be dealt with the question of inclusion versus exclusion of African knowledge from the academic knowledge debate. That is why this work follows two different aims that can be accomplished by different methodological strategies only.

What is especially interesting regarding this methodological handling of the dataset is the fact that the research question could not be addressed and fulfilled the way it is if it had been for leaving the analyses of such a huge dataset to the conventional way of doing computations and the very restricted set of traditional digital tools that can be found in usual office software. The insight should be carried out beyond the reach of Global and Area Studies that custom-tailored digital tools, forged in an interdisciplinary setting, can enable broader research questions. The e-Humanities pose a lot of interesting questions especially to the knowledge-centered parts of the Computer Sciences. In our case the question prevails how to model library data in a most fitting way for specialists of certain fields other than the library sciences. How to pave the way and build a common technological ground for similar projects?

---

numerous publishing initiatives during the last two decades to publish literature in African countries themselves.

[3]The research question is explained in detail at http://www.uni-leipzig.de/~ral/gchuman/en/units/graduate-school-global-and-area-studies/transnationalization-and-regionalization/doktoranden?tx_wecstaffdirectory_pi1[curstaff]=86

[4] For some of the results of the German part of the analysis see [MSH14b].

[5] http://catalogue.bnf.fr/

2 Datasets and Challenges

The raw datasets were compiled by the BnF from their catalog repository and according to four rather broad criteria:

- at least one of the authors of a publication has to be somehow connected to the African cultural region according to the denoted country or language code in the general catalogue ("catalogue général")
- the publications linked with these people became relevant only if published in the form of books/monographs
- the publications had to be published after 1950
- the publications had to be published in France[6]

To accomplish a corresponding filtering, the BnF first compiled a list of authors who have been assigned an African nationality and one of those for whom an African language skill cannot be ruled out[7]. For those authors of potential interest all those books (including their bibliographical data) were selected in a second step, for which the previously extracted authors acted as primary or (one of the) secondary authors. Both lists (51.266 authors and 64.574 books) were made available through a data set which is based on the INTERMARC format[8]. This textual format works with a "zone"-based mechanism for structuring entries[9] which can hold by themselves multiple values which are indicated by different "splitter sequences" like "$a". In the Excel-based output format some of those multi-values were already resolved into their own columns, while some remained combined. The individual records originate from different internal systems. This leads to the heterogeneous nature of the dataset with different levels of data granularity and completeness.

The retrieved lists use numerical identifiers for publications and authors but do not provide identifiers or other special notions for other entity types such as cities or publishers. Those are compiled on a textual level only and seem to have been manually collected without being fully normalized since spelling and expression variants persist throughout many fields.[10]

From this originates the need to account for the various common tasks of heterogeneous data integration such as data cleaning and normalisation.

---

[6] Excluded are books which are not published in France but that were bought by the BnF as well as books published in francophone countries other than France that also fall under the National Library's collective order (e. g. some African countries, (francophile) Canada and Switzerland. Some additional African authors (those with French citizenship that otherwise would have been excluded from the analysis) could be found via their language code.

[7] That includes people with an African language code (likely Africans) as well as people generically classified as "multilingual".

[8] http://www.bnf.fr/en/professionals/intermarc_format_eng.html

[9] Zones are fixed 3-digit numbers that denote a common field type. An example for possible sub-values for such fields and their delimiters can be seen here for zone 260, the publisher's address: http://www.bnf.fr/documents/pb-RIMB08_260.pdf

[10] This leads to the question about who actually collects/makes and gives the codes in national bibliographies and in libraries respectively.

The BnF provides their whole catalog dataset as linked open data in the RDF format[11] and explains the process of Linked Data publication in depth in [SWMD13]. This complete data dump (containing over 38 GB worth of uncompressed XML files) was loaded and inspected. It showed, that this dataset does not constitute a real advantageous alternative over our Excel files for several reasons:

- Dealing with the inconvenient size of the dataset in line with the need for a pre-filtering according to the above-mentioned criteria would have been an additional task for us.
- We saw problems of communicating the "raw" data model which consists of vocabularies from 20 different namespaces to non-computer-scientists in order to define a common starting ground. It proved too complicated to find clear transformation steps towards the research questions in this setting.
- While multi-valued fields are properly split (which gives us an evaluation basis for our splitting), no substantial normalisation steps seem to be taken. Full text field content stays full-text. Orthographic variation is not resolved.
- Semantic enrichment and linking in the RDF data set do not focus on the domain of our research question.

The last point is really important: While there exists for example a comprehensive geographical linking of books and places on the FRBR[12]-level of "Manifestations", its semantics seem to lie solely in the description and referencing of the work's subject matter and not in the organisational aspects (e.g. the publication location of the edition) that we are interested in. Even more interesting entities such as publishers are not linked or assigned a unique identifier.

But even with an extended linking of all relevant aspects, there could remain uncertainties concerning the achieved accuracy and completeness of the "semantification" process that may render it impractical to build accurate scientific methods on top of such a (semi)automatic external preprocessing. After all it may be better to perform such crucial steps by oneself to keep full control. So we refrained from using the RDF data set in our analyses. Yet it can be seen as a remarkable and much appreciated step of the BnF, that this substantial amount of internal data is provided within such an open technological and legal framework.

3 Implementation

For the cleaning and normalisation of entries as well as their reliable interlinking according to common properties there exists a plentitude of possible technological resources but still no definitive methodology or toolset. The consumer-specific way of handling and enriching heterogeneous data is in our case quite comparable to the idea of "self-service Linked Government Data" as presented in [MCP12], where raw publicly

[11] http://data.bnf.fr/semanticweb-en, subject to a special license: http://data.bnf.fr/licence
[12] Functional Requirements for Bibliographic Records, see http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

available data sources are enriched, linked and re-published by the data consumer. Its proposed workflow is based on Open Refine[13] (formerly Google Refine), a very capable tool for data cleaning and linking.

Such standard methods and tools are a big help to fix the issues in raw datasets. But the interconnected nature of a bibliographical dataset[14] renders them less useful, as they see the datasets primarily as independent tabular structures. Yet the biggest problem with such tools is, that all enhancement operations like normalization, splitting and merging as well as re-linking are irreversible and permanent (with the exception of linear undo/redo functionality), and lack options to properly document the intention behind them. This poses their most prominent disadvantage for an explorative academic scenario where hypotheses have to be tested thoroughly and in many cases have to be abandoned later on in favour of new ones, without starting over again on all the other aspects, too.

Besides that category of programs, specialized transformation and integration tools for library (meta) data exist, see for example Metafacture[15]. Their focus lies in the definition of static transformation workflows for changing datasets, whereas we need a flexible explorative system for a static dataset where we mainly aim at filtering data than to completely transform all entries into another format.

So we decided to build a completely independent toolbox to be able to design all aspects according to the scientific quality requirements for the Global Studies. The toolbox consists of a data store and integrated process chain and will be extended with a web-based visual analytics interface. While developing, a prototype-driven process allows for an exploration of the peculiarities of the dataset prior to fixing the complete strategy for approaching the research questions.

The process consists of several steps. Pre-Processing consists of converting all record entries to single linkable representants yielding the original field values. Then the field values where fixed parsing can be applied (like multi-valued fields) are transformed to represent specific human-interpretable values of determinable semantic. Finally there is an iterative process of creating machine-readable, unified, normalized and queryable values. The needed data store has to support short feedback cycles (preferably without demanding a fixed schema), it should be capable of modelling a “natural” networked view on the data so that normalization is not only lexically motivated (e.g. truncated full text search) but can derive new identity of entities and semantic concepts (publishers, genres and the like) by adding new entities to the datastore.

According to good experiences in previous smaller projects we chose the Property Graph model as fitting representation and the de-facto leading Graph Database Neo4J[16] as our backend. We decided to use JRuby[17] as programming environment since it enables us to

[13] http://openrefine.org/
[14] Different entity types such as people, places, subjects and organisations all contribute to the dataset in an individual way.
[15] https://github.com/culturegraph/metafacture-core/
[16] http://www.neo4j.org/
[17] http://jruby.org/, a Ruby runtime based on the Java Virtual Machine

use Java libraries (such as an embedded Neo4J engine) while we can at the same time benefit from Ruby's concise syntax and other handy options, e.g. the ability to extend the language with dynamic programming concepts to act as Domain Specific Language. The datasets are in the Office Open XML Format and can be read with the library ("ruby gem") roo[18]. Since this mode of accessing the data is cell-based and therefore very slow we decided to first transform the Excel files into a raw tabular JSON format to have a fast way to easily load the table into memory on demand.

The conversion of those raw tables into a graph then consists of multiple steps. The splitting of multi-fields and linking of books and authors via their IDs is a purely rule-based mechanism. Then follows an "interdisciplinary explorative rule mining" process, where a growing and alterable list of steps is employed to record transformations . The effects of those transformations can then be assessed in an analytics frontend. The work is still ongoing and can therefore only be briefly outlined at the moment:

The existing graph gives insights into the distribution of several field values, link patterns and other properties. The researchers have to decide if such patterns represent

- useful filter criteria,
- candidates for the merging of datasets,
- or just insignificant statistics

Based on that decision a set of graph queries for reading and transforming data and then altering the graph is built and successively extended. When some rules are later on deemed unsuitable, a reprocessing has to take place. We managed to improve the performance of Neo4J for that task significantly from hours to minutes by using "unforced" transactions[19]. This speed makes the re-processing of changed rulesets quite convenient. In addition, a partial re-processing from fixed-state copies can be introduced.

Using these query collections we can not only alter the existing entries but also add new entries for discovered entities and link them to the basic data sets according to field values. It will become far easier to then find a place of publication, a publishing house, a translation or a book of a particular genre if all variables in the writing have been attributed a new single qualifier which at the same time combines all the former qualifiers to one (such as one graph node for the differently written names of one publishing house). This process goes hand in hand with data exploration and gradually improves the dataset while keeping track of all performed actions. The goal behind this setup is to have a flexible environment not only to support our own research purposes but also to be able to take part in larger efforts of Knowledge Federation. So we took care to ensure that our graph based model can be easily exported into popular formats such as RDF or Topic Maps. For best practices in that field see for example the also JRuby-based approach in [BJSM10].

[18] https://github.com/Empact/roo
[19] http://components.neo4j.org/neo4j-kernel/1.9.M01/apidocs/org/neo4j/kernel/TransactionBuilder.html#unforced()

4 Towards a new Methodology

When we consider a simple example where we want to determine if a book is a novel or not, we can imagine plenty of different indicators that contribute to that decision: There are directly accessible properties such as the word "roman" attached to the title field. This is a quite common practice but still gives results that are far from being complete. Additionally we can include indirect approximations. If an author has written ten books and nine of them are already classified as novels we may want to interpolate this class for the tenth book too. This may apply even more for publishers specialised in a certain field. Finally there is the option to define certain context-sensitive rules: One could hypothesize that books with more than 2 authors are most likely not novels. This hypothesis can be tested beforehand using the already classified works and then be applied to define exceptions to other rules, revert a previous classification, notify the user or the like.

In general such a combination of several incomplete heuristics is a valuable tool in an explorative filtering workflow. The stepwise refinement of classifications gradually reduces the uncertainty of the final quantitative aggregations while it possibly also narrows down the selection of objects of interest to a manageable portion for manual qualitative analyses. To support this (methodologically rather experimental) course of action, there must be flexible tools and data-centric prototypes that are developed in close dialogue between Humanities scholars and computer scientists.[20]

All taken steps (data cleaning and normalisation followed by classification, inference and filtering and finally querying and aggregation) should be available through an integrated toolbox. In most cases all of them can be expressed in terms of simple graph queries and value assignments. Those queries make explicit and therefore document the process of data transformation and selection. They are ensuring reproducible final results while providing flexible means for further revision and alteration of intermediate processing steps – and that at purely computational cost and not with manual effort. By introducing such tools and data processing methods into the Humanities and Social Sciences, new questions can be posed and successively answered through data analysis. As already stated above, this methodological toolset is beneficial for quantitative studies as it makes possible comprehensive inquiries on large datasets as well as for qualitative studies for which proper examples can be picked in a representative way.

5 Preliminary results

The conducted work up to this point is characterized by an explorative approach to get to know the datasets better and to decide on the applicability and fitness of quantitative methods in the context of the research questions. Can we identify information on sticking points, continuities, discontinuities and breaks in the period of the study,

[20] If sufficiently funded and after evaluation, this may then lead to being able to reproducing the method in other interdisciplinarily oriented projects. Similar research questions could henceforth be addressed with an easier methodological starting point.

especially with regard to the chosen thematic, geographic and temporal profiles of the published material but also with regard to personal, lingual and editorial dominances?

For example how big is the influence of certain publishers or institutions in the whole publishing process? What kind of publishing houses supported the publishing of African academic literature throughout the years and which role did publications in their original language have in this process in comparison to translations? It will also be interesting to get to know which books from what kind of authors were published and where the authors came from. Can we identify changes in thematic, lingual and geographic orientation during the time period for example? In this respect, results of the quantitative analysis will show the conjunctures of the whole transfer process in matters of the progress in time, the chosen topics and the privileged original languages but also relating to the authors themselves and further actors such as translators and publishers.

This is directly linked to new methodological questions: Should the analyses be split by time periods, for example, to show geographical, lingual and thematic dominions as well as correlations between them? Regarding figure 1 that shows how *L'Harmattan* has constantly increased its share in the total numbers of the books in the given list in comparison to Karthala that seems to have a rather constant share of the books, another methodological question follows: Should large editorial players, that quantitatively determine the numbers, be analysed individually later on with respect to the research question to prevent the insight into representative results from being altered by internal publisher policies and positions towards African authors? It is an important lesson that such analyses can only grow to full potential when guided and directed by explorative data analyses. If conducted in an interdisciplinary way this will become even more fructuous.

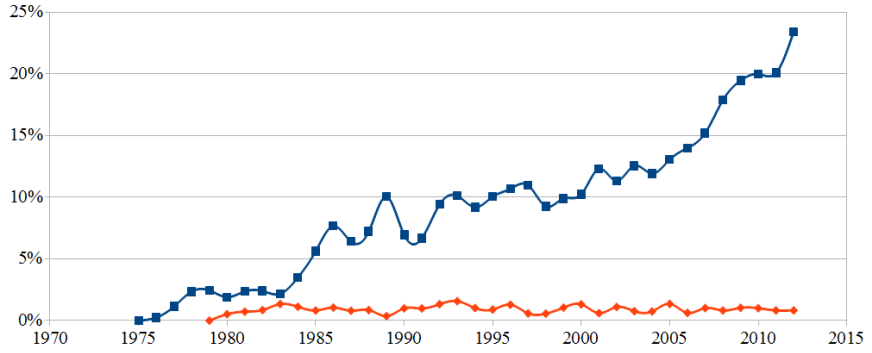


Figure 1: Percentage of all included books which are published by L'Harmattan (blue) and Karthala (orange), calculated by merging over 60 different renderings and typos of the publishers' and their spin-offs' names

In this project phase graph queries have proven to allow for quick interaction with the enhanced dataset, giving smaller result sets in mere milliseconds while returning graph-global aggregations after few seconds. The included full text indices for node properties allow an almost instantaneous display of time series for the occurrence of certain keywords in titles, notes and other such textual fields. This runtime performance allows

for the later creation of query interfaces to power a user-friendly interactive toolbench in the spirit of Visual Analytics, see [KKEM10].

About 45% of all listed books have a DDL class assigned. An analysis of the actually used categories (see Figure 2) shows as expected an uneven density and depth throughout the categories.[21]

By first intuition the categorisation space can be quite easily divided into segments belonging to one of the following three classes:

- definitive rejections (e.g. works from the natural sciences or works of fiction)
- definitive acceptance of categories (because they obviously lie in line with the research questions)
- a few borderline cases that soften the "accept"-category (e. g. books classified within the spectrum of medicine but oriented to the social aspects of the field rather than to the medical) or the "reject"-category respectively (e. g. books classified as belonging to the arts but treating its history and interpretation rather than being illustrated books)

All should be easily recognisable and determinable and form larger regions of common classes. Yet in practice almost every category yields strong potential to belonging to the "borderline" class, rendering the DDC (although partially very fine-grained) a tool which is still too coarse and detached from the research question as to be a solitary filter criterion. One reason for this is that in the Humanities and Social Sciences there is much impetus to reflect on or to analyze phenomena within a large spectrum of socially relevant topics and particular communities. Therefore they also contribute written (meta)literature to that field which in many cases has to be classified with the same category as its subject matter literature.

What is important to note so far is that in the field of the Social Sciences as well as the Global and Area Studies, it is very advantageous to make use of comparable programs that facilitate quantitative analyses the way it was shown. After all, the results of this analysis will show the most exact number - that is to say the quantity of books that was systematically possible to capture - of the presence of African academic literature from the Social Sciences and the Humanities published in France since the 1950s.

Of course, the results need to be counterproofed in a second step according to their relevance for the rather qualitatively oriented second research question but it would not have been possible to get the results in the first place if it had not been for the use of such an informatic tool to handle the sheer amount of data. It goes without saying that it was important to have generated rather too many results in the first place than too little and to therefore check the data in a second step according to its relevance.

[21] Since the DDC combines enumerative and faceted principles some of the emergent patterns may be representations of common semantics and some only co-incident.

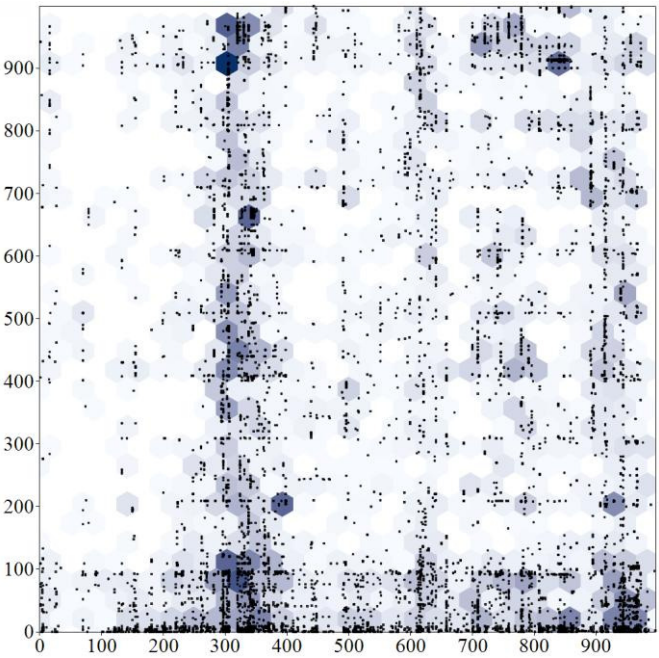


Figure 2: Graphical representations of occurring DDC numbers (first three digits as combined number on x, second three on y; background: hexbin choroplet map for easier density estimation)

It is also worth it to note that the method as such, even if that much effective, reproduces blurs as the search runs go in line with predefined parameters. As was stated above we should ask who actually collects/makes and gives the codes in national bibliographies and in libraries respectively. These codes are assigned manually and therefore according to a judgment of the person who actually completes the data record, which is why we have to keep in mind that this is one of the inevitable but reproduced blurs of the study.

For the computer sciences and their growing engagement in the e-Humanities it is important to take away from such projects the lesson on how crucial it is to offer flexible, normalized, disambiguated and entity-centric models of domain knowledge to researchers alongside semi-automatic methods for their research-centric unification and enrichment.

References

[Bg06] Bgoya, W.: Introduction – Scholarly Publishing: An Overview. In (Mlambo, A., Ed.): African Scholarly Publishing Essays, Oxford 2006; P: 5 ff.

[BJSM10] Bleier, A.; Jähnichen, P.; Schulze, U.; Maicher, L.: The Praxis of Social Knowledge Federation. In (Karabeg, D.; Park J. Eds.): Self-Organizing Collective Mind, Second International Workshop on Knowledge Federation, Dubrovnik 2010

[Es06] Espagne, M.: Jenseits der Komparatistik. Zur Methode der Erforschung von Kulturtransfers. In (Mölk, Ulrich, Ed.): Europäische Kulturzeitschriften um 1900 als

Medien transnationaler und transdisziplinärer Wahrnehmung. Bericht über das zweite Kolloquium der Kommission „Europäische Jahrhundertwende. Literatur, Künste, Wissenschaften um 1900 in Grenzüberschreitender Wahrnehmung“, Göttingen 2006.

[Es12] Espagne, M.: Comparison and Transfer: A Question of Method. In (Middell, M.; Roura, L., Eds.): Transnational Challenges to National History Writing, Basingstoke 2012.

[KKM10] Keim, D.; Kohlhammer, J.; Ellis, G.; Mansmann, F. (Eds.): Mastering The Information Age - Solving Problems with Visual Analytics. Eurographics, 2010

[KS03] Kaelble, H.; Schriewer J. (Eds.): Vergleich und Transfer. Komparatistik in den Sozial-, Geschichts- und Kulturwissenschaften, Frankfurt am Main/New York 2003.

[Lü08] Lüsebrink, H-J.: Interkulturelle Kommunikation. Interaktion, Fremdwahrnehmung, Kulturtransfer, Stuttgart/Weimar, 2nd edition, 2008.

[MCP12] Maali, F.; Cyganiak, R.; Peristeras, V.: A Publishing Pipeline for Linked Government Data. In (Simperl, E.; Cimiano, P.; Polleres, A.; Corcho, O.; Presutti, V. Eds.) The Semantic Web: Research and Applications, LNCS 7295. Springer Berlin Heidelberg 2012; P: 778-792

[Mi01] Middell, M.: Von der Wechselseitigkeit der Kulturen im Austausch. Das Konzept des Kulturtransfers in unterschiedlichen Forschungsrichtungen. In (Andrea Langer, Ed.): Metropolen und Kulturtransfer im 15./16. Jahrhundert, Prag/Krakau/Danzig/Wien 2001.

[MSH14b] Middell, M.; Steinbach-Hüther, N.: La présence de la littérature académique de l’Afrique en Allemagne. In (Espagne, M.; Lüsebrink, H.-J., Eds.): Transferts Culturels (exact title still to be given, to be published in 2014; Paris: Karthala)

[WZ03] Werner, M.; Zimmermann, B.: Penser l’histoire croisée. Entre empirie et réflexivité, in: Annales, 58/1. (2003); P: 7-36.

[SWMD13] Simon, A.; Wenz, R.; Michel, V.; Di Mascio, A.: Publishing Bibliographic Records on the Web of Data: Opportunities for the BnF. In (Cimiano, P. et al. Eds.) ESWC 2013, LNCS 7882; P: 563–577

BPM im Öffentlichen Sektor

(BPM@PublSect)

Workshop zum „BPM im Öffentlichen Sektor“

Organisatoren

Konrad Walser / André Göbel

Berner Fachhochschule / Hochschule Harz
E-Government-Institut / FB Verwaltungswissenschaften
Morgartenstrasse 2a / Am Domplatz 16
CH-3000 Bern 22 / D-38820 Halberstadt
konrad.walser@bfh.ch / agoebel@hs-harz.de

Das Geschäftsprozessmanagement nimmt im Öffentlichen Sektor eine immer größere Bedeutung ein. Die Gründe mögen wie folgt lauten: E-Government, E-Health, E-Education, Effizienzdruck, fehlende Mittel des Öffentlichen Sektors sowie erwünschte Effektivitäts- sowie Wirkungssteigerungen des Handelns im Öffentlichen Sektor. Dies macht die Thematisierung unterschiedlicher Facetten des BPM im Öffentlichen Sektor besonders interessant. Spannend ist dabei die Frage, welches Steuerungsmodell im Öffentlichen Sektor vorherrscht und was dieses Steuerungsmodell für Einflüsse auf die Entwicklung und Ausprägung des BPM im Öffentlichen Sektor hat.

Dies ist die zweite Ausführung des Workshops zum BPM im Öffentlichen Sektor, der von der GI-Fachgruppe Verwaltungsinformatik ausgerichtet wird. Vor einem Jahr wurde der Workshop alleine für die Öffentliche Verwaltung durchgeführt. Dieses Jahr wurde das „Einzugsgebiet“ für mögliche Beiträge auch auf das Gesundheitswesen und Non-Profit-Unternehmen des Öffentlichen Sektors ausgedehnt. In der Tat hat sich dies bewährt. Neu wurden auch Beiträge aus dem Krankennhausbereich und aus dem Bereich Universitäten eingereicht.

Es ist uns gelungen ein spannendes Set an Beiträgen und ein spannendes Programm zusammenzustellen. Die Themen reichen von der BPMN-Unterstützung von WTO-Ausschreibungen bis zur Thematisierung von Prüfungsprozessen sowie der Modellierung von Drittelprozessen an Universitäten (Entwicklung eines entsprechenden Referenzprozessmodells). Wieter wird eine Prozess-Fallstudie eines privaten Unternehmens im Bereich Supply Chain Management im Hypothekarbereich in Relation zu einem großen E-Government-Projekt dargestellt. Weiter wird die Geschäftsprozessdokumentation in föderalen Strukturen thematisiert, die Standardisierung im BPM-Bereich in föderalen Strukturen im Hinblick auf eine wissenschaftliche Fundierung untersucht sowie die Ausprägung des Geschäftsprozessmanagements in der Wirtschaftsförderung in einem deutschen Bundesland thematisiert. Interessant ist ferner ein Beitrag zur Ergänzung von Geschäftsprozessen im E-Government basierend auf den Errungenschaften des Semantic Web (Linked Data). Geschäftsprozessmanagement kann nicht ohne Mitarbeiter eingeführt werden. Ein weiterer Beitrag thematisiert Mitarbeiterwiderstände und -einbezug in der Polizeiorganisation eines deutschen Bundeslandes.

Effective Business process documentation in federal structures

René Schumann, Serge Delafontaine, Caroline Taramarcaz, Florian Evéquoz

Institute of Information Systems
HES-SO Valais / Wallis
Rue de Technopole 3
3960 Sierre
rene.schumann@hevs.ch, serge.delafontaine@hevs.ch
caroline.taramarcaz@hevs.ch, florian.evequoz@hevs.ch

Abstract: The Swiss eGovernment strategy promotes business process orientation of the public administration. Therefore public administrations have been strongly encouraged to document their business processes. The eCH standards have defined the foundations for the modeling. But in particular in smaller cantons and municipalities the business process identification and modeling can be considered as a serious effort, relative to the organization size. In this article we outline means for the efficient process identification and modeling. Also we discuss the short- and long-term benefits that public administration can obtain from modeling their business processes.

1 Introduction

The Swiss eGovernment strategy [Bun07], adopted by the federal council in 2007, is a process-oriented strategy. The goal of the strategy is to allow the administration, enterprises and citizen to carry out their transactions with the authorities by electronic means. Moreover, the strategy explicitly promotes business process orientation of the public administration. Therefore, there is a strong incentive for public administration to modernize their internal procedures and transition from a purely functional organizational to a business process management approach. To attain this goal, the eCH association[1] plays a crucial role. Created as a public-private partnership, its mission is to define standards and best practices in eGovernment, that will be enforced and implemented by public administrations at all three federal levels (federal, cantonal, municipal) in Switzerland.

To be able to implement cross-level electronic procedures (like for example, the building permit procedure which requires the involvement of several services of the public administration at different levels, usually municipal and cantonal), administrations need a way to understand one another's way of handling administrative procedures. In other words, process documentation in a common language is required on all levels of the administration. The eCH association has defined already a number of standards for process documentation

[1] http://www.ech.ch

for the public administration. For instance, they have fixed BPMN 2.0 as a common language for describing processes [eCH11]. Also eCH has published modeling conventions [eCH13]. This offers some advantages, e.g. in contrast to the National Prozessbibliothek, as the process models, are based on a common base.

Even though Switzerland is considered to be a small country (ca. 8 Mio. inhabitants), its administrative structure is rather fine-grained. Switzerland's federal structure consists out of 26 cantons, which again are formed by ca. 2400 municipalities. Due to the federal principle, and the strong emphasis of local autonomy, processes for providing similar services can be designed differently in different places. Thus, there is no *one fits all* process model that can be used for all entities offering a particular type of service. Each canton and each municipality is responsible to identify their process models and also for the way how they have organized their processes. Modeling the processes in a way that is conform to the eCH standards, requires a particular modeling expertise. But it has to be stated that for a large number of municipalities cannot provide this expertise. In particular, as a lot of Swiss municipalities are small. There exist about 1000 (out of about 2400) municipalities with less than 1000 inhabitants. These small municipalities often employ only 2-3 full time equivalent persons in administrations, which typically do not have the required expertise for modeling their processes in an appropriate way. But also in larger municipalities or on the cantonal level, which in principle have sufficient resources, the modeling of processes should not be done by each and every entity, as a) this would be inappropriate resource intensive, and b) for most administrations the process models will be very similar, and therefore the modeling will be redundant.

For the successful gathering of process models on the level of cantons and municipalities an efficient process identification is key. In particular for municipalities the resource investment is relatively large, and can only be considered as a long term investment. Thus, it is not only important to perform the process identification and modeling effective, but also find ways how to create additional values for municipalities and their citizens based on the process models, short-term.

Within this article we are addressing these issues. We outline here best-practices that have been identified as useful, with a particular focus also to be applicable in smaller organizations. Our focus in this article to give an overview about the field and aspects that needs to be considered, to enable an effective process documentation.

In Section 2 we are going to discuss means to identify and model processes effectively. Nevertheless, we will not put a particular strong emphasis to this aspect, as this has been discussed in the literature quite extensively. We will refer there to the existing body of research, see e.g. [Har07, Sil11]. Also there exists a market for service providers that offer their expertise to companies and administrations.

In Section 3 we will put a special focus on the aspect of sharing process models. Sharing is an promising mean to avoid redundancy and waste of resources in the modeling of processes. We will discuss this point in more details as we consider this aspect of business process management (BPM) as one that has not been explored in to depth, until now. Interesting projects have been established, e.g. in Germany the *Nationale Prozessbibliothek*[2]

[2] http://www.prozessbibliothek.de

or in Switzerland the *Prozessaustauschplattform*[3]. Within this section we will first discuss required means for quality management of process models (Section 3.1). Then we are going to outline the process platform ech-bpm, in the way it is currently developed and deployed in Switzerland (Section 3.2).

In Section 4 we are going to discuss benefits of process modeling. We will pointing out effects have been observed and (partially) documented in the literature. But will focus on ideas how the created process models could be used to create additional benefits.

A particular challenge in this research is the methodology to ensure that the expected results, in terms of a more efficient process identification and modeling, can be realized by the means proposed here. In particular for aspect like process sharing are rather new, and not much data has been collected, yet. The platforms for sharing process models are currently being under development or just have prototypically deployments. Nevertheless, the value of sharing information and models is in general not doubted. Also a direct comparison between two municipalities is hard to achieve. Another approach would be a comparison within one administrative entity before and after the introduction of the means we have discussed here, but this is neither feasible, as the major part of process identification and modeling is a singular event. After process models exists they need to be maintained and partially optimized, but the required efforts is often magnitudes lower than for the initial process identification and modeling. What can be done is the comparison between the effort of *early adopters* of process modeling, i.e. the administrations that started very early with the process identification and modeling, with the efforts for identifying and modeling the processes by the *late followers*, i.e. the administrations that will identify and model their processes after the infrastructure for sharing process models has been created and is well-established, and therefore also provide a rich base of information. As it has become clear by now, we are currently in the phase of the *early adopters* and pioneers with whom we are currently collaborating to create the process platform. Thus, the article lacks a proper empirical evaluation of the gains in effective process identification and modeling. Nevertheless, we are convinced that the means discussed in this article provide an additional value to the research community, and that the empirical evaluation needs to be done later, when the infrastructure has been established.

2 Effective identification and modeling of business processes

When it comes to identifying and modeling business processes, we find it very important to keep in mind that these activities are the requirements of the later stages of analyzing, transform and optimize the processes within a company or administration [Sil11].

[3] `http://www.ech-bpm.ch`

2.1 Method selection for identifying and modeling process models

In the literature a variety of methods for identifying and modeling business processes can be found. There are, nonetheless, points that always have been found in different methods and can be considered as core elements. Also some have been identified as best practice to ensure the effectiveness of the identifying and modeling of the business process. A collection of best-practices for modeling business processes is presented in [Sil11]. According to [Har07] the core elements of the modeling methods are the following.

1. Identify the processes to model
2. Prepare the collect of the information
3. Collect the information about the processes
4. Create the processes models
5. Validate the processes models

As mentioned before, within this papers we do not focus on the creation of process models, themselves. There exist a large body of research, and we have just stated two sources, which already provides meta-analysis of effective business process gathering and model creation. We complement this general guidelines with experience we obtained in our own projects, in particular with the public administration in municipalities in Switzerland, to validate that this guidelines lead to a practical procedures for creating process documentations are useful for public administrations. We outline our experience in the reminder of this section.

2.2 Process documentation projects - lessons learned

In the context of the eCH process platform project, described in greater details in Section 3.2, a goal was to acquire process documentation from administrations early on in the course of the project. This was meant as a concrete practical application of eCH standards for BPM and as a way to learn about the challenges administration will face when documenting processes. We need this hand-on knowledge to come up with practical guidelines for effectively documenting processes in public administrations. Although those results qualify as *lessons learned* and are not the result of robust scientific experiments, we believe they provide a qualitative value to the BPM practitioner's community in the public administration.

Two municipalities were involved in our process documentation projects, namely the city of Gossau (SG) and the city of Dübendorf (ZH). Dübendorf provided 14 process documentations in a first project. For that project, the municipality had 2 employees in charge, and 2 external consultants. As the municipality was already ISO 9001:2008 certified, they had process documentation at hand, although not in BPMN. Gossau provided 54 processes

modeled from scratch by two collaborators, without any external help (consultants). Objectives of the process documentation projects were to obtain processes that would showcase the correct use of all eCH standards related to BPM. Therefore, the projects were realized in two stages: (1) modeling and (2) quality assurance of the models. The modeling was done primarily by public servants, and the quality assurance was outsourced to consulting companies well aware of the relevant eCH standards. A final human validation step was taken by the ech-bpm project team when preparing the processes for publication on the ech-bpm platform. Deliverables of the process documentation projects included field reports written by the consultants in charge of the quality assurance and addressed to the municipality and the ech-bpm project team. Those documents describe some problems found in the process models. In particular, they hint to the fact that business process models are mainly communication tools, meant to serve as a documentation that can be provided to all stakeholders (current collaborators, new employees, IT service providers, etc.). Therefore they should be well understandable. This is even more true when those processes are meant to be published in an open process library. We summarize next the main recommendations we could make from these experiences:

- *Follow modeling conventions*. Dübendorf reported that they quickly agreed on conventions to coordinate the independent modeling of four collaborators, two of them being external. Moreover, they organized common sessions, where they handled such cases as granularity of modeling and exceptions.
- *Avoid ambiguities* in naming elements: for example activities within a process that bear the same name as the whole process should be avoided. Similarly, abbreviations should be avoided or at least well documented. The majority of findings in the quality insurance at Gossau fell in this category.
- *Anonymization*: names of persons or external companies should be anonymized.
- Dübendorf reported that the *BPMN skills* in the participating public administration needs to be already good (i.e. not beginner-level) in order to understand some of the eCH standards (e.g. the modeling conventions [eCH13]).

3 Sharing Models

The value of sharing information has been documented, even before terms like shared-economy have become popular. Even though sharing processes and process related information is not completely new, it has been less explored in commercial environment. Only for supporting actions for the value creations processes, standards and best practices have been documented and shared in commercial settings, e.g. ITIL for managing the IT [BÏ3]. Processes and procedures within the core value creation processes have been documented for internal purposes, but have been considered as secret or at least unaccessible from outside. In case of public administration the situation differs in at least two aspects.

First: The public administration has to provide a large number of different services, which are all considered as public tasks, ranging from domains like defense, education or social

welfare, depending on the type of administration several hundreds different services have been identified [DES14]. So while in company the number of activities considered to be part of the core competences are often only a few dozen, the number in public administrations is magnitudes larger.

Second: Most processes of the core administration services are not secret. The value creation of a public service is, mostly public available. Note that for instances the processes in the German process repository (Nationale Prozessbibliothek) are only accessible for the German public administration. But the processes in the Swiss process platform are publicly available as it follows the open data principles. In principle it is easier in the public administration to collect business processes in repositories and share them. Process repositories usually provide functionalities like storing business processes, categorizing and querying and versioning. More advanced functionalities, often at the level of research prototypes, include similarity search, variants management, merging, refactoring, requirements mining or support for the reuse of process patterns [DRH11, Awa07, CJ07, YDG10, ZZ10, WRMR11]. The use of repositories for public processes can therefore facilitate the research about process repositories, which have, up to now, mainly used within companies or their service providers. Potential fields of research can be the field of repositories itself and also the quantitative and comparative research on real-world processes, which often requires to have a larger number of processes available.

3.1 Quality management of process models

Before process models should be shared it is important to make sure they are of a high quality. Otherwise the value of sharing process models is diminished, or even worst becomes negative, as false information is shared and eventually applied in several places. Quality assurance is feasible for the Swiss process platform, as a uniform process representation is facilitated by the eCH standards, in particular [eCH11, eCH13].

Modeling constraints for process models can origin from different sources, like the eCH standard [eCH13] or other sources, e.g. [Sil11, GL10]. The overall quality of a process model depends on two aspects a) its correct content and b) its formal correct representation. Both needs to be validated within the quality management process. Our goal here is to automate the validation of the formal correct representation of a process model. This is done by the *formal validation*. Within the formal validation three different aspects have to be validated.

1. **Schema validation**, i.e. making sure the model is a valid BPMN 2.0 model
2. **Style validation**, i.e. making sure the modeling conventions have been respected
3. **Structural validation**, i.e. trying to find structural problems, like deadlocks and lacks of synchronization

Note that the formal validation explicitly does not address the content of the model. The *content validation*, i.e. checking if the process is modeled correctly, cannot be automa-

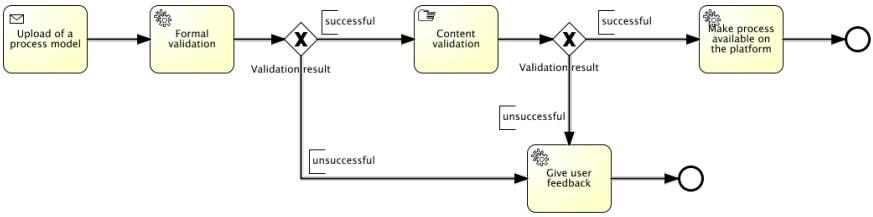


Abbildung 1: Quality management process of the Swiss process library (simplified)

ted, as it would require that the validation tool already has a correct representation of the process, which is not the case, and also would make modeling obsolete. The *content validation* needs to be done by a human expert that has insight knowledge about the process, so she can judge if the model reflects the process in reality correctly.

It becomes clear that the task of the human expert is crucial for the overall quality management process, in which the expert can be considered as a *scarce* resource. Thus, her time needs to be used carefully. Therefore it is required that only formally correct process models are presented to the expert, so she can concentrate on the content validation, and is not distracted by focusing on models with structural or style problems. Based on this principle the quality management process of the Swiss process library, depicted in Figure 1 has been designed. Note, we consider that after the user receives feedback, he has to upload a new version of the process model, and the quality management process is restarted from the beginning. To implement the formal validation it needs to be decomposed into single checks that can be automated. We will discuss this here in brief. We have implemented all these validation checks in form of web services, which allows a simple integration into the quality management process, but also offer the option to use the services for different purposes, or to add or exchange single checks if necessary. Due to space constraints we are not going into much detail here. The validation tool can be accessed via `http://val2pro.hevs.ch`[4].

3.1.1 Schema validation

The first important aspect of the formal validation is to assure that the process model is a valid BPMN 2.0 model. To validate this, we check, if the XML representation of the model complies with the XML schema defined in the BPMN 2.0 standard [OMG11]. If this check failed, the process is not a standard BPMN 2.0 process model, and subsequently does not comply with eCH standards [eCH11] and cannot be shared on the platform. If the first check is passed, the structure and the style of the process model needs to be validated.

[4]Fast validation is recommended, as the full validation uses extensively queries in dictionaries with consumes a lot of time; performance improvement is planned

3.1.2 Structural validation

We base our structural validation on the heuristics presented by Gruhn and Laue [GL10]. They have derived 6 patterns that could lead to problems. A process model will be checked if non of the patterns as defined in [GL10] can be found in the process model. We used *Refined Process Structure Tree* (RPST) introduced in [VVK08]. The goal of RPST is to find a unique, modular and as fine as possible decomposition into a hierarchy of sub-workflows of a given workflow. The graph-based process description is translated into a block-based process description. The idea is to create the RPST of the process and then use the different blocks of this tree to analyze the underlying structure of the process. This allows the detection of deadlocks and lack of synchronization problems. For example we investigate blocks that begins and ends with a gateway to control if the both gateways have the same type. If this is not the case it indicates that a structural problem could be persistent. To compute the RPST from the BPMN diagrams we have written a parser to use the jbpt library[5].

3.1.3 Style validation

The translation of the modeling conventions [eCH13] is probably the most challenging aspect, as the current eCH standard is addressing descriptive process models [Sil11]. Thus, the target audience of the standards are primarily modelers and less tool providers, as it is relatively vague and needs to be operationalized.

Up to now, we have identified and implemented 72 different style validation tests. All these tests are implemented by analyzing elements in the XML representation of the process model. Analysis are typically done by parsing the XML representation for the occurrence of particular patterns or anti-patterns. We cannot give more details on these patterns here, due to space limitations. To parse the XML we used a Java library named jsoup[6]. It is a Java/HTML Parser, but works very well with XML files, too. This library use CSS, DOM, and jquery-like methods to provide a convenient API for extracting and manipulating data

3.2 The Process Exchange Platform in Switzerland

The eCH process platform[7] is primarily targeted towards public servants involved in business process management, e.g. assuming the role of Chief Process Officers (CPO), in their organization. The vision of the platform as defined by the eCH association is that it should foster the use of BPM in Swiss administrations, accelerate the creation of skills amongst public servants, and foster the reuse of existing processes in other administrations to avoid redundant developments (i.e. harmonization). To implement this vision, the platform is designed around three main parts: (1) Resources, (2) Community, and (3) the Process Library, which is its core component. We briefly describe hereafter the different parts.

[5] `https://code.google.com/p/jbpt/`
[6] `http://jsoup.org`
[7] `http://www.ech-bpm.ch`

The Resources are typically documents describing standards and best practices for implementing BPM in Switzerland. Those resources are mainly prescriptive eCH documents and additional didactic material. They are made available for everyone, curated on the platform, and presented in a didactic fashion. The Community part of the platform encompasses all the features meant to foster the active participation of the actors of BPM in Switzerland. There, people can share their experiences, ask for advice, comment existing processes, standards or methodologies, and get in touch with one another. It is also foreseen that the content of the platform (Resources and Process Library) is maintained up-to-date thanks to the active participation of the community. Finally, the process library is the heart of the process platform. It contains the business process descriptions of the public administrations in Switzerland. For each process, a BPMN 2.0 model is available, as well as a textual description following a common template. Additional metadata for each process is also given, e.g. name of the administration responsible for the process, contact person, classification of the process according to the Swiss reference cartography framework (eCH-0145 Aufgabenlandkarte) [eCH14], and so on. Navigation in the library will follow a faceted search approach, taking benefit of the rich metadata available and semantic technologies. Moreover, an API will be developed to allow administrations to upload their processes easily on the platform. This API will be implemented by vendors of BPM modeling tools and will enable *one-click uploads* of processes along with associated metadata to the process library.

The platform is currently under design and development. However, a scaled-down version of it was already published in October 2013 under `http://www.ech-bpm.ch`. This beta-version contains information about the project, minimal content in the Resources section (e.g. eCH standards) and Community section (e.g. crowdsourcing platform for collecting ideas and requirements for the development of the final process platform). At the core of the platform, the Process Library contains around 80 processes from two public administrations (Cities of Gossau and Dübendorf). On the actual ech-bpm beta platform, the processes from those pilot projects are not yet stored in a fully-fledged process repository, but as static content pages. However, this gives a glimpse of what the final platform will offer. Process diagrams can already be viewed and downloaded. Textual descriptions of processes obtained from pilot documentation projects are provided along with relevant metadata. But navigation and search are still kept minimal in this beta version of the platform. Between its launch in October 2013 and April 2014, the website received about 3500 visitors and more than 16,000 hits, indicating a strong interest on the part of the BPM actors concerned with eGovernment in Switzerland.

4 Benefits of modeling business process

Up to now, we have mainly pointed on the aspect that modeling business processes is an effort. This is, of course only one aspect. Modeling business processes is not done without a cause. Benefits from documenting business processes have been commonly documented in various domains, and we have already mentioned some goals of documenting business processes above. In the following we are going to highlight two different aspects. First,

the benefits of having created and available a documentation of the business processes. Documentation activities often allow to reflect and optimize business processes, but also the availability of process documentation can be useful for communication or teaching purposes.

We are also pointing out that it is necessary to provide benefits out of the BPMN 2.0 models. In particular we will provide an example how to create short-term benefits for municipalities and their citizens, out of the XML representation of business process models.

We are not addressing process modeling for the purpose of automating business processes here. The maturity of BPM in the administration is currently at a level, where processes get mainly modeled for documenting purposes. This is also reflected in the eCH standards. As pointed out the modeling conventions for process models currently addressing purely descriptive BPMN [eCH13]. Also proper analysis of business processes and corresponding planning of required resources are currently seldom found in public administrations.

4.1 Value of documentation

As McGregor stated "You cant change what you cant see [McG05] BPM documentation allows an organization to gain the knowledge , or, to take McGregors words, to *see*, how she works from the inside, first when modeling and then when exploiting the models.

That knowledge can be used to obtain some return on investment, e.g. by analyzing and optimizing the processes. But as Chow et al. [CMR07] pointed out, a government can expect different ROI than just money. Responsiveness, compliance, reaction times and low amount of process failures can be improved with the help of well documented business process models.

Chow et al [CMR07] proposes three distinct values delivered by the use of BPM, that are:

1. Efficiency: Deliver more, better, faster and cheaper than your current alternative.
2. Control: Consistently knowing the current status and outcome of your processes.
3. Agility: Ability to adapt quickly to changing world conditions.

The efficiency can be measured by metrics like speed or capacity. It reduces operational costs, improve productivity, impact resource utilization and improve the services and the overall quality. The control can be measured by metrics like the rate of non-compliance or the Service Level Agreement (SLA) failure rate. It ensures the compliance, improve visibility and manages the outcomes of the processes. The agility can be measured by the speed to create and change processes. It improves the capacity to change with the clients demands.

The value of process documentation can also be compared with measures of the success of a process modeling project. A case-study was conducted by [BGR05], examining 9

different modeling projects. Based on empirical data, they present success measures of a process modeling project:

- Model quality, i.e. the extent to which all desirable properties of a model are fulfilled to satisfy the needs of the model users in an effective and efficient way
- User satisfaction, i.e.the extent to which the model users believe process modeling fulfills the objectives that underlay the modeling project]
- Individual impact, i.e. how process modeling has influenced the Process stakeholders; those who have a role in the processes being modeled.
- Process impacts, i.e. the overall effect of process modeling on the processes modeled.
- Project efficiency, i.e. the process modeling projects ability to maximize the obtained outcomes in relation to the invested resources.

4.2 Using process models to provide additional value and services

We are aware that the definition of the process models creates significant efforts for the public authorities in Switzerland, in particular for the municipalities. Even though we have outlined in this article a review of techniques that can reduce these efforts, still it will create extra efforts for the involved entities. But most of the positive effects of the availability of process models can only been realized in a long-term time horizon. In particular for the municipalities, for which the efforts is relatively bigger than for a canton or the federal state level, it is important to provide means to also have short-term benefits out of their models. Thus we must stimulate the usage of process models that provide short-term benefits. We hereby consider providing more and new information or services to ease the interaction with the public administration. In this article we show one example how process models can enable better services to the citizens of a municipality, and also have the potential to reduce efforts and costs within the administration itself, short-term.

We consider that for each activity within the process model the required data in- and outputs are annotated in the BPMN 2.0 model. Based on this annotated process model we can compute the required documents to initiate a process instance. A document is required, iff it is a required input of an activity and has not been created as an output by a preceding activity. These required inputs are presented in form of a checklist. This checklists could be directly given to the citizens interested in this service, e.g. by providing them online on the web page of the municipality. A person that wants to obtain a service (delivered by executing a process instance) can use this checklist to tick-off all inputs and prepare their request better, before they go to the administrative office. This could also decrease the process execution time and the workload on the public servants, as the efforts for requesting documents from the citizen is reduced.

In the following we give an example of a process shown in Figure 2. We have created this example, by adding the in- and output annotations to an existing process we obtained from

#	Activity name	Inputs	Outputs
1	Inform all employees of HR	Termination	-
2	Inform the mayor	-	Mail
3	Check if the employee has informed his superiors	-	-
4	Create a termination confirmation	-	Confirmation
5	Send the information to the employee	Confirmation	Confirmation
6	Create a checklist and tracking the ongoing discharge	-	Checklist
7	Create a draft of a work certificate	-	Draft certificate
8	Revise work certificate	Draft certificate	Certificate
9	Get signature of Mayor and head of office	Certificate	Certificate
10	Arrange exit interviews for personal service and Mayor	Certificate	Certificate
11	Give the work certificate to the employee	Certificate	Certificate
12	Perform the exit interview	-	-
13	Archive the dossier	-	-

Tabelle 1: In- and outputs of the example process

the Swiss Process Library[8]. The activities are annotated with the in- and outputs as shown in Table 1.

As it can be seen, the execution of the process has intended side-effects. There exist outputs by process activities that are not *consumed* by subsequent activities. An example of such a desired site effect is that a confirmation is send to the employee. The resulting checklist is shown in Figure 3.

5 Outline

Within this article we have pointed out that the modeling of business processes in the public administration can provide benefits a) while modeling the processes, b) short-term benefits by valorizing the existing process models, and c) long term, by enabling a more efficient collaboration between all levels of the federal administration in Switzerland. But, in particular for smaller cantons or municipalities the identification and the modeling of the business processes has to be considered as a significant extra effort. Therefore the efficiency in documenting business processes and also the collaboration between administrations, by sharing process models needs to be enhanced, in particular as public budgets are shrinking. We have presented techniques to increase the efficiency in identify and mo-

[8] We have translated it, for a harmonized use of language within this article.

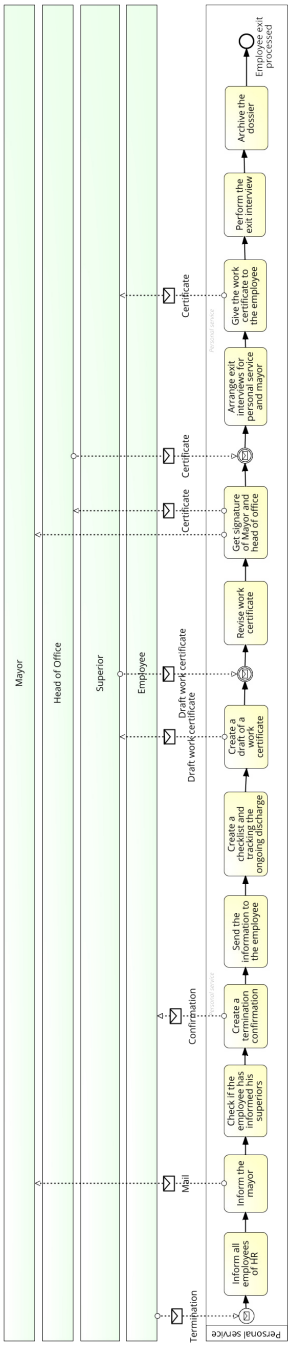


Abbildung 2: Process model for handling a termination of an employee

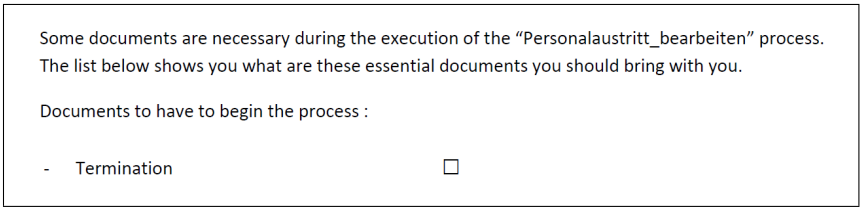


Abbildung 3: Automatically generated checklist from the process model, shown in Figure 2

del business process. Also we have outlined that the value of process libraries and the sharing of business processes has a particular role in BPM in the domain of eGovernment. Also we have discussed means how the efficiency gains in the public administrations can be empirical measured, which requires a long term comparative study, so sound empirical results can only be available in a few year, after the infrastructure is established.

In particular the empirical research will be in the focus of one branch of future research that needs to be done, to demonstrate that the infrastructure (process maps, the process library , the quality management process) offers concrete additional values for the public administration. Also we are currently investigating how repositories can be used to increase efficiency in searching for existing process models and how existing models can be used in the process of creating new process models, and support this process with semantic annotated process models, and a process designer that take advantage of existing model repositories.

Acknowledgement

This work has been supported by the following projects: eGov Innovation Center, funded by HES-SO Call SmartCity, grant AGP-32679; eCH-Prozessaustauschplattform, National Prioritized Project B1.13, funded by eGovernment Schweiz Aktionplne 2012-14, grant 9439; Val2Pro, funded by RSCO-ISNet, grant ES-ISNET13-05.

Literatur

[Awa07] A. Awad. Bpmn-q: A language to query business processes. In *Proceedings of EMISA. 2007*, Seiten 115 – 128, 2007.

[BÏ3] Roland Böttcher. *IT-Service Management mit ITIL - 2011 Edition*. Heise, Hannover, 3rd. Auflage, 2013.

[BGR05] Wasana Bandara, Guy G Gable und Michael Rosemann. Factors and measures of business process modelling: model building through a multiple case study. *European Journal of Information Systems*, 14(4):347–360, 2005.

[Bun07] Geschäftsstelle E-Government Schweiz Bundesrat. E-Government-Strategie Schweiz, 2007.

[CJ07] K. Choi und M. Jang. An xml-based process repository and process query language for integrated process management. *Knowledge and Process Management*, 14(4):303 – 316, 2007.

[CMR07] Linus Chow, Charles Medley und Clay Richardson. *BPM and Service-Oriented Architecture Teamed Together: a Pathway to Success for an Agile Government*, Seiten 33 – 54. Workflow Management Coalition, 2007.

[DES14] Serge Delafontaine, Florian Evéqouz und René Schumann. Approaches for Mapping E-Government Services. In Dagmar Lück-Schneider, Thomas Gordon, Siegfried Kaiser, Jörn von Lucke, Erich Schweighofer, Maria A. Wimmer und Martin G. Löhe, Hrsg., *Gemeinsam Electronic Government ziel(gruppen) gerechet gestalten und organisieren. Proceedings Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2014*, Jgg. 229 of *Lecture Notes in Informatics*, Seiten 97 –109. Gesellschaft fr Informatik, 2014.

[DRH11] Remco Dijkman, Marcello La Rosa und Reijers Hajo. Managing Large Collections of Business Process Models - Current Techniques and Challenges. *Computers in Industry*, 63(2):91 – 97, 2011.

[eCH11] eCH. eCH-0074 Geschäftsprozesse grafisch darstellen - Der Einsatz von BPMN aus Geschäftssicht., 2011.

[eCH13] eCH. eCH-0158: BPMN-Modellierungskonventionen fr die öffentliche Verwaltung, 2013.

[eCH14] eCH. eCH-0145: Aufgabenlandkarte der öffentlichen Verwaltung der Schweiz, 2014.

[GL10] Volker Gruhn und Ralf Laue. A heuristic method for detecting problems in business process models. *Business Process Management Journal*, 16(5):806 – 821, 2010.

[Har07] P. Harmon. *Business Process Change A guide for Business Managers and BPM and Six Sigma Professionals*. Morgan Kaufmann Publishers, 2nd. Auflage, 2007.

[McG05] Mark McGregor. *Where BPM is headed beyond best practice in BPM and Next Practice*. Meghan Kiffer Pr, 2005.

[OMG11] OMG. Business Process Model and Notation (BPMN), 2011.

[Sil11] Bruce Silver. *BPMN Method and Style*. Cody-Cassidy Press., 2nd. Auflage, 2011.

[VVK08] Jussi Vanhatalo, Hagen Vlzer und Jana Koehler. The Refined Process Structure Tree. In Marlon Dumas, Manfred Reichert und Ming-Chien Shan, Hrsg., *Business Process Management*, Jgg. 5240 of *Lecture Notes in Computer Science*, Seiten 100 – 115. Springer, 2008.

[WRMR11] B. Weber, M. Reichert, J. Mendling und H. A. Reijers. Refactoring large process model repositories. *Computers in Industry*, 62(5), 2011.

[YDG10] Zhiqiang Yan, Remco Dijkman und Paul Grefen. Fast business process similarity search with feature-based similarity estimation. On the Move to Meaningful Internet Systems. In Robert Meersman, Tharam Dillon und Pilar Herrero, Hrsg., *On the Move to Meaningful Internet Systems: OTM 2010*, Jgg. 6426 of *Lecture Notes in Computer Science*, Seiten 60 – 77. Springer, 2010.

[ZZ10] Xulin Zhao und Ying Zou. A Business Process Driven Approach for Generating Software Architecture. In *Quality Software (QSIC), 2010 10th International Conference on*, Seiten 180 – 189. IEEE Press, 2010.

CARES: Requirements Specification with BPMN 2.0 in WTO Procurement

Daniel Hadrian and Florian Evequoz

HES-SO // Valais Wallis
Institute of Business Information Systems
TechnoPôle 3
3960 Sierre
daniel.hadrian@rigertconsulting.ch
florian.evequoz@hevs.ch

Abstract: The WTO procurement procedure for IT infrastructures and services involves the creation of software requirements specifications (RS). This task is challenging for the public administration, as few aids and guidelines are available. But RS is important in the WTO procedure because it serves as a basis for the evaluation of bidders. On the other hand, BPMN is increasingly applied in administrations to document business processes. Taking the situation of the Swiss administration as a reference, we introduce an original methodology to specify requirements from business processes in BPMN 2.0, based on standardized phrase patterns that are attached to BPMN elements. Our methodology also provides a template structure and guidelines for producing RS that can be included in WTO-conform Request for Proposals. Our methodology originates from the situation in Swiss administration, but it is generic enough to be applied in other public administrations as well as in private companies. We also conduct user interviews to identify the actual tasks and challenges faced by users when specifying requirements. Findings from the interviews suggest that communication issues between stakeholders (business users, analysts, IT specialists) are the most pregnant. Based on results and analysis of user interviews, we design a software prototype that aims at solving most problems faced by users when producing requirements specifications. The main advantage of our methodology is that administrations can reuse existing BPMN process documentation as a basis for specifying requirements in the WTO procedure.

1 Introduction

Public procurement of IT infrastructures and services is set by common agreements for countries that are members of the World Trade Organization (WTO). The WTO rules and regulations are meant to promote competition, guarantee the right usage of public funds, and ensure fair and non-discriminatory treatment and equal opportunity between the competing bidders. Above a certain cost threshold in the procurement of IT infrastructures and services, the WTO rules and regulations are applied.

Early in the procurement procedure, the public administration wanting to procure must issue a Request For Proposals (RFP). In Switzerland, the legal basis that governs public procurements [Lo94] gives precise guidelines once the request for proposal has been launched. However, as stated by [Ga07], there is a lack of instructions regarding the form and the content a request for proposal should have. Even HERMES 5, the new version of the official Swiss federal government IT-project methodology (eCH-0054) [Ec13a] does not contain the template of a RFP. Moreover, previous research we conducted by interviewing WTO procurement specialists in Swiss administration tend to confirm this lack of templates, and point out the crucial role of Requirements Specifications within a RFP (as the main basis for the evaluation of the bidders) and the challenges associated with gathering and specifying requirements [Ha11]. Therefore, we believe that a methodology for producing WTO-conform requirement specifications would be a great help for Swiss administrations, but also to other administrations conforming to WTO.

Apart from the procurement theme, public administrations face increasing challenges and opportunities in the E-Government domain, increasing the capabilities but somewhat also modifying the culture of the public sector. In particular, the Swiss Strategy for E-Government recognizes the importance of business processes towards a successful implementation of E-Government [Fi07]. Indeed BPMN 2.0 has been adopted as a national standard for describing business processes (eCH-0140) [Ec13b]. We know that BPMN aims at bridging the gap between business people, business analysts and IT specialists. Those three groups of stakeholders are also key players in the creation of requirements specification. Therefore we think the BPMN knowledge in public administration can be leveraged upon, in order to ease the creation of requirements specifications.

In this article, we introduce a methodology, supported by a software prototype called CARES (Computer Aided Requirements Engineering Software) that eases the creation of requirements specifications based on a BPMN 2.0 documentation of processes. The requirements specification produced can then be used by a public administration as part of a RFP in a public procurement procedure.

In our work, we use the context of the Swiss Government as a reference. In Switzerland, BPMN has been declared a standard for process documentation in public administration [Ec13b], which provides an ideal use case for the methodology we present here. However our findings could be applied to other administrations as well as to the private sector. We also found it most important to primarily target public administrations because of the high regulation in this particular field. The main difference between Switzerland and the EU-member are to be found in the threshold values for public procurement. For EU members, the threshold starts at 134,000 Euros for state administrations and at 207,000 Euros for local and regional authorities and other public sector authorities. In Switzerland the threshold is set at 250,000 Francs (200’000 Euros) for state administrations and is federaly regulated by the cantons.

The article is structured as follows. We first present related work in requirement specifications. Then, we introduce our methodology for requirements specifications

based on BPMN documentation of processes. In section 4, we describe the results of interviews we conducted to enhance our understanding of the tasks and problems users face when specifying requirements, and confronting our methodology to the practice. In section 5, we present the design of a software prototype supporting our methodology and incorporating the main findings from user interviews. We conclude by presenting perspectives and future works envisioned.

2 Related Work

To the knowledge of the authors, existing research to date does not specifically address the use of BPMN in requirements specification. Practitioners in the fiels of requirements engineering that we interviewed as part of preliminary research [Ha11] were also not aware of existing use of BPMN in that context. Specifying requirements from a business point of view is generally part of a Goal-Orientend Requirements Engineering (GORE) approach. Several GORE frameworks for specifying strategic requirements exist. KAOS, i* and B-SCP are probably the most famous. Knowledge Acquisition in autOmated Specification of software systems (KAOS) [La01] enables the computation of requirements from Goal Diagrams. i* (pronounced "i star") or i* framework describes the dependency between actors and their motivations, which is useful for early requirements [Ko06]. B-SCP: Business-Strategy, Context and Processes is based on three different frameworks, namely i*, role activity diagrams and Jackson context diagrams [BK06]. A common feature of those frameworks is that they use their own abstractions for representing the business from which the requirements are derived. They do not rely e.g. on existing process documentation nor do they capitalize an a standard notation as BPMN.. UML Use cases are also used for requirement specifications. In contrast to processes modeled in BPMN, UML provides system use-cases rather than business use-cases.

Best practices for software requirements specification are presented in IEEE-830-1998 [Ie98]. In particular, this document considers three categories of methods, languages and tools for requirements specification ([Ie98] section 4.3.2.3). The process-based approach is one of them. In a process-based approach to requirements specification, requirements are organized „into hierarchies of functions that communicate via data flows“. The approach we present in this article falls in the process-based category.

More specifically, in Switzerland, several guidelines for public procurement also exist [Gu11] [Li12]. Neither provides a template nor guidelines for requirements specifications.

This short survey of the related work shows that no existing approach to date explicitly builds upon a BPMN documentation of business processes to prepare requirements specifications.

3 A BPMN-Based Methodology for Requirements Specification

In this section, we introduce our methodology for requirement specifications based on BPMN documentation of processes. This methodology has three stages. First, process modeling in BPMN, then actual specification of requirements alongside business processes, and finally production of the requirement specification document with the help of a template. We describe the methodology hereafter in a prescriptive fashion.

3.1 Process Modeling

The business processes serve as a basis to specify the actual requirements. Therefore they have to be modeled precisely, using modeling conventions to minimize interpretability bias as well as to maximize coherence. They must stand in the context of the organization in order to show the perimeter of the system that will have to support them. In order to depict the context of the organization and to show the perimeter of the future system, we recommend building up a process architecture (see Figure 1), either top-down or bottom-up by applying the principles explained in the relevant eCH-standards (eCH-0145, eCH-0138, eCH-0139 and eCH-0140, see www.ech.ch).

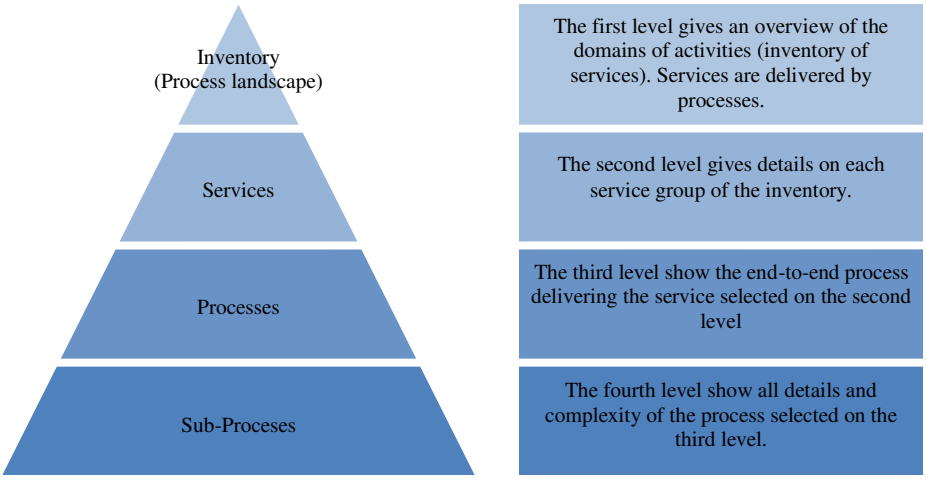


Figure 1 – Levels of process documentation pyramid

The first two levels of the process documentation pyramid in Figure 1 serve only the purpose of putting the processes into context with the whole organization and to show the system perimeter. The third level of the architecture reveals an end-to-end process. The goal of this representation is to show what happens in that process, what are the possible issues and what participants and/or roles are involved. Figure 2 below shows an

example of a process model in BPMN on this level of the pyramid (left) and the detailed business use-case, in form of a BPMN sub-process (right, fourth level).

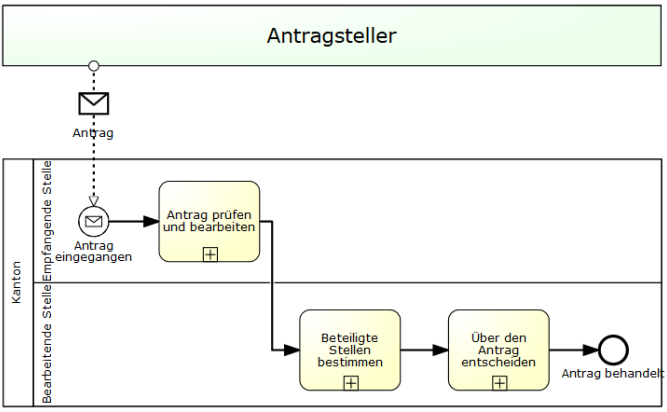


Figure 2 –End-to-end process in BPMN with three business use-cases;

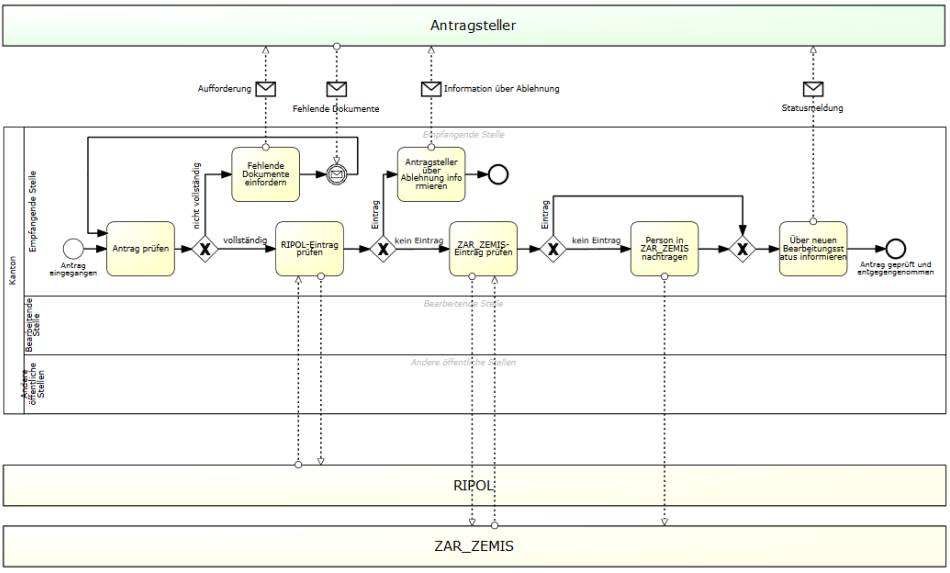


Figure 3 - Detailed use-case "Receive and check enquiry"

3.2 Requirements specification

Requirements are specified on the fourth level of the pyramid, namely the detailed business use-cases. We propose to attach each requirement to an activity, a message flow or a message of the BPMN model at this level (business use-case level). In order to

express the requirement, we use the standardized phrase pattern from [PR11] as shown in Figure 4.

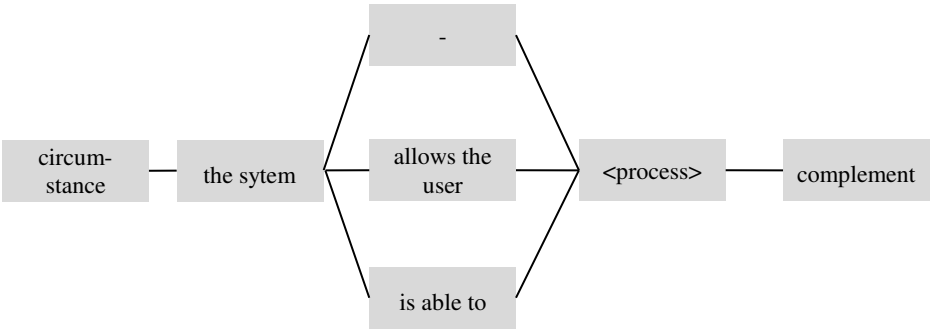


Figure 4 - Standardized phrase pattern

The following table shows a few examples of requirements expressed with the pattern.

Shape name	**Req. ID**	**Specification**
Check inquiry	FRT-001	The system is able to call the business rule relative to the integrity of an inquiry (FRR-001)
	FRT-002	In case the inquiry is incomplete, the system allows the user to send a message to the inquirer containing the list of the missing documents.
Invoice generation	FRT-002	The system is able to check if the inquirer has a record in RIPOL (FRR-002)
	FRR-002	The system is able to call the business rule relative to the RIPOL records

Table 1. Interface specifications of the sub-conversation “communication”

3.3 Requirement Specification Template

Once the requirements have been specified and attached to the business processes in BPMN, they need to be formalized in a requirement specification document. This document will be attached to the RFP of a WTO-conform procurement process. We introduce hereafter a template for this document, inspired by [Ie98] but modified to fit the needs of our original approach. Text in italic font is the chapter structure of the document. We comment chapters directly after they appear, when we propose significant modifications or additions to [Ie98], and provide recommendations for authors.

1. Introduction
    1.1. Purpose
    1.2. Scope
    1.3. Definitions, acronyms, and abbreviations
    1.4. References
    1.5. Overview *

This section shall: describe the content of the other section of the RS; explain how the RS is organized; inform how the business processes have been organized and have been modeled; describe the BPMN shapes used, the meaning of the symbols, the modeling conventions used; describe how the requirements' IDs are built and how to do forward and backward tracing of them.

2. Overall description
    2.1. Product perspective*

In addition to what is mentioned in [Ie98] we intent to call this section "system perimeter" as stated in [Eb10] and [PR11]. The "system perimeter" can be depicted by a BPMN 2.0 **conversation diagram** and some additional artifacts that BPMN 2.0 provides. Using conversation diagrams in this situation has been done prior by [Ri11].

    2.2. Product functions*

"This subsection of the SRS should provide a summary of the major functions that the software will perform. For example, an RS for an accounting program may use this part to address customer account maintenance, customer statement, and invoice preparation without mentioning the vast amount of detail that each of those functions requires." [Ie98]

To achieve this granularity of detail we propose to use a **process landscape or process map** of the business processes the application will have to sustain. In this chapter the process landscape can be detailed to what we will call use-cases later on.

    2.3. User characteristics
    2.4. Constraints
    2.5. Assumptions and dependencies
3. Specific requirements*

This chapter describes precisely each and every requirement the system will have to fulfill. It is based on the chapter 2.1 that describes the system perimeter or product perspective and on the use-case diagrams from chapter 2.2. Each use-case will be detailed within a **single end-to-end BPMN diagram**. The requirements will be written down and split into two main categories, namely functional requirements (FR) and nonfunctional requirements (NR). The authors of the RS are free to make sub-classes for each FR and NR like for instance requirements related to interfaces, performance, compliance, etc.

The authors will take care of avoiding redundancy among requirements. If for instance the chapter 2.1 of the RS already specifies all the interfaces of the system, they shall not be mentioned in this section any more.

The sections in this chapter can be organized the following way:

3.1. End-to-*end process 1 "Name of the Process"*
  3.1.1. Use-*case 1 "Name of the use-case"*
  3.1.2. Use-*case 2 "Name of the use-case"*
3.2. End-to-*end process 2 "Name of the Process"*
  3.2.1. Use-*case 1 "Name of the use-case"*
  3.2.2. Use-*case 2 "Name of the use-case"*
3.3. etc

4 Users' Point of View in Requirements Specification

In order to enhance the methodology presented above and design a tool supporting it, we wanted to complete our understanding of the actual tasks and problems of users when preparing requirement specifications. Therefore we conducted interviews with selected professionals active in Requirements Engineering in the public sector. This type of professionals would be the target user group of our methodology.

4.1 Methodology for User Interviews

We interviewed five people working both in the private and the public sectors across Switzerland. These interviews were done as contextual inquiries, where participants shared information about their work, roles and tasks, being conducted in a semi-structured way by the interviewer [BH97] [Sa13]. All the interviews were recorded. There were five participants: two Requirement and Process engineers from public administration who were involved in WTO procurement, two business analyst from an IT company who acts as a vendor for the public administration, one participant was the CEO of a startup company involved in business consultancy for WTO procurement. Our areas of interest in the interviews were:

- Work and Roles of the user
- Requirement specification and Request for proposal document
- Weighting of the requirements
- Collaboration between different stakeholders or people involved
- Use of BPMN and other modeling languages

Having defined our methodology prior to the interviews, we were explicitly interested in confronting it with real users' workflows and problems when specifying requirements. Therefore, we classified the statements of our participants into three broad categories:

- Problems,

- General important statements
- Process- or work-related terminologies

The different statements gathered from the recording were analyzed in a brainstorming session, with the help of an affinity diagram [BH97]. In particular, we discussed the user problems in details and generated solutions wherever possible. We present below the main problems reported and other findings useful in the context of our methodology, that we extracted from the affinity diagram.

4.2 Main Problems in Requirements Specifications

We summarize here the pain points our participants reported most often in the process of producing a requirements specification. Although the number of participants in our contextual inquiries was low, we believe these problems may inform further research.

Communication issues. Extracting actual requirements from interviews conducted with business people is seen as the biggest challenge. Several participants reported that business specialists have a hard time describing exactly the processes they follow and expressing precisely what they would need (i.e. the requirements). Also, attaching requirements to specific parts in a process can be difficult. Software vendors might also face difficulties when trying to understand requirements stated in RFP.

Requirements gathering, in relation with communication issues above, is perceived as a time-consuming task, accounting for more than 50% of the work towards producing the RS.

Expressing requirements in a standardized form. For one participant, requirements expressed in free text are hard to manage, especially if the list is long. Standardization (e.g. using phrase patterns) would be appreciated. However, another participant pointed out a weakness of standardized form for requirements: business specialists need to be educated to use them correctly.

Tracing requirements coming from different sources is an important challenge noted by several participants. In particular, merging requirements in a later phase should leave the initial requirements accessible. There should therefore be a way to trace back to the original requirements.

Weighting requirements often involves negotiation between stakeholders, many approvals and manual tasks, as it directly influences on the evaluation of the bids later in the procurement process. Therefore, it is a difficult and time-consuming task.

4.3 Other Findings Relevant to our Methodology

The results of the affinity diagram also gave valuable indications on how to enhance our methodology to accommodate the users needs. Among these indications we can mention:

- There are overall requirements that are valid for the whole project before we get to functional and nonfunctional requirements
- All the requirements must be framed as questions or declarations allowing a YES/NO answer by the bidders.
- Requirements have to be expressed directly by the end-users of the IT service or good being procured, before they are refined into WTO-proof requirements. The refinement must be backwards traceable for the end-user so he knows where his initial requirement was incorporated.
- Besides functional and non-functional requirements, interface-related requirements shall have their own category
- Requirements have to be weighted and the user must be able to set the weighting scale according to his needs
- The normalized phrase pattern by [PR11] is suitable for WTO requests for proposal but other phrase patterns might be relevant like agile phrase patterns for SCRUM or Xtreme Programming (e.g. user stories)

5 CARES Software Prototype

Based on our methodology and incorporating the lessons we learned from user interviews, we designed a tool to support the methodology and implemented it as a prototype. This tool was coined "CARES" for 'Computer Aided Requirements Engineering Software'. We wanted this software to be able to reuse existing BPMN 2.0 business processes and attach functional and nonfunctional requirements to process elements. Moreover, it must offer features to administrate, trace, weight and export the requirements in order to incorporate them into a consistent request for proposal. We stated our own initial requirements for CARES that would support our methodology as follows:

- CARES is meant to be a requirements specification tool only (i.e. not a modeling tool)
- CARES has export functions to Microsoft Excel and Word (Requirements Specifications document in Word, and Requirements List in Excel)
- CARES is able to access and re-use existing BPMN 2.0 diagrams from an existing business process management system (BPMS)
- CARES shall be web-based (SaaS)

5.1 Designing CARES

Based on the affinity analysis, we consolidated all problems-solutions pairs into categories and functions, and we derived the information architecture from them. This architecture is depicted hereafter in three different tree structures; each tree showing one of the main functions: project management, requirements management, collaboration.

5.2 Information architecture – Projects

Doing requirements specification for a request for proposal is a long lasting activity and must therefore be handled as a project. CARES must be able to handle different projects for different requests for proposal. Furthermore a project is composed of files (protocols, pictures, etc.), process models, requirements (in the form of BPMN shape attributes). These will be organized through folders (Figure 5)

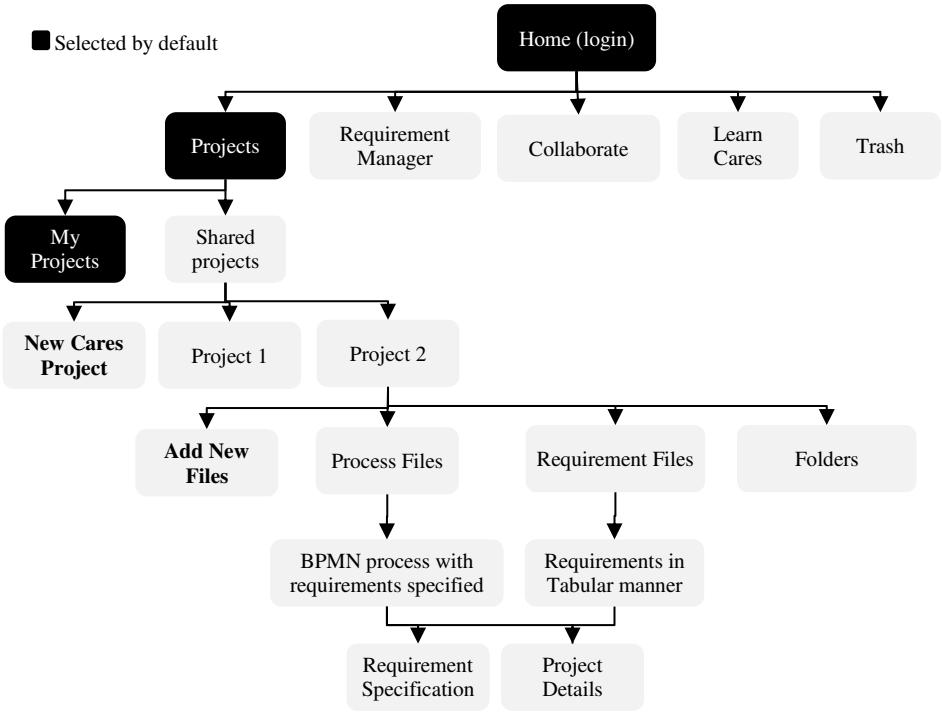


Figure 5 - Information architecture of CARES - projects

5.3 Information architecture – Requirements manager

As stated earlier in the findings for CARES, requirements must be expressed by the user, before they are refined into WTO-proof requirements. The requirements manager function is tailored to respond to this need. Through a tabular view, the requirements engineer can filter, add, delete, search, edit, save, print and export the requirements. Moreover, our interviews have shown that users face difficulty while working with large set of requirements. The requirements manager aims at easing this aspect (Figure 6).

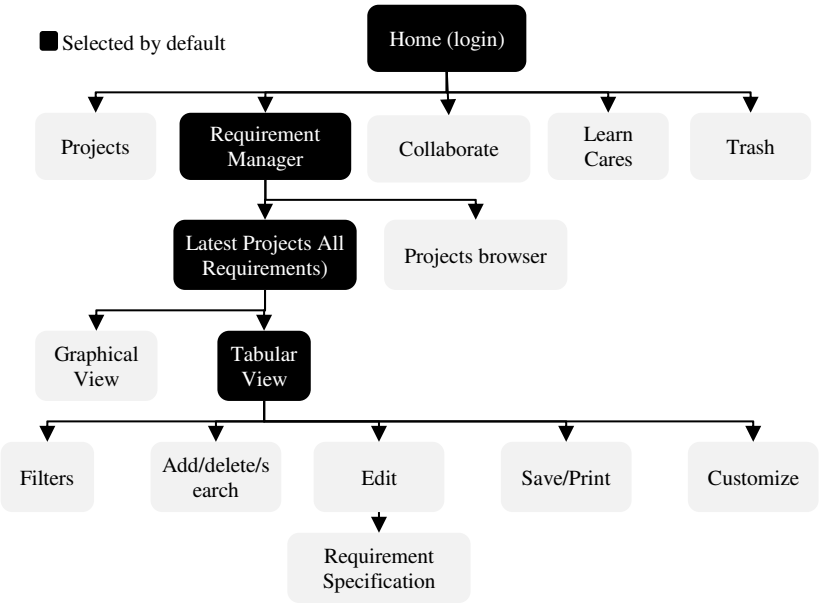


Figure 6 - Information architecture of CARES - Requirements manager

5.4 Information architecture – Collaboration

The collaboration function is targeted at ubiquity. The requirements specification project could be led by the requirements engineer or project manager and collaborators could share their thoughts, approve or comment requirements and participate in live discussions via the web application. Figure 7 depicts the information architecture for the collaboration function.

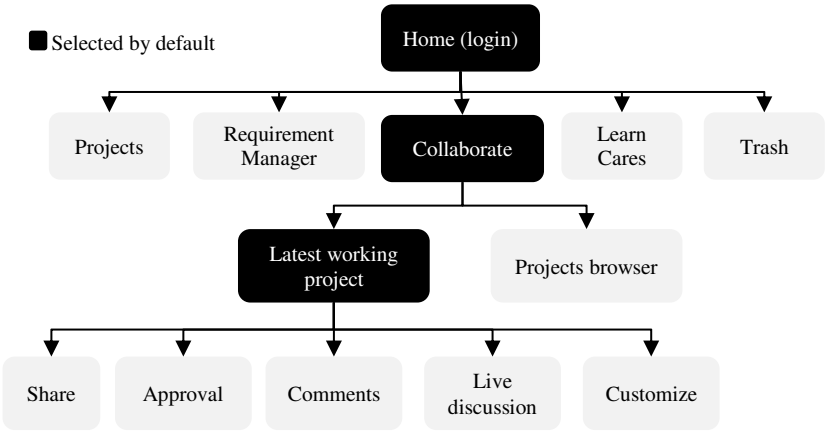


Figure 7 Information architecture CARES – Collaboration

5.5 CARES Prototype

As we already knew, there was no business process related notation specifically tailored for RS. Chances were little that we would find an existing tool that would do what we wanted with few adjustments. Nevertheless we conducted a comparative analysis of the Requirements Engineering tools we were able to find on the market to see if they fitted our requirements. We compared Sparx Enterprise Architect, the ARIS Platform, Visual Paradigm for UML along with Agilian Enterprise Architecture and Visure Requirements. The comparative tool analysis turned out that none of these tools were either simple to use nor did they feature the required flexibility to accommodate our methodology easily. Therefore we chose to implement our own SaaS solution to support our methodology. Our solution builds upon the BPM modeling tool of Signavio (www.signavio.com). It uses its REST and Mashup APIs to access the underlying process repository of Signavio. We added a layer of interface dedicated to the Requirements Specification functionalities that were implemented as described above in Section 5. Screenshots from the prototype are presented in Figure 8.

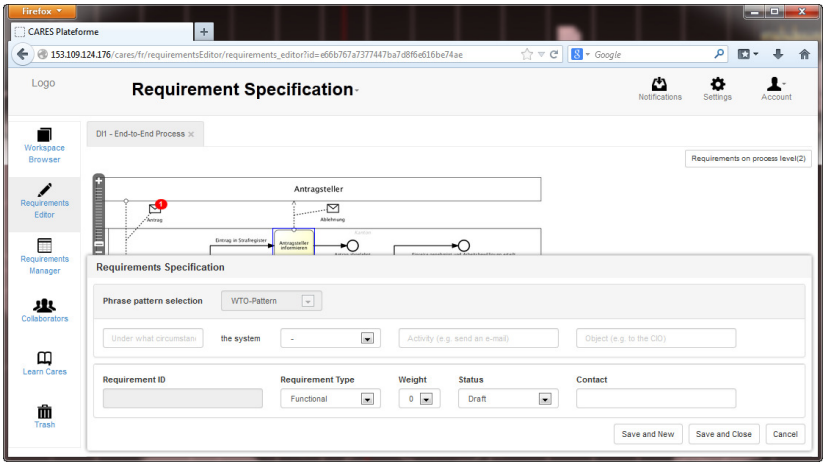


Figure 8 Screenshot of CARES prototype: specifications of requirements on a BPMN diagram with phrase patterns

6 Conclusion and Perspectives

In this article we presented an original methodology to specify requirements from business processes in BPMN 2.0, based on standardized phrase patterns that are attached to BPMN elements. Our methodology also provides a template structure and guidelines for producing requirement specifications that can be included in WTO-conform Request for Proposals. The main advantage of our methodology in the context of WTO

procurement procedure is that administrations can reuse existing BPMN process documentation as a basis for specifying requirements. Moreover, although our methodology originated from the needs and situation of the Swiss public administration, it can be applied for specifying requirements based on BPMN in other countries' public administration as well as in private companies.

We also conducted user interviews to identify the actual tasks and challenges faced by users when specifying requirements. Findings from this interviews suggest communication issues between stakeholders (business users, analysts, IT specialists) are the most pregnant in requirements specifications. Our approach using BPMN and standardized phrase patterns aims at lowering those issues. Based on results and analysis of user interviews, we designed a software prototype that aims at solving the most important problems faced by users.

Future works include evaluation and testing of the prototype with users. Perspectives for future development of CARES include:

- Evaluation of the proposals submit by the vendors
- Extensive collaborative features in CARES
- Graphical/Tree view of the requirements within the requirements manager
- Merge function to merge similar requirements during the refinement phase
- Requirement Specification for agile project management

Acknowledgements

The research was funded by RCSO-ISNet. The authors wish to thank Vikas Luthra, Simon Martin and Alain Duc who were involved in the user research, design and implementation of the prototype, as well as the participants in our user interviews. Sections 4 and 5 of this article are partially taken from an internal unpublished report written by Vikas Luthra in the framework of this project.

References

[BH97] Beyer, H., & Holtzblatt, K.: Contextual design: defining customer-centered systems. Elsevier 1997.

[BK06] Bleistein, Steven, K. C.: B-SCP: a requirements analysis framework for validating strategic alignment of organizational IT based on strategy, context, and process. Sydney, 2006.

[Br89] Brooks, F.: Mythical man-month: essays on software engineering. Dorset House Publishing, New York 1989.

[Eb10] Ebert, C.: Systematisches Requirements Engineering.dpunkt.verlag, Heidelberg, 2010.

[Ec12] Verein eCH: eCH-0073 Vorgabe zur Beschreibung von Leistungen und Prozessen. Version 2.0, www.ech.ch, 2012.
[Ec13a] Verein eCH: eCH-0054 HERMES Projektmanagement Methode, Version 2.0, Entwurf, www.ech.ch, 2013
[Ec13b] Verein eCH: eCH-0140 Vorgaben zur Beschreibung und Darstellung von Prozessen der öffentlichen Verwaltung der Schweiz. Version 1.0, www.ech.ch, 2013
[De91] US Department of Defense: Software Technology Strategy. Washington, 1991
[Fi07] Federal IT Steering Unit (FITSU): Swiss Strategy on E-Government, Bern, 2007
[Ga07] Galli, P. et al.: Praxis des öffentlichen Beschaffungsrechst (2nd edition Ausg.). Zürich, Basel, Genf: Schulthess, 2007
[Gu11] Conférence Romande des Marchés Publics: Guide Romand pour les marchés publics, Abgerufen am 02. 01 2012 von http://www.vd.ch/fileadmin/user_upload/organisation/dinf/sg-dinf/guide_romand/AA_Pr%C3%A9face_Guide_romand.pdf, 2011
[Ha11] Hadrian D.: Requirements Engineering à l'aide de BPMN 2.0 dans le cadre d'appels d'offres selon l'OMC. Master thesis. HES-SO Master, 2011.
[Ie98] IEEE: IEEE Recommended Practice for Software Requirements Specifications, The Institute of Electrical and Electronics Engineers, Inc., New-York, 1998
[Ko06] Koliadis George, A. V.: Combining i* and BPMN for Business Process Model Lifecycle Managmeent. University of Wollongong (UOW), School of Information Technonolgy and Computer Science (SITACS), 2522Gwynneville NSW, 2006.
[La01] Lamsweerde, v.: Goal-oriented requirements engineering: A guided tour. International Joint Conference on Requirements Engineering (S. 249-263). IEEE Conference, Toronto, 2001.
[Li12] Cantonal Aministration, Bern, Switzerland: LIBIT. Guide sur l'adjudiciton des marchés publics, Abgerufen am 25. 01 2012 von http://www.fin.be.ch/fin/fr/index/informatik/informatik/rechtliche_grundlagen/libit_2.html., Bern, 2011
[Lo94] Swiss Confederation: LMP, Loi sur les marchés publics, LMI, Loi sur les marchés intérieurs, AMP, Accord sur les marchés publics. Berne, 1994.
[PR11] Pohl, K., & Rupp, C.: Basiswissen Requirements Engineering (3. korrigierte Auflage Ausg.), dpunkt.verlag GmbH, Heidelberg, 2011.
[Ri11] Ritter, Daniel, et al.: Building a Business Graph System and Network Integration Model Based on BPMN, in Remco Dijkmann, Jörg Hofstetter et Koehler Jana. Business Process Model and Notation, Third International Workshop on BPMN, Proceedings. Luzern, S. 154-159.Springer, 2011.
[Sa13] Saini, A., Nanchen, B., & Evequoz, F.: Putting the Customer Back in the Center of SOA with Service Design and User-Centered Design. In Service-Oriented and Cloud Computing (pp. 94-103). Springer Berlin Heidelberg, 2013.
[We10] Wesfall, L.: Software Requirements Engineering: What, Why, Who, When, and How. University of Texas, 2010.

Terravis – Large Scale Business Process Integration between Public and Private Partners

Walter Berli[1], Daniel Lübke[2], Werner Möckli[1]

[1]SIX Terravis AG
Brandschenkestrasse 47, CH-8002 Zürich
walter.berli@six-group.com
werner.moeckli@six-group.com

[2]innoQ Schweiz GmbH
Gewerbestrasse 11, CH-6330 Cham
daniel.luebke@innoq.com

Abstract: Supply Chain Management is a very popular and already mature topic between companies in which business processes are coordinated across company borders. However, there are also business processes that span not only companies but also public services. One example is the mortgage business in which banks and insurances are involved as private partners but interact with land registries and notaries which are public partners. Terravis is a platform that integrates the business processes between these parties and offers a high business value by being capable of doing end-to-end processing without media conversions. Such an effort is nevertheless complicated due to the number of interacting parties, different mindsets and different goals. Within this paper we discuss challenges integration projects face and describe experiences about how these challenges can be addressed.

1 Introduction

One long-term trend in the manufacturing industries is Supply Chain Management: Supply Chains are optimized between different companies. These networks of companies have become larger and more complex over time. They can be seen as the successor architectures for Large Scale Business Process Integration that build networks along business processes that span many organizations (both private and public) and in which typically no organization can dominate the other (in contrast to supply chains in which usually the OEM dominates.)

One example for the need of such an integration is the Swiss mortgage business spanning banks, notaries and land registries, from which Terravis[1] emerged. Terravis is a platform run by SIX Terravis Ltd.[2] that offers the following services:

[1] Terravis is a registered product of SIX Terravis Ltd.
[2] SIX Terravis Ltd. is owned by SIX Group Ltd. and is supervised by the Swiss Federal Office of Justice

- Portal for land register and geographical data
- Electronic transactions with focus on the business between public administration (land register, commercial registers and notaries) and economy (mortgage institutes [banks, insurance companies, pension funds], geometers, institutions with high dependencies on land register data (e.g. telecom companies, Swiss Confederation, Swiss Railway SBB and others)
- Fiduciary management of register mortgages for banks

Within this paper, we want to explain the challenges, possible solutions and our experiences in Terravis as an example for a large-scale business process integration project. This paper is structured as follows: The next section explains the BPM Life-Cycle. In section 3, the Business Process Evaluation with the business motivation for Terravis is explained, followed by the Business Process Analysis with related experiences are presented in section 4. In section 5, the same is done for Business Process Design & Modeling. Section 6 will present the resulting solution architecture. In section 7 the main challenges and the counter-measures are summarized before a conclusion and outlook is given.

2 BPM Life-Cycle and Related Work

One way to optimize business processes is following the Business Process Management (BPM) approach. BPM is a cyclic approach that tries to continuously improve business processes. There are many different variations of this cycle, but we will use the one by Weske [We12] as shown in figure 1 within the context of this paper.

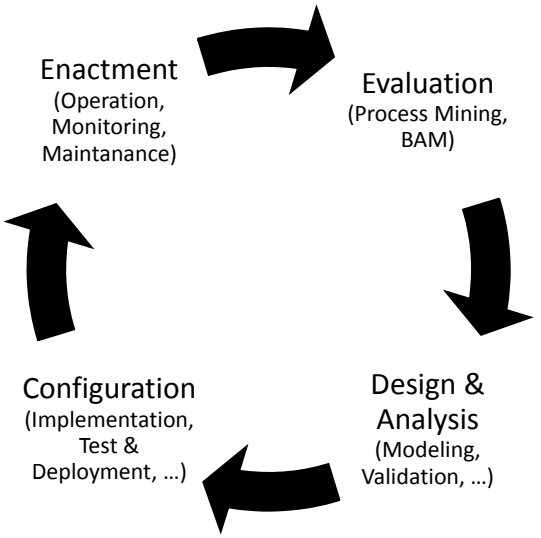


Figure 1: Business Process Lifecycle according to [We12]

The general approach is to find the potential for optimization in the currently executed business processes. Once found, attempt to improve them, implement the new business processes and analyze them further.

The following sections are structured according to this pattern: Section 3 starts with a short explanation of the evaluation of the old paper processes followed by the analysis of those processes. In section 4, the design is described and in section 5, we will present the configuration with some pieces that are related to enactment.

Terravis itself cannot endure the BPM Life-Cycle alone, because Terravis depends on the Business Process Evaluation of partners. Partners gain experience with the newly implemented business processes and forward their new requests to Terravis, who then change the business processes accordingly.

3 Business Process Evaluation

3.1 Mortgage Business before Terravis

Switzerland is organized as a federation[3]: The 26 cantons are responsible for the land registries and notaries within their area. With the introduction of the civil code [CC] in

[3] Swiss political system, explained in the citizen portal by the Swiss Confederation (https://www.ch.ch/de/politisches-system-schweiz/ - visited on 15 May 2014)

1912, a national legal foundation for the land registries and commercial registers were introduced and the associated competencies were transferred to the cantons.

3.1.1 Notary Systems in Switzerland

When the civil code was introduced, there was no need for harmonization of national laws for notaries. The consequences were:

- Each of the 26 cantons introduced a legal base for the notaries individually,
- In Switzerland three general types of organizations were established:
    - **Cantons with independent civil law notaries:** The notary is exclusively entitled to seal and stamp contracts of transfer of land ownership and to convey to the land register,
    - **Cantons with office notaries (Amtsnotariat):** The land register official is also notary with similar competences as a civil law notary
    - **Cantons with a mix of independent civil law notaries and office notaries:** Both systems can be present in one canton.

There is an exclusivity of the notaries in every canton for stamping land register contracts in the canton that appointed them. The consequence of that delimitation is a heterogeneous practice leading to different contracts, procedures, and forms.

Until 2012, there was no legal base in Switzerland that allowed

- electronic registration of notarial contracts
- electronic execution of notarial contracts

As a consequence, there was no industry IT solution or any standard other than Mac- or Windows-PCs using standard text processing (e.g. Microsoft Word), internet connection and accounting software.

3.1.2 Land Registry System in Switzerland

The civil code defines the constitutive law of the land register: parcel, ownership, legal rights and the dependency on the official survey. The mortgage notes are one of the rights to be registered with and only issued by the land register. The national regulation defined to a certain extent how a land register is organized, under which conditions a register entry can be accepted or has to be refused. The land register was kept in large heavy books where the inscriptions were done in proper structured form.

In Switzerland, the cantons began to digitize the land register data in the 1990s. Instead of setting up a nationwide project, the Swiss Confederation allowed any applying canton to realize its land registry solution as it saw fit, which resulted in five different software

solutions over the past 20 years. In 2007, the Swiss regulation defined a standard data model and interfaces for land registries for the first time.

3.1.3 Mortgage borrowing in Switzerland

Mortgage lending is one of the main pillars of the Swiss banking industry. Over decades, home owners and the economy have profited from low interest rates and steadily rising home prices when applying for hypothecary credits secured by mortgage notes. Even though the market volume exceeded 700 billion Swiss Francs by 2011[4] [SNB], the communication and business processes were never successfully standardized, nor digitized, throughout Swiss-wide between banks, notaries, and land registries. Because of the complexity of the business, the large number of involved partners, lack of standards and non-existing electronic interfaces, no one envisioned that a solution such as Terravis could be achieved.

4 Business Process Analysis

4.1 Project eGRIS

In 2001, the Swiss Confederation initiated a project called eGRIS with the aim of a nationwide portal for land register data for its citizens. As a result, in 2007, the Swiss government enacted the Technical Regulation for Land Registers in Switzerland [TGBV], which for the first time in Switzerland, defined the following aspects: a standard data model for land registers (eGRSIDM, eGRIS Data Model), a unique identifier for land (E-GRID, electronic Parcel ID), a unique identifier for right in the land register (EREID, electronic Rights ID), unique identifiers for persons registered in the land register (UID, AHVN13), a standard interface for geometers to land register (AVGBS) and a standard interface for exporting land register information (GBS).

In 2007, in response to budget cuts, the project management suggested to open the scope and invited the Swiss banks to join the eGRIS project. The aim of the project eGRIS was adjusted to the requirements of the banks (represented by the Commission for Financial Standardization[5]) and other authorized participants of the economy. This included the portal with integrated interfaces and the standardized, media-break-free electronic transactions between banks, land registry offices and notaries. In addition, the project management of the Swiss Confederation intended to setup a centralized data pool for land register data. The cantons refused the data pool and withdrew from eGRIS.

As a result, the Federal Department of Justice, SIX[6] (representing the Swiss banks) and the cantons agreed in 2009 that SIX was to take the lead of the project eGRIS, control the project finances and implement it together with the cantons. The new eGRIS steering

[4] Schweizerische Nationalbank – Statistische Publikationen (http://www.snb.ch/de/iabout/stat/statpub/bchpub/stats/bankench - visited on 15 May 2014)
[5] Swiss Commission for Financial Standardisation (www.sksf.ch - visited on 15 May 2014), a board of the Swiss Banks Association (www.swissbanking.ch - visited on 15 May 2014)
[6] SIX Group Ltd. is the main infrastructure for the Swiss financial industry

committee representing SIX, the banks and the cantons agreed on the following approach:

- Deliverables, time-line and milestones: the portal and electronic transactions to be realized by SIX in agreement with the cantons and the banks by end of 2014.
- SIX sets up an operating company.
- Every stakeholder (SIX, banks and cantons) carries their own project costs.
- A central data pool and the long term safeguarding of land register data are not part of the deliverables.

The Confederation withdrew from the project lead, but remained in the project organization without voting rights.

4.2 Stakeholders' Motivations

There were a large number of stakeholders involved in developing Terravis. Every group was involved in different stages of the development and therefore, the motivations were dynamic.

4.2.1 Banks

Banks are constantly looking for opportunities to increase the efficiency of their business operations. On one hand, there is a limited potential to improve the internal processes and one the other hand, the operations of the credit divisions is the less optimized in all Swiss banks.

The banks are looking to improve the integration between business processes of other stakeholders. With Terravis, there was the chance to increase efficiency, once all intended processes were implemented, by up to 50 percent.

4.2.2 Land Registry Offices/Cantons

The organizational integration of the land registries varies from canton to canton. The responsibilities are legally defined on the Confederation level in the land register regulations [GBV]. On a political level, the responsibilities are defined by the cantonal government (except cantons Lucerne and Zurich where the land register is incorporated in the Cantonal Court) and on the functional level, they are defined by the inspector.

Due to the fact that the cantons did not feel to be involved by the eGRIS project management headed by the Swiss Government between 2001 and 2009, they appreciated that SIX respected their requirements:

- Autonomy of the cantons regarding the land registries
- The data sovereignty still remains with the cantons
- No centralized data pool
- Decentralized solution regarding portal and electronic transactions

- Involving the cantons as equal partners in the project steering committee of eGRIS.

4.2.3 Notaries

Involving the notaries in project eGRIS was a demanding task. The notaries of the cantons, along with office notaries, were represented by their land registries and the respective notary inspector. The cantonal notary associations of 13 cantons are organized in the Swiss Notary Association (SNV)[7]. When SIX launched the eGRIS project together with the cantons in 2009, the SNV pursued its own project. The aim of that project was to develop an electronic notary register, a digital archive for notaries and electronic conveyance to the land register offices, based on a platform that the Federal Office of Justice promised to develop.
The motivation of the notaries was heterogeneous. In general, they wanted to ensure legal certainty and privacy, especially for notaries' customers. As stakeholders, the notaries could imagine to digitize the communication with the clients, the cantonal authorities (tribunal, land register offices and commercial register offices) and institutional partners (banks etc.), but they want to keep status quo, i.e. no change to the way the notary business was being done.

4.2.4 Raising Stakeholder Awareness

There were several key success factors in the eGRIS project. The project management convinced the opinion leaders by identifying the benefits and by disclosing the costs. Furthermore, it was important to raise the awareness of the legal departments of the cantons to check if the legal base exists and if not, to immediately establish a legal project.

4.2.4 Trust-building measures

There was no legal obligation of the land registers and the notaries to introduce Terravis; therefore, SIX could only succeed by taking several trust-building measures. The difficulties that the project management faced in the planning was that they didn't know most of the opinion leaders and all the dependencies; therefore, an interactive and intuitive proceeding was chosen.
The success of the project can be accredited to the communication of our own values from the very beginning, involving the stakeholders in the definition of the electronic processes and a fair requirement management. SIX led the project with transparency and honesty to all involved at all times.

4.3 Elicitation Techniques

Since the Swiss Confederation has no legal power to enforce a solution on all the cantons (autonomy of the cantons), the business process definitions and requirements needed to be elicited, verified, and agreed upon by a large number of stakeholders.

[7] Swiss Notary Association represents the notaries of 13 cantons (www.schweizernotare.ch - visited on 15 May 2014)

The majority of the representatives in the project eGRIS are experts in the land register business and legal questions, but only a few have an IT background. This added the challenge of discussing the migration from paper-based processes to electronic ones at the same time.

A number of the canton representatives and notaries are retained in private companies such as SIX Terravis that are outside of the public administration and are offering services related to land registries and notaries.

While the measures described in section 3.2.4 address political problems, process elicitation was a huge problem, too. Every party only knew their "local" activities in the process, i.e. a notary knew how to draft a contract, but was unaware of the activities and dependencies before and after him. Establishing a global view of the business process was thus a challenging task like it is in other projects [Lü06]. In order to elicit the paper-based process and define the new electronic process, several elicitation techniques were used:

- **Stakeholder Workshops:** SIX organized several individual workshops with representatives of each stakeholder (banks, land register, notary). The aim and agenda of each workshop was proposed by the project management and adjusted if required by the participants. The definition of a process was discussed on drawings of business processes (very high level view, also called Level 0 or 1) to get an understanding of the discussed process and the business rules that apply to it.

- **Prototypes:** The results of the stakeholders' feedback were taken over into the next workshops. After a first common understanding was reached, SIX developed a prototype to visualize the results and to clarify additional details. The prototype was quickly developed as a BPEL [Jo07] process and used the BPMS' task facilities consisting of BPEL4People [Ag07] and the product's task manager. The user interfaces were not pretty, but contained all data to be changed or displayed. With this information, the standard cases of a business process could be conducted step by step and every party could give concrete feedback about the process flow and the necessary data.

5 Business Process Design & Modeling

The results of the workshops and the prototypes needed to be documented. While these two steps are interrelated in practice, the problems and solutions are distinct. The results related to modeling and documentation are discussed in this section.

Different stakeholders needed to be served by the process documentation. On the one hand, the people actually conducting or supervising the business needed to accept the new process definitions. On the other hand, implementers and IT departments needed to change their systems in order to interact with Terravis.

Right from the beginning, the decision was made to split the business aspects from the technical aspects into separate documents.

5.1 Process Documentation for Business Stakeholders

The main problem with the documentation for business stakeholders was understandability. The processes with many variants and differences due to differences in the notary systems are complex. In addition, the exchanged data structures are complex, too.

Most of the stakeholders from the public sector were not used for formal business process notations like BPMN [Ag11] or EPCs [NR02]. As a consequence, business process diagrams in those notations failed as a means of communication during workshops. High-level diagrams, which essentially are choreography diagrams but without a standard syntax, are used of the standard notations. While being very similar in expression, they could be understood easily by the non-technical stakeholders. All of these overview diagrams are followed by a free-text description of the process steps and business rules. Data structures are represented as lists of forms with placeholders.

5.2 Process Documentation for Technical Stakeholders

The technical documentation consists of service documentation and process documentation. Only the latter is relevant to the main topic of this paper.
Initially, the technical process documentation consisted of UML sequence diagrams depicting exactly one out of many scenarios of a process. The messages between the objects corresponded to SOAP messages that had to be exchanged between Terravis and a partner system. While this is helpful to get an introduction in to the process, complex processes were illustrated in dozens of diagrams that still could not cover all possible process variants. Especially when discussing processes with system implementers and IT departments, this view was missing. As a result, all processes were documented as BPMN diagrams that carry lot of technical details while still being abstract with regard to Terravis' internal implementation details like internal services.

6 Implementation & Execution

After having a coarse-grained specification, implementation work could be started. Because there are many non-functional properties in a large-scale business process integration system, a robust architecture is needed. Within this section we will explain the main architecture drivers of a large-scale business process integration platform like Terravis before explaining the high-level architecture.

6.1 Architecture Drivers

The main architectural drivers are derived from the environment consisting of many partner systems plus some additional business demands that are of cross-cutting concern:

- **Interoperability:** Banks, Notaries, and Land Registries deploy a variety of software solutions and operating systems. Thus, a central business process platform must be accessible in a platform-independent way.

- **Scalability to thousands of partner systems:** Terravis did not start with a big bang. Instead canton after canton and bank after bank was and is going to be connected. Nevertheless, the architecture needs to be prepared to deal with 1000+ partner systems in the final environment.

- **Fault Tolerance:** Besides performance and network aspects, the huge amount of partner systems also inflicts imminent failures. If all surrounding systems are available 99% of the time, the central business process platform can only expect all 1000 surrounding systems to be available only at 0.0043% of the time (0.99^{1000}) which is less than 23 minutes over a year. Failure tolerance and being able to resume processes after systems come back online are crucial.

- **Multi-Version Support:** Surrounding systems change and get updated. However, the business process platform also changes and delivers new features and will offer new service interfaces for added functionality. Because deployment windows are different between all partners, the central process platform needs to support a set of old service versions that allows partners to upgrade within a defined time-window and not in parallel with the central platform.

- **Short Time-to-Market:** Building a huge platform takes time, especially if requirements are not clear and evolve quickly. It is therefore a good idea to design an architecture that can be built iteratively and still address conceptually all necessary architecture drivers. Effort spend in architecture and implementation needs to be balanced and interleaved. Also a quick delivery of features allows the central platform to gain trust with its partners instead of letting them wait for a long period of time.

- **Process Monitoring/Data Integrity:** In eGovernment scenarios, the central platform needs to be able to demonstrate when and who has placed an offer/contract and when requests were sent to partners.

- **Confidentiality:** While most of the land register data is available publicly, business processes contain personal data and additional business-relevant information. Providing confidentiality for this information is crucial also with regard to building trust with partners.

- **Legally Binding Digital Signatures:** Very often when a business process platform is used to conduct business with administration offices, the question of legally binding signatures arise. In Switzerland, the law for digital signatures [ZertES] defines the requirements for digital signatures that are equivalent to hand-written ones.

6.2 Resulting Architecture

The Terravis Architecture is shown in figure 2. At the core, is a Business Process Management System (BPMS). It is responsible to coordinate the business processes and offers time-stamping of performed activities and resuming of faulting process instances due to service invocation errors. Also having this component that allows deployment of processes modeled in BPEL or BPMN, it provides a high abstraction of business process implementations and thus higher productivity and the ability to implement features faster.

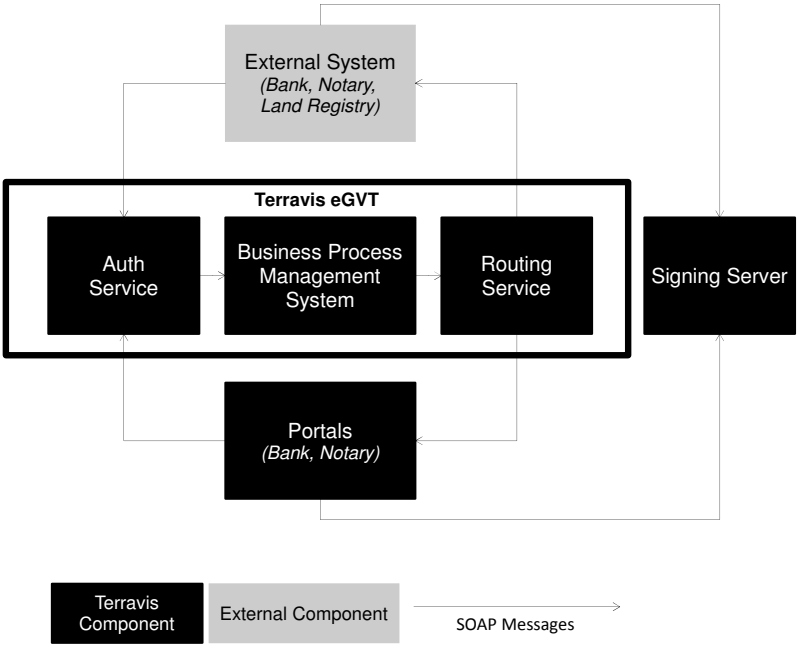


Figure 2: Terravis Architecture

SOAP/HTTPS was chosen as the integration technology. SOAP is platform-independent by using XML serialization and offers a good means of describing the service contract by using WSDL. In addition, most platforms can generate wrapper classes for the XML Schema. For confidentiality reasons, all SOAP connections from and to Terravis are secured by Two-Way-SSL. Both systems are authenticated by certificates and the communication channel is encrypted.

The BPMS does not manage all aspects of the SOAP messaging itself. Instead, it is surrounded by two Enterprise Service Bus (ESB) components: an incoming authorization service and an outgoing routing service. The incoming authorization service checks permissions of a partner to start a process instance or participate in an

already running process instance. If permission is granted, then the authorization service will route the incoming message to the correct process implementation. The routing service takes outgoing SOAP messages from process instances and checks their destinations. It then delivers the messages to the intended partner while using the correct proxies, SSL certificates etc.

Both ESB components can also convert messages between different service versions. All services are versioned and the version is indicated in their namespaces. By matching the version of the handled message with the supported version(s) of the target, the correct transformation is chosen. This functionality has been added after the second service version went productive and is one example of an already-planned, but later implemented, architecture decision of an important architecture driver.

The ESB components and the BPMS are supported with internal services for ID Generation, Document Generation etc. Together they build the eGVT (elektronischer Geschäftsverkehr Terravis), i.e. the core back-end for the business processes.

As the first institution to offer central digital signatures, Terravis offers a signature server that can be used for signing PDFs that are exchanged between banks, notaries, and land registries. The offered solution is completely ZertES-compliant, but will not be described in further detail as this is out of scope of this paper.

In the beginning, the envisioned architecture integrated directly with banking systems and notary systems. However, only land registries were able to change their systems and implement the necessary services quickly. To offer an intermediate solution, Terravis developed a portal that can be used by the banks and notaries to conduct their business as long as their systems are not ready. The portals use the same service interfaces like the external systems; there is no difference between the portal and a “real” external system from the process point-of-view. While this solution is good and allows the quick roll-out of the whole solution, it creates a new problem: At a certain point in time, banks will need to switch from the portal to their systems. This will make advanced routing necessary that is based on the process start date, the process type and the business partner.

Terravis has considerably grown in size (number of processes and process complexity) in a very short time-frame. A sample set of metrics that can be used to judge size and complexity is published [Lü13].

6.3 Development Process and Tooling

Besides the product/software itself, it is important to be able to roll out changes quickly after implementation. Besides basics like a central build server with a nightly build, automated deployments are one piece of the necessary infrastructure; however, no tool is good if used incorrectly. We have positive experiences with iterative development with short iterations and production releases. Terravis tries to deploy a new feature version every month.

7 Predominant Problems & Solutions

The project management of SIX had a complicated start due to the unusual constellation with SIX from the private sector setting up a new service in a public sector domain. As a consequence, a very high priority was to build up trust between SIX and the partners as well as between the partners themselves. Doing workshops and trying to moderate between parties and making sure to reach a common understanding was one part. Delivering quickly and without major hassles was the other part. Both parts, although one technical and one political and therefore not closely related, needed to go together to succeed.

This also included Change Management on all levels: Evangelizing end-to-end processing without papers and media transformations is a huge cultural shift from a 100% paper-driven business. SIX needed to educate and demonstrate feasibility of the envisioned solution. But by also being part of a huge organization with very established services, Terravis was a start-up in a huge organization that created friction and needed persuasion and introduction of new concepts, processes and tools.

Because Switzerland is divided in cantons and the business processes and conventions are very different across them, there are few if any people who understand the business processes end-to-end and can specify them. The result are many change request due to imperfect specifications and incorrect assumptions. The development process, the architecture and the development team needed to be able to deal with this. One solution is to support only the 80-90% cases, i.e. a certain process not with all variations but with the important and most frequently used ones. By doing this approach the costs-to-benefit ratio is much better and development can proceed much quicker because the remaining variants are usually very hard to specify and very complex. If such a variant arose, it could still be handled on paper. Also business mistakes are handled out of the process. This reduces complexity and allows quicker development and concentration on business benefits. Experiences with corner cases and frequent errors can be used to improve process models later.

A huge technical problem is quality assurance (QA). Efficient QA is necessary for doing quick deployments for fixes or frequent releases for new features. However, deployment speed is constrained by testing efforts that increase exponential with added features. Business processes are heavily tested with unit tests [Lü07] by using BPEL Unit [ML06]. However, whole system tests, including the portals, is still a manual and time-consuming task.

The data model of land registry data is also very complex because it needs to support history-versioning and the storage of historic, unstructured, or semi-structured data that have been valid under the different cantons legislature. Storing and parsing this data, as defined by the GBDBS Schema[8], is complicated and error-prone.

[8] Grundbuchschnittstelle, (http://www.ejpd.admin.ch/content/dam/data/wirtschaft/grundbuch___egris/tgbv_anhang3.pdf - visited on 15 May 2014)

As already discussed the sheer amount of partner systems creates a set of problems on its own: multi-versioning of services, routing of messages and fault tolerance are only the major ones. Consistently following the 'separation of concerns' paradigm is the only way to limit impact, isolate logic and reducing complexity in such a system.

8 Conclusions & Outlook

8.1 Conclusion

Terravis is one of the first platforms that offer large-scale integration of many partners across different domains. We expect those scenarios to increase, especially when both public and private parties participate in processes.

While such projects are technically challenging, they also have many non-technical success factors. Gaining trust and building concrete requirements engineering are two of them. On the technical side, a solid architecture and the development process with a supporting tool-chain to roll out new features quickly are crucial.

8.2. Outlook

Terravis has grown from its initial scope and now supports also integration into the Swiss Register of Commerce and the trustee functions of mortgages at SIX. Considering that current users are very pleased with Terravis, users and stakeholders with similar businesses (e.g. involving notaries or land registries) also want to benefit from the offered integration.

To keep the current development pace, new options for reducing testing efforts, while increasing test coverage, are continuously explored and implemented.

We expect Terravis to grow larger and other platforms for different domains to emerge. These platforms, support the exchange of structured information in contrast to emails and unstructured PDFs, which boost efficiency and allow all involved partners to integrate their processes tightly without media interruptions.

8.2. Relevance of other countries with Switzerland

Even though each country has developed its own legal framework for land ownership and conveying, Terravis is the sole provider of a straight through solution for the mortgage business and conveying.

Terravis is developed as a product that can be adapted to the needs of any jurisdiction by changing the parameters. Seeing that Switzerland provides one of the most complex structures, the platform can handle any business case as long as the process is standardized and structured. An implementation of the platform in all member states of the EU enables a single market for mortgages in Europe.

References

[Ag07] Agrawal, A.; Amend, M.; Das, M.; Ford, M.; Keller, C.; Kloppmann, M.; König, D.; Leymann, F.; Müller, R.; Pfau, G.; Plösser, K.; Rangaswamy, R.; Rickayzen, A.; Rowley, M.; Schmidt, P.; Trickovic, I.; Yiu, A. & Zeller, M.: WS-BPEL Extension for People (BPEL4People), Version 1.0, 2007.

[Ag11] Aggarwal, A.; Amend, M.; Astier, S.; Barros, A.; Bartel, R.; Benitez, M.; Bock, C.; Brown, G.; Brunt, J.; Bulles, J.; Chapman, M.; Cummins, F.; Day, R.; Elaasar, M.; Frankel, D.; Gagné, D.; Hall, J.; Hille-Doering, R.; Ings, D.; Irassar, P.; Kieselbach, O.; Kloppmann, M.; Koehler, J.; Kraft, F. M.; van Lessen, T.; Leymann, F.; Lonjon, A.; Malhotra, S.; Menge, F.; Jeff Mischkinsky, D. M.; Moffat, A.; Mueller, R.; Nijssen, S.; Ploesser, K.; Rivett, P.; Rowley, M.; Bernd Ruecker, T. R.; Samoojh, S.; Shapiro, R.; Saxena, V.; Schanel, S.; Scheithauer, A.; Silver, B.; Srinivasan, M.; Toulme, A.; Trickovic, I.; Voelzer, H.; Weber, F.; Westerinen, A. & White, S. A.: Business Process Model and Notation (BPMN), Version 2.0 Object Management Group (OMG), 2011.

[CC] Schweizerisches Zivilgesetzbuch – SR 210 – Version of 1 July 2013

[GBV] Grundbuchverordnung – SR 211.432.1 – Version of 1 January 2012

[Jo07] Jordan, D.; Evdemon, J.; Alves, A.; Arkin, A.; Askary, S.; Barreto, C.; Bloch, B.; Curbera, F.; Ford, M.; Goland, Y.; Guízar, A.; Kartha, N.; Liu, C. K.; Khalaf, R.; König, D.; Marin, M.; Mehta, V.; Thatte, S.; van der Rijn, D.; Yendluri, P. & Yiu: A. Web Services Business Process Execution Language Version 2.0, 2007.

[Lü06] Lübke, D.: Transformation of Use Cases to EPC Models Proceedings of the EPK 2006 Workshop, Vienna, Austria, 2006.

[Lü07] Lübke, D.: Unit Testing BPEL Compositions in Baresi, L. & Nitto, E. D. (Eds.) Test and Analysis of Service-Oriented Systems Springer, 2007.

[Lü13] Lübke, D.: BPEL Metrics in Large Scale Process Integration – A Time-Line, Technical Report available at http://www.daniel-luebke.de/files/20130601-BPEL_Metrics_in_Large_Scale_Process_Integration.pdf, 2013, last visited 2014-05-08

[ML06] Mayer, P. & Lübke, D.: Towards a BPEL Unit Testing Framework TAV-WEB '06: Proceedings of the 2006 Workshop on Testing, Analysis, and Verification of Web Services and Applications, Portland, USA, ACM Press, 2006, 33-42.

[NR02] Nüttgens, M. & Rump, F. J.: Syntax und Semantik Ereignisgesteuerter Prozessketten (EPK) Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen – Promise 2002, Gesellschaft für Informatik, 2002, 64-77.

[We12] Weske, M.: Business Process Management - Concepts, Languages, Architectures, Springer Verlag Berlin, 2012.

[TGBV] Technische Verordnung des EJPD und des VBS über das Grundbuch– SR 211.432.11 – Version of 1 February 2013

[ZertES] Bundesgesetz über Zertifizierungsdienste im Bereich der elektronischen Signatur – SR 943.03 – Version of 1 August 2008

Geschäftsprozessmanagement in der Wirtschaftsförderung. Status Quo zu Notwendigkeiten, Anwendungen und Hemmnissen.

André Göbel[+] Emanuel Hesse[*]

Hochschule Harz
Fachbereich Verwaltungswissenschaften
Domplatz 16
38820 Halberstadt
[+]agoebel@hs-harz.de [*]ehesse@hs-harz.de

Abstract: Der Beitrag stellt auf Basis einer ersten empirischen Untersuchung den Status Quo des Geschäftsprozessmanagements (GPM) in der kommunalen Wirtschaftsförderung dar. Obwohl GPM in der öffentlichen Verwaltung ein Instrument von wachsender Bedeutung darstellt, wurde der Bereich kommunaler Selbstverwaltung hierzu bislang wissenschaftlich nicht evaluiert. Die vorliegende qualitative Voruntersuchung bietet einen ersten Erkenntnisgewinn zur Wirtschaftsförderung.

1 Einführung

Die kommunale Wirtschaftsförderung ist Gegenstand verschiedener wissenschaftlicher Disziplinen. Geographen und Wirtschaftsgeographen im Besonderen untersuchen vor allem örtliche sowie wirtschaftsstrukturelle Zusammenhänge, wogegen die interdisziplinär wirkenden Verwaltungswissenschaften sich den aufbauorganisatorischen und normativen Gesichtspunkten widmen. Juristische und makroökonomische Aspekte bereichern ebenfalls die multiperspektivische Forschung zur Wirtschaftsförderung. Bei so großer Interdisziplinarität ist es fast schon verwunderlich, dass das Geschäftsprozessmanagement (GPM), also der genauen Auseinandersetzung mit den ablauforganisatorischen Bedingungen innerhalb der Wirtschaftsförderung, bislang in allen Disziplinen nicht besetzt wurde. Die Ursachen liegen vermutlich im aufbauorganisatorischen Wesen der kommunalen Wirtschaftsförderung: Innerhalb der kommunalen Selbstverwaltung sind wirtschaftsförderliche Tätigkeiten vollkommen freiwillig und werden daher standortabhängig unterschiedlich sowie meist in kleinstorganisatorischer Form durch Einzelpersonen erbracht (vgl. Abschnitt 2). Daraus folgend waren die ablauforganisatorischen Forschungsanreize bislang begrenzt, wodurch auch keine Analyse des GPMs in der kommunalen Wirtschaftsförderung stattfand.

Aus diesem Grund wurde die nachfolgend beschriebene Voruntersuchung durchgeführt. Die Ergebnisse dieser Vorstudie sind Ausgangslage für eine bundesweite Untersuchung zum Status Quo des GPMs in der kommunalen Wirtschaftsförderung, die in Kürze fol-

gen wird. Ziel der hier dargestellten Erstuntersuchung war die Erfassung des Bewusstseins für GPM in der Wirtschaftsförderung. Der gewählte qualitative Forschungsaufbau dieser Vorstudie erlaubt eine erste thematische Näherung mit folgenden Forschungsfragen:

- Welches Begriffsverständnis zu GPM ist erkennbar?
- Werden konkrete Ansätze für GPM betreiben?
- Wie werden praktische Erfahrungen zur GPM-Anwendung reflektiert?
- Können Hemmnisse und Perspektiven zu GPM benannt werden?

Der vorliegende Beitrag stellt die thematische Genese sowie die ersten Forschungsergebnisse dar. Nach einer theoretischen Einführung zur kommunalen Wirtschaftsförderung werden die Erkenntnisse zum Status Quo des Geschäftsprozessmanagements (GPM) in der kommunalen Wirtschaftsförderung sowie im Weiteren auch deren Reflexion aus Sicht der befragten Experten behandelt. Der Abschluss bietet einen Ausblick auf die weitere Bearbeitung und den Umgang mit Forschungsdesideraten.

2 Theoretische Grundlagen und Problemstellung

2.1 Entwicklung von Geschäftsprozessmanagement

In der Betriebswirtschaft avanciert Prozessmanagement zu einem wichtigen Instrument, um den stetig steigenden Herausforderungen gerecht zu werden. Veränderte Kundenerwartungen, zunehmende Komplexität der Aktivitäten und letztendlich die Erfolgsziele des Unternehmens verlangen ein kontinuierliches Geschäftsprozessmanagement. Erste Überlegungen zu einer prozessorientierten Denkweise lassen sich aus der Organisationslehre ableiten. Nordsieck betonte bereits in den dreißiger Jahren die Notwendigkeit einer ablauforientierten Organisationsgestaltung [No32]. Trotz der frühen theoretischen Grundlagen zur Prozessorientierung sah die unternehmerische Praxis bis zu Beginn der achtziger Jahre anders aus. Unternehmen verfolgten weiterhin den klassischen Ansatz der Organisationsgestaltung, bei dem die Ablauforganisation der Aufbauorganisation untergeordnet war [Al05, S. 79], wie es auch heute noch in der öffentlichen Verwaltung zu beobachten ist. Erst durch die Arbeit von Gaitanides rückten in der Betriebswirtschaft die Prozesse wieder in den Vordergrund [Ga83]. Einen Schritt weiter ging der Ansatz des Business Process Reengineering in den neunziger Jahren von den Autoren Hammer und Champy [HC94]. Sie forderten ein generelles Umdenken und eine radikale Neugestaltung sämtlicher Geschäftsprozesse. Ab dem Jahre 2000 stand zunehmend der Kunde im Fokus der Geschäftsprozesse. „Über Geschäftsprozesse wird das Denken und Handeln der gesamten Organisation auf Kunden ausgerichtet. Je effizienter die Geschäftsprozesse die Kundenforderungen und –erwartungen erfüllen, umso zufriedener sind die Kunden und umso erfolgreicher ist das Unternehmen“ [SS13, S. 59]. Inzwischen umfasst das Thema Geschäftsprozessmanagement verschiedene Disziplinen aus Themenbereichen wie beispielsweise Qualitätsmanagement (QM), Changemanagement (CM), Customer Relationship Management (CRM), Wissensmanagement (WM), Projektma-

nagement (PM) und viele weitere Managementinstrumente mehr. Die verschiedenen Blickwinkel der jeweiligen Fachrichtung erschweren eine einheitliche Definition.

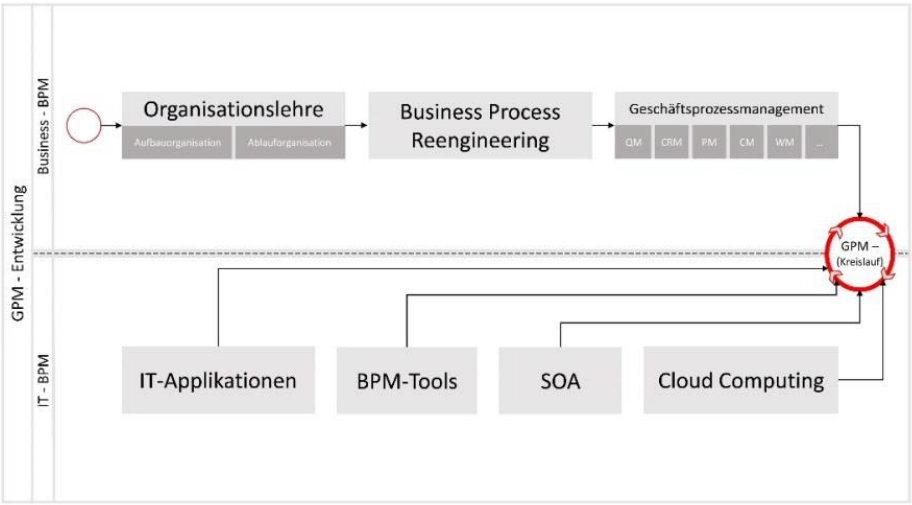


Abbildung 1: Entwicklung von Geschäftsprozessmanagement (eigene Darstellung 2014)

Eine weitere Möglichkeit, den Begriff Geschäftsprozessmanagement genauer abzugrenzen, ist die Unterscheidung nach Scheer zwischen betriebswirtschaftlichen BPM (Business-BPM) und technologischem BPM (IT-BPM) [SKJ06, S. 6]. Insbesondere die IT hat in den letzten Jahren die Entwicklung des Geschäftsprozessmanagement geprägt. Die Anwendungsvielfalt reicht von IT-Applikationen (ERP-Systeme, EAI-Plattformen) und BPM-Tools bis hin zu serviceorientierten Architekturen (SOA) oder dem Einsatz von Cloud Computing. Die verschiedenen Technologien verfolgen im weitesten Sinne das Ziel, die Effizienz zu steigern. Grundlage dafür ist das Vorhandensein von Geschäftsprozessen. Unter Berücksichtigung des IT-Einsatzes kann folgende Definition herangezogen werden: Business Process Management umfasst „alle Aktivitäten, um die modellbasierten automatisierten Geschäftsprozesse (samt manuellen Aktivitäten) eines Unternehmens (und unternehmensübergreifend) stets optimal ablaufen lassen zu können*"* [WS09, S. 1]. Für eine vorerst abschließende Betrachtung muss auf den Begriff des Managements eingegangen werden. Dieser umfasst „Führung, Organisation und Controlling zur zielorientierten Steuerung von Geschäftsprozessen" [SS13, S. 6]. Aus einer kontinuierlichen Steuerung bzw. dem Managen der Prozesse ergibt sich der Geschäftsprozessmanagement-Kreislauf nach Allweyer [Al05, S. 91].

In der öffentlichen Verwaltung, aber auch in den Verwaltungswissenschaften selbst, stellt Geschäftsprozessmanagement ein junges Untersuchungsfeld dar. Verschiedene Herausforderungen wie z. B. der demografische Wandel, die angespannte öffentliche Finanzsituation oder eine fehlende Kundenorientierung und weitere Aspekte des gesellschaftlichen Wandels haben die Entwicklung hin zu einer prozessorientierten Verwaltung geprägt. Insgesamt ist die Thematik den Bemühungen einer Verwaltungsmodernisierung zuzuordnen [BAF09, S. 9ff.]. Es ist dabei nicht unüblich, dass sich die

öffentliche Verwaltung betriebswirtschaftlicher Ansätze bedient und versucht, diese adaptiv zu übertragen [Lü12, S. 3]. Gerade in diesem Zusammenhang trifft die öffentliche Verwaltung auf eine besondere Herausforderung und es bedarf einer gesonderten Betrachtung. Zum besseren Verständnis und zur Positionierung von Geschäftsprozessen in der öffentlichen Verwaltung wird hier das Geschäftsprozess-Referenzmodell nach Walser herangezogen [Wa14, S. 85–87]. Aus diesem Modell lassen sich die verschiedenen Aktions- und Spannungsfelder ableiten, die sowohl für die Entwicklung in der öffentlichen Verwaltung als auch im spezifischen Umfeld der kommunalen Wirtschaftsförderung relevant sind. Zum einen agieren sowohl in der öffentlichen Verwaltung als auch mit dieser unterschiedliche Akteure mit unterschiedlichen Aufgaben, wodurch sich sehr differenzierte Prozessverläufe ergeben. Zum anderen sind strategische Elemente zwar vorhanden, müssen aber aufgrund der unterschiedlichen Interessen separat betrachtet werden. Diese besonderen Gegebenheiten und Abhängigkeiten erschweren die Einführung und Entwicklung von Geschäftsprozessmanagement in der öffentlichen Verwaltung. Ebenfalls relevant ist, wie in der Privatwirtschaft auch, die Organisationsform. Die starke Ausrichtung an der Aufbauorganisation erschwert eine Prozessorientierung [BAF09, S. 30–31].

Als wesentliche Treiber zu einer verstärkt prozessorientierten Organisation in der öffentlichen Verwaltung, sind die Bemühungen der Kommunalen Gemeinschaftsstelle für Verwaltungsmanagement zum Aufbau eines Prozessregisters [KGSt14] als auch das Vorhaben zur Nationalen Prozessbibliothek zu nennen [NPB14]. Ferner untersuchte aktuell eine gemeinsame Studie von IMTB und PwC den Status Quo von Prozessmanagement in den öffentlichen Verwaltungen [NG14]. Zentrales Ergebnis ist, dass rund 80 % der Teilnehmer Prozessmanagement einsetzen, allerdings zum Großteil nur in ausgewählten Organisationseinheiten. Der Entwicklungsstand wird eher durchschnittlich bewertet. Diese empirische Erkenntnis sowie die Erkenntnisse aus aktuellen Literaturbeiträgen zum Geschäftsprozessmanagement in der kommunalen Verwaltung, waren Motivation für eine nähere Untersuchung zum Geschäftsprozessmanagement in der kommunalen Wirtschaftsförderung.

2.2 Ziel und Struktur der Wirtschaftsförderung

Kommunale Wirtschaftsförderung wird als ein Instrument der kommunalen Wirtschaftspolitik eingesetzt [Ku08a, S. 19] und dient somit der strukturpolitischen Einflussnahme unterhalb der staatlichen Ebene [HKL10, S. 262]. Der Betrieb erfolgt in der Regel direkt durch die zuständige Gebietskörperschaft [DR12, S. 28] im Rahmen der kommunalen Selbstverwaltung (Art. 28, Abs. 2 GG; [FHP13, S. 5]) und dient dem Bereich der allgemeinen Daseinsvorsorge [RR03, S. 9]. Das Ziel der Wirtschaftsförderung ist die Stärkung des eigenen Wirtschaftsstandorts durch die gezielte Beeinflussung verschiedener Standortfaktoren zur Steigerung privatwirtschaftlicher Prosperität in der Kommune.

Die Organisationsform der kommunalen Wirtschaftsförderung ist aufgrund der Freiwilligkeit der Aufgabenerledigung in den deutschen Kommunen höchst unterschiedlich realisiert, sodass sich keine einheitlichen Schlussfolgerungen auf das innerbetriebliche Geschäftsprozessmanagement in den Wirtschaftsförderungen schließen lassen. Für die

Aufgabenerledigung können Städte und Gemeinden „die Organisationsform eines Amtes oder eines Fachbereiches, eines Eigenbetriebes, oder einer eigenen Gesellschaft wählen. Die Wirtschaftsförderung kann allein oder in interkommunaler Kooperation oder unter Einbindung Dritter erbracht werden“ [FHP13, S. 5f.]. Bei den letztgenannten Organisationsmodellen erfahren die Geschäftsprozesse eine weitere Komplexitätssteigerung, weil die Arbeit der Wirtschaftsförderung hierbei meist unter Mitwirkung örtlicher Unternehmen geführt wird, bei der die Kommune jedoch eine Mehrheitsbeteiligung hält [Ha04, S. 65]. Die unterschiedlichen Strukturen belegen, dass es keine zwingend vorteilhafte Rechtsform für den Wirtschaftsförderungsbetrieb gibt [Gö12, S. 18]. Betreibermodelle mit einer Ausgliederung der Aufgaben verfolgen den Leitgedanken, hierdurch eine grundsätzliche Flexibilisierung der kommunalen Aufgabenerfüllung zu erreichen [Ei01, S. 71–76]. In allen Analysen wird stets die notwendige Optimierung der Zusammenarbeit zwischen den Akteuren am Standort zum Gelingen von Wirtschaftsförderungsaktivitäten betont. Dabei ist zu beobachten, dass die Wirtschaftsförderung überwiegend als eigene Dienststellen der Gebietskörperschaft (Dezernat, Amt, Abteilung, Fachbereich) betrieben werden und seltener in Form von Stabsstellen bei den (Ober-)BürgermeisterInnen ([HKL10, S. 262]; [FHP13, S. 6]). Somit scheint sich nach aktueller Anforderung die Prozessverankerung der Aufgabenerledigung innerhalb der örtlichen Behörden als sinnvoller Gestaltungsrahmen etabliert zu haben[1].

Zudem hat die kommunale Wirtschaftsförderung seit Mitte der 1990er Jahre einen starken Wandel durch ein heterogenes Aufgabenwachstum, veränderte Kundenerwartungen sowie durch weitere veränderte innere und äußere Rahmenbedingungen erfahren ([HKL10, S. 262]; [Ac12, S. 279f.]). Die Aufgabenbewältigung in den Wirtschaftsförderungen ist zu einer hochkomplexen Querschnittsaufgabe innerhalb der kommunalen Selbstverwaltung avanciert, deren Erledigung „die Einbindung einer Vielzahl von lokalen und überlokalen Akteuren erfordert (Verwaltung und Rat; Stadtplanung, Umwelt, Baugenehmigung, Verkehrsplanung, Liegenschaften, Kämmerei, Denkmalpflege, Gewerbeaufsicht; Arbeitsverwaltung; Industrie- und Handelskammer, Handwerkskammer, Gewerkschaften, Europäische Union (EU) u. a.)“ [HKL10, S. 262]. Darüber hinaus kennzeichnet sich der Wandel der kommunalen Wirtschaftsförderung durch steigenden Einsatz von Informationstechnologien (IT) zur Aufgabenbewältigung.

Durch diese Strukturen wird deutlich, dass die Beschäftigen der kommunalen Wirtschaftsförderung in politische, verwaltungsbürokratische und technische Prozesskomplexe eingebunden sind, deren Orchestrierung jedem Beschäftigten im Sinne seiner Aufgabenerfüllung gelingen muss. Aufgrund der oben dargestellten personellen Situation wird deutlich, dass allein schon fehlende Dokumentationen der wichtigsten Ablaufprozesse in der Wirtschaftsförderungsorganisation zu spontanen Leistungsausfällen gegenüber den Unternehmen führen können, wenn der direkte Ansprechpartner der Wirtschaftsförderung zum Beispiel bei Krankheit vertreten werden muss. Aus der Summe der Einflussfaktoren schlussfolgernd, wurde die These gefasst, dass professionelles Geschäftsprozessmanagement in der kommunalen Wirtschaftsförderung eine gleichermaßen wachsende Bedeutung erfahren haben muss.

[1] Vgl. zur umfassenden Auseinandersetzung der Wirtschaftsförderungsorganisation auch Bl03 und Gö12.

3 Forschungsmethodik

Für eine erste Annäherung an den Themenkomplex Geschäftsprozessmanagement in der Wirtschaftsförderung wurde ein qualitativer Forschungsaufbau für die hier zugrunde liegende Vorstudie gewählt. Im Rahmen von leitfadengestützten, teils narrativ geprägten Experteninterviews mit Wirtschaftsförderern wurden folgende Bereiche evaluiert: a) allgemeines und begriffliches Verständnis von GPM, b) aktuelle Ansätze für GPM in der Wirtschaftsförderung, c) praktische Erfahrungen zur GPM-Anwendung und Reflexion dieser sowie d) Hemmnisse und Perspektiven zu GPM in der Wirtschaftsförderung.

Aufgrund des qualitativen Designs wurde eine geschichtete Zufallsstichprobe mit disproportionaler Auswahl nach Schumann [Sc12, S. 94–96] von 18 Wirtschaftsförderungen gewählt, um so ein möglichst differenziertes Meinungsbild einfangen zu können. Die Repräsentativität steht bei dieser Vorgehensweise nicht im Vordergrund, auch eine Gewichtung der Aussagen wird nicht verfolgt. Die Interviews wurden im Zeitraum vom 17. März bis 4. April 2014 nach vorheriger Terminvereinbarung mit den Experten telefonisch geführt und aufgezeichnet. Anschließend erfolgte die Transkription durch „bewusst einfache und schnell erlernbare Transkriptionsregeln, die die Sprache deutlich ‚glätten' und den Fokus auf den Inhalt des Redebeitrages setzen" ([DP13, S. 20ff], in Anlehnung an [Ku08b, S. 27]).

lfdn	Interviewpartner	Standortdaten	Organisationsdaten	
	Position	**Einwohner**	**Mitarbeiter**	**Form**
T01	Agenturleitung	25.000-50.000	1 bis 2	Teil eines Amtes
T02	Geschäftsführer	über 100.000	3 bis 5	sonstiges
T03	Leiter	unter 25.000	1 bis 2	Amt und privatrechtlich
T04	Leiter	25.000-50.000	1 bis 2	Stabsstelle
T05	Geschäftsführer	50.000-100.000	3 bis 5	Amt und privatrechtlich
T06	Leiter	über 100.000	mehr als 5	Amt
T07	Mitarbeiter	unter 25.000	1 bis 2	Stabsstelle
T08	Geschäftsführer	über 100.000	mehr als 5	Amt und privatrechtlich
T09	Fachdienstleiter	über 100.000	1 bis 2	Amt
T10	Leiterin	50.000-100.000	1 bis 2	Stabsstelle
T11	Mitarbeiterin	25.000-50.000	1 bis 2	Teil eines Amtes
T12	Mitarbeiterin	21.531	1 bis 2	Stabsstelle
T13	Fachdienstleiter	über 100.000	3 bis 5	Stabsstelle
T14	Leiter	25.000-50.000	3 bis 5	Stabsstelle
T15	Geschäftsführer	über 100.000	mehr als 5	Amt und privatrechtlich
T16	HA-Leiter	über 100.000	mehr als 5	Teil eines Amtes
T17	Leiter	50.000-100.000	1 bis 2	Stabsstelle
T18	Mitarbeiter	über 100.000	mehr als 5	Amt und privatrechtlich

Abbildung 2: Übersicht der Interviewpartner. Die Zufallsstichprobe erreichte folgende Bundesländer: Baden-Württemberg, Bayern, Brandenburg, Bremen, Hessen, Niedersachsen, Nordrhein-Westfalen und Sachsen-Anhalt; Median Mitarbeiterzahl: 3 (Eigene Erhebung 2014)

Zu Beginn der Interviews wurden die Experten über die Art und Weise der offenen Befragung aufgeklärt und Ihre Zustimmung zum Mitschnitt für eine anschließende anonymisierte Auswertung eingeholt. Aus der Befragung ergaben sich die nachfolgend erläuterten Erkenntnisbereiche, darunter das Begriffsverständnis zum Geschäftsprozessmanagement sowie Aussagen über die Notwendigkeit, aktuelle An-

wendungen, Hemmnisse und Erfahrungen, ergänzt um Reflexionen und die Einschätzung der künftigen GPM-Entwicklung in der kommunalen Wirtschaftsförderung.

4 Ergebnisse der Vorstudie

4.1 Verständnis

Die Gesamtauswertung der Interviews zeigt eine sehr diffuse Begriffsverortung sowie eine schwammige Ein- und Abgrenzung des Themenfeldes durch die Experten. Die Bandbreite der verschiedenen Aussagen wird nachfolgend in vier Typen klassifiziert: Null-Typus ohne GPM-Verständnis; Typus mit GPM-Grundverständnis; fortgeschrittener Typus und professioneller Typus. Zu den jeweiligen Verständnistypisierungen werden nachfolgend eine Erläuterung und ein Ankerbeispiel aus den Interviews aufgeführt.

Ein vereinzelt aufgetretener Typus beschreibt Wirtschaftsförderer ohne thematisches Verständnis zum Geschäftsprozessmanagementbegriff „Null-Typus“: *„…ich habe nicht* die geringste Ahnung, was Sie sich d*arunter vorstellen.“* (T08, 01:46)

Ein weiterer Typus zeigt bereits ein „Grundverständnis“ des GPM-Begriffs. Hinweise hierauf bilden Reflexionen der Ablaufgestaltung und deren Modellierungsmöglichkeiten sowie Hinweise auf Standardisierungsaspekte und eine Objektbezogenheit des eigenen „Geschäftsfelds“. Dieser Typus erkennt, dass ein Ereignis einen Prozess auslöst, der durch eine bestimmte und ggf. standardisierte Vorgehensweise zu einem Ergebnis führt (Input > Output). Des Weiteren zeigt dieser Typus kein oder nur ein begrenztes Verständnis der komplexen Zusammenhänge des Geschäftsprozessmanagements: *„Ge*schäftsprozessmanagement heißt für mich, wie sich der Ablauf gestaltet, um einen typi*schen Vorgang, in dem jeweiligen Geschäftsfeld umsetzen zu können.“ (T0*2, 00:48)

Darüber hinaus konnte ein weiterer Typus mit einem „fortgeschrittenen“ Begriffsverständnis identifiziert werden. Dieser Typus zeigt ein klares Verständnis der Komplexität des Geschäftsprozessmanagements, benennt Steuerungsaspekte und differenziert in Innen- und Auswirkung verschiedener Prozesse, inklusive der abstrakten Erkenntnis von Rollen und Akteuren: *„Geschäftsprozessmanagement ist die Steuerung aller Abläufe und Aktivitäten innerhalb unserer Wirtschaftsförderung […] [und] umfasst die interne*n Abläufe oder Prozesse der Wirtschaftsförderung sowie auch die Prozesse, die nach a*ußen gehen.“* (T03, 01:03) Ferner auch als weiterer Anker: *„*Prozesse betreffen den Kontakt zum Bürger, der eben auch Kunde im örtlichen Einzelhandel, im Handwerk, im Dienstl*eistungsbereich ist.“ (T01, 02:35)*

Ein darauf aufbauender Typus weist ein „professionelles“ Begriffsverständnis über Geschäftsprozessmanagement auf. Dieser erweitert den vorhergehenden „fortgeschrittenen“ Typus um Reflexionen zu tangierenden Aspekten des Controllings von Geschäftsprozessen oder vereinzelt auch zum Einflussbereich des Wissensmanagements: *„Es geht [im* GPM] darum, dass man klar definiert, wie Geschäftsprozesse vonstattengehen und dafür sorgt, dass die beteiligten Mitarbeiter und Kollegen Informationen darüber haben und

dass man Kontrollmechanismen hat, um Abweichungen zu erkennen." (T18, 00:45) Lediglich eine Einzelnennung reflektiert Geschäftsprozessmanagement auch als zyklische Vorgehensweise, sodass im Typus des „professionellen" Begriffsverständnisses subsummiert wird: *„Geschäftsprozessmanagement bedeutet Planung, Steuerung und Kontrolle der Abläufe innerhalb der Wirtschaftsförderung." (T06, 02:02)*

Aus den Erkenntnissen der Interviewdaten sowie der hier vorgeschlagenen Typisierung ist zu schließen, dass noch kein einheitliches Begriffsverständnis zum Geschäftsprozessmanagement vorherrscht. Auch wird Geschäftsprozessmanagement nur selten als integrierter Handlungsgegenstand der Wirtschaftsförderung angesehen.

4.2 Notwendigkeit

Vor dem Hintergrund der eingangs skizzierten Sonderrolle (Aufgaben, Organisationsform und -größe) der Wirtschaftsförderung stellt sich die Frage, ob Geschäftsprozessmanagement von den Wirtschaftsförderern überhaupt für notwendig angesehen wird. Die Antworten in den Experteninterviews zeigen zwar ein sehr differenziertes Bild, lassen sich jedoch in drei Kategorien untergliedern. Zunächst ist eine Gruppe der Aktiven erkennbar, die den Einsatz von Geschäftsprozessmanagement durchweg für notwendig betrachtet und Prozessmanagement im unterschiedlichen Umfang betreibt. Als auslösende Treiber in der Wirtschaftsförderung werden verschiedene Gründe genannt.

Zum einen wird die Notwendigkeit darin erkannt, der besonderen Verortung der Wirtschaftsförderung gerecht zu werden, das heißt Strukturen zu schaffen, die die aufbauorganisatorische und ablauforganisatorische Komplexität aufgrund der verschiedenen Anforderungen und Interessen reduzieren: *„Geschäftsprozessmanagement ist in der* Wirtschaftsförderung sinnvoll, weil wir eine Schnittstellenfunktion zwischen Interessenten (Unternehmer), dem Umfeld (Bürger) und Auftraggebern aus der Politik und der Verwaltung sind. Also unterschiedlichste Interessensgruppen. Und wenn wir da nicht gewisse Strukturen drin haben, dann kann man in diesem Sumpf relativ schnell versin*ken." (T03, 05:31)* Zum anderen ist die Wirtschaftsförderung nicht nur ein Akteur in einem dynamischen Umfeld, sondern befindet sich selbst in einem dynamischen Wandel durch komplexer werdende äußere und innere Rahmenbedingungen der Wirtschaftsförderungstätigkeit. Wie vielfach durch die Experten angemerkt wurde, ist auch aus diesem Grund eine prozessorientierte Vorgehensweise in Verbindung mit ergänzenden Konzepten wie dem Changemanagement[2] anzuwenden: „Wir sind gerade komplett im Umbruch und das wird sicherlich noch mal einen besonderen Einfluss haben auf neue Prozesse, aber auch auf das veränderte Rollenverständnis von Mitarbeitern." (T15, 05:05)

Neben der Gruppe, die aktiv Prozessmanagement betreibt, gibt es auch Wirtschaftsförderungen die aufgrund verschiedener und teils mangelhafter Rahmenbedingungen zwar die Notwendigkeit für Prozessmanagement bestätigen, jedoch dieses nicht aktiv umsetzen können, wie folgendes Ankerbeispiel treffend verdeutlicht. *„Das wäre sicherlich ein* Wunsch, das [GPM] zu tun. Da unsere Wirtschaftsförderung, aber viel zu klein aufge-

[2] Changemanagement im Kontext des Geschäftsprozessmanagements: Changemanagement „unterstützt Führung und Mitarbeiter, die Notwendigkeit von Veränderungen zu erkennen und zu akzeptieren sowie Veränderungen umzusetzen und zu stabilisieren" [SS13, S. 532].

stellt ist, haben wir gar keine Möglichkeit darüber nachzudenken, das wirklich in ordentlichen Geschäftsprozessen abzubilden, und dann auch noch zu analysieren und zu verändern. Sondern hier ist es einfach Tagesgeschäft. Man muss hier reagieren. Agieren *und viel analysieren ist kaum möglich."* (T11, 02:21)

Die interviewten Wirtschaftsförderer äußerten mehrfach das Interesse, sich der Thematik Prozessmanagement zu nähern. Oft wird aber die Möglichkeit sich dem Themenbereich ausführlich zu widmen, durch fehlende Kapazitäten (vgl. Abschnitt 3.4 „Hemmnisse") verhindert. Neben den Befürwortern gibt es noch als dritte Gruppe die Passiven, bei denen Geschäftsprozessmanagement keine Rolle spielt: *„Wir haben es einfach noch nicht gebraucht." (T07, 03:06)*

4.3 Anwendung

Die im Abschnitt theoretische Grundlagen vorgestellte Einteilung in geschäftliches und technisches BPM zeigt sich auch in der praktischen Umsetzung innerhalb der Wirtschaftsförderung. Es konnten Einsatzbereiche mit einem betriebswirtschaftlichen Bezug zum Geschäftsprozessmanagement identifiziert werden, aber auch Projekte, die IT-getrieben sind. Gleich mehrfach wurde als Praxisbeispiel die Einführung eines CRM-Systems mit der Prozesserfassung in Verbindung gebracht: *„*Wir benötigen GPM für den Bereich Customer Relationship Management, ein CRM-System, wo wir Daten zu Unternehmen erfassen und wo wir Daten zu Vorgängen mit Unternehmen erfassen. Also welche Briefe und welche E-Mails wir versendet haben, welche Informationen wir gesammelt haben, in welchen Projekten bestimmte Teilnehmer beteiligt waren, aber auch wen wir zu welcher Veranstaltung *einladen." (T05, 02:52)*

Hierbei sind die Wirtschaftsförderungen gezwungen, ihre Abläufe zu dokumentieren, damit das System auf Basis von Prozessmodellen konfiguriert werden kann. Die genannten betriebswirtschaftlichen Anwendungsbeispiele stehen häufig im Zusammenhang mit den unterschiedlichen Managementdisziplinen. Besonders auffällig sind dabei Themen aus dem Bereich des Qualitäts- und Wissensmanagements, wobei Geschäftsprozessmanagement als Unterstützung fungiert. Insbesondere soll durch die Prozesserfassung ein besserer Überblick entstehen, hier verbunden mit dem Ziel eines attraktiveren Angebots: *„Um einfach den Überblick zu behalten. Wenn Sie anfangen immer mehr Dinge zu m*achen, immer mehr Aktivitäten. Um die Wirtschaftsförderung davon zu lösen, nur derjenige zu sein, der über Fördergelder berichtet oder Fördergelder beantragt und bereitstellt. Als moderne Wirtschaftsförderung muss ich wirklich mehr Aktivitäten und Mehrwerte schaffen. (T03, 06:39)

Auf die Frage, warum Geschäftsprozessmanagement in der Wirtschaftsförderung durchgeführt wird, ist das Handeln der Experten überwiegend als eine spontane Reaktion auf ein konkretes Problem oder auf Forderungen von außen zu bewerten: *„Wir sind einfach* der Frage nachgegangen, wenn immer mal wieder Beschwerde darüber geführt wird, dass die Kollegen nicht gut zu erreichen sind, woran das liegt*." (T16, 04:43)*

Die Interviews zeigen, dass spezielle BPM-Methoden in der Regel nicht angewandt werden. Die Prozessauswahl und –erfassung erfolgt zufällig und wird nur teilweise dokumentiert. Der Einsatz spezieller BPM-Tools spielt beim Großteil der Wirtschaftsförde-

rungen überhaupt keine Rolle. Lediglich eine Wirtschaftsförderung setzt Visio als ein Basisinstrument für Prozessdokumentation ein und eine andere denkt über die Anschaffung einer BPM-Software nach. Selbst eine strukturierte Vorgehensweise lassen die wenigsten Experten erkennen. Die aktuelle Umsetzung verdeutlicht folgendes Ankerbeispiel: *„*Wir haben keine spezielle Software, wo wir dann Prozesse abbilden. [...] Das ist noch mehr so der händische Weg. Dass man wirklich per Hand ordnet, schaut wie ist die *Situation, wo wollen wir hin, wie können wir das erreichen" (T09, 03:03)*

4.4 Hemmnisse

In einer Gesamtbetrachtung der Hemmnisse, welche die Nutzung des Geschäftsprozessmanagements unter den aktiven und passiven Experten behindern, werden verschiedene Ursachen genannt. Hierbei sind die bereits angesprochenen fehlenden Ressourcen der klar dominierende Faktor, sowohl bei der Anwendung als auch für den praktischen Einstieg. Die Ursachen sind auf das besondere Wesen der Wirtschaftsförderung zurückzuführen: *„Und da bleibt dann oft die Sache auf der Strecke. Allein schon vor dem Hintergrund, dass man eben oft als Einzelkämpfer unterwegs ist." (T04, 15:06)*

Ein weiterer Hinderungsgrund bei der Umsetzung von Geschäftsprozessmanagement wird in der erforderlichen Aufgabenflexibilität gesehen: *„Die große Individualität ist* eigentlich das größte Hindernis. Dass man eigentlich nie weiß, wann man seinen Geschäftsprozess fertig hat oder ob er nicht im nächsten Moment schon wieder völlig überholt ist." (T05, 08:28)

Die hohe Individualität aufgrund spezieller Kundenwünsche erschwert die Ableitung standardisierter Prozesse in der Wirtschaftsförderung. Gleichzeitig ist dieses Vorhaben mit einem hohen Aufwand verbunden: „Der Aufwand, den man treiben müsste, um eine Geschäftsprozessoptimierung *[...]* wirklich voranzubringen, steht keiner Relation voraus, hinsichtlich zum Ergebnis. Ich habe ganz wenige Standardgeschäftsprozesse, die ich optimieren kann." (T02, 07:55)

Ergänzen lassen sich die Hemmnisse durch Aspekte des Changemanagements, die von den Experten vor allem auch bei IT-Vorhaben genannt werden: „Man rennt nicht immer bei jedem Kollegen, den man in Prozesse einbeziehen will, offene Türen ein. Da gibt es Leute, die sich fürchten, dass man ihnen was wegnehmen will, dass man Kompetenzen beschneiden will, dass man überwachen will. Von daher ist es nicht immer nur so, dass da alle sagen: hey juhu, endlich organisiert man sich mal. Es gibt auch Menschen die haben Angst, vor Veränderungen." (T16, 09:56)

4.5 Perspektive

Im Rahmen der Interviews äußerten sich die Experten auch zur Einschätzung der künftigen Notwendigkeit des Einsatzes von Geschäftsprozessmanagement in der kommunalen Wirtschaftsförderung. Die Meinungen der Wirtschaftsförderer reichen von vollster Zustimmung bis hin zur skeptischen Differenzierung, die eine explizite Abgrenzung von spezifischen Wirtschaftsförderungstätigkeiten oder den Einsatz von Geschäftsprozessmanagement in Abhängigkeit zur institutionellen Größe der Wirtschaftsförderung ver-

langen. Erstaunlich an der Bandbreite der Auskünfte war das Fehlen einer völligen Negation der künftigen Notwendigkeit. Unisono ergab die Sicht der Experten, dass professionelles Geschäftsprozessmanagements künftig auch für die kommunale Wirtschaftsförderung eine Notwendigkeit darstellen wird.

Als besondere Gründe für diese Notwendigkeit wurden in den Interviews verschiedene Angaben gemacht, welche die aufbauorganisatorische, fachliche und prozessuale Komplexität der Wirtschaftsförderungstätigkeit auf kommunaler Ebene besonders verdeutlicht. Genannte Beispiele betrafen im Allgemeinen die qualitative Erwartungshaltung von Unternehmen und der Politik sowie den wachsenden Medieneinsatz zur Fallbearbeitung bei gleichzeitig steigender Interaktionskomplexität. Im Speziellen wurden darüber hinaus Steuerungsziele zur Qualitätsoptimierung, potenziell verbesserte Außenwirkungen und eine Steigerung der Kundenorientierung im Sinne einer Aufgaben- und Ablaufprofessionalisierung verstanden. Ferner wird die Notwendigkeit für Geschäftsprozessmanagement auch mit dem Auftrag zur Dienstleistungsfähigkeit für die Wirtschaft, im Sinne der Wirtschaftsförderung als professionelles Bindeglied zwischen Behörde und Unternehmen, verstanden: *„Geschäftsprozessmanagement wird eine relativ große B*edeutung haben, weil kommunale Wirtschaftsförderung letztendlich eine freiwillige Leistung ist. Je mehr ich meine Prozesse transparent machen kann, je mehr ich meine Prozesse auch besser messbar machen kann und je mehr ich meine Prozesse dokumentiere, desto besser kann ich zeigen, was wir eigentlich machen.*" (T03, 11:23)*

Ein sehr konkreter Anwendungsbezug für Geschäftsprozessmanagement wird retrospektiv vor allem bei der Einführung und im späteren Betrieb von Kundendatenmanagementsystemen als auch von Dokumentenmanagementsystemen gesehen. Aber auch die Nutzung im Bereich des Portfoliomanagements und der Strategiefindung spiegelt sich in verschiedenen Aussagen wider und wird als lohnende Perspektive für Geschäftsprozessmanagement in der Wirtschaftsförderung genannt. Das Ziel wird hierbei in transparenten Controllingmechanismen zur Angebotsqualität und Nachfragequantität gesehen, ergänzt durch qualitative Bewertungen der Angebotsnotwendigkeit im Kontext des Standortwettbewerbs. Dazu wird ergänzt: *„Die Rahmenbedingungen werden sich ändern* und die Möglichkeiten, Rahmenbedingungen zu gestalten, werden sich stark erhöhen. Wirtschaftsförderung muss diesem Schritt standhalten und muss auch vor diesem Hintergrund die Möglichkeit einer optimalen Gestaltung von Geschäftsprozessen Rechnung *tragen." (T12, 17:55)*

Darüber hinaus verdeutlichen zwei Argumente aber auch eine gewisse Grundskepsis gegenüber der Notwendigkeit von professionellem Geschäftsprozessmanagement in kommunalen Wirtschaftsförderungen. Zum einen wird aufgrund der häufig einzelfallgetriebenen Tätigkeiten bezweifelt, dass durch Geschäftsprozessmanagement positive Effekte zur Professionalisierung erzielt werden können. Zum anderen wird aufgrund mangelnder Ressourcen und der hohen Komplexität der mögliche Nutzen gegenüber dem zu betreibenden Aufwand in Frage gestellt, sodass die Probanden den Einsatz von Geschäftsprozessmanagement in der kommunalen Wirtschaftsförderung eher skeptisch beurteilen: *„Ein durchstrukturiertes, institutionalisiertes Prozessmanagement innerhalb* der Wirtschaftsförderung, stelle ich mir so extensiv wie man es in vielen anderen Bereichen vorfindet, nur schwer vor." *(T06, 07:53)*

4.6 Erfolgsfaktoren

In Teilen wurden durch die Experten mit bekundetem Erfahrungswissen zum Geschäftsprozessmanagement in der Wirtschaftsförderung auch Reflexionen ihrer GPM-Aktivitäten genannt. Ein häufig bekundetes Phänomen ist das Ignorieren der Notwendigkeit für eigenes Geschäftsprozessmanagement, bis die gegebenen Umstände radikale und meist politisch diktierte Prozessveränderungen definieren. In deutlich milderer aber sehr anschaulicher Form, umschreibt dies ein Wirtschaftsförderer mit folgenden Worten: *„Sie kennen ja das alte Bei*spiel von demjenigen, der in den Wald geht und mit der *stumpfen Axt hakt. Und jetzt fragt man ihn: ‚warum hakst du mit der stumpfen Axt?' ‚Ja* weil ich keine Zeit habe zum Schärfen.*' Er braucht halt das Zehnfache an Zeit dafür.* Und so ähnlich ist es im täglichen Ablauf auch. Irgendwann ist halt die Axt so stumpf *[…], dass man sagt, Mensch jetzt muss ich mir aber wirklich Gedanken machen, wie ich das optimieren kann." (T02, 03:46)*

Der empirische Befund zeigt sehr deutlich, dass im Allgemeinen die Einrichtung eines Geschäftsprozessmanagements in der kommunalen Wirtschaftsförderung als sinnvoll erachtet wird. Es wird jedoch von proaktiv erfahrenen Experten in der vorliegenden Analyse darauf hingewiesen, dass vor allem das Maß beziehungsweise der sinnvolle Umfang des Geschäftsprozessmanagements ein ausschlaggebender Erfolgsfaktor für entsprechende Optimierungen in der Wirtschaftsförderung darstellt. Vor allem für ad hoc entstehende Prozesse mit Projektcharakter (einzigartige, innovative, einmalig auftretende Arbeitsanforderungen) empfiehlt sich aus Sicht der Befragten kein GPM-Ansatz, wie folgendes Ankerbeispiel verdeutlicht: *„Ja, Geschäftsprozessmanagement ist in der Wir*tschaftsförderung sinnvoll. Man darf das nur sicherlich nicht übertreiben, weil wir natürlich i*mmer in einem Bereich sind, wo viele Aufgaben kurzfristig neu hinzukommen."* (T06, 05:25)

Darüber hinaus führen einige Experten mit Langzeiterfahrung beim Einsatz von Geschäftsprozessmanagement in der Wirtschaftsförderung auch evaluiertes positives Feedback ihrer Kunden an, wie zum Beispiel: *„Ja [GPM in der Wirtschaftsförderung ist* sinnvoll]. Also die Unternehmen geben die Rückmeldung, dass die Wirtschaftsförderung *im Vergleich zu vor zehn Jahren […] heute ganz anders positioniert ist und dass unsere* Außenwirkung wesentlich professioneller ist, [z.B. durch] die Rückmeldung, wie Projek*te abgelaufen sind."* (T03, 05:31)

Als spezifische Erfolgsfaktoren werden vor allem ausreichende personelle und finanzielle Kapazitäten für die Einführung und den Betrieb eines Geschäftsprozessmanagements in der kommunalen Wirtschaftsförderung angeführt. Auch das Aufzeigen von gezielten Verbesserungen durch eine ganzheitliche Prozessanalyse und –modellierung wird als Erfolgsfaktor, insbesondere im Dialog mit weiteren Stakeholdern (Politik, Verwaltung, Unternehmen und anderen) der Wirtschaftsförderungsinstitution, genannt: *„Geschäfts*prozessmanagement ist absolut sinnvoll und zielführend. Weil nur so, kann jeder für sich selbst, die Art des [Ressourcen-]Einsatzes messen. Um zu sagen, was verwende ich denn *in meiner Zeit, auf welche Dinge und wo […] verschwende ich Zeit, […] um [dadurch] halt mehr Effizienz zu erreichen und mehr Zielorientierung für bestimmte Inhalte." (T15,* 10:50)

Diese Darstellung wurde in Einzelfällen auch durch Qualitätsmanagementaspekte ergänzt. Demgegenüber stehen jedoch klare Ressourcenprobleme für den Betrieb eines professionellen und ganzheitlichen Geschäftsprozessmanagements. Vor allem kleinstorganisierte Wirtschaftsförderungsinstitutionen sehen berechtigter Weise keinen Mehrwert für internes Geschäftsprozessmanagement, welches sich ausschließlich auf ihren eigenen Bereich bezieht: *„Geschäftsprozessmanagement ist immer sinnvoll, wenn die Wirtschaftsförderung mehr als ein oder zwei Beschäftigte umfasst, […] um do*rt [in der Kundenbetreuung] den Informationsfluss zu organisieren, zum Beispiel in der Grundstücks*betreuung." (T16,* 07:34)

Auch die starke Projektcharakteristik vieler Wirtschaftsförderungsaktivitäten bestärkt die befragten Experten in ihrer Zurückhaltung gegenüber einem internem Geschäftsprozessmanagement, wie folgendes Ankerbeispiel verdeutlicht: *„Man muss sich auf der* anderen Seite aber auch fragen: W*elchen Nutzen habe ich davon. […] Denn so ein Pr*ozessmanagement muss ja auch gepflegt werden. [Aufgrund vieler Einzelfälle bin] ich eigentlich ständig damit beschäftigt meine Prozesskette immer wieder neu, sauber zu definieren. Nur um beim nächsten Mal wieder zu merken, dass sie trotzdem nicht funkti*oniert." (T05, 06:48)*

Insbesondere die fehlenden Ressourcen schränken auch jene Wirtschaftsförderungen mit positiven Erfahrungen bei der Realisierung eines internen Geschäftsprozessmanagements ein, weil dieses dann häufig nicht nachhaltig geführt beziehungsweise gepflegt werden kann. Hierdurch werden die positiven Effekte dann nicht mehr nutzbar. Aus diesem Grund empfiehlt ein Experte folgende pragmatische Vorgehensweise: *„Ein g*ewisses Volumen muss vorliegen, dass wir Geschäftsprozesse bildlich, graphisch, lyrisch auch entsprechend dokumentieren. Tagesgeschäft, [also] Prozesse mit einer geringen Komplexität, Prozesse die unhandliche Erfahrung aufweisen, Prozesse die wir nach *unserer Meinung sehr schnell steuern und abwickeln können, dokumentieren wir nicht."* (T12, 11:33). Ein weiterer Proband ergänzt zum GPM-Einsatz anschaulich: *„Der g*esamtbürokratische Aufwand behindert uns natürlich bei der Arbeit. Sonst ist eigentlich eher nichts [dagegen einzuwenden]. Dadurch sind eher Dinge besser geworden, wenn *man sich darüber Gedanken gemacht hat."* (T18, 03:51)

5 Reflexion, Limitation und Forschungsdesiderate

Die hier zugrunde gelegten empirischen Befunde lassen für den Kreis der Befragten folgenden Zustand für GPM in der Wirtschaftsförderung skizzieren: Geschäftsprozessmanagement in der Wirtschaftsförderung ist reaktiv, eher zufällig und spontan, häufig ohne vorgegebene Strukturen und mit sehr limitierten Ressourcen ausgestattet. Unter Berücksichtigung der verschiedenen Reifegradmodelle lässt sich das GPM in der Wirtschaftsförderung einem Anfangsstadium zuordnen ([AK09, S. 7–8]; [Bu11, S. 10–11]).

Es ist erneut der explorative Charakter der durchgeführten Vorstudie zu betonen. Nachvollziehbare Rückschlüsse und repräsentative Aussagen zum Entwicklungsstand sind aufgrund des Samplings der Stichprobe nicht möglich. Auch ergaben die Ergebnisse keine direkten Hinweise auf eine Abhängigkeit der Anwendung von GPM in Relation

zur Organisationsgröße oder –struktur. Hierfür ist eine quantitative Folgeuntersuchung notwendig. Für die weitere empirische Erforschung wird folgende zentrale Annahme festgehalten: Die spezielle Typologie der Wirtschaftsförderung erschwert GPM-Vorhaben.

Im Fokus weiterer Aktivitäten stehen daher die Untersuchung der Struktur und Rahmenbedingungen der Wirtschaftsförderung, um deren Einfluss auf die Entwicklung von GPM näher bestimmen zu können. Dieses ist auch Grundlage für die Herleitung von Standardprozessen. Des Weiteren ist die Untersuchung von Fallbeispielen und die Ermittlung von Good Practices geplant, um den Wirtschaftsförderungen praxiserprobte Empfehlungen im Umgang mit Geschäftsprozessmanagement aufzeigen zu können.

6 Zusammenfassung

Auf Basis von Expertenbefragungen wurden in diesem Beitrag Erkenntnisse zum Status Quo des Geschäftsprozessmanagements in der kommunalen Wirtschaftsförderung aufgezeigt. Die Untersuchung hat verdeutlicht, dass es verschiedene Anknüpfungspunkte für Geschäftsprozessmanagement gibt. Die Ergebnisse zeigen aber auch, dass GPM in der Wirtschaftsförderung noch am Anfang steht und dass es bei der Anwendung große Unterschiede gibt. Die Entwicklung wird offensichtlich durch verschiede Faktoren beeinflusst. Hieran sollen weitere Forschungsaktivitäten anknüpfen. Geplant sind zunächst die Untersuchung von Fallbeispielen und eine bundesweite quantitative Befragung, um die in diesem Beitrag aufgezeigten Entwicklungstendenzen zu verallgemeinern.

Literaturverzeichnis

[Ac12] Ackermann, G.: Erfolgskontrolle in der kommunalen Wirtschaftsförderung. Analyse und Modifikation einer fragwürdigen Forderung, 2012, 2012.

[AK09] Allweyer, T.; Knuppertz, T.: EDEN – Reifegradmodell Prozessorientierung in Unternehmen, 2009.

[Al05] Allweyer, T.: Geschäftsprozessmanagement. Strategie, Entwurf, Implementierung, Controlling. W3L-Verl., Herdecke, Bochum, 2005.

[BAF09] Becker, J.; Algermissen, L.; Falk, T.: Prozessorientierte Verwaltungsmodernisierung. Springer, Berlin, Heidelberg, 2009.

[Bl03] Blume, L.: Kommunen im Standortwettbewerb. Nomos-Verl.-Ges., Baden-Baden, 2003.

[Bu11] Regierungsprogramm Vernetzte und transparente Verwaltung. Projekt Prozessmanagement. Reifegradanalyse, 2011.

[DP13] Dresing, T.; Pehl, T.: Praxisbuch Transkription. Regelsysteme, Software und praktische Anleitungen für qualitative ForscherInnen. Marburg, 2013.

[DR12] Dallmann, B.; Richter, M.: Handbuch der Wirtschaftsförderung. Praxisleitfaden zur kommunalen und regionalen Standortentwicklung. Freiburg, 2012.

[Ei01] Eichhorn, P.: Öffentliche Dienstleistungen. Reader über Funktionen, Institutionen und Konzeptionen. Nomos, Baden-Baden, 2001.

[FHP13] Fuchs, T.; Hansen, C.; Pickenäcker, B. A.: Wirtschaftsförderung. Winkler & Stenzel, 2013.
[Ga83] Gaitanides, M.: Prozessorganisation. Entwicklung, Ansätze und Programme prozessorientierter Organisationsgestaltung. Vahlen, München, 1983.
[Gö12] Göbel, A.: Verwaltung als Standortfaktor für Unternehmen. Osnabrück, 2012.
[Ha04] Haug, P.: Kommunale Wirtschaftsförderung. Eine theoretische und empirische Analyse. Kovac, Hamburg, 2004.
[HC94] Hammer, M.; Champy, J.: Business reengineering. Die Radikalkur für das Unternehmen. Campus Verlag, Frankfurt, 1994.
[HKL10] Henckel, D.; Kuczkowski, K.; Lau, P.: Planen - Bauen - Umwelt. Ein Handbuch. VS-Verl., Wiesbaden, 2010.
[KGSt14] Prozessregister KGSt. https://www.kgst.de/login/anmelden.dot, 05.05.2014.
[Ku08a] Kulke, E.: Wirtschaftsgeographie. Schöningh, Paderborn, München, Wien, Zürich, 2008a.
[Ku08b] Kuckartz, U.: Qualitative Evaluation. Der Einstieg in die Praxis. In Qualitative Evaluation, 2008b.
[LGK14] Lück-Schneider, D. et al. (Hrsg.): GI Edition Proceedings Band 229 - Gemeinsam Electronic Government ziel(gruppen)gerecht gestalten und organisieren. Köllen, Bonn, 2014.
[Lü12] Lück-Schneider, D.: Geschäftsprozessmanagement in der öffentlichen Verwaltung. Eine Einführung. Neuhofen, 2012.
[NG14] Naumann, R.; Grabner, W.: Prozessmanagement als wesentlicher Faktor der Herausforderungen des demographischen Wandels in der öffentlichen Verwaltung. Ein Kooperationsprojekt von PwC und IMTB, 2014.
[No32] Nordsieck, F.: Die schaubildliche Erfassung und Untersuchung der Betriebsorganisation. Poeschel, Stuttgart, 1932.
[NPB14] Nationale Prozessbibliothek. http://www.prozessbibliothek.de/, 05.05.2014.
[RR03] Reschl, R.; Rogg, W.: Kommunale Wirtschaftsförderung. Standortdialog und Standortentwicklung in Kommunen und Regionen. Sternenfels, 2003.
[Sc12] Schumann, S.: Repräsentative Umfrage. Praxisorientierte Einführung in empirische Methoden und statistische Analyseverfahren, 2012.
[SKJ06] Scheer, A.-W. et al. (Hrsg.): Agilitat durch ARIS Geschäftsprozessmanagement. Jahrbuch Business Process Excellence 2006/2007. Springer, 2006.
[SS13] Schmelzer, H. J.; Sesselmann, W.: Geschäftsprozessmanagement in der Praxis. Kunden zufrieden stellen - Produktivität steigern - Wert erhöhen. Hanser, München, 2013.
[Wa14] Walser, K.: Ordnungsrahmen zum Einsatz des Geschäftsprozessmanagements und des Dokumentenmanagements in der Öffentlichen Verwaltung. In (Lück-Schneider, D. et al. (Hrsg.)): GI Edition Proceedings Band 229. Köllen, Bonn, 2014, S. 81–95ff.
[WS09] Weißenberg, N.; Stemmer, M.: Moderne IT-Plattformen für Geschäftsprozessmanagement und Portale. Vergleichende Bewertung der Toolsuiten von IBM, IDS Scheer & SAP sowie lntalio & Liferay. Kurzfassung, Dortmund, 2009.

Semantic based auto-completion of business process modelling in eGovernment

Eliane Maalouf, Maria Sokhn

Institute of information systems
University of applied sciences of western Switzerland, Valais-Wallis
Technopôle 3
3960 Sierre
eliane.maalouf@hevs.ch
maria.sokhn@hevs.ch

Abstract: So far, the semantic business process modelling is gaining in importance mainly due to the advances in the domain of semantic web technologies, which help enhancing, the modelling issues of the business process management life cycle. The linked data approach becomes interesting to apply on Business Processes to enrich them with information from diverse sources and link them to those sources. The aim of our project is to enhance the phase of process modelling of Swiss eGovernment public processes by suggesting to the modeller process fragments for auto-completion from a semantic repository. Created processes are linked and semantically enriched with ontology concepts related to the domain of public sector and eGovernment that will enhance query results in the repository as well as improve process documentation. We focus on linking the processes to the public sector standards.

1 Introduction

Business Process Management is the discipline concerned with business processes optimization. The BPM life cycle is generally divided into four phases: design, deployment/configuration, execution and analysis. Process modelling in the design phase plays a central role in this life cycle. Process models assist in managing complexity by providing insight and by documenting procedures. Van der Aalst survey on Business Process Management (BPM) [WAM03] states that the notion of a process model is foundational for BPM. He mentions that a process can be modelled from control-flow perspective (modelling the ordering of activities), resource perspective (roles, organizational units, authorizations, etc.), data perspective (decisions, data creation, forms, etc.), time perspective (durations, deadlines, etc.) and the function perspective (activities and related applications). In a recent work, the authors of [Ni13] stated that existing process knowledge in order to simplify the modelling process has not yet been applied. Van der Aalst categorized the main activities related to the management of large process model collections as search (given a query a set of models is returned), merge (combine a set of models into a single one), cluster (identify a set of related models), unify/refactor (improve the models by aligning them, removing redundancies, etc.), and convert (map from one notation to another). Van der Aalst mentions that a general problem affecting all activities is the use of informal text. The same activity may be

labelled differently in different process models thus hindering the detection of redundancies and inconsistencies. To avoid this problem, he suggests using a common ontology. In his opinion, semantic technologies aim to address obvious limitations related to string edit distances and linguistic similarity. However, in practice, few process model collections use a common ontology. Therefore, in most cases, semantic annotations still need to be added to process models before being able to use semantic technologies.

Work on Semantic Business Process Management have proliferated in the last few years, following the European Framework 6 funded project SUPER[1]. This project defined a Semantic Business Process life cycle adding automation to the bi-directional translation between the business and the systems spheres. The fundamental approach is to represent both the business perspective and the systems perspective of enterprises using a set of ontologies, and to use machine reasoning for carrying out or supporting the translation tasks between the two worlds. The project resulted in a semantic services architecture connecting semantic modelling and configuration tools, semantic execution and analysis tools and semantic repositories over a semantic service bus. It also resulted in a set of ontologies to support the design, resumed by [Di07] as follows:

- Upper Process Ontology (UPO), defining top-level concepts such as task, goal and condition
- Business Process Modelling Ontology (BPMO), extending the UPO into a full process ontology, providing abstractions over different business process modelling notations such as BPMN and EPC
- sBPMN , sEPC and sBPEL – ontologised versions of subsets of the BPMN, EPC and WS-BPEL respectively. sBPEL is additionally enriched with extensions from the Web Services Modelling Ontology (WSMO) for goal-oriented discovery, mediation and execution of services

The SUPER project allowed automatic discovery of services using semantic web technologies and linked them to process activities to ease configuration before execution. Our approach to the problem is different in the sense that we focus on the business process design phase only as will be described later on.

2 Swiss E-Government Public Business Processes Context

Switzerland is investing more and more in E-Government initiatives to modernize the electronic platforms to support citizens' access to services. The "eCH association" is responsible for elaborating the standards for cyber administration adoption in Switzerland. Those standards are emitted as recommendations and are freely available [eCHa]. The standard documents eCH-0140 and eCH-0158 defined the Business Process Model and Notation (BPMN 2.0) as the recommended notation to model Swiss administration public processes and defined some best practices for modelling inspired from the notation [eCHb] [eCHc].

[1] http://www.ip-super.org/content/view/45/69/

To complete those standardization efforts, the Swiss confederation along its State Secretariat for Economic affairs (SECO) started multiple priority projects to lead the research in cyber administration. One of those projects is the "B1.13: eCH-platform for process exchange between communities and cantons". The aim of this project is to ease knowledge transfer on business processes from one entity to another, share documentation about processes, surpass the languages barriers and offer standard procedures to external partners among other goals. This platform will play the role of a global process documentation platform and process repository. The different municipalities and cantons that form the administration body of the country provide processes. The platform is currently in its Beta-Version and accessible through http://www.ech-bpm.ch. Since it is a process documentation platform, execution data about processes are not collected and configuration of processes does not fall in its requirements. The modelling environment, subject of this paper, can be seen as the modelling tool on top of this public repository and provides process editing interface to the modellers. Hence, for modelling we are only concerned by the process model elements and its documentation without configuration details.

In their study about "Semantic Business Process Management for e-Government", Liu et al. [Zh13] concluded that in order to improve the degree of automation processing in business process management, techniques from semantic web such as ontology languages and reasoners have to be transferred to the business process world. An improved semantic management is thought to fix some of the inherent problems related to semantic deficiencies in government documents, inconsistencies of semantic information among different agencies, complex change management activities and poor process management practices. The authors mention that those problems seem to be more complex in Switzerland given its structure of 26 cantons and multiple municipalities with different degrees of independence in decision-making and the use of four different official languages.

3 Background

3.1 Business Process Modelling Notation

The BPMN language provides a graphical representation for the process designers to organize and model their business processes. Its recent 2.0 version defines the formal semantics of its elements. BPMN graphical objects are divided into three main categories: Flow Objects (Activities, Gateways and Events), Connecting Objects (Sequence Flows) and Swimlanes (Pools and Lanes) [Ch12]. Additionally, BPMN2.0 introduces extension capabilities to Flow Objects and Artifacts to allow process designers to express additional features of process models.

The authors of [Ch12] provide a formal definition of the BPMN meta-model syntax and specifications written in natural language. The ontology is called BPMN 2.0 Ontology, written in the Web Ontology Language (OWL) and has 85 classes, 178 object properties

and 59 data properties. This ontology will be further used to assert the structural elements of the process model.

3.2 Linked Data and Linked BPM

Linked Data is a set of best practices for publishing and interconnecting structured data on the web. The authors of [Ch09] provide an overview of its principles. Linked Data provides explicit links between data from diverse domains with the goal of increasing machine-readability of published data. Linked Data is published using the Resource Description Framework where URIs refer between various entities on the web.

The work of F.Gao in [Fe11] provided a case to implement the concepts of Linked Data with BPM. It presented BPMN2.0 extension examples to link a process to different views of the Architecture of Information System (ARIS) framework: organization view, functional view, data view, control view and control/service view. The process model in this framework is the controlling dispatcher of resources and actions. Extensions to BPMN2.0 are typically created and annotated in the design phase. The views are modelled with linked data, they are interlinked with each other and can be further linked to the external open data. BPMN 2.0 defines Extension element that can extend Flow Objects (e.g. Sub-processes, Activities, Tasks) and Relation element that can extend Artefacts (e.g. Group, Data Object, Text Annotation). Extensions in the different views are suggested as follows:

- Organization view: defines the roles, the participants, the organizational entities and the relationships between them. BPMN2.0 Activities should be annotated with organizational information like: who is responsible with executing the activity (individual or group), which organization/group does he belong to and what role does he have. BPMN 2.0 PartnerRole, Participant and PartnerEntity should be connected to concepts from dedicated organizational ontologies for roles, participants and organizational entities respectively.
- Data view: describes the information objects, the concepts in the messages exchanged by process tasks. Data encapsulated in BPMN ItemAwareElement should be annotated with concepts from domain specific ontologies.
- Control view: describes the control flow of business processes. BPMN Activity, Event and EventDefinition should carry information about occurrence time and other execution related data.
- Function view: defines the functions required to satisfy the objective of the enterprise. BPMN Activity should be defined as a domain specific business function instead of just a label.

3.3 Semantic Business Process Modeling with BPMN

Process modelling in eGovernment has been mainly been performed with standard modelling languages such as Activity Diagrams (AD) [Uml], Business Process Modelling Notation (BPMN) [Bpmn], or Event Driven Process Chains (EPC) [Sc00]. These approaches do not take into account the context of the eGovernment process

where activities may depends on legal regulations. Some works have been conducted to brdge this gap such as the PICTURE approach proposed by [Jo07]. This approach represents not only isolated processes but a complete landscape of the organisation taking into account technical and organisational measures. In their work, the authors of [Vo13] proposed a modelling approach that consists of building blocks within BPMN.

Requirements for a semantic modelling environment, were described by [Di07], we retain the following:

- Adding semantic annotation of existing process models, i.e. adding references to ontology elements.
- Storing the semantic process models into a Semantic Business Process Library
- Querying of the library for discovery of existing semantic process models or fragments for reuse through auto-completion, which will decrease the effort and time required for modelling of new processes

In their work the authors of [Ag09] define interesting scenarios for the annotated business models: decision making support, reuse BP patterns, reuse BP models/fragments, auto-completion, querying policies, goal-based BP analysis. In order to enable those scenarios, the paper suggested the use of organizational ontologies: Organizational Structure Ontologies (OSO), Organizational Units Ontology (OUO), Business Roles Ontology (BROnt), Business Functions Ontology (BFO), Business Resources Ontology (BRO), Business Goals Ontology (BGO). At the moment of writing of this paper, we were not aware of organizational ontologies representing the Swiss public administration.

In [Ch09], the authors present three types of assertions that can be carried out on annotated business processes: BP model type assertions to store information on the type of a graphical object; BP model structural assertions to store information on how the graphical objects are connected and BP model semantic assertions to represent annotations of graphical objects.

The authors of [MP07] provide a model for semantically annotating business process models stored in a repository to enhance process lookup, navigation in the repository and understanding of process models. It suggests annotating processes with the following perspectives: functional (process type, process area), behaviour (timing of activities), organizational (actor, organization level), informational (exchange process phase) and business process context (process relationship, business context, and goal).

Given the time allocated for this project, missing organization and domain ontologies cannot be engineered. We favour the use of an annotation scheme similar to the last model presented above combined with the ARIS views annotation model.

3.4 Business Process Auto-completion

The author in [Th07] suggests to complete processes at modelling time with fragments following a defined set of rules: Constraint rules concerning the integrity of semantic business process models; Event-Condition-Action rules having a syntax described by IF,

THEN, AND, OR and Dynamic Rules applied during process modelling by the modeller.

On the other hand, [Ag10] presents a modelling environment that takes the user's modelling context and the modelling history of a community of users into account. Rankings of the retrieved processes from the repository are based on similarity between the query and the model retrieved, patterns observed in other users' preferences and implicit user feedback.

Born et al. [Ma09] proposes a weighted average of three auto completion criteria. The criteria are process-context based analysis, pre-and post-condition analysis and non-functional property analysis.

In our project, we will base auto-completion algorithm on similarity with the process part that was modelled so far based on structural similarity constructed by querying with SPARQL the business process RDF definition. Context information are also included in the modelling interface to narrow the lookup space. Those annotations will be described later on.

4. Aim

From the context information provided above, we notice that managing a large process repository is a complex task especially in the case of the Swiss administration. First, managing a large amount of processes requires automations at some leves of the life cycle to avoid time consuming. Second, this administration is multi-leveled (cantons, communes) with relative degrees of freedom in decision making added to that the use of four national languages. The main issue of this context can be described as followed:

- Each commune can model its processes in its own language (i.e. German, French, Italian or Romansh). The level of language comprehension by the analyst hinders sharing knowledge about processes in the repository.
- Modelling procedures using BPMN is not an easy task especially for small communes who do not necessarily have a specialized analyst in process modelling. Humans modelling new processes do not have access to existing models in the repository through the modelling tool to lookup similar processes or fragments that can be reused.
- Business Processes are labelled and annotated using natural language that introduces inconsistencies in the usage of the language (e.g. the same activity can be named differently by different modellers). This makes the handling of the processes by machines more complex to extract context information.
- Documentation of processes is limited to the explicit information that can be added by a human and mapping a process to other processes or domain is yet to be done manually. In large repositories this element is crucial.

We aim to prove the concept that semantic web technology and linked Data can help address the mentioned issues in an efficient way.

5. Method

In order to attain the objective described earlier, we target the development of a semantic process modeling environment on top of a process repository containing a set of 50 processes from a local commune. In order to address the issues above the tool makes available multiple functionalities briefly described below. The development is taking place in iterations, at this stage of the work, some functionalities may be more detailed than others.

5.1 Modelling Interface

We favoured the reuse of existing Open source modelling tool over redeveloping a new one. “Signavio Core Components” code is being used for the initial development phase since it supports BPM 2.0 elements, offers a user-friendly interface and is based on JavaScript. The code is light and does not ship with any necessary prior database configuration, which leaves the choice open for us. New features can be developed in the form of plugins. Since this code is integrated in multiple other open source tools like jBPM[2] and Activiti[3] plugins developed in our platform can be easily reintegrated in others.

5.2 RDF Representation of Processes

When a new process is added to the repository, we store a representation of it in Resource Description Framework format in a triplestore. Based on the BPMN 2.0 Ontology, an XSLT transformation file is used to transform the BPMN process. The initial process file is saved without modification in a separate traditional transactional database.

For example, we might get the following triples to describe the process file we are working on:

bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa rdf:type bpmn20base:Process .
bpmnFile:sid-2ad23ad3-68de-4eed-b3cb-907162de7da7 bpmn20base:rootElements bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa .
bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa bpmn20base:id "sid-ae602356-3fea-4054-abde-6a72aeba08fa"^^xsd:string .
bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa bpmn20base:isClosed "false"^^xsd:boolean .
bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa bpmn20base:isExecutable "false"^^xsd:boolean .
bpmnFile:sid-ae602356-3fea-4054-abde-6a72aeba08fa bpmn20base:processType bpmn20base:None .

[2] https://www.jboss.org/jbpm
[3] http://activiti.org/

5.3 Process Auto-completion

The auto-completion functionality in the modeller environment is being developed as an extension plugin. When the modeller adds an element in the interface an event is caught and a query is formed to the SPARQL endpoint that will query the structure of the processes for a similar construct. To narrow the query space filtering can be done based on annotation elements.

5.4 Process Translation and Vocabulary Management

The BPMN 2.0 standard enables the serialization of processes in XML format. Tools that support BPMN 2.0 support also serialization. In the XML representation of processes, name attributes of elements are the labels provided by the modeller for the given element. In our work, we only consider translating these attributes. For this proof of concept implementation we care to translate form German to French. Given the time we have on this project, implementing an advanced machine translation algorithms is not possible. We aim to offer an initial help to the modeller in the translation. The responsibility falls back on the human to confirm that the tool expressed correctly the concept in the destination language. In order to enable translation, we form a dictionary in RDF from a subset of the multilingual terminology database TERMDAT[4], a database for recording Swiss legal and administrative terminology along other terminology from the public sector. Then we populate GATE[5] Gazetteers with the dictionary concepts. The translation procedure would go as follows:

- The user activates the translation function for a process
- The process is annotated with GATE based on the Gazetteers. GATE enables the lookup of words in a given text using the Gazetteers and those matches are linked to dictionary concepts. This method only works to find exact matches.
- In no match is found, an external web translation API is called and requested a translation for the missing match. Meanwhile, partial matches are also looked up in the dictionary.
- The results from the local dictionary, the API are presented to the user is a form along the partial matches found. The user needs to confirm the translated text coming from the API or modify it.
- Once the user confirms the presented translation, this translation is added to the dictionary and will become the preferred translation to be used in subsequent translations when needed. A new process file is created where labels are in the destination language. If the user does not modify an API translation, the original text will save in the new process file to avoid having uncontrolled labels.

We favoured the construction of the dictionary with TERMDAT since the e-CH 0158 standard [eCHc] recommends the use of a standardized vocabulary and advised the use of this terminology database. To implement this recommendation furthermore, we

[4] http://www.bk.admin.ch/themen/sprachen/00083/00854/index.html?lang=en
[5] https://gate.ac.uk/

modify the modelling tool to allow the suggestion of words from the dictionary to label process elements and annotations.

5.5 Repository Querying

We defined a set of lookup queries to implement in order to question the repository for relevant processes. The queries are being constructed with SPARQL. The result is an identification of processes that answer this query. The queries are then displayed to the interface from the relational database. The queries being implemented are: lookup a match of a label in the repository (results include equivalent match in other languages), lookup based on a graphical input query, lookup for processes that start with an event E1 and followed by Activity A1, lookup all processes owned/executed/published by a person/role/entity, lookup for all processes that include a given fragment or sub-process, lookup all processes that include interaction between two named participants, lookup all process related to a given service/mission, lookup all process written in a given language.

The query interface will enable the user to enter a graphical query by constructing a part of the process he is looking for as well as a form to enter more semantic annotations to be looked up.

5.6 Linking Processes

In order to define interesting meta-data to add to processes in the repository of the B1.13 project, business consultants defined a set relevant to the domain. A basic ontology will be constructed using those concepts. These meta-data will be used to annotate the process with context information. A subset of those annotations included and their relation to the ARIS framework: process type (functional), total time (functional), owner (organizational), required documents (data), etc. Only annotations that are related to specific elements and not the whole process can be attached following the extension model presented by of F.Gao in [Fe11]

Two required annotation types can be linked to an external ontology: mission id and service id. More precisely, the eCH association emitted eCH-070, a standard listing the administration services [eCHd]. This standard was transformed into an ontology and is published under http://logd.ch/voc/service.html . For missions, there is a similar standard that is not published yet for which an internal ontology was developed and will be linked to the processes in our proof of concept.

6. Conclusions and future work

The current stage of the development is working on the interfaces, to integrate the modules of auto-completion and the translation presentation. In parallel, the translation module is being developed and dictionary management is being set. The results that will come out of the first tests will be compared to graph based techniques currently used in

modelling tools. Links to more data sources will be prepared, especially to link participants in processes to ontology instances in Friend of a Friend (FOAF) ontology for example.

In order to showcase the potential of Linked Data with business processes, a set of electronic public sector processes (e.g., eService suivi patrouilles) were annotated with information about the commune that offers it (e.g., Fribourg), the standard to which it is related (e.g., 400-10-012-006), the development language with which it was developed (e.g., Java). Then a visualization of this database was done using Linkurious. Such visualization is interesting in the case of a process repository where we try to make sense of the processes' relations between each other and their relations to the standards that regulate them.

References

[Uml] Object Management Group. UML 2.0 Supers tructure Specification. http://www.omg.org/cgi-bin/doc?formal/05-07-04

[Bpmn] Object Management Group. BPMN Final Adopted Specification 1.0. http://www.bpmn.org/Documents/OMG%20Final%20Adopted%20BPMN%2010%20Spec%2006-02-01.pdf

[Sc00] Scheer, A.-W.: ARIS - Business Process Modeling. 3 edn. Springer Publishing, Heidelberg et al. (2000)

[Jo07] Jörg Becker, Daniel Pfeiffer, Michael Räckers: PICTURE - A new Approach for Domain-Specific Process Modelling. CAiSE Forum 2007

[Vo13] Volker Schmitz, externer Berater des Projekts FIM, Föderales Informationsmanagement: Das FIM Vorhaben und seine Berührungspunkte zur XÖV Standardisierung 6. XÖV Anwenderkonferenz Bremen am 06. November 2013

[WAM03] Wil M. P. Van Der Aalst, Arthur H. M. Ter Hofstede, and Mathias Weske. 2003. Business process management: a survey. In *Proceedings of the 2003 international conference on Business process management* (BPM'03), Wil M. P. Van Der Aalst, Arthur Ter Hofstede, and Mathias Weske (Eds.). Springer-Verlag, Berlin, Heidelberg, 1-12.

[Di07] Dimitrov, M.; Simov, A.; Stein, S. & Konstantinov, M. (2007), A BPMO Based Semantic Business Process Modelling Environment., *in* Martin Hepp; Knut Hinkelmann; Dimitris Karagiannis; Rüdiger Klein & Nenad Stojanovic, ed., 'SBPM' , CEUR-WS.org, .

[eCHa] «eCH E-Government Standards,» Available: http://www.ech.ch/vechweb/page?p=categoryList&site=/documents/Standards/A%20bis%20Z [Last accessed on 06.05.2014].

[eCHb] F. Geschäftsprozesse, «eCH-0140 Vorgaben zur Beschreibung und Darstel-lung von Prozessen der öffentlichen Verwaltung der Schweiz,». Available:

http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0140&documentVersion=1.00. [Last accessed on 06.05.2014].

[eCHc] F. Geschäftsprozesse, «eCH-0158: BPMN-Modellierungskonventionen für die öffentliche Verwaltung,» 26 06 2013. Available: http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0158&documentVersion=1.0. [Last accessed on 06.05.2014].

[Zh13] Zhan Liu et al., Semantic Business Process Management for e-Governement, Proceedings of the IADIS International Conference WWW/INTERNET 2013, ISBN 978-989-8533-16-6.

[Ch12] Christine Natschläger-Carpella, Extending BPMN with Deontic Logic ISBN 978-3-8325-3225-3 313 Seiten, Erscheinungsjahr: 2012

[Ch09] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3): 1-22 (2009)

[Fe11] Feng Gao et al., Extending BPMN2.0 to Enable Links between Process Models and ARIS Views Modeled with Linked Data, Lecture Notes in Business Information Processing, 2011, p41-52

[Ag09] Agata Filipowska, Monika Kaczmarek, Marek Kowalkiewicz, Ivan Markovic, and Xuan Zhou. 2009. Organizational ontologies to support semantic business process management. In Proceedings of the 4th International Workshop on Semantic Business Process Management (SBPM '09), Martin Hepp, Knut Hinkelmann, and Nenad Stojanovic (Eds.). ACM, New York, NY, USA, 35-42. DOI=10.1145/1944968.1944975 http://doi.acm.org/10.1145/1944968.1944975

[Ch09] Chiara Di Francescomarino, Chiara Ghidini, Marco Rospocher, Luciano Serafini, Paolo Tonella, Semantically aided business process modelling, The Semantic Web - ISWC 2009. Lecture Notes in Computer Science Volume 5823, 2009, pp 114-129

[MP07] Mturi Elias, and Paul Johannesson. A context-based process semantic annotation model for a process model repository. Business Proc. Manag. Journal 19(3):404-430 (2013)

[Th07] Thomas Hornung , Agnes Koschmider , Andreas Oberweis, A.: Rule-based Autocompletion Of Business Process Models. In: CAiSE Forum 2007

[Ag10] Agnes Koschmider, Andreas Oberweis, Designing Business Processes with a Recommendation-Based Editor. Handbook on Business Process Management International Handbooks on Information Systems 2010, pp 299-312

[Ma09] Matthias Born et al. Auto-completion for Executable Business Process Models. Business Process Management Workshops. Lecture Notes in Business Information Processing Volume 17, 2009, pp 510-515

[eCHd] http://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0070

[Ni13] Nico Clever et al. Towards auto-suggested process modeling – prototypical development of an auto-suggest component for process modeling tools. 5th International Workshop on Enterprise Modelling and Information Systems Architectures, 2013.

Drittmittelprojekte in Hochschulen – Entwicklung eines fallstudienbasierten Referenzprozessmodells

Stefan Gröger, Matthias Schumann

Georg-August-Universität Göttingen
Professur für Anwendungssysteme und E-Business
Platz der Göttinger Sieben 5
37073 Göttingen
sgroege@uni-goettingen.de
mschuma1@uni-goettingen.de

Abstract: In Drittmittel (DriMi)-Projekten müssen eine Vielzahl von administrativen Tätigkeiten ausgeführt werden. Diese Tätigkeiten bieten diverse Ansatzpunkte für IT-Unterstützung. Hierdurch könnten Wissenschaftler von administrativen Tätigkeiten entlastet werden, so dass sie sich auf ihre Forschung konzentrieren können. Um geeignete Lösungen zu konzipieren, muss jedoch zunächst untersucht werden, wie solche DriMi-Projekte genau ablaufen. In diesem Beitrag wird daher ein Referenzprozessmodell (RPM) auf Basis von vier Fallstudien erarbeitet. Dieses liefert eine Ausgangsbasis für organisatorische Schwachstellen- und IT-Potenzialanalysen, so dass die oben genannte Situation verbessert werden kann.

1 Einleitung

Das gestiegene DriMi-Volumen stellt sowohl Wissenschaftler als auch die administrativen Einrichtungen einer Hochschule (HS) vor Herausforderungen [St13]. So müssen sich Wissenschaftler bspw. immer mehr administrativen Aufgaben widmen, wodurch weniger Zeit für Forschungsaktivitäten verbleibt [GS14a]. Aus diesem Grund wünschen sich Wissenschaftler eine Prozessunterstützung von der Projektidee bis zur Publikation [KE12]. In der zentralen HS-Verwaltung kommt es hingegen mit den derzeitigen Verfahrensweisen zu Kapazitätsengpässen [GS13]. Fallstudien belegen, dass häufig eine redundante Daten- und Dokumentverwaltung existiert, Medienbrüche an Abteilungsgrenzen vorzufinden sind und einheitliche IT-Lösungen zum Management von DriMi-Projekten kaum eingesetzt werden [EKB11, EKB13]. Innerhalb des DriMi-Prozesses scheinen somit Optimierungspotenziale vorhanden zu sein, die sich einerseits auf organisatorische Verbesserungen im Ablauf und andererseits auf IT-Unterstützung beziehen. Um dies zu belegen bzw. die Aussage auf Allgemeingültigkeit zu prüfen, gilt es zunächst ein Referenzprozessmodell abzubilden, welches den derzeitigen Ist-Stand dokumentiert. Anhand des Modells lassen sich anschließend Analysen bzgl. Verbesserungen durchführen und ein Soll-Prozess ableiten. Da derzeit noch kein Referenzprozessmodell existiert [GS14b], welches den Ist-Stand beschreibt und die DriMi-

Projektabläufe hochschulübergreifend vereint, besteht das Ziel dieses Beitrags darin, ein solches zu entwickeln. In die Modellentwicklung können extern verfügbare Prozessbeschreibungen nicht einfließen, da diese nicht in der notwendigen Modellierungstiefe vorliegen (vgl. bspw. [Un08]). Die zugrunde liegende Forschungsfrage lautet:

> *Wie ist der Prozess der DriMi-Projektbeantragung und -bewirtschaftung in Hochschulen allgemein gestaltet?*

Der Beitrag ist folgendermaßen aufgebaut: Nach dieser Einleitung werden im zweiten Abschnitt die Grundlagen dargeboten und anschließend die Forschungsmethodik (Abschnitt 3) erläutert. Im vierten Abschnitt wird das induktiv erstellte Referenzprozessmodell präsentiert, welches im darauf folgenden Abschnitt 5 diskutiert wird. Der Beitrag schließt mit einem Fazit in Abschnitt 6.

2 Grundlagen

2.1 Drittmittel

Eine einheitliche Definition des Begriffs Drittmittel (DriMi) ist in der Literatur nicht vorhanden [Ka08, Ha09, Mi10]. In diesem Beitrag wird daher die weitverbreitete Definition des Statistischen Bundesamtes verwendet. Dieses versteht unter dem Begriff DriMi „solche Mittel, die zur Förderung von Forschung und Entwicklung sowie des wissenschaftlichen Nachwuchses und der Lehre zusätzlich zum regulären Haushalt (Grundausstattung) von öffentlichen und privaten Stellen eingeworben werden“ [St11]. Grundsätzlich können drei verschiedene Abläufe von DriMi-Projekten unterschieden werden [GS13]: (1) Im Standardfall wird ein Antrag verfasst und an einen Geldgeber zur Beurteilung verschickt. Anschließend kommt es zur Bewilligung des Projektes oder es wird abgelehnt. Die beiden anderen Fälle können als erweiterte Fälle bezeichnet werden. Sie beziehen sich auf DriMi-Projekte, die ein zweistufiges Antragsverfahren aufweisen. Hierbei muss zunächst eine Skizze verfasst und durch den Geldgeber beurteilt werden. Es wird entschieden, ob ein (Voll-)Antrag gestellt werden darf. Die beiden erweiterten Fälle unterscheiden sich hinsichtlich des Merkmals *Kooperation*: Während im (2) Einzelprojektfall mit zweistufigem Antragsverfahren lediglich ein Wissenschaftler an einem DriMi-Projekt arbeitet, arbeiten im (3) Verbundprojektfall mehrere Wissenschaftler in einem Projekt zusammen.

2.2 Referenzprozessmodell

Ein Referenzprozessmodell (RPM) ist eine spezielle Art eines Referenzmodells (RM) [CWS13]. Daher wird zunächst dargelegt, was unter einem RM verstanden wird: Ein RM “is an information model used for supporting the construction of other models” [Th06]. Ein RM ist somit durch seine Allgemeingültigkeit gekennzeichnet [Br80] – was jedoch nicht heißt, dass es universell ist. RM besitzen zwar eine gewisse Allgemeingültigkeit, müssen jedoch nicht zwangsläufig der unternehmensspezifischen Situation entsprechen [FL07, Be02b] – d. h. aus einem RM können Instanz-spezifische Modelle ab-

geleitet werden [Re99]. Diese Eigenschaften lassen sich auf RPM übertragen. In einem RPM werden somit allgemeingültige Abläufe dargestellt, die von den Abläufen spezifischer Organisationen generalisiert wurden und einen Empfehlungs-/Dokumentationscharakter besitzen [SL12, FL07, FL04]. Die RPM müssen dabei nicht sämtliche Alternativen bzw. Aktivitäten enthalten, sondern es geht vielmehr darum, eine einheitliche und gemeinsame Basis zur Orientierung zu schaffen.

3 Forschungsmethodik

3.1 Vorgehen bei der Referenzmodellierung

Das Vorgehen bei der Erstellung des RPM lehnt sich an BECKER ET AL. an [BKR12, BAF09, Be02a]. Demnach ist zunächst ein (1) *Modellierungsziel* festzulegen, dann die (2) *Modellierungsmethodik* auszuwählen und schließlich das (3) *Referenzmodell zu erstellen*. (1) Das *Modellierungsziel* besteht darin, den DriMi-Prozess zu dokumentieren und damit eine Ausgangsbasis für Prozessanalysen zu schaffen, welche helfen können, die Probleme im Rahmen von DriMi-Projekten zu bewältigen (vgl. zu Problemen bspw. [GS14b]). (2) Zur *Modellierungsmethodik* gehören drei Unterelemente [KTG97]: Als *Modellierungstechnik* kommt zum Erreichen des o. g. Ziels ein Flussdiagramm infrage [Gi01]. Daher wurde zur Dokumentation der zu erhebenden Prozessmodelle BPMN als *Modellierungsmethode* gewählt. Als *Modellierungswerkzeug* fand Microsoft Visio Verwendung. (3) Die *Erstellung des Referenzmodells* erfolgte ebenfalls in drei Schritten [SL12]: Zunächst galt es die (3a) *Ist-Modellierung vorzubereiten*. Da ein induktives Ist-Modell zu erstellen war (aufgrund des definierten Ziels), musste eine geeignete Datenerhebungsmethode gewählt werden. In der Literatur wird hierfür die Expertenbefragung unter Zuhilfenahme vorhandener Dokumentationen als geeignet beschrieben [SL12]. Dies geht mit dem Verständnis einer Fallstudie einher [Yi09], weshalb diese Methode zur Erhebung der Daten verwendet wurde (zum Fallstudiendesign vgl. Abschnitt 3.2). Der zweite Schritt der *Referenzmodellerstellung* bestand in der (3b) *Erhebung und Dokumentation der Ist-Prozesse* (BPMN-basiert). Hierfür wurden sämtliche im Rahmen der Beantragung und Bewirtschaftung eines DriMi-Projektes relevanten Aktivitäten erfasst und in Prozessmodelle überführt. Aufgrund der rekursiven Vorgehensweise bei der Datenerhebung wurden die Prozessmodelle hierbei kontinuierlich evaluiert. Als Ergebnis dieses zweiten Schrittes entstanden pro Fallstudie drei separate Prozesse, welche die Abläufe der in Abschnitt 2.1 dargelegten Fälle von DriMi-Projekten je Fallstudie widerspiegeln. Im dritten Schritt der *Referenzmodellerstellung* wurden diese Prozessmodelle (3c) *konsolidiert* – d. h. sie wurden zu einem integrierten Modell zusammengeführt (zum Vorgehen vgl. Abschnitt 3.3).

3.2 Fallstudiendesign

Das Fallstudiendesign lehnt sich an Eisenhardt [Ei89], Myers [My09] und Yin [Yi09] an. Aufgrund des o. g. Modellierungsziels ist die Fallstudie explanativ ausgerichtet [Yi09]. Insgesamt wurden vier Fallstudien erhoben und anschließend zu einem RPM

konsolidiert. Es liegt somit ein *holistic multiple-case Design* vor [Yi09]. Die Fallstudienauswahl erfolgte auf Basis des *Most Different Systems Designs (MDSD)* [Li75, PT70]. Beim MDSD werden in ihren Eigenschaften sehr unterschiedliche Untersuchungsobjekte ausgewählt [Mo08]. Es wird davon ausgegangen, dass bei vorhandenen Gemeinsamkeiten hinsichtlich der Erkenntnisse aus den sehr unterschiedliche Fällen eine Übertragbarkeit auf weitere Fälle möglich ist [Ja13, Kl09, Mo08]. Der Ansatz kann somit induktiv eingesetzt werden und verspricht selbst bei einer relativ geringen Anzahl an Fallstudien eine gute Eignung [Kl09, Pi06]. Um diese möglichst verschiedenen Fallstudien erheben zu können, wurde mehrere Vergleichsvariablen herangezogen (z. B. Mitarbeiter/Studierendenzahl, Organisationsform, DriMi-Volumen). Die Fallstudien beziehen sich auf HS mit einer Mitarbeiterzahl zwischen 2.000 und 13.000 sowie einer Studierendenzahl zwischen 12.000 und 31.000. Die Organisationsform der HS reicht von sehr zentral bis zu sehr dezentral – d. h. einerseits werden die administrativen Tätigkeiten in DriMi-Projekten fast vollständig in der Zentralverwaltung der HS erledigt, andererseits übernehmen diese Aufgaben eher die Institute und die zentrale Verwaltung agiert prüfend. Das DriMi-Volumen erstreckt sich von 30 Mio. € bis 230 Mio. €. Im Rahmen der Fallstudien wurden primäre und sekundäre Daten verwendet [My09]. Die primären Daten wurden in semi-strukturierten Interviews erhoben [My09]. Sekundäre Datenquellen waren bereitgestellte Informationsblätter und Dokumentationen. Als Interviewpartner dienten Mitarbeiter verschiedener HS-Einrichtungen (z. B. Finanzabteilung). Die einzelnen Fallstudien wurden zyklisch angefertigt: Nach einem Erstinterview (ca. 120 Min.) wurden sukzessiv mehrere Detailinterviews (ca. 60 Min.) durchgeführt. Anschließend wurde ein Erstinterview in der nächsten HS durchgeführt, woran sich wiederum Detailinterviews schlossen. Danach wurden die Erkenntnisse der zweiten auf die erste Fallstudie zurückgespiegelt und erneut Detailinterviews geführt. Gleichermaßen wurde mit Fallstudie drei und vier verfahren. Ziel dieses Vorgehens war es, die erhobenen Prozessmodelle jeweils weiter zu verfeinern bzw. zu evaluieren. Wie bereits dargelegt, entstanden hieraus für jede der vier Fallstudien drei HS-spezifische Prozessmodelle (Standard-, Einzelprojekt- und Verbundprojektfall, siehe Abschnitt 2.1).

3.3 Vorgehen bei der Konsolidierung der Prozessmodelle

Der Grunddanke des Vorgehens bei der Konsolidierung richtet sich erneut nach dem MDSD [Li75, PT70] und dem in Abschnitt 2.2 definierten Verständnis eines RPM. Im Rahmen der Konsolidierung wurden die in den Fallstudien erhobenen Prozesse auf Aktivitätsebene zusammengeführt. Dies geschah in Anlehnung an ein Vorgehen zur Taxonomie-Entwicklung [NVM13]:

Zunächst wurde die erste Fallstudie ausgewählt und sukzessiv jede einzelne Aktivität in der Taxonomie ergänzt. Danach wurde die zweite Fallstudie selektiert und ebenfalls alle Aktivitäten sukzessiv durchgearbeitet. Hierbei wurde jedoch jeweils geprüft, ob die selektierte Aktivität bereits in der Taxonomie enthalten ist. Falls ja, wurde dies in der Taxonomie bei der vorhandenen Aktivität vermerkt und damit eine Zuordnung der Aktivität vorgenommen. Falls die selektierte Aktivität nicht in der Taxonomie enthalten war, wurde die Taxonomie um ein Element erweitert ohne eine Zuordnung vorzunehmen. Gleichermaßen wurde mit der dritten und vierten Fallstudie verfahren. Bei

diesem Vorgehen wurden die Tätigkeiten jeweils inhaltlich geprüft und analysiert, um dadurch unter den HS nicht einheitlich verwendete Begriffe für gleiche Tätigkeiten zu vereinheitlichen. Letztlich entstand eine vollständige Liste aller Aktivitäten aus der erkennbar ist, welche Aktivitäten in mehreren bzw. allen Fallstudien vorkommen. Nach dem MDSD sind genau diese Aktivitäten auf weitere Fälle übertragbar und somit in das RPM aufzunehmen. Alle entfallenden Aktivitäten gelten als HS-spezifisch und sind bei der Überführung des RPM in ein HS-spezifisches Modell zu ergänzen [Be02b].

4 Ergebnisse der Referenzmodellierung

Insgesamt konnte aus den Prozessen eine Taxonomie mit 178 Aktivitäten erzeugt werden, von welchen 111 Aktivitäten in das RPM übertragen wurden (basierend auf dem MDSD, vgl. Abschnitt 3.3). Die Auswertung der Daten aus den Fallstudien bzw. der Taxonomie ergab eine dreistufige Hierarchie. Auf oberster Ebene (Ebene 1) wird nach den Phasen eines DriMi-Projektes gegliedert. Innerhalb dieser Phasen müssen verschiedene Teilprozesse absolviert werden (Ebene 2), welche wiederum verschiedene Aktivitäten enthalten (Ebene 3). In den folgenden Abschnitten werden die einzelnen RPM separat für die Phasen eines DriMi-Projektes beschrieben (Ebene 1). Dabei wird jeweils zwischen den drei Fällen eines DriMi-Projektes unterschieden (vgl. Abschnitt 2.1). Darüber hinaus sind folgende Punkte anzumerken: (1) Um zu verdeutlichen, wie viele der betrachteten HS eine bestimmte Aktivität ausführen, wurde eine dreistufige Farbskala gewählt (dunkelgrau=alle HS, hellgrau=mind. drei HS, schwarz=weniger als drei HS). Diese Skala gilt abhängig vom jeweiligen Fall eines DriMi-Projektes. (2) Der Rahmen der Teilprozesse bzw. Aktivitäten gibt an, ob dieser/diese obligatorisch oder optional ist. Dabei wird ein obligatorisches Element immer ausgeführt, wenn der hierarchisch übergeordnete Teilprozess ausgeführt wird, wohingegen ein optionales Element übersprungen werden kann. (3) Auf Phasenebene (Ebene 1) wird davon ausgegangen, dass alle Phasen sukzessiv absolviert werden. Daher sind die Antrags-, Durchführungs- und Abschlussphase als obligatorisch markiert. Die Skizzenphase wird als optional angesehen, da nicht alle DriMi-Projekte eine solche Phase besitzen.

4.1 Skizzenphase

Die Abbildung 1 veranschaulicht das RPM der Skizzenphase. Diese besteht aus drei Teilprozessen. Die *Beantragung einer Anschubfinanzierung* für ein DriMi-Projekt ist hierbei optional, da die Wissenschaftler selbst entscheiden, ob Sie für eine Skizzenanfertigung zusätzliche Mittel benötigen. Ist dies gegeben, müssen diese einen „Antrag stellen“, welcher von einer zentralen Einrichtung „zu prüfen ist“ und anschließend wird „entschieden, ob Mittel genehmigt werden“. Davon abhängig wird die Aktivität „Mittelzuweisung“ (optional) ausgeführt. Außerdem kann bei der „Prüfung“ oder „Beurteilung“ auch entschieden werden, dass eine „Überarbeitung erforderlich“ ist (optional). In diesem Fall wird nach der Überarbeitung nochmals geprüft. Im nächsten Teilprozess wird die *Skizze angefertigt*. Hierbei ist eine „Beratung“ durch eine zentrale Einrichtung optional. Danach ist die *Skizze beim Geldgeber einzureichen*. Die zu erledigenden Tätigkeiten sind hierbei alle optional. Dies liegt darin begründet, dass während der „finanziel-

len Prüfung“ der Skizze ein Mangel festgestellt werden kann, der eine Einreichung verhindert und somit werden die anderen Aktivitäten nicht ausgeführt (unwahrscheinlicher Fall). Die Aktivität „finanzielle Prüfung“ muss jedoch ebenfalls als optional markiert werden, da diese nicht bei allen Verbundprojekten durchgeführt wird.

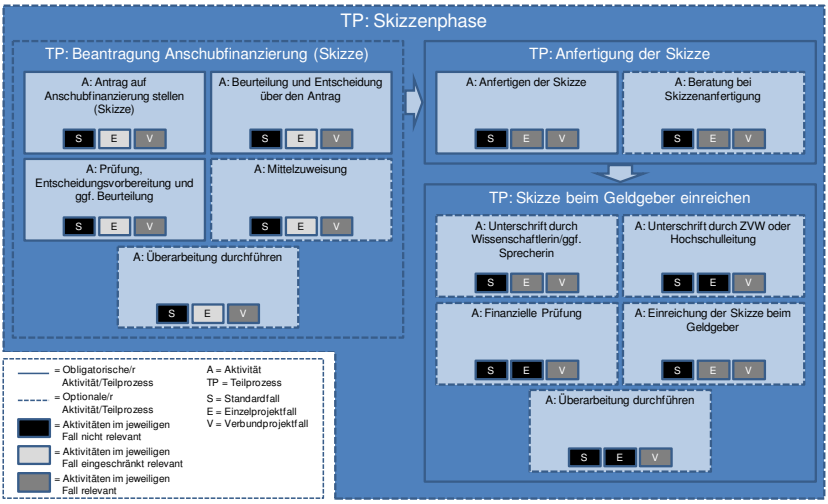


Abbildung 1: Referenzprozessmodell der Skizzenphase

4.2 Antragsphase

Die Antragsphase besteht aus sechs Teilprozessen (vgl. Abbildung 2), wobei sich der Teilprozess *Anfertigung DriMi-Anzeige* nicht eindeutig in diese Phase einordnen lässt, da ein Teil der HS angaben, diesen Teilprozess erst nach der Bewilligung auszuführen (Durchführungsphase). Die Erläuterung dieses Teilprozesses erfolgt hier jedoch in der Antragsphase – sie kann aber gleichermaßen auf die Durchführungsphase übertragen werden. Ansonsten verlaufen die Teilprozesse *Beantragung der Anschubfinanzierung* und *Antrag beim Geldgeber einreichen* fast identisch zu ihren Pendants aus der Skizzenphase. An einigen Stellen ändert sich ggf. das Routing, welches im RPM jedoch außer Acht gelassen wurde. Für die *Anfertigung des Antrags* gilt, dass wenn der Wissenschaftler während des „Anfertigens des Antrags“ die „Beratung“ durch eine zentrale Einrichtung in Anspruch nimmt (optional), von der beratenden Einrichtung i. d. R. zur Verwaltung und Dokumentation eine „papierbasierte Projektakte angelegt“ (optional) wird. Die beiden Teilprozesse *Durchführung Begehung* und *Vereinbarung über Grundausstattung* sind nur im Verbundprojektfall relevant. Da sie hier jedoch nicht zwangsläufig vorkommen müssen, werden die Prozesse als optional markiert. Als Begehung wird eine persönliche bzw. vor Ort Begutachtung des DriMi-Projektes bezeichnet, welche bei koordinierten Projekten der DFG durchgeführt wird. Eine *Vereinbarung über Grundausstattung* ist notwendig, wenn ein bestimmtes Objekt (z. B. eine Maschine) für die Durchführung eines DriMi-Projektes benötigt wird – dieses jedoch nicht durch den Geldgeber finanziert werden soll. Die HS muss sich daher eine andere Finanzierungsquelle für das Objekt suchen.

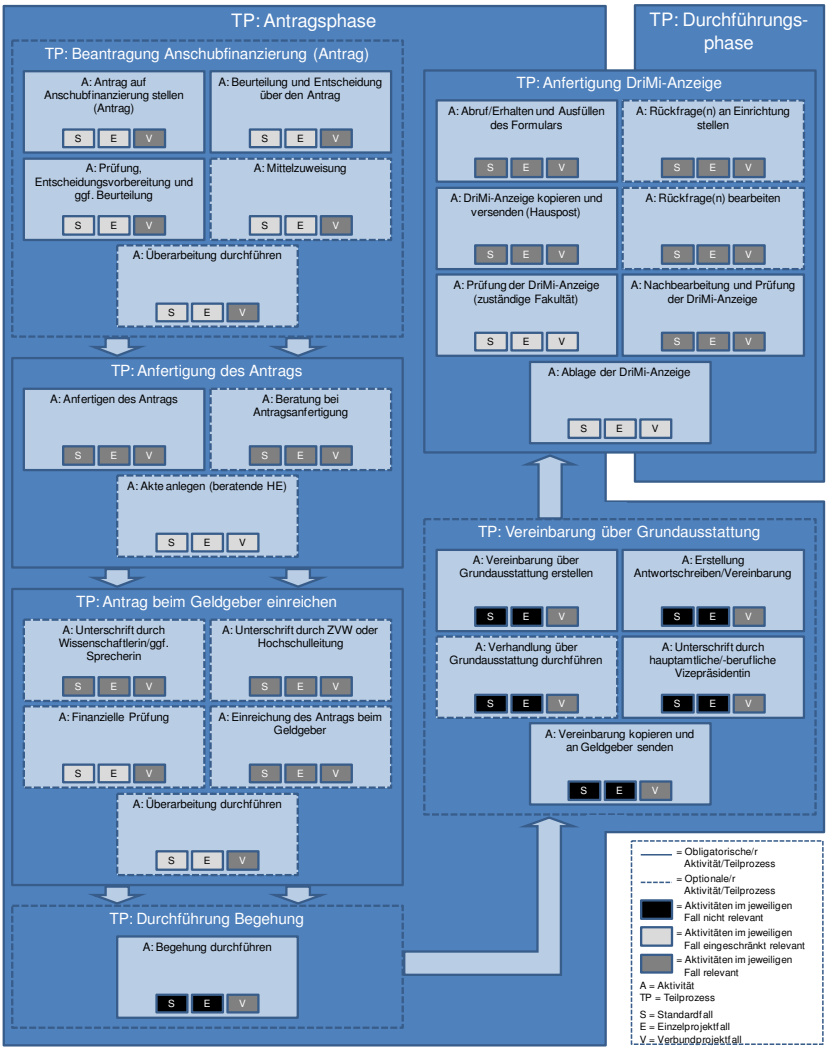


Abbildung 2: Referenzprozessmodell der Antragsphase

Muss eine *Vereinbarung über Grundausstattung* getroffen werden, wird von der HS und vom Geldgeber eine „Vereinbarung erstellt“ und anschließend entscheidet die HS, ob mit dem zuständigen Ministerium für Wissenschaft eine „Verhandlung durchgeführt“ werden soll (optional), in der über eine finanzielle Beteiligung an den Kosten diskutiert wird. Unabhängig von dieser Entscheidung muss die HS den Geldgeber darüber informieren, ob die Vereinbarung akzeptiert wird, was durch die „Erstellung eines Antwortschreibens“ mit anschließender „Unterschrift durch einen Vizepräsidenten“ und dem „Senden der Vereinbarung an den Geldgeber“ erfolgt. Nach § 25 des Hochschulrahmengesetztes müssen die DriMi-einwerbenden Einrichtungen das Einwerben der DriMi gegenüber der zentralen HS-Verwaltung anzeigen. Dies erfolgt im Teilprozess *Anfertigung DriMi-Anzeige*. Hierbei wird ein „Pdf-Formular auf einer

Webseite abgerufen und ausgefüllt“ und anschließend „per Hauspost versendet“. I. d. R. erfolgt dann zunächst eine „Prüfung in der Fakultät“, welche die DriMi-Anzeige dann zur „Nachbearbeitung und Prüfung“ an die zentrale HS-Verwaltung weiterleitet (papierbasiert). Die DriMi-Anzeige wird nach erfolgreicher Prüfung lokal „abgelegt“ und archiviert. Falls eine der Prüfungen (Fakultät, zentrale HS-Verwaltung) nicht erfolgreich verläuft, können „Rückfragen an die einwerbende Einrichtung gestellt werden“ (optional), welche diese „beantwortet“ (optional).

4.3 Durchführungsphase

Die Durchführungsphase besteht aus neun Teilprozessen, von denen vier Teilprozesse in Abbildung 3 als Platzhalter dargestellt sind. Zwei der vier Teilprozesse sind ausführlich in Abbildung 4 zu finden (*Anfertigen des ZVN*, *Prüfung durch Geldgeber (ZVN)*). Die anderen beiden Teilprozesse (*Personaleinstellung, Materialeinkauf*) werden im Rahmen dieses Artikels nicht betrachtet, da diese wenig DriMi-Spezifität aufweisen. Da der Teilprozess *Anfertigung DriMi-Anzeige* bereits in Abschnitt 4.2 erläutert wurde, wird mit der Beschreibung des Teilprozesses *Einrichtung DriMi-Konto* begonnen. Nach Bewilligung eines Projektes müssen zunächst verschiedene Buchungsobjekte im hierfür verwendeten Anwendungssystem (AS) eingerichtet werden. Das Einrichten beginnt mit dem „Anlegen einer papierbasierten DriMi-Akte“ in der zuständigen zentralen HS-Einrichtung. Anschließend werden „ein Innenauftrag/eine Kostenstelle“ und „ein Fond/ eine Finanzstelle“ im AS angelegt sowie die „Zuordnung der Buchungsobjekte vorgenommen“ (Innenauftrag zu Fonds). Ein Anschreiben für die Bewilligungsempfänger (papierbasiert) dient dazu, die HS-Einrichtung über die angelegten Buchungsobjekte zu informieren (Innenauftragsnummer). Dieses wird zunächst „erstellt“ und anschließend „kopiert, lokal archiviert und an die HS-Einrichtung weitergeleitet“. Diese „nimmt das Anschreiben zur Kenntnis und legt es lokal ab“. Der Teilprozess *Aushandeln der internen Kostenträger* ist lediglich bei Verbundprojekten relevant. Außerdem ist dieser nur notwendig, wenn es eine *Vereinbarung über Grundausstattung* gab (vgl. Abschnitt 4.2), weshalb er als optional gekennzeichnet ist. Wenn ein Geldgeber bestimmte Objekte (z. B. eine Maschine) als Grundausstattung betrachtet, dann muss eine andere Finanzierungsquelle gesucht werden. Innerhalb der HS werden daher alle ggf. „zu beteiligenden Organisationseinheiten kontaktiert“ (z. B. Präsidium, Fakultät) und anschließend wird über eine „Kostenbeteiligung verhandelt“. Der Teilprozess *Einholung eines Rechtsmittelverzichtes* (RMV) muss nur bei Mitteln von Bundesministerien oder der DFG absolviert werden – er ist daher ebenfalls als optional anzusehen. Im Ablauf dieses Teilprozesses wird zunächst von der zuständigen zentralen Einrichtung das „Formular (RMV) per Hauspost an den DriMi-empfangenden Wissenschaftler gesendet“. Dieser prüft und „erkennt die Bewilligungsbedingungen durch seine Unterschrift an“. Anschließend wird der „RMV kopiert und lokal abgelegt“ sowie zur zentralen Einrichtung zurückgesendet. Hier erfolgt eine „rechtsverbindliche Unterschrift“ und der „RMV wird zum Geldgeber gesendet“. Ab dieser Stelle kann nun offiziell mit dem DriMi-Projekt begonnen werden.

Wie die Abbildung 3 zeigt, können nun *Mittelabrufe*, *Personaleinstellungen* und *Materialeinkäufe* durchgeführt werden. Diese Teilprozesse sind mehrfach und parallel ausführbar (je nach Bewilligung). Da jedoch nicht zwangsläufig in DriMi-Projekten *Personal*

eingestellt oder *Material eingekauft* werden muss, sind diese beiden Prozesse als optional anzusehen.

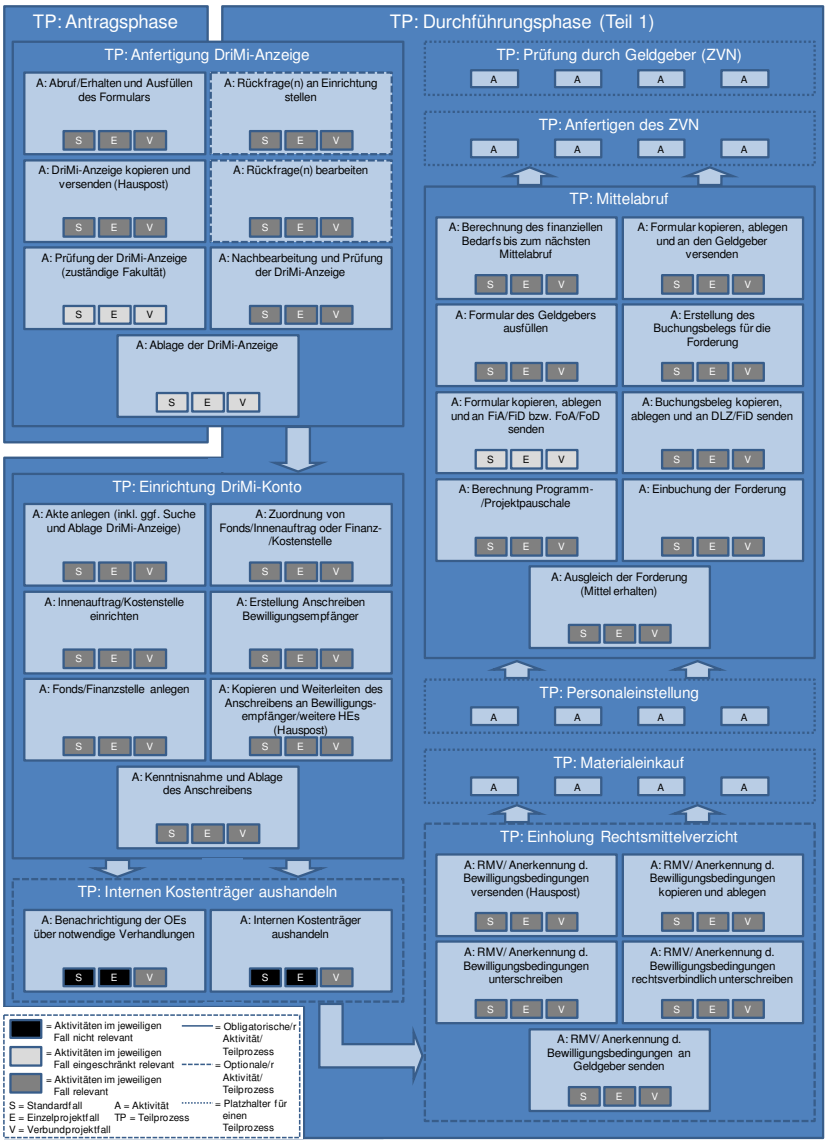


Abbildung 3: Referenzprozessmodell der Durchführungsphase (Teil 1)

Ein *Mittelabruf* verläuft folgendermaßen: Zunächst „berechnet (i. d. R. der Wissenschaftler) den finanziellen Bedarf an Mittel“ und „füllt auf Basis der errechneten Zahlen das zugehörige Formular des Geldgebers aus“. Dieses wird dann „kopiert und lokal archiviert und anschließend an die für DriMi-Projekte zuständige zentrale Einrichtung gesendet“. Die zentrale Einrichtung „berechnet nun die Programm-/ Projektpauschale“

und vervollständigt das Formular. Danach wird das Formular erneut „kopiert, lokal abgelegt und das Original an den Geldgeber gesendet“. Anschließend wird ein „Buchungsbeleg erstellt“, über welchen die für Buchungen zuständige Organisationseinheit aufgefordert wird, eine „offene Forderung (angeforderte Mittel) in das AS einzubuchen“. Dazu wird der Buchungsbeleg nach der Erstellung ebenfalls „kopiert, lokal abgelegt und weitergeleitet“. Nachdem die Zahlung des Geldgebers erfolgte, wird dann ein „Ausgleich der Forderung“ vorgenommen.

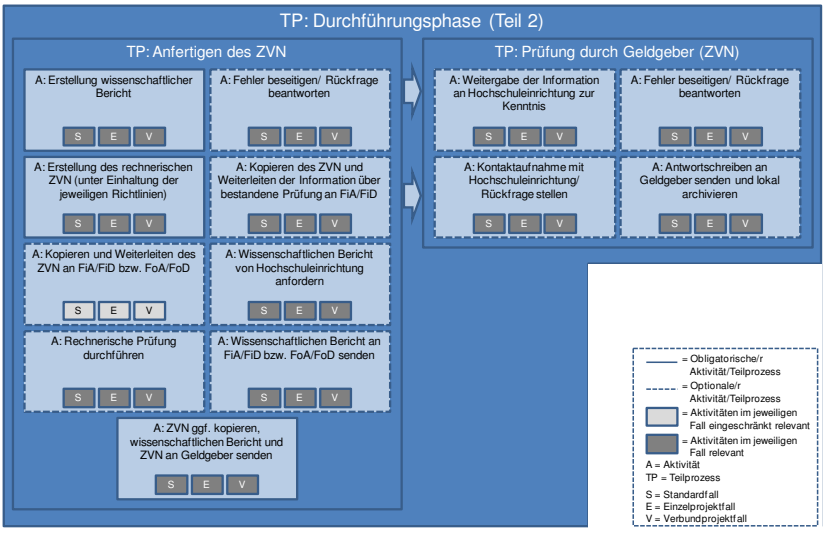


Abbildung 4: Referenzprozessmodell der Durchführungsphase (Teil 2)

Nach einer gewissen Zeit (abhängig vom Geldgeber) muss ein *Zwischenverwendungsnachweis (ZVN) erstellt* werden. (vgl. Abbildung 3 u. Abbildung 4). Hierzu werden ein „wissenschaftlicher Bericht angefertigt“ sowie der „rechnerische ZVN erstellt“. I. d. R. läuft der ZVN über eine prüfende Einrichtung (z. B. zentraler DriMi-Bereich). Dies ist jedoch nicht zwangsläufig gegeben, weshalb fast alle weiteren Aktivitäten dieses Teilprozesses als optional zu markieren sind. Lediglich das „Senden des ZVN und des wissenschaftlichen Berichtes an den Geldgeber“ ist noch obligatorisch. Falls jedoch der ZVN zentral zu prüfen ist (z. B. wenn dies der Geldgeber explizit fordert), wird der „ZVN kopiert, lokal abgelegt und zur zuständigen zentralen Einrichtung gesendet“ (postalisch). Dort folgt die „rechnerische Prüfung“ des ZVN, wobei ggf. Fehler entdeckt werden, die anschließend „zu beseitigen sind“. Falls die „rechnerische Prüfung“ durch die interne Revision durchgeführt wird, „kopiert diese wiederum den ZVN für die eigenen Unterlagen und leitet die Information, dass die Prüfung erfolgreich verlaufen ist, an den zentralen DriMi-Bereich weiter“. Da der geprüfte ZVN vom zentralen DriMi-Bereich zum Geldgeber gesendet wird, „fordert der dort tätige Mitarbeiter den wissenschaftlichen Bericht an“. Nachdem dieser an den zentralen DriMi-Bereich „gesendet“ wurde, werden beide „Dokumente zum Geldgeber geschickt“. Im letzten Teilprozess der Durchführungsphase wird nun der *ZVN durch den Geldgeber geprüft*. In diesem Teilprozess sind nur die Aktivitäten der HS berücksichtigt, da diese im Fokus dieses Beitrags stehen (vgl. Abbildung 4). Aus diesem Grund sind auch sämtliche Aktivitäten des Teil-

prozesses als optional zu markieren, denn diese müssen nur ausgeführt werden, wenn der Geldgeber einen Fehler im ZVN entdeckt. In diesem Fall wird die HS benachrichtigt, dass entweder (1) eine Verrechnung mit dem nachfolgenden ZVN möglich ist oder (2) der ZVN korrigiert werden muss. Im ersten Fall wird der „Wissenschaftler vom zentralen DriMi-Bereich über die notwendige Verrechnung informiert“ und anschließend ist der Teilprozess abgeschlossen. Im zweiten Fall erfolgt eine „Kontaktaufnahme“ zwischen zentralem DriMi-Bereich und Wissenschaftler, der „Fehler wird gemeinsam beseitigt“ und ein „Antwortschreiben (inkl. korrigiertem ZVN) zum Geldgeber gesendet“.

4.4 Abschlussphase

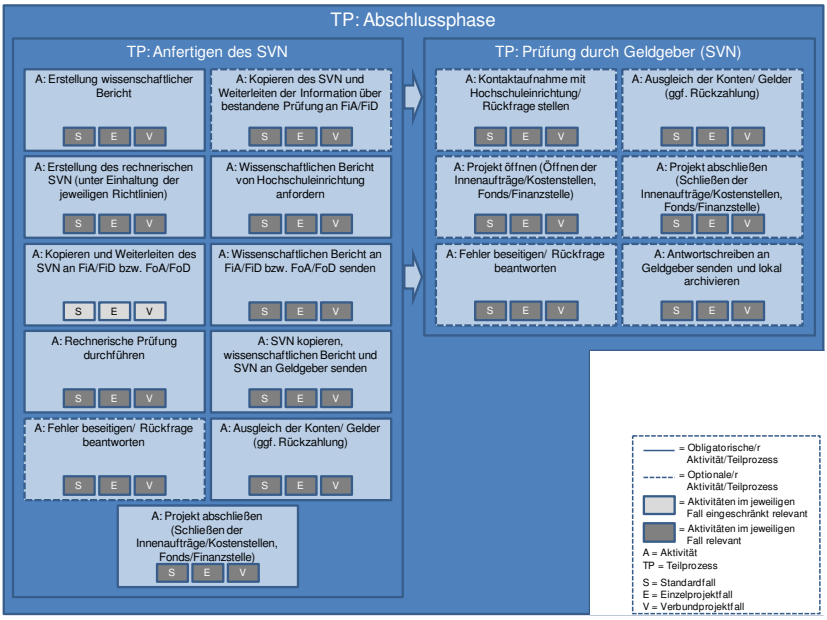


Abbildung 5: Referenzprozessmodell der Abschlussphase

Die Abschlussphase eines DriMi-Projektes besteht aus zwei Teilprozessen (vgl. Abbildung 5). Beide Teilprozesse ähneln ihren Pendants aus der Durchführungsphase sehr deutlich, weshalb an dieser Stelle lediglich auf die Unterschiede eingegangen wird. Zunächst ist festzuhalten, dass beim *Anfertigen des Schlussverwendungsnachweises (SVN)* immer eine zentrale Einrichtung beteiligt wird, weshalb im Vergleich zum Teilprozess *Anfertigen des ZVN* vier Aktivitäten obligatorisch werden. Des Weiteren kommen zwei neue Aktivitäten hinzu: Bevor der SVN zum Geldgeber gesendet wird, müssen „Kontendifferenzen ausgeglichen“ werden – d. h. wenn bspw. zu viele Mittel abgerufen wurden, werden diese an den Geldgeber erstattet. Danach wird das „Projekt offiziell abgeschlossen“ und damit die angelegten Buchungsobjekte (z. B. Innenauftrag, Fonds) im AS abgegrenzt. Diese Aktivitäten sind auch im Teilprozess *Prüfung durch den Geldgeber (SVN)* neu enthalten. Falls ein Geldgeber bei der Prüfung des SVN einen Fehler entdeckt, muss im Vergleich zum Pendant aus der Durchführungsphase zusätzlich

das „Projekt im AS wieder geöffnet“ werden, so dass die Buchungsobjekte bearbeitet werden können. Nachdem der „Fehler beseitigt“ ist, werden die „Konten entsprechend ausgeglichen“ und das „Projekt wieder geschlossen“. Unabhängig davon entfällt in diesem Teilprozess die Aktivität „Weitergabe der Information an den Wissenschaftler“ (vgl. Abbildung 4), da im Fall eines SVN keine Verrechnung mehr möglich ist.

5 Diskussion

Zur Interpretation des RPM werden die drei Hierarchiestufen herangezogen, nach denen das RPM aufgebaut wurde (Phasen-, Teilprozess-, Aktivitätsebene; vgl. Abschnitt 4). Für jede Stufe wird (1) ein Vergleich nach den verschiedenen Fällen eines DriMi-Projektes (Standard-, Einzelprojekt-, Verbundprojektfall; vgl. Abschnitt 2.1) und (2) ein Vergleich bzgl. der verschiedenen Fallstudien in den HS vorgenommen: Der Vergleich der verschiedenen Fälle eines DriMi-Projektes auf Phasenebene zeigt kaum Unterschiede auf – lediglich die Skizzenphase (inkl. enthaltener Teilprozesse und Aktivitäten) kommt im Einzel- und Verbundprojektfall im Vergleich zum Standardfall hinzu. Dies ist jedoch bereits durch die Definition der Fälle vorgegeben (vgl. Abschnitt 2.1). Ein Vergleich der verschiedenen Fallstudien auf Phasenebene zeigt keine Unterschiede auf – auf dieser Ebene laufen somit die Projekte in allen HS gleich ab. Wenn die verschiedenen Fälle eines DriMi-Projektes auf Teilprozessebene verglichen werden, ergeben sich einige Unterschiede. Zunächst ist festzuhalten, dass ein Unterschied bzgl. der Teilprozesse der Skizzenphase zu finden ist. Wenn ein DriMi-Projekt keine Skizzenphase besitzt, werden selbstverständlich auch die zugehörenden Teilprozesse nicht ausgeführt. Weitere Unterschiede beziehen sich nur auf Verbundprojekte – der Standard- und der Einzelprojektfall verlaufen nach der Skizzenphase auf allen Ebenen identisch. Im Verbundprojektfall werden zusätzlich die Teilprozesse *Durchführung Begehung*, *Vereinbarung über Grundausstattung*, *Internen Kostenträger aushandeln* sowie z. T. der Teilprozess *Beantragung Anschubfinanzierung* relevant. Der Vergleich der Fallstudien zeigt auf Teilprozessebene keine Unterschiede. Auf Aktivitätsebene ändert sich die Situation. So wird aus den RPM deutlich, dass die farbliche Skalierung für die verschiedenen Fälle eines DriMi-Projektes immer einheitlicher wird. Dies bedeutet, dass eine Annäherung der verschiedenen Prozessverläufe in den späteren Phasen (Durchführungs-, Abschlussphase) stattfindet und somit die zu erledigenden Tätigkeiten im Verlauf eines DriMi-Projektes unabhängig vom vorliegenden Fall zunehmend übereinstimmen. Eine einheitliche Unterstützung über alle Fälle eines DriMi-Projektes ist somit eher in den späteren Phasen denkbar. Ebenfalls wird aus den RPM deutlich, dass die Aktivitäten beim Vergleich der vier Fallstudien untereinander im Projektverlauf auch immer ähnlicher werden (erkennbar an den zunehmenden dunkelgrauen Markierungen). Dieser Aspekt wird durch die erstellte Taxonomie weiter gestützt: Aus den RPM wurden Aktivitäten ausgeschlossen, die nicht mindestens in drei HS zu finden waren (vgl. Abschnitt 4). Eine deskriptive und phasenbezogene Auswertung der Taxonomie bzgl. der Ausschlüsse von Aktivitäten ergibt, dass in der Skizzen- und der Antragsphase jeweils ca. 53 % der Aktivitäten nicht in das RPM übernommen wurden, wohingegen dies in der Durchführungsphase nur 30 % und in der Abschlussphase sogar nur ca. 19 % der Aktivitäten waren. D. h. wiederum, dass die späteren Phasen eines DriMi-Projektes in den verschiedenen HS sehr ähnlich ablaufen und somit deutlich weniger HS-Spezifika

vorhanden sind. Für das RPM bedeutet diese Tatsache, dass die Aussagekraft bzw. die Genauigkeit des RPM für die späteren Phasen höher ist. Außerdem lässt sich eine bessere Standardisierbarkeit der Prozesse in den späteren Phasen ableiten. Nichtsdestotrotz sind auch in den früheren Phasen eines DriMi-Projektes die Kernaktivitäten bzw. -abläufe in den HS ähnlich – es werden nur am Rand des Prozesses einzelne HS-spezifische Aktivitäten ergänzt.

Unabhängig davon heben die RPM Wiederholungsmuster in Teilprozessen hervor. So ist deutlich zu erkennen, dass die zu erledigenden Tätigkeiten des Teilprozesses *Beantragung Anschubfinanzierung (Skizze)* mit denen des Teilprozesses *Beantragung Anschubfinanzierung (Antrag)* übereinstimmen. Ein ähnliches Bild zeigt sich hinsichtlich der *Anfertigung des Antrags bzw. der Skizze*, der *Einreichung beim Geldgeber* sowie der *Anfertigung* und *Prüfung des ZVN bzw. SVN.* Wenn folglich einer dieser Teilprozesse verbessert bzw. unterstützt wird, können hiervon ggf. weitere Teilprozesse profitieren. Besonders wirkungsvoll ist selbstverständlich auch die Unterstützung von Prozessen bzw. Aktivitäten mit iterativem Charakter (z. B. *Mittelabruf*).

Letztlich stellt sich dann natürlich die Frage, was überhaupt unterstützt bzw. verbessert werden kann. Die RPM zeigen diverse Medienbrüche, bei denen Dokumente elektronisch erzeugt, ausgedruckt, ausgefüllt, versendet und anschließend wieder elektronisch erfasst werden müssen, auf (z. B. *Anfertigung DriMi-Anzeige*). Außerdem sind Teilprozesse zu finden, bei denen Prozessschritte auf mehrere Beteiligte verteilt sind, obwohl das Erledigen der Schritte von einzelnen Personen wesentlich sinnvoller wäre, da so Koordinationsaufwand entfällt (z. B. *Einrichten DriMi-Konto*). Darüber hinaus liegen viele administrative und Dokument-intensive Prozesse vor, in denen sowohl Papier- als auch digitale Dokumente verwendet und jeweils lokal abgelegt werden. Der Prozess wäre an vielen Stellen wesentlich schlanker, wenn eine einheitliche Datenbasis vorliegen würde, auf die alle Projektbeteiligten zugreifen könnten. Dies sind nur erste Ansatzpunkte für Verbesserungspotenziale. Eine systematische Potenzialanalyse anhand der RPM steht in zukünftiger Forschung noch aus.

6 Fazit

In diesem Beitrag wurde ein RPM für die Abwicklung von DriMi-Projekten in HS auf Basis von Fallstudien erstellt. Um die in Abschnitt 1 dargelegten Probleme bei der Abwicklung von DriMi-Projekten zu bewältigen (vgl. hierzu auch [GS14b]), sind die dargelegten Prozesse in zukünftiger Forschung hinsichtlich organisatorischer Schwachstellen systematisch zu untersuchen sowie hinsichtlich IT-Unterstützungspotenziale zu prüfen. Ein Gesamtvergleich der drei verschiedenen Fälle eines DriMi-Projektes (Standard-, Einzelprojekt-, Verbundprojektfall; vgl. Abschnitt 2.1) zeigt, dass in den frühen Phasen eines DriMi-Projektes (Skizze, Antrag) mehr Unterschiede zwischen den einzelnen Fällen vorhanden sind als in den späteren Phasen (Durchführung, Abschluss). Dies bedeutet, dass sich die zu erledigenden Tätigkeiten in den verschiedenen Fällen eines DriMi-Projektes immer weiter annähern. Bei der Erstellung des RPM zeigte sich außerdem, dass auf Ebene der Teilprozesse kaum Unterschiede zwischen den Fallstudien vorhanden sind (lediglich geänderte Reihenfolgen) und daher konnten alle identifizierten

Teilprozesse in das RPM aufgenommen werden. Bzgl. der Aktivitäten veränderte sich die Situation. Hier konnten nur ca. 62 % der erhobenen Aktivitäten in das RPM übertragen werden. Dies zeigt einerseits, dass gewisse HS-spezifische Aktivitäten und Abläufe vorhanden sind – andererseits ist jedoch auch ein Großteil der Abläufe identisch. Insofern schafft das entwickelte RPM eine gute Ausgangsbasis für die obengenannte zukünftige Forschung. Unabhängig vom Inhalt leistet der vorliegende Artikel außerdem einen methodischen Beitrag. Im Vorfeld dieser Untersuchung wurde eine umfangreiche Literaturrecherche betrieben, um ein Vorgehensmodell für die Erstellung eines fallstudienbasierten RPM zu identifizieren. Ein Modell, welches die Zusammenführung mehrerer Prozesse in einem RPM beschreibt, wurde dabei jedoch nicht gefunden. Das in diesem Artikel erarbeitete und verwendete Vorgehen schließt diese Lücke und ist somit für eine breite Gruppe von Wissenschaftlern von Interesse.

Literaturverzeichnis

[BAF09] Becker, J.; Algermissen, L.; Falk, T.: Prozessorientierte Verwaltungsmodernisierung - Prozessmanagement im Zeitalter von E-Government und New Public Management, Berlin, 2009.

[Be02a] Becker, J. et al.: Konfigurative Referenzmodellierung. In (Becker, J.; Knackstedt, R. Hrsg.): Wissensmanagement mit Referenzmodellen. Konzepte für die Anwendungssystem- und Organisationsgestaltung. Heidelberg, 2002; S. 25–144.

[Be02b] Becker, J. et al.: Nutzung von Referenzmodellen für die Einführung von Workflowmanagment am Beispiel der Produktionsplanung und -steuerung. In (Becker, J.; Knackstedt, R. Hrsg.): Wissensmanagement mit Referenzmodellen. Heidelberg, 2002; S. 239–255.

[BKR12] Becker, J.; Kugeler, M.; Rosemann, M. Hrsg.: Prozessmanagement: Ein Leitfaden zur prozessorientierten Organisationsgestaltung. Berlin, 2012.

[Br80] Bretzke, W.-R.: Der Problembezug von Entscheidungsmodellen. Tübingen, 1980.

[CWS13] Czarnecki, C.; Winkelmann, A.; Spiliopoulou, M.: Reference Process Flows for Telecommunication Companies. In Business & Information Systems Engineering, 2013, 5; S. 83–96.

[Ei89] Eisenhardt, K. M.: Building Theories from Case Study Research. In Academy of Management Review, 1989, 14; S. 532–550.

[EKB11] Ebeling, B.; Klages, M.; Breitner, M. H.: IT-gestütztes Management von Drittmitteln im Rahmen des Forschungsmanagements an deutschen Hochschulen. In (Heiß, U. et al. Hrsg.): Proceedings Informatik 2011, Bonn, 2011; S. 452–466.

[EKB13] Ebeling, B.; Köpp, C.; Breitner, M. H.: Diskussion eines Prototyps für das dezentrale Management von Forschungsressourcen an deutschen Hochschulinstituten: Wirtschaftsinformatik Proceedings 2013, 2013; S. 343–357.

[FL04] Fettke, P.; Loos, P.: Referenzmodellierungsforschung. In Wirtschaftsinformatik, 2004, 46; S. 331–340.

[FL07] Fettke, P.; Loos, P.: Perspectives on Reference Modeling. In (Fettke, P.; Loos, P. Hrsg.): Reference Modeling for Business Systems Analysis. Hershey, 2007; S. 1–21.

[Gi01] Giaglis, G. M.: A Taxonomy of Business Process Modeling and Information Systems Modeling Techniques. In International Journal of Flexible Manufacturing Systems, 2001, 13; S. 209–228.

[GS13] Gröger, S.; Schumann, M.: IT-Unterstützung zur Verbesserung der Dritt- und Sondermittelbewirtschaftung an Hochschulen - State of the Art, Göttingen, 2013.

[GS14a] Gröger, S.; Schumann, M.: Managing third-party funding projects at German state universities - A theoretical deduction of design principles for implementing an IT-artifact: ECIS Proceedings 2014, Tel Aviv, 2014; S. 1–15.
[GS14b] Gröger, S.; Schumann, M.: IT-gestützte Abwicklung von Drittmittelprojekten - Stand der Forschung und Praxis. In Verwaltung & Management, 2014, 20; S. 102–112.
[Ha09] Hamdan, M.: Drittmittelforschung in der Medizin: eine Gratwanderung zwischen strafbarer Vorteilsannahme und verfassungsgarantierter Forschungsfreiheit, Frankfurt am Main, 2009.
[Ja13] Jahn, D.: Einführung in die vergleichende Politikwissenschaft, Wiesbaden, 2013.
[Ka08] Kaiser, R.: Drittmittel, Sponsoring und Fundraising - rechtskonforme Finanzierung öffentlicher Aufgaben oder Einstieg in die Korruption?, Zürich, 2008.
[KE12] Küsters, U.; Erben-Russ, M.: Forschungsinformationssysteme bei Fraunhofer. In (Bittner, S.; Hornbostel, S.; Scholze, F. Hrsg.): Forschungsinformationssysteme in Deutschland, Berlin, 2012; S. 37–40.
[Kl09] Klug, H.: Erfolgsfaktoren bei der Umstellung von Informationssystemen an Hochschulen: Wirtschaftinformatik Proceedings 2009, Wien, 2009; S. 473–482.
[KTG97] Kettinger, W. J.; Teng, J. T. C.; Guha, S.: Business Process Change: A Study of Methodologies, Techniques, and Tools. In MIS Quarterly, 1997, 21; S. 55.
[Li75] Lijphart, A.: The Comparable-Cases Strategy in Comparative Research. In Comparative Political Studies, 1975, 8; S. 158–177.
[Mi10] Misera, S.: Drittmittelforschung - Chancen, Risiken und Praxisprobleme, Bonn, 2010.
[Mo08] Moodie, G.: From vocational to higher education. Berkshire, 2008.
[My09] Myers, M. D.: Qualitative research in business and management. London, 2009.
[NVM13] Nickerson, R. C.; Varshney, U.; Muntermann, J.: A method for taxonomy development and its application in information systems. In European Journal of Information Systems, 2013, 22; S. 336–359.
[Pi06] Pickel, G.: Methodisches Design der Studie. In (Pickel, G. et al. Hrsg.): Osteuropas Bevölkerung auf dem Weg in die Demokratie: Repräsentative Untersuchungen in Ostdeutschland und zehn osteuropäischen Transformationsstaaten. Wiesbaden, 2006; S. 19–30.
[PT70] Przeworski, A.; Teune, H.: The logic of comparative social inquiry, New York, 1970.
[Re99] Reiter, C.: Toolbasierte Referenzmodellierung - State of the Art und Entwicklungstrends. In (Becker, J.; Rosemann, M.; Schütte, R. Hrsg.): Referenzmodellierung. State of the art und Entwicklungsperspektiven. Heidelberg, 1999; S. 45–68.
[SL12] Schwegmann, A.; Laske, M.: Istmodellierung und Istanalyse. In (Becker, J.; Kugeler, M.; Rosemann, M. Hrsg.): Prozessmanagement: Ein Leitfaden zur prozessorientierten Organisationsgestaltung. Berlin, 2012; S. 165–194.
[St11] Statistisches Bundesamt: Statistisches Jahrbuch 2011. https://www.destatis.de/DE/Publikationen/StatistischesJahrbuch/StatistischesJahrbuchKomplett.pdf?__blob=publicationFile, 2012-04-14.
[St13] Statistisches Bundesamt: Bildung und Kultur - Finanzen der Hochschulen (Fachserie 11 Reihe 4.5). bis 2011. https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/BildungKulturFinanzen/FinanzenHochschulen2110450117004.pdf?__blob=publicationFile, 2013-06-21.
[Th06] Thomas, O.: Understanding the Term Reference Model in Information Systems Research: History, Literature Analysis and Explanation. In (Bussler, C. J.; Haller, A. Hrsg.): Business Process Management Workshops, Berlin, 2006; S. 484–496.
[Un08] Universität Bielefeld: Ergebnispräsentation Sollkonzeption. https://www.uni-bielefeld.de/Universitaet/Ueberblick/Organisation/Verwaltung/Dez_F/Finanzprozesse/Uni_Bielefeld_Ergebnisprasentation_Sollkonzeption_080630.pdf, 2014-06-01.
[Yi09] Yin, R. K.: Case study research - Design and methods. Los Angeles, Calif, 2009.

Untersuchungen zur Beherrschbarkeit von Variabilität in Prüfungsprozessmodellen deutscher Hochschulen

Oliver Arnold, Ralf Laue

Fakultät Physikalische Technik/Informatik
Westsächsische Hochschule Zwickau
Dr.-Friedrichs-Ring 2A
08056 Zwickau
oliver.arnold@fh-zwickau.de
ralf.laue@fh-zwickau.de

Abstract: Prozesse im öffentlichen Sektor weisen häufig große Gemeinsamkeiten, aber auch bedeutende Unterschiede auf Grund örtlicher Gegebenheiten und Vorschriften auf. Ein Beispiel hierfür sind die von uns untersuchten Prüfungsprozesse an deutschen Hochschulen. Die geeignete Darstellung dieser Variabilität in entsprechenden Modellen bietet die Basis für die effektive Unterstützung der Prozesse mit Mitteln der Informationstechnik. Der vorliegende Beitrag stellt zwei bekannte Ansätze zur Modellierung von Prozessmodellvariabilität vor, wendet diese exemplarisch auf die in sechs verschiedenen Studiengängen erhobenen Prüfungsprozessmodelle an und geht dabei auf Vor- und Nachteile dieser Ansätze ein.

1 Einleitung

Ein Geschäftsprozessmodell (GPM) beschreibt eine Reihe von Aktivitäten, die eine Organisation zum Erreichen eines bestimmten Zieles ausführen muss. Im öffentlichen Sektor könnte ein GPM z.B. die Schritte beschreiben, die für das Ausstellen einer bestimmten Genehmigung oder die Durchführung einer öffentlichen Ausschreibung nötig sind. Dazu definiert das GPM eine Reihe von Aktivitäten oder Funktionen der beteiligten Menschen oder IT-Systeme. Ein wesentlicher Bestandteil des GPM ist die Modellierung des Kontrollflusses, d.h. der zeitlich-logischen Abhängigkeiten zwischen den Aktivitäten. Auch die nötigen Ressourcen sowie die verarbeiteten Informationen sind meist Teil des GPM.

In der Praxis ist nun häufig zu beobachten, dass der gleiche grundlegende Geschäftsprozess je nach vorhandenem Umfeld (dem sogenannten Prozesskontext) unterschiedlich abläuft. In [Vo12] beispielsweise wird beschrieben, wie in zehn niederländischen Gemeinden die gleichen grundlegenden durch Gesetze vorgegebenen Verwaltungsprozesse auf immer wieder unterschiedliche Art und Weise umgesetzt werden.

Der richtige Umgang mit dieser GPM-Variabilität (GPMV) war in den letzten Jahren Gegenstand zahlreicher wissenschaftlicher Untersuchungen [Va13]. Wird keine beson-

ders auf die Behandlung von Variabilität spezialisierte Modellierungsmethode eingesetzt, bieten sich zwei Wege an, um Prozessvarianten zu modellieren. Diese werden nachfolgend beschrieben und bewertet.

Eine Möglichkeit stellt die Modellierung der verschiedenen Prozessvarianten in einem gemeinsamen GPM dar. Dies wird in der Literatur zur GPMV gemeinhin als Single-Model-Ansatz bezeichnet [KY12][HBR10]. Dabei entsteht in der Regel ein großes und komplexes Modell mit zahlreichen Verzweigungen, die eigentlich der Unterscheidung zwischen den verschiedenen Prozesskontexten dienen. Beispielsweise könnte in einer Behörde ein bestimmter Prozess mit einem Online-Formular unterstützt werden, in einer anderen Behörde hingegen muss der entsprechende Antrag persönlich in der Behörde abgegeben werden. An einem bestimmten Punkt im Prozessablauf wird dann eine Entscheidung zwischen diesen beiden Vorgehensweisen modelliert. Die Entscheidung wird in derselben Einrichtung (zumindest solange der Prozesskontext stabil bleibt) immer wieder gleich ausfallen. Dennoch müsste im Modell die Entscheidung bei jedem Prozessdurchlauf immer wieder getroffen werden. Solche eigentlich der Unterscheidung zwischen verschiedenen Prozessvarianten dienenden Verzweigungen werden als „Variationspunkte“ bezeichnet. Im Single-Model-Ansatz werden sie jedoch wie „normale“ (zur Laufzeit des Prozesses zu treffenden Entscheidungen) behandelt [HBR09].

Die zweite Möglichkeit wäre die Modellierung der verschiedenen Prozessvarianten in je einem eigenen separaten Modell. Dies wird in der einschlägigen Literatur als Multi-Model-Ansatz bezeichnet [KY12][HBR10]. Dabei wird das Prozessmodell kopiert und manuell angepasst. Da dennoch viele Teile des Prozessmodells in allen oder zumindest in vielen Varianten vorhanden sind, entsteht bei dieser Art der Prozessmodellierung viel Redundanz. Eine Kopplung der einzelnen Modelle ist häufig nur über die Namensgleichheit der Bezeichner gegeben. Oft erfolgt mit der Zeit eine Degeneration der Modelle, da Verbesserungen dann ggf. nur an einzelnen Varianten vorgenommen werden [HBR09].

Beide konventionelle Ansätze führen somit zu unbefriedigenden Ergebnissen. Im vorliegenden Artikel werden nun ebenfalls Prozesse aus dem öffentlichen Sektor für eine Untersuchung möglicher Ansätze zur Beherrschbarkeit von GPMV herangezogen. Dabei handelt es sich um die Prüfungsprozesse an verschiedenen sächsischen Hochschulen, wie Kapitel 2 weiter ausführt. Im darauffolgenden Kapitel 3 werden zwei gängige Ansätze zur Beherrschbarkeit von GPMV näher vorgestellt. Kapitel 4 beschreibt die erhobenen beispielhaften Prozessmodelle und das Vorgehen zur Erhebung der Modelle. Im Anschluss geht Kapitel 5 auf die in den Modellen festzustellende Varianz ein. Schließlich beschreibt Kapitel 6 exemplarisch die Anwendung der beiden in Kapitel 3 vorgestellten Ansätze auf die Beispielmodelle, und Kapitel 7 beschließt den Artikel mit einer Zusammenfassung und Ausblick.

2 Variabilität in Prüfungsprozessen

Die Durchführung von Prüfungen an sächsischen Hochschulen muss sich an den rechtlichen und prozessualen Rahmenbedingungen orientieren, die durch die entsprechenden Prüfungsordnungen der Hochschulen abgesteckt werden [AHS11].

Der durch die Prüfungsordnungen vorgegebene Prozessablauf soll durch Informations- und Workflowsysteme unterstützt werden. Dabei ist zu beachten, dass die Prüfungsordnungen der verschiedenen Hochschulen und Fakultäten u.a. aufgrund ihrer gemeinsamen Grundlage im Gesetz über die Freiheit der Hochschulen im Freistaat Sachsen. (Sächsisches Hochschulfreiheitsgesetz – SächsHSFG) zahlreiche Gemeinsamkeiten, aber auch beachtenswerte Unterschiede aufweisen. Damit entsprechen diese Prüfungsprozesse dem in Kapitel 1 beschriebenen Muster eines Basisprozesses mit zahlreichen Varianten. Diese unterscheiden sich in einigen Fällen nur durch verschiedene Ausprägungen von Geschäftsregeln bzw. Prozessvariablen (z.B. konkrete Studiendauer oder konkreter Zeitraum, nachdem eine erste Wiederholungsprüfung durchgeführt werden kann). Bei anderen Varianten kann es zu einem weitreichenden Umbau der Prozessstruktur kommen, indem ganze Teilprozesse hinzugefügt, entfernt oder innerhalb des Modells verschoben werden (siehe z.B. der in Kapitel 6 beschriebene um die Möglichkeit des Vorschlags einer Prüfergruppe erweiterte Prozess vor einer Einzelprüfungsleistung).

Die Modellierung dieser Prüfungsprozesse mit einem für GPMV geeigneten Ansatz würde es ermöglichen, ein GPM zu erschaffen, das dann kontextorientiert auf die einzelnen Hochschulen und Fachgebiete mit ihren spezifischen Prüfungsmodalitäten angepasst werden kann. Ferner ist es gerade im Bereich der Prüfungsprozesse notwendig, eine Rückverfolgbarkeit (engl. traceability) der einzelnen Prozessbausteine in die zugrundeliegenden Prüfungsordnungen zu formalisieren. Wird dabei berücksichtigt, dass Prüfungsordnungen wiederum auf einer gemeinsamen gesetzlichen Grundlage beruhen und dass die angesprochenen Gesetze und Ordnungen sich mit der Zeit ändern und somit in verschiedenen Versionen vorliegen, ergibt sich ein komplexes Netz bzgl. der Rückverfolgbarkeit. Werden darauf aufbauend nun formale Verknüpfungen zur Rückverfolgbarkeit etabliert, könnte auch eine zumindest semiautomatische Prüfung auf Einhaltung gesetzlicher Regelungen (engl. compliance) entwickelt werden.

Die Bereitstellung eines konfigurierbaren Referenzmodells hat den Vorteil, dass für die einzelnen Hochschulen der Aufwand bei der Erstellung von Prozessmodellen verringert wird. Wenn diese Prozessmodelle die Grundlage für Anforderungsdokumente bei der Entwicklung oder Anpassung von Software zur Unterstützung der modellierten Prozesse bilden, kann sich folglich der Prozess der Softwareeinführung verkürzen. Hinzu kommt, dass davon ausgegangen werden kann, dass geprüfte Referenzmodelle eine hohe Qualität aufweisen [Be13].

3 Ansätze zur Beherrschbarkeit von Geschäftsprozessmodellvariabilität

In [To13] werden zwei grundlegende Ansätze im Bereich der GPMV beschrieben. Dies ist zum einen der sogenannte Verhaltensansatz (engl. behavioural approach). Ausgangspunkt für die Variantenbildung ist hier ein konfigurierbares Gesamtmodell (z.B. mit Platzhaltern). Somit sind alle Mitglieder einer GPM-Familie mit allen Gemeinsamkeiten und Unterschieden (d.h. das Verhalten aller möglichen Varianten) in einem gemeinsamen Modellartefakt abgebildet.

Der zweite grundlegende Ansatz wird als Strukturansatz (engl. structural approach) bezeichnet. Dabei wird zunächst ein "Basismodell" modelliert, das zugleich auch selbst eine der möglichen Varianten darstellen kann. Zur Umsetzung der Variabilität werden dann Änderungsoperationen auf die Struktur dieses Basismodells angewendet. Dadurch werden die anderen Varianten aus dem Basismodell abgeleitet.

Zwei häufig zitierte Vertreter dieser Ansätze werden in den folgenden beiden Unterkapiteln näher vorgestellt.

3.1 C-EPC

In [RA07] wird der Modellierungsansatz der Configurable Event-Driven Process Chain (dt. Konfigurierbare Ereignisgesteuerte Prozesskette) beschrieben. Er erweitert die weitverbreitete Modellierungsmethode der EPK um konfigurierbare Elemente. Diese konfigurierbaren Elemente dienen der Umsetzung der Variationspunkte. Das so entstehende konfigurierbare Referenzmodell beinhaltet somit sowohl die Gemeinsamkeiten als auch die Unterschiede der GPM-Familie. Alle zugehörigen Informationen sind somit in einem einzigen Artefakt (dem konfigurierbaren Referenzmodell) enthalten. Damit ist die C-EPC ein Vertreter des sogenannten Verhaltensansatzes. Nach [To13] handelt es sich dabei um den bekanntesten und meistzitierten Vertreter dieser Ansatzfamilie zur Beherrschbarkeit von GPMV.

In C-EPCs sind sowohl Funktionen als auch logische Konnektoren konfigurierbar. Konfigurierbare Funktionen können entweder für die entstehende Variante eingeschlossen (im Modell dargestellt durch „ON"), immer übersprungen („OFF") oder nur unter bestimmten Bedingungen übersprungen („OPT") werden.

Konfigurierbare Konnektoren können auf einen konkreten Konnektortyp abgebildet werden. Dabei kann durch Konfiguration aber nur eine weitere Einschränkung des möglichen Verhaltens vorgenommen werden. Beispielsweise könnte ein konfigurierbarer XOR bzw. ODER-Konnektor derart konfiguriert werden, dass in der resultierenden Variante immer genau eine der nachfolgenden Sequenzen ausgeführt wird (d.h. die Auswahl wird bereits zur Konfigurationszeit getroffen). Andererseits kann ein AND-Konnektor nicht weiter eingeschränkt werden. Er verlangt immer, dass alle ausgehenden Kontrollflüsse ausgeführt werden.

Durch logische Ausdrücke können auch Abhängigkeiten zwischen den zur Verfügung stehenden Konfigurationsoptionen zum Ausdruck gebracht werden. Neben zwingenden Abhängigkeiten, die in den Modellen als „Requirement“ gekennzeichnet sind, können auch optionale Abhängigkeiten im Sinne von Empfehlungen modelliert werden, die wiederum in den Modellen als „Guideline“ bezeichnet werden.

3.2 Provop

In [HBR10] wird ein weiterer Ansatz zur Umsetzung von GPMV beschrieben. Der Ansatz trägt die Bezeichnung Provop, was für PROcess Variants by OPtions steht. Der Ansatz geht davon aus, dass jedes Prozessmodell aus einem anderen abgeleitet werden kann, indem eine Reihe von Änderungsoperationen auf ein zugrundeliegendes Basismodell angewendet werden. Damit ist Provop ein Vertreter des sogenannten Strukturansatzes, bei dem eine Prozessfamilie durch die schrittweise Modifizierung der Struktur einer bestimmten Prozessvariante entsteht.

Bei den grundlegenden Änderungsoperationen handelt es sich um das Einfügen (INSERT), Entfernen (DELETE) und Verschieben (MOVE) von Teilen eines GPMs sowie dem Verändern (MODIFY) von Prozesselementattributen. Die Operationen werden auf sogenannte Anpassungspunkte (engl. adjustment points) angewendet, die dem Eingang bzw. Ausgang eines Aktivitätsknotens bzw. eines logischen Konnektors entsprechen.

Die konkreten Operationen (bzgl. eines konkreten Prozesses) werden zwecks besserer und konsistenter Nutzung zu Optionen gruppiert, die immer zusammen angewendet werden. Außerdem können Beziehungen zwischen Änderungsoptionen festgelegt werden (beispielsweise der gegenseitige Ausschluss zweier Optionen oder dass zwei Optionen voneinander abhängen und somit nur gemeinsam auf ein Prozessmodell angewendet werden können).

Die Anwendung der Optionen, d.h. die eigentliche Variantenbildung, resultiert aus der Auswertung von Kontextvariablen in Kontextregeln. Eine solche Kontextvariable könnte die für die Variante gültige Prüfungsordnung sein, wie das Beispiel in Kapitel 6 zeigen wird. Darüber hinaus ist es auch möglich, Kontextvariablen an definierten Stellen dynamisch zur Laufzeit auszuwerten und so weitere Optionen anzuwenden. Zwischen den Basisprozess und den Varianten besteht eine feste Beziehung, so dass auch Änderungen am Basisprozess an die Varianten propagiert werden.

4 Beschreibung der Prozessmodelle und Vorgehen bei der Modellierung

Im Zuge unserer Untersuchungen wurden Prozessmodelle von sechs untersuchten Studiengängen erhoben. Wie Tabelle 1 zu entnehmen ist, wurde dabei von drei verschiedenen sächsischen Hochschulen jeweils der Studiengang BWL einbezogen, um zunächst eine bessere Vergleichbarkeit zu erreichen. Dazu kam im nächsten Schritt ein weiterer Studiengang, der sich inhaltlich deutlich von BWL unterscheidet. Im Sinne der Untersuchung

von Prozessvariabilität sollte durch diese Wahl eine möglichst hohe Varianz zwischen den Modellen erreicht werden.

Hochschule	**BWL**	**Anderer Studiengang**
Westsächsische Hochschule Zwickau	Betriebswirtschaft (B.A.)	Gebärdensprachdolmetschen (Diplom)
TU Bergakademie Freiberg	Betriebswirtschaftslehre (B.Sc.)	Gießereitechnik (B.Sc.)
Universität Leipzig	Betriebswirtschaftslehre (Management Science) (M.Sc.)	Lehramt Allgemein Staatsexamen (M.Sc.)

Tabelle 1: Übersicht der untersuchten Studiengänge

Die Prozessmodelle wurden mit dem frei verfügbaren Modellierungswerkzeug bflow* Toolbox (http://www.bflow.org/) in Form von ereignisgesteuerten Prozessketten modelliert. Ausgangspunkte war die Analyse der Prüfungsordnung des Betriebswirtschafts-Studiengangs an der Westsächsischen Hochschule Zwickau (WHZ). Dazu wurde zunächst die betreffende Prüfungsordnung Paragraph für Paragraph gesichtet und in einzelne Prozessmodellfragmente umgesetzt. Dabei wurde bei jeder einzelnen Funktion notiert, in welchen Paragraphen die Durchführung dieser Aktivität geregelt ist. Die Fragmente wurden dann in einem zweiten Arbeitsschritt zu den in Abbildung 1 ersichtlichen Teilmodellen kombiniert.

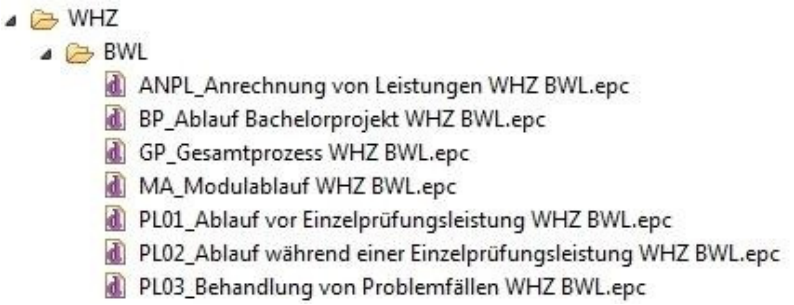


Abbildung 1: Teilprozessmodelle für Studiengang Betriebswirtschaft an der WHZ

Parallel zur Erstellung dieser Modelle wurde zunächst ein umfangreiches Textdokument erstellt, in dem die Unterschiede zwischen den Studiengängen herausgearbeitet wurden. Dieses Textdokument bildete dann die Grundlage für die Umsetzung der Unterschiede in sogenannte Varianzfragmente, also kleine Prozessmodellfragmente, die die Unterschiede im Prozessverlauf darstellen.

Das dann vorliegende Prozessmodell zum Betriebswirtschafts-Studiengang an der WHZ diente nun als Grundlage für die Erstellung der anderen Modelle und wurde somit zum

Basismodell, anhand dessen die Variabilität festgestellt wird. Dieses Basismodell wurde fünfmal kopiert für die anderen Studiengänge. In diese Kopien wurden dann die Varianzfragmente an passender Stelle eingearbeitet und farblich markiert. Das grundlegende Vorgehen zum Schaffen der Datengrundlage für die weiteren Untersuchungen entspricht somit den in Kapitel 1 erwähnten Multi-Model-Ansatz, bei dem für jede Variante jeweils ein neues vollständiges Modell erstellt wird.

5 Beschreibung der Varianz in den erhobenen Prozessmodellen

Abbildung 2 zeigt die Anzahl der festgestellten Unterschiede zwischen den Modellen der untersuchten Prüfungsprozesse. Außerdem erkennbar ist die Reihenfolge der Ableitung der einzelnen Prozessmodelle. Ausgangspunkt war jeweils das Modell des BWL-Studiengangs an der WHZ. Davon ausgehend wurden die Unterschiede zu den beiden BWL-Studiengängen an der TU Bergakademie Freiberg (TUBAF) sowie der Universität Leipzig analysiert. Die beiden anderen Studiengänge für Freiberg und Leipzig wurden dann wiederum aus diesen beiden Modellen abgeleitet. Lediglich das Modell für den Studiengang Gebärdensprachdolmetschen an der WHZ wurde direkt aus dem entsprechenden Modell für den BWL-Studiengang an der WHZ abgeleitet.

Deutlich erkennbar ist hierbei, dass die Unterschiede zwischen den verschiedenen Hochschulen zahlenmäßig größer ausfallen (im Durchschnitt 10,5 Varianzen) als zwischen verschiedenen Studiengängen der gleichen Hochschule (im Durchschnitt 6,33 Varianzen), auch wenn diese inhaltlich sehr unterschiedlich ausgerichtet sind. Dieses Ergebnis lässt sich schon aufgrund der begrenzten Menge an untersuchten Prozessen nicht verallgemeinern, entspricht aber doch der Erwartung, da Prüfungsprozesse in erster Linie durch die zentralen Verwaltungsvorgaben an den einzelnen Hochschulen geprägt sind. Dennoch ist auch ein deutlicher Einfluss der Besonderheiten der einzelnen Studiengänge erkennbar.

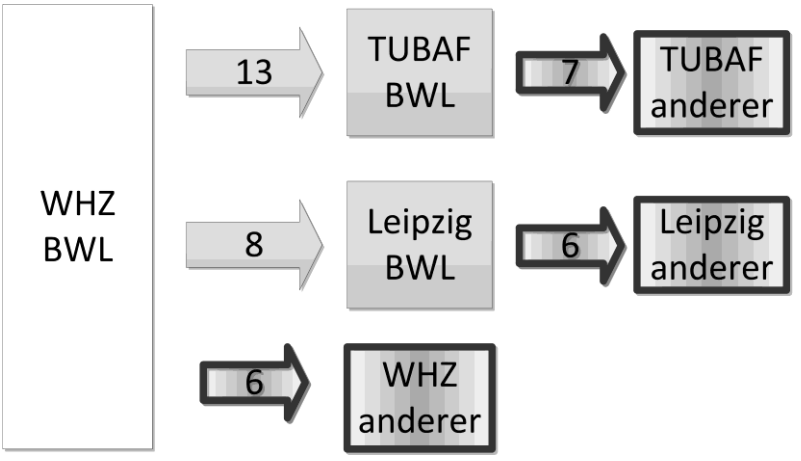


Abbildung 2: Varianzabstände zwischen den erhobenen Modellen

Einige der festgestellten Unterschiede sind möglicherweise auch darauf zurückzuführen, dass in einigen Hochschulen bestimmte Prozessbestandteile in den Prüfungsordnungen verankert sind, während andere Hochschulen darauf verzichten, diese Abläufe in die Prüfungsordnungen aufzunehmen, obwohl die gelebte Realität ggf. sehr ähnlich ist. Da die vorliegenden Prozessmodelle nur auf Basis der Analyse der Prüfungsordnungen der ausgewählten Studiengänge entstanden, deutet dies ggf. auf eine Diskrepanz bzgl. der Auffassung zur Notwendigkeit der rechtsverbindlichen Festlegung bestimmter Abläufe hin.

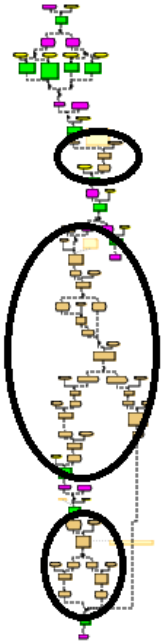


Abbildung 3: Umfang der Prozessunterschiede am Beispiel des Ablaufs vor einer Einzelprüfungsleistung im BWL-Studiengang TUBAF

Weiterhin ist zu beobachten, dass die 40 festgestellten Variabilitätspunkte insgesamt verhältnismäßig umfangreiche Auswirkungen auf die Prozessmodelle beinhalten. Im Durchschnitt führt jeder Variabilitätspunkt zum Hinzufügen, Verändern oder Entfernen von 12,7 Prozessmodellelementen. Die kleinste Varianz umfasst 2 Änderungen, die größte 53. Deutlich wird dies auch in Abbildung 3. Sie zeigt ein Modell des Ablaufs einer Einzelprüfungsleistung im BWL-Studiengang an der TUBAF. Jeweils durch Ellipsen markiert sind die Unterschiede zur entsprechenden Teilprozessmodellvariante des BWL-Studiengangs an der WHZ.

6 Anwendung von C-EPC und Provop auf die Prüfungsprozessvarianz

Die beiden in Kapitel 3 vorgestellten Ansätze zur Beherrschbarkeit von GPMV sollen nun exemplarisch auf einen beispielhaften Variabilitätspunkt in den erhobenen Prüfungsprozessmodellen angewandt werden.

Abbildung 4 zeigt einen Variabilitätspunkt im Ablauf vor einer Einzelprüfungsleistung im Studiengang Gebärdensprachdolmetschen an der WHZ. Die zugehörige Prüfungsordnung sieht in §19 vor, dass der Studierende für eine mündliche Prüfung einen Prüfer bzw. eine Prüfergruppe vorschlagen kann. Diese Möglichkeit besteht laut der entsprechenden Prüfungsordnung im BWL-Studiengang der WHZ nicht. In der Abbildung sind die hinzugekommenen Prozessmodellelemente durch einen Kasten hervorgehoben.

Abbildung 4: Variabilitätspunkt im Prozessmodell für den Studiengang Gebärdensprachdolmetschen an der WHZ

In Abbildung 5 wird die Umsetzung dieses Variabilitätspunkts mit dem C-EPC-Ansatz gezeigt. Die Funktion „Prüfer bzw. Prüfergruppe vorschlagen“ wird zur konfigurierbaren Funktion (erkennbar an der dicken Umrandung). Verbunden wird dieses konfigurierbare Element mit einem sogenannten „Requirement“, das spezifiziert, unter welchen Umstän-

den diese konfigurierbare Funktion Teil des Prüfungsablaufs wird. In diesem Fall wird diese Funktion immer aktiviert, wenn im Prozesskontext definiert ist, dass es sich um den Studiengang Gebärdensprachdolmetschen (GSD) der WHZ handelt. Handelt es sich um einen anderen Studiengang, wird diese Funktion zusammen mit den vor- und nachgelagerten Ereignissen sowie der verbundenen Organisationseinheit aus dem Prozessablauf entfernt. Dies entspricht dem in (Rosemann and van der Aalst 2007) beschriebenen Verhalten.

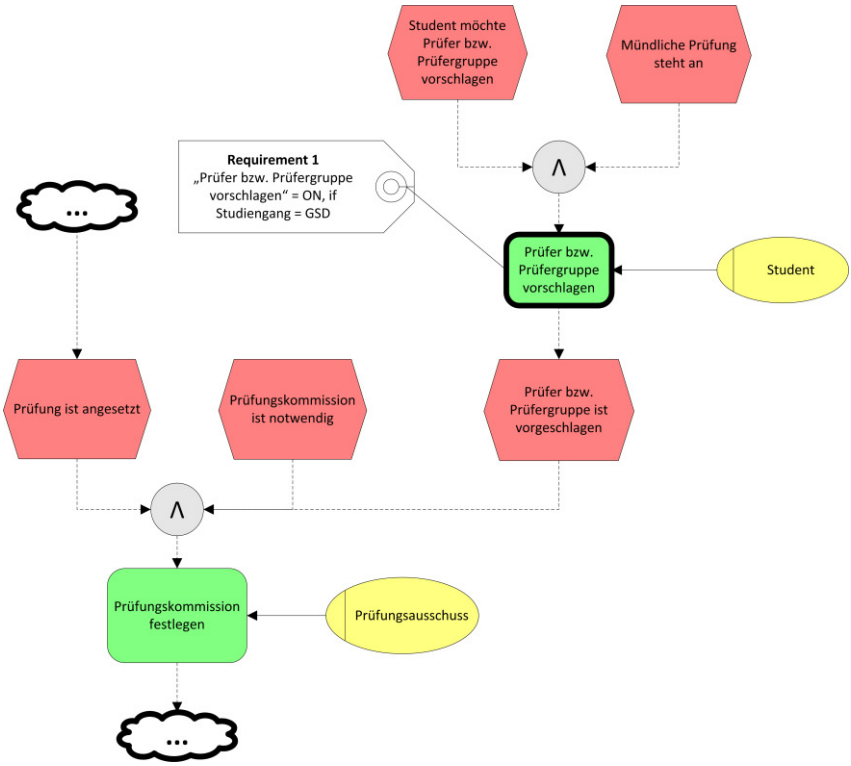


Abbildung 5: Umsetzung des Variabilitätspunkts mit dem C-EPC-Ansatz

Erkennbar ist bei der Anwendung des C-EPC-Ansatzes, dass eine wesentlich kompaktere Darstellung eines Prozessmodells und seiner Varianten möglich ist als beim in Kapitel 1 beschriebenen Single-Model-Ansatz. Dies liegt vor allem daran, dass das Konzept der Varianten im Modell explizit dargestellt wird. Es erfolgt eine deutliche Trennung zwischen Konfigurationsentscheidungen und normaler Prozesslogik. Dennoch bleibt das resultierende Modell nach wie vor sehr groß. Bereits im vorliegenden Fall muss das konfigurierbare Referenzmodell alle 40 Variabilitätspunkte der sechs untersuchten Studiengänge beinhalten. Wird später ein weiterer Studiengang hinzugefügt, muss das C-EPC-Modell überarbeitet werden, was ggf. zu Konsistenzproblemen bzgl. der zuvor abgeleiteten Variantenmodelle führen kann.

In [To13] wird bzgl. der Eignung von C-EPC (bzw. dem Verhaltensansatz allgemein) argumentiert, dass gerade das Zusammenfassen aller nötigen Informationen in einem einzigen Modell den Ansatz für die Umsetzung kleinerer Modelle geeigneter macht. Der Verhaltensansatz ist hier dem Strukturansatz überlegen. Der Strukturansatz würde nämlich diese Informationen über mehrere Artefakte verstreuen, was nach Einschätzung der Autoren ungünstig für die Übersicht ist.

Abbildung 6 zeigt die Umsetzung des gleichen Variabilitätspunkts mit dem Provop-Ansatz. Dabei wird ausgehend von dem im oberen Teil der Abbildung sichtbaren Basisprozess beschrieben, wie von diesem Varianten abgeleitet werden können. Der Basisprozess stellt hierbei selbst auch eine Variante dar. In diesem Fall die Variante für den GSD-Studiengang. Die Variationspunkte, an denen jeweils Änderungen eingebracht werden können, sind mit einem ♦-Symbol und einem Buchstaben gekennzeichnet. Auf diese Variationspunkte wird dann in den Änderungsoptionen Bezug genommen, um festzulegen, wo im Prozessmodell die Änderungen vorgenommen werden sollen. Änderungsoption 1 beispielsweise baut den Basisprozess gemäß der für die Option gültigen Kontextregel (Studiengang ≠ „GSD“) um, wenn es sich um einen anderen Studiengang als GSD handelt. Dementsprechend wird die Funktion „Prüfer bzw. Prüfergruppe vorschlagen“ zusammen mit den zugehörigen Elementen entfernt.

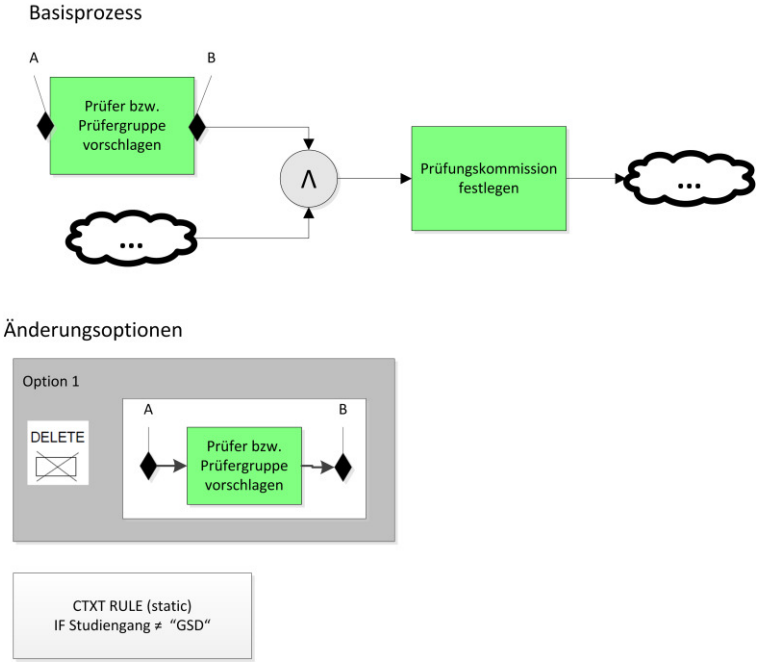


Abbildung 6: Umsetzung des Variabilitätspunkts mit dem Provop-Ansatz

Ein deutlicher Vorteil des Strukturansatzes (zu dem Provop gehört) gegenüber dem Verhaltensansatz liegt darin, dass nicht vorab alle Varianten bekannt sein müssen, um den Basisprozess und die Änderungsoptionen für die Erstellung bestimmter Varianten defi-

nieren zu können. Allerdings muss auch hier bei Einführung neuer Änderungsoptionen zur Erstellung neuer Varianten sorgfältig auf Wechselwirkungen mit bestehenden Änderungsoptionen geachtet werden.

[To13] argumentieren, dass Provop im Vergleich zu C-EPC insbesondere bei größeren Modellen den überlegenen Ansatz darstellt, da hier durch entsprechende Abstraktionsmechanismen (Separieren der Änderungsoptionen vom eigentlichen Modell etc.) das Verständnis für die Nutzer der Modellierungsmethode steigt. Diese Aussage macht auch deutlich, dass es nicht per se einen überlegenen Ansatz gibt, sondern dass hier durchaus auch die Beschaffenheit des Geschäftsprozesses und andere Umweltfaktoren eine Rolle spielen können.

7 Zusammenfassung

Sowohl C-EPC als auch Provop sind mögliche Ansätze zur Beherrschbarkeit von GPMV. Die C-EPC basiert auf den Grundgedanken, das gesamte Verhalten aller Varianten in einem gemeinsamen konfigurierbaren Referenzmodell zu vereinen. Was bei einer geringen Zahl von Varianten zunächst vorteilhaft ist, wird mit steigender Zahl der Varianten unhandlicher. Schon das Zusammenführen der sechs Variantenmodelle der untersuchten Prüfungsprozesse führt zu einem großen und komplexen Modell.

Der Provop-Ansatz hingegen ermöglicht es, ausgehend von einer festgelegten Basisvariante schrittweise die weiteren Prozessmodelle durch Verändern der Struktur des Modells abzuleiten. Somit müssen bei der Anwendung dieses Ansatzes im Wesentlichen immer die Unterschiede zwischen zwei Varianten betrachtet werden, was die Komplexität der Modellierung beträchtlich reduziert und das Vorgehen auch unempfindlicher gegenüber dem späteren Hinzufügen einer weiteren Variante macht. Dennoch ist auch die Anwendung des Provop-Ansatzes auf reale Prozessmodelle mit einem beträchtlichen Modellierungsaufwand verbunden.

Einige Zahlen sollen das verdeutlichen. Die Modelle der sechs Prüfungsprozesse (jeweils zusammengesetzt aus den in Abbildung 1 dargestellten sieben Teilmodellen) umfassen im Durchschnitt 556 Modellelemente (Funktionen, Ereignisse sowie logische Konnektoren). Die Gesamtzahl der in den verschiedenen Varianten hinzugefügten Elemente liegt bei 284. Wären diese auf alle fünf vom Ausgangsmodell abgeleiteten Modelle gleichverteilt, würden jeweils 56,8 Modellelemente hinzukommen. D.h. selbst in den betrachteten (eigentlich durch die gemeinsame Grundlage im SächsHSFG wohldefinierten) Prozessen kommen in jeder Variante durchschnittlich etwa 10,2% variable Elemente hinzu. Aufgrund dessen sowie des bereits in Kapitel 5 beschriebenen mitunter recht großen Umfangs der hinzukommenden Prozessfragmente stellt sich auch die grafische Darstellung der GPMV-Modelle als schwierig dar.

Zusammenfassend ist festzustellen, dass die explizite Modellierung von Modellvariabilität den Vorteil bietet, dass Änderungen (etwa aufgrund gesetzlicher Vorgaben) schnell und zuverlässig für alle Modellvarianten umgesetzt werden können. Der etwa mit der C-EPC-Methode verfolgte Ansatz, alle Varianten in einem Diagramm zu modellieren, stößt

schon bei den sechs von uns untersuchten Modellvarianten schnell an die Grenzen der Handhabbarkeit. Ansätze wie Provop, bei denen die Variabilität durch die Unterschiede zu einem Basismodell ausgedrückt werden, schneiden hier besser ab, setzen aber dennoch sehr gute Modellierungskenntnisse und einen nicht zu unterschätzenden Arbeitsaufwand voraus.

In unserer weiteren Forschung beabsichtigen wir, Methoden zur Modellierung von Modellvarianten weiter zu untersuchen und deren Handhabbarkeit zu verbessern.

Die Arbeit wurde im Rahmen der Initiative „Bildungsportal Sachsen" durch den Arbeitskreis E-Learning der Landesrektorenkonferenz Sachsen mit Mitteln des SMWK gefördert.

Literaturverzeichnis

[AHS11] Altvater, P.; Hamschmidt, M.; Stratmann, F.: Prozessorientierung in Hochschulen – Mehr als Tools und Referenzmodelle. In HIS:Forum Hochschule 12. 2011; S. 1–24.

[Be13] Becker, J.; Clever, N.; Holler, J.; Püster, J.; Shitkova, M.: Semantically Standardized and Transparent Process Model Collections via Process Buliding Blocks. In eKNOW 2013: The Fifth International Conference on Information, Process and Knowledge Management, Nice, 2013. S. 172–177.

[HBR09] Hallerbach, A.; Bauer, T.; Reichert, M.: Issues in Modeling Process Variants with Provop. In Business Process Management Workshops - BPM 2008 International Workshops, Milano 2008. Springer, Berlin Heidelberg 2009; S. 56–67.

[HBR10] Hallerbach, A.; Bauer, T.; Reichert, M.: Configuration and Management of Process Variants. In Handbook on Business Process Management. Springer, Berlin Heidelberg 2010; S. 237–255.

[KY12] Kumar, A.; Yao, W.: Design and Management of Flexible Process Variants Using Templates and Rules. In Computers in Industry 63(2), 2012; S. 112–130.

[Pa11] Pascalau, E.; Awad, A.; Sakr, S.; Weske, M.: On Maintaining Consistency of Process Model Variants. In Business Process Management Workshops - BPM 2010 International Workshops and Education Track, Hoboken 2010. Springer, Berlin Heidelberg 2011; S. 289–300.

[RA07] Rosemann, M.; van der Aalst, W.M.P.: A Configurable Reference Modelling Language. In Information Systems Databases: Their Creation, Management and Utilization. Elsevier Science Ltd, Oxford, 2007; S. 1–23.

[To13] Torres, V.; Zugal, S.; Weber, B.; Reichert, M.; Ayora, C.; Pelechano. V.; A Qualitative Comparison of Approaches Supporting Business Process Variability. In Business Process Management Workshops, BPM 2012 International Workshops, Tallinn, 2012; Springer, Berlin Heidelberg 2013; S. 560–572.

[Va13] Valenca, G.; Alves, C.; Alves, V.; Niu, N.: A Systematic Mapping Study on Business Process Variability. In International Journal of Computer Science & Information Technology 5(1), 2013; S. 1-21.

[Vo12] Vogelaar, J.; Verbeek, H.M.W.; Luka, B.; van der Aalst, W.M.P.: Comparing Business Processes to Determine the Feasibility of Configurable Models: A Case Study. In Business Process Management Workshops, BPM 2011 International Workshops, Clermont-Ferrand, 2011. Springer, Berlin Heidelberg 2012; S. 50–61.

Mitarbeiterwiderstände und -partizipation bei der Einführung von BPM – Ein Multi-Level-Ansatz am Beispiel des Landespolizeiamtes Schleswig-Holstein

Dirk Frosch-Wilke[1]; Rudolf Gaspary[2]

[1]Institut für Wirtschaftsinformatik
Fachhochschule Kiel
Sokratesplatz 2
24149 Kiel
[2]Sachgebiet 123
Qualitätsmanagement, Controlling, methodische Beratung
Landespolizeiamt Schleswig-Holstein
Mühlenweg 166
24116 Kiel
dirk.frosch-wilke@fh-kiel.de
rudolf.gaspary@polizei.landsh.de

Abstract: Die Beteiligung der Mitarbeiter als Instrument eines integrierten Prozessmanagements ist allgemein anerkannt. In diesem Beitrag werden unterschiedliche Beteiligungsformen von Mitarbeitern und deren Auswirkung auf die Transparenz, die Akzeptanz und die Prozessqualität bei der Einführung des Business Process Managements (BPM) beim Landespolizeiamt Schleswig-Holstein dargestellt. Ausgangspunkt der Mitarbeiterbeteiligung ist hierbei ein Multi-Level-Ansatz zur Analyse wahrgenommener Widerstände. Mit Hilfe einer solchermaßen vertieften Analyse von Mitarbeiterwiderständen lassen sich zielgerichtet geeignete Beteiligungsformen von Mitarbeitern der öffentlichen Verwaltung in einem BPM-Projekt ermitteln, die zu verbesserten Prozessen führen können.

1 Einleitung

Schon seit vielen Jahren wird das Business Process Management (BPM) in der öffentlichen Verwaltung als ein Managementkonzept begriffen, das Verwaltungshandeln effektiver und bürgernäher (z.B. in Verbindung mit dem eGovernment) gestalten kann. Dabei greift man überwiegend auf bewährte und anerkannte Verfahren und Methoden des BPM in der Privatwirtschaft zurück. Dabei gelten jedoch in der öffentlichen Verwaltung besondere Rahmenbedingungen, so dass das Prozessmanagement dort noch keinen durchgängigen „Siegeszug“ angetreten hat [VF10]. Als eine besondere Rahmenbedingung, die die Einführung des Prozessmanagements in der öffentlichen

Verwaltung betrifft, werden die elementaren Veränderungen, die für die Mitarbeiter[1] gerade auf der Ebene der Sachbearbeiter mit der Einführung des BPM eintreten können, genannt [VF01].

Organisatorische Veränderungen erzeugen in der Regel zunächst einmal Widerstände bei den Mitarbeitern, da Veränderungsprozesse die Transformation von Bekanntem zu Unbekanntem beinhalten [BH01]. Aus diesem Grund wird dem Thema „Mitarbeiterwiderstand“ im Rahmen des Change Managements eine große Bedeutung beigemessen, da Widerstände von Mitarbeitern ein wesentlicher Misserfolgsfaktor von organisatorischen Veränderungen sind (siehe u.a. [Hä08],[Pi00], [St96]). Widerstände von Mitarbeitern gegen Veränderungen, wie sie z.B. durch die Einführung des BPM in der öffentlichen Verwaltung auftreten, werden oftmals sehr stark aus einer personenzentrierten Sicht (z.B. Befürchtungen, Missverständnisse) heraus betrachtet, und es wird nicht hinreichend der organisatorische Kontext (z.B. Vorgesetzte, Kollegen/innen, Kommunikationsformen, Fähigkeit der Organisation zum Wandel) beachtet. Jedoch sind gerade diese kontextuellen Faktoren sowohl hilfreich für die Erklärung von Widerständen als auch ergeben sich hieraus Ansatzpunkte, negative Einstellungen gegen Veränderungen in positive zu überführen (siehe u.a. [FFD08], [KK00]). Daher plädieren Autoren wie Bouckenooghe [Bo09] oder Ford et al. [FFD08] für einen Multi-Level-Ansatz, der neben den personalen auch den organisatorischen Kontext beim Umgang mit Widerständen bei Veränderungsprozessen berücksichtigt.

In diesem Beitrag greifen wir diesen Multi-Level-Ansatz auf und zeigen, wie unter Beachtung verschiedener organisatorischer und sozialer Faktoren, Mitarbeiterwiderstände durch adäquate Formen der Mitarbeiterbeteiligung bei der Prozesserhebung und –verbesserung, abgebaut und in positive Einstellungen gegenüber dem BPM in der öffentlichen Verwaltung überführt werden können. Als Fallbeispiel dient dabei eine Prozessanalyse und –verbesserung in einem Sachgebiet innerhalb des Landespolizeiamtes von Schleswig-Holstein.

Der Aufbau der Arbeit folgt in wesentlichen Teilen dem Design-Science-Research-Ansatz von Hevner et al. [He04]: Im **Kapitel 2** erfolgt die allgemeine Problembeschreibung, indem wir auf Mitarbeiterwiderstände bei organisatorischen Veränderungen bei Einführung des BPM eingehen. Die Forschungsfrage sowie der Stand der Forschung zur multi-perspektivische Betrachtung von Mitarbeiterwiderständen und verschiedenen Formen der Beteiligung von Mitarbeitern im Rahmen des Change Managements werden ebenfalls in diesem Kapitel behandelt. Das **Kapitel 3** stellt dann das BPM-Projekt bei der Landespolizei Schleswig-Holstein vor. Die Multi-Level-Analyse als Instrument des Umgangs mit Mitarbeiterwiderständen bei der Einführung des BPM in der öffentlichen Verwaltung wird exemplarisch in **Kapitel 4** vorgestellt. Die auf der Multi-Level-Analyse beruhende Auswahl geeigneter Mitarbeiterbeteiligungsformen in diesem Projektkontext, wird hier ebenfalls beschrieben. Das **Kapitel 5** schließt diesen Beitrag mit einer ersten Evaluation der erzielten Ergebnisse und einem Ausblick ab.

[1] Aus Gründen der Lesbarkeit verwenden wir in diesem Beitrag nur die männliche Form.

2 Widerstände und Partizipation bei organisatorischen Veränderungen

2.1 Widerstand bei organisatorischen Veränderungen

Widerstand gegen Veränderungen wird vordringlich aus der Perspektive derjenigen betrachtet, die Veränderungen in einer Organisation umsetzen wollen (engl. change agents). Demzufolge wird Mitarbeiterwiderstand auch als eine Hürde im Rahmen der Einführung des BPM bezeichnet. Ferner ergibt sich bei dieser eindimensionale Betrachtung, dass Widerstand als ein objektiv vorhandener Sachverhalt verstanden wird, und dabei nicht in Erwägung gezogen wird, dass Widerstand lediglich eine Interpretation von Verhalten von Mitarbeitern durch die „change agents" ist (vgl. [FFD08]). Auch wird dann nicht hinreichend betrachtet, welchen Einfluss die „change agents" und deren Verhalten und Interaktionen sowie die Vorgehensweise im Veränderungsprojekt zur Existenz und zum Ausmaß von Widerstand beitragen.

Letztlich kann eine solch einseitige Sichtweise von Widerstand dazu führen, dass die möglichen rationalen Gründe, die Ursache des Widerstands sind, nicht erkannt und dann auch nicht im Sinne der Organisation in den Veränderungsprozessen berücksichtigt werden (vgl. [SJM02], [KL04]).

Daher bedarf es einer mehrdimensionalen Betrachtung (Multi-Level-Ansatz) bei der Betrachtung von Widerständen, die berücksichtigt, dass

a) Widerstand nicht ein objektiv vorhandener Sachverhalt sein muss,

b) „change agents" sowie die Organisation und Art der Durchführung eines BPM-Einführungsprojektes können ursächlich für Widerstände sein und

c) das Widerständen ein Potential zur verbesserten Implementierung des BPM innewohnen kann.

2. 2 Mitarbeiterpartizipation

Die aktive Beteiligung von Mitarbeitern an Veränderungsprozessen greift die im vorherigen Abschnitt beschriebene Multi-Level-Analyse bei der Betrachtung von Widerständen auf. So soll durch die Beteiligung der Mitarbeiter z.B. eine höhere Transparenz geschaffen werden und somit den Mitarbeitern eine bessere Einschätzung ermöglicht werden, die Veränderungen des eigenen Arbeitsumfelds nach Einführung des BPM beurteilen zu können. Ferner können mit der Beteiligung der Mitarbeiter Potentiale für eine verbesserte Implementierung des BPM gehoben werden. Auch treten mögliche Widerstände offensichtlicher zu Tage, werden dadurch auch objektiver und die Ursachen können besser erkannt werden.

Die Intensität der Beteiligung der Mitarbeiter an einem BPM-Einführungsprojekt kann je nach Ausgangssituation und Aufgabenstellung unterschiedlich sein. Die Abbildung 1 gibt einen Überblick über mögliche Beteiligungsformen.

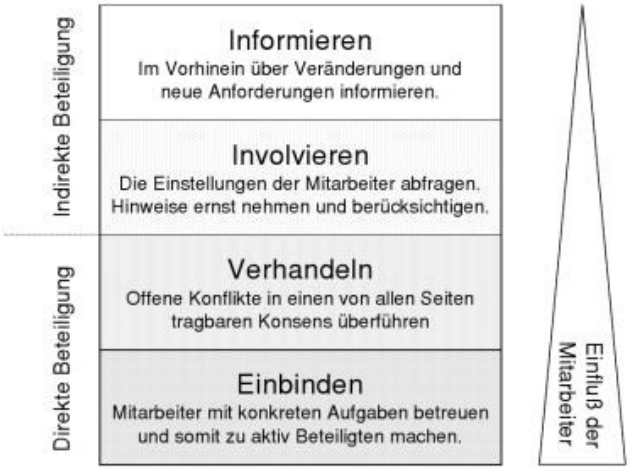


Abbildung 1: Formen der Mitarbeiterbeteiligung [Er00]

Gemäß des Multi-Level-Ansatzes bedarf die Auswahl angemessener Formen der Mitarbeiterbeteiligung zur Reduktion von Widerständen gegen organisatorische Veränderungen zunächst einer gründlichen Analyse dessen, was überhaupt als Widerstand wahrgenommen wird. Abhängig vom Ergebnis dieser Analyse sollte über die Häufigkeit und Intensität der Mitarbeiterbeteiligung entschieden werden. Ein genereller „je-mehr und je-häufiger desto besser"-Ansatz ist unserer Erfahrung nach nicht zielführend, da er zu einer Überlastung der Mitarbeiter führen kann, zumal diese durch die Stelleneinsparungen der vergangenen Jahre, über wachsenden Leistungsdruck, Arbeitsverdichtungen und erhöhte Verfügbarkeitsanforderungen klagen [GdP13].

3 BPM-Projekt beim Landespolizeiamt Schleswig-Holstein

3.1 Organisatorischer Rahmen

Das Landespolizeiamt (LPA) ist das polizeiliche Führungs- und Logistikzentrum auf Landesebene in Schleswig-Holstein. Das LPA

- unterstützt das Innenministerium bei der Wahrnehmung seiner Grundsatzaufgaben,
- führt die regionalen Polizeidirektionen und die Wasserschutzpolizeireviere,
- übernimmt die Einsatzführung bei herausragenden Anlässen,
- regelt den polizeilichen Aufgabenvollzug,

- versorgt die Landespolizei mit Dienstleistungen insbesondere in den Bereichen Personalwirtschaft, Verwaltung, Technik und Haushalt
- nimmt die Stabs- und Führungsaufgaben für die Bereiche der Verkehrs- und der Wasserschutzpolizei wahr.

3.2 Ziele und Aufgaben des Prozessmanagement-Projektes

Mit Hilfe des Prozessmanagement sollen die Strukturen der Landespolizei Schleswig-Holstein untersucht werden.

Primäres strategisches Ziel ist die Verbesserung der Geschäftsprozesse, um diese zu verschlanken und Effizienzsteigerungen innerhalb der Organisation zu realisieren.

Ein weiteres Ziel des BPM-Projektes ist die Identifikation von weiteren IT-Unterstützungsmöglichkeiten in den Prozessen des LPAs.

Ursprünglich sollten in diesem Projekt alle Bereiche der Landespolizei Schleswig-Holstein betrachtet werden. So sollte es keine Bereiche geben, die a-priori von einer Untersuchung ausgenommen werden. Es wurde nämlich mit einem erheblichen Maße an Widerstände in allen Ebenen der Polizei gerechnet. Da jedoch noch keine Erfahrungen mit den Methoden und Konzepten des Prozessmanagements bei der Landespolizei vorlagen und es auch galt, zunächst einmal Know-how aufzubauen, wurde der Fokus des Projektes zunächst auf das Landespolizeiamt innerhalb der Landespolizei gelegt.

Die Abteilungen des LPAs wurden gebeten, Organisationseinheiten zu benennen, die sich für eine Untersuchung mit den Methoden des Prozessmanagements eignen und bereit sind, sich aktiv einzubringen.

Vor diesem Hintergrund wurden die folgenden operativen Ziele definiert:

- Identifizierung der Geschäftsprozesse in Organisationseinheiten ausgewählter Fachlichkeiten,
- Modellierung dieser Geschäftsprozesse,
- Analyse der Geschäftsprozesse,
- Modellierung von Soll-Prozessen für den Fall erkannter Verbesserungspotentiale,
- Umsetzung der Soll-Prozesse und -nach Ablauf einer Halbjahresfrist- Evaluierung der modifizierten Geschäftsprozesse.

Eine Voraussetzung für die Prozessanalyse beim LPA war, zunächst eine Qualifizierung von Mitarbeitern vorzunehmen, die geeignet schienen, eine Prozessmodellierung zu moderieren, zu begleiten und umzusetzen. Die Organisationseinheit für Qualitätsmanagement und methodische Beratung ist als fachliche Einheit des LPA mit der Untersuchung beauftragt worden. Die qualifizierende Unterstützung wurde von der Fachhochschule Kiel übernommen.

Die Forschungsfragen, die sich im Rahmen dieses Projektes stellte, war: Inwiefern kann die multi-perspektivische Betrachtung und Analyse von Mitarbeiterwiderständen zur Auswahl geeigneter Formen der Mitarbeitbeteiligung beitragen und inwiefern wirken sich letztere dann auf das Projektergebnis aus?

4 Vorgehensweise und Ergebnisse

Im Folgenden wird exemplarisch anhand der Prozessaufnahme, -modellierung und –verbesserung die Vorgehensweise beschrieben, der ein Multi-Level-Ansatz zum Umgang mit Mitarbeiterwiderständen zugrunde liegt. Als Beispiel dient dabei die Arbeit innerhalb eines Sachgebiets des LPA, dass aus Gründen der Anonymisierung lediglich mit Sachgebiet 1 (SG1) bezeichnet und nicht weiter beschrieben wird. Diese Anonymisierung hat keine Auswirkungen auf die hier beschriebenen Erkenntnisse, da diese unabhängig von dem konkreten Tätigkeitsfeld des SG1 sind.

4.1 Analyse der Ausgangssituation

Die Leitung des SG1 war sehr daran interessiert, dass eine Prozessanalyse innerhalb ihres Zuständigkeitsbereiches durchgeführt wird. Hintergrund dafür war, dass erhebliche Effizienzsteigerungspotentiale innerhalb des SG1 durch Prozessverbesserungen vermutet wurden. Viele der Verfahren im SG1 sind „historisch gewachsen“ und werden zum Teil je nach den beteiligten Personen und - bei Sachgebietsübergreifenden Prozessen - anderen involvierten Dienststellen der Landespolizei unterschiedlich gehandhabt.

Aufgrund ihres Interesses an dem Projekt kann die Leitung des SG1 als „change agent“ (vgl. Kapitel 2) bezeichnet werden. Durch das starke Interesse der SG1-Leitung zur Teilnahme an dem Projekt, sind bei den Sachbearbeitern im SG1 aber Befürchtungen geweckt worden, dass ihr Sachgebiet einen Beitrag zum geplanten Stellenabbau bei der Landespolizei leisten soll. Dies in einer Situation, in der die Mitarbeiter aktuell über hohe Arbeitsbelastungen klagen und gleichzeitig in der jüngeren Vergangenheit gegen ihre tariflichen Eingruppierung protestiert haben, da diese ihrer Ansicht nach nicht ihren tatsächlichen Aufgaben und Verantwortlichkeiten entspricht.

Somit kann die Situation in diesem Sachgebiet vor Beginn der Prozessanalyse als durchaus angespannt, bezeichnet werden. Von der Leitung des SG1 wurde an das BPM-Projektteam kommuniziert, dass mit Widerständen der Mitarbeiter zu rechnen ist

4.2 Art der Mitarbeiterbeteiligung

Nach der Rückmeldung der Leitung des SG1 an das Projektteam wurde die Ausgangslage wie oben beschrieben aufgenommen. Anschließend wurde diese mit Hilfe des Multi-Level-Ansatz zur Beurteilung der kommunizierten Mitarbeiterwiderstände (vgl. Kap.2.1) analysiert:

a) Die als Widerstände benannten Reaktionen der Mitarbeiter des SG1 auf die Ankündigung der Prozessanalyse basierten auf Äußerungen, die vor allem Skepsis, Unerfahrenheit und Unsicherheit ausdrückten. Diese sind bedingt durch die aktuelle Situation im SG1 (siehe Kap. 4.1) und sind somit nicht primär als Widerstände gegen das BPM-Projekt des LPAs zu werten.

b) Bei der engagierten Interessenbekundung der Leitung des SG1 an einer Prozessanalyse haben sich die Mitarbeiter des SG1 nicht hinreichend beteiligt gefühlt und die Ziele, die damit verfolgt werden sollten, waren ihnen nicht hinreichend bekannt. Auch verstärkte der ursprüngliche BPM-Kontext bzgl. die angekündigten Stelleneinsparungen bei der Landespolizei Schleswig-Holstein die Befürchtung, dass Stellen im SG1 gestrichen werden sollten.

c) Einige Mitarbeiter zeigten sich in der Vergangenheit durchaus kritisch in Bezug auf organisatorische und personalpolitische Gegebenheiten. Allerdings konnte bei vielen Mitarbeitern keine „Dienst-nach-Vorschrift"-Mentalität festgestellt werden, sondern eher ein Interesse, dass die notwendigen Arbeiten erledigt werden, auch wenn sie nicht unmittelbar zu den eigenen Aufgaben gemäß Stellenplatzbeschreibung gehören. Vor dem Hintergrund dieser Einstellungen der Mitarbeiter und der großen Arbeitsverdichtung im SG1 kann a-priori ein Interesse an der Realisierung von Effizienzsteigerungen und organisatorischen Verbesserungen im SG1 bei den Mitarbeitern unterstellt werden.

Diese mehrdimensionale Betrachtung der „Widerstände" im SG1 diente als Basis für eine begründete Auswahl geeigneter Arten der Mitarbeiterbeteiligung. Die Analyse zeigte, dass nicht per se von Widerständen gegenüber dem BPM auszugehen ist, sondern die Mitarbeiter eher unsicher sind, da sie die Konsequenzen für ihr eigenes Arbeitsumfeld nicht abschätzen können. Ferner zeigte sich auch, dass die Mitarbeiter ein großes Interesse an der Mitgestaltung und Verbesserung der in ihrem Bereich stattfindenden Verwaltungsverfahren haben.

Daher wurden zwei Arten der Mitarbeiterbeteiligung gewählt (vgl. Abbildung 1):

1. Indirekte Beteiligung - Information

   Die Mitarbeiter wurden über die Ziele des BPM-Projektes informiert. Insbesondere wurden dadurch die Befürchtungen im Hinblick auf einen Stellenabbau entkräftet. Auch wurden die Mitarbeiter darüber informiert, dass ihre aktive Mitwirkung an der Erhebung und Verbesserung der Prozesse im SG1 gewünscht ist.

2. Direkte Beteiligung – Einbinden

   Die Mitarbeiter wurden intensiv bei der Aufnahme aller Ist-Prozesse im SG1 beteiligt. Auch die auf Basis dieser Ist-Aufnahme entstandenen Prozessmodelle wurden zusammen mit den Mitarbeitern des SG1 geprüft. Ferner waren die Mitarbeiter aufgefordert, in jeder Phase der Prozessaufnahme und Modellprüfung, auf mögliche Prozess-Schwachstellen hinzuweisen.

4.3 Umsetzung der Mitarbeiterbeteiligung

Nach einem Vorgespräch mit der Leitung des SG1 bei dem über die konkreten Ziele des BPM-Projektes informierte wurde, erfolgte auch eine entsprechende Information der Mitarbeiter des SG 1. Auch zu Beginn eines ersten (eintägigen) Workshops mit den Mitarbeitern wurden von Seiten der BPM-Projektleitung die Ziele des Projektes ausführlich erläutert.

Ziele dieses Workshops waren zum einen die Information der Mitarbeiter und zum anderen sollten die Mitarbeiter als „Prozessexperten" bei der Erhebung der Ist-Prozesse und dem Erkennen von Prozessschwachstellen mitwirken. Somit diente dieser Workshop sowohl der Herstellung von Transparenz über das Vorhaben als auch der Einbindung der Mitarbeiter, da diese beiden Arten der Mitarbeiterbeteiligung als die geeignetsten nach der Multi-Level-Analyse der Widerstände identifiziert werden konnten.

Insbesondere durch die Beteiligung der Mitarbeiter als Experten im Rahmen des Workshops konnten „Unsicherheiten" gegenüber dem Prozessmanagement abgebaut werden, da die Mitarbeiter aktiv mitgestalten konnten.

Im Rahmen dieses Prozesses wurden unter Zuhilfenahme von Karten und rudimentären, an die BPMN 2.0 Notation (siehe z.B. [GL13]) angelehnten Symbolen, die Ist-Prozesse „modelliert" (vgl. Abb. 2[2]). Auf diese Weise wurde BPMN 2.0 zumindest ansatzweise den Mitarbeitern bekannt gemacht.

[2] Die Abbildungen 2 und 3 sollen lediglich die Vorgehensweise/Methodik nicht aber die jeweiligen Inhalte der Prozessdokumentation veranschaulichen.

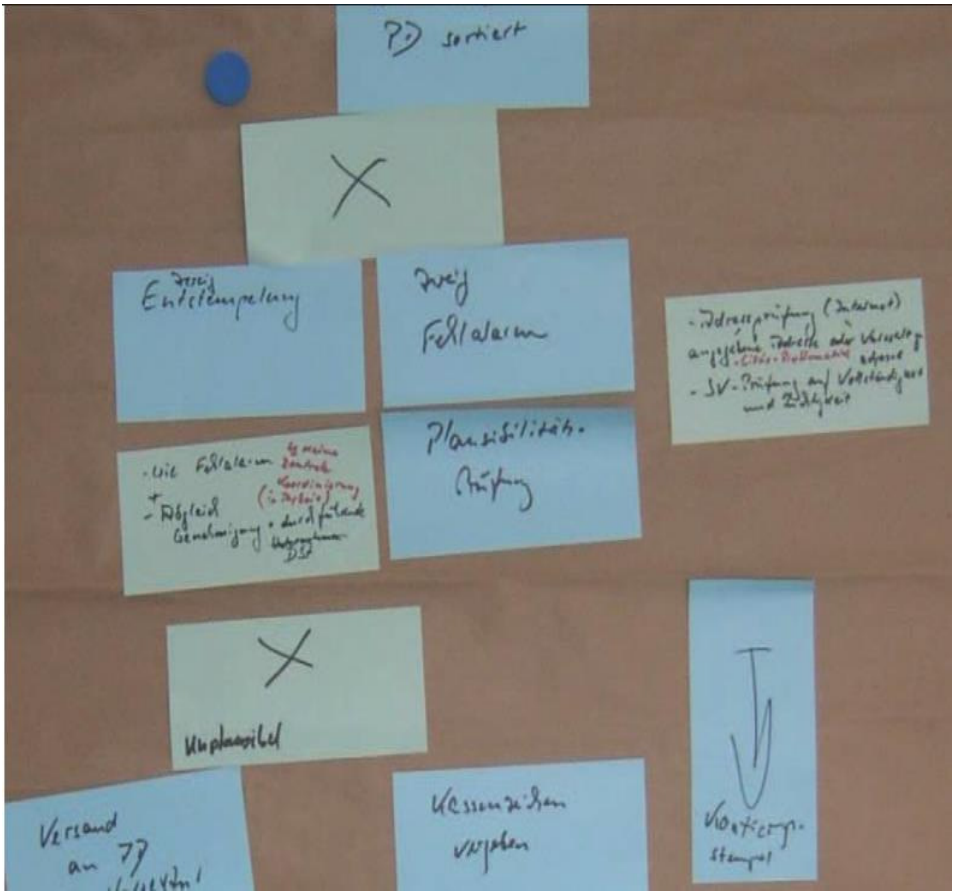


Abbildung 2: Erste Prozessdokumentation

Hintergrund dieser Vorgehensweise war, dass die Mitarbeiter auch im Rahmen der Qualitätssicherung der Prozessmodelle beteiligt werden sollten, und das Projektteam sich für BPMN 2.0 als Modellierungssprache für die Dokumentation der Prozesse entschieden hatte. So wurden auch die im Rahmen des Workshops erhobenen Prozesse des SG1 mit Hilfe der BPMN 2.0 modelliert (vgl. Abb. 3)

Gleichzeitig wurden im Rahmen der Prozesserhebung von den Mitarbeitern zahlreiche Prozessverbesserungsvorschläge herausgearbeitet. Die Mitarbeiter zeigten sich hierbei sehr aktiv und interagierten auch untereinander intensiv. Die gemachten Prozessverbesserungsvorschläge wurden protokolliert.

Im weiteren Projektverlauf wurden diese Prozessverbessungsvorschläge bei der Definition der Soll-Prozesse herangezogen.

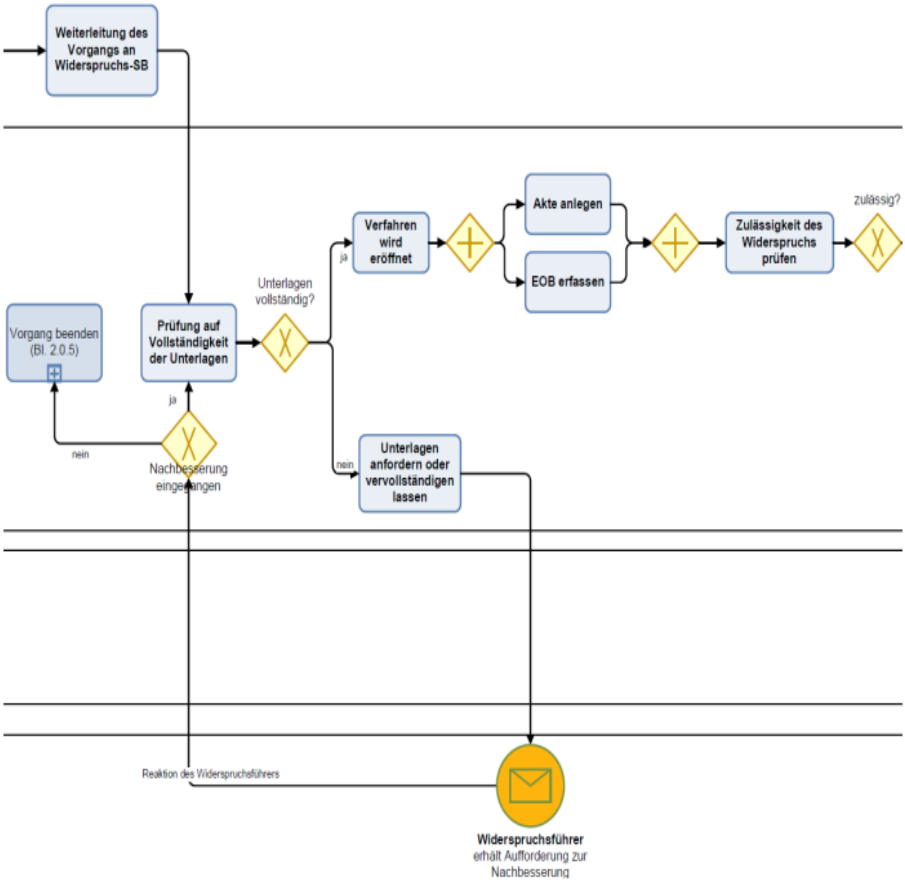


Abbildung 3: Prozessdokumentation mittels BPMN 2.0 (Ausschnitt)

Die Ist- als auch Soll-Prozessmodelle wurden mit den Mitarbeitern im Detail durchgesprochen. Dabei machte es sich positiv bemerkbar, dass bereits beim Prozessaufnahme-Workshop die Mitarbeiter für ein Prozessmodellierung sensibilisiert wurden. Daher konnte der Zeitaufwand für die Einarbeitung in die Modellierungssprache BPMN 2.0 reduziert werden, der ansonsten gerade bei Mitarbeitern in der öffentlichen Verwaltung, die in der Regel über keine Erfahrungen mit semi-formalen Sprachen verfügen, oft sehr hoch ist [LH13].

Im weiteren Vorgehen haben die von den modifizierten Prozessen betroffenen Mitarbeiter eigeninitiativ Modellierungsdetails vorgeschlagen. So konnte identifizierte Doppelarbeit eliminiert werden und die Prozesse verschlankt werden. Eine Nacherhebung 6 Monate nach der Einführung der modifizierten Sollprozesse hat die Richtigkeit des Vorschlags gezeigt und zur Akzeptanz des Tools BPM beigetragen.

5 Zusammenfassung und Ausblick

In diesem Beitrag wurde ein Multi-Level-Ansatz zur Beurteilung von Mitarbeiterwiderständen verwendet, um geeignete Mitarbeiterbeteiligungsformen bei einem BPM-Projekt innerhalb der Landespolizei Schleswig-Holstein zu identifizieren.

Dabei zeigte sich, dass das beobachtbare Mitarbeiterverhalten, das insbesondere von „change agents“ als Widerstand gegen organisatorische Veränderungen wahrgenommen wurde, unterschiedliche Ursachen haben kann. So resultiert der „scheinbare“ Widerstand nicht unbedingt aus den Zielen des BPM-Projektes, sondern ist oftmals durch die organisatorische und persönliche Historie, die die Mitarbeiter in der öffentlichen Verwaltung erfahren haben, begründet. Bei einer angemessenen Beteiligung von Mitarbeitern an einem BPM-Projekt können sogar Potentiale für Prozessverbesserungen erzeugt werden.

Die obigen Erkenntnisse beruhen auf Erfahrungen, die die Autoren bei einem BPM-Projekt der Landespolizei Schleswig-Holstein machen konnten. Daher kann diese Arbeit lediglich den Anstoß für eine umfassendere Untersuchung geben. Letztere ist notwendig, um validieren zu können, ob sich die hier beschriebenen Erkenntnisse auf BPM-Projekte in der öffentlichen Verwaltung verallgemeinern lassen und der Multi-Level-Ansatz somit ein geeignetes Instrumentarium darstellt. Hierbei sollten verschiedene organisatorische Kontexte sowie unterschiedliche Mitarbeiterbeteiligungsformen systematisch untersucht werden.

Literaturverzeichnis

[BH01] Bovey, W.H.; Hede, A.: Resistance to organizational change: the role of cognitive and affective processes. In: Leadership and Organizational Development Journal, Vol. 22, 2001, S. 372-382.

[Bo09] Bouckenooghe, D.: Change recipants’ attitudes toward change: A review study, Vlerick Leuven Gent Working Paper Series, No.2009/14, 2009.

[Er00] Erdmann, J.: Integriertes Prozessmanagement, Hamburg, 2000.

[FFD08] Ford, J.D.; Ford, L.W.; D’Amelio, A.: Resistance to change: The rest of the story (2008). In: Academy of Management Review, Vol. 33, 2008, S. 362–377.

[GD13] Göpfert, J.; Lindenbach, H.: Geschäftsprozessmodellierung mit BPMN 2.0, Oldenbourg Wissenschaftsverlag, 2013.

[GdP13] Gewerkschaft der Polizei Schleswig-Holstein: Uns steht das Wasser jetzt schon bis zum Hals, 2013, http://www.gdp.de/gdp/gdpsh.nsf/id/DE_GdP-Protest-zum-Stellenabbau (zuletzt abgerufen am: 12.04.2014)

[Hä08] Häring, K.: Die Rolle der Führungskräfte in Veränderungsprozessen. In (Litzcke, S.M.; Nolte, R., Hrsg): Change Management – Theorie und Paxis, Brühl, 2008, S. 67-79.

[HL13] Laue, R.; Hogrebe, F.: Zur Verständlichkeit graphischer Symbole in Geschäftsprozessmodellierungssprachen, In (Horbach, M. (Hrsg.)): INFORMATIK 2013 – Informatik angepasst an Mensch, Organisation und Umwelt, Springer, 2013, S. 693-705.

[KK00] Klein, K.J.,; Kozlowski, S.W.J.: (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. Organizational Research Methods, Vol. 3, 200, S.211–236.

[KL04] Knowles, E. S.; Linn, J. A.:. The promise and future of resistance and persuasion. In (Knowles, E. S.; Linn, J. A. (Hrsg.)): Resistance and persuasion, NJ, 2004, S. 301–310.

[Pi00] Piderit, S. K.: Rethinking resistance and recognizing ambivalence: A multidimensional view of attitudes toward an organizational change. In: Academy of Management Review, Vol. 25, 2000, S. 783–794.

[SJM02] Schulz-Hardt, S.; Jochims, M.; Frey, D. 2002. Productive conflict in group decision making: Genuine and contrived dissent as strategies to counteract biased information seeking. In: Organizational Behavior and Human Performance, Vol. 88, 200, S. 563–586.

[St96] Strebel, P.: Why do employees resist change?. In: Harvard Business Review, May-June 1996, S. 86-92.

[VF10] Voigt, J.; Frühauf, L.: Prozessmanagement in der öffentlichen Verwaltung. In (Jochem, R.; Mertins, K.; Knothe, T.; Hrsg.): Prozessmanagement – Strategien, Methoden, Umsetzung, Symposion, 2010, S. 353-376.

Das BPM-Standardisierungssystem in der Öffentlichen Verwaltung der Schweiz und dessen Theorie-basierte Evaluierung

Konrad Walser

Berner Fachhochschule, E-Government-Institut
Morgartenstrasse 2a/Postfach 305
CH-3000 Bern 22
konrad.walser@bfh.ch

Abstract: Im vorliegenden Beitrag wird der Versuch unternommen, das Vorgehen zur BPM-Standard-Artefakteentwicklung nach eCH einer wissenschaftlichen Fundierung zu unterziehen. Überdies leistete der Beitrag einen Kurzbeschrieb des aktuell impliziten BPM-Standardisierungssystems der ÖV in der Schweiz, wodurch dieses explizit gemacht wird. Weiter wurden verschiedene Theoriebausteine zur Entwicklung des BPM-Standardisierungssystems (Darstellung) und zur Begründung von dessen Zweckmäßigkeit zur Anwendung gebracht. Im Weiteren folgt eine Diskussion der Verbesserungen der Prozessorientierung auf Basis aktueller Entwicklungen und ausgehend vom BPM-Standardisierungssystem und dessen wissenschaftlicher Fundierung. Im Weiteren zeigt dieser Beitrag, dass aus evidenzorientierter empirischer Sicht noch immer eine geringe BPM-Verbreitung im kommunalen Bereich in der Schweiz vorhanden ist. Ein weiteres Thema stellen im vorliegenden Beitrag überdies Klarstellungen zur aktiven Gestaltbarkeit der BPM-Standardisierungssystems im Hinblick auf mehr Wirksamkeit und verbesserter Umsetzung des BPM in der ÖV dar.

1 Einleitung

1.1 Problemstellung

Dieser Beitrag nimmt Bezug auf frühere Beiträge ([WSB10], [WS11]) in denen die Konkretisierung eines BPM-Standardisierungssystems der Öffentlichen Verwaltung (ÖV) Schweiz dargestellt wird. Der vorliegende Beitrag geht dabei einen Schritt weiter und konkretisiert modellhaft das aktuelle BPM-Standardisierungssystem an sich und untersucht theoriebasiert dessen aktuell implizite Konzeption, d.h. macht dieselbe explizit. Dabei handelt es sich um eine Untersuchung im Nachhinein und nicht um eine Konzeption zum Voraus. Verwaltungen präsentieren sich typischerweise als miteinander zusammenarbeitende und somit kollaborative (prozessorientierte) Institutionen oder Elemente eines Staatsystems (Verbundsystem), deren horizonta-

le, vertikale und interoperable interne und externe Zusammenarbeit über Aufgabenverteilungen geregelt ist ([LSS10], S. 32 ff.). Die Kollaboration macht an den Grenzen der Verwaltung nicht halt. Die Verwaltung kann einerseits als von den Lieferanten bis zu den Kunden der Verwaltung reichende Leistungsverwaltung verstanden werden. Die Verwaltung kann andererseits als politische Verwaltung verstanden werden, in welcher in intensivem Dialog mit Stakeholdern Politiken entwickelt, beschlossen, verworfen, implementiert, evaluiert sowie terminiert werden ([Wa12a], [Wal12b]). Überdies stehen Verwaltungen als Teil der Exekutive in intensiven Austauschbeziehungen zu Legislativen und Judikativen. Die Zusammenarbeit erfolgt in allen geschilderten Tätigkeitsdomänen der Verwaltung typischerweise über drei oder mehr föderale Ebenen (Vgl. [Li05]; Föderalismusprinzip). Die Arbeitsteilung (Subsidiaritätsprinzip [Li05]) determiniert die Arbeit in den Verwaltungseinheiten und die Integration zwischen den Verwaltungseinheiten bei der Verwaltungs-übergreifenden Zusammenarbeit. Verwaltungsaufgaben sind gesetzlich determiniert. Gesetze wiederum basieren auf Verfassungen. In den Verfassungen wird auch das angesprochene Subsidiaritätsprinzip in seiner Länder- oder Staats-spezifischen Ausprägung konkretisiert [20]. Das Thema Geschäftsprozessmanagement ist heute in Verwaltungen ein sowohl theoretisch wie praktisch diskutiertes Thema ([AIP10], [BAF09], [Be03], [BEK10], [BS10], [GH10], [Ha05], [LMK10], [WSB10], [WT03]), auch etwa im Vergleich zwischen Deutschland, der Schweiz und Österreich ([AWL12], [AWL13]). Dies beruht auf der zunehmenden Durchdringung des Verwaltungshandelns mit Informationstechnologie und der damit verbundenen Vernetzung ([Sc08], [SP06]). In der Schweiz entwickelte sich ausgehend von der Gründung von www.ech.ch (Verein zur Förderung von E-Government-Standards) auch eine Standardisierungs-Fachgruppe zum Geschäftsprozessmanagement (eCH-Fachgruppe GPM[1]). Ausgehend von der Leitung der Fachgruppe, die beim Informatiksteuerungsorgan des Bundes ISB angesiedelt ist, sowie der Zusammensetzung der Fachgruppe aus BPM-Beratern, -Werkzeuganbietern, -Verwaltungspraktikern und Mitgliedern von Hochschulen verläuft die Diskussion aktuell dahingehend, dass die bereits bestehenden Hilfsmittel und Standards zum BPM systematisch weiter entwickelt und ergänzt werden. Dabei konnte ein Bedarf in Richtung BPM-Maturitätsmodell, -Organisationshandbuch und -Modellierungskonventionen, Social Media und Front Office festgestellt werden. Dieser Bedarf wird über entsprechende (Teil-)Arbeitsgruppen nun systematisch gedeckt. Ebenfalls steht immer wieder die Frage im Vordergrund, wie mit den entwickelten Hilfsmitteln, Werkzeugen und Standards die Bildung einer BPM-Community in der Schweiz unterstützt werden kann, um dem BPM-Thema in der schweizerischen Verwaltung zum Durchbruch zu verhelfen. Ein erster Wurf und Prototyp einer Community- und Austauschplattform zum BPM-Thema scheiterte. Nun ist ein zweiter Anlauf im Gang.[2] Dabei handelt es sich um den Aufbau der Prozessaustauschplattform mit Community-Building-Funktion [eCoJ], das priorisierte Vorhaben B1.13 von E-Government Schweiz.[3] Als physische Plattform für den Austausch in der Community entwickelt und etabliert sich aktuell eine Konferenz (BPM@ÖV201X) zum Thema BPM. Diese bietet die Möglichkeit, sich vor Ort über

[1] Vgl. hierzu http://verwaltungsmodernisierung.ning.com/ (Aufruf per 2014-06-14).

[2] Vgl. zur neuen Plattform http://www.ech-bpm.ch/de (Aufruf per 2014-06-08).

[3] Vgl. http://www.ech.ch/vechweb/page?p=page&site=/Projekte%20ffo/Prozessaustausch plattform (Aufruf per 2014-06-08).

die neusten BPM-Entwicklungen der ÖV im In- und Ausland in Kenntnis zu setzen und auszutauschen. Ergänzend dazu und halbjährlich alternierend besteht das Tagungsformat BPM4EGOV.[4] Die Ausgaben 2013 und 2014 der BPM@ÖV boten eCH-Arbeitsgruppen eine Plattform zur Vorstellung von deren Arbeiten.[5] Eine Herausforderung für die Zukunft betrifft die Frage, in welcher Form sich das Prozessmanagement der ÖV weiter entwickelt oder gar einen Durchbruch erreicht. Es scheint, wie wenn die New Public Management Initiativen der Vergangenheit nicht zur erhofften Veränderung von funktional gegliederten Verwaltungsorganisationen in Richtung leistungsorientierter Verwaltungswertschöpfungseinheiten geführt haben [OA11]. Zu stark ist die Tendenz in der ÖV, an der traditionellen „siloorientierten" ‚Zuständigkeit' der Verwaltungsorganisation festzuhalten. Auch im Nachgang zum New Public Management und im Rahmen der E-Government-Aktivitäten ist ebenfalls keine breite Umsetzung von Prozessorientierungen und -implementierungen in der ÖV der Schweiz zu beobachten (vgl. hierzu [Sc14] und Auszüge dazu weiter unten).

1.2 Zielsetzung

Im vorliegenden Beitrag erfolgt ergänzend zu [WS11] eine kurze Aufdatierung zu aktuell in Entwicklung befindlichen eCH-BPM-Standards und -Hilfsmittel. Ferner geht der vorliegende Beitrag der Frage nach, in welcher Beziehung der standardisierte Entwicklungsprozess zu Hilfsmitteln und Standards von eCH zu relevanten wissenschaftstheoretischen Positionen steht.

1.3 Methodisches Vorgehen zur Evaluation der eCH-BPM-Standardisierungsplattform

Die Fachgruppe eCH für das Geschäftsprrozessmanagement (GPM) stellt eine Expertengruppe dar, welche im Sinne des Prinzips PPP/Public-Private-Partnership funktioniert. Die Fachgruppe ist aktuell Schirmorganisation mehrerer Arbeitsgruppen, die sich mit der konkreten Arbeit an Hilfsmitteln und Standards beschäftigen. Dabei stellen Experten von Anbietern, Beratern, Verwaltungsmitarbeitern und Hochschulbeteiligten die Community dar, in welcher einerseits Entwicklungen von Artefakten wie Standards und Hilfsmittel sowie deren Diskussion und Vorverabschiedung zuhanden öffentlicher Anhörungen erfolgen. Die Gremien von eCH erhalten eine weitere Möglichkeit zu reviewen und last but not least erfolgt eine

[4] Vgl. zu Letzterem www.bpm4egov.ch (Aufruf per 2014-06-07).

[5] Vgl. zu BPM@ÖV2013 und BPM@ÖV2014 folgende Links:
BPM@ÖV2013 Anwendertag: www.wirtschaft.bfh.ch/de/forschung/e_government_institut/veranstaltungen/bpmoev2013_anwendertag.html (Aufruf per 2014-04-07) mit eCH-Fachgruppenbeteiligung.
BPM@ÖV2013 International Day mit wissenschaftlichem Track: http://www.wirtschaft.bfh.ch/de/forschung/e_government_institut/veranstaltungen/bpmoev2013_wissenschaftliche_konferenz.html (Aufruf per 2014-04-07)
BPM@ÖV2014 Anwendertag: http://www.wirtschaft.bfh.ch/de/forschung/e_government_institut/veranstaltungen/bpmoev2014_swiss_day.html (Aufruf per 2014-04-07).

Anhörung innerhalb der gesamten eCH-Community.[6] Aus methodisch-wissenschaftlicher Sicht gelangt im Rahmen der entsprechenden Standardisierungen ein Mix aus folgenden theoretischen Ansätzen zum Einsatz: Design Science Ansatz, Method Engineering ([HMP04], [PTR07]) und Wahrheitstheorien ([Fr07], [La07]). Diese Ansätze werden im vorliegenden Beitrag zur Evaluation des aktuellen Standardisierungsvorgehens eingesetzt, ebenso wie rare Hinweise aus empirischen Untersuchungen. Im Folgenden wird vertiefter auf die Zusammenhänge zwischen wissenschaftlich-theoretischen Konstrukten und deren Anwendbarkeit innerhalb des BPM-Standardisierungssystems eingegangen.

2 Theoretische Grundlagen zum BPM in der Öffentlichen Verwaltung

Die Autoren Dunleavy et al. (2005) [DM05] diskutieren generell die Verwaltungstätigkeitsveränderung basierend auf IT-Technologieeinführung in der ÖV und der erforderlichen Veränderung zur Governance. Die Autoren Gulledge/Sommer (2002) [GS02] diskutieren die radikalen erforderlichen Veränderungen, welche durch die Geschäftsprozessorientierung in der ÖV erforderlich sind. Die Gründe für eine Implementierung des BPM in der Öff. Verwaltung lauten in Anlehnung an [GS02] wie folgt: Dokumentierung zum Verständnis von Prozessen, Accountability dank Prozesszuständigkeiten, Managen der Prozess-Performance, Verbesserung der Produktequalität durch Prozessimplementierung und damit Qualitäts- und Performance Management. [Kl04] meint betreffend BPM in der ÖV das Folgende: „Comparing the frameworks and guidelines provided by central European and U.S. governmental units, we find information integration and process integration as prominent concepts to guide interoperability efforts, but they seem to point to different directions“. [On04] merkt zum BPM in der ÖV an: „Process management is a managerial approach characterised by the focus on business processes that can provide an important contribution to the management of public sector organisations“. Im Übrigen gilt für seinen Beitrag: „A model for the analysis of the factors enabling the implementation of process management in the public sector is proposed. [We] […] employ process management for the effective management of inter-organisational relations in the public sector”. Laut [Wi02] last sich Prozessmanagement als eine mögliche Teildimension in einem dreidimensionalen Framework verstehen: Progress of a Public Service (Public Services, Processes and Workflow), Abstraction Layers, Different Views (Process Perspective) zur Realisierung eines online one-stop-governments. Nach [Wi02] gilt im Übrigen: „Traunmüller and Lenk [TL00] suggest a differentiation of four perspectives to understand concepts and processes of public administrations: citizens and customers, process (reorganization), (tele)cooperation and knowledge“. [PW03] stellen fest: „Recently, process modelling and process reorganisation have been recognised as being of utmost importance for making e-government implementations success. Due to the high complexity of governmental

[6] Vgl. zum eCH-Vorgehen der Standardisierung eCH-0001 [eCH01] eCH-Standardisierungsprozess: www.ech.ch sowie eCH-0003 [eCH03].

processes and organisational structures, appropriate modelling methodologies and tools are, however, not really available yet. In our contribution, we describe the needs for a comprehensive Business Process Management methodology and toolkit targeted for the public sector“. [PTG09] thematisieren verschiedene Ansätze des BPM in der ÖV um ihr eigenes Domänenmodell zu entwickeln (vgl. daran angelehnt auch [Wa13a], [Wa13b]), das Government Enterprise Architecture Model GEA. [SJ07] thematisieren das Thema Change Management in Relation zum BPM im E-Government-Kontext wie folgt: „E-government is becoming extremely important, however, it cannot be successfully implemented without changes in business processes that are performed inside governmental institutions. Public sector has some specifics, which make renovation projects considerably different. This makes it clear, that radical changes in the execution of business processes and in the organizational structures are not suitable because they are not possible for many reasons. Therefore, classic business process change methodologies have to be adapted“. [BAN03] fassen ihre Schlussfolgerungen zum BPM in der ÖV wie folgt zusammen: „Process oriented analysis and optimisation of administrative procedures are key prerequisites for the successful organisational and technical restructuring of municipal administration in the move to Electronic Government. The exploitation of the full potential of information and communications technology can only be achieved through structured processes. The complexity of process models requires both a systematic preparation, and a methodical approach to the implementation of process oriented E-Government projects. Therefore this article provides a procedure model for process oriented organisation design, underlined by a case study which describes an optimisation project of the building permission procedure in the German municipality of Emsdetten“. Und überdies kommen [BAN06] zu folgendem Schluss: “There is a lack of process orientation in public administrations. Additionally, existing processes are regularly not applicable to e-government. Therefore, e-government projects in practice are not always able to fully implement transactional processes. One of the main reasons for the lack of process orientation is that there are few BPR methodologies applied and verified in public administrations“. Interessante Einblicke in unterschiedliche Prozesscharakteristika und deren Implikationen präsentieren [TRG06]. Zusammenfassend können die verschiedenen Positionen aus der Literatur wie folgt kommentiert werden: Zur Standardisierung ist wenig zu finden. Die Positionen in der Literatur beziehen sich hauptsächlich auf das Change Management, Enabler Faktoren für die Einführung des BPM in ÖV, etc. Verschiedene Positionen der Literatur thematisieren auch die Problematik der Breite der Aufgaben und insbesondere auch die föderale Struktur für interoperable Prozesse in der ÖV, die das BPM erschweren. Außerdem wird konstatiert, dass ein Mangel an Prozessorientierung vorherrsche. Zudem ist davon auszugehen, dass die angelsächsische Community und Verwaltung einen stärker Management-orientierten Ansatz in Anlehnung an das NPM-Paradigma vertritt als die übrige Community und Verwaltung. Im zuletzt kokmmentierten Bereich wird thematisiert, dass Prozessmanagement Chefsache sei. Forschungsfragen lauten damit wie folgt: Was bringen Standards im E-Government in föderalen Konstellationen bezüglich BPM? Wie kann die Erarbeitung von Standards unterstützt und systematisch angegangen werden. Lassen sich die in der Literatur thematisierten BPM-Probleme der ÖV konstruktiv lösen? Gibt es einen theoretischen Unterbau, über welchen das BPM in der ÖV fundiert werden kann.

3 Wissenschaftliche Fundierung des BPM-Standardisierungssystems in der Öffentlichen Verwaltung der Schweiz

3.1 Theoriebasierte Fundierung des BPM-Standardisierungssystems

Im Folgenden wird zunächst die methodische Fundierung konkretisiert. Ansätze, die zur Fundierung herangezogen werden, sind der Design-Science-Ansatz (bezieht sich auf das eigentliche Erarbeiten der Standards) ([HMP04]; [PTR07]) sowie die Begründung mittels Wahrheitstheorien ([Fr07]; [La07]). Zu Letzteren gehören Korrespondenztheorie, Kohärenztheorie sowie Konsenstheorie. Bei der Entstehung und Entwicklung des BPM-Standardisierungssystems wird auf der Basis des Ansatzes von [PTR07] vorgegangen. Deren methodisches Vorgehen sieht sechs Schritte sowie Iterationen vor, um dadurch (mehr) Rigorosität in den Problemlösungs-orientierten Entwicklungsprozess von Artefakten zu bringen. Dieses sieht sechs Schritte vor: 1. Identify problem and motivation; 2. Define objectives of a solution, what would a better artifact accomplish?; 3. Design and develop artifact; 4. Demonstration, find suitable context, use artifact to solve problem; 5. Evaluation, observe how efficient and effective the design is; iterate back; 6. Communication, scholarly publications, professional publication. Dieser Beitrag beurteilt nicht selber die in eCH entwickelten Artefakte. Dies ist hinsichtlich deren Verabschiedung durch die eCH-Gremien und die BPM-Community (hauptsächlich in der Schweiz) erfolgt. Der zum Teil breit intendierte Einsatz von eCH-Standards nimmt den zweiten Punkt vorweg. Die Artefakte wurden und werden entwickelt und zur breiten Diskussion erst innerhalb von eCH und danach in der entsprechenden schweizerischen BPM-Community bereitgestellt. Ebenfalls erfolg(t)en Vorstellungen und Kommentierungen an entsprechenden Konferenztracks (BPM@ÖV201X). Es ist das Ziel, dass die Standards in der Praxis Nutzen stiften. Erweiterte empirische Evidenz dafür ist noch zu erarbeiten. In zwei studentischen Arbeiten, welche durch den Autor betreut wurden, wurde z.B. der eCH-Standard eCH-0158 Modellierungskonventionen für die ÖV für KMU's angepasst und wird dort eingesetzt. Ausgehend von der Korrespondenztheorie lassen sich folgende Kriterien für die Rechtfertigung der Gültigkeit anführen: Übereinstimmung mit der (wahrgenommenen) Realität bzw. Erfahrung, Empirische Bestätigung, Nachweis der erfolgreichen Umsetzung oder Implementierung in der Praxis z.B. anhand von qualitativen und quantitativen empirischen Untersuchungen. Eine Wertung sieht ausgehend vom Begründungszusammenhang BPM-Standardisierungssystem wie folgt aus. Eine Übereinstimmung mit der Realität kann dadurch sichergestellt werden, dass sich die eCH-Fachgruppe GPM aus verschiedensten Personen mit den folgenden Erfahrungen zusammensetzt: BPM-Forschungs-, -Schulungs-, -Beratungs- oder -– Umsetzungserfahrung. Diese unterschiedlichen Perspektiven fließen in die eCH-Fach- und -Arbeitsgruppen sowie deren Arbeit an Artefakten (Standards und Hilfsmittel) ein. Eine empirische Bestätigung ergibt sich wie folgt: Es ist noch unsicher, ob das BPM in der ÖV einen effektiven Durchbruch erreicht. Indes erfahren die Standards z.B. bei Anbietern und Beratern zunehmend Beachtung (das zeigt sich u.a. dadurch, dass sie in der Standardisierungsgruppe mitwirken). Der Nachweis der erfolgreichen Umsetzung oder Implementierung in der Praxis kann analog zum Design-Science-

Ansatz durch Präsentationen an Konferenzen und Veranstaltungen bestätigt werden. Dies erfolgt sowohl für die Standard-Dokumente von eCH sowie für deren Anwendung (vgl. Hinweise zu BPM@ÖV2013 und BPM@ÖV2014 weiter oben). Ausgehend von der Kohärenztheorie lassen sich folgende Kriterien für die Rechtfertigung anführen: Erforderlich ist die Stimmigkeit des BPM-Standardisierungssystems im Rahmen der Theorie. Es ist mindestens eine widerspruchsfreie Einbettung in anerkanntes Wissen zum BPM, dessen Organisation und technische Umsetzung sicherzustellen. Die Stimmigkeit mit bzw. die Verwendung von bestehenden Konzepten (z.B. aus Verwaltungswissenschaft/New Public Management [SP06]) ist erforderlich. Ferner ist eine Wertung ausgehend vom in Entwicklung befindlichen BPM-Standardisierungssystem anzugehen. Es ist eine Stimmigkeit bezüglich bestehender Konzepte oder Theorien erforderlich. Bislang fehlt eine weithin anerkannte konzeptionelle Grundlage in Form eines holistischen und praktisch gelebten Verwaltungs-Referenzprozessmodells (etwa analog zu [Wa12]). Jedoch existieren aus schweizerischer Sicht partielle Architekturreferenzmodelle etwa für das E-Government (eCH-0122 bis eCH-0125: E-Government Architekturen Vertrieb, Produktion, Kommunikation). Dafür bildet das Dokument eCH-0126 zusammen mit der E-Government-Strategie der Schweiz (von 2007) die auf die Leistungsverwaltung beschränkte Grundlage für die BPM-Positionierung in der Schweiz. Ausgehend von der Konsenstheorie lassen sich folgende Kriterien für die Rechtfertigung der Gültigkeit anführen: Ein (u.a. empirischer) Nachweis der breiten Zustimmung für die aufgestellten Theorien/Modelle/Artefakte in der jeweiligen Forschungs- und Anwendungs-Community ist erforderlich: Im vorliegenden Fall die eCH-BPM-Community. Ein Konsens wird im Kreis der anerkannten Experten in der Standardisierungsarbeit sowie in der Anhörung sichergestellt (Vgl. [eCH01]; [eCH03]). Durch die Bezugnahme auf vorhandene und verwandte Ansätze wird zum einen – gemäß der Kohärenztheorie der Wahrheit – eine möglichst weitgehende widerspruchsfreie Einbettung des BPM-Standardisierungssystems in anerkanntes Wissen angestrebt. Zum anderen wird dadurch ein – zumindest in Teilen vorhandener – Konsens unterstellt (Konsenstheorie der Wahrheit). Durch die verschiedenen eCH-internen und -externen Anhörungen wird (teilweise und je nach kommentierendem Publikum) Praktikabilität sichergestellt. Wo möglich soll empirische Evidenz herangezogen werden (vgl. dazu [FT12], [HFL10], [OA11], [Sc14]).

3.2 eCH-Standardisierungsvorgehen und dessen Evaluation auf Basis der erwähnten Theorien

Der erwähnte eCH-Standardisierungsprozess gemäß [eCH01] und [eCH03] sieht wie folgt aus: Der Expertenausschuss verabschiedet eCH-Standards. Der Vorstand verabschiedet den eCH-Standardisierungsprozess. Der oben erwähnte und im Folgenden dargestellte Standardisierungsprozess von eCH umfasst den gesamten Lebenszyklus, d.h. von der Anmeldung bis zum Rückzug eines Standards, mit allen notwendigen Teilprozessen, Zwischenschritten und Reviews. Er wird dabei als geradliniger Prozess beschrieben. Dieser richtet sich an vergleichbaren Prozessen im internationalen Umfeld aus (z.B. an Internet Engineering Taskforce (IETF)). Mit der Standardisierung dieses Prozesses werden folgende Ziele erreicht: Der Standardisie-

rungsprozess konzentriert sich auf Standards zur Unterstützung von eGovernment durch Informations- und Kommunikationstechnologien (IKT). Neben bereits identifizierten Standards können auch Themen zur Standardisierung beantragt werden (Initialisierung von Fachgruppen). Ein effizienter Ablauf der Standardisierung wird unterstützt. Die resultierenden Standards sind offen und zeitgerecht. Sie berücksichtigen bereits vorhandene nationale und internationale Standards. Die Kriterien für Annahme, Verabschiedung und Rückzug eines Standards sind geregelt. Der Prozess der Standardisierung sieht wie in der Abbildung 1 dargestellt aus und umfasst die Prozessschritte wie folgt: Kickoff/Erarbeitung, Formale Qualitätssicherung 1, Inhaltliche Qualitätssicherung 1, Öffentliche Konsultation, Einarbeitung und Rückmeldungen, Formals Qualitätssicherung 1, Inhaltliche Qualitätssicherung 2, Genehmigung-/Ablehnung, Publikation und (regelmäßige) Überprüfung.

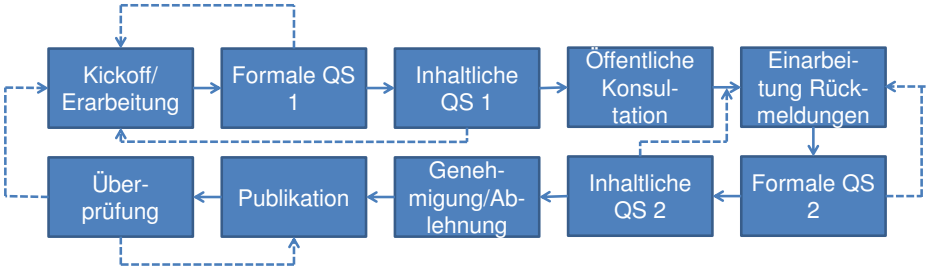


Abbildung 1: eCH-Standardisierungsprozess nach eCH-0003 [eCH03].

Ausgehend von den unter Kapitel 3.1 dargestellten theoretischen Bausteinen zur Fundierung des BPM-Standardisierungssystems der Schweiz wird im Folgenden auf die möglichen Bezüge zwischen dem Standardisierungsprozess und den theoretischen Bausteinen zur Evaluation des BPM-Standardisierungssystems eingegangen. Design-Science-Ansatz: Es werden die Schritte gemäß [PTR07] angegeben und aufgeführt und unmittelbar danach wird entsprechend kommentiert. 1. Identify problem and motivation: Rückblickend betrachtet steht hinter jedem der erwähnten Standards zum BPM von eCH eine Motivation und ein Problembereich. Z.B. im Bereich Modellierungskonventionen die Angabe eines reduzierten Shape-Sets und die Konkretisierung von Do's und Dont's im BPMN-Modellierungskontext; im Bereich der Maturitätsmessung, dass ein standardisiertes Vorgehen/Werkzeug für die Messung der Prozess- und Organisationsmaturität bezüglich BPM besteht, um Baselines zu definieren, basierend auf denen im Sinne der laufenden Verbesserung Verbesserungsoptionen gemanaged werden können. 2. Define objectives of a solution, what would a better artifact accomplish?: Vgl. hierzu den ersten Punkt zur Problemdefinition. 3. Design and develop artifact: Dies erfolgt in Abbildung 1 im Prozessschritt Erarbeitung. 4. Demonstration, find suitable context, use artifact to solve problem: Dies ist im eCH-Standardisierungsprozess nicht explizit vorgesehen, allerdings können die (inhaltlichen und formalen) Qualitätsssicherungen 1 und 2 sowie die öffentliche Konsultation dazu beitragen, dass eine Problemlösung effektiv adressiert wird. Dies muss zugegebenermassen aber nicht so sein. Es besteht die Gefahr, dass prophylaktisch Standards entwickelt werden zu einem Problem, das noch nicht besteht, da in der Regel die Problemlösungseffizienz des Standards/Hilfsmittels

erst in der eigentlichen Problemlösung (praktischen Nutzung) resultiert. 5. Evaluation, observe how efficient and effective the design is; iterate back: Die Evaluationsmöglichkeit ist im Rahmen der Qualitätssicherungen und Konsultationen gegeben. Iterationen sind in Abbildung 1 ebenfalls ersichtlich. Zudem besteht im Rahmen der Nutzung des Standards die Möglichkeit, bei eCH Requests for Changes einzugeben. Damit können die Standards und Hilfsmittel laufend verbessert werden. 6. Communication, scholarly publications, professional publication: Publikationen von Standards, Hilfsmittel, etc. zum BPM in der ÖV und im E-Government erfolgen über die Webseite www.ech.ch, auch was die Konsultation betrifft und Schlussfolgerungen in Artikeln wie dem Vorliegenden. Ausgehend von den unter 3.1. gemachten Überlegungen zu Korrespondenz-, Kohärenz- sowie Konsenstheorien können hier ausgehend von der Präsentation des Prozessmodells zur Standardisierung basierend auf [eCH03] im Folgenden weitergehende Evaluationen gemacht werden: Korrespondenztheorie – Übereinstimmung mit wahrgenommener Realität: Dies kann im Standardisierungsprozesse nach [eCH03] am ehesten über die Konsultationen erfolgen, aber auch über die eigentliche Nutzung, die zu Change Requests führen kann. Ansonsten gilt, was weiter oben zum Bezug zur Korrespondenztheorie erwähnt ist. Kohärenztheorie – Widerspruchsfreie Einbettung in anerkanntes Wissen: Vgl. hierzu die Theoriezusammenfassung in Kapitel 2. Es bleibt zu evaluieren, was ausgehend davon Widerspruchsfreiheit sein kann. Ansonsten gilt, was oben zum Bezug zur Kohärenztheorie erwähnt ist. Konsenstheorie – Nachweis breiter Zustimmung. Auch dies ist am ehesten über die Konsultationen abrufbar. Überdies erweist sich über die Nutzung in praktischen Kontexten, ob die Zustimmung groß ist. Ansonsten gilt, was weiter oben zum Bezug zur Konsenstheorie erwähnt ist. Ein Konsens muss überdies auch erreicht werden über die Zusammenarbeit in den entsprechenden Fachgruppen. Dass es überhaupt zur Dokumentation von Standards und Hilfsmittel kommt hat mit dem Konsens in der entsprechenden Gruppe zu tun. Die weitergehende Konsenserhebung ist über die Konsultationen leistbar. Zusammenfassend kann wie folgt gewertet werden. Die Anwendung eines Standards hängt von verschiedenen Faktoren ab. Einerseits muss für das Problem ein Bewusstsein vorhanden sein, damit eine Problemlösung in Form eines BPM-Standards/Hilfsmittels angenommen wird. Andererseits muss in einer Verwaltungsorganisation eine gewisse Maturität vorhanden sein, dass erkannt wird, dass Prozessmanagement in einer Verwaltung ein Thema ist. Dies wiederum führt zur Frage der Führung. Die Führung wiederum führt zur Frage, ob klare Zielsetzungen zur Implementierung und zum Management von Geschäftsprozessen vorhanden sind. Im Gegensatz zur theoretischen Schlusswertung zeigen wenige Auszüge aus einer empirischen Studie [Sc14] geringe Evidenz für eine breite Durchsetzung des Prozessmanagements in der ÖV der Schweiz (Kommunen). In der Studie zum BPM in allen deutsch- und rätoromanisch-sprachigen Schweizer Gemeinden wird ersichtlich, bei einem n von 191, dass eine sehr geringe Bekanntheit von Standards wie z.B. von eCH Tatsache ist. Zwar wird das Prozessmanagement in der Umfrage von [Sc14] durch die Kommunen als wichtig bezeichnet, alleine die Umsetzung des Prozessmanagements scheint marginal. 67% von n = 191 bezeichnen BPM als wichtig. Bei 20% von n = 191 sind BPM-Projekte umgesetzt, bei 7% sind entsprechende BPM-Projekte in Planung. 7% von n = 191 setzen umfassend Prozessmanagement ein, bei 43% werden einzelne Prozesse grafisch modelliert.

6 Zur Umsetzungsproblematik des Geschäftsprozessmanagements in der Öffentlichen Verwaltung aus Schweizer Sicht

Wie bereits erwähnt hat sich das Paradigma des New Public Managements nicht oder nur partiell durchgesetzt. Eine Idee daraus war die Strukturierung der Verwaltung auf Basis von Verwaltungsleistungen und auf den Kunden hin. Dies hätte zur Folge haben müssen, dass die Prozessorientierung und -organisation gegenüber der Aufbauorganisation einen Primat hätte erhalten sollen. Bekanntlich ist diese Entwicklung vor allem deshalb ausgeblieben, weil die Verwaltung Ressort-orientiert arbeitet und weil die Zuständigkeit nach wie vor dominant im Vordergrund steht (Bürokratiemodell) ([OA12], [SP06]). Mit der Einführung/Diskussion von E-Government ist diese Diskussion wieder aufgeflammt. Die Durchsetzung des Primats der Prozessorganisation und -orientierng vor der Aufbauorganisation hat es auch im Rahmen der E-Government-Diskussion schwer. Entsprechend ist die flächendeckende Umsetzung der Prozessorientierung kein Thema. Vielmehr sind es wenige z.T. kleine Leuchtturmprojekte, welche erwähnt werden können (vgl. hierzu auch [WS11]). Neuerdings ergeben sich neue Ansatzpunkte, wie mit der Prozessorientierung erneut vorwärts gemacht werden kann, und zwar über die Aufgabenorientierung, die im HRM2 – Harmonisiertes Rechnungsmodell 2 begründet liegt ([KDK07]). Die Konten-Nummerierung wurde mit dem (funktionalen) Aufgabenraster der Verwaltungen gekoppelt. Dies erlaubt, dass ausgehend vom Kontenplan die Leistungen an den Aufgaben festgemacht werden können, was wiederum eine Kopplung mit Prozessen ermöglicht. Aus der BPM-Standardisierungsplattform lassen sich im Sinne einer zunehmend bewussteren Gestaltung desselben folgende Verbesserungs- und Optimierungsmöglichkeiten erwähnen: Es ist ein Koordinationskonzept zu den verschiedensten Community-Plattformen erforderlich. Die Plattform NING zur Verwaltungsmodernisierung kann wohl abgelöst werden, wenn die Prozessaustauschplattform steht. Erforderlich ist wohl auch gesetzgeberischer Input, damit das Paradigma der Kundenorientierung mehr Bedeutung vor der funktionalen Organisation der Verwaltung erhält. Die Vervollständigung des Set's an Maßnahmen inklusive Bereitstellung der Prozessaustauschplattform kann (muss aber nicht) zu einem weiteren großen BPM-Entwicklungsschritt in der ÖV führen. Ein zentrales Teil fehlt in den eCH-Dokumenten weitgehend, und zwar die Leistungsart, welche dem Kunden dabei behilflich ist, sein Leistungsbündel zusammenzustellen und die entsprechenden Aufträge in die unterschiedlichen Verwaltungssilos zu triggern.

7 Zusammenfassung und Ausblick

Im vorliegenden Beitrag erfolgt der Versuch, das Vorgehen zur Standard-Artefakteentwicklung nach eCH einer wissenschaftlichen Fundierung zu unterziehen. Überdies leistet der Beitrag einen Kurzbeschrieb des aktuell impliziten BPM-Standardisierungssystems der ÖV in der Schweiz, wodurch dieses explizit gemacht wird. Weiter werden verschiedeneTheoriebausteine zur Entwicklung des BPM-Standardisierungssystems (Darstellung) und zur Begründung von dessen Zweckmäßigkeit zur Anwen-

dung gebracht. Im Weiteren folgt eine Diskussion der Verbesserungen der Prozessorientierung auf Basis aktueller Entwicklungen und ausgehend vom BPM-Standardisierungssystem und dessen wissenschaftlicher Fundierung. Im Weiteren zeigt dieser Beitrag, dass aus evidenzorientierter empirischer Sicht noch immer eine geringe BPM-Verbreitung im kommunalen Bereich in der Schweiz vorhanden ist. Ein weiteres Thema stellen im vorliegenden Beitrag überdies Klarstellungen zur aktiven Gestaltbarkeit der BPM-Standardisierungssystems im Hinblick auf mehr Wirksamkeit und verbesserter Umsetzung des BPM in der ÖV dar. Im Sinne eines Ausblicks ist weitere Forschungsarbeit zur Verifizierung eines Nutzens eines BPM-Standardisierungssystems für die ÖV erforderlich und es können künftig auf Basis der geschilderten Darstellung Entwicklungsideen konkreter gestaltet werden. Im Sinne der thematisierten externen Evidenz ist überdies weiter rigorose empirische Forschung zu Erfolg und Misserfolg der Prozessorientierung und -Implementierung auf Basis von eCH-Standards und auf Basis des BPM-Standardisierungssystems in der Schweiz zu betreiben. Überdies ist u.a. auch den Gründen für die aktuell schwache Prozess(management)orientierung in der ÖV Schweiz nachzugehen. Es sind Erfolgsfaktoren für die positive Entwicklung desselben auszuloten. Überdies ist die Diskussion darüber weiter zu führen, welche Möglichkeiten sich anhand des BPM-Standardisierungssystems zur Verbreiterung des BPM-Ansatzes in der ÖV ergeben.

Literaturverzeichnis

[AIP10] Algermissen, L.; Instinsky, M.; Pähler vor der Holte, N. (2010): Prozessmanagement im Zeichen von Dienstleistungen: Eine methodische Unterstützung zur Umsetzung der Dienstleistungsrichtlinie, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 65-76.

[AWL12] Ahrend N., Walser K., Leopold H. (2012): Comparative Analysis of the Implementation of Business Process Management in Public Administration in Germany and Switzerland. In: Proceedings der PoEM 2012 - The 5th IFIP WG8.1 Working Conference on the Practice of Enterprise Modelling, Rostock

[AWL13] Ahrend, N.; Walser, K.; Leopold, H. (2013): Case Study of the Implementation of Business Process Management in Public Administration in Germany, Switzerland and Austria. Erscheint in Proceedings ECEG2013 vom 13./14. Juni 2013, Como.

[BAN03] Becker, J.; Algermissen, L.; Niehaves, B. (2003): Processes in e-Government Focus: A Procedure Model for Process Oriented Reorganisation in Public Administrations on the Local Level. In: Electronic Government – Lecture Notes in Computer Science 2739 (2003), S. 147-150.

[BAN06] Becker, J.; Algermissen, L.; Niehaves, B.: A procedure model for process oriented e-government projects. In: Business Process Management Journal 12 (2006) 1 S. 61-75.

[Be03] Behjat, S. (2003): Wertschöpfungsprozesse der Öffentlichen Verwaltungen als Grundlage von e-Government. deposit.d-nb.de/cgi-bin/dokserv?idn=970740263 &dok_ var=d1&dok_e xt=pdf&filename=970740263.pdf (Aufruf per 2009-08-17; erstellt per 2003).

[BEK10] Brüggemeier, M.; Ecker, K.-P.; Knopp, M.; Schilling, P.; Steffens, P.; Tschichholz, M. (2010): FRESKO – die effiziente Prozessketten-Verbindung zwischen Unternehmen und Verwaltungen, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 40-52.

[BS10] Brüggemeier, M.; Schulz, S. (2010): Datenpointernetzwerk: Informationsintegration für eine vernetzt arbeitende, transparentere und weniger spürbare Verwaltung der Zukunft, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 17-28.

[DM05] Dunleavy, P.; Margetts, H.; New Public Management Is Dead—Long Live Digital-Era Governance. In: JPART 16 (2005), S. 467-494.

[eCH01] eCH-0001: eCH-Standardisierungsprozess. www.ech.ch/alfresco/guestDownload/attach/workspace/SpacesStore/f686f8f1-a7dc-4760-a96f-5f91b0d82a76/STAN_d_DEF_2005-10-04_eCH-0001%20Standardisierungsprozess.pdf (Aufruf per 2014-06-08).

[eCH03] eCH-0003: Leitfaden zur Genehmigung von Anträgen. www.ech.ch/alfresco/guestDownload/attach/workspace/SpacesStore/62e65721-032c-40d7-9de6-991e22d27b0ca/STAN_d_DEF_2014-06-06_eCH-0003_V9%201_Leitfaden%20zur%20Genehmigung%20von%20Anträgen.pdf (Aufruf per 2014-06-08).

[eCoJ] eCH (ohne Jahr): eCH-Prozessaustauschplattform für Gemeinden und Kantone. www.ech.ch/alfresco/guestDownload/attach/workspace/SpacesStore/951c4b82-734f-4f83-891a-c437ec51d6e2/Kurzbeschrieb_B1_13.pdf (Aufruf per 2013-05-05).

[Fr07] Frank U (2007): Ein Vorschlag zur Konfiguration von Forschungsmethoden in der Wirtschaftsinformatik. In: Lehner F, Zelewski S (Hrsg): Wissenschaftstheoretische Fundierung und wissenschaftliche Orientierung der Wirtschaftsinformatik. GITO, Berlin, S. 158–185.

[FT12] Fischer, M.; Telesko, R. (2012): Geschäftsprozessmanagement in der Öffentlichen Verwaltung – Eine Umfrage zum aktuellen Stand. In: eGov Präsenz (2012) 1, S. 44-45.

[GH10] Guthier, T.; Hünemohr, H. (2010): eBundesrat: Referenzprozess für die elektronische Vorgangsbearbeitung in Bundesratsangelegenheiten, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 88-96.

[GS02] Gulledge, T.R.; Sommer, R.A.; Business Process Management: public sector implications. In: Business Process Management Journal 8 (2002) 4, S. 364-376.

[Ha05] Hach, H. (2005): Evaluation und Optimierung kommunaler E-Government-Prozeße. www.zhb-flensburg.de/dißert/hach/dißertation-hhach-veroeffentlichung.pdf (Aufruf per 2008-10-24; erstellt per 2006).

[HFL10] Houy, C.; Fettke, P.; Loos, P. (2010): Empirical research in business process management – analysis of an emerging field of research. In: Business Process Management Journal 16 (2010) 4, S. 619-661.

[HMP04] Hevner, A. R.; March, S. T.; Park, J.: Design Science in Information Systems Research, MIS Quarterly, Vol. 28, No. 1, March 2004, S. 75-105.

[KdK07] KdKF (2007): Handbuch Harmonisiertes Rechnungslegungsmodell für die Kantone und Gemeinden – HRM2, Konferenz der Kantone – Finanzdirektoren. www.frischisuhr.ch/doc uments/232-handbuch_hrm_d.pdf (Aufruf per 2013-05-05; erstellt per 2007-01-19).

[Kl04] Klischewski, R.: Information Integration or Process Integration? How to Achieve Interoperability in Administration. In: Electronic Government – Lecture Notes in Computer Science 3183 (2004), S. 57-65.

[La07] Lange, C. (2007): Ein Bezugsrahmen zur Beschreibung von Forschungsgegenständen und -methoden in Wirtschaftsinformatik und Information Systems – ICB Research Report. www.wi-inf.uni-due.de/FGFrank/documents/Arbeitsberichte_ICB/No1.pdf (Aufruf per 2013-03-31).

[Li05] Linder, W. (2005): Schweizerische Demokratie – Institutionen, Prozesse, Perspektiven. Haupt, Bern.

[LMK10] Licker, J.; Mayer, A.; Kaiser, S. (2010): Modulare Einführung für eine bedarfsorientierte Unterstützung der Vorgangsbearbeitung, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeins. Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 77-87.

[LSS10] Lenk, K.; Schuppan, T.; Schaffroth, M. (2010): Organisationskonzept für das E-Government Schweiz, noch nicht veröffentlichtes White Paper des Informatikstrategieorgans des Bundes, 2010.

[OA11] Ohne Autor (2011): Umfrage BPM in der Öffentlichen Verwaltung. www.bpm4egov.ch/fileadmin/dowloads/6.Netzwerktreffen-Sept11/2011_BPM_Umfrage_Bericht_Auswertung. pdf (Aufruf per 2013-05-12; erstellt per 2011-09).

[OA12] Ohne Autor (2012): Debate: New Public Management. In: Swiss Political Science Review 1 (2012) 4, S. 143-166.

[On04] Ongaro, E.: Process management in the public sector - The experience of one-stop shops in Italy. In: The International Journal of Public Sector Management 17 (2004) 1, S. 81-107.

[PTG09] Peristeras, V.; Tarabanis, K.; Goudos, S.K.: Model-driven eGovernment interoperability: A review of the state of the art. In: Computer Standards & Interfaces 31 (2009), S. 613-628.

[PTR07] Peffers, K.; Tuunanen, T.; Rothenberger, M.S.; Chatterjee, S. (2007/2008): A Design Science Research Methodology for Information Systems Research. In: Journal of Management Information System 24 (2007/2008) 3, S. 45-77.

[PW03] Palkovits, S.; Wimmer, M.A.: Processes in e-Government – A Holistic Framework for Modelling Electronic Public Services. In: Electronic Government - Lecture Notes in Computer Science2739 (2003), S. 213-219.

[Sc08] Schaffroth, M. (2008): Interoperabilität und Geschäftsprozessmanagement im E-Government, in: eGov Präsenz (2008) 2, S. 46-49.

[Sc14] Schlatter, U. (2014): Prozessmanagement in den Gemeinden der deutsch- und rätoromanisch-sprachigen Schweiz. www.igict.ch/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/user_upload/ig_ict/content/Auswertung_zur_IWI-Gemeindeumfrage_2014.pdf&T=1404908347&hash=1642318f59971c0673047786cbdad8539aee628e (Aufruf per 2014-06-07).

[SJ07] Stemberger, M.I.; Jaklic, J.: Towards E-government by business process change - A methodology for public sector. In: International Journal of Information Management 27 (2007), S. 221–232

[SP06] Schedler, K.; Proeller, I. (2006): New Public Management, Haupt/UTB, Bern.

[TL00] Traunmüller, R.; Lenk, K. (2000): e-Commerce as a stimulus for e-Government. In: Proceedings of the XIII Bled Conference on Electronic Commerce: The End of the Beginning, Bled (2000).

[TRG06] Tochtermann, K.; Reisinger, D.; Granitzer, M.; Lindstaedt, S.N. (2006): Integrating ad hoc processes and standard processes in public administrations. In: Wimmer, M.

A.; Traunmüller, R. (Hrsg.). Knowledge transfer across Europe: 4th Eastern European e|Gov Days and 5th e|Gov Days; OCG Schriftenreihe Band 203, Wien.

[Wa12a] Walser, K. (2012a): Architectural Principles for E-Government Business and Application Architectures based on an E-Government Business Process Reference Model. In: Konjovic, Z.; Milosavljevic, B.; Markovic, M. (Eds.): MeTTeG12 – Proceedings of the 6th International Conference on Methodologies, Technologies and Tools Enabling e-Government – Belgrade, Serbia – July 3-5, 2012. University of Novi Sad, Faculty of Technical Sciences, S. 112-124.

[Wa12b] Walser, K. (2012b): The deconstruction and reconfiguration of municipalities – a better approach than municipal mergers? – The information system point of view. In: Konjovic, Z.; Milosavljevic, B.; Markovic, M. (Eds.): MeTTeG12 – Proceedings of the 6th International Conference on Methodologies, Technologies and Tools Enabling e-Government – Belgrade, Serbia, University of Novi Sad, S. 115-137

[Wi02] Wimmer, M.A.: A European perspective towards online one-stop government: the eGOV project. In: Electronic Commerce Research and Applications 1 (2002) 1, S. 92-103.

[WS11] Walser, K.; Schaffroth, M. (2011): BPM and BPMN as Integrating Concepts in eGovernment – the Swiss eGovernment BPM Ecosystem. In: Proceedings S-BPM One 2010, Karlsruhe, S. 106-120.

[WSB10] Walser, K.; Schaffroth, Bagnoud, L. (2010): Verwaltungsprozessmanagement unter Verwendung des eCH-0096 BPM Starter Kits, in: Wimmer, M.A.; Brinkhoff, U.; Kaiser, S.; Lück-Schneider, D.; Schweighofer, E.; Wiebe, A. (Hrsg.): Vernetzte IT für einen effektiven Staat – Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010 in Koblenz, GI-Edition Lecture Notes in Informatics, S. 52-64.

[WT03] Wimmer, M.; Traunmüller, R. (2003): Geschäftsprozessmodellierung in E-Government: eine Zwischenbilanz, eGov days 2003, www.egov.ocg.at (06-02-2004).

Big Data Technologien zur Analyse und Steuerung komplexer cyber-physischer Systeme

(CPSData)

Big Data Technologien zur Analyse und Steuerung komplexer cyber-physischer Systeme[1]

Abstract: Neue Kommunikationstechnologien und immer leistungsfähigere Hardware im Bereich eingebetteter Systeme haben in den letzten Jahren zu einer massiven Zunahme von vernetzten cyber-physischen Systemen (CPS) geführt. Solche aus Softwaremodulen, Netzwerk- und Kommunikations- und physischen Hardwarekomponenten bestehenden Systeme werden die Grundlage für die technologische Bewältigung vielartiger gesellschaftlicher Herausforderungen bilden. Nachhaltige Elektromobilitäts- und Verkehrskonzepte, intelligente und adaptive Stromnetze, selbstoptimierende Plattformen für die Gebäudeautomatisierung oder individualisierte und hoch-flexible Produktionsprozesse im Kontext Industrie 4.0 bezeichnen nur einige Herausforderungen. Wesentlich für die Bewältigung dieser Handlungsfelder wird neben der standardisierten nahtlosen Vernetzung die intelligente Nutzung von CPS Daten. Voraussetzung hierfür sind entsprechende Verfahren zum Datenmanagement, zur Analyse und Speicherung sowie zur echtzeitfähigen Verarbeitung der Daten.

Big Data Technologien, wie die Echtzeitverarbeitung von Datenströmen oder die Parallelisierung der Datenanalyse durch Map-Reduce in Verbindung mit NoSQL und In-Memory Datenspeichern versprechen Werkzeuge für die skizzierten Problemfelder zu liefern. Nur mit Hilfe solcher Werkzeuge scheint eine Beherrschung der immensen produzierten Datenmengen und damit auch eine Nutzung aktueller CPS zu deren voller Effektivität und Effizienz überhaupt möglich.

Der CPSData Workshop soll Wissenschaftlern, Entwicklern, Systembetreibern und anderen Interessierten ein Forum bieten, Fragestellungen an der Schnittstelle zwischen CPS und den Big Data Technologien zur Datenauswertung zu diskutieren sowie neue Problemfelder und Forschungsrichtungen zu identifizieren.

Themenschwerpunkte

- Neue Datenspeichertechnologien
- NoSQL und In-Memory Datenbanksysteme
- (verteiltes) Datenmanagement
- Parallele Datenstromverarbeitung
- Echtzeitfähigkeit von Datenverarbeitungssystemen
- Neue Metriken und Bewertungsrahmen für Datenmanagement in CPS
- Datenfilterung, semantische Annotation und Auswertung heterogener Daten
- Datenspeicherungs- und Retrieval Strategien

[1] http://www.twt-gmbh.de/cpsdata

- Internet of Things, cyber-physische Systeme
- Daten-basierte Regelung von Sensor- und Aktornetzwerken
- Industrie 4.0
- Cybersecurity
- Humanzentrierte CPS
- Einsatzgebiete und Anwendungsfälle für datenzentrierte cyber-physische Systeme
- Value-added Services, die auf den CPS Daten aufsetzen
- Rollen und unternehmensübergreifende Integrationsansätze auf Basis datenintensiver CPS
- Anforderungs- und Architekturmanagement für datenintensive CPS

Organisationskomitee: Dr. Michael Ditze (TWT GmbH), Dr. Sebastian Hudert (TWT GmbH), Dr. Markus Pfeil (TWT GmbH), Dr.-Ing. Frank Golatowski (Universität Rostock), Dr. Christoph Niedermeier (Siemens AG), Dr. Harald Schöning (Software AG)

Programmkomitee: Dr. Thomas Bangemann (ifak), Dr. Alexey Cheptsov (HLRS), Prof. Dr. Torsten Eymann (Universität Bayreuth), Dr. Victor Fäßler (TWT GmbH), Olaf Fischbach (TWT GmbH), Dr. Stephan Grimm (Siemens AG), Kai Hackbarth (ProSyst Software GmbH), Dr. Holmer Hemsen (TU Berlin), Dr.-Ing. Thorsten Henkel (Fraunhofer SIT), Dr.-Ing. Bastian Koller (HLRS), Prof. Dr. Heiko Krumm (TU Dortmund), Norman Natzke (Carano GmbH), Dr. Alexander Paar (TWT GmbH), Prof. Dr. Achim Rettberg (Offis), David Schumm (TWT GmbH), Dr. Thilo Stadelmann (ZHAW Zürich), Dr. Hans-Peter Steiert (Daimler AG), Dr. Jürgen Tacken (Orga Systems), Dr. Alexander Viehl (FZI), Dr. Philipp Wieder (GWDG), Wolfgang Ziegler (Fraunhofer SCAI)

Software Engineering for Software-Intensive Cyber-Physical Systems

Ilias Gerostathopoulos, Jaroslav Keznikl, Tomas Bures, Michal Kit, Frantisek Plasil

Faculty of Mathematics and Physics
Charles University in Prague
Malostranske Namesti 25
11800 Prague, Czech Republic
{iliasg, keznikl, bures, kit, plasil}@d3s.mff.cuni.cz

Abstract: In software-intensive cyber-physical systems (siCPS) the interplay of software control with the physical environment has a prominent role. Nowadays, siCPS are expected to (i) effectively deal with the issues of distribution, scalability, and environment dynamicity, (ii) control their emergent behavior, and, at the same time, (iii) be versatile and tolerant in face of changes and threats. Although approaches that individually meet the above requirements of siCPS already exist, their synergy in a comprehensive software engineering framework is far from trivial. In this paper, we pinpoint the important characteristics of engineering siCPS in an attempt to show that they introduce distinct challenges to traditional software engineering. We argue that this can be addressed by a synergy and adaptation of existing models and abstractions, show our proposal towards such a synergy, and discuss its implications.

1 Introduction

Cyber-physical systems (CPS) are systems of collaborating elements which closely interact with their environment by sensing and actuating. Typically, CPS are characterized by being decentralized, distributed, and heterogeneous.

With the proliferation of smart embedded and mobile devices (smart phones, intelligent cars, etc.) and wireless networks, there is a further trend of CPS becoming large-scale pervasive systems, which combine data from various sources to control real-world ecosystems (e.g., intelligent traffic control, which gathers data about traffic from cars and other sensors in a city and uses them to navigate cars, control the traffic lights, and manage parking allocation). An important feature of these systems is that they are adaptive in order to adjust to situations in the physical environment, and they exhibit emergent behavior (i.e., behavior that comes about as the joint product of behaviors and interactions of many elements of the system). These CPS are also highly dependent on software – they are software-intensive systems [HRW08]. This means that software is by far the most important and most complex constituent of modern CPS.

Continuous dependable operation of CPS is particularly important as the close connection to the physical environment frequently renders the functionality of CPS safety-critical (e.g., operation of the traffic lights in the intelligent traffic control). In addition to being dependable, the software of CPS has to be able to adapt to changing situations in the physical environment. Ideally, it should possess some self-awareness and self-healing properties to cope with not fully anticipated situations.

Along the lines above, in this paper we consider a class of CPS that are software-intensive and, at the same time, distributed at a large scale, inherently dynamic, self-adaptive, self-aware, exhibiting emergent behavior, and safety-critical. It is also important to note that these CPS are targeted by on going research agendas (e.g., EU framework Horizon 2020). We will refer to these CPS as *software-intensive CPS* – siCPS. We argue that siCPS have a number of specifics, which prevent to fully employ traditional software models and software engineering methods. This calls for tailored models and software engineering abstractions that address and potentially take advantage of the specifics of siCPS [Le08]. In fact, siCPS reach the threshold when it is disputable whether we are still dealing with tailored traditional software engineering or whether we are encountering a new paradigm in computing.

As the particular contribution of this paper, (i) we overview these specifics (Section 2) and analyze how they can be addressed by a synergy and adaptation of existing software models and software engineering abstractions (Section 3). On the basis of this, (ii) we give a practical example of such a synergy (Section 4), namely DEECo, an ensemble-based component system [Bu13]. Finally, based on the lessons learned with DEECo, (iii) we discuss potential challenges stemming from the interplay of the models and abstractions in such a synergy (Section 5).

2 Software Engineering Specifics of siCPS

The large-scale physical distribution and interconnectedness within the physical environment makes siCPS rather specific in terms of software engineering (SE). In this section, we overview these specifics from the perspective of SE assumptions and opportunities.

2.1 SE Assumptions Violated in siCPS

A number of assumptions that are typically presumed in traditional SE of general-purpose software systems (GPSS) are violated in siCPS. The assumptions build on the fact that a lot of complexity related to networking and the environment can be considered low-level in GPSS and abstracted away by the operating system and middleware. Of course, even in traditional SE some key assumptions may be violated when developing GPSS with special needs (e.g., high-availability, open-endedness). Nevertheless, siCPS stand out by the large number of such violated assumptions.

Therefore, below we identify and discuss a number of assumptions in traditional SE of GPSS that we deem to have a significant simplifying effect on software development but – according to our experience – cannot be preserved in engineering siCPS:

A1 *Static physical structure* – Even though data and code are subject to mobility in GPSS, the physical nodes where the code is running are typically stationary. In siCPS, the physical substratum is continuously evolving, as nodes move in the physical environment. The fundamental challenge is how to map the ever-changing substratum to the network of computational nodes so that stringent requirements on the desired services are always met.

A2 *Location obliviousness* – The cost and profit of reaching a particular node is typically not significantly influenced by its physical location. This independence facilitates the creation of open-ended and dynamic distributed GPSS and is generally considered an asset. In siCPS, locality of peer nodes is a fundamental design constraint, since physical proximity directly affects reachability and connectivity on one hand and functional correctness on the other.

A3 *Reachability (clique connectivity)* – GPSS typically rely on the Internet network stack for the underlying communication protocols (Internet-based systems [Fr07]). This means that with high probability any node can successfully establish point-to-point communication links with any other node in the system. In siCPS there is no such guarantee, as nodes often operate over dynamic networks lacking a permanent infrastructure, such as mobile ad-hoc networks (MANETs). This limitation imposes a fundamental constraint in the design of siCPS, since nodes are expected to operate in full autonomy, even detached from their peers.

A4 *Stable connections* – In most GPSS, on top of being able to reach and connect to remote subsystems, connections are typically considered stable. This is manifested in the handling of communication errors in such systems: errors are considered exceptions and have to be handled accordingly. In siCPS, errors in communication are the rule, not the exception. Thus, they can no more be handled as exceptions. The property of unstable connectivity has to be acknowledged and ideally be reflected in the employed SE abstractions.

A5 *Availability of global state* – Reasoning over the global state of a distributed system is a requirement for many applications. Although techniques exist for traditional distributed GPSS (e.g., distributed consensus), they are not directly applicable to siCPS because of the loose connectivity among the nodes. Also, since the local state in siCPS evolves continuously with the physical environment, attaining global state is generally infeasible.

A6 *Marginality of real-time aspects* – GPSS typically do not impose hard real-time constraints on their operation and communication. When time matters (e.g., Internet-based video streaming applications), it is mostly because late responses may impede system performance rather than correctness. In siCPS, the passage of time becomes a central feature of system behavior and design, since stringent notion of time is fundamental for measuring, predicting and controlling properties of the physical environment.

A7 *Crisp consistency* – In traditional distributed GPSS, there is a crisp notion of data consistency – the data is either consistent or not (this includes also eventual consistency etc.). On the other hand, in siCPS, where strict distributed synchronization becomes too expensive, such interpretation of consistency is not desirable. Rather, in siCPS it is important to quantify and/or guarantee the degree of (in)consistency [Al14].

A8 *Controlled dynamism* – Many GPSS are dynamic in the sense that they dynamically adapt to changes and recover from malign states. This kind of dynamism, though, is typically a result of actions initiated by the system itself or its administrator. On the contrary, in siCPS, dynamism is inherent, imposed by the

physical environment itself. Thus, siCPS need to detect and recover from contingent and often unforeseen situations in their environment in a non-disruptive way and without supervision (they have to be self-aware and autonomic).

A9 *Focus on reactive behavior* – Outputs of a GPSS are typically reactions to explicit stimuli, such as service requests and internal/external events (e.g., computation is initiated as a response to user input). Instead of waiting for an event, siCPS have to operate proactively in order to react to and also perform changes based on properties that are either sensed or predicted. Relying on simple (e.g., rule-based) reaction patterns is insufficient, since it may lead to oscillations and instability.

A10 *Stateful communication* – GPSS usually assume stateful communication in the communication protocols they employ. This enables effective synchronization among distributed components. Moreover, since stable connections are assumed (A4), errors are treated as exceptional and detected and solved via explicit error recovery. In siCPS, stateful communication does not scale. In fact, extreme network dynamism, typical for siCPS, may incur recurrent error recovery.

2.2 SE Opportunities in siCPS

As pointed out in Section 2.1, none of the discussed assumptions can be generally presumed in siCPS. This makes it a non-trivial challenge to develop siCPS by applying traditional SE methods. However, it would be wrong to perceive all specifics of siCPS as impeding their development, since they may provide opportunities for getting around the violated assumptions. In this perspective, it is desirable to take advantage of such siCPS specifics instead of aiming at adapting traditional SE methods, e.g., building a complex middleware to provide a traditional programming model.

To pinpoint this idea, we have compiled a list of specifics, which we believe can be advantageously exploited in addressing the violated assumptions. Although not complete, we believe this list gives an important research direction for siCPS design methods:

O1 *Physical mobility* – Devices used in siCPS span from stationary to portable and mobile ones. Computational nodes deployed on mobile devices can carry information while moving. This contributes to the overall connectedness of the system, as a mobile node covers a much bigger physical area while moving, and can effectively spread the information in the area and connect otherwise disconnected network partitions. For example, a vehicle moving along a street segment can aggregate temperature data measured from sensors positioned in the tarmac along its route (which themselves cannot reach any external network), and publish the data on a remote server, or spread it to other vehicles in the vicinity.

O2 *Physical locality* – The fact that devices in siCPS are physically close provides a natural way to partition the system into subsystems based on geographical location. This is, again, special to siCPS; general-purpose systems are rarely partitioned based on physical location, because of the otherwise useful assumption on location obliviousness. Having such a natural partitioning can be easily exploited to achieve high levels of scalability.

O3 *Location-dependency of data* – Data in siCPS are often location-dependent, meaning that the value of certain measureable system attributes depend on the physical location of the sensors that provide the raw data. This dependency, in combination with the physical proximity of sensor nodes, allows for data sharing and reuse among nearby nodes and has the potential to contribute to system robustness (in face of sensor failures, etc.).

O4 *Physical laws in data evolution* – Since siCPS operation typically involves sensing physical-environment properties (e.g., position, battery capacity, temperature), one can take advantage of the physical laws that govern the evolution of the values of such properties to estimate/predict their real values. In effect, a value that is slightly stale can still be used, if certain safety bounds on its evolution in time can be established [Al14]. As an example, consider a wireless-based adaptive cruise control system: a stale value of the front vehicle's position can still be used by the rear vehicle's cruise control, since it is possible to estimate the actual position based on the maximum and minimum vehicle acceleration, typically provided by car manufacturers.

3 Approaches that Partially Reflect the Specifics of siCPS

There are no comprehensive methods or supporting models that address the specifics of siCPS in their entirety, as far as our research has indicated. Nevertheless, our experience shows that some SE approaches target these specifics at least partially. In this section we provide a short overview of such approaches (summarized in Table 1), with the goal to later show how they can be combined in a comprehensive framework.

Agent-based systems. In order to deal with dynamicity in siCPS, one can be inspired by *autonomous agents*. This abstraction brings conceptual autonomy to the loosely coupled system parts. Each part is designed to operate with a partial view of the whole system, beneficial when the global state is not available (A5). For example, in the Belief-Desire-Intention (BDI) architectural model [RG95], agents maintain a *belief* about the rest of the system to guide their autonomous decisions. In addition, *multi-agent systems* [SL08] feature the concepts of agent roles and groups, which bring the autonomy to architecture organization and allow building *self-organized systems* that do not rely on the assumptions of controlled dynamism (A8) and static physical structure (A1). An important problem is that industrial agent implementations do not translate the conceptual autonomy and the other useful agent notions (goals, intentions, roles, groups) into proper software engineering constructs that satisfy real-life requirements of autonomous behavior. In particular, they still rely on the assumption of relatively stable bindings between the agents (A4), which is not plausible in most siCPS.

Ensemble-based systems. Another important specific of siCPS is the opportunistic fashion of operation in a dynamic environment at a massive scale. To this end, the paradigm of *attribute-based communication* in *ensemble-based systems* has recently gained attention [De13]. Here, the target of communication is determined according to the values of its attributes rather than by a direct identifier. This paradigm can be exploited to model a best-effort, dynamic coordination of components, effectively dealing with cases when the assumptions of static physical structure (A1), reachability

Assumption:	Agent-based systems	Ensemble-based systems	MANET & gossip protocols	Real-time & control systems	DEECo
A1 Static physical structure		+	+		+
A2 Location obliviousness		+	+		+
A3 Reachability		+	+		+
A4 Stable connections	-		+	-	+
A5 Availability of global state	+	+			+
A6 Marginality of real-time aspects				+	(+)
A7 Crisp consistency		-		+	(+)
A8 Controlled dynamism	+	+	+	-	+
A9 Focus on reactive behavior			+	+	+
A10 Stateful communication				+	+

Table 1: Assumptions from Section 3 and DEECo: lifting "+", partially lifting "(+)", and specific reliance upon "-".

(A3), and controlled dynamism (A8) are violated. However, the application of this paradigm typically relies on explicit and crisp handling of data consistency (A7).

MANET and gossip protocols. At the network layer, extensive research in the areas of *mobile ad-hoc networks* (MANETs) has resulted into a number of routing protocols (see [NPD12] for a comprehensive review), which are able to operate over infrastructure-less dynamic networks. In MANETs, each node acts both as a host and as a router. Node mobility results in dynamically changing network topology. As such, MANET protocols lift the assumption of static physical structure (A1) and work even when the reachability assumption (A3) is violated, thus becoming very relevant to siCPS. Moreover, MANET protocols lift the assumption of location obliviousness (A2), as they enable position-based packet routing [MWH01] (sometimes called *geocast* routing). A promising synergy for siCPS is to combine geocast protocols at the network layer with gossip protocols at the data dissemination layer, effectively enabling proactive, opportunistic communication (A9) in MANETs [Fr07]. Integration of gossiping brings a remedy in cases of unstable connections (A4) and inherent dynamism (A8).

Real-time and control systems. As to strong interaction with physical environment, many techniques already exist in the domain of *embedded real-time systems* [Bu05] and *software control systems* [Pa12]. Such techniques promote proactive behavior (A9) and focus on real-time attributes (A6). They employ control feedback loops, which continuously maintain the operational normalcy (stability) of a system by adequate scheduling of periodic tasks. These techniques stand as a promising way to handle data outdatedness in absence of crisp consistency interpretation (A7) in siCPS, by effectively setting the bounds that define the range of normal system operation. Communication in embedded real-time systems is also typically stateless (A10); consider, e.g., data publishing on CAN bus. Nevertheless, real-time analysis and design typically rely on the assumption of predictable environment, which itself relies on controlled dynamism (A8) and stable connections (A4) assumptions.

4 DEECo: A Synergy

In order to evaluate the potential for a synergy of the approaches discussed in Section 3, as a particular example we present DEECo [Bu13, Ke12] – an *Ensemble Based Component System* that we have proposed specifically for architecting siCPS.

In DEECo, we take the approach of adopting component-based development (CBD) as the basic substratum on top of which we embed selected SE approaches from Section 3. CBD employs reuse, encapsulation and separation of concerns in order to manage the complexity of building and maintaining large applications [CL02]. In CBD, and thus also in DEECo, systems are built around well-defined architectures based on a composition of components, which themselves are seen as encapsulated, reusable, and substitutable entities.

In the remainder of this section, we describe the individual constituents of the DEECo component model with focus on how we approached the synergy. We refer the interested reader to [Bu13] for a detailed technical description of DEECo and for the formal semantics of DEECo. Also, a Java implementation is available[1].

4.1 Component

Adopting the ideas of agent-based and self-adaptive systems, the concept of *component* in DEECo is centered on the features of autonomy, self-adaptation, and belief (A5). Specifically, a component is an autonomous, encapsulated, and composable software entity constituting its own state and behavior.

As is typical for software agents, component state is expressed in terms of *knowledge* (e.g., line 3 in Figure 1). Note that in DEECo, all the data accessible to a component is referred to as knowledge. In alignment with the BDI architectural model, knowledge of a component comprises both the private component state (e.g., calendar) and the component's belief about the rest of the system (e.g., parkingAvailability). In slight difference from traditional BDI approach, rather than being updated explicitly by the component itself, the belief is updated automatically (by the execution environment, Section 4.3) as a result of component composition (Section 4.2). This decision further stresses the component's autonomy and separation of concerns.

The behavior of a component is represented by a set of *processes* (e.g., lines 4-7 in Figure 1). Following the notions of control systems and self-adaptive systems, a process is essentially a feedback loop, continuously and proactively maintaining the operational normalcy of a component (A9). At the same time, each process executes concurrently, independently of the other processes, i.e. it atomically reads its inputs, executes its body, and atomically writes its outputs. A process operates strictly upon the knowledge of the corresponding component; it may thus interact with other components only through the (externally updated) belief (A4, as there is no "direct" communication among components).

[1] https://github.com/d3scomp/JDEECo

```
1.  component Vehicle
2.    knowledge:
3.      calendar, parkingAvailability, plan, ...
4.    process computePlan(in calendar, in parkingAvailability, out plan):
5.      function:
6.        plan ← JourneyPlanner.computePlan(calendar, parkingAvailability)
7.      scheduling: periodic( 5000ms )
8.    ...
9.
10. component ParkingLot
11.   knowledge:
12.     position, availability, ...
13.   process monitorAvailability(out availability):
14.     …
15.
16. // updates Vehicle's belief about availability of all ParkingLots along the route
17. ensemble UpdateAvailabilityInformation:
18.   coordinator: Vehicle
19.   member: ParkingLot
20.   membership:
21.     ∃ event ∈ coordinator.calendar: distance(member.position, event.position) < TRESHOLD
22.   knowledge exchange:
23.     coordinator.parkingAvailability ← members.reduce(member.availability)
24.   scheduling: periodic( 2000ms )
```

Figure 1: Example of a DEECo component and ensemble definition in a DSL.

4.2 Component Composition

For component composition we adopt the approach of ensemble-based systems and multi-agent systems by employing autonomic self-organization of components into component *ensembles* (in multi-agent systems called *groups*). This self-organization is based on a declarative representation of a component's membership in an ensemble, based on the component's context (A1 and A3). In order to distinguish in which ensemble the membership is being decided upon, every ensemble has a *coordinator*. Membership in an ensemble with a given coordinator is based on whether a component is able to assume the role of a *member* w.r.t. the coordinator. This is expressed technically via a *membership condition*, which decides whether two given components can form a coordinator-member pair. Following the idea of attribute-based communication, the membership condition is defined upon the attributes (i.e., knowledge exposed for this purpose) of the components in question (e.g., line 21 in Figure 1). Note, that the ensemble definition is generic and determines ensemble instantiation for each group of components meeting the membership condition (w.r.t. particular coordinator). Also, a component can be a member or coordinator of multiple ensembles at the same time. Within an ensemble, we adopt the idea of stateless, proactive communication employed in control systems and gossip-based systems (A9 and A10). Specifically, the communication takes the form of stateless *knowledge exchange*. Its objective is to update the belief of the components within the ensemble recurrently and proactively, based on a given prescription (e.g., line 23 in Figure 1). This form of communication aligns well with the proactive, cyclic execution of component processes. Note, that the statelessness and proactivity make knowledge exchange suitable for cases of faulty connections (A4) and inherent dynamism (A8).

4.3 Execution Environment

The main task of the DEECo execution environment is performing knowledge exchange in a distributed setting. For this purpose, we combine the protocols for geographical routing in MANETs with gossip protocols so as to enable location-aware communication of belief (A2) in mobile ad-hoc environments (A1 and A3) with unstable connections and inherent dynamism (A4 and A8). Specifically, the execution environment proactively advertises the knowledge of a (source) component to all the other potentially-interested (target) components via a geocast protocol. Then, in case the source and target components meet the membership condition of an ensemble, the execution environment updates the belief of the target component according to the knowledge exchange prescription of the ensemble.

Adopting the approach of embedded real-time systems, the execution environment is also responsible for execution of component processes and activities related to knowledge exchange in a (soft) real-time fashion (A6 and partially A7), featuring both periodic and event-based scheduling.

5 Discussion of Implications

Engineering siCPS with the basic building blocks (autonomous components, ensembles) offered by the proposed synergy in DEECo offers several advantages, but also poses new challenges. As seen in Table 1, DEECo addresses all of the identified challenges of A1-A10, which we deem a step forward. Certainly, there could be other assumptions, e.g., predictability of underlying platform and global synchronization of beliefs, which still remain to be addressed. Building on our experience in applying the ensemble-based component system paradigm to two real-life case studies, namely the intelligent vehicle navigation [Bu13] and the firefighter tactical coordination [Bu14], this section discusses the implications of merging different methods.

Exploitation of the opportunities. A close synergy of geocast MANET protocols and attribute-based communication, and an integration of membership evaluation and routing in particular, allows exploiting new opportunities based on physical locality (*O2*) and location-dependency of data (*O3*) (i.e., membership can effectively exploit physical location). Further, the proactive gossip-based advertisement of belief enables exploiting the physical mobility (*O1*). The cyclic and real-time nature of component processes also facilitates use of models that estimate/predict the safety bounds of knowledge evolution [Al14]. This is done by exploiting the physical laws that govern the evolution of certain knowledge values (*O4*).

Components as autonomous agents. Borrowing the ideas of belief and autonomous operation from agent-based systems and coupling them with the encapsulation and deployment facilities of component-based systems results into a dependable platform for robust component-based agent implementations. The robustness is achieved by grafting such “agents” with implicit component binding and communication. Contrary to other agent-based frameworks, the autonomous components thus do not communicate directly, e.g., via sending messages; instead, component knowledge serves as a communication medium. A component’s belief, i.e., the part of its knowledge that reflects knowledge of

other components, plays a role of "smart" sensors and actuators. For instance, a belief could represent a "smart" sensor providing "positions of up to 10 closest parking lots, which are available". All in all, a component's belief is updated externally – via knowledge exchange handled by execution environment.

Stateless interaction. Adopting the idea of attribute-based communication in component interaction has many advantages when considering that components in siCPS recurrently appear and disappear and form dynamic groups of best-effort coordination. At the same time, having no means of direct component binding and addressing makes it challenging – but certainly not impossible, as we have observed – to realize some forms of protocol-based communication. This is essential in certain interactions, e.g., reserving of a parking place by a specific vehicle at a specific parking lot. Stateless interaction dictates knowledge design in a way that it is always possible to reconstruct the state of the session from the knowledge, e.g. by assigning each parking reservation request a globally unique identifier (GUID), so that a reservation response could refer to it.

Embedded feedback loops. When designing siCPS, special means have to be provided for feedback loops. By building on the ideas of control and real-time systems, DEECo embeds the feedback loop operation both at design time and runtime. Systems based on feedback loops typically require a description of operational normalcy, usually in terms of periodic scheduling of tasks. However, the adoption of this idea needs a paradigm change in the design process, to explicitly focus on the normalcy that each process is expected to maintain as opposed to goals to be achieved [Ke13].

Decentralized operation. Coupling best-effort data dissemination of MANET protocols with attribute-based communication and decentralized system operation can result in situations when different parties act based on inconsistent local beliefs – so-called split-brain scenarios. For instance, a component can believe itself to be a member of an ensemble, while the ensemble's coordinator does not recognize this situation (or vice-versa). This behavior is in a way inevitable, however it has to be accounted for in the design, e.g., by making components only weakly synchronized or by relying on an underlying network or physical environment to provide some guarantees (thus making these split-brain situations temporary with an upper bound for duration).

Ensembles as component connectors. The duality between components and ensembles resembles the classical problem of components and connectors – especially whether connectors are only special types of components and what functionality should be in connectors and what functionality should be in components. In particular, this holds when connectors comprise complex adaptation logic. In DEECo though, this problem is partially remedied by distinguishing that (i) although stateful, a component has a direct access solely to its local knowledge, (ii) an ensemble embodies only stateless exchange of knowledge among its member components. This is a strong conceptual difference pushing ensembles into the role of simple connectors and components into the role of entities performing the actual computation and data aggregation.

Parallel process execution. The physical world is inherently concurrent. Software engineering abstractions for engineering siCPS have to deal with concurrency by allowing execution of processes in parallel. This leads to challenges with regard to the handling of shared resources, which, if not dealt with, can result into race conditions,

deadlocks, etc., effectively jeopardizing the safety of the system. Similar to actor-based design, where the exchanged data are considered immutable, DEECo avoids introducing any dedicated synchronization constructs. Rather, it employs the simple semantics of atomically operating over knowledge while applying the rule of single-writer for each knowledge field. The downside of the approach is that it sometimes leads to the necessity of having a special "aggregation" process in a component, which merges data coming from different sources (similar situation happens in actor-based approaches as well). However, this seems a reasonable price to pay for preventing race conditions by design.

6 Related Work

Since CPS is an emerging class of systems, there are multiple research efforts trying to shed light on the state of the art and the challenges ahead [KK12, Sh09]. Unfor- tunately, not as many solutions are proposed, especially when considering guidance via proper software engineering abstractions specific to CPS. Our work highlights the problems in CPS software engineering, while, at the same time, we propose solutions to these problems and evaluate their implications. In the same spirit, in [DLS12], Derler, Lee and Vincentelli focus on the challenges with modeling CPS caused by the intrinsic heterogeneity, concurrency, and sensitivity of such systems. Backed up by a hybrid-system-modeling environment called Ptolemy II, their approach emphasizes determinism and predictability in modeling and simulations of safety-critical CPS. In [Le08], Lee reviews the requirements/specifics of CPS and identifies the absence of timing behavior in core abstractions in computing as the main impediment in developing future CPS. In our work, we focus on the subset of CPS that is software-intensive, where structural models and systematic engineering methods become more relevant.

Our aim at a synergy can be compared to frameworks proposed for self-adaptive/self-organizing systems, e.g., [DFR10], and autonomic agent-based systems, e.g., [LPH04]. In [DFR10], Di Marzo Serugendo, Fitzgerald and Romanovsky propose a synergy of self-organization, agent-inspired autonomy and rule-based reasoning into a service-oriented architectural framework. Their approach is centered around the concepts of self-describing components, component metadata and interaction policies executed at runtime, resembling the concepts of components, component knowledge and ensembles, respectively. In [LPH04], Liu, Parashar and Hariri present a component-based framework for autonomic agents building on agent-based middleware infrastructure. The difference from these and other similar approaches lies in the fact that we deal with the specifics of siCPS, where unreliable communication and extreme dynamism loom large.

7 Conclusion

Building software for software-intensive cyber-physical systems (siCPS) is far from trivial. In this paper, we attempted to pinpoint the challenges and pitfalls associated with applying traditional software engineering (SE) methods in siCPS and to show how these challenges can be met by a comprehensive synergy and adaptation of existing SE models, methods and abstractions. This we exemplified on the DEECo component model. The evaluation of the proposed synergy in DEECo outlines a number of

interesting research topics in terms of addressed and waiting-to-be-addressed issues, such as design based on maintaining operational normalcy.

References

[Al14] Al Ali, R. et al.: Architecture Adaptation Based on Belief Inaccuracy Estimation. In: Proc. WICSA'14, Sydney, Australia, 2014. IEEE, 2014; pp. 87-90.

[Bu05] Buttazzo, G. et al.: Soft Real-Time Systems: Predictability vs. Efficiency. Springer, 2005.

[Bu13] Bures, T. et al.: DEECo: An Ensemble-based Component System. In: Proc. CBSE'13, Vancouver, Canada, 2013. ACM, 2013; pp. 81-90.

[Bu14] Bures, T. et al.: Adaptation in Cyber-Physical Systems: from System Goals to Architecture Configurations. Tech. Rep. D3S-TR-2014-01, Charles University.

[CL02] Crnkovic, I.; Larsson, M.: Building Reliable Component-Based Software Systems. Artech House, Inc., Norwood, MA, USA, 2002.

[De13] De Nicola, R. et al.: A Language- Based Approach to Autonomic Computing. In: Formal Methods for Components and Objects, Springer, 2013; pp. 25-48.

[DFR10] Di Marzo Serugendo, G.; Fitzgerald, J.; Romanovsky, A.: Meta-Self: An Architecture and a Development Method for Dependable Self-* Systems. In: Proc. 25th ACM Symp. on Applied Computing, Sierre, Switzerland, 2010. ACM, 2010; pp. 457–461.

[DLS12] Derler, P.; Lee, E. A.; Sangiovanni-Vincentelli, A.: Modeling Cyber-Physical Systems. In: Proceedings of the IEEE, 100(1):13-28, Jan 2012.

[Fr07] Friedman, R. et al.: Gossiping on MANETs: The Beauty and the Beast. In: ACM SIGOPS Operating Systems Review, 41:67–74, 2007.

[HRW08] Hölzl, M.; Rauschmayer, A.; Wirsing, M.: Software-Intensive Systems and New Computing Paradigms. In: Engineering of Software-Intensive Systems: State of the Art and Research Challenges, Springer-Verlag, 2008; pp. 1-44.

[Ke12] Keznikl, J. et al.: Towards Dependable Emergent Ensembles of Components: The DEECo Component Model. In: Proc. of WICSA/ECSA'12. IEEE, 2012; pp. 249–252.

[Ke13] Keznikl, J. et al.: Design of Ensemble-based Component Systems by Invariant Refinement. In: Proc. of CBSE'13, Vancouver, Canada, 2013. ACM, 2013;pp. 91-100.

[KK12] Kim, K.-D.; Kumar, P.R.: Cyber-Physical Systems: A Perspective at the Centennial. In: Proceedings of the IEEE, 100 (Special Centennial):1287-1308, May 2012.

[Le08] Lee, E. A.: Cyber Physical Systems: Design Challenges. In: Proc. ISORC'08, Orlando, Florida, USA, 2008. IEEE, 2008; pp. 363–369.

[LPH04] Liu, H.; Parashar, M.; Hariri, S.: A Component-Based Programming Model for Autonomic Applications. In: Proc. ICAC'04, New York, USA, 2004. IEEE; pp. 10-17.

[MWH01] Mauve, M.; Widmer, A.; Hartenstein, H.: A Survey on Position-based Routing in Mobile Ad Hoc Networks. In: IEEE Network, 15(6):30-39, Nov 2001.

[NPD12] Natesapillai, K.; Palanisamy, V.; Duraiswamy, K.: A Review of Broadcasting Methods for Mobile Ad Hoc Network. In: International Journal of Advanced Computer Engineering, Serial Publications, New Dehli India, Sep 2012.

[Pa12] Patikirikorala, T. et al.: A Systematic Survey on the Design of Self-Adaptive Software Systems using Control Engineering Approaches. In: 2012 ICSE Workshop on Soft. Eng. for Self-Adaptive and Self-Managing Systems, Jun 2012. ACM; pp. 33–42

[RG95] Rao, A. S.; Georgeff, M. P.: BDI Agents: From Theory to Practice. In: Proc. of the 1st Int. Conf. on Multi-Agent Systems, 1995; pp. 312–319.

[Sh09] Sha, L. et al.: Cyber-Physical Systems: A New Frontier. In: Machine Learning in Cyber Trust, Springer US, 2009; pp. 3–13.

[SL08] Shoham, Y.; Leyton-Brown, K.: Multiagent Systems: Algorithmic, Game- Theoretic, and Logical Foundations. Cambridge University Press, 2008.

Ein abstraktes SystemC-Modell zur Analyse und Leistungsabschätzung des europäischen Zugsicherungssystems ETCS

Benjamin Beichler, Alexander Nitsch, Frank Golatowski, Christian Haubelt

Institut für Angewandte Mikroelektronik und Datentechnik
Universität Rostock
Richard Wagner Str. 31
18119 Rostock-Warnemünde
{benjamin.beichler; alexander.nitsch;
frank.golatowski; christian.haubelt} @uni-rostock.de

Abstract: In diesem Beitrag wird ein SystemC-Modell der Geschwindigkeits- und Abstandsüberwachung aus dem European Train Control System (ETCS) vorgestellt. Dieses Modell dient als Ausgangspunkt für die frühzeitige Abschätzung der Leistungsfähigkeit des Systems und die in der Berechnung entstehenden Datenmengen. Hierfür wurde eine neuartige Methode entwickelt, welche es bei minimalen Anpassungen am Anwendungsmodell erlaubt, schnell unterschiedliche Entwurfsalternativen zu explorieren. Dabei werden SystemC-Prozesse in Abhängigkeit von Scheduling-Entscheidungen gestartet, deren Kommunikationsverhalten aufgezeichnet und anschließend Ausführungs- und Kommunikationszeiten simuliert.

1 Einleitung

Der europäische Bahnsektor ist durch unterschiedliche Sicherungssysteme gekennzeichnet, was den grenzüberschreitenden Verkehr von Schienenfahrzeugen durch Inkompatibilitäten erschwert. Zur Verbesserung der Interoperabilität ist das European Train Control System (ETCS) als einheitliches europäisches Zugsicherungssystem entstanden. Neben den vorrangigen Zielen, die Kosten für den Betrieb und die Instandhaltung zu reduzieren sowie Interoperabilität zu schaffen, hat das ETCS zum Ziel, die Streckenkapazität und -geschwindigkeit weiter zu erhöhen. Dazu ist es notwendig, eine Vielzahl von Sensor-, Strecken- sowie Verkehrsdaten zu erfassen und zu verarbeiten. Als sicherheitskritisches Cyber-Physikalisches-System unterliegt das Zugsteuerungssystem strengen Echtzeitanforderungen, was eine sehr hohe Rechenleistung der Onboard-Unit voraussetzt. Die Anwendung von Multicore-Technologien bietet eine Möglichkeit, das weiter ansteigende Datenaufkommen durch parallele Datenverarbeitung zu beherrschen. Ein wesentlicher Bestandteil des ETCS-Standards ist die Geschwindigkeits- und Abstandsüberwachung. Für diese komplexe Komponente wird mit der Modellierungssprache SystemC ein abstraktes, ausführbares Systemmodell entwickelt, wodurch gleichzeitig dieser Teil der ETCS-Anforderungsspezifikation formalisiert wird. Dieser Ansatz unterscheidet sich von

bereits verfügbaren Modellen, da er Performance-Abschätzungen für Hardware-/Softwaresysteme des Zugüberwachungsrechners zu einem sehr frühen Zeitpunkt des Systementwurfs gestattet. Die entwickelte Methode erlaubt bei minimalen Anpassungen am Anwendungsmodell, schnell unterschiedliche Entwurfsalternativen zu explorieren. Dabei werden SystemC-Prozesse in Abhängigkeit von Scheduling-Entscheidungen gestartet, deren Kommunikationsverhalten aufgezeichnet und anschließend Ausführungs- und Kommunikationszeiten simuliert. Mit der Verwendung der Open-Source-Modellierungssprache SystemC sind die Ergebnisse dieser Arbeit in das BMBF/ITEA-Projekt openETCS eingeflossen. Das Projekt hat zum Ziel, Open-Source-Modelle und Werkzeuge für das Zugsicherungssystem ETCS zu entwickeln, um eine Langzeitverfügbarkeit und Wartbarkeit zukünftiger Systeme zu gewährleisten. Der Rest des Beitrages strukturiert sich wie folgt: Im zweiten Abschnitt werden verwandte Arbeiten diskutiert. Abschnitt drei behandelt die grundlegenden Zusammenhänge der Geschwindigkeits- und Abstandsüberwachung des ETCS-Standards bevor im vierten Teil die Performance-Evaluationsmethodik vorgestellt wird. Der fünfte Abschnitt dieser Publikation zeigt die Implementierungsdetails des Anwendungsmodells zur Geschwindigkeits- und Abstandsüberwachung. Abschnitt sechs stellt die gewonnen Ergebnisse der Performance-Evaluation des Anwendungsmodells vor. Eine Zusammenfassung erfolgt in Abschnitt sieben.

2 Stand der Technik

Seit der Verabschiedung des ETCS-Standards wurden mehrere Arbeiten veröffentlicht, in denen verschiedene Aspekte der ETCS-Spezifikation untersucht wurden. Viele der Arbeiten beschäftigen sich mit den Echtzeiteigenschaften und der Ausfallsicherheit der Kommunikationsverbindung zwischen Zug und streckenseitigen Einrichtungen. In [zHHS13, JzHS98, ZH05, HJU05] werden verschiedene Petri-Netz-Erweiterungen genutzt, um die funktionalen Eigenschaften sowie stochastische Sicherheiten der Kommunikation zu untersuchen. Die Modellierung und Berechnung der Abstands- und Geschwindigkeitsüberwachung von ETCS wurde in [BT11, Fri10] bearbeitet, wobei sich diese Arbeiten auf die funktionalen Eigenschaften der Berechnung konzentrieren und eine anwendungsspezifische Modellierungsmethodik verwenden. Im Rahmen dieser Publikation wird die in [ACC12] standardisierte Modellierungssprache SystemC verwendet, um einerseits die funktionalen Aspekte zu beschreiben und andererseits auch eine Analyse der Performance des Berechnungssystems vorzunehmen. Im Rahmen der Performance- und Energieverbrauchsanalyse wurden für SystemC-Modelle einige Methoden entwickelt. Auf dem abstrakten Electronic System Level (ESL) werden beispielsweise Simulations-Traces [WHL05] zur Abschätzung von Performance oder Leistungsaufnahme genutzt. Diese Arbeit nutzt einen ähnlichen Ansatz wie die in [SFH+06, SHT09] vorgestellten Virtual Processing Components (VPC). Im Gegensatz zu den genannten Publikationen wird die Modellierung des Anwendungsmodells nicht auf aktororientierten Modellen umgesetzt, sondern auf Anwendungsmodellen, die direkt SystemC benutzen. Um Performance-Analysen auf der abstrakten System-Ebene zuzulassen, wurden die Zeitannotationen mit wenigen Veränderungen am Anwendungsmodell umgesetzt. Außerdem wurde die Konformität zur Referenzimplementierung des SystemC-Simulators beibehalten. Der Ansatz aus

[ASH+11] zeigt einige Ähnlichkeiten zur vorgestellten Methode, in dem ein zyklenapproximiertes Verfahren mit TLM-Modellen vorgestellt wird. Die Kalibrierung des Modells wird in [ASH+11] sehr feingranular durch eine Analyse des Quellcodes mit Hilfe eines Compilers vorgenommen, was allerdings eine Neukompilierung nach jeder Änderung der Konfiguration nach sich zieht. Die Methodik in diesem Beitrag arbeitet auf einem abstrakteren Niveau und benötigt keine Neukompilierung.

3 Geschwindigkeits- und Abstandsüberwachung im ETCS

Auf Grund der niedrigen Haftreibung zwischen Rad und Gleis benötigt das Bremsen eines Schienenfahrzeuges besondere Aufmerksamkeit. Im Vergleich zu einem PKW hat ein Zug eine viel geringere Bremsverzögerung und damit einen sehr langen Bremsweg, welcher teilweise im Kilometerbereich liegt. Da Gefahrenstellen deshalb nicht einsehbar sind, wird der Triebfahrzeugführer vom ETCS unterstützt, welches die Überwachung der genauen Geschwindigkeit und Position eines Zuges sicherstellt. Dadurch werden zum einen Fahrinformationen und zum anderen sicherheitskritische Bremsbefehle generiert, um eine größtmögliche Fahrsicherheit zu gewährleisten. Das Kapitel 3.13. *„Geschwindigkeits- und Abstandsüberwachung“* der Anforderungsspezifikation des ETCS-Standards [UNI12] beschreibt die dazu notwendigen Funktionseinheiten und Berechnungen.

Abbildung 1: Geschwindigkeits- und Abstandsüberwachung im ETCS, Funktionseinheiten: Acceleration due to Gradient **(AdtG)**, Calculation of Deceleration **(CoD)**, Calculation of Brake Build Up Time **(CoBBUT)**, Most restrictive Speed Profile **(MrSP)**, Determination of Supervised Targets **(DoST)**, Determination of Brake Deceleration Curves **(DoBDC)**, Supervision Limits **(SL)**, Speed and Distance Monitoring Commands **(SaDMC)**; die Datenabhängigkeiten sind in Abbildung 4 dargestellt

Wie in Abbildung 1 zu sehen, werden für die Berechnungen zugseitige und streckenseitige Eingangsparameter sowie Odometriedaten (momentane Geschwindigkeit/Position) benötigt. Die zugseitigen Eingaben sind jene Daten, welche der Zug lokal zur Verfügung stellt. Diese beinhalten im Wesentlichen: Bremsmodelle, Status der Bremsen, Korrekturfaktoren für Bremsmodelle, zugspezifische Parameter, das Traktionsmodell und feste Werte. Für die Berechnung des Bremsweges muss die Leistungsfähigkeit der Bremsen vorliegen. Ihre mathematische Entsprechung kann zum einen in Treppenfunktionen und zum anderen in den branchenüblichen Bremshundertstel dargestellt sein. Der Standard sieht vier verschiedene Bremsarten vor. Neben den grundsätzlich immer vorhandenen pneumatischen Bremsen können zusätzlich noch Nutzbremsen, Wirbelstrombremsen oder Magnetschienenbremsen vorhanden sein, die in

unterschiedlichen Kombinationen zur Anwendung kommen. Die konkrete Bremskombination hängt darüber hinaus vom aktuellen Zustand der Bremse ab. Ein weiterer wichtiger Aspekt ist die Berücksichtigung der Zeitspanne, die vergeht, bis eine Bremse ihre vollständige Bremskraft entfaltet. Im Betrieb können die Bremsmodelle drei unterschiedlichen Verwendungszwecken dienen. Im Notfall wird die vollständig verfügbare Bremsleistung berechnet, was den sicheren Betrieb eines Zuges garantiert. In den anderen beiden Fällen werden nicht alle Eingangsparameter und Korrekturfaktoren verwendet, weil es sich nicht um sicherheitskritische Berechnungen handelt. Korrekturfaktoren für Bremsmodelle sind notwendig, weil sich das Bremsverhalten je nach Streckenzustand (trocken, nass oder vereist) verändert. Die Bremsbeschleunigung unterliegt einer statistischen Streuung, welche vom Zughersteller über empirische Testszenarien ermittelt und mit einer statistischen Sicherheit angegeben wird. Zusätzlich sind die Zuglänge, die maximale Geschwindigkeit und Achslast sowie die Zugkategorie von großer Bedeutung. Die Zugkategorie bestimmt in diesem Fall, welche Geschwindigkeitsrestriktion zur Anwendung kommt. Das Traktionsmodell beschreibt den Einfluss der Traktionsbeschleunigung des Motors auf die Bremswirkung eines Triebfahrzeugs, wenn eine automatische Zwangsbremsung erfolgt, da die Traktion nicht beliebig schnell deaktiviert werden kann. In der ETCS-Spezifikation werden die erforderlichen Algorithmen beschrieben. Für die Algorithmen sind verschiedene Konstanten, sogenannte „Fixed Values“ festgelegt. Beispielsweise wird dort die nominale Reaktionszeit des Triebfahrzeugführers festgeschrieben.

Die streckenseitigen Eingangsinformationen sind der Fahrbefehl, streckenseitige und andere Geschwindigkeitsbeschränkungen, das Höhenprofil und sonstige Streckenbeschaffenheiten sowie streckenspezifische Vorgabewerte. Der sogenannte Fahrbefehl, engl. „Movement Authority“, ist die Erlaubnis eines Zuges, sich in einem bestimmten Streckenabschnitt bewegen zu dürfen. Aus der Beschaffenheit der Schiene und der Umgebung ergeben sich verschiedene Geschwindigkeitseinschränkungen. Diese lassen sich in statische und dynamische Beschränkungen einteilen und werden von den streckenseitigen Einrichtungen an den Zug übermittelt. Die statischen Geschwindigkeitsrestriktionen ergeben sich aus der Geometrie des Streckenabschnitts, die dynamischen lassen sich auf vorübergehende Unsicherheiten auf der Strecke, der möglichen Zugkategorie sowie einer spezifischen Achslast zurückführen. Im Höhenprofil sind Informationen über Gefälle und Steigungen der Strecke enthalten. Neben den Streckeneigenschaften, aus denen Geschwindigkeitsbeschränkungen resultieren, gibt es noch andere Streckenbeschaffenheiten, die den Zugbetrieb direkt beeinflussen. Es können sich Einschränkungen für die Nutzung von Spezialbremsen ergeben. Darüber hinaus können weitere Meldungen existieren, wie zum Beispiel Hinweise auf verringerte Haftung in gewissen Streckenabschnitten. Streckenspezifische Vorgabewerte regeln im Allgemeinen die Standard-Werte für verschiedene Berechnungen oder Einstellungen und können für jede Strecke unterschiedlich definiert sein.

Für die Berechnung der Bremsbefehle und Führerstandsanzeigen sind im ETCS-Standard verschiedene Module beschrieben (Abbildung 1), die im Folgenden kurz dargestellt werden. Das Modul „Acceleration due to Gradient“ (AdtG) errechnet die durch ein Höhenprofil der Strecke verursachte Beschleunigung des Zuges. Notwendige

Operationen in dem Modul sind die Berechnung einer Zuglängenkompensation und die Berechnung der Beschleunigung anhand der Zugmasse, wobei die Rotationsmasse der Räder eine wichtige Rolle spielen. Das Modul „Calculation of Deceleration“ (CoD) fast die Bremsmodelldaten für die Notbremse und die verschiedenen Einflüsse auf die Beschleunigungsfunktionen zu einer Funktion zusammen, woraus sich eine positionsabhängige Bremsbeschleunigungsfunktion ergibt. Die Zeitspanne, die die Bremsen benötigen, um ihre vollständige Bremskraft zu entfalten, wird in dem Model „Calculation of Brake Build Up Time (CoBBUT) realisiert. Auf Grund der Vielzahl unterschiedlicher Geschwindigkeitsbeschränkungen besteht die Notwendigkeit, zu jeder Position die jeweils restriktivste Höchstgeschwindigkeit auszugeben. Diese Funktionalität wird im Modul „Most restrictive Speed Profile“ (MrSP) erzielt. Die Geschwindigkeits- und Abstandsüberwachung muss simultan verschiedene Ziele, sogenannte „Targets“ überwachen. Das Modul „Determination of Supervised Targets“ (DoST) unterhält eine Liste dieser Ziele mit einer genauen Position und der dazu assoziierten Geschwindigkeit. Bremskurven sind in der Geschwindigkeits- und Abstandsüberwachung sehr wichtige Elemente. Sie ermöglichen die Vorhersage des Bremsweges in Abhängigkeit von einer bestimmten Geschwindigkeit und werden im Modul „Determination of Brake Deceleration Curves“ (DoBDC) berechnet. Die im Fokus dieser Arbeit stehende Bremskurve ist die „Emergency Brake Deceleration Curve“ (EBD). Sie repräsentiert die Bremskurve im Falle einer Notbremsung. Die Informationen aus allen vorhergehenden Funktionaleinheiten werden im Modul „Supervision Limits“ (SL) genutzt, um die nötigen Daten zum Einleiten von Maßnahmen der Geschwindigkeits- und Abstandsüberwachung zu errechnen, wofür im Standard verschiedene Überwachungsgrenzen definiert sind. Als letzte Instanz der Geschwindigkeits- und Abstandsüberwachung berechnet das Modul „Speed and Distance Monitoring Commands“ (SaDMC)} konkrete Zugbefehle, wie automatische Bremsungen, aber auch die Anzeigen für den Triebfahrzeugführer. Im Wesentlichen werden die überwachten Ziele mit der aktuellen Position und Geschwindigkeit verglichen und anschließend geeignete Maßnahmen eingeleitet.

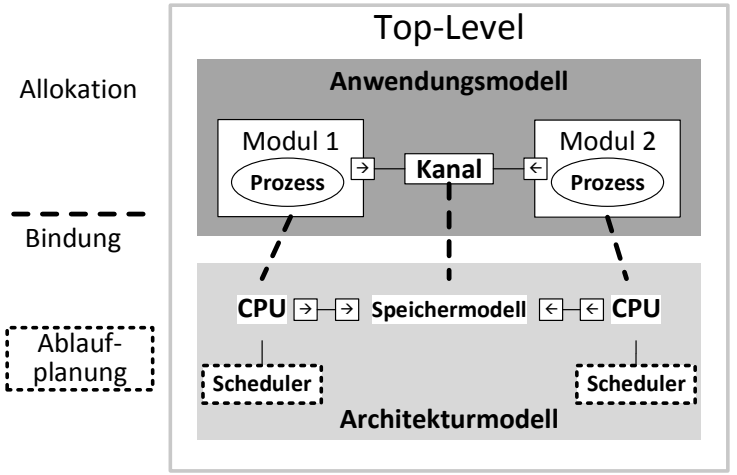


Abbildung 2: Mapping des Anwendungsmodells auf ein Architekturmodell

4 Performance-Evaluation

Im Entwicklungsprozess von eingebetteten Systemen ist es wichtig, schon frühzeitig Abschätzungen der Leistungsfähigkeit des gesamten Systems durchführen zu können. Dies ist insbesondere nötig, um eine Bewertung der getroffenen Entwurfsentscheidungen vornehmen zu können. Da die Entscheidungen auf Systemebene den größten potentiellen Einfluss auf die spätere Leistungsfähigkeit haben, ist es das Ziel, mit der vorgestellten Methode Aussagen über die Leistungsfähigkeit bereits auf dieser Ebene treffen zu können. Die Modellierungssprache SystemC hat sich als de-facto Standard für die Beschreibung und Simulation von Hardware-/Software-Systemen entwickelt. Durch die Flexibilität von SystemC in der Verwendung auf verschiedensten Abstraktionsebenen eignet es sich besonders auch für abstrakte Systembeschreibungen.

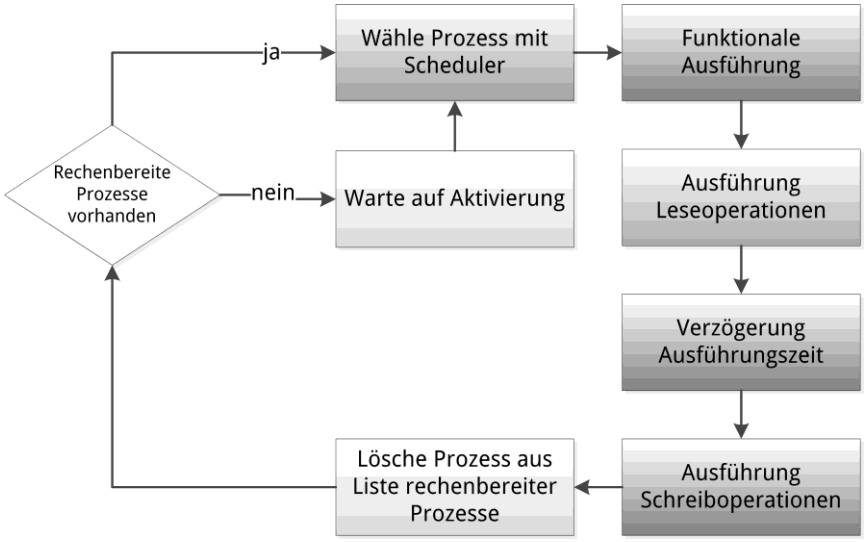


Abbildung 3: Simulationsschleife zur Performance-Evaluation

Für die Performance-Analyse eines Anwendungsmodells wird ein separates abstraktes Architekturmodell entwickelt, welches die Aspekte der Berechnungs- und Kommunikationsdauer simuliert. Die Abbildung 2 zeigt eine solche Kombination aus Anwendungs- und Architekturmodell. Dadurch können die Entwurfsentscheidungen der Allokation, Bindung, und Ablaufplanung einzelner Komponenten modelliert werden. Die Entwurfsentscheidungen werden dabei auf verschiedene programmatische Aspekte abgebildet:

- Allokation: Erzeugung von Objekt-Instanzen
- Bindung: Aufruf von Objekt-Methoden auf den Architektur-komponenten
- Ablaufplanung: Erzeugung eines Scheduler-Objektes, welches aus der Liste rechenbereiter Prozesse den auszuführenden Prozess auswählt

Nach der Elaboration dieses Top-Level-Modells wird während der Simulation innerhalb jeder abstrakten Recheneinheit die in Abbildung 3 gezeigte Schleife ausgeführt. Die ursprünglichen verhaltensbeschreibenden SystemC-Prozesse werden nicht direkt vom SystemC-Simulatorkern ausgeführt, sondern innerhalb der Recheneinheiten. Nach der Aktivierung eines SystemC-Prozesses des Anwendungsmodells durch ein Eingangssignal wird dieser nicht direkt ausgeführt, sondern bis zur geplanten Ausführung der Berechnungseinheit verzögert. Nach der Auswahl durch den Scheduler erfolgt die funktionale Ausführung des Anwendungsprozesses, in welcher gleichzeitig die Kommunikationsvorgänge des Anwendungsprozesses aufgezeichnet werden. Eine Blockierung des Prozesses unterstützt die derzeitige Implementierung nicht, weshalb das vorgestellte Verfahren momentan lediglich `SC_METHOD`s, die über `sc_signal`s kommunizieren, als Anwendungsmodell unterstützt. Während der funktionalen Ausführung schreitet die Simulationszeit nicht voran. Nach dieser funktionalen Phase werden die einzelnen Phasen des Prozesses zeitlich annotiert. In der ersten Phase werden alle Lesevorgänge simuliert, indem die aufgezeichneten lesenden Kommunikationsvorgänge in Blocking-TLM-Transaktionen übersetzt und innerhalb des Prozesses der Berechnungseinheit ausgeführt werden. Darauf folgt die Simulation der Ausführungszeit des Anwendungsprozesses, die im Wesentlichen einen `wait()`-Aufruf enthält. Im Anschluss werden die aufgezeichneten schreibenden Kommunikationsvorgänge in Blocking-TLM-Transaktionen übersetzt und ausgeführt. Erst innerhalb dieser Phase werden die erzeugten Ergebnisse, welche auf die Kommunikationskanäle geschrieben wurden, sichtbar und können weitere Prozesse aktivieren. Die dafür nötigen Änderungen am Anwendungsmodell sind so gestaltet, dass einerseits möglichst viele Freiheiten in der Benutzung von SystemC gestattet sind und andererseits möglichst wenige Änderungen für die Performance-Evaluation im Anwendungsmodell nötig sind.

```
...
SC_CTOR(test_module){
SC_METHOD_ANNOTATED(compute,10,SC_NS);
Sensitive << in1;
...
```

Listing 1: Änderung im Anwendungsmodell

Die beiden wesentlichen Änderungen sind die Änderungen der Kommunikationskanäle zwischen den Modulen und die Deklaration der verhaltensbeschreibenden Prozesse. Die Ersetzung der Kommunikationskanäle findet im Top-Level des Modells statt und ist durch die klare Trennung zwischen Interface und Implementierung dieser Komponenten einfach vorzunehmen. Durch die neuen Kanäle ist das Aufzeichnen der Kommunikationsaufrufe leicht möglich. Die Änderung der Deklaration von Prozessen in Anwendungsmodulen ist im Listing 1 zu sehen. Neben der Änderung des Makronamens wird als zusätzlicher Parameter eine Zeitkonstante übergeben. Diese wird als Ausführungszeit dieses Prozesses genutzt. Neben einer statischen Zeitangabe ist prinzipiell auch die Übergabe eines `sc_time`-Objektes möglich, um eine variable Ausführungszeit zu ermöglichen. Dieses Makro erlaubt dadurch eine gesteuerte Ausführung der Prozesse. Offensichtlich kann es bei diesem Ansatz zu Ungenauigkeiten in der Simulation des Zeitverhaltens kommen, da Datenwerten in der funktionalen Ausführung früher gelesen werden, als dies in einer echten Implementierung der Fall wäre, in welcher die Daten erst nach der Ausführung der lesenden TLM-Transaktionen

vorliegen würden. Diese zeitliche Entkopplung zwischen funktionaler und zeitlicher Simulation lässt aber eine schnelle und einfache Analyse der dynamischen Kommunikationseinflüsse der Komponenten zu. Die weitere angestrebte Verwendung von non-blocking-TLM bei gleichzeitiger Verwendung von „temporal decoupling" kann das Konzept noch sinnvoll erweitern.

5 Anwendungsmodell für die Geschwindigkeits- und Abstandsüberwachung im ETCS

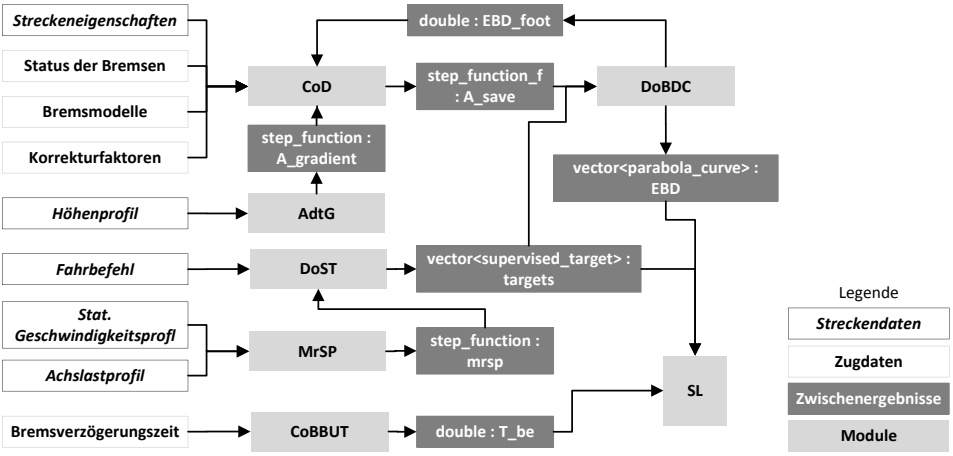


Abbildung 4: Übersicht der SystemC-Module

Dieser Abschnitt beschreibt den konkreten Aufbau und die Umsetzung des entwickelten SystemC-Modells. Die Abbildung 4 zeigt die im Rahmen dieses Beitrags implementierten Module sowie deren Datenabhängigkeiten. Die Implementierung der Module orientiert sich stark an der ETCS-Spezifikation [UNI12]. Alle vorgegebenen Partitionierungen, Benennungen, und Datentypen der Komponenten des Anwendungssystems werden aus der Spezifikation übernommen.

Es ergibt sich außerdem der Bedarf an speziellen Datentypen für die Berechnung in der Geschwindigkeits- und Abstandsüberwachung. Dies sind Repräsentationen für abschnittsweise definierte Funktionen. Für die spezielle Anwendung werden drei Typen von Funktionen benötigt und innerhalb von C++-Klassen implementiert:

- abschnittsweise definierte konstante Funktionen (Treppenfunktion) mit einer unabhängigen Variable: `step_function`
- abschnittsweise definierte konstante Funktionen (Treppenfunktion) mit zwei unabhängigen Variablen: `step_function_f`
- abschnittsweise definierte parabelförmige Funktion: `parabola_curve`

Im Wesentlichen werden alle diese Datenstrukturen mit dem assoziativen Container `std::map` umgesetzt. Als Schlüssel werden die Funktionsargumente des Beginns eines Abschnittes genutzt. Die gespeicherten Daten sind die Funktionswerte. Für Treppenfunktionen mit zwei unabhängigen Variablen werden als Funktionswerte Treppenfunktionen mit einer unabhängigen Variable gespeichert. Die Klasse `parabola_curve` speichert für jeden Abschnitt die Werte s_{begin}, v_{begin} und a.

Die Funktionswerte eines Abschnitts werden anhand des Geschwindigkeits-Weg-Gesetzes der gleichmäßig beschleunigten Bewegung berechnet, also mit der Formel:

$$v = \sqrt{2 \cdot a \cdot (s - s_{begin}) + v_{begin}^2}$$

Eine der Herausforderungen der Algorithmen der Geschwindigkeits- und Abstandsüberwachungen ist die Verarbeitung dieser Datentypen. In vielen Fällen ist es nötig, über Teile des Definitionsbereiches der Funktionen zu iterieren, weshalb verschiedene Hilfsfunktionen für diese Aufgaben implementiert wurden.

Es wurden darüber hinaus folgende generelle Entscheidungen für die SystemC-Modellierung des gesamten Anwendungsmodells getroffen:

- alle Prozesse innerhalb des Modells werden als Method-Prozesse modelliert
- alle Verbindungen zwischen den Modulen des Modells werden mit Signalen realisiert
- das System wird nur über die Dateneingänge aktiviert

Mit diesen Entscheidungen modelliert dieses Anwendungsmodell ein reaktives System, welches permanent auf Eingangswerte reagiert und keinen gemeinsamen Takt besitzt. Das Verhalten des Systems wird nur mit SystemC-Method-Prozessen realisiert, um die Datentransformationen zu beschreiben und effektiv zu simulieren. Diese Vorgehensweise erlaubt eine kompakte abstrakte Beschreibung des Systems, welche aber keine zeitliche Annotation der Vorgänge vornimmt.

6 Testergebnisse

Anhand der in dieser Publikation vorgestellten Methodik zur Performance-Evaluierung wurden mit dem Anwendungsmodell zur Geschwindigkeits- und Abstandsüberwachung von ETCS verschiedene Messungen durchgeführt. In Abbildung 5 sind die drei Szenarien mit Allokations-, Bindungs-, und Ablaufplanungsentscheidungen zu sehen. In allen Architekturen wurden ein FIFO-Scheduler sowie ein zentraler Speicher verwendet. Es wurde nur die Anzahl der Berechnungseinheiten und die nötige Bindung der Aufgaben an die Berechnungseinheit variiert.

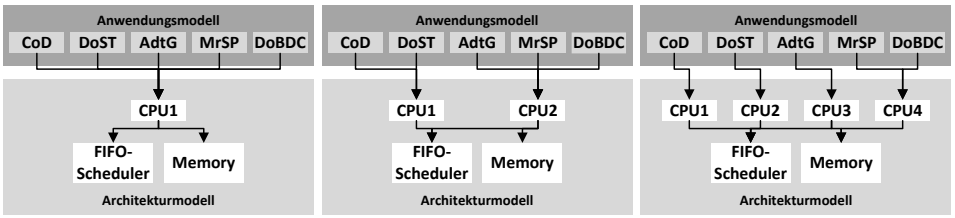


Abbildung 5: Getestete Architekturen für das Anwendungsmodell

Für die Kalibrierung des Modells wurde das Anwendungsmodell mit der Laufzeitanalyse-Programmsammlung Valgrind untersucht, welche in [NS07] vorgestellt wird. Die dadurch erreichte grobe Schätzung der benötigten Maschinenzyklen wurde mit einer Maschinenzyklusdauer von 1 ns kombiniert, welches mit einem Testdatensatz die Simulationsergebnisse in Tabelle 1 erzeugt hat. Es wurde ein zentraler Speicher mit einer einfachen Implementierung des TLM-Blocking-Interfaces genutzt, welcher aber durch beliebige andere TLM-Modelle ersetzt werden kann.

Anzahl CPU	Ausführungszeit [ms]	Deltazyklen
1	11,027	273
2	10,813	298
4	14,686	303

Tabelle 1: Ergebnisse der Performance-Evaluation

Die Tabelle 1 zeigt die ermittelten Werte für einen kompletten Berechnungszyklus der Geschwindigkeits- und Abstandsüberwachung. Durch den erhöhten Kommunikations- und Simulationsaufwand mit mehreren Recheneinheiten steigt die Anzahl der Deltazyklen mit steigender Anzahl der Berechnungseinheiten. Im Gegensatz zur möglicherweise vermuteten Verkürzung der simulierten Berechnungszeit mit mehreren Berechnungseinheiten steigt die Berechnungszeit mit 4 Berechnungseinheiten im Vergleich zur Verwendung einer Berechnungseinheit. In diesem konkreten Beispiel lassen sich diese Werte durch den FIFO-Scheduler erklären, welcher keine Datenabhängigkeiten beachtet, sodass einige Berechnungen unnötigerweise mehrfach ausgeführt werden müssen. Für messbare Ergebnisse der Simulationsperformance wurde die gleiche Berechnung zyklisch bis zu einer Simulationszeit von 100 Sekunden ausgeführt. Auf einem Linux-System mit 2,26 GHz Intel Core2Duo Prozessor beendet das reine Anwendungsmodell die Simulation nach 11,22 s, wohingegen die Simulation mit Performance-Evaluation nach 42,87 s endet. Diese zu erwartende Verlängerung entsteht im Wesentlichen durch die hinzukommenden Kontext-Wechsel der Prozesse in den Berechnungseinheiten sowie den zusätzlich auszuführenden Programmteilen der Performance-Evaluation, die innerhalb des reinen Anwendungsmodells durch die Verwendung der `SC_METHOD`-Prozesse entfallen.

7 Zusammenfassung

Im Rahmen dieser Arbeit konnte ein Modell für die Berechnung der Geschwindigkeits- und Abstandsüberwachung erstellt werden, wodurch eine abstrakte und ausführbare Darstellung der ETCS-Spezifikation entstanden ist. Auf der Grundlage dieses SystemC-Modells ist es möglich, durch die erfolgte Analyse der Spezifikation Ungenauigkeiten und potentielle Fehlerquellen zu identifizieren. Die durchgeführten Tests haben gezeigt, dass dieser Ansatz Performance-Abschätzungen für Hardware-/Softwaresysteme des Zugüberwachungsrechners zu einem sehr frühen Zeitpunkt des Systementwurfs ermöglicht. Obwohl die momentane Implementierung der Performance-Evaluation noch Einschränkungen aufweist, erlaubt die entwickelte Methode bei minimalen Anpassungen am Anwendungsmodell, schnell unterschiedliche Entwurfsalternativen zu explorieren.

Danksagung

Diese Arbeit wurde teilweise vom BMBF im Rahmen des ITEA-Projekts "openETCS" unter dem Kennzeichen 01IS12021K gefördert.

Literaturverzeichnis

[ACC12] Design Automation, Standards Committee und Ieee Computer: IEEE Std 1666-2011, IEEE Standard for Standard SystemC® Language Reference Manual, Band 2011. 2012, ISBN 9780738168012.

[ASH+11] S. Abdi, G. Schirner, Yonghyun Hwang, D.D. Gajski und Lochi Yu: Automatic TLM Generation for Early Validation of Multicore Systems. Design Test of Computers, IEEE, 28(3):10–19, May 2011, ISSN 0740-7475.

[BT11] B.Vincze und G. Tarnai: Development and analysis of train brake curve calculation methods with complex simulation. Advances in Electrical and Electronic Engineering, 5(1-2):174–177, 2011.

[Fri10] B. Friman: An algorithm for braking curve calculations in ERTMS train protection systems. Advanced Train Control Systems, Seite 65, 2010.

[HJU05] H. Hermanns, D.N. Jansen und Y.S. Usenko: A comparative reliability analysis of ETCS train radio communications, February 2005. AVACS Technical Report No. 2.

[JzHS98] L. Jansen, M. M. zu Hörste und E. Schnieder: Technical issues in modelling the European Train Control System. Proceedings of the workshop on practical use of coloured Petri Nets and Design /CPN 1998, Seiten 103–115, 1998.

[NS07] N. Nethercote und J. Seward: Valgrind: a framework for heavyweight dynamic binary instrumentation. ACM Sigplan Notices, 2007

[SFH+06] M. Streubühr, J. Falk, Ch. Haubelt, J. Teich, R. Dorsch und T. Schlipf: Task-accurate performance modeling in SystemC for real-time multi-processor architectures. In: Design, Automation and Test in Europe, 2006. DATE '06. Proceedings, Band 1, Seiten 2 pp.-, 2006.

[SHT09] M. Streubühr, Ch. Haubelt und J. Teich: System level performance simulation for heterogeneous multi-processor architectures. Workshop on Rapid Simulation and Performance Evaluation: Methods and Tools (Rapido), 2009.

[UNI12] UNISIG: SUBSET-026 – System Requirements Specification. SRS 3.3.0, ERA, 2012.

[WHL05] T. Wild, A. Herkersdorf und G. Y. Lee: TAPES—Trace-based architecture performance evaluation with SystemC. Design Automation for Embedded Systems, 10(2-3):157–179, September 2005, ISSN 0929-5585.

[ZH05] A. Zimmermann und G. Hommel: Towards modeling and evaluation of ETCS real-time communication and operation. Journal of Systems and Software, 77(1):47–54, 2005.

[zHHS13] M.M. zu Hörste, H. Hungar und E. Schnieder: Modelling Functionality of Train ControlSystems using Petri Nets. Towards a Formal Methods Body of Knowledge for Railway Control and Safety Systems, Seite 46, 2013.

Cyber-physische Systeme im OP-Saal - Ein Machbarkeitsnachweis -

Martin Kasparick, Frank Golatowski, Dirk Timmermann

Institut für Angewandte Mikroelektronik und Datentechnik
Fakultät für Informatik und Elektrotechnik
Universität Rostock
Richard-Wagner-Straße 31
18119 Rostock
martin.kasparick@uni-rostock.de
frank.golatowski@uni-rostock.de
dirk.timmermann@uni-rostock.de

Abstract: In einem OP-Saal befinden sich eine große Anzahl von Medizingeräten und Informationssystemen. Eine herstellerübergreifende Kommunikation bzw. Interoperabilität ist derzeit kaum bzw. gar nicht möglich. Dieses Problem adressiert die vorliegende Arbeit. Es wird von einer serviceorientierten Kommunikationsarchitektur ausgegangen und gezeigt, wie heutige Medizingeräte ohne entsprechende Hard- und/oder Software-Schnittstellen integriert werden können. Anhand einer prototypischen Umsetzung eines Teils eines endoskopischen Arbeitsplatzes wird die Machbarkeit des vorgestellten Ansatzes gezeigt. Ausgehend von derart vernetzten cyber-physikalischen Systemen (CPS) im OP-Saal ergeben sich erstmals Möglichkeiten der Big-Data-Analysen der anfallenden Daten, da diese nicht mehr lokal auf die Geräte beschränkt sind, sondern global verknüpft werden können.

1 Einleitung

In einem heutigen OP-Saal befinden sich eine Vielzahl medizinischer Geräte zur Überwachung und Beeinflussung der Vitalfunktionen, zur Durchführung des operativen Eingriffs sowie Systeme zur Dokumentation und Einsichtnahme in Patientendaten etc. Nach heutigem Stand agieren diese Geräte nicht vernetzt und somit ohne Interaktion und Datenaustausch untereinander. Lediglich Geräte desselben Herstellers nutzen teilweise Insellösungen; eine herstellerübergreifende Kommunikation ist aufgrund unterschiedlicher Hardwareschnittstellen und proprietärer Protokolle nicht möglich. In der vorliegenden Arbeit wird gezeigt, dass das beschriebene Problem mit einer SOA-basierten Architektur gelöst werden kann.

Cyber-Physical Systems (CPS) werden nach Lee und Seshia [LS11] als die Integration der Berechnungen (computation) in die physikalischen Prozesse definiert. Dabei überwachen und steuern eingebettete Computer und Netzwerke die physikalischen Prozesse. Dies geschieht typischerweise mittels Regelkreisen, wobei die physikalischen Prozesse die Be-

rechnungen beeinflussen und umgekehrt. Für ein solches System ist es nicht ausreichend die physikalischen und rechnenden Komponenten einzeln zu betrachten, sondern deren Zusammenspiel. Ein moderner Operationssaal stellt demnach ein solches CPS dar. Lee et al. [LSC+12] betrachten Medical Cyber-Physical Systems (MCPSs) sogar als eigene Klasse von CPSs. Die Begründung liegt in der Kombination von eingebetteter Software zur Steuerung der Geräte, der (zukünftigen) Netzwerkfähigkeit der Geräte, der komplexen physikalischen Welt und den Auswirkungen auf die Patientensicherheit. Die Komplexität bezieht sich hierbei vor allem auf den menschlichen Körper als physikalische Welt.

Auf den einzelnen Geräten im OP-Saal fällt eine große Menge von Daten an. Bezogen auf das Big Data 3-V-Modell (Data Volume, Data Velocity und Data Variety) [Lan01] [Bey11] stellt das OP-Umfeld vor allem eine Herausforderung in Bezug auf die Geschwindigkeit (Velocity) und Mannigfaltigkeit (Variety) der Daten dar. Für viele Aspekte, wie z. B. die Gerätesteuerung, müssen Echtzeitanforderungen eingehalten werden. Außerdem unterscheiden sich die anfallenden Daten von Geräte- über Vitalparameter bis hin zu Patienten-, Befund- und Bilddaten stark. Aus Sicht der Big Data Analysen besteht ein enormes Potential zur Steuerung der medizinischen Geräte an sich sowie etwa im Hinblick auf die Qualitätssicherung oder die Weiterentwicklung der medizinischen Methoden und Vorgehensweisen. Dies kann aber nur erschlossen werden, wenn eine Interoperabilität der Medizinprodukte untereinander, bis hin zu den Krankenhausinformationssystemen hergestellt werden kann. Die vorliegende Arbeit thematisiert die Herausforderung, einen solchen interoperablen Datenaustausch zu ermöglichen. Zur Realisierung werden neue Methoden der Gerätekommunikation zum Einsatz kommen. Bedingt durch die langen Lebenszyklen von Medizinprodukten ist es zwingend notwendig, die heutigen Systeme in die neue Infrastruktur einzubinden. Daher liegt ein Hauptfokus dieser Arbeit auf dem Thema der Integration von nicht SOA-kompatiblen Systemen. Die Integration wird anhand eines Teils eines medizinischen Endoskopie-Arbeitsplatzes gezeigt.

Die vorliegende Arbeit ist folgendermaßen gegliedert: Zunächst wird auf einige Grundlagen und den Stand der Technik eingegangen. Es folgt eine Beschreibung der SOA-basierten Gesamtarchitektur. In Kapitel 4 wird das Einbinden von nicht DPWS-kompatiblen Systemen beschrieben. Danach wird die prototypische Umsetzung dargestellt. Zum Abschluss werden eine Zusammenfassung und ein Ausblick gegeben.

2 Grundlagen und Stand der Technik

2.1 DPWS

Um Web Services auf eingebetteten Systemen mit beschränkten Ressourcen nutzen zu können wurde das Geräteprofil DPWS (Devices Profile for Web Services) [OAS09] entwickelt. DPWS setzt auf SOAP (Simple Object Access Protocol) auf und stellt eine Untermenge der verfügbaren Web Service Standards dar.

Auf dem Gerät, welches gesteuert werden soll bzw. welches Daten zur Verfügung stellen soll, wird die DPWS-Server-Seite implementiert. Auf der Steuereinheit bzw. Datensenke

wird die DPWS-Client-Seite implementiert. Bei DPWS ist vor allem die serverseitige Implementierung ressourcenschonend. Soll die Kommunikation zwischen den eingebetteten Geräten erfolgen, müssen beide Seiten implementiert werden.

DPWS nutzt den Standard WS-Discovery [OAS09] zum Auffinden von Geräten (Discovery). WS-MetadataExchange und WS-Transfer werden zum XML-basierten Austausch von Geräte- und Servicebeschreibungen genutzt. Zum Suchen von Diensten senden Clients sogenannte Probe-Nachrichten per UDP-Multicast. Die Geräte, die einen entsprechenden Dienst anbieten, antworten mit einem Probe Match via Unicast. Zur Auflösung der Geräteadresse, die als Teil der Probe Match-Nachricht übermittelt wird, werden Resolve-Nachrichten versandt. Die entsprechende Resolve Match-Nachricht enthält eine gültige Transportadresse. Zusätzlich zum beschriebenen Verfahren ist im DPWS-Standard vorgesehen, dass sich neue Geräte mit einer sogenannten Hello-Nachricht bekannt machen und mit einer Bye-Nachricht abmelden. Beides erfolgt per UDP-Multicast.

Für die Funktionalität des Abonnierens und Versendens von Ereignissen (Events) wird der Standard WS-Eventing genutzt. Dieser ermöglicht es, dass ein Client vom Server Benachrichtigungen u. a. über Werteänderungen erhält, ohne ein aktives Polling zu betreiben.

2.2 DPWS im OP-Saal

Eine Analyse des Standes der Technik präsentieren Mauro et al. in [MSLK09]. Forschungslücken im Bereich der SOA-Integration medizinischer Geräte werden ausgehend von anderen Anwendungsgebieten, z. B. der industriellen Automatisierung, identifiziert.

Verschiedene Projekte beschäftigen sich mit der Nutzung von gerätenaher SOA im OP. Im Zuge der Projekte smartOR [iMuB] und orthoMIT [BORN09] wurde ein SOA-basiertes Framework für den OP entwickelt. Dieses wird u. a. von Ibach et al. [IBR10] [IBS+12] und Benzko et al. [BINR10] beschrieben. Es kommen standardisierte Protokolle und Interfacebeschreibungen zum Einsatz. Als Basistechnologie wird DPWS eingesetzt.

Ein medizinisches Geräteprofil auf Basis von DPWS wird von Pöhlsen et al. [PSS+09] vorgeschlagen. Wesentlich sind hierbei das Geräteprofil MDPWS (Medical Devices Profile for Web Services) und das klinische Integrationsprofil BICEPS (Basic Integrated Clinical Environment Protocol Specification) [Sch13]. Die technischen Hintergründe werden im Whitepaper von Schlichting und Pöhlsen [SP14b] ausführlich dargestellt. Es erfolgte eine Einreichung als Standardisierungsvorschlag [SP14a] bei den 11073/HL7 Standard Weeks.

Im Zuge des DOOP-Projekts [doo] (Dienst-orientierte OP-Integration) wurden Lösungsstrategien für eine Plug-and-Play-artige Medizingerätevernetzung vorgestellt. Im Fokus steht das SOA-Paradigma zur Integration von Medizingeräten im OP-Saal und auf der Intensivstation. [GBM12]

Die beschriebenen (Vor-)Projekte konnten den Stand der Technik nicht soweit vorantreiben, dass eine herstellerunabhängige Gerätevernetzung in heutigen OP-Sälen realisiert werden kann. Daher wurde das BMBF geförderte Projekt „OR.NET - Sichere dynami-

sche Vernetzung in Operationssaal und Klinik“[1] [orn12] [KBB13] initiiert, das Partner aus den Bereichen der Wissenschaft, Industrie und Standardisierung vereint. Das Ziel ist eine standardisierte, herstellerübergreifende und sichere Interoperabilität der Medizinprodukte und IT-Systeme. Außerdem werden Themen, wie die sichere Bedienung durch die Optimierung der MMI (Mensch-Maschine-Interaktionen) sowie Zulassungsfragen bearbeitet. Die Gerätevernetzung erfolgt nach dem Prinzip einer SOA mittels DPWS. [orn14]

2.3 SOA-Integration von nicht SOA-kompatiblen Systemen

Mit dem Pattern des „Legacy Wrappers“ zeigt Erl [ER09] konzeptionell die Integration von Legacy-Systemen in eine SOA. Das Gerät wird als Service gekapselt, sodass es für den Servicekonsumenten transparent ist und Daten ausschließlich über die entsprechenden Services ausgetauscht werden können. Abbildung 1 verdeutlicht die Funktionsweise des Patterns. Durch die enge Kopplung der Servicelogik sowohl zur Legacy-API, also auch zum Servicevertrag kann eine Kopplung des Servicenutzers und der Implementierung vermieden werden. So kann das SOA-Grundprinzip der losen Kopplung umgesetzt werden.

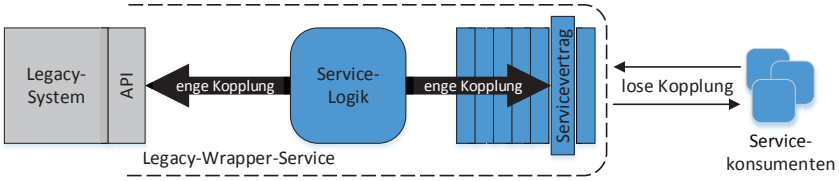


Abbildung 1: Legacy-Wrapper-Service nach Erl [ER09] (Seite 442).

Mit der Service-Oriented Device Architecture (SODA) stellen de Deugd et al. [dDCK+06] ein System zur serviceorientierten Kapselung von Geräten vor. Ein Gerät wird mit zwei verschiedenen Adaptern ausgestattet. Der Device-Adapter realisiert die Kommunikation zu den (proprietären) Geräte-Schnittstellen und Protokollen. Auf der anderen Seite enthält der Device-Adapter eine Geräte und Szenario spezifische Logik, um ein abstraktes Servicemodell des Gerätes zur Verfügung zu stellen. Der Bus-Adapter stellt die Verbindung zur SOA her, in dem das abstrakte Servicemodell in spezifische SOA-Binding-Mechanismen überführt wird.

3 Gesamtarchitektur

Die Architektur des Gesamtsystems muss die Kommunikation sowohl medizinischer Geräte untereinander als auch mit den Informationssystemen des Krankenhauses[2] ermög-

[1] Die vorliegende Arbeit entstand im Zuge des OR.NET-Projekts.

[2] wie z. B. KIS (Krankenhausinformationssystem) oder PACS (Picture Archiving and Communication System)

lichen. Die Notwendigkeit zur Integration der heute im Betrieb befindlichen Geräte ergibt sich aus den langen Lebenszyklen und hohen Kosten von Medizinprodukten.[3] Daher ist die möglichst einfache und kostengünstige Integration eine Grundvoraussetzung für eine Marktakzeptanz der neuartigen Kommunikationsprinzipien. Die Abbildung 2 zeigt einen (vereinfachten) Architekturentwurf, wie er im OR.NET-Projekt verfolgt wird. Die Kommunikation erfolgt nach den Prinzipien einer Serviceorientierten Architektur (SOA) und nutzt hierfür die konkrete Umsetzung des Devices Profile for Webservices (DPWS). Die Webservicekommunikation wird sowohl zwischen den einzelnen medizinischen Geräten als auch zum sogenannten Konnektor IS[4] genutzt.

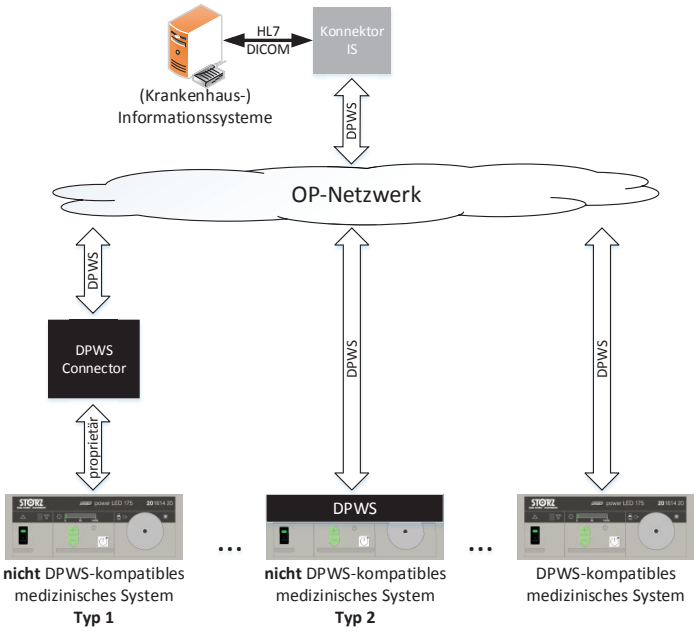


Abbildung 2: Gesamtarchitektur der DPWS-basierten Vernetzung von OP-Geräten. (Vereinfachte Darstellung des Architekturentwurfs im OR.NET-Projekt [orn12].)

Die im unteren Bereich von Abbildung 2 dargestellten medizinischen Systeme repräsentieren die Geräte, die sich im OP-Saal befinden. Die Unterscheidung zwischen DPWS-kompatiblen und nicht DPWS-kompatiblen Systemen ergibt sich aus der Art der Einbindung der Geräte in den DPWS-basierten Kommunikationsverbund. Ein nicht DPWS-kompatibles System vom Typ 1 verfügt weder über die entsprechende Hardwareschnittstelle noch über einen DPWS-Stack. Daher wird zur Einbindung zusätzliche Hardware benötigt. Der Typ 2 von nicht DPWS-kompatiblen Systemen verfügt bereits über eine

[3]Die im Zusammenhang mit der Integration von Systemen in eine neue Umgebung gebräuchliche Bezeichnung des „Legacy Systems“ ist ggf. irreführend. Der Begriff „Legacy“ bezieht sich nicht auf die medizinische Hauptfunktionalität der Geräte, sondern lediglich auf ihre Kommunikationsfähigkeit. Daher werden im Verlauf der Arbeit meist Bezeichnungen wie „nicht SOA- / DPWS-kompatibles System“ genutzt.

[4]IS wird als Abkürzung für Informationssystem genutzt.

IP-basierte Netzwerkschnittstelle, sodass lediglich die Firmware erweitert werden muss. In Abschnitt 4 dieser Arbeit wird detailliert auf die Einbindung von derartigen Systemen eingegangen. Als DPWS-kompatible Systeme werden solche bezeichnet, die zum Zeitpunkt der Erstinbetriebnahme bereits hardware- und softwareseitig in der Lage sind an der DPWS-Kommunikation teilzunehmen.

Der Konnektor IS bildet die Schnittstelle zu den Informationssystemen des Krankenhauses. Zwischen dem Konnektor IS und den Informationssystemen werden typischerweise standardisierte Protokolle wie HL7 oder DICOM genutzt. Durch die direkte Anbindung der medizinischen Geräte an die Informationssysteme können große Mengen von Daten persistent gespeichert werden. Außerdem steht ausreichend Rechenleistung zur Verfügung, um Methoden der Big Data für Analysen dieser Daten zu nutzen.

Aspekte von Security und Safety sollen nicht Gegenstand der vorliegenden Arbeit sein. Mechanismen zur Zuordnung von Geräten zu Patienten bzw. medizinischen Eingriffen, Authentifizierung und Autorisierung etc., die sicherstellen, dass der Informationsaustausch von Patienten- und Steuerdaten nur zwischen den „richtigen“ Geräten möglich ist, sind in der Architektur enthalten.

4 Einbindung von nicht DPWS-kompatiblen Systemen

In der vorliegenden Arbeit wird ein Blackbox-Ansatz genutzt, um die Lücke zwischen den existierenden OP-Geräten mit ihren unterschiedlichen (Hardware-) Schnittstellen und proprietären Protokollen und der zukünftig standardisierten webservicebasierten Kommunikation zu schließen. Für nicht DPWS-kompatible Systeme vom Typ 1 (siehe Abschnitt 3) stellt die Blackbox ein System aus Hard- und Softwarekomponenten dar. Im Fall eines Typ 2 Systems reduziert sich die Komplexität entsprechend auf Softwarekomponenten. Im Folgenden wird vor allem auf die Umsetzung für Typ 1 Systeme eingegangen. Die softwareseitige Umsetzung des vorgestellten Systems kann aber analog auf Systeme des zweiten Typs übertragen werden.

Hardwareseitig muss sowohl eine Schnittstelle zum bestehenden herstellerspezifischen System zur Verfügung gestellt werden, als auch eine IP-basierte Schnittstelle um an der Webservice-Kommunikation teilnehmen zu können.

Die Blackbox-Software besteht aus zwei wesentlichen Teilkomponenten: Legacy-Adapter und DPWS-Adapter. Der Legacy-Adapter implementiert einerseits die Schnittstellen und Protokolle des einzubindenden Gerätes, andererseits die geräte- und anwendungsspezifische Logik (vergleiche auch de Deugd et. al [dDCK+06]). Des Weiteren kann eine optionale Abstraktionsschicht als Subkomponente integriert werden, die eine Kapselung von als schützenswert eingestuften firmenspezifischen Lösungen vornimmt. Diese Abstraktionsschicht kann sowohl oberhalb, als auch unterhalb der Subkomponente erfolgen, die die spezifische Anwendungslogik enthält. Der DPWS-Adapter (nach de Deugd et al. auch Bus-Adapter) realisiert die Einbindung des Systems in die mittels DPWS realisierte Serviceorientierte Architektur. Abbildung 3 illustriert schematisch den Aufbau der Blackbox.

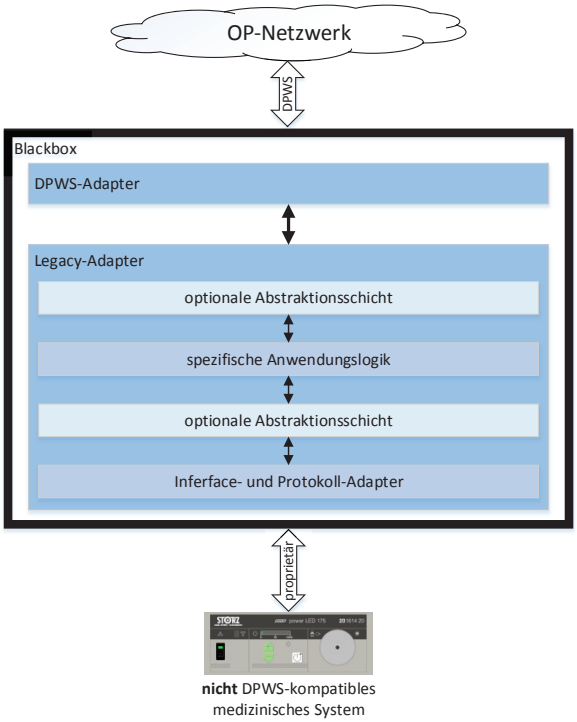


Abbildung 3: Modell einer Blackbox zur Integration von nicht DPWS-kompatiblen Systemen.

5 Prototypische Umsetzung

Die beschriebene Blackbox-Architektur zum Einbinden von nicht DPWS-kompatiblen Systemen wurde in einem prototypischen Demonstrator evaluiert. Abbildung 4 illustriert den Gesamtaufbau, der im Folgenden beschrieben wird. Eine schematische Darstellung zeigt Abbildung 5.

Im rechten Teil der Abbildung 4, gekennzeichnet mit a_1) - a_3), befinden sich zwei Geräte eines endoskopischen Arbeitsplatzes, die mit einem DPWS-Konnektor ausgestattet sind. Dies ist zum einen eine endoskopische Kamera (Karl Storz Image 1), zum anderen eine dazugehörige Kaltlichtquelle (Karl Storz PowerLED 175). Für die Kamera ist ein Service implementiert, der es ermöglicht ein Standbild abzurufen. Die Kaltlichtquelle stellt Services zum Auslesen der aktuellen Lichtintensität in Prozent sowie zum Erhöhen und Verringern der Intensität bereit. Weiterhin kann ein Eventing-Service abonniert werden, der bei einer Änderung des Intensitätswertes ausgelöst wird. Die Blackbox zur Integration dieser beiden Systeme in die DPWS-basierte Kommunikation wird inklusive der beschriebenen Abstraktionsschicht realisiert, die firmenspezifisches Know-how effektiv kapselt. Daher wird der geräteseitige Teil des DPWS-Konnektors (Hardware und die Softwarekom-

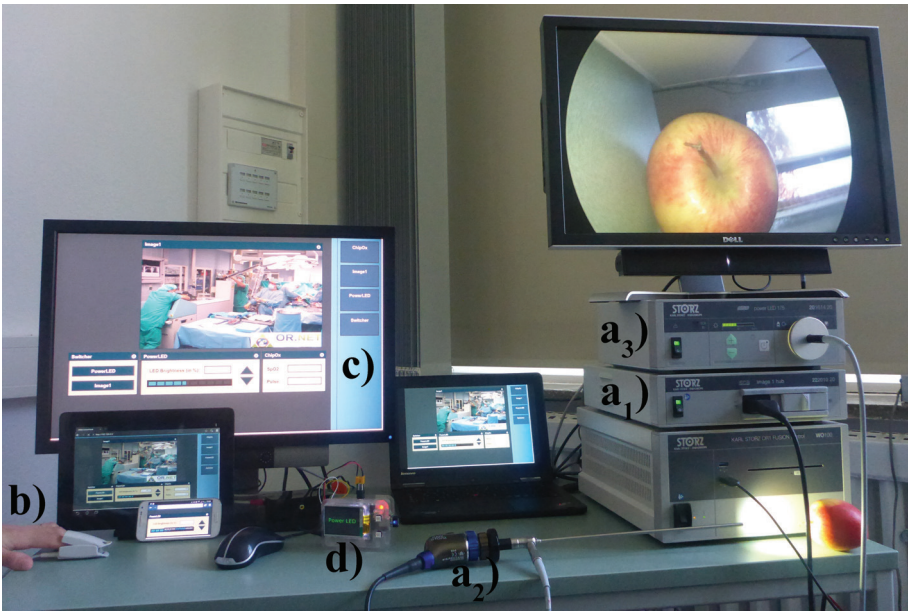


Abbildung 4: Demonstration einer prototypischen Umsetzung der DPWS-Integration von medizinischen Geräten. a_1) Endoskopkamera mit a_2) Kamerakopf und a_3) Kaltlichtquelle; b) Pulsoximeter-Fingerclip; c) Dashboard; d) Eingabeschalter.

ponente Legacy-Adapter) nicht beschrieben. Die DPWS-Adapter Softwarekomponente ist auf Basis des JMEDS[5] [Mat13] DPWS-Stacks realisiert.

Weiterer Bestandteil des Demonstrators ist ein Pulsoximeter (corscience ChipOx), dessen Fingerclip in der linken unteren Ecke von Abbildung 4, mit der Markierung b), zu sehen ist. Ein Pulsoximeter ermittelt die Vitalparamter Puls und Sauerstoffsättigung. Für dieses Gerät fungiert ein Raspberry Pi Board als DPWS-Konnektor Blackbox. Die geräteseitige Schnittstelle wird über den UART des Boards realisiert. Die Softwarekomponenten Legacy- und DPWS-Adapter sind in Java implementiert. Es wird ebenfalls der JMEDS DPWS-Stack genutzt. Es sind Services zum manuellen Lesen der aktuellen Werte von Puls und Sauerstoffsättigung realisiert, sowie Eventing-Services, die über die Änderungen von Puls bzw. Sauerstoffsättigung informieren.

Alle beschriebenen Geräte stellen außerdem einen für Medizingeräte üblichen Heartbeat zur Verfügung. Dieser informiert periodisch über die Präsenz der Geräte, sodass Ausfälle nach einer definierten Zeitschranke erkannt werden können. Die Realisierung erfolgt mithilfe von Eventing-Services.

Zur Anzeige und Steuerung der Geräte- und Vitaldaten dient ein Dashboard, das in Abbildung 4 mit c) hervorgehoben ist. Dieses Mensch-Maschine-Interface ist webbasiert umgesetzt, sodass es von beliebigen Geräten mit einem Webbrowser angezeigt und genutzt wer-

[5] Java Multi Edition DPWS Stack

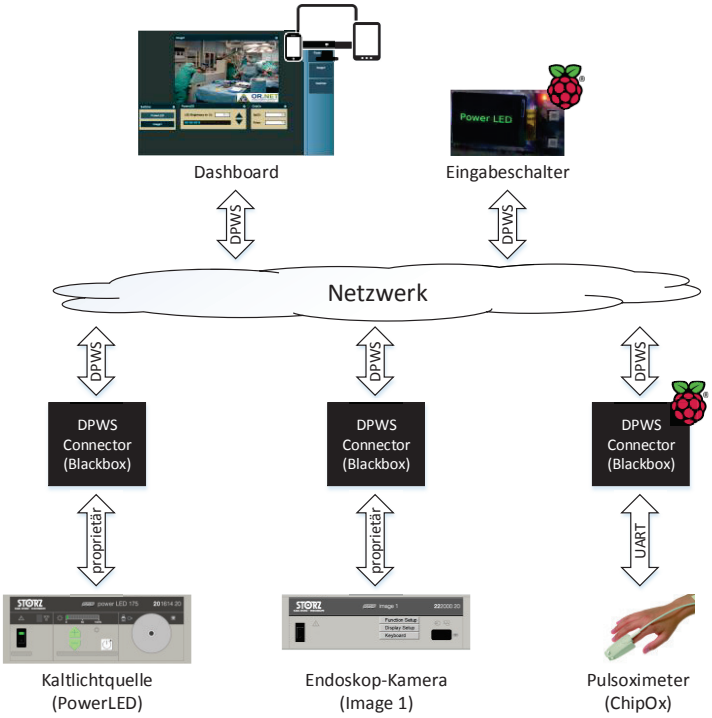


Abbildung 5: Schematische Darstellung des prototypischen Demonstrators.

den kann. Wie im mittleren und linken Teil von Abbildung 4 zu erkennen ist, werden exemplarisch ein Standard-PC, Laptop, Android-Tablet und -Smartphone genutzt. Der Webserver, der das Dashboard bereitstellt, implementiert die DPWS-Client-Seite. Es können so die beschriebenen Services genutzt werden. Als weiteres Steuergerät fungiert ein Raspberry Pi basierter Schalter mit zwei Tasten, der ebenfalls an der DPWS-Kommunikation teilnimmt. Dieser ist in Abbildung 4 mit d) gekennzeichnet. Über das Dashboard kann entschieden werden, ob der Schalter die Intensität der Kaltlichtquelle erhöht bzw. verringert oder ob durch einen Tastendruck ein Standbild von der Endoskopkamera abgeholt und auf dem Dashboard angezeigt werden soll.

Der vorgestellte Prototyp demonstriert wichtige Aspekte des zukünftigen OP-Saales: Einerseits wird eine herstellerunabhängige Vernetzung gezeigt. So erfolgt z. B. die Steuerung der Kaltlichtlichtquelle wahlweise über einen auf Raspberry Pi Basis realisierten Schalter oder über das webbasierte Dashboard, sodass beliebige Geräte genutzt werden können. Aus diesem Einsatz von Standardhardware kann abgeleitet werden, dass eine Kommunikation zwischen spezialisierten Herstellern in der DPWS-basierten Umgebung keine Herausforderung mehr darstellen wird. Ein weiterer Aspekt ist die dynamische Vernetzung. Durch die Möglichkeit mit einer Mensch-Maschine-Schnittstelle mehrere Geräte zu steuern, kann die Anzahl der Eingabegeräte reduziert werden. Dies führt zu einer geringeren

Kabelanzahl im OP, sodass beispielsweise das Risiko des Stolperns oder hygienische Probleme reduziert werden können.

Des Weiteren ermöglicht das vorgestellte Konzept des Dashboards neue Steuerungskonzepte und Möglichkeiten im Bereich der Telemedizin. In einem realen OP-Saal können deutlich mehr Daten und Parameter angezeigt und verändert werden, als im Prototyp gezeigt. Jeder beteiligte Akteur an einer Operation kann genau jene Informationen angezeigt bekommen, die benötigt werden. Diese können sich stark unterscheiden, beispielsweise zwischen Chirurgen und Anästhesisten. Durch die Zusammenführung aller Daten, die in vollen Umfang an Spezialisten an anderen Orten weitergegeben werden können, wird der Aspekt der intraoperativen Telemedizin neue Impulse erfahren.

6 Zusammenfassung und Ausblick

Die vorliegende Machbarkeitsstudie stellt einen Blackbox-Ansatz zur Integration von nicht DPWS-kompatiblen (medizinischen) Systemen in eine serviceorientierte Kommunikationsumgebung vor. Dies ermöglicht eine herstellerunabhängige und dynamische Vernetzung. Die Blackbox beinhaltet einen Legacy-Adapter, der die proprietäre Schnittstelle zum Gerät realisiert und einen DPWS-Adapter, der die Web-Service-Kommunikation implementiert. In einem prototypischen Demonstrator konnte die Funktionsfähigkeit unter Beweis gestellt werden.

Zukünftige Arbeiten werden sich mit der Verbesserung der Echtzeit-Eigenschaften der DPWS-Kommunikation beschäftigen. Dies ist sowohl für die Steuerung der medizinischen Geräte notwendig, als auch von Relevanz für den Big Data Aspekt der Velocity (Geschwindigkeit). In Bezug auf die semantische Interoperabilität wird mithilfe der Normenfamilie ISO/IEEE 11073 gearbeitet. Diese ist vor allem für den Bereich der Personal-Health-Geräte definiert. OP-Geräte sind derzeit nicht berücksichtigt, sodass eine Erweiterung angestrebt wird. Ebenso werden die Bestrebungen der Standardisierung des Kommunikationsprotokolles vorangetrieben und Aspekte der Zertifizierbarkeit von Medizinprodukten mit SOA-basierter Kommunikation betrachtet.

In einem OP-Saal fallen eine große Menge Daten an, die nur mit Hilfe einer (herstellerunabhängigen) Vernetzung genutzt werden kann. Mit Big Data Methoden können diese gezielt analysiert werden. So könnten beispielsweise relevante Daten für die Dokumentation der Operation gefiltert werden, um den Ärzten und dem OP-Personal die Arbeit zu erleichtern. Auch Aspekte der Qualitätssicherung und Optimierung der Abläufe sind relevante Fragestellungen, bis hin zur Weiterentwicklung der medizinischen Forschung an sich.

Acknowledgement

Diese Arbeit ist im Projekt OR.NET entstanden. OR.NET wird unter dem Förderkennzeichen 16KT1238 (im Förderprogramm IKT 2020) durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert. Der prototypische Demonstrator wurde mit Unterstützung der Firma KARL STORZ GmbH & Co. KG (Tuttlingen, Deutschland) entwickelt. Weiterer Dank gilt Priv.-Doz. Dr. med. Klaus F. Wagner (Chefarzt der Klinik für Anästhesiologie und Intensivmedizin des Klinikums Südstadt in Rostock) für die Beratung in medizinisch technischen Fragen.

Literatur

[Bey11] M. Beyer. Gartner Says Solving „Big Data“ Challenge Involves More Than Just Managing Volumes of Data, 2011. URL: http://www.gartner.com/newsroom/id/1731916. Datum: 18.05.2014.

[BINR10] J. Benzko, B. Ibach, M. Niggemeyer, K. Radermacher. A novel SOA based approach towards the integration of a robotic system into a modular surgical work system and IT network. In *Computer assisted radiology and surgery : proceedings of the 24rd international congress and exhibition*, Seiten 193–194, Geneva, Switzerland, 2010. Heinz U. Lemke. Heidelberg : Springer, 2010.

[BORN09] C. Buschmann, J. A. K. Ohnsorge, K. Radermacher, F. U. Niethard. Das Projekt orthoMIT. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 52(3):287–296, 2009. doi: 10.1007/s00103-009-0790-z. URL: http://dx.doi.org/10.1007/s00103-009-0790-z.

[dDCK+06] S. de Deugd, R. Carroll, K. E. Kelly, B. Millett, J. Ricker. SODA: Service Oriented Device Architecture. *Pervasive Computing, IEEE*, 5(3):94–96, 2006. doi: 10.1109/MPRV.2006.59.

[doo] DOOP-Projekt (Dienst-orientierte OP-Integration). URL: http://www.doop-projekt.de/. Datum: 22.05.2014.

[ER09] T. Erl, S. Roy. *SOA Design Patterns*. Prentice Hall, 2009. doi: 10.1016/j.artmed.2009.05.004.

[GBM12] D. Gregorczyk, T. Bußhaus, R. Mildner. Strategien zur Konstruktion von Medizingeräteprofilen. *White Paper*, 2012. URL: http://www.doop-projekt.de/tl_files/ebooks/medizingeraeteprofile/files/assets/downloads/publication.pdf.

[IBR10] B. Ibach, J. Benzko, K. Radermacher. OR-Integration based on SOA - Automatic detection of new Service Providers using DPWS. In *Proceedings of the 24rd international Congress and Exhibition: Computer assisted Radiology and Surgery (CARS)*, Seiten 195–196, Geneva, Switzerland, 2010. Heinz U. Lemke. Heidelberg : Springer, 2010.

[IBS+12] B. Ibach, J. Benzko, S. Schlichting, A. Zimolong, K. Radermacher. Integrating medical devices in the operating room using service-oriented architectures. *Biomedizinische Technik. Biomedical engineering*, 57(4):221–228, August 2012. doi: 10.1515/bmt-2011-0101. URL: http://www.degruyter.com/view/j/bmte.2012.57.issue-4/bmt-2011-0101/bmt-2011-0101.xml.

[iMuB] DGBMT Innovationen in Medizintechnik und BioEngineering. DGBMT - Innovationen in Medizintechnik und BioEngineering - smartOR. URL: http://www.vde.com/de/fg/DGBMT/Arbeitsgebiete/Projekte/Seiten/Medizintechnik-Biomedizintechnik-Wissenschaft-Forschung-Studium-Innovationen-smartOR.aspx.

[KBB13] C. Kücherer, J. Benzko, T. Bußhaus. Forschungsprojekt OR.NET: Abschied von Insellösungen. *Deutsches Ärzteblatt 2013; 110(46)*, 4/2013, 2013. URL: http://www.aerzteblatt.de/pdf.asp?id=149231.

[Lan01] D. Laney. 3D data management: Controlling data volume, velocity and variety. *Application Delivery Strategies - META Group Research Note*, 6, 2001.

[LS11] E. A. Lee, S. A. Seshia. *Introduction to Embedded Systems - A Cyber-Physical Systems Approach*. LeeSeshia.org, first edition (version 1.08). Auflage, 2011. URL: http://leeseshia.org.

[LSC+12] I. Lee, O. Sokolsky, S. Chen, J. Hatcliff, E. Jee, B. Kim, A. King, M. Mullen-Fortino, S. Park, A. Roederer, K. K. Venkatasubramanian. Challenges and Research Directions in Medical Cyber-Physical Systems. *Proceedings of the IEEE*, 100(1):75–90, Januar 2012. doi: 10.1109/JPROC.2011.2165270.

[Mat13] Materna GmbH. JMEDS (Java Multi Edition DPWS Stack) — Free Development software downloads at SourceForge.net, 2013. URL: http://sourceforge.net/projects/ws4d-javame/.

[MSLK09] C. Mauro, A. Sunyaev, J. M. Leimeister, H. Krcmar. Service-orientierte Integration medizinischer Geräte - eine State of the Art Analyse. *Proceedings of Wirtschaftsinformatik 2009 (WI 2009), Business Services: Konzepte, Technologien, Anwendungen*, Band 1:119–128, 2009.

[OAS09] OASIS. Devices Profile for Web Services Version 1.1 - OASIS Standard, 2009. URL: http://docs.oasis-open.org/ws-dd/dpws/1.1/os/wsdd-dpws-1.1-spec-os.html. Datum: 18.05.2014.

[orn12] Homepage: OR.NET - Sichere dynamische Vernetzung in Operationssaal und Klinik, 2012. URL: www.ornet.org. Datum: 20.05.2014.

[orn14] Broschüre: OR.NET - Sichere dynamische Vernetzung in Operationssaal und Klinik, 2014. URL: http://www.ornet.org/wp-content/uploads/2014/04/Broschüre_final_2.0.pdf. Datum: 22.05.2014.

[PSS+09] S. Pöhlsen, S. Schlichting, M. Strähle, F. Franz, C. Werner. A Concept for a Medical Device Plug-and-play Architecture Based on Web Services. *SIGBED Rev.*, 6(2):6:1—-6:7, 2009. doi: 10.1145/1859823.1859829. URL: http://doi.acm.org/10.1145/1859823.1859829.

[Sch13] S. Schlichting. BICEPS - Ein Protokoll zur Integration von Medizinprodukten am klinischen Arbeitsplatz. In *GMDS 2013. 58. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS)*, Lübeck, Germany, 2013.

[SP14a] S. Schlichting, S. Pöhlsen. An architecture for distributed systems of medical devices in high acuity environments - A Proposal for Standards Adoption. In *HL7, 2014. 11073/HL7 Standards Week*, San Antonio, Texas, USA, 2014.

[SP14b] S. Schlichting, S. Pöhlsen. An architecture for distributed systems of medical devices in high acuity environments - A technical whitepaper. *White Paper*, 2014. URL: http://sourceforge.net/projects/opensdc/files/framework/1.0-BETA_02/Documentation/TechnicalWhitepaper.pdf/download.

Sicherstellen semantischer Integrität im Big Data Umfeld – Problemstellung und Lösungsansatz

T. Hoppe[1,2], H. Eisenmann[2], A. Viehl[1], O. Bringmann[1,3], C. Hennig[1,2]

[1]FZI Forschungszentrum Informatik
Haid-und-Neu-Str. 10-14
76131 Karlsruhe
{hoppe,viehl,hennig}@fzi.de

[2]Airbus DS GmbH
88039 Friedrichshafen
harald.eisenmann@astrium.eads.net

[3]Universität Tübingen
Sand 13
72076 Tübingen
bringmann@informatik.uni-tuebingen.de

Abstract: In diesem Beitrag wird die Notwendigkeit des konsequenten Einsatzes von IT-gestützten Modellen im Systems Engineering, exemplarisch aus Sicht der Raumfahrtindustrie, motiviert und dabei auf aktuell existierende Anforderungen und Probleme eingegangen. Aus wissenschaftlich-technischer Sicht wird dabei auf Problemstellungen aus dem Bereich Big Data, insbesondere auf problematische Aspekte zur Etablierung einer konsistenten Semantik auf der Modellierungsebene und zur Laufzeit, fokussiert. Es wird ein Ausblick auf einen entwickelten Ansatz zur Addressierung der identifizierten Herausforderungen skizziert, welcher auf einem methodischen Vorgehensmodell und relevanten technischen Lösungsbausteinen basiert.

1 Motivation

Im Ingenieursbereich haben Modelle vielseitiger Art und Weise eine lange Tradition. Dazu zählen unter anderem Zeichnungen auf Papier, abstrahierte Miniaturnachbauten, bis hin zu voll funktionsfähigen Simulationsmodellen. In den letzten Jahrzehnten verzeichnen insbesondere IKT-gestützte Modelle einen rasanten Anstieg bei der Nutzung im Systems Engineering. Dies ist vor allem einer steigenden Unterstützung durch entsprechende Modellierungssoftware zu verdanken. Diese ermöglicht sowohl eine effektive und effiziente Erstellung von Modellen als auch eine einfache Anpassung bei notwendigen Änderungen. Außerdem wird eine Wiederverwendung und teilweise projektspezifische Anpassung ermöglicht. In diesem Kontext bilden Modelle vereinfachte Nachbildungen eines Systems, die der Lösung einer bestimmten Aufgabe dienen. Diese Aufgaben können bspw. kommu-

nikativer, analytischer oder demonstrativer Natur sein. Die eingesetzte Modellierungstechnologie hängt von den Anforderungen an das Modell selbst, der Erstellungskomplexität eines Modells, sowie weiteren Faktoren wie kommerziellen Aspekten ab.

Modellbasierte Lösungen, die den gesamten Entwicklungsprozess umfassen, sind heutzutage für jede Ingenieursdisziplin separat zu finden. Damit ist jeder Schritt von der Anforderungsspezifizierung, über Analyse, Design, Entwicklung, Produktion und Verifikation umfassend durch softwarebasierte Modelle unterstützt. Die Erfahrungen, die mit modellgetriebenem Engineering (engl. Model-Driven Engineering) im Softwarebereich gemacht wurden, sind in das klassische Systems Engineering transferiert worden. Daraus ist das modellbasierte Systems Engineering (engl. Model-Based Systems Engineering (MBSE) [FGS07]) entstanden. Für die Umsetzung von MBSE sind folgende Schritte notwendig:

1. Alle Modelle, die während des Systementwicklungsprozesses zum Einsatz kommen, müssen identifiziert und spezifiziert werden.
2. Für jedes Modell müssen die Schnittstellen mit dem Entwicklungsprozess und innerhalb der Domäne spezifiziert werden.
3. Um einen Informationsaustausch zwischen dem domänenspezifischen Modell und dem domänenübergreifendem Modell, dem sogenannten Systemmodell, zu ermöglichen, müssen die Schnittstellen miteinander verbunden werden.

Solch ein generelles Systemmodell bildet die Grundlage sowohl für den Informationsaustausch zwischen den einzelnen Domänen als auch für überblickhaftes Informationsmanagement auf Systemebene während des gesamten Systemlebenszyklus. Weiterhin muss das Systemmodell die Modellierung aller benötigten Abstraktionsniveaus und Entwurfsphasen unterstützen und von jeder am Entwicklungsprozess beteiligten Disziplin, den Kunden und den Zulieferern akzeptiert werden. Die einzelnen Disziplinen haben sich durchgesetzte, reife Werkzeuge, die alle domänenspezifischen Aspekte des Systems abbilden können und auch den Datenfluss innerhalb der Disziplin unterstützen. Eine Lösung, die den gesamten MBSE-Anwendungsbereich vom Erfassen der Anforderungen bis zur Auslieferung eines Systems umfasst, erweist sich als erheblich komplexer.

In diesem Beitrag werden die konkreten Herausforderungen von MBSE in der Raumfahrt vorgestellt und ein möglicher Lösungsweg skizziert. Dazu stellt das folgende Kapitel die aktuellen Probleme mit MBSE aus Sicht der Raumfahrtindustrie detailliert dar. Im dritten Kapitel wird ein Lösungsansatz präsentiert, mit dem die beschriebenen Probleme angegangen werden können. Das vierte Kapitel fasst die Probleme und den Lösungsansatz zusammen.

2 Aktuelle Probleme mit MBSE aus Sicht der Raumfahrtindustrie

In Teilen der Raumfahrtindustrie wird MBSE bereits angewendet [EMdK09], [dKEB10], [Eur12]. Allerdings müssen die folgenden Herausforderungen noch angegangen werden.

Diese liegen an der Schnittstelle der Themen Cyber Physical Systems und Big Data.

- Eine Vielzahl am Systementwicklungsprozess beteiligten Stakeholdern, wie Auftraggeber, direkt am System arbeitenden Engineering-Disziplinen, und Zulieferern
- Verwendung eigener Frameworks für die domänenspezifische Modellierung der Stakeholder
- Massiver Daten– und Wissensaustausch zwischen den beteiligten Stakeholdern
- Hohe Komplexität der Datenmodelle, die es erlauben, interdisziplinäre Abhängigkeiten zu erfassen und zu überprüfen
- Projektspezifische Anpassungen im Kontext geringer Stückzahlen innerhalb eines Projektes

In der Raumfahrt sind viele der zum Einsatz kommenden Systeme Prototypen, da die Einsatzzwecke oftmals sehr vielfälter Natur sind. Weiterhin werden die anfallenden Arbeiten oftmals geographisch voneinander getrennt erledigt. Als Beispiel sei die Rosetta Mission genannt, an der insgesamt 50 Firmen aus weltweit 14 Ländern beteiligt waren. Allein in Europa waren etwa 1000 Personen mit der Entwicklung und dem Bau von Rosetta beschäftigt. Es mussten 11 verschiedene wissenschaftliche Instrumente in die Raumsonde integriert werden, wozu charakteristische Systemparameter wie Dimensionen, Massen, Energiebedarf und Interfaces wie Befestigungspunkte und Kommunikationsprotokolle zwischen dem jeweiligen Hersteller eines Instrumentes und dem Hersteller der zentralen Komponenten der Raumsonde ausgetauscht werden mussten [FHT04].

Spezifisch für die Raumfahrt ist weiterhin, dass Tests unter realen Bedingungen, wie sie im Weltraum vorherrschen, nicht oder nur sehr eingeschränkt durchführbar sind. Weiterhin können kleinste Fehler bereits zum Scheitern einer Mission führen. Deshalb müssen die Systeme sehr fehlertolerant und teils redundant ausgelegt sein. Daraus resultiert ein noch höheres Maß an Komplexität.

2.1 Semantik auf Modellierungsebene

Die konzeptuelle Datenmodellierung dient als Grundlage für den Entwurf des im vorherigen Kapitels vorgestelltem Systemmodells. Dabei wird der Umfang der zu verwaltenden Daten und deren Abhängigkeiten untereinander von allen Beteiligten in einem konzeptuellen Datenmodell (engl. Conceptual Data Model (CDM)) erfasst. Das CDM fungiert dabei auch als Grundlage für einen Abgleich zwischen den Modellierern des CDMs und den beteiligten Personen aus den Engineeringdomänen. Weiterhin dient das CDM als Referenzimplementierung aller Werkzeuge, die später mit dem Systemmodell arbeiten werden [KOB07], [Oli07].

Bei, wie in der Raumfahrt üblich, projektspezifischen Zusammensetzungen von Domänen, Zulieferern und Kunden ist es deshalb essentiell, dass das CDM sowohl klar definierte

Grundbausteine vorgibt, als auch projektspezifische Anpassungen unterstützt, damit nicht für jedes Projekt ein neues CDM entworfen werden muss.

2.2 Semantik zur Laufzeit

Die Semantik, die in den Modellen der einzelnen Disziplinen, Zulieferern und Kunden steckt, muss zwischen diesen ausgetauscht werden. Dafür wird, wie im vorausgegangenen Abschnitt beschrieben, das CDM verwendet. Damit ergeben sich eine Vielzahl an Problemen, die im Folgenden erläutert werden:

Diskrepanz zwischen modellierter und implementierter Semantik Mit dem aktuellen Stand der Technik kann die vollständige Semantik der Daten nicht im CDM definiert werden, da entweder die verwendete Modellierungssprache semantisch nicht reich genug ist oder keine Generierungsmöglichkeiten für die definierten Bedingungen zur Verfügung steht. So ist Ecore beispielsweise semantisch schwach, bietet aber mit dem Eclipse Modeling Framework [SBMP08] eine starke Codegenerierungsplattform. Faktenbasierte Ansätze wie Object Role Modeling [HMM08], [Hal11] bieten zwar eine semantisch reichhaltigere Modellierung, aber die Transformation dieser Modelle in eine ausführbare Applikation ist bisher nur sehr eingeschränkt unterstützt [Hop14], [SV07], [PVN12]. Manchmal stehen auch Performanceansprüche an die zu generierende Applikation einer Transformation im Weg. Resultierend daraus wird die Semantik, die nicht aus dem CDM in die Implementierung übertragen werden konnte, direkt in der Implementierung umgesetzt. Sie ist somit nur implizit in der Implementierung vorhanden und nicht explizit im Modell spezifiziert.

Für den Nutzer eines MBSE–Werkzeugs ergibt sich somit eine Diskrepanz zwischen der Semantik, die bezogen auf das CDM erwartet wird, und der, die tatsächlich implementiert ist.

Datenaustausch zwischen domänenspezifischen Werkzeugen Viele Disziplinen setzen Standardsoftware für den modellbasierten Entwurf ein. Diese stammt von unterschiedlichen Herstellern, setzt auf verschiedenen Technologien auf, realisiert andersartige Paradigmen oder weist sonstige semantische Unterschiede auf. Aufgrund der Vielzahl von Disziplinen und externen Partnern im Raumfahrtbereich muss ein massiver Datenaustausch zwischen einer Vielzahl von heterogenen Werkzeugen stattfinden. Eine Übersicht über eine solche Infrastruktur ist schematisch in Abbildung 1 gegeben.

Der Datenaustausch ist allerdings durch die fehlende Anpassbarkeit der Standardsoftware problematisch. Deshalb wird in Teilen der Raumfahrtindustrie ein zentrales System eingesetzt und der Datenaustausch mit jedem domänenspezifischen Werkzeug muss in Form einer Modelltransformation vom applikationsspezifischen in das zentrale Datenmodell definiert werden. Bei der Umsetzung dieser Transformationen kann es zu semantischen Fehlinterpretationen einzelner Aspekte eines der beteiligten Datenmodelle kommen. Dadurch kann ein Import der Daten in das zentrale System und ein Export zum gleichen Werkzeug unproblematisch sein. Allerdings kann ein Import von Daten aus einem Werkzeug und

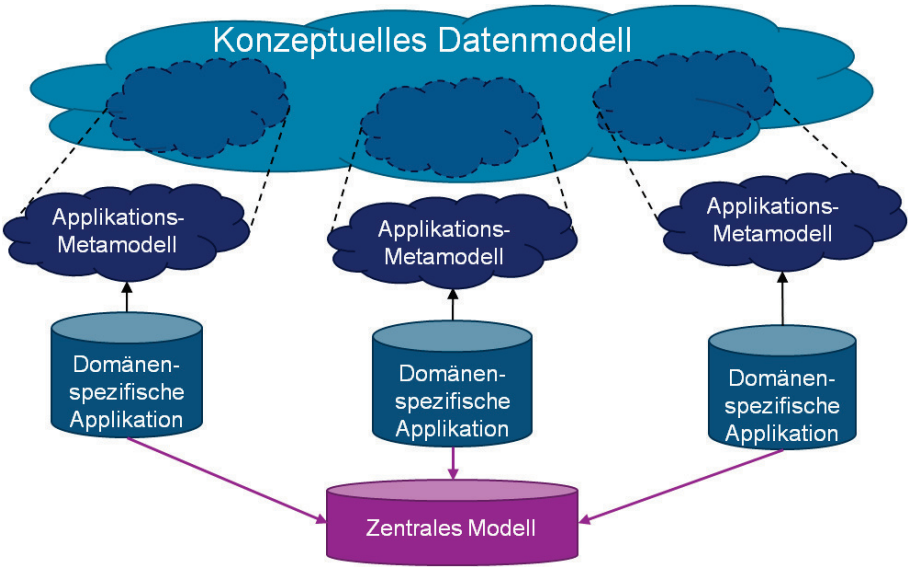


Abbildung 1: Anbindung domänenspezifischer Standardsoftware an eine zentrale Applikation. Jede der Domänen hat ihr eigenes Datenmodell welches vom jeweiligen Werkzeug implementiert wird. Die Summe aller systemübergreifend relevanten Daten ist im konzeptuellen Datenmodell festgehalten und bildet die Grundlage für das zentrale Model.

ein Export zu einem anderen Werkzeug in falsch interpretierten oder strukturierten Daten enden.

Ein Parameter ist mehr als ein Wert Die Komplexität der Datenstrukturen lässt sich anhand der zu betrachtenden Aspekte eines Parameters veranschaulichen. Dieser durchfährt im Ingenieursbereich oftmals eine Entwicklung von einer geschätzten Annahme hin zu einem Wert, der auf Daten basiert und immer weiter verfeinert wird, bis er letzendlich gemessen werden kann. Hierbei müssen sowohl Grenzwerte als auch Zugriffsrechte verwaltet werden. Des Weiteren müssen die einzelnen Werte eines Parameters jederzeit für alle bisherigen Phasen des Systemlebenszyklus zugreifbar sein. Außerdem werden Parameter teilweise kommentiert, um beispielsweise Beobachtungen, Wertherkunft oder Auffälligkeiten zu annotieren. Insbesondere Werte, die von Zulieferern stammen, müssen auf ihre Kompatibilität mit den in der Gesamtarchitektur festgelegten Grenzen überprüft werden und eventuelle Abweichungen müssen in beiderseitigem Übereinkommen korrigiert werden.

Datenverfeinerung während des Entwicklungsprozesses In der Raumfahrtindustrie ist der Systementwicklungsprozess in Phasen aufgeteilt, die jeweils einen fest definierten Systemabstraktionsgrad beschreiben. Die Daten einer Phase werden in der darauffolgenden Phase verfeinert. Im Bedarfsfall muss aber auch auf die Daten aus zurückliegenden Phasen zurückgegriffen werden, um beispielsweise die Auswirkungen von Änderungen zu

ermitteln. Außerdem können die Daten aus früheren Phasen unter Umständen für mehrere Projekte verwendet werden, wohingegen die Daten der späten Phasen projektspezifisch sind. Während sich die Daten von Phase zu Phase weiterentwickeln, müssen auch die Validierungsbedingungen angepasst werden, denn diese sind zum Teil ebenso phasenabhängig.

Die Ermittlung, welche Daten einer Phase in der nächstfolgenden auf welche Art und Weise verfeinert wurden, ist angesichts der Mannigfaltigkeit der verwendeten Datenmodelle und deren hochkomplexer innerer Struktur mit einer Vielzahl von Bedingungen keine triviale Aufgabe. Insbesondere dann, wenn eine Referenz aus der einen Phase in eine Vielzahl von Referenzen in der nächsten Phase aufgebrochen wird, um beispielsweise die Unterschiede der einzelnen Instanzen zu modellieren.

Projektspezifische Anpassungen des Datenmodells zur Laufzeit Im Raumfahrtbereich werden häufig nur Prototypen und Kleinstserien mit geringen Stückzahlen gefertigt. Dementsprechend sind kosteneffiziente, projektspezifische Anpassungen des Datenmodells zur Laufzeit notwendig. Dabei wird ein gewisser Rahmen durch das implementierte Datenmodell vorgegeben, welcher dann sowohl erweitert als auch beschnitten werden kann. Dabei beziehen sich die Erweiterungen auf Klassen, Attribute, Referenzen und auch die Vererbungshierarchie.

Multidomänendaten erfordern eine angepasste Visualisierung Die Daten, die aus einer einzelnen Disziplin stammen, können sehr gut durch die disziplinspezifischen Modelle dargestellt werden, die eine Analyse der Daten ermöglichen. Allerdings erfordert die Integration von Daten aus einer Vielzahl von Disziplinen eine passende Visualisierung. Diese sollte sowohl eine Einschränkung der darzustellenden Daten erlauben als auch ein automatisiertes Erstellen des passenden Layouts beinhalten.

Ontologische Ansätze zur Wissenserweiterung Die Datenintegration aus verschiedenen Quellen ermöglicht auch weitere Überprüfungen der Daten insbesondere bezüglich der Konsistenz zwischen den einzelnen Quellen. Somit kann beispielsweise eine Instanz im Datenmodell, die einen Wassertank im mechanischen Design darstellt, mit einer anderen Instanz im Datenmodell, die den gleichen Wassertank im thermischen Design repräsentiert, verbunden werden.

Aufgrund dieser Verbindungen ergeben sich sowohl Notwendigkeiten als auch Ausschlüsse, die im Datenmodell spezifiziert werden müssen, da zum Beispiel Aktoren und Sensoren per Definition disjunkt sein müssen. Ontologische Konzepte, die den Typ eines Objektes anhand der ihm zugewiesene Attribute bestimmen, sind Teil aktueller Forschungen [Hen12]. Sie wurden bisher unter anderem aufgrund der Komplexität, die sich aus der Vielzahl von möglichen Kombinationen ergibt, aber noch nicht implementiert.

Zentrales versus dezentrales Datenmanagement Für den Umgang mit den Daten zur Laufzeit ergeben sich zwei unterschiedliche Ansätze. Zum einen könnten die Daten durch eine zentrale Applikation verwaltet werden und zum anderen könnte jedes im MBSE–

Prozess involvierte Werkzeug die Verwaltung seiner Daten selbst organisieren. Zur Datenverwaltung zählen in beiden Fällen die Versionierung, das Definieren und Ausführen von Validierungsbedingungen, das Verwalten von projektspezifischen Anpassungen, das Bereitstellen von Daten für eine andere Applikation und Weiteres.

Bei einem zentralen Ansatz, wie er im Raumfahrtbereich aktuell angewendet wird, müssen die einzelnen Funktionen nur einmal realisiert werden und arbeiten auf dem zentralen Datenmodell. Des Weiteren werden Adapter benötigt, die die applikationsspezifischen Daten zwischen der spezifischen Form und der des zentralen Datenmodells transferieren. Diese realisieren den Austausch mit den Werkzeugen der involvierten Disziplinen und Partner. Bei einem dezentralen Ansatz müssen die Funktionen für jedes Datenmodell jeder beteiligten Applikation realisiert werden. Weiterhin muss der Datenaustausch zwischen allen involvierten Werkzeugen wechselseitig durch Adapter realisiert werden.

Der zentrale Ansatz bietet den Vorteil, dass neue Werkzeuge oder Partner nur eine Schnittstelle mit dem zentralen Datenmodell haben und nicht wie beim dezentralen Ansatz mit jedem Partner haben, mit dem Daten ausgetauscht werden müssen. Weiterhin bietet der zentrale Ansatz den Vorteil einer zentralen Datenverwaltung, wodurch das Abrufen von Daten nicht an die jeweiligen Quellen gebunden ist. Ebenfalls kann sichergestellt werden, dass alle Daten früherer Entwicklungsphasen gegebenenfalls wieder zur Verfügung stehen.

3 Entwicklung eines Lösungsansatzes

Die im vorherigen Kapitel beschriebenen Probleme sind miteinander verzahnt, so dass die unabhängige Betrachtung einzelner Lösungen für jedes der identifizierten Probleme nicht zu einer zufriedenstellenden Gesamtlösung führt. Vielmehr müssen die unterschiedlichen Dimensionen der Problemstellung identifiziert und die jeweiligen Teilbereiche analysiert werden. Die dadurch gewonnenen Erkenntnisse dienen zum einen dem Problemverständnis und zum anderen bilden sie die Grundlage für die Ausarbeitung einer Lösungsstrategie.

Um die beschriebenen Probleme anzugehen werden vier Dimensionen betrachtet:

Probleme In dieser Dimension werden die aktuellen Unzulänglichkeiten bei der Umsetzung von MBSE detailliert erfasst.

Funktionen Diese Dimension umfasst alle notwendigen Funktionen für die Umsetzung von MBSE.

Forschungsstand In dieser Dimension soll der aktuelle Stand der Forschung und Technik in Bezug auf die zu realisierenden Funktionen und die auftretenden Probleme eingeordnet werden.

Testumgebung Diese Dimension umfasst das Vorgehen, um einzelne Blöcke von Problemen in der Tiefe zu analysieren und mögliche Lösungen zu evaluieren.

Für jeden der zu untersuchenden Teilbereiche müssen vier zentrale Punkte mithilfe der Testumgebung herausgearbeitet werden:

1. Der aktuelle Stand der Technik in der Industrie muss erfasst werden.
2. Es müssen ein oder mehrere Lösungsansätze umgesetzt werden und miteinander verglichen werden.
3. Dabei ist es wichtig, die Lösungen sowohl isoliert als auch im Gesamtkontext zu betrachten, um mögliche Abhängigkeiten respektive spezielle Konfigurationen ausfindig zu machen, unter denen eine bestimmte Lösung schlechter oder besser funktioniert.
4. Die vorhanden Lösungen müssen im Kontext einer Gesamtlösung gegeneinander abgewogen werden.

Die im vorherigen Kapitel beschriebenen Probleme können in mehrere Blöcke aufgeteilt werden, die zwar nicht unabhängig voneinander sind, aber jeweils ihren individuellen Schwerpunkt haben. Diese Blöcke werden im Folgenden detailliert beschrieben, wobei die vorher genannten Dimensionen im jeweiligen Kontext erläutert werden.

3.1 Analyse des Status quo

Das Fundament für die Entwicklung von Lösungen für die in Kapitel 2 beschriebenen Probleme wird mithilfe einer Testumgebung gelegt, die den derzeitigen Stand widerspiegelt. Darauf aufbauend können dann Lösungen für die identifizierten Probleme entwickelt und evaluiert werden.

Um den derzeitigen Stand abzubilden, muss ein zu definierendes Datenmodell entworfen und implementiert werden. Des Weiteren müssen alle notwendigen Funktionalitäten, die für die Durchführung von MBSE notwendig sind, umgesetzt werden. Außerdem müssen die Schnittstellen zu externen Werkzeugen simuliert werden, um später die Qualität des Datenaustausches bewerten zu können.

Diese Testumgebung dient weiterhin dazu Evaluierungskriterien zu definieren, sodass die Lösungen, die entwickelt werden, auch bewertet werden können und somit untereinander vergleichbar sind. Dieser Prototyp dient weiterhin dazu, einen repräsentativen Testdatensatz zu entwickeln. Die Evaluierungsergebnisse, die mit dieser Testumgebung basierend auf dem zu definierenden Referenzdatensatz erstellt wurden, bilden die Grundlage und stellen eine untere Schranke dar, die von allen Lösungen mindestens erreicht werden muss. Nur dann sind diese als besser zu bewerten als der aktuelle Stand.

3.2 Semantisch starke konzeptuelle Datenmodelle zur Laufzeit

Das Ziel dieser Testumgebung ist, die Auswirkungen eines semantisch starken konzeptuellen Datenmodells auf die Applikation zu untersuchen. Dabei sind auch die Auswirkungen auf Codegenerierungsmechanismen zu analysieren. Die Durchsetzung der im Datenmodell definierten Bedingungen ist der Schlüssel zum Erfolg für dieses Szenario. Bestimmte Realisierungen könnten aus Performancegründen nicht praktikabel sein, da sich alles im Big Data Umfeld, mit einer großen Anzahl an Bedingungen und Instanzen, auf denen diese ausgewertet werden müssen, bewegt. Mit den gewonnenen Daten soll unter anderem evaluiert werden, was geeignete Technologien und Beschreibungsformen für semantisch starke Bedingungen sind und inwiefern sich neue Implementierungen bestehender Technologien finden lassen, die hinreichende Performance bieten. In diesem Hinblick sind z.B. die partielle oder ereignisbasiert Auswertung von Bedingungen interessant.

3.3 Grenzen von Datenmodellanpassungen zur Laufzeit

Wie in Kapitel 2 beschrieben, sind projektspezifische Anpassungen des Datenmodells innerhalb der Applikation essentiell. In dieser Testumgebung soll analysiert werden, wo die Grenzen für solche Anpassungen liegen und wie diese umgesetzt werden können.

Dazu muss eine Modellierungsplattform für die Anpassungen zur Verfügung gestellt werden, um eine grafische Repräsentation der Anpassungen für Diskussionsrunden und für die Dokumentation zu ermöglichen. Dabei muss auch sichergestellt werden, dass die Anpassungen konform zum statischen Teil des Datenmodells sind. Anschließend muss für die definierten Anpassungen eine Codegenerierung angeboten werden, um das vorhandene Werkzeug mit den definierten Anpassungen zu erweitern.

Des Weiteren soll das Verfeinern von Daten in dieser Testumgebung behandelt werden. Dabei steht vor allem das Abbilden von Daten einer Phase des Systementwicklungsprozesses auf eine vorherige Phase im Mittelpunkt. Dadurch hat man jederzeit Zugriff auf die Daten aller vorausgegangenen Phasen und kann Auswirkungen von Veränderungen analysieren.

3.4 Integration von Applikationen bei zentralen und dezentralen Ansätzen

Diese Testumgebung behandelt zwei Aspekte. Zum einen soll der Datenaustausch analysiert und verbessert werden und zum anderen soll der derzeitige zentrale Ansatz mit einem dezentralen Ansatz verglichen werden.

Der Datenaustausch zwischen einem domänenspezifischen Werkzeug und der zentralen Applikation erfordert eine bidirektionale Transformation der Daten zwischen domänenspezifischen Datenmodellen und zentralem Datenmodell. Dabei ist es essentiell, dass keines der Datenmodelle Raum für Interpretationen lässt, da dies sonst bei Nutzung der Daten

durch eine dritte Applikation zu Fehlern führt. In diesem Zusammenhang sollte auch der Einsatz eines Standards wie OSLC [Wor10] für die Datenübertragung untersucht werden. Außerdem sollte die Software, die die Transformation durchführt, dahingehend erweitert werden, dass jederzeit ermittelt werden kann, welche Daten mit welchem Werkzeug ausgetauscht wurden. Dies würde feingranulare Änderungsmitteilungen für die angebundenen Werkzeuge ermöglichen. Damit könnten Simulationen vereinfacht werden, indem unter Umständen diese nur für Teilbereiche eines Systems neu ausgeführt werden müssen.

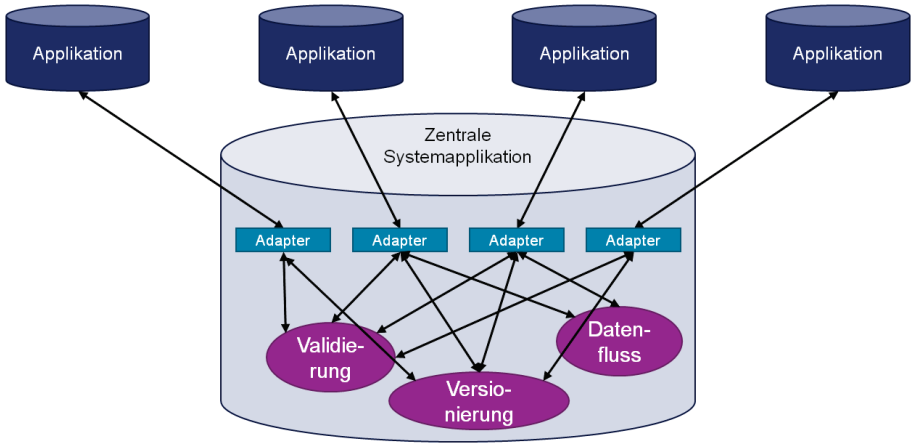


Abbildung 2: Ein zentraler Ansatz für MBSE wie er im Moment verwendet wird. Die Werkzeuge sind mit der zentralen Applikation verbunden und die Funktionalitäten sind beispielhaft abgebildet.

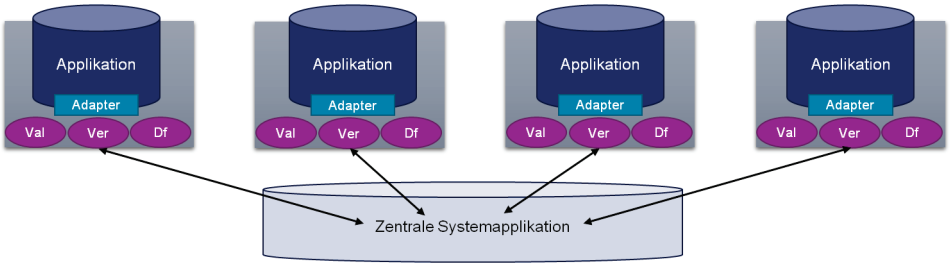


Abbildung 3: Ein dezentraler Ansatz, der die Funktionalitäten spezifisch für jedes Datenmodell eines Werkzeugs umsetzt und die zentrale Applikation nur ein Datencontainer ohne weitere Funktionalitäten ist. Die beispielhaft dargestellten Funktionalitäten sind Validierung (Val), Versionierung (Ver) und Datenfluss (Df).

Der zweite Teil dieser Testumgebung widmet sich der Frage, welcher Ansatz besser ist – zentral oder dezentral. Dazu muss zuerst ein dezentraler Ansatz umgesetzt werden. Anschließend können die Vor– und Nachteile beider Ansätze miteinander verglichen werden. Der zentrale Ansatz ist in Abbildung 2 und der dezentrale Ansatz ist in Abbildung 3 illustriert.

3.5 Datenorganisation im Hinblick auf den Datenfluss

In dieser Testumgebung steht das Management der Daten im Vordergrund. Als erstes soll untersucht werden, wie sich der Datensatz in mehrere Teile aufbrechen lässt. Dies ist essentiell, da ein Datensatz, der aus einer einzelnen Datei bestehen würde, nicht handhabbar ist, da er viel zu groß für den Speicher wäre. Dabei können verschiedene Aspekte betrachtet werden. Dazu zählen eine Aufteilung bezüglich der Produktstruktur, des Datenflusses oder der Art der Daten (Sensor, Aktor, Konfiguration, Realisierung,...).

Ein weiterer Punkt, der in dieser Testumgebung betrachtet werden soll, ist das gemeinsame Nutzen von Attributen durch mehrere Objekte. Dabei sollen die Vor– und Nachteile zweier möglicher Strategien herausgearbeitet werden: Zum einen könnte jedes Attribut nur virtuell geteilt werden. Zum anderen könnte der Wert für jeden, der darauf zugreift, dupliziert werden und Änderungen müssten synchronisiert werden.

Des Weiteren soll das Management der Metadaten beleuchtet werden. Das sind die Daten, die für die Durchführung des Prozesses notwendig aber nicht Teil des zu entwickelnden Systems sind.

3.6 Visualisierung heterogener Daten

Das Vereinigen von Daten aus verschiedenen Werkzeugen erfordert auch neue Methoden der Darstellung, welche in dieser Testumgebung näher betrachtet werden sollen.

Klassische tabellarische Ansichten ermöglichen nur geschulten Nutzern neue Zusammenhänge in den Daten zu erkennen. Für einen gezielten Erkenntnisgewinn sind Modelle viel besser geeignet. Die Darstellung der Daten in den einzelnen Werkzeugen kann sehr unterschiedlich sein. Doch wie soll die Visualisierung für die vereinigten Daten aussehen? Gibt es Grenzen? Welche Visualisierungsmethoden und –stile sind hilfreich? Diese Fragen müssen beantwortet werden um passende Visualisierungen zu finden.

Dabei sollte dem Nutzer auch die Möglichkeit geboten werden, die darzustellenden Referenzen und Attribute einzugrenzen, um bestimmte Aspekte eines Systems bewusst auszublenden oder hervorzuheben. Für die automatische Erzeugung der Visualisierung dieser Modelle ist ein automatisierter Layoutmechanismus essentiell.

4 Zusammenfassung

Modellbasiertes Systems Engineering (MBSE) wird in Teilen der Raumfahrtindustrie bereits erfolgreich eingesetzt. Die auftretenden Probleme und Schwächen sind allerdings nicht von der Hand zu weisen und müssen angegangen werden, um die Vorteile, die MBSE bietet, besser ausnutzen zu können.

Es gibt bereits Ansätze, die die Probleme auf konzeptueller Ebene angehen [Hen12],

[Hop14]. Allerdings treten viele Schwachstellen erst zur Laufzeit zu Tage. Die zentralen Aspekte sind:

- Auftritt einer semantischen Diskrepanz zwischen einem konzeptuellen Modell und der implementierenden Applikation
- Datenaustausch mit domänenspezifischen Standardwerkzeugen im Big Data Umfeld
- Mehrschichtige und komplexe Parameter
- Projektspezifische Anpassungen des Datenmodells zur Laufzeit
- Visualisierung heterogener Daten aus verschiedenen Quellen
- Wissenserweiterung und logische Ausschlüsse basierend auf vorhandenen Daten
- Zentrales versus dezentrales Datenmanagement

Um die genannten Probleme nachhaltig zu lösen reicht eine isolierte Betrachtung einzelner Probleme nicht aus, da diese miteinander verzahnt sind. Deshalb wird in diesem Beitrag ein Ansatz vorgestellt, der das Gesamtproblem in einzelne Teilbereiche herunterbricht, die jeweils im Gesamtkontext betrachtet werden. Zentrales Element des Lösungsansatzes sind kontrollierbare Testumgebungen, die eine Evaluierung des Ausgangsproblems und der entwickelten Lösungen ermöglichen. Dabei gilt es zunächst den aktuellen Stand zu erfassen und die Probleme zu verstehen. Anschließend können die einzelnen Teilbereiche analysiert und potentielle Lösungen entworfen sowie evaluiert werden. Dabei müssen auch untereinander auftretende Abhängigkeiten von Lösungsansätzen analysiert werden.

Literatur

[dKEB10] Hans Peter de Koning, Harald Eisenmann und Massimo Bandecchi. Evolving standardization supporting model based systems engineering. 2010.

[EMdK09] Harald Eisenmann, Juan Miro und Hans Peter de Koning. MBSE for European Space-Systems Development. *INCOSE Insight*, 12(4):47–53, 2009.

[Eur12] European Space Agency (ESA), http://vsd.esa.int/. *Virtual Spacecraft Design*, 2012.

[FGS07] Sanford Friedenthal, Regina Griego und Mark Sampson. INCOSE Model Based Systems Engineering (MBSE) Initiative. *INCOSE MBSE Track*, 2007.

[FHT04] Horst Fiebrich, JE Haines und Ferdinando Tonicello. Power System Design of the Rosetta Spacecraft. *AIAA Paper*, 5535:2004, 2004.

[Hal11] T. Halpin. Fact-Orientation and Conceptual Logic. In *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, Seiten 14–19. IEEE, 2011.

[Hen12] Christian Hennig. Evaluating Ontologies for Use in Model-Based Systems Engineering. Diplomarbeit, Technische Universität München, 2012.

[HMM08] T.A. Halpin, A.J. Morgan und T. Morgan. *Information modeling and relational databases*. Morgan Kaufmann, 2008.

[Hop14] Tobias Hoppe. Fact-Based Modeling for Eclipse Applications. Master's thesis, Hasso-Plattner Institut an der Universität Potsdam, 2014.

[KOB07] John Krogstie, Andreas Lothe Opdahl und Sjaak Brinkkemper. *Conceptual Modelling in Information Systems Engineering*. Springer, 2007.

[Oli07] Antoni Olivé. *Conceptual modeling of information systems*. Springer, 2007.

[PVN12] B. Piprani, S. Valera und G. M. Nijssen. Metamodel for Fact Based Information Model Registration. Bericht, Metadata Standards ISO/IEC JTC1 SC32 WG2, 2012.

[SBMP08] D. Steinberg, F. Budinsky, E. Merks und M. Paternostro. *EMF: eclipse modeling framework*. Addison-Wesley Professional, 2008.

[SV07] S. Sosunovas und O. Vasilecas. Tool-supported method for the extraction of OCL from ORM models. In *Business Information Systems*, Seiten 449–463. Springer, 2007.

[Wor10] OSLC Core Specification Workgroup. OSLC core specification version 2.0. *Open Services for Lifecycle Collaboration, Tech. Rep*, 2010.

Big Data im Störmanagement

Sebastian Schmitz, Julian Krenge

Informationsmanagement
FIR e.V. an der RWTH Aachen
Campus-Boulevard 55
52074 Aachen
Sebastian.Schmitz@fir.rwth-aachen.de
Julian.Krenge@fir.rwth-aachen.de

Abstract: Störungen in der Produktion sind die Ursache weitreichender Auswirkungen auf den Wertschöpfungsprozess. Mit der Vierten Industriellen Revolution (Industrie 4.0) und der damit verbundenen Informatisierung der Produktion nimmt die Generierung von Informationen über den gesamten Produktionsprozess weiter zu. Die vielfältigen und heterogenen Daten stellen die Unternehmen vor verschiedene Herausforderungen. Besonders das Fehlen eines Konzeptes für ein ganzheitliches Störungsmanagement verhindert die zielgerichtete Nutzung der generierten Daten. Big Data stellt hier einen technischen Enabler dar, der eine Antizipation und Früherkennung von Störungen im Produktionsprozess noch bevor diese eintreten ermöglicht und somit als Grundlage für ein proaktiven Störungsmanagement dient. Eine Entstörung kann somit früher und gezielter erfolgen.

1 Einleitung

Die zunehmende inner- und überbetriebliche Vernetzung einzelner Prozesse begründet die wirtschaftlich hohe Bedeutung von Störungen. Unabhängig von den Ursachen haben Störungen weitreichende Folgen, die sich bspw. für einen kritischen Maschinenausfall auf bis zu 50.000$/min. belaufen [AT06]. Je früher eine Störung erkannt wird, desto geringer sind die auftretenden Produktionsausfälle und die damit verbundenen Kosten. Störungen sind dabei definiert als zeitlich befristete Zustände in der Wertschöpfungskette, in denen durch das Einwirken von Störgrößen auf die Produktionsfaktoren und deren Kombinationsprozess eine unmittelbar festgelegte Abweichung vom optimalen Prozessverlauf und dessen Ergebnis entsteht [He94]. Störungen haben verschiedenen Ursachen, die sich nach unterschiedlichen Kriterien kategorisieren lassen. Eine mögliche Kategorisierung ist das Ishikawa-Diagramm, bei dem die Störungen nach den Einflussfaktoren Mensch, Maschine, Material, Methode und Umwelt sortiert werden [KB03].

Der Ablauf des Störungsmanagements lässt sich generell in zwei Phasen einteilen. Zunächst erfolgt eine Identifizierung und Analyse der Störung. Dabei kann es sich

sowohl um eine bereits aufgetretene Störung, als auch um präventive Maßnahmen handeln. Diese erste Phase hat das Ziel, die Ursache der Störung und deren Auswirkung auf das Gesamtsystem zu bestimmen. Darauf aufbauend lassen sich in der zweiten Phase Gegenmaßnahmen ableiten und durchführen. Ziel dieser zweiten Phase ist die effektive Auflösung der Störung.

2 Stand des Wissens

2.1 Bestehendes Störungsmanagement

Bedingt durch den Anstieg erwartbarer Störungen steigt die Notwendigkeit eines effizienten Störungsmanagements. Ein frühzeitiges Erkennen der Störung ist aufgrund der Vielzahl potentieller Störungen und oftmals fehlender Kenntnisse über die Ursachen sehr komplex [MK13]. Daher liegt der Fokus des Störungsmanagement in der Minimierung der Störauswirkung und weniger in der Prävention von Störungen. Viele kleine und mittelständige Unternehmen (KMU) verfügen aber über kein systematisches Störungsmanagement.

Aufgrund der Vielzahl auftretender Störungen in der Produktion ist der Faktor Mensch, also die Einbindung der Mitarbeiter und ihres impliziten Wissens, eine wichtige Komponente in bestehenden Störungsmanagementsystemen. Die Identifikation der Störungen erfolgt in diesen Fällen also durch die Mitarbeiter, die vor dem Hintergrund ihres Erfahrungswissens auf Basis der aktuellen Informationslage Entscheidungen bezüglich der Gegenmaßnahmen treffen. Umfassende Maßnahmen zur Bewertung und Ausbau von Kompetenzen der Mitarbeiter bewirken qualitativ hochwertigere Entscheidungen, um die Störung zu beseitigen [MK14]. Die Einbeziehung intelligenter Technologien zum präventiven Erkennen der Störungen werden nur unzureichend genutzt.

2.2 FMEA

Die Fehlermöglichkeits- und Einflussanalyse (FMEA) stellt eine analytische Vorgehensweise zur proaktiven System- und Risikoanalyse dar. FMEA beschriebt das Erkennen der Störungen, die Bewertung des Risikos sowie die Analyse des Eintrittes. Die FMEA-Methode ermöglich das Aufdecken vorhandener Schwachstellen, um systematisch Fehlerzusammenhänge zu bündeln [GJ04]. In einem aktiven Störungsmanagement ist FMEA eine geeignete Methode, Störungen frühzeitig zu erkennen und effektive Gegenmaßnahmen einzuleiten.

2.3 Big Data Lösungen in der prädiktiven Instandhaltung

In der prädiktiven Instandhaltung wird mit Algorithmen und auf Basis historischer Daten die Ausfallwahrscheinlichkeit von Komponenten bestimmt. Sensoren ermöglichen über die Analyse von Mustern die Kenntnis über den Zustand [Br13].

Die Möglichkeit der Verarbeitung großer Datenmengen zu Informationen wird unter dem Begriff „Big Data“ zusammengefasst. Hinter dem Begriff verbergen sich die 4 “V“. Volume beschreibt die steigende Datenmenge, die verstärkt durch vernetze Maschinen und Sensorik gewonnen werden. Diese Daten liegen in heterogenen Strukturen (Variety) vor, da Daten unterschiedlicher Quellen bereitgestellt werden. Die unterschiedlichen Quellen erzeugen eine Unsicherheit, sodass ein weiteres Attribut die Zuverlässigkeit (Veracity) der Daten ist. Die hohe Geschwindigkeit (Velocity), mit der die Daten verarbeitet werden müssen, wird durch eine hohe Reaktionsgeschwindigkeit begründet [VJ12]. Für Unternehmen werden die Daten als zusätzliche Ressource zunehmend bedeutender [Bi13].

Der Einsatz von Big Data in der Instandhaltung ermöglicht ein aktives Störmanagement, bei dem die Störungen präventiv erkannt und Gegenmaßnahmen eingeleitet werden können.

2.4 Complex Event Processing

Event-Driven Architecture (EDA) ist ein ereignisorientierter Architekturstil für Unternehmensanwendungen mit dem Ziel, die Agilität, die Reaktionsfähigkeit und die Echtzeitfähigkeit der Geschäftsprozesse eines Unternehmens zu erhöhen. Die Softwarekomponente CEP (Complex Event Processing) als mögliches Element einer EDA beschreibt die direkte und echtzeitfähige Verarbeitung von Unternehmensereignissen [BD10]. Die strukturierte Analyse der Daten ermöglicht das Erkennen von Mustern,. die sich aus der zeitlichen, räumlichen oder kausalen Korrelation von Ereignissen ergeben. Eine direkte Weiterverarbeitung macht eine Speicherung der feilgranularen Daten in einer Datenbank unnötig. Die gewonnen Informationen lassen sich bedarfsgerecht an den richtigen Abnehmer weiterleiten [Lu12].

Complex Event Processing lässt sich wegen dieser genannten Eigenschaften als Big Data-Werkzeug verstehen [Ha13]. Durch Big Data ergeben sich neue Herausforderungen an das Complex Event Processing bezüglich des Volumens und der Vielfalt der Daten, die jedoch auch die Möglichkeit umfasst, den Mehrwert der Ressource Daten zu nutzen.

3 Big Data-basiertes Störungsmanagement für produzierende Unternehmen

Dieses Paper soll ein Konzept für ein Störungsmanagement für produzierende Unternehmen auf Basis von Big Data-Technologie vorstellen. Dieses wird im Folgenden skizziert.

3.1 Gesamtbild eines übergreifendes Störungsmanagements

Um ein übergreifendes Störungsmanagement für die gesamte Produktion einzuführen, sind zunächst sowohl die Fachbereiche als auch die IT zu involvieren. Bisher konnten Optimierungen in der Produktion vor allem durch die Expertise der Fachbereiche vorangetrieben werden, die IT war nach Abschluss der Konzeptionsphase vor allem mit der informationstechnischen Umsetzung der entwickelten Konzepte beteiligt. Doch aufgrund der hohen erforderlichen IT-Kompetenz beim Einsatz von Big Data-Systemen wandelt sich dieser Aspekt. Die IT muss hier als aktiver Sparringspartner gesehen werden [HS12] und bereits früh in die Konzeptentwicklung für das Störungsmanagement eingebunden werden. Zusammenfassen sollten also sowohl die Stakeholder aus den Fachbereichen und aus der IT eingebunden werden.

Ferner müssen bei der Konzeption zwei unterschiedliche Aspekte des Störungsmanagements betrachtet werden. Zum einen muss die Erhebung der Daten geplant und implementiert werden. Doch dies muss in Zusammenarbeit mit der Erarbeitung des Konzepts zur Behebung der Störungen erfolgen. Es sollte sich an dieser Stelle also nicht blind auf die Technologie verlassen werden: Die zu behandelnden Störungen müssen klar formuliert sein, die zur Erkennung dieser Störungen notwendigen Informationsbedarfe müssen bekannt sein und das Konzept zur Entstörung muss formuliert sein. Das heißt, Störungen werden bezüglich ihrer Bedeutung für den Produktionsablauf und ihrer tatsächlichen Entstörbarkeit priorisiert, für diese Störungen werden dann die konkreten Informationsbedarfe abgeleitet. Nur so kann eine zielgerichtete Einführung der IT zur Störungsidentifikation durchgeführt werden. Zusammenfassend müssen also sowohl Analyse als auch Behebung der Störung betrachtet werden, siehe Abbildung 1.

		Stakeholder	
		Fachbereiche	IT-Abteilung
Phase	Analyse	Was sind die benötigten Informationen zur Erkennung der Störungen?	Sind die benötigten Daten effizient erhebbar?
	Behebung	Welche Störungen können wie entstört werden?	Wie können die Störungen aus den Daten identifiziert werden?

Abbildung 1: Fragestellung bei der Einführung von Störungsmanagement in der Industrie auf Basis von Big Data

Bei Auswahl einer Big Data-Lösungen sind weiterhin unterschiedliche Fokussierungen festzulegen, um das technische Grobkonzept festzulegen [Li14]. Dabei ist zunächst festzulegen, ob die Auswertung der Daten historisch oder zeitnah erfolgen soll. Eine historische Auswertung stellt dabei oft den initialen Schritt dar, kann aber für viele Anwendungsbereiche hinreichend sein, wenn zum Beispiel Kundenverhalten analysiert

wird. Für den hier skizzierten Anwendungsfall des Störungsmanagements hingegen ist eine zeitnahe Auswertung erforderlich, da die Störungen möglichst schnell behoben werden sollen, um ihren Einfluss auf das Gesamtsystem zu minimieren [Vo11].

Im Bereich der Datenanalyse ist ebenfalls zwischen einer Fokussierung auf automatisierte oder manuelle Vorgehensweise zu unterscheiden [LD14]. Die manuelle Analyse wird dabei durch einen Experten durchgeführt und eignet sich vor allem für die Konzeptionsphase oder selten durchgeführte Auswertungen. Beim Störungsmanagement hingegen ist eine automatische Analyse notwendig, um die zahlreichen und unterschiedlichen Störungen erkennen und in Folge bearbeiten zu können.

Eine dritte grundsätzliche Konzeptentscheidung ist die zwischen aktiven Datenabfragen und passiven Ereignisse [LD14]. Von aktiven Datenabfragen wird gesprochen, wenn die Big Data-Analyse durch einen Nutzer oder durch ein anderes Ereignis angestoßen wird. Dies ist besonders in frühen Umsetzungsphasen zu empfehlen, wenn die automatisch identifizierten Störungen noch durch zusätzliche Analysen verifiziert werden sollen. Im Störungsmanagement sind jedoch langfristig passive Ereignisse notwendig, d. h. die Big Data-Analyse wertet fortwährend die Daten der Produktionsstätte aus und informiert automatisch die Entstörungsprozesse.

3.2 Skizze der Lösung: Fehleranalyse

Die automatische Erfassung von Maschinen- und Betriebsdaten ist auf Fertigungsebene mittlerweile Standard [GG13]. Im Rahmen von Industrie 4.0, dessen Ziel eine intelligente Vernetzung zur Verfügung stehender Datenquellen ist, wird die Anzahl der generierten Daten weiter zunehmen [An11]. Aufgrund der Tatsache, dass bestehende IT-Systeme nicht in der Lage sind, diese Daten zu verarbeiten, werden diese Daten nicht verwendet [Kü13]. Im Produktionskontext können diese Daten eingesetzt werden, um Produktionsausfälle zu verhindern. Die Produktionsausfälle haben sehr unterschiedliche Ursachen, wie beispielsweise Lieferengpässe oder Maschinendefekte. Eine valide Datengrundlage ermöglicht eine rechtzeitige Reaktion, sodass die Störauswirkungen gering sind oder im Vorfeld vermieden werden.

Trotz der intelligenten Technologien ist der Mensch durch seine assoziativen und sensorischen Fähigkeiten weiterhin ein wichtiger Faktor im Störungsmanagement [Ga13]. Vor allem menschliche Fehler in der Produktion, hervorgerufen durch Stress oder fehlende Motivation, lassen sich gar nicht oder nur sehr schwierig mit Sensorik erfassen. Mit der CES-Technologie stehen Möglichkeiten zur Verfügung, mit denen bereitgestellte Meta-Daten, wie beispielsweise Sprache oder Text, erfasst werden können. Durch den Abgleich mit einem Pool möglicher Optionen, wird eine Abweichung von der erwartbaren Handlung analysiert.

Die Integration intelligenter Sensorik und dem Faktor Mensch ermöglicht einen neuen Ansatz im Störungsmanagement, indem Störungen antizipiert werden. Dazu müssen Maschinendaten, Umgebungsdaten sowie Metadaten der Mitarbeiter erfasst und miteinander in Relation gesetzt werden, um mögliche Störungen des Produktionsprozesses frühzeitig zu erkennen. Es ist zu erwarten, dass Mitarbeiter

Abweichungen hinsichtlich der Maschinensteuerung oder der Geräuchentwicklung beobachten, die so durch Sensoren nicht erfassbar sind bzw. ohne die Kontextinformation des Mitarbeiters keinen ausreichenden Hinweis auf eine Störung liefern. Damit ist die Grundlage für ein proaktives Störungsmanagement gelegt. Unternehmen können aufgrund der soliden Datengrundlage früher reagieren und damit Kosten und Risiken senken. Verbunden damit ist eine echtzeitfähige Verarbeitung der Daten. Big-Data ist als Lösung geeignet, die heterogenen und unstrukturierte Daten zielgerecht zu verarbeiten. Das Complex Event Processing ist ein wichtiger Baustein dieses proaktiven Störungsmanagements, um Verknüpfungen und Muster zwischen den Daten herzustellen. Weiterhin müssen Ereignismuster erkannt werden, bevor die Störungen auftreten. Störungen an Maschinen können sich z.B. durch ein erhöhtes Vibrationsniveau an der Maschine oder ein verändertes Geräuschlevel in der Maschinenumgebung ankündigen.

Die Daten spielen bei der Störungserkennung eine übergeordnete Rolle. Aus einer fachlichen Perspektive müssen daher zunächst die benötigten bzw. involvierten Informationsobjekte identifiziert werden. Ferner ist es erforderlich zu ermitteln, welche Informationstechnologien erforderlich sind, um die relevante Informationsobjekte zu erzeugen. Außerdem werden Attribute definiert, die Aussagen über die Informationsqualität erlauben und als Referenzwert für die Beurteilung der Güte dienen. Für das Störungsmanagement kann der Referenzwert genutzt werden, um die Wahrscheinlichkeit des Störungseintretens zusätzlich zu bewerten. Für die Entscheidungsunterstützung ist eine Aussage über die Informationsqualität ebenfalls von großer Bedeutung, um Entscheidungen basierend auf schlechten Informationen zu vermeiden.

Bestehende Complex Event Processing Systeme stehen vor der Herausforderung, dass die Muster manuelle erzeugt und deklarativ beschrieben werden, sodass die Muster nur bei exakter Übereinstimmung erkannt werden. Ein semantisch-basiertes CEP Management ermöglicht, dass Muster schlussgefolgert, aufgegriffen, angepasst und weiterentwickelt werden. Dieses selbstlernende Monitoring kann zur Verfeinerung der relevanten Ereignismuster eingesetzt werden. Die komplexen Ereignismuster werden regelbasiert beschrieben und auf dieser Basis werden ereignisgetriebene Berechnungen durchgeführt. Eine proaktive Mustererkennung ermöglicht, Störungen anhand von Teilmustern zu identifizieren. Dieser Prozess benötigt die Verwendung von Hintergrundwissen. Es besteht die Herausforderung, kontinuierlich veränderte Muster auf Basis von strukturellen Eigenschaften und Nutzungsinformationen zu erstellen. Eine transparente Darstellung der Muster für den Benutzer ermöglicht, diese erzeugten Muster kontinuierlich zu verbessern. Damit können unterschiedliche Vorhersagen in Bezug auf mögliche Störungen getroffen werden, sodass das Unternehmen in kritischen Situationen proaktiv reagieren können.

Wird über ein Muster eine Störung erkannt, wird diese Situation einer proaktiven Online FMEA-Analyse unterzogen, die Auskunft über die Eintrittswahrscheinlichkeit und die Auswirkung der erkannten möglichen Störung bestimmt. Für Störungen mit einer hohen Eintrittswahrscheinlichkeit und weitreichenden Auswirkung werden automatisch Gegenmaßnahmen vorgeschlagen. Hierfür werden zunächst die einzelnen Ebenen

hinsichtlich ihrer Rahmenbedingungen und Parameter untersucht und basierend darauf geeignete Optimierungsmodelle aufgestellt. Weiterhin werden für die einzelnen Ebenen geeignete Lösungsverfahren konzipiert und implementiert, die dynamisch, ausgehend von der aktuellen Ist-Situation neue Maßnahmen zur Behebung einer Störung generieren. Ausgehend von einem zur Verfügung stehenden Maßnahmenkatalog werden die zur Verfügung stehende Reaktionen abgeleitet. Bei der Wahl der Reaktion darf nicht nur die direkte Auswirkung auf die Störung, sondern muss auf den gesamten Prozess betrachtet werden. Eine umfassende Complex Online Optimization ermöglicht eine dynamische Anpassung der Produktionspläne, wie beispielsweise dem Vorziehen von Prozessschritten. Zur Bewertung der Reaktion lassen sich Kennzahlen, die die Zusatzkosten und Verzögerung, definieren.

Die abgeleitete Reaktion auf die Störung muss in einer engen Verknüpfung mit der CEP-basierten Störungserkennung stehen, um den späteren Erfolg der Reaktion ableiten zu können. Wird die Störung trotz Behebung weiterhin erkannt, kann dies bedeuten, dass die Reaktionsmaßnahme nicht erfolgreich war.

3.3 Skizze der Lösung: Störungsbehebung

Mit der proaktiven online FMEA ergibt sich eine Möglichkeit die Störung hinsichtlich der Eintrittswahrscheinlichkeit und der Auswirkung zu bewerten. Werden antizipierte und reale Störungen betrachtet, können Störungen in großer Anzahl auftreten. Bei den antizipierten Störungen ist dabei zusätzlich eine Bewertung, wann die Störung eintritt, wichtig. Auf Grundlage dessen lässt sich die Priorität der Störung bewerten, sodass ein Mitarbeiter die Störung beheben kann.

Um die Mitarbeiter zielgerichtet einzusetzen, muss zunächst eine automatische Auswahl getroffen werden. Eine Störung kann unterschiedliche Ursachen besitzen, beispielsweise fehlendes Material oder eine defekte Maschine. Ist die Ursache der Störung analysiert, lässt sich kontextbasiert der richtige Personenkreis ausmachen, der die Störung beheben kann. So wird die Meldung über die Störung im Fall von fehlendem Material an die interne Logistik, im Fall eines Maschinenausfalls an Mitarbeiter der Instandhaltung weitergeleitet.

Um möglichst effektiv auf die Störung zu reagieren, wird der dazu passende Mitarbeiter gesucht. Kriterien, nach dem die Auswahl erfolgt, sind die Erfahrung mit dieser speziellen Störung, der momentanen Verfügbarkeit und dem aktuellen Aufenthaltsort des Mitarbeiters. Im Mitarbeiterprofil wird die allgemeine berufliche Erfahrung sowie verschiede Zertifikate gespeichert. Damit bestimmt sich, welche Mitarbeiter generell für die Behebung der Störung genutzt werden können. Die weitere Auswahl erfolgt über die momentane Verfügbarkeit. Zudem kann der aktuelle Aufenthaltsort des Mitarbeiters bewertet werden. Besonders in großen Anlagen, müssen die Mitarbeiter unter Umständen weite Strecken zurücklegen. Bei der Auswahl des Mitarbeiters muss auf die Datensicherheit bzw. den Datenschutz besonders großer Wert gelegt werden, da hier mit sehr sensiblen Daten gearbeitet wird. Ein mögliches Umsetzungskonzept ist die Daten nicht auf einzelne Mitarbeiter, sondern nur auf Gruppen von Mitarbeitern zu beziehen.

Zudem muss bei der Lokalisierung der Mitarbeiter nicht der genaue Ort betrachtet werden, es reicht das Areal zu kennen, in dem der Mitarbeiter sich momentan aufhält.

Nach der Auswahl eines Mitarbeiters werden diesem kontextbasierte Informationen bereitgestellt, die zur Behebung der Störung relevant sind. Derzeit werden die Informationen innerhalb Fabrikanlagen oft nicht bedarfsgerecht übermittelt, sodass die Mitarbeiter nur unzureichende Informationen erhalten. Auf dessen mobilen Endgerät lassen sich Detailinformationen zu der identifizierten Störung und der angestrebten Lösung anzeigen. Gleichzeit kann dem Mitarbeiter Informationen der Aufenthaltsort benötigter Objekte, beispielsweise Werkzeug oder Ersatzteile mitgeliefert werden. Mit dieser kontextbasierten Informationsdistribution werden die Reaktionsprozesse auf Störungen effizienter und weniger fehleranfällig.

3.4 Skizze der Lösung: Zwischenfazit

Mit Hilfe von Big Data ergibt sich ein neuer Ansatz im Störungsmanagement. Die Analyse aller zur Verfügung stehender Daten ermöglicht, potentielle Störungen zu identifizieren, bevor die Störung Auswirkung auf den Produktionsprozess hat. Abbildung 2 fasst den Ablauf zusammen.

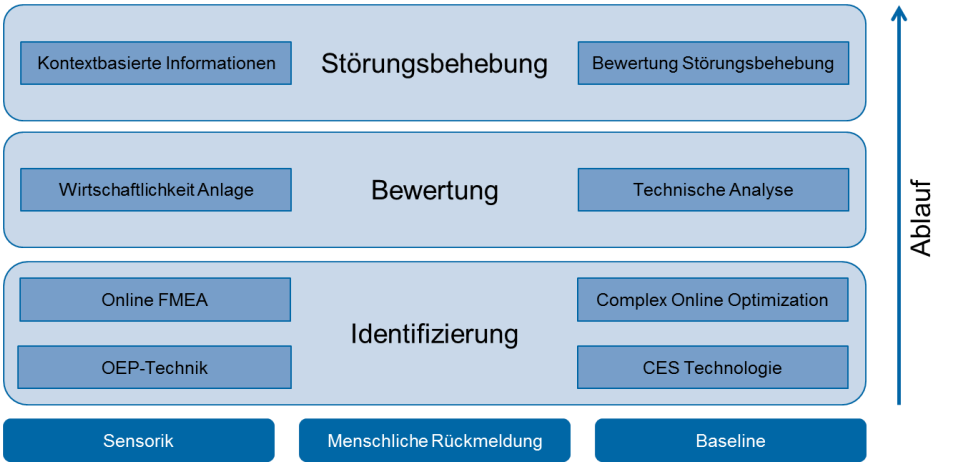


Abbildung 2: Aufbau des Störungsmanagements

Big Data-Lösungen stehen in der praktischen Anwendungen noch verschiedenen Herausforderungen gegenüber, wie LIGGESMEYER, DÖRR UND HEIDRICH aufführen. Im Folgenden sollen diese Herausforderungen auf praktischer und wissenschaftlicher Ebene für die hier skizzierte Lösung diskutiert werden.

4.1 Praktische Herausforderungen

Eine wesentliche praktische Herausforderung ist die Existenz einer geeigneten technischen Infrastruktur [LD14] einerseits zu Erhebung der Daten, andererseits zur Ausführung der Big Data-Analysen. Während der letztere Aspekt für Unternehmen in der Regel aufgrund der bereits bestehenden betrieblichen Anwendungssysteme einfach zu stemmen ist, die Betriebsdatenerfassung hingegen muss in den meisten Unternehmen erweitert werden. Das heißt, eine informationstechnische Anbindung aller zu überwachenden Entitäten mit einem hinreichend Informationsumfang muss gewährleistet sein. Big Data bietet hier einen deutlichen Vorteil, da unstrukturierte Daten verschiedener Quellen genutzt, und somit die eventuell notwendigen Investition in Sensorik oder Terminals in der Produktionsstätte minimiert werden können.

Weiterhin müssen für Big Data-Lösungen die Anwendungsfälle klar definiert sein [LD14]. Dies gilt im Störungsmanagement insbesondere für die zu behandelnden Störungen, wie bereits oben dargestellt. Der übergeordnete Anwendungsfall des industriellen Störungsmanagements hingegen zeigt großes Potenzial.

Eine weitere praktische Herausforderung für Big Data-Lösungen ist die Standardisierung der verschiedenen Aspekte [LD14]. Das heißt, um Big Data-Lösungen in der Industrie zur Anwendung zu verhelfen, müssen diese mit den etablierten Standards, de-facto Standards und Normen harmonieren. In der produzierenden Industrie herrscht eine hoher Grad an Standardisierung bspw. durch wissenschaftliche Referenzprozessmodelle und Datenübertragungsstandards wie EDIFACT. Da das Störungsmanagement im ersten Schritt vor allem für einzelne Produktionsstätte Anwendung finden wird, besteht hier eine besondere Gefahr von der Entwicklung proprietärer Lösungen. Dies muss für Störungsmanagement adressiert werden, die hier skizzierte Lösung bildet dabei nur den ersten Schritt.

4.2 Wissenschaftliche Herausforderungen

Auf wissenschaftlicher Ebene ist die erste Herausforderung die Entwicklung geeigneter Entscheidungsmodelle [LD14]. Die proaktive online-FMEA, skizziert im vorliegenden Beitrag, stellt dabei genau diese Entscheidungsunterstützung für die praktische Anwendung dar. Die Datenvisualisierung der Ergebnisse der Big Data-Analysen ist eine weitere wesentliche Herausforderung, gerade in der Kombination mit dem Bedarf an Entscheidungsmodellen [LD14]. Diese muss allerdings für konkrete Störungen durchgeführt werden und findet daher im Zuge der Konzeption zur Entstörung statt. Ein Beispiel ist unten gegeben.

Eine weitere Herausforderung ist die Daten- und Informationsqualität [LD14]. Dies soll im vorliegenden Ansatz durch die Einbindung der menschlichen Komponente adressiert werden. D. h. es werden die zwar unstrukturierten, aber sehr wertvollen Rückmeldungen der Mitarbeiter in der Produktionsstätte eingebunden. Die Einbindung von benutzerzentrierten Informationen verschärft aber die Problematik der Datennutzungskontrolle [LD14]. Diese Problematik ist in der hier skizzierten Lösung noch nicht abschließend beantwortet.

5 Beispiel: Erhöhter Energieverbrauch aufgrund Verschleiß

Der diskutierte Ansatz soll am Beispiel einer Werkzeugmaschine gezeigt werden. Eine Hauptursache für den Ausfall von Maschinen und Geräten ist Verschleiß [Po09]. Unter Verschleiß wird der fortschreitenden Materialverlust aus der Oberfläche eines festen Körpers verstanden, der durch Kontakt- und Relationsbewegung hervorgerufen wird [CG07]. Studien ergeben, dass durch Verschleiß mehr Energie für die selbe Aufgabe benötigt wird [CN04].

Smart Meter Sensoren ermöglichen, den momentanen Energiebedarf einer Werkzeugmaschine zu erfassen. Dieser lässt sich mit den Vergangenheitswerten für den selben Herstellungsprozess abgleichen. Steigt der momentane Verbrauch im Vergleich zu den historischen Werten an, kann dies auf Verschleiß zurückgeführt werden, vergleiche Abbildung 3.

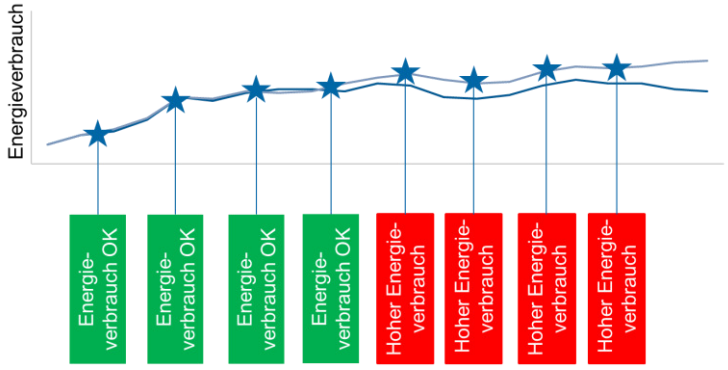


Abbildung 3: Auswirkung des Verschleiß auf den Energieverbrauch

Die Identifizierung der genauen Ursache erfolgt über weitere Sensorik, wie beispielsweise Vibrationssensoren an der Maschine, oder die Erfahrung und Einschätzung der Produktionsmitarbeiter. Zudem erfolgt ein Abgleich mit historischen Werten. Somit lässt sich die Ursache des erhöhten Energiebedarfes feststellen, beispielsweise ein Verschleiß am Keilriemen des Antriebsstrangs.

Mithilfe der Online-FMEA wird die Auswirkung des derzeitigen Verschleißes (erhöhte Energiekosten, Abfall der Qualität), sowie die Eintrittswahrscheinlichkeit und die Auswirkung, wenn beispielsweise der Keilriemen reißt und die Maschine damit vorrübergehen ausfällt, berechnet. Ein Vergleich mit weiteren vorliegenden Störungen bestimmt die Priorität dieser Störung. Gleichzeitig kann eine Mitarbeitergruppe ausgewählt werden, der in Lage ist, den Reparaturauftrag an der Maschine vorzunehmen. Ein freier Mitarbeiter sieht auf seinem mobilen Endgerät priorisiert alle Störungen in seinem Anwendungsbereich, wie Abbildung 3 illustriert

Abbildung 4: Priorisierung der Störungen

6 Fazit

Big Data bietet im Störungsmanagement große Potentiale, um Störungen frühzeitig zu erkennen und diese zu beheben, bevor die Störung negative Auswirkung auf die Produktion hat. Intelligente IKT-Strukturen liefern vielfältige Daten über die Abläufe und Zustände auf der Fertigungsebene. Big Data schafft die Voraussetzung diese unterschiedlichen Daten zu analysieren und Zusammenhänge zu erkennen. Eine wichtige Komponente ist die Weiterentwicklung des Complex Event Processings, sodass die Muster selbstständig verbessert werden. Trotz der großen Potentiale und der technischen Fortschritte hinsichtlich der Datenanalyse steht die Einführung eines ganzheitlichen Störungsmanagements vor vielfältigen Herausforderungen. Grund sind vor allem fehlenden organisatorischen Strukturen und eine fehlende Standardisierung der Daten. Mit Big Data ist die Grundlage gelegt, ein ganzheitliches Störungsmanagement einzuführen. Eine konsequente Umsetzung ermöglicht den Unternehmen stabile Prozesse und eine verbesserte Liefertermintreue.

Literaturverzeichnis

[An11] Anicic, D. (2011). Event Processing and Strem Reasoning with ETALIS. Karlsruhe.

[At06] ATS. (2006). Stopped Production Costs $22,000 per minute: Studie.

[Bi13] Bitkom. (2013). Management von Big Data Projekte: Leitfaden. Berlin.

[Br13] Brücher, C. (2013). Rethink Big Data. Mitp Professional: Verlagsgruppe Hüthig Jehle Rehm.

[BD10] Bruns, R., & Dunkel, J. (2010). Event-Driven Architecture: Softwarearchitektur für ereignisgesteuerte Geschäftsprozesse. Berlin: Springer.

[CN04] Choi, Y., Narayanaswami, R., & Chandra, A. (2004). Tool wear monitoring in ramp cuts in end milling using the wavelet transform. International Journal of Advanced Manufacturing Technologies, 419–428.

[CG07] Czichos, H., Gerschwiler, K., Habig, K. H., Kleinlein, E., Klocke, F., Knoll, G., . . . Woydt, M. (2007). Tribologie-Handbuch: Reibung und Verschleiss (3., Aufl). Studium. Wiesbaden: Vieweg, F.

[GG13] Ganschar, O., Gerlach, S., Hämmerle, M., Krause, T., & Schlund, S. (2013). Produktionsarbeit der Zukunft: Industrie 4.0. Stuttgart.

[GJ04] Goebbels, S., & Jakob, R. (2004). Geschäftsprozess-FMEA: Fehlermöglichkeits- und Einfluss-Analyse für IT-gestützte Geschäftsprozesse (1. Aufl). Düsseldorf: Symposion.

[Ha13] Han, Y.-H. (2013). Ubiquitous information technologies and applications: CUTE 2012. Lecture notes in electrical engineering: Vol. 214. Dordrecht, New York: Springer.

[He94] Heil, M. (1994). Entstörung betrieblicher Abläufe. TU München, München.

[HS12] Holzwart, G., & Seufert, A. (2012). Datenexplosion ist nicht nur ein Technikthema. Computerwoche, (33-34), 6.

[KB03] Kamiske, G. F., & Brauer, J.-P. (2003). Qualitätsmanagement von A bis Z: Erläuterungen moderner Begriffe des Qualiätsmanagements (4., aktualisierte und erg. Neuaufl). München, Wien: Hanser.

[Kü13] Kümmerlen, R. (2013). Big Data: 10 Fakten zur Datenanalyse.

[LD14] Liggesmeyer, P., Dörr, J., & Heidrich, J. (2014). Big Data in Smart Ecosystems. Informatik Spektrum, 37(2), 105–111. doi:10.1007/s00287-014-0772-x

[Lu12] Luckham, D. C. (2012). Event processing for business: Organizing the real time strategy enterprise. Hoboken, N.J: Wiley.

[MK14] Mayer, G., Knüppel, K., Möhwald, H., & Nyhuis, P. (2014). Kompetenzorientiertes Störgrößenmanagement: Mitarbeiterkompetenzen sinnvoll in das Störungsmanagement integrieren. Productivity Management, 19(1), 15–18.

[MK13] Meyer, G., Knüppel, K., Busch, J., Jakob, M., & Nyhuis, P. (2013). Effizientes Störgrößenmanagement: Ansatz zur Kategorisierung von Störgrößen in der Produktion. Productivity Management, 18(5), 49–52.

[Po09] Popov, V. L. (2009). Kontaktmechanik und Reibung: Ein Lehr- und Anwendungsbuch von der Nanotribologie bis zur numerischen Simulation. Berlin: Springer.

[VJ12] Velten, C. & Janata, S. (2012). Datenexplosion in der Unternehmens-IT: Wie Big Data das Business und die IT verändert.

[Vo11] Vogel-Heuser, B. (2011). Erhöhte Verfügbarkeit und transparente Produktion ([Online-Ausg.]). Embedded systems: Vol. 2. Kassel: Kassel Univ. Press.

[We12] Werdich, M. (2012). FMEA - Einführung und Moderation: Durch systematische Entwicklung zur übersichtlichen Risikominimierung (inkl. Methoden im Umfeld). FMEA - Einführung und Moderation,

Organic Data: Ein sicheres, dezentralisiertes Big Data Konzept

Sebastian von Mammen, Carsten Grenz, Jörg Hähner
Organic Computing
Universität Augsburg
Eichleitnerstr. 30
86159 Augsburg

Sabine Timpf
Geoinformatik
Universität Augsburg
Alter Postweg 118
86159 Augsburg

{sebastian.von.mammen,carsten.grenz,joerg.haehner}
@informatik.uni-augsburg.de
sabine.timpf@geo.uni-augsburg.de

Daniel Loebenberger
Dept. for Computer Security
Bonn-Aachen Int. Center for IT
Dahlmannstr. 2
53113 Bonn

Stefan Mandl, Oleksandr Kozachuk
EXASOL AG
Neumeyerstr. 48
90411 Nürnberg

daniel@bit.uni-bonn.de
{stefan.mandl, oleksandr.kozachuk}@exasol.com

Abstract: In diesem Papier stellen wir ein Konzept vor für die sichere, dezentralisierte Sammlung und Verwertung großer Datenmengen. Die Kernidee ist eine selbstorganisierte hierarchische Organisation cyber-physikalischer "Organic Data"-Knoten (ODNs), die fähig sind, Daten lokal zu speichern, zu verarbeiten und gezielt unter Nachbarknoten zu kommunizieren. Das resultierende Netzwerk ist eine sich anpassende virtuelle Struktur, in der Daten aggregiert, verarbeitet und gespeichert werden. Der präsentierte Ansatz schließt Business-to-Consumer-Modelle gestützt auf potentiell personenbezogenen Datensätzen explizit mit ein. Neben der allgemeinen Darstellungen des Konzepts und weiterführender Erläuterungen hinsichtlich zugrundeliegender Technologien, werden konkrete Benchmarks für seine Evaluation präsentiert.

1 Einführung

Die massenhafte Verarbeitung und Speicherung von Daten erfordert ein besonderes Augenmerk auf den Schutz der Privatsphäre [BF10, KTGH13]. Bestehende Big Data Anwender wie z.B. Google und Facebook werden diesem Anspruch keinesfalls gerecht. Ihre Geschäftsmodelle und ihre DV-Infrastruktur zielen darauf ab, sämtliche Nutzerdaten zentral und weitestgehend ungefiltert zu horten[1], was beispielsweise die kürzlich offengelegten unautorisierten Zugriffe seitens der amerikanischen und britischen Sicherheits-

[1] www.datacenterknowledge.com

behörden in hohem Maß begünstigte [Hol13].

Wir begegnen dieser Herausforderung mit *Organic Data*, einem dezentralen, selbstorganisierenden Ansatz zu Big Data, der dem Missbrauch persönlicher Daten vorbeugt und gleichermaßen effizient und skalierbar ist. Anstatt die Daten verteilter Quellen zentral zu aggregieren, werden sie von lokalen Netzwerkknoten gezielt angefragt und verarbeitet, um dann gegebenenfalls als verallgemeinertes Faktum einer Wissensbasis hinzu gefügt oder als erkanntes Ereignis an andere Systemknoten weitergemeldet zu werden. Die Empfänger verfahren genauso, nur auf einem höheren Abstraktionsniveau, sodass Wissen über das beobachtete System umfassend und in hohem Maße abstrahiert gesammelt wird. Der grundlegenden Frage der Partitionierung der Datenbestände in einem verteilten Datenbanksystem wird durch die cyber-physikalische Netzwerkstruktur begegnet; sie ergibt sich, wie im nächsten Abschnitt genauer erläutert, automatisch durch das Zusammenspiel von (authorisierten) Benutzeranfragen, der initialen Gewinnung von Daten und ihrer weiterführenden Verarbeitung und Kommunikation.

Organic Computing Algorithmen [MSSU11] sollen die Topologie und die Kommunikation im Netzwerk "selbst"-organisieren, um sie einerseits den dynamischen Abstraktionen anzupassen und andererseits, um zu jedem Zeitpunkt Redundanz und Ressourcenverbrauch der Netzwerkknoten zu optimieren. Die Knoten des Organic Data Netzwerk sammeln, lernen, kanalisieren Daten und bedienen außerdem lokale, hochperformante Datenbanken.

Aufgrund der dezentralen Organisation von Organic Data muss im Anwendungskontext speziell auf Datensicherheit, Visualisierung und Benutzerfreundlichkeit eingegangen werden. Nutzung von Verschlüsselungstechniken, Authentisierungs- und Anonymisierungsverfahren, sowie die Verteilung auf mehrere Knoten und die möglicherweise vielschichtige Verarbeitung sollen unberechtigte Zugriffe auf die Daten verhindern. Die eingesetzte Sicherheitstechnik soll insbesondere Datenschutz auf hohem und quantifizierbarem Niveau gewährleisten: privacy by design. Ein Visual Analytics System soll es ermöglichen, Informationen der Netzwerkknoten einzusehen und höher- und tieferliegende Ebenen der Netzwerk- und Abstraktionshierarchie interaktiv zu steuern. Eine Herausforderung dabei liegt in der notwendigen dynamischen Anpassung des der Visualisierung zugrundeliegenden Datenmodells an den Abstraktionsgrad. Die Interaktion soll auf intelligente Art und Weise eine explorative Analyse der Daten unterstützen, ohne den Datenschutz zu verletzen.

Weiterhin ist dieser Beitrag wie folgt gegliedert. In Sektion 3 nehmen wir Bezug auf den Stand der Forschung aus verschiedenen, für Organic Data relevanten Blickwinkeln und erläutern das Wechselspiel der benannten Aspekte. In Sektion 4 werden Möglichkeiten dargelegt, um konkrete Implementierungen des Organic Data Ansatzes zu evaluieren. Wir schließen den Beitrag mit einer Zusammenfassung in Sektion 5.

2 Datenfluss im Organic Data Netzwerk

Anstatt wie bei herkömmlichen Big Data Ansätzen, Quelldaten zentral zu aggregieren, sollen sie durch ein Netzwerk von *Organic Data Nodes* (ODNs) geschleust werden. Organic

Data Nodes bieten eine Laufzeitumgebung für Softwareagenten, die die Kommunikation im Netzwerk koordinieren, empfangene Daten filtern, abstrahieren und lokal speichern. Dabei soll die Vertraulichkeit und Anonymität der Daten an jedem Punkt des Systems gewährleistet werden. Zugriffe auf die lokalen Daten verschiedener ODNs sollen aufgrund eingehender und ausgehender Datenströme erteilt werden. Wir wollen untersuchen, inwieweit diese Klassifizierung automatisch oder durch möglicherweise verteilte Autoritäten erfolgen muss.

2.1 Selbstorganisation nach Lernzielen

Die ODNs sollen sich mittels entsprechender Methoden des Organic Computing selbstorganisieren, sodass sich der Datenfluss aus der jeweiligen Problemstellung ergibt. Unser algorithmischer Entwurf sieht vor, dass eine bestimmte Anzahl von ODNs die Informationen von einer großen Anzahl an Quellen bezieht, verarbeitet und an einige wenige Senken weitergibt. Der relative Informationsgewinn entscheidet, ob der Datenfluss zwischen Knotenpaaren zu- oder abnimmt. Durch diesen Mechanismus kann der Anwender die Problemstellung definieren, indem er beschreibt, welche Zustände oder Prozesse ihn interessieren, also großen subjektiven Informationsgewinn versprechen (bspw. eine schnell wachsende Anzahl von Verkehrsteilnehmern pro Flächeneinheit). Der Anwender kann diese Bewertungen kontinuierlich verändern, um Abstraktion und Datenfluss im Organic Data Netzwerk zu verfeinern. Auch kann sich das Netzwerk durch dieses Prinzip ständig an neuartige Beobachtungen (Anomalien) oder Veränderungen des Netzwerks (bspw. durch Ausfall oder Hinzunahme von Quellen, Senken oder ODNs) anpassen.

Die resultierende Kommunikationstopologie des Organic Data Netzwerks korreliert mit den Lern- bzw. Abstraktionsprozessen. Entsprechend spiegelt sie sich auch in der lokal umgesetzten Datenpersistenz wider: Jeder ODN speichert, je nach Privacy-Modell und Anwendungsdomäne, einkommende Daten (auf erster Knotenebene u.a. Rohdaten) und/oder abstrahierte, ausgehende Daten. Ein in hohem Grad verteilter, hierarchisch organisierter Datenbestand und der Einsatz moderner kryptographischer Verfahren erschwert nicht nur den unerlaubten Zugriff Dritter, sondern ermöglicht auch die nahtlose Skalierung hochperformanter In-Memory Datenbanksysteme.

2.2 Hierarchischer Zugriff

Der effiziente und authentifizierte Zugriff auf den gesamten Datenbestand des Organic Data Netzwerks - von hochgradig abstrahierten Informationen bis hinab zu Rohdaten (falls vorhanden) - erfolgt durch das Zusammenspiel dreier Komponenten.

Datenbankorganisation Anfragen, die auf höherer Ebene nicht bedient werden können, werden über die unmittelbaren ODN-Quellen abgewickelt.

Authentisierungsmechanismen Zugriffe werden bezüglich der Datengranularität (bis zu

Rohdateneinsichtnahme) und der Vollständigkeit (bis zur Beschreibung des Gesamtsystems) separat authentifiziert. Dazu soll eine globale Schlüsselhierarchie eingeführt werden, die den Zugriff auf die einzelnen Datenblöcke koordiniert.

Visual Analytics Methoden Eine hierarchische Visualisierung der ODNs erlaubt es, (vertikal) in einzelne Knoten einzutauchen und dadurch die Granularität der abgebildeten Daten zu erhöhen. Auf gleichbleibender Abstraktionsebene ermöglicht es eine graphbasierte Darstellung, die Zusammenhänge der abstrahierten Daten sowie der ODNs (horizontal) abzubilden.

3 Stand der Forschung

Organic Data vereint Aspekte aus mehreren Forschungsbereichen, um ein integriertes, skalierbares System zur Verfügung zu Stellen. Organic Computing Ansätze dienen als algorithmische Grundlage der selbständigen Organisation von Netzwerktopologie und Datenbeständen. Ansätze aus dem Bereich verteiltes, hierarchisches Lernen werden gebraucht, um Daten in einem dezentralen System verarbeiten und verwalten zu können. Mit der dezentralen, hierarchischen Organisation gehen darüberhinaus besondere Anforderungen an den Schutz der Daten, an die Skalierbarkeit der Speicherung sowie die Visualisierung und Anwenderfreundlichkeit einher.

3.1 Selbstorganisierende Netzwerklösungen

Die Selbstorganisation von Computernetzwerken steht im Mittelpunkt verschiedener Organic Computing Lösungen [MSSU11], welche häufig Natur-inspirierten Algorithmen verwenden. Diese Lösungen sind skalierbar, robust und bieten eine hohe Verfügbarkeit des Systems, indem sie ein globales Management zu Gunsten verteilter in-network Managementalgorithmen aufgeben [Dre06].

Im Bereich von vernetzten ad-hoc Netzwerken spielt häufig die Topologiebildung zur Laufzeit eine wichtige Rolle. So wird z.B. das Finden kurzer Wege beim Routing mit Hilfe von virtuellen Ameisen gelöst [GSB02]. Auch die Optimierung von Ablageorten von Daten zur Laufzeit im Netzwerk kann mit Hilfe von Schwarmalgorithmen gelöst werden [MWT11]; so auch in peer-to-peer Systemen [FLMM10]. Des Weiteren können die eingesetzten Netzwerkprotokolle selbst zur Laufzeit dynamisch angepasst werden [TZHMS10, THH10].

Ein Anwendungsbeispiel für selbstorganisierende Netzwerke stellen aktive Smart Camera Systeme dar, die aus Kameras und Sensoren mit unterschiedlicher Ausprägung und Mobilitätseigenschaften bestehen können [GJH+12]. Planungsalgorithmen nutzen die Mobilität von Knoten zur Lösung von Aufgaben aus und haben somit direkten Einfluss auf die Netztopologie [WGH11].

3.2 Verteiltes und hierarchisches Lernen

Die ODNs werden Muster im Stil eines Feed-Forward Netzwerks einkommende Daten lernen und die resultierenden Lernhypothesen weiterleiten. Es konnte gezeigt werden, dass ein derart verteilter Lernansatz redundante, unsichere Quelldaten zusammenführen kann, um die Lernhypothese zu verbessern [CS93]. Die parallele Verarbeitung erlaubt außerdem bestimmte (globale) Lösungen schneller zu finden als mit zentralisierten Systemen [SWA08]. Falls das Problem nicht von beliebigen Blickwinkeln aus erlernt werden kann, bzw. falls keine invariante Partitionierung des Suchraums möglich ist, kann man zumindest sicherstellen, dass man nur jene Muster lernt, die über den gesamten Suchraum Gültigkeit besitzen [PH94]. Immens steigende Datenvolumen motivierten die Entwicklung verteilter Lernmethoden. Mittlerweile werden sie wegen ihrer Robustheit und Effizienz v.a. im Kontext drahtloser Sensornetzwerke diskutiert [PKP06].

Die Zusammenführung verteilten Wissens reicht von der Aggregation (gefilterter) Teilergebnisse, über ihre Akkumulation, bis hin zum Lernen auf den bisherigen Lernhypothesen (Meta-Learning oder hierarchisches Lernen) [vMSDJ11, vMSSJ12, vMS13]. Bei numerischen Werten bietet es sich an, Cluster zu lernen, bspw. mittels künstlicher neuronaler Netze [SvMJ10] oder genetischer Algorithmen [SvMJ11]. Regelhafte Repräsentationen können durch die Verallgemeinerung ihrer Konditionen und die Aggregation ihrer Aktionen zusammengeführt werden [SDv$^+$ed].

Jeder ODN wird mittels lokal akkumulierten oder generierten Wissens (Lernhypothesen) den Informationsgehalt einkommender Datenströme bewerten. Diese Bewertungen liegen der Selbstorganisation im Organic Data Netzwerk zugrunde. Außerdem können durch das Traversieren des Organic Data Netzwerks einerseits abstraktere, andererseits umfassendere Lernhypothesen oder Muster generiert werden, die dem Anwender schließlich präsentiert werden.

3.3 Security/Privacy

Die Kommunikation der ODNs soll mit modernen Verfahren Ende-zu-Ende verschlüsselt werden. Das dazu nötige Schlüsselmaterial wird mittels einer Public-Key-Infrastructure (PKI) bereitgestellt. Hierfür eignet sich eine moderne Blockchiffre wie der Advanced Encryption Standard (AES) [fip01]. Dieser wird in gängigen Sicherheitsprotokollen des Netzwerkverkehrs standardmäßig eingesetzt, bspw. IPsec [KS05] oder TLS [DR08, Eas11].

Einem Angreifer muss es unmöglich sein, einen eigenen Knoten in das Netzwerk einzubringen. Auf Netzwerkebene eignen sich hierfür Message-Authentication-Codes wie CBC-MACs oder CMACs [BR05, nis01]. Auch asymetrische Authentisierungverfahren, wie der Digital Signature Standard (DSS) [NIS09], können je nach Anforderung gezielt eingesetzt werden. Letztere erfordern eine dedizierte Schlüssel-Infrastruktur, die beispielweise auf X.509 Zertifikaten beruhen kann, siehe [CSF$^+$08].

Die einzelnen Datenquellen dürfen nicht rück-verfolgbar sein. Dies kann beispielsweise dadurch erfolgen, dass anstelle der Identität einer Datenquelle ein randomisierter Hashwert

übermittelt wird. Hierfür können Hash-Verfahren, wie der Secure Hashing Algorithm (SHA-3) [BDPA13], eingesetzt werden. Diese haben die Eigenschaft, dass aus der Berechnung des Hashwerts keine Rückschlüsse über die Eingabe gezogen werden können.

Die Integration der dargelegten Verfahren ist insbesondere wegen der dezentralen und inhärent dynamischen Organisationsstruktur des Organic Data Ansatzes von großem Interesse.

3.4 Skalierbarkeit von Big Data Lösungen

Obwohl aktuelle High-End Ethernet Karten bereits einen Durchsatz von 56 Gigabit/Sekunde bieten (in beide Senderichtungen) und sich somit der Durchsatz im Netzwerk in großen Schritten dem Durchsatz des Hauptspeichers annähert, skaliert Kommunikation in einem Netzwerk nicht beliebig. Physikalische Limitierungen, bspw. durch Switches, erschweren es, eine annähernd gleichbleibende Kommunikationsgeschwindigkeit bei steigender Knotenanzahl zu garantieren. Auch die Datenspeicherung unterliegt klaren Einschränkungen. EXASolution, eine massivparallele und spaltenbasierte, relationale in-memory Datenbank kann diese Limitierungen teilweise kompensieren und eine schnelle Analyse großer Datenmengen (100TB bei Clustern von ungefähr 100 Knoten) dennoch ermöglichen. Zudem können Strategien der Datenpartitionierung und -verteilung Zugriffe drastisch beschleunigen. Der Zugriff auf eine große Anzahl ausgelagerter Partitionierungen (Shards) kann durch ein entsprechendes Betriebssystem, bspw. EXAClusterOS, gewährleistet werden, das einen Rechner-Verbund zu einem logischen Knoten zusammenführt [KMN+13]. Doch selbst dann wächst die Wahrscheinlichkeit ungünstiger Verteilungen mit der Anzahl der Shards [CAA08, RS10]. Selbstorganisiertes Clustering sowie die Formation hierarchischer Strukturen von ODNs versprechen die Skalierbarkeit weiter drastisch zu erhöhen. Bei einem moderaten System von 5000 Clustern (derartige Infrastrukturen mit ca. einer halben Millionen Servern sind beispielsweise bei Google schon länger im Betrieb), erreicht man bereits den Exabyte Bereich.

3.5 Visual Analytics

Die Transformation von Rohdaten in eine aussagekräftige Visualisierung erfolgt auf Basis einer Modellbildung um wichtige von unwichtigen Rohdaten unterscheiden zu können [KAF+08]. Die Modellbildung ist Teil eines Datenabstraktionsprozesses [TVPE92, Tim99], dessen Ergebnis zum Wissenszuwachs in der explorativen Analyse [DMK05] beitragen soll. Sind die ODN Teil eines Geosensornetzwerks [Duc13], so kann die raum-zeitliche Komponente der Daten für weitere Analysen [AA06] sowie zur Qualitätskontrolle genutzt werden. Die interaktive Visualisierung der (Geo)Daten [AA99, WGK10] folgt dem Mantra der Informationssuche: Überblick – Wichtiges herausheben – Zoom und Filter – Details auf Anfrage[CMS99]. In den Schritten *Zoom und Filter* sowie *Details auf Anfrage* ist der semantische Zoom wichtig, der die Präsentation von Datendetails auf Anordnung

vom Nutzer unter Berücksichtigung der Anforderungen bzw. des Ziels der Datenanalyse umsetzt [Mod97]. Damit ändert sich mit der Zoomstufe nicht nur die Detailliertheit der Information sondern auch die Art bzw. der Typ der dargestellten Information.

3.6 Nutzerinteraktion

Nutzer interagieren mit visuellen Darstellungen auf sehr komplexe Weise – die Qualität der Benutzerschnittstelle hängt häufig davon ab, wie sehr deren Funktionalität mit den Analyseintentionen des Nutzers übereinstimmt bzw. ob die Schnittstelle einzelne Operationen (Toolbox) oder vollständige Analyseabläufe (Processing) unterstützt [BRN12, Tim03]. Untersuchungen zur Funktion von Operationen führten zu einer Unterscheidung in hochwertige (Filter, Relate, Aggregate) und niederwertige (Zoom, Pan) Operationen in der Informationsvisualisierung [Pla05]. MacEachren und Kraak unterscheiden drei verschieden Achsen: Interaktionsebene (intensiv bis oberflächlich), Zielgruppe (einzelne Person bis Öffentlichkeit), sowie Aufgabenstellung (Informationsaustausch bis Wissenskonstruktion) [MK01].

4 Benchmarks

Der integrative Charakter von Organic Data bedingt, dass eine Implementierung hinsichtlich verschiedener Kriterien evaluiert werden muss. Im Folgenden wird konkret auf Effizienz, Selbstorganisation, Datensicherheit und Zugänglichkeit eingegangen.

4.1 Benchmark Velocity

Je nach Anwendung müssen Big Data Lösungen unterschiedlichen Ansprüchen bzgl. der Datenverarbeitungsgeschwindigkeit genügen. Offenbar wird dieser Unterschied wenn man bspw. mittels Mobilitätsdaten einen idealen Standort für eine neue Supermarktniederlassung ermitteln oder, im Gegensatz dazu, bei steigender Unfallgefahr den Verkehrsfluss dirigieren möchte. Im Allgemeinen müssen im ODN die Senken bei beliebigen Quelldaten zumindest so schnell bedient werden können, wie für die jeweilige, anwendungstypische Entscheidungsfrequenz nötig. Wir müssen deshalb einerseits zeigen, dass die verwendeten Algorithmen prinzipiell echtzeitfähig sind. Außerdem müssen wir die Auswirkungen lokaler Verarbeitungsalgorithmen und Entscheidungen der ODNs auf das Laufzeitverhalten des Gesamtsystems analysieren. Da sich netzwerkspezifische Kenngrößen, wie z.B. Latenzzeiten, aus einer Vielzahl von Parametern ergeben, bspw. der Netzwerktopologie, dem Datenfluss, sowie lokaler Performanzspezifikationen, werden wir neben ausführlichen Tests anhand der Demonstratoren auf bewährte simulationsbasierte Analyseverfahren zurückgreifen.

4.2 Benchmark Self-X Properties

Es soll ein Design Space für Organic Data Systeme charakterisiert werden, in dem klassische Netzwerkeigenschaften, wie z.B. Bandbreite, Stabilität, den erforschten Netzeigenschaften von Organic Data Networks, z.B. Selbstkonfiguration, -heilung und -optimierung (self-x properties), gegenüber gestellt werden. Eine maßgebliche Frage wird die Abbildung von "Privacy-by-Design" Parametern, also der selbstorganisierte Schutz von Daten, in diesen Designspace haben. Die entwickelten Algorithmen sollen in diesen Designspace eingeordnet und bzgl. ihrer Performanz am Labordemonstrator evaluiert werden. Zukünftigen Nutzern des Systems wird dieser Designspace zusammen mit den entwickelten Algorithmen eine Richtlinie darstellen, anhand derer er sein System auf die konkreten Privacy-Anforderungen und Netzwerkausgestaltung anwenden kann.

4.3 Benchmark Security/Privacy

Um die Sicherheit des resultierenden Organic Data Netzes zu validieren, wird überprüft in wie weit die Aspekte Verschlüsselung, Authentisierung und Anonymität erreicht wurden. Die Verschlüsselung wird dadurch verifiziert, dass die resultierenden Datenströme ununterscheidbar von echtem Zufall sind. Dazu werden unter anderem standardisierte statistische Verfahren zur Überprüfung eingesetzt. Authentisierung wird validiert, indem gezeigt wird, dass ein empfangenes Datenpaket nur von dem erwarteten Sender der Nachricht stammen kann. Der Einsatz moderner Sicherheitsverfahren, wie TLS zur Sicherung der Kommunikation zwischen verschiedenen ODNs sollte dies im allgemeinen gewährleisten. Es ist jedoch im Kontext des Organic Data Netzes erforderlich dies gezielt zu verifizieren. Um Anonymität zu gewährleisten, müssen einzelne Datenquellen voneinander ununterscheidbar sein. Konkret muss gezeigt werden, dass jedes verarbeitete Rohdatenpaket ebenso von mindestens einer weiteren Quelle stammen kann. Damit lässt sich nicht rückverfolgen welche Quelle genau die Rohdaten zur Verfügung gestellt hat: privacy by design.

4.4 Benchmarks Visual Analytics & HCI

Um die Effektivität abstrakter Darstellungsmodi zu analysieren, möchten wir Nutzerbewertungen verschiedener Abstraktionsebenen bei gleichbleibender Analyseaufgabe sammeln. Die Dynamisierung der Visualisierung zur Abbildung sich verändernder Muster und der Adaption der Organic Data Netzwerkstruktur kann durch den Gebrauch von Metriken fr Veränderungsraten bemessen werden. Bei der Evaluation der Nutzerinteraktionen als auch deren Umgebungen kommen klassische Methoden wie Zeitmessung, Messung der Anzahl Klicks, Videoanalyse und Befragung zum Einsatz [KEM07]. Die Bewertungskriterien müssen erweitert werden, um die Effektivität horizontaler und vertikaler Darstellungs- und Explorationsmodi zu quantifizieren.

5 Zusammenfassung & Ausblick

In diesem Beitrag haben wir Organic Data als dezentrales Konzept für Big Data dargestellt. Wir haben den Datenfluss in einem Organic Data Netzwerk erläutert und die Spezifikation einzelner Knoten in diesem Netzwerk (Organic Data Nodes) beschrieben. Der notwendigen Integration eines neuen Ansatzes wie Organic Data in den Anwendungskontext haben wir durch Erläuterungen hinsichtlich der Datensicherheit sowie der Usability (Visual Analytics und HCI) Rechnung getragen. Um einerseits die Herausforderungen des Konzepts zu unterstreichen und um andererseits konkrete Implementierungen zu unterstützen, haben wir außerdem konkrete Vorschläge für Benchmark-Metriken unterbreitet. Unsere Ausführungen definieren die Eckpfeiler von Organic Data. Um eine tiefgreifende Integration der verschiedenen Aspekte systematisch umzusetzen, würden wir uns auf den Design Science Ansatz besinnen und schrittweise die Details der verschiedenen verwobenen Entwicklungszyklen herausarbeiten und miteinander in Einklang bringen - von der Anforderungsanalyse zum Feldversuch, vom existierenden Grund- und Expertenwissen zu dessen konkreter Erweiterung und schließlich vom konkreten Systementwurf bis zum marktreifen Prototypen [Hev07].

Literatur

[AA99] G.L. Andrienko und N.V. Andrienko. Interactive maps for visual data exploration. *International Journal for Geographical Information Science*, 13(4):355–374, 1999.

[AA06] N. Andrienko und G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data*. Springer Berlin Heidelberg, 2006.

[BDPA13] Guido Bertoni, Joan Daemen, Michaël Peeters und Gilles Van Assche. The Keccak sponge function family. online: `http://keccak.noekeon.org/`, January 2013.

[BF10] David Bollier und Charles M Firestone. *The promise and peril of big data*. Aspen Institute, Communications and Society Program Washington, DC, USA, 2010.

[BR05] John Black und Phillip Rogaway. CBC MACs for Arbitrary-Length Messages: The Three-Key Constructions. *JC*, 18(2):111–131, 2005.

[BRN12] D. Burkhardt, T. Ruppert und K. Nazemi. Towards process-oriented Information Visualization for supporting users. In *Interactive Collaborative Learning (ICL), 2012 15th International Conference on*, Seiten 1–8, 2012.

[CAA08] Jean-Daniel Cryans, Alain April und Alain Abran. *Criteria to Compare Cloud Computing with Current Database Technology*, Jgg. 5338, Seiten 114–126. Springer Berlin Heidelberg, 2008.

[CMS99] S.K. Card, Jock Mackinlay und B. Shneiderman. *Readings in Information Visualization. Using Vision to think*. Morgan Kaufmann Publishers, San Francisco, CA, 1999.

[CS93] Phillip K Chan und Salvatore J Stolfo. Toward parallel and distributed learning by meta-learning. In *AAAI workshop in Knowledge Discovery in Databases*, Seiten 227–240, 1993.

[CSF+08] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley und W. Polk. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, May 2008. RFC 5280.

[DMK05] Jason Dykes, A.M. MacEachren und M.-J. Kraak. *Exploring Geovisualization*. Elsevier Pergamon Press, Amsterdam, 2005.

[DR08] T. Dierks und E. Rescorla. The Transport Layer Security (TLS) Protocol, Version 1.2, August 2008. RFC 5246.

[Dre06] Falko Dressler. Self-Organization in Ad Hoc Networks: Overview and Classification. Bericht, University of Erlangen, Dept. of Computer Science 7, Erlangen, 2006.

[Duc13] Matt Duckham. When Computing Happens Somewhere. In *Decentralized Spatial Computing*, Seiten 3–32. Springer Berlin Heidelberg, 2013.

[Eas11] D. Eastlake, 3rd. Transport Layer Security (TLS) Extensions: Extension Definitions, January 2011. RFC 6066.

[fip01] *Federal Information Processing Standards Publication 197 - Announcing the ADVANCED ENCRYPTION STANDARD (AES)*, 26 November 2001. Publication 197.

[FLMM10] Agostino Forestiero, Emilio Leonardi, Carlo Mastroianni und Michela Meo. Self-Chord: A Bio-Inspired P2P Framework for Self-Organizing Distributed Systems. *IEEE/ACM Transactions on Networking*, 18(5):1651–1664, Oktober 2010.

[GJH+12] C. Grenz, U. Jänen, J. Hähner, C. Kuntzsch, M. Menze, D. d'Angelo, M. Bogen und E. Monari. CamInSens - Demonstration of a Distributed Smart Camera System for In-Situ Threat Detection. In *Distributed Smart Cameras (ICDSC), 2012 Sixth International Conference on*, Seiten 1–2, 2012.

[GSB02] M. Günes, U. Sorges und I. Bouazizi. ARA - The Ant-Colony Based Routing Algorithm for MANETs. In *Proceedings. International Conference on Parallel Processing Workshops*, Seiten 79–85. IEEE Comput. Soc, 2002.

[Hev07] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

[Hol13] Martin Holland. NSA-Überwachungsskandal: PRISM, Tempora und Co. - was bisher geschah. online: `http://heise.de/-1909702`‘, July 2013.

[KAF+08] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer und Guy Melançon. Visual Analytics: Definition, Process, and Challenges. In Andreas Kerren, JohnT. Stasko, Jean-Daniel Fekete und Chris North, Hrsg., *Information Visualization*, Jgg. 4950 of *Lecture Notes in Computer Science*, Seiten 154–175. Springer Berlin Heidelberg, 2008.

[KEM07] A. Kerren, A. Ebert und J. Meyer. *Human-centered Visualization Environments*. Springer Berlin Heidelberg, 2007.

[KMN+13] Alfons Kemper, Tobias Mühlbauer, Thomas Neumann, Angelika Reiser und Wolf Rödiger. Bericht vom Herbsttreffen der GI-Fachgruppe Datenbanksysteme. *Datenbank-Spektrum*, 13(1):65–66, 2013.

[KS05] S. Kent und K. Seo. Security Architecture for the Internet Protocol, December 2005. RFC 4301.

[KTGH13] Dominik Klein, Phuoc Tran-Gia und Matthias Hartmann. Big Data. *Informatik-Spektrum*, Seiten 1–5, 2013.

[MK01] Alan M. MacEachren und Menno-Jan Kraak. Research Challenges in Geovisualization. *Cartography and Geographic Information Science*, 28(1):3–12, 2001.

[Mod97] D. Modjeska. Navigation in Electronic Worlds: A Research Review, Technical Report. Bericht, Computer Systems Research Group, University of Toronto, 1997.

[MSSU11] Christian Müller-Schloer, Hartmut Schmeck und Theo Ungerer, Hrsg. *Organic Computing - A Paradigm Shift for Complex Systems*. Autonomic Systems. Birkhäuser Verlag, 2011.

[MWT11] Hannes Mühleisen, Tilman Walther und Robert Tolksdorf. Data Location Optimization for a Self-Organized Distributed Storage System. In *Proeedings of the Third World Congress on Nature and Biologically Inspired Computing*, Seiten 176–182. IEEE Press, 2011.

[nis01] NIST Special Publication 800-38B: Recommendation for Block Cipher Modes of Operation; The CMAC Mode for Authentication, 2001.

[NIS09] NIST. FIPS 186-3: Digital Signature Standard (DSS). Bericht, Information Technology Laboratory, National Institute of Standards and Technology, June 2009.

[PH94] FJ Provost und DN Hennessy. Distributed machine learning: scaling up with coarse-grained parallelism. In *Proceedings/... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology*, Jgg. 2, Seite 340, 1994.

[PKP06] Joel B Predd, SB Kulkarni und H Vincent Poor. Distributed learning in wireless sensor networks. *Signal Processing Magazine, IEEE*, 23(4):56–69, 2006.

[Pla05] C. Plaisant. Information Visualization and the Challege of Universal Usability. In J. Dykes, A.M. MacEachren und M.-J. Kraak, Hrsg., *Exploring Geovisualization*, Seiten 53–82. Elsevier Ltd., 2005.

[RS10] Kurt Rohloff und Richard E. Schantz. High-performance, massively scalable distributed systems using the MapReduce software framework: the SHARD triple-store. In *Programming Support Innovations for Emerging Distributed Applications*, PSI EtA '10, Seiten 4:1–4:5, New York, NY, USA, 2010. ACM.

[SDv$^+$ed] Abbas Sarraf Shirazi, Timothy Davison, Sebastian von Mammen, Jörg Denzinger und Christian Jacob. Adaptive Agent Abstractions to Speed Up Spatial Agent-Based Simulations. *Simulation Modelling Practice and Theory*, 2013 (submitted).

[SvMJ10] Abbas Sarraf Shirazi, Sebastian von Mammen und Christian Jacob. Adaptive Modularization of the MAPK Signaling Pathway Using the Multiagent Paradigm. In *Parallel Problem Solving from Nature – PPSN XI*, Jgg. 6239 of *Lecture Notes in Computer Science*, Seiten 401–410. Springer Verlag, Krakow, Poland, 2010.

[SvMJ11] Abbas Sarraf Shirazi, Sebastian von Mammen und Christian Jacob. Hierarchical Self-Organized Learning in Agent-Based Modeling of the MAPK Signaling Pathway. In *CEC 2011, IEEE Congress on Evolutionary Computation*, Seiten 2245–2251, New Orleans, Louisiana, 2011. IEEE Press.

[SWA08] Padhraic Smyth, Max Welling und Arthur U Asuncion. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems*, Seiten 81–88, 2008.

[THH10] Sven Tomforde, Björn Hurling und Jörg Hähner. Dynamic Control of Mobile Ad-hoc Networks-Network Protocol Parameter Adaptation using Organic Network Control. In *ICINCO (1)*, Seiten 28–35, 2010.

[Tim99] Sabine Timpf. Abstraction, Levels of Detail, and Hierarchies in Map Series. In Christian Freksa und DavidM. Mark, Hrsg., *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science*, Jgg. 1661 of *Lecture Notes in Computer Science*, Seiten 125–139. Springer Berlin Heidelberg, 1999.

[Tim03] S. Timpf. Geographic Activity Models. In Matt Duckham, Michael Goodchild und Michael F. Worboys, Hrsg., *Foundations of Geographic Information Science*, Seiten 241–254. CRC Press, 2003.

[TVPE92] Sabine Timpf, Gary S. Volta, David W. Pollock und Max J. Egenhofer. A conceptual model of wayfinding using multiple levels of abstraction. In A.U. Frank, I. Campari und U. Formentini, Hrsg., *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, Jgg. 639 of *Lecture Notes in Computer Science*, Seiten 348–367. Springer Berlin Heidelberg, 1992.

[TZHMS10] Sven Tomforde, Ioannis Zgeras, Jörg Hähner und Christian Müller-Schloer. Adaptive control of sensor networks. In *Autonomic and Trusted Computing*, Seiten 77–91. Springer, 2010.

[vMS13] Sebastian von Mammen und Jan-Philipp Steghöfer. *The Computer after Me*, Kapitel Bring it on, Complexity! Present and future of self-organising middle-out abstraction. World Scientific Publishing, submitted 2013.

[vMSDJ11] Sebastian von Mammen, Jan-Philipp Steghöfer, Jörg Denzinger und Christian Jacob. Self-organized Middle-Out Abstraction. In Christian Bettstetter und Carlos Gershenson, Hrsg., *Self-Organizing Systems*, Jgg. 6557 of *Lecture Notes in Computer Science*, Seiten 26–31, Karslruhe, Germany, 2011. Springer Verlag.

[vMSSJ12] Sebastian von Mammen, Abbas Sarraf Shirazi, Vladimir Sarpe und Christian Jacob. Optimization of Swarm-based Simulations. *ISRN Artificial Intelligence*, (Article ID 365791):12, 2012.

[WGH11] Michael Wittke, Carsten Grenz und Jörg Hähner. Towards Organic Active Vision Systems for Visual Surveillance. In Mladen Berekovic, William Fornaciari, Uwe Brinkschulte und Cristina Silvano, Hrsg., *ARCS*, Jgg. 6566 of *Lecture Notes in Computer Science*, Seiten 195–206. Springer, 2011.

[WGK10] M. Ward, G. Grinstein und D. Keim. *Interactive Data Visualization: Foundations, Techniques, and Application*. A K Peters, 2010.

Realizing the Predictive Enterprise through Intelligent Process Predictions based on Big Data Analytics: A Case Study and Architecture Proposal

Julian Krumeich[1], Jens Schimmelpfennig[2], Dirk Werth[1], Peter Loos[1]

[1]Insitute for Information Systems (IWi),
German Research Center for Artificial Intelligence (DFKI),
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
{Julian.Krumeich, Dirk.Werth, Peter.Loos}@dfki.de

[2]Software AG,
Altenkesseler Str. 17, 66115 Saarbrücken, Germany
Jens.Schimmelpfennig@softwareag.com

Abstract: Today's globalized economy forces companies more than ever to constantly adapt their business process executions to present business situations. Companies that are able to analyze the current state of their processes and moreover forecast its most optimal progress as well as proactively control them based on reliable predictions will be a decisive step ahead competitors. The paper at hands examines, based on a case study stemming from the steel manufacturing industry, which production-related data is currently collectable using state of the art sensor technologies forming a potential foundation for a detailed situation awareness and derivation of accurate forecasts. An analysis of this data however shows that its full potential cannot be utilized without dedicated approaches of big data analytics. By proposing an architecture for implementing predictive enterprise systems, the article intends to form a working and discussion basis for further research and implementation efforts in big data analytics.

1 Introduction

1.1 Vision of the Predictive Enterprise

In order to meet the steadily rising market requirements in today's globalized world, companies are more than ever forced to adapt their business processes to present process and business situations. The rising digitalization of the real world, driven by the era of Internet of Things, allows for unprecedented insights into current business process situations [Kr14a]. In the future, only those companies which are able to analyze their business operations based on this rapidly growing mass of data, predict the best proceeding process flow, and proactively control their processes with this knowledge will survive in increasingly competitive markets. Such a company sketches the vision of a "Predictive Enterprise" as the next stage in the evolution of real-time enterprises within the age of data as a decisive competitive asset [Kr14a, Lu06].

1.2 Motivation and problem definition

Today's corporate practice is beyond the realization of this vision. The potential of massive amounts of data in companies which can already be collected by state of the art sensor technologies, has only received scant attention in business process management (BPM) [Un12]. Frequently, this is due to the lack of technical capabilities to analyze big data in a timely manner and to derive the right decisions. However, business analysts agree that companies are going to face existential difficulties if they are not able to discover or even anticipate problems and defects in business process executions in time [PG11]. Thus, BPM software has to provide means to analyze process executions, forecast their further progress and consequently anticipate potential problems early enough to proactively enact appropriate countermeasures. Yet, current approaches and techniques in predictive analytics rarely take into account context situations, e.g. through the utilization of technologies like Complex Event Processing (CEP), as a basis to derive forecasts. Most predictions remain purely stochastic involving the risk of computing highly-likely forecasts of process progressions that turn out to be entirely unlikely in the present, perhaps exceptional, situation. This calls for a dedicated consideration of given situations respectively events to compute reliable forecasts. Even though available data sets will grow continuously due to enhanced sensing capabilities and provide unprecedented insights into process situations, this entails the need for dedicated big data analytics able to exploit the full potential out of such data.

1.3 Research contribution and applied research method

By means of a case study (see sec. 2 and for preliminary work [Kr14b, Kr14c]), the paper at hand analyzes which process and context data a sample company can currently collect by state of the art sensor technology forming a potential basis for a predictive planning and control of manufacturing processes. In doing so, the paper basically follows case study research, which has been applied in Information Systems research for almost two decades [DP03]. In particular, the paper employs the methodology proposed by Benbasat et al. [BGM87]: as the unit of analysis the steel bar production line at Saarstahl AG, a major German steel producer, was chosen. The two research questions applied to the case study are “What type of data is currently available in industry processes using state-of-the-art sensor technology to realize event-based predictions?” and “Why is it a ‘Big Data’ challenge to analyze this data appropriately?” The results of the selected case study are considered to be generalizable to other manufacturing enterprises. Since the chosen research design is a single-case study, it is particular appropriate to revelatory cases, which is typical for big data analytics cases as a relatively new phenomenon in information systems research. The case study data was collected by the central department of information and communication technology of the company. As data selection methods, interview techniques were applied and physical artifact—sensor networks—investigated.

By analyzing the collected data, the paper points out that current state of the art sensor technologies already provide data that has to be considered through the lens of big data (see sec. 2.3). To make sufficient business value out of it, especially in terms of process

predictions, dedicated big data analytics are required. For this, the paper derives seven requirements to be considered for a system implementation to realize process predictions based on big data analytics (see sec. 3.1). Moreover, the article proposes a general system architecture (see sec. 3.2) intending to form a working and discussion basis for further research and implementation efforts in big data analytics (see sec. 5).

2 Case study: Big data in the steel bar production at Saarstahl AG

2.1 Brief description and case study settings

This case study analyzes a production part within the steel bar production at Saarstahl AG, a major German steel producer, and provides types and sizes of sensor data that is currently collectable throughout the production processes. In the considered production area, half a million tons of steel are produced annually. In order to meet customers' specific quality requirements for various existing end products, Saarstahl AG conducts comprehensive quality checks within its production line. These include diameter controls with laser, surface testing by magnetic particle testing, checks for internal steel errors by ultrasound, and a variety of temperature and vibration measurements. All of these techniques provide continuous sensor data at the lowest system level (L1). In addition, other sensor systems in production (ambient and positioning sensors) are installed to monitor the control of steel bars via a material flow tracking system (L2-system level). Based on this basic data and the available customer orders, the production planning and control system calculates a rough schedule (L3 to L4 system level).

2.2 Current problem statement

Steelmaking processes are significantly influenced by internal and external events that lead to deviations in production quality. Some examples are fluctuating material properties of pig iron or production-induced deviations due to physical processes taking place. This entails a need for frequent adaptations of production instances and subsequently extensive re-planning of underlying production plans. For instance, if a standard further processing of steel provides the steps peeling, heat treatment and finally sawing of steel bars, then the crude steel cannot pass these steps if the sensor network used in the process detects a bending, as an example of a production deviation.

In this case, this intermediate either needs to be post-processed in a dressing and straightening machine to comply with quality requirements, or the processing of the intermediate can continue for another customer with lower quality criteria. Therefore the production planning for this process instance is obsolete and needs to be re-initiated in order to meet quality promises and deadlines to all customers, e.g., steel bars need to be factored into the normal process flow after conducting such an ad hoc processing step. Since the production planning systems compute an almost full capacity for the following days in batch mode, the simple insertion of this process instance into the running production flow may contradict to the planned execution of the production plan.

Therefore, in case a production deviation is detected during the production process, an intelligent recalculation of the entire production plan would be necessary in real-time to ensure an optimal solution from a global planning perspective. If such production influencing events were even anticipatable, corresponding countermeasures could be initiated before its occurrence.

Currently, the data provided by the sensor network, which is integrated into the production processes, exceeds the volume that is analyzable so far. The information and control systems used and the analysis techniques available on the market do not allow the producer to monitor and control the entire production process in real-time. Moreover, no future states and events, such as looming production deviations, can be predicted on time. Hence, so far the control of production processes is rather done in a reactive way; yet, the sketched vision of a predictive enterprise has not been realized.

2.3 Data characterization, challenge of Big Data and requirement derivation

In the following, sample data obtained from the applied sensor networks at Saarstahl AG are described according to the Big Data characteristics proposed by Gartner [BL12]. If the entire sensor networks within the production process is considered—which would be necessary for a comprehensive production planning and control on an L3-/L4-systems' level—, the Big Data challenge will be many times higher.

An example from the sensor network illustrates the immense volume of data in monitoring the production process. In the rolling mills 31 and 32 there are two optical surface test sensors that can continuously provide real-time data for the detection of surface defects during the rolling process. Basically, this allows to take into account the varying customer demands for a particular surface quality. The unit can already prototypically detect errors and differentiate the types of errors. This optical testing generates several hundred terabytes of data annually (volume). Currently, only a sporadic reactive analysis of these data is possible. Also, other context data from the sensor network and the systems settled on L-2 or L-3 level can currently not be linked due to the volume of data to be analyzed. While these systems could in principle detect production deviations in batch mode, this takes too long to allow timely reactions.

While this is just one example of very large resulting data of individual sensors in a particular section of the factory, another example illustrates the high data diversity (variety), which is continuously collected by different sensors and sensor networks at various points throughout the production process. This places high demands on an analysis by big data principles. For instance, the further processing of steel bars as of now already provides half a million of sensor data records, which reflect one production area to a particular context. In the next couple of months the sensor performance will be advanced, such that over 1.5 million sensor data will be available on L1 and L2 level. According to the principles of CEP; however, only the identification of relevant events in this torrent of both homogeneous and heterogeneous data as well as their correlation allows to derive patterns and deviations. This is possible only by using highly structured, technical knowledge. At this point, the basic claim to a scalable solution becomes clear,

since sensor networks should be flexibly expandable, but must also allow analyses and forecasts within a required time frame. The company plans to increase sensing in this subsection to an output of more than five million records, which underlines the need for scalability. Thus, in terms of the analysis of these large and diverse data, the responding time is crucial, since speed is a decisive competitive factor in the analysis (velocity / analytics). Classic reporting or batch processing would be significantly too slow, so that so-called high velocity technologies must be performed in near-real-time analyses. For the purposes of the outlined vision of predictive enterprises, it is also crucial to conduct accurate forecasts of the process sequences. Each day, an average of one terabyte of video data is recorded in a single subsection of the plant. However, a pure video analysis method is not sufficient for predictive analytics methodologies. In the existing system, it has been shown that only some production deviations could be detected by this classical approach. In addition, there is no feedback for process optimization. Therefore process data need to be included in the model formation and forecasting. Here, as outlined, over one million data sets are incurred in the coming months. For analyzing the dependencies among process and video data, data from a long period of time must be used for model training. In this case, the data volume may rapidly exceed 50 terabytes. For a real-time adaptive prediction, on average one-tenth of the data should be used. At present, however, such a number of data can hardly be processed in real-time. A direct compression of the data is impossible because of its variety to be considered.

As of now, due to its big data characteristics, the production process is away from the envisioned optimum in terms of planning and control. Technically, Saarstahl AG can integrate further sensor technology into its production processes to achieve a better monitoring. However to increase the demanded business value, current analytical methods take too long to complete an analysis.

3 Towards a system architecture proposal for realizing intelligent process predictions based on Big Data analytics

3.1 Requirement analysis

In past, datasets sensed from manufacturing processes were rather small covering only insufficiently operational context situations; hence, leading to imprecise forecasts. While in principle, it was possible to expand the data base, this information gathering has proven to be too complex, too expensive and not accomplishable in a timely manner. Conventionally, predictions had been computed using stochastic means; consequently, percentage values indicate which outcome or result a process will have. Such stochastic means however neither take into account nor reflect the current context situation in which a manufacturing process takes place resulting in inaccurate forecasts (cf. [Kr14b] for a comparison of stochastic and event-based predictions). Today, manufacturing companies still perform insufficient analyzes and forecasts based on their collected sensor data, even though it will positively influence their economic and ecological performance [Un12]. This is in contrast to industries such as insurance or banking that have fully implemented predictive analytics in their business models [MCD13]. Yet,

recent technological progress in the fields of Internet of Things and Cyber-Physical Systems enable to equip production processes relatively cost-neutral by sensors. This allows to measure internal and external production process parameters in a previously unprecedented level of detail. Thereby the technical foundation has been created to establish and continuously enrich a data base allowing for highly-accurate predictions.

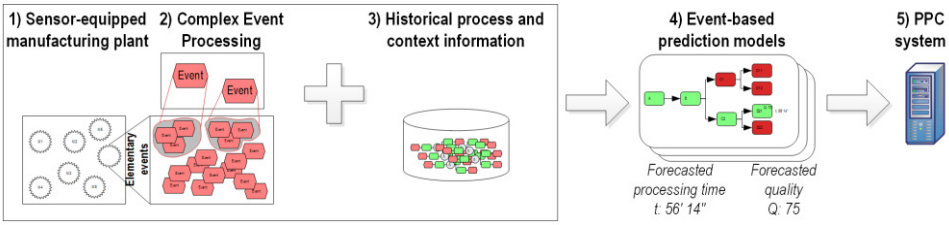


Figure 1: Basic concept for an event-based process forecast and control

As the use case analysis showed, the overall challenge of computing accurate predictions has shifted from being able to capture current context situations—as the baseline for deriving forecasts—to being able to manage and adequately consider this mass of data. In addition, the use case illustrated that the mass of data is progressively growing possessing the necessity of having a scalable platform for integrating sensor technology (cf. Fig. 1, 1). This leads to the first requirement of the system architecture.

Requirement 1. Provide scalable means for extending sensors throughout production processes and to store the mass of collectable data in descriptive process and context models.

In addition to collecting and storing process data in historical description models—as a knowledge base for process predictions (cf. Fig. 1, 3)—, it is compulsory to investigate current real-time conditions of processes through analyzing their current events and context situations. Due to the wide variety of heterogeneous sensors, which are used in enterprises, this is particularly a challenge of big data analytics in terms of data variety (cf. use case analysis in sec. 2.3). Eventually, these streams of atomic data have to be searched for patterns in real-time, e.g. by utilizing CEP technology (cf. Fig. 1, 2).

Requirement 2. Provide means for detecting and filtering of complex events within tremendous streams of atomic sensor data.

Such CEP allows to correlate current conditions with historical ones as the baseline for deriving event-based predictions. Through this correlation, event-based forecast models can be derived and repeatedly adapted (see Fig. 1, 4). As a technical infrastructure, a platform must be available, which combines a batch-oriented analysis with that of a distributed stream mining analysis. Whereas batch-oriented methods are important for training prediction models; stream-oriented ones are compulsory for the actual real-time analysis of incoming data streams. In particular, for the automatic analysis of image and video sensors—as outlined in the use case—dedicated algorithms needs to be available in order to derive structured information from unstructured data. To realize a real-time correlation of complex events with historical process data lead to the third requirements.

Requirement 3. Provide real-time data storage capabilities to correlate and analyze big data collections and streams in terms of high volume, high variety and high velocity.

Based on such correlations, prediction models comprising forecasts about the future progress of business processes can be computed, which allows a proactive reaction to predicted problems (cf. Fig. 1, 4). Several prediction techniques and algorithms can be applied, all of which have to be capable of processing big data and to cope with real-time requirements, which is another challenge in terms of big data analytics. In this regard, sophisticated CEP techniques are required predicting likelihoods of the occurrence of future atomic events that will eventually trigger complex ones. These probability assumptions will realize a predictive complex event detection.

Requirement 4. Derive and continuously adapt (event-based) prediction models.

Based on event-based prediction models forecasts of substantially higher accuracy can be computed, since predictions are not purely based on stochastic, but instead, the actual current state is decisive for the computation (see Fig. 1, 4). Hence, process progressions can be forecasted and certain deviations from planned and required process execution objectives can be proactively detected leading to corresponding system responses.

Requirement 5. Create alerts as responses to predicted deviations from planned process objectives based on calculated forecasts.

Within computed prediction models not only one single possible future process progress will be forecasted, but several multiple ones whose occurrence are depended both on non-influenceable events as well as on influenceable actions. Thus, recommendations or automatic decisions and actions should be provided to positively influence and control the outcome of business processes in accordance to specific process objectives. This will allow to realize a proactive incident mangement in contrast to a reactive incident handling as a prerequisite of a predictive enterprise.

Requirement 6. Derive recommendation and automatic decisions for mitigation actions.

As a result of intelligent algorithms it should further be determined whether changes enacted within one process instance—as a response to detected deviations, defects and problems—will impact other running instances that in turn will affect recommendations and automatic actions. These automatic or manually triggered actions based on real-time event-based predictions eventually realize an intelligent proactive process planning and control (PPC, cf. Fig. 1, 5) and thus the vision of a predictive enterprise.

Requirement 7. Enact sophisticated proactive process adaptations on the basis of computed recommendations and decisions.

3.2 Architecture proposal

This section outlines a functional description of components required for realizing a system for event-based process prediction based on big data analytics (cf. Fig. 2). On the

basis of the previously derived requirements, the architecture comprises three inherent layers that are typically considered within analytical processes [Ak13]: a descriptive (cf. requirements 2, 3), a predictive (cf. requirement 4) and a prescriptive layer (cf. requirements 5, 6). On top of them, an adaptation layer (cf. requirement 7) allows for intelligent actions incorporated into business process management engines as responses to prescriptive decisions. As the foundation, an integration layer (cf. requirement 1) realizes the system's physical interweaving into production process and facilities.

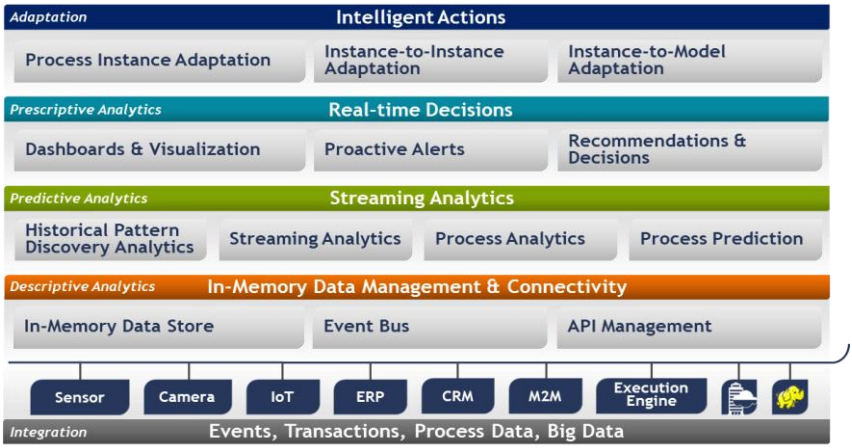


Figure 2: Functional building blocks for a predictive enterprise architecture

Integration layer L0: Event, transactions, process data, Big Data. The foundation of this system is a solid integration platform that connects the system to a company's existing IT infrastructure. In an industry 4.0 context additional adapters for sensor and IoT object integration are required. Due to the wide variety of heterogeneous sensors, which are used throughout enterprises, an initial classification into an ontology-based enterprise data model is required, which also forms the cornerstone to depict context situations in detail. Such a model should consist of at least three subtypes: an enterprise resources model, a business process context model and an enterprise external stimuli model. The use of these models realizes the semantic interoperability for the obtained context data and thus enables in principle an automated analysis.

Descriptive analysis layer L1: In-memory data management & connectivity. Due to the high volume of data and the velocity in which it is generated, an in-memory data management platform is utilized, allowing for distributed in-memory data management with extremely low, predictable latency and fast data access (microseconds) for real-time data handling. An in-memory data store will act as a central point of coordination, aggregating and distribution. Besides data management, events such as alerts or system messages are communicated using an event bus to which components can publish and subscribe. To manage the diverse data sources and connected enterprise systems, an API management component is introduced.

Predictive analysis layer L2: Streaming analytics. Real-time data accessible via the in-memory data management platform can be preprocessed. In particular, for the

intelligent evaluation of image and video sensors as outlined in the use case, special algorithms have to be developed in order to derive structured information from unstructured multimedia data. The results are fed back to the in-memory data store. The aggregated data is used for both ex-post and ex-ante analysis. From an ex-post viewpoint, historic data can be analyzed for pattern detection and correlated with respective business process behaviors. Based on these patterns, real-time event detection, such as deviations of production progress from an expected state, can be learned, optimized and applied to monitor real-time data streams. Data analysis components can communicate detected patterns to the event bus on which a complex event processing engine (CEP) is operated to correlate business process relevant events from data analysis results. These can then be used for process prediction.

Prescriptive analysis layer L3: Real-time decisions. In order to enable process owners to make qualified real-time decisions, all relevant data needs to be aggregated and visualized appropriately. Here, dashboarding functionalities similar to current business activity management solutions can be applied. In addition to that, process owners must be notified proactively if a decision is required or when a deviation in the current state of a process instance is detected. Besides pure visualization and notification, a recommendation is generated based on historic process analysis. To understand why a recommendation was made, drill-down functionalities allow to navigate to previous process instance information and enable process users to make qualified decisions.

Adaptation layer L4: Intelligent actions. Based on the data gathered and the resulting process prediction, business processes can either be adapted on an instance base (process instance adaptation) by adjusting the current process execution, or by optimizing the entire process type (instance-to-model). However, adaptations in a process instance can lead to necessary adaptations in other correlated process instances such as supporting or following processes (instance-to-instance). Here, the process owner is also supported. Once the adaptations were decided on, a governance process ensures a consistent transition of changes back into the process execution system(s).

4 Related work

4.1 Process data and context data acquisition and its real-time analysis

In order to detect context situations and progress of ongoing processes, data must be continuously collected. Technically, this can be done by means of physical and virtual sensors, which are often connected in a wireless network. Physical sensors can measure for instance pressure, temperature or vibration and recognize complex structures such as audio signals [SLT07]. Additional data is obtained from IT systems by evaluating for example log files or exchanged messages at the lowest system level [Ch10].

The next step is to analyze this mass of collected data. A common approach is that atomic events, which are detected by sensors, are being first cumulated to more complex ones, which is called CEP. CEP combines methods, techniques and tools to analyze mass of events in real-time and to obtain higher aggregated information from this most

elementary data. Such approaches can be found for example in [WDR06], who apply CEP on real-time data streams, and with particular respect to RFID-based data streams in [Wa06]. Often ontologies are used to identify semi-automatically the semantic context of events [TL11]. Operations on data of atomic and complex events require algorithms that are well-adapted to the process context as well as on big data characteristics.

A well-known approach for big data analytics is MapReduce, which decomposes a problem into independent sub-problems that are distributed and solved by so-called mappers. However, traditional forms of MapReduce follow a batch approach, which is inapplicable for data streams. In this regard, the field of research of stream mining and stream analytics has been formed recently [Br11].

4.2 Process forecast calculation and simulation

Predictive analytics refers to the forecast of prospectively occurring events. The necessary mathematical modelling can be done by using historical data. A characteristic example of a simple forecast calculation is the moving average. Such simple models only work if they are mostly independent of external influencing events, which is rarely the case. Especially in business processes several dependencies and influencing factors exist. Thus, modern statistical methods are required to recognize dependencies and patterns in large amounts of data, such as decision trees, univariate and multivariate statistics and data warehouse algorithms [LFC11].

Approaches combining predictive analytics with BPM are coined as Business Process Intelligence (BPI) (cf. [LFC11] for a discussion on existing definitions). Since it is feasibly to increase the context awareness in the era of Internet of Things, it is promising to use data collected by sensors to increase the accuracy of business process predictions. Reference [Fü12] present a first conceptual framework for combining CEP with predictive analytics. Also [JMM12] provide first conceptual works; however, they state that "[e]vent-based or event-driven business intelligence" approaches are only rudimentary researched and find limited integration in software tools. Even if [SMJ13] already combines BAM with BPI, the question on how findings from this connection can be used for adapting and optimizing business processes is still open.

4.3 Business process adaptation and optimization

In traditional BPM approaches, processes are typically adapted and improved on type level at design time after passing a business process lifecycle. This ex post handling—as it is regularly applied in business process controlling and mining—however causes considerable time delay. Since an aggregation of process execution data has to be done in a first place, before processes can eventually be optimized at type level, problematic process executions have already been completed and can no longer be optimized. Thus, scientists from the process mining community explicitly call for a stronger "operational support" [va12]. This requires the existence of immediate feedback channels to the business process execution. Without the possibility of such feedback loops, only ex-post

considerations are possible. At this level of run time adaptation and optimization of business processes, first approaches already exist [JMM12].

In the scope of commercial BPM systems, there are some early adopters characterized by the term Intelligent Business Process Management Suites [Si12]. For example, IBM's Business Process Manager provides means to analyze operations on instance level in real-time, to control them and to respond with further process steps in an ad-hoc manner [IBM13]; yet, a fully automation cannot be attested. On the market of real-time CEP, further vendors are present; e.g. Bosch Software Innovations have implemented CEP functionality into BPM solutions [Si12]. However, to consider them as intelligent BPM systems, a more powerful automated analysis of the variety of events must be implemented as a basis to realize event-based predictions and process control [Kr14a].

5 Summary and outlook

The paper at hands examined, based on a case study stemming from the steel manufacturing industry, which production-related data is currently collectable using state of the art sensor technologies forming a potential foundation for a detailed situation awareness and derivation of accurate process forecasts. An analysis of this data however showed that its full potential cannot be utilized without dedicated approaches of big data analytics. By proposing an architecture for implementing predictive enterprise systems, the article provides a working and discussion basis for further research and implementation efforts in big data analytics.

Acknowledgement

This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IS12050 (project IDENTIFY).

References

[Ak13] Akerkar, R: Advanced data analytics in business. In (Akerkar et al., Eds.): Big data computing. CRC Press, Boca Raton, FL, 2013; pp. 373-400.

[BGM87] Benbasat, I.; Goldstein, D. K.; Mead, M.: The case research strategy in studies of information systems. In: MIS Quarterly, vol. 11, no. 3, 1987; pp. 369-386.

[Br11] Brito et al.: Scalable and low-latency data processing with stream MapReduce. In: Proc. IEEE Third Int. Conf. Cloud Computing Technology and Science (CLOUDCOM '11). IEEE Computer Society, 2011; pp. 48–58.

[BL12] Beyer, M. A.; Laney, D.: The importance of "big data": A definition. In: https://www.gartner.com/doc/2057415, 2012; accessed 2014-07-10.

[Ch10] Choi, J.: RFID context-aware systems. In (Turcu et al., Eds.): Sustainable radio frequency identification solutions. InTech, Rijeka, 2010; pp. 307–330.

[DP03] Dubé, L.; Paré, G.: Rigor in information systems positivist case research: current practice, trends and recommendations. In: MIS Quarterly, vol. 27, no. 4, 2003; pp. 597-636.

[Fü12] Fülop et al.: Predictive complex event processing: A conceptual framework for combining complex event processing and predictive analytics. In: Proc. Fifth Balkan Conf. Informatics (BCI '12), ACM, 2012; pp. 26–31.

[IBM13] IBM (Eds.): IBM Business Process Manager Standard. In: http://www-03.ibm.com/software/products/en/business-process-manager-standard, 2013, accessed 2014-07-10.

[JMM12] Janiesch, C.; Matzner, M.; Müller, O.: Beyond process monitoring: A proof-of-concept of event-driven business activity management. In: Business Process Management Journal, vol. 18, no. 4, 2012; pp. 625–643.

[Kr14a] Krumeich, J.; Weis, B.; Werth, D.; Loos, P.: Event-driven business process management: Where are we now? – A comprehensive synthesis and analysis of literature. In: Business Process Management Journal, vol. 20, no. 4, 2014; pp. 615-633.

[Kr14b] Krumeich, J.; Jacobi, S.; Werth, D.; Loos, P.: Towards planning and control of business processes based on event-based predictions. In: Proc. 17th Int. Conf. Business Information Systems (BIS 2014), LNBIP vol. 176, Springer, 2014; pp. 38-49.

[Kr14c] Krumeich, J.; Jacobi, S.; Werth, D.; Loos, P.: Big Data Analytics for Predictive Manufacturing Control – A Case Study from Process Industry. In: Proc. 3rd IEEE Int. Congress on Big Data; in press.

[LFC11] Linden, M.; Felden, C.; Chamoni, P.: Dimensions of business process intelligence. In: Business Process Management Workshops (BPM 2010), LNBIP vol. 66, Springer, 2011; pp. 208–213.

[Lu06] Lundberg, A.: Leverage complex event processing to improve operational performance. In: Business Intelligence Journal, vol. 11, no. 1, 2006; pp. 55–65.

[MCD13] Minelli, M.; Chambers, M.; Dhiraj, A.: Big data, big analytics: Emerging business intelligence and analytic trends for today's business. John Wiley & Sons, Hoboken, NJ, 2013.

[PG11] Pettey, C.; Goasduff, L.: Gartner says between now and year-end 2014, overlooked but easily detectable business process defects will topple 10 global 2000 companies. In: http://www.gartner.com/newsroom/id/1530114, 2011, accessed 2014-07-10.

[SMJ13] Schwegmann, B.; Matzner, M.; Janiesch, C.: A method and tool for predictive event-driven process analytics. In: Proc. 11th Int. Conf. Wirtschaftsinformatik (WI2013), University of Leipzig; pp. 721–735.

[SLT07] Stäger, M.; Lukowicz, P.; Tröster, G.: Power and accuracy trade-offs in sound-based context recognition systems. In: Pervasive and Mobile Computing, vol. 3, no. 3, 2007; pp. 300–327.

[TL11] Taylor, K.; Leidinger, L.: Ontology-driven complex event processing in heterogeneous sensor networks. In: 8th Extended Semantic Web Conf. (ESWC 2011), LNCS vol. 6644, Springer, 2011; pp. 285–299.

[Un12] Unni, K.: Steel manufacturing could use more sensing and analysis. In: http://www.sensorsmag.com/process-industries/steel-manufacturing-could-use-more-sensing-and-analysis-10249, 2012, accessed 2014-07-10.

[va12] van der Aalst et al.: Process mining manifesto. In: Business Process Management Workshops (BPM 2011), LNBIP vol. 99, Springer, 2012; pp. 169–194.

[Wa06] Wang, F.; Liu, S.; Liu, P.; Bai, Y.: Bridging physical and virtual worlds: Complex event processing for RFID data streams. In: 10th Int. Conf. Extending Database Technology (EDBT 2006), LNCS vol. 3896, Springer, 2006; pp. 588–607.

[WDR06] Wu, E.; Diao, Y.; Rizvi, S.: High-performance complex event processing over streams. In: Proc. Int. Conf. Management of Data (SIGMOD '06), ACM, 2006; pp. 407–418.

Konzept und Methodik zur Entwicklung einer mobilen digitalen Persönlichkeit für nutzergerechte Anwendungen

David Schumm[1], Anna Julia Spanke[1], Kirsten Dreier[1], Victor Fäßler[1], Birger Becker[2], Hartmut Schmeck[2]

[1] TWT GmbH Science & Innovation
Ernsthaldenstraße 17
70565 Stuttgart
David.Schumm@twt-gmbh.de
Julia.Spanke@twt-gmbh.de
Kirsten.Dreier@twt-gmbh.de
Victor.Faessler@twt-gmbh.de

[2] FZI Forschungszentrum Informatik
Haid-und-Neu-Str. 10-14
76131 Karlsruhe
Birger.Becker@fzi.de
Hartmut.Schmeck@fzi.de

Abstract: Wir stellen die digitale Persönlichkeit vor, in der ein Nutzer individuelle Bedürfnisse und Eigenschaften hinterlegen kann. Entsprechend der Freigaben des Nutzers können diese Informationen verschiedenen Geräten und Diensten zur Verfügung gestellt werden, damit sich diese an den Nutzer anpassen können. Durch die Bereitstellung der digitalen Persönlichkeit als Dienst für Smartphones oder Wearables unterstützt sie einen Nutzer in der Mobilität und macht ihn zu einer Komponente im Gefüge der cyber-physischen Systeme. Wir schlagen eine interdisziplinäre Herangehensweise vor, die Ansätze zur Adaptivität von Systemen mit kommunikationswissenschaftlichen Ansätzen zur Nutzerakzeptanz, psychologischen Ansätzen zur Erfassung von Persönlichkeitsmerkmalen sowie Sensorik zur Ermittlung der Verfassung eines Nutzers verknüpft. Ein Beitrag besteht in der Vorstellung potentieller Anwendungsszenarien der digitalen Persönlichkeit im Bereich der Mobilität und im Gebäude-Energiemanagement. Ein weiterer Beitrag ist eine 2-stufige Methodik zur Entwicklung eines Persönlichkeitsmodells. Diese Methodik dient als Ausgangsbasis für die zukünftige Realisierung ausgewählter Szenarien.

1 Einleitung

Entwicklungen der vergangenen Jahre zeigen, dass sich technische Lösungen nur selten am individuellen Nutzer, seinen Bedürfnissen und an seiner individuellen Situation orientieren [Gr04] [Ja05]. Unser Ansatz einer mobilen digitalen Persönlichkeit bezweckt diese Lücke zu schließen. In der digitalen Persönlichkeit sind die individuellen Bedürfnisse und Eigenschaften eines Nutzers digital erfasst und werden Anwendungen sowie Geräten über Abfrageschnittstellen zugänglich gemacht. Sicherheitsmechanismen sollen dem Nutzer Kontrolle über die Informationen geben, die er preisgeben möchte. Die Verwaltung der digitalen Persönlichkeit obliegt damit ausschließlich dem Nutzer, der partielle Freigaben veranlassen und zurücknehmen kann und damit selbstbestimmt mit Anwendungen und Geräten interagiert. Eingerichtete Vertrauensbeziehungen zwischen

dem Nutzer und verschiedenen Anwendungen und Geräten, wie bspw. einem Navigationssystem, sollen die Datennutzung und die zweckmäßige Verwendung regeln. Die digitale Persönlichkeit soll dem Nutzer die Kontrolle über seine Präferenzen geben und ihm eine reaktive Möglichkeit zur Steuerung cyber-physischer Systeme bieten.

Dieser Beitrag ist wie folgt strukturiert: Im ersten Kapitel motivieren wir unseren Ansatz. Im zweiten Kapitel stellen wir den Stand der Technik im Bereich der Adaptivität von Anwendungen vor und leiten daraus Anforderungen und eine Zielstellung ab. Im dritten Kapitel zeigen wir Anwendungsgebiete für die mobile digitale Persönlichkeit auf. Das vierte Kapitel diskutiert die Methodik, nach der wir unseren Ansatz realisieren wollen. Kapitel fünf fasst den Beitrag zusammen und gibt einen Ausblick.

2 Stand der Technik

2.1 Forschungsansätze

Moderne Anwendungen und Technik bieten verschiedene Möglichkeiten zur Anpassung durch einen Benutzer, bspw. hinsichtlich der Lautstärke, Schriftgröße oder dem Helligkeitskontrast. In der Forschung gibt es verschiedene Ansätze, Benutzerschnittstellen in möglichst vielen Eigenschaften adaptiv zu gestalten. Das EU FP7-Projekt MyUI[1] hat z.B. eine Bibliothek[2] bewährter Anpassungsmöglichkeiten (Patterns) entwickelt und deren Nutzen wissenschaftlich fundiert. Überaus interessant werden diese Ergebnisse, wenn man sie mit Ansätzen zur Erfassung der Eigenschaften eines Nutzers in Verbindung bringt und den Nutzer stärker in die Anwendungen integriert. Das EU FP7-Projekt di.me[3] entwickelt eine Architektur und Implementierung, mit der Nutzer persönliche Informationen lokal oder in der Cloud verwalten und in Anwendungen wie Facebook integrieren können [TSH11]. In einem Verbund aus mehreren Projekten wurde im VUMS Cluster in diesem Kontext eine weitreichende Aufstellung vorgenommen, welche Eigenschaften eines Nutzers relevant sein könnten, um ihn digital zu erfassen [VU12].

Wir müssen zwischen digitaler Identität [Wi05] [Ca04], der Persönlichkeit, als auch der gegenwärtigen Verfassung eines Nutzers unterscheiden. Die digitale Identität eines Nutzers, also „wer genau“ eine Anwendung bedient, ist für eine nutzergerechte Adaptivität nicht unbedingt ausreichend. Entscheidend sind die wesentlichen Bedürfnisse, die den Nutzer kennzeichnen, denn diese unterscheiden sich individuell stark. Die Persönlichkeit eines Nutzers beschreibt Dimensionen individueller Unterschiede in Bezug auf kognitive und emotionale Zustände sowie Verhaltensmuster [Te08]. Handlungen und Reaktionen scheinen zumindest teilweise von der zugrundeliegenden Persönlichkeit gesteuert zu werden. Persönlichkeitsmerkmale sollten demzufolge auch einen Einfluss darauf haben, wie sich bspw. eine Person im Straßenverkehr normalerweise verhält. Die Verfassung einer Person bezeichnet einen vorübergehenden Zustand oder eine Stimmung, die ebenfalls Auswirkungen auf das situative Verhalten einer Person hat.

[1] MyUI Project: http://www.myui.eu
[2] MyUI Pattern Browser: http://myuipatterns.clevercherry.com
[3] Integrated Digital Me Userware: http://www.dime-project.eu

Mehrere Forschungsprojekte haben sich mit digitalen Identitäten befasst. Das Network of Excellence FIDIS[4] untersucht neben gesellschaftlichen Auswirkungen auch die Sicherheit von digital repräsentierten Nutzern. Das EU FP6 Projekt PRIME[5] und dessen FP7 Nachfolgeprojekt PrimeLife[6] erforschen Identitätsmanagementsysteme und unterstützende Technologien, um Datenschutz und Datenkontrolle zu gewährleisten. Diese Projekte betrachten dabei die Spur an Daten, die ein Nutzer über längere Dauer an verschiedenen Stellen hinterlässt. Das EU FP7 Projekt PICOS[7] untersucht ebenfalls das digitale Identitätsmanagement, legt aber den Schwerpunkt auf Sicherheit und Datenschutz innerhalb von digital organisierten Gemeinschaften.

Nutzermodellierung stellt einen Ansatz dar, für eine bestimmte Anwendung relevante Eigenschaften eines Nutzers digital abzubilden, um die Nutzung einer Anwendung oder eines Geräts zu erleichtern. Kobsa [Ko01] stellt in diesem Kontext eine Übersicht verschiedener Herangehensweisen an die Nutzermodellierung vor. Kobsa hat vorausgesagt, dass mobile Nutzermodelle und deren Einsatz zur Personalisierung von Anwendungen und Geräten in Zukunft eine Rolle spielen werden. Allerdings sind die Präferenzen und Modelle heutzutage noch immer über zahlreiche Anwendungen und Geräte verstreut, wie z.B. in Web Browsern, in sozialen Netzwerken, in Navigationssystemen, in einzelnen Apps und Desktop-Programmen. Heckmann [He06] liefert die Austauschsprache UserML und ein weitreichendes Inventar an Nutzereigenschaften, das neben Kontaktinformationen und demographischen Informationen auch psychologische Aspekte zur Erfassung der Persönlichkeit und Verfassung eines Nutzers enthält und damit eine solide Grundlage für verschiedene Anwendungsszenarien bildet. Allerdings besteht nach unserem Wissen bisher keine standardisierte Schnittstelle zur Abfrage von Nutzereigenschaften, die Zugriffsmethoden, ein Austauschformat und ein Eigenschaftsinventar definiert. Armac et al. [Ar09] greifen den Ansatz der Nutzermodellierung in Smart Environments (Zuhause, Auto, Arbeitsplatz, Hotel) auf und erläutern ähnliche Szenarien zur Anpassung der Umgebung an die Bedürfnisse des Nutzers unter Berücksichtigung von Datenschutz, beziehen allerdings die Persönlichkeit und die Verfassung des Nutzers nicht mit ein. Die Nutzermodellierung kommt unserem Verständnis einer digitalen Persönlichkeit sehr nahe, ist jedoch überwiegend von ingenieurwissenschaftlichen Prinzipien getrieben. Unsere interdisziplinäre Herangehensweise besteht darin, Psychologen, Kommunikationswissenschaftler, Informatiker und Ingenieure zusammenbringen, um die bestehenden Konzepte auf spezifische Anwendungsszenarien anzupassen, soweit erforderlich zu erweitern und in alltagstauglicher Form in die Praxis zu bringen.

Eine Patentrecherche hat angrenzende Erfindungen ergeben: Ourega [Ou10] beschreibt, wie die kontrollierte Weitergabe von Teilen eines Persönlichkeitsprofils in einem sozialen Netzwerk realisiert werden kann. Garijo Mazario et al. [Ga08] thematisieren eine Lösung zum Austausch, Zugriff und zur Verarbeitung digitaler Informationen einer Person, wie z.B. E-Mails, Aufgaben, Kontakte, Geschmack und Vorlieben. Diese Art der Informationsverarbeitung wird auch als Personal Information Management (PIM) bezeichnet. EssentialPIM [As13] als Beispiel einer PIM-Lösung bietet sowohl lokale

[4] Future of Identity in the Information Society: http://www.fidis.net/
[5] Privacy and Identity Management for Europe: https://www.prime-project.eu/
[6] PrimeLife: http://primelife.ercim.eu/
[7] Privacy and Identity Management for Community Services: http://www.picos-project.eu/

(Desktop) als auch mobile Anwendungen (Apps) an, mit denen persönliche Daten entweder zentral in der Cloud des Anbieters verwaltet oder lokal auf einem (mobilen) Gerät gespeichert werden.

Was den Bereich der Mobilität anbelangt, die einen Schwerpunkt unserer Arbeit bildet, sind sowohl wissenschaftliche Erkenntnisse als auch Produktinnovationen im Bereich Fahrerassistenz entwickelt worden. In dem BMBF-Projekt Invent[8] wurde unter anderem das Fahrerverhalten und die Mensch-Maschine-Kommunikation im Kontext aktiver Sicherheit und Fahrerassistenz betrachtet. Ein aktueller Trend in der Automobilindustrie besteht in einer engeren Integration des Nutzers (genauer, des sog. „Digital Native“) in das Auto mittels des Smartphones [Me13]. Dadurch ist es bspw. möglich, personalisierte Radiosender zu spielen, welche den Musikgeschmack des Fahrers einbeziehen.

2.2 Anforderungsanalyse

Viele kommerzielle Anbieter von Diensten wie z.B. sozialer Netzwerke oder auch Online-Partnervermittlungen gewinnen detaillierte digitale Persönlichkeitsprofile durch Techniken wie Data Mining und die Anwendung statistischer Methoden auf große Datenmengen (Big Data). Allerdings stehen diese Nutzerprofile nicht dem eigentlichen, informationsbereitstellenden Nutzer zur freien Verfügung und die Profile sind darüber hinaus nur bedingt gegen Zugriffe durch Dritte wie Werbepartner dieser Firmen geschützt oder werden sogar gezielt durch diese genutzt. Der Zugriff auf diese sensible Informationssubstanz wird also nicht vom Nutzer geregelt, sondern vom Anbieter.

Es bestehen nach unserem Wissen aktuell keine Standards oder offene Schnittstellen für eine integrierte und komfortabel nutzbare, aber vom Nutzer kontrollierte Verwaltung der digitalen Persönlichkeit. Wie bereits dargestellt, gibt es EU-weite Bestrebungen, um Lösungsansätze dafür zu entwickeln. Jedoch gibt es noch immer einen Mangel an pragmatischen Techniken und überzeugenden Lösungen für Modelle zur Erfassung der digitalen Persönlichkeit eines Nutzers, deren sicherer Bereitstellung als zugriffsgeschützter Dienst zur Unterstützung der informationellen Selbstbestimmung sowie deren Integration in cyber-physische Systeme zur Anpassung an einen bestimmten Nutzer. Obgleich die Komplexität der Anwendungen und der Einstellungsfreiraum zunehmen, fehlen dem Nutzer einfach zu bedienende Möglichkeiten, seine Bedürfnisse an eine Anwendung korrekt und schnell zu formulieren.

Es fehlen Anwendungen, die auf einen Menschen als gewachsene und sich verändernde Persönlichkeit eingehen [AZ09] [JLZ09] [HZR10]. Navigationssysteme bspw. haben sich als essentielle Hilfestellung im hektischen Verkehr entwickelt. Sie differenzieren aber nicht, welchen Typ von Fahrer sie unterstützen sollen [JZ11]. Auch in der Haustechnik steigen die Möglichkeiten der digitalen Unterstützung und Steuerung, der konkrete Nutzer wird allerdings nur unzureichend mit einbezogen [Zi12]. So ermöglichen Hausautomationssysteme eine komfortable Steuerung der elektrischen Geräte in einem Haus, jedoch richten sich „intelligente Abläufe“ noch nicht automatisch an der Verfassung der Menschen aus, die sich zu einem bestimmten Zeitpunkt in dem Haus aufhalten.

[8] Intelligenter Verkehr und nutzergerechte Technik: http://www.invent-online.de/de/projekte.html

3 Szenarien nutzergerechter Anwendungen

Im Folgenden stellen wir Szenarien nutzergerechter Anwendungen mit dem Schwerpunkt auf Mobilität und Gebäude-Energiemanagement vor, in denen der Nutzer nicht als statische Kenngröße betrachtet wird, sondern seine gegenwärtige Verfassung und Persönlichkeit einbezogen werden. Für alle Szenarien gilt die Annahme, dass der Nutzer entscheidet, welche Daten er den Systemen zur Verfügung stellen möchte. Je nach Szenario kommen unterschiedliche Anwendungen, Systeme und Informationsbausteine aus der digitalen Persönlichkeit zum Tragen. Den Szenarien liegt das in Abbildung 1 illustrierte Prinzip der digitalen Persönlichkeit zugrunde.

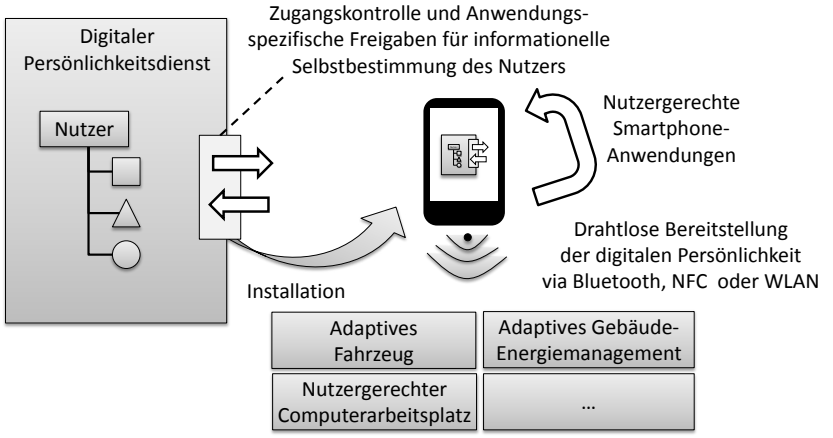


Abbildung 1: Prinzip und Anwendungsbereiche der mobilen digitalen Persönlichkeit.

3.1 Anpassung der Komfort-Systeme in Fahrzeugen

Die mobile digitale Persönlichkeit erfasst mehrere Arten von Nutzerdaten:

- physische Eigenschaften des Nutzers, z.B. Größe und Gewicht,
- allgemeine und ggf. tagesformabhängige Präferenzen des Nutzers im Hinblick auf Temperatur und Beleuchtung,
- permanente Einschränkungen des Nutzers, z.B. Allergien,
- temporäre Einschränkungen des Nutzers, z.B. Schnupfen oder Kopfschmerz,
- Kontextinformationen, z.B. Uhrzeit, Helligkeit, gefahrene bzw. zu fahrende Wegstrecke, anstehende Termine und Tagesplanung, Positionsbestimmung etc.
- kognitive Fähigkeiten, z.B. Hör- und Sehfähigkeit, Hand-Auge-Koordination und Kontextwechselfähigkeit,
- emotionale Zustände bzw. States, z.B. Wut, Stress, Traurigkeit, Fröhlichkeit,
- grundlegende Persönlichkeitseigenschaften, z.B. Neurotizismus, sowie deren Subskalen, z.B. Ängstlichkeit, Impulsivität.

Diese Informationen dienen zur Anpassung von Komfort-Systemen im Fahrzeug. Denkbar ist bspw. die Regelung der Belüftung ohne externe Luftzufuhr bei Pollenflug oder in

einem Tunnel. Des Weiteren kann die Sitzposition im Fahrzeug an die physischen Bedürfnisse und Einschränkungen des Fahrers angepasst werden und so etwa bei Müdigkeit auf dem Weg nach Hause eine andere Sitzposition eingestellt werden als auf dem Weg zur Arbeit.

3.2 Anpassung des Infotainments in Fahrzeugen

Die mobile digitale Persönlichkeit erfasst mehrere Arten von Nutzerdaten:

- Präferenzen des Nutzers im Hinblick auf die Art der aufbereiteten Informationen, z.B. ob zusätzliche Informationen zur Route gewünscht werden oder ob nur minimale Informationen zur Route angezeigt werden sollen,
- Präferenzen in der Navigation, z.B. Vermeidung von hohen Brücken, langen Tunneln oder das Vermeiden des Links-Abbiegens, Vermeidung von Staus etc.,
- Präferenzen des Nutzers im Hinblick auf den Umfang von Nachrichten, z.B. ob der Nutzer umfangreich oder sehr knapp informiert werden möchte,
- Präferenzen des Nutzers zum Inhalt der Nachrichten (Wirtschaft, Sport etc.),
- präferierte Musikrichtungen und Lieblingsbands des Nutzers mit Bezug auf Lieblingssender und die Sortierung von Sendern,
- präferierte Programme im Hinblick auf die Dauer des Programms (u.a. Wechsel von Nachrichten und Musik),
- Unterhaltungswünsche je nach Tageszeit, gefahrener Strecke, Terminen etc.,
- Präferenzen bei Unterhaltungsangeboten im autonomen Fahren.

Die Daten dienen zur Bereitstellung eines individuellen Infotainment-Programms und tragen zur Unterhaltung des Nutzers im Fahrzeug bei.

3.3 Anpassung an Fahrgemeinschaften

Fortgeschrittene Systeme können ggf. die digitalen Persönlichkeiten aller Mitfahrer einbeziehen. Wir schlagen daher den Einsatz eines „Kollektivs digitaler Persönlichkeiten“ zur Fahrzeuganpassung unter Berücksichtigung mehrerer Mitfahrer vor. Denn: Sofern nur eine Anpassung an den Fahrer stattfindet, könnte ein unerwünschtes Verhalten des Fahrzeugs auftreten – z.B. hört der Fahrer leidenschaftlich gerne Heavy Metal, befindet sich aber auf einer Dienstreise mit Kollegen, die diese Leidenschaft nicht teilen. Dieses Szenario der Fahrgemeinschaften ist nicht auf Pkw per se beschränkt, sondern lässt sich prinzipiell auch in Bussen oder Zügen dazu einsetzen, eine bestimmte Gruppendynamik zu erkennen (z.B. Unzufriedenheit in einem bestimmten Waggon) und ggf. Gegenmaßnahmen (z.B. Musik, Temperaturregelung, Belüftung) zu ergreifen.

3.4 Anpassung der Mensch-Maschine-Schnittstellen (HMI) in Fahrzeugen

Die mobile digitale Persönlichkeit erfasst mehrere Arten von Nutzerdaten:

- Physiologische Einschränkungen des Nutzers wie z.B. schlechtes Sehvermögen. Auf Basis der eingegeben Daten zeigt bspw. eine Bedienoberfläche große Sym-

bole oder eine größere Schrift. Bei einer Rotgrünblindheit kann auf problematische Farben in der Bedienoberfläche verzichtet werden. Auch kognitive Einschränkungen wie z.B. Sprachverständnis oder Kontextwechselfähigkeit könnten auf diese Weise berücksichtigt werden.
- Kulturelle Eigenschaften eines Nutzers: Zum Beispiel ist die Farbe „Rot“ nicht universell eine Signalfarbe.
- In Bezug auf die Sprachausgaben ist je nach Nutzer ggf. ein männlicher oder weiblicher Sprecher mit langsamer oder schneller Sprechweise bevorzugt.
- Vorhandene technische Kenntnisse des Nutzers können berücksichtigt werden.
- Auch Persönlichkeitseigenschaften und Verhaltensdispositionen, wie z.B. Gewissenhaftigkeit bzw. ein gewisser Sinn für Ordnung könnten die Anpassung der Schnittstelle mitgestalten.

Die eingegeben Daten dienen zur Anpassung der Mensch-Maschine-Schnittstelle in Fahrzeugen.

3.5 Vorbereitung einer Fahrt bzw. der Ankunft am Ziel

Dieses Szenario sieht die Nutzung der digitalen Persönlichkeit zur Fahrtvorbereitung und zur Vorbereitung der Ankunft vor. Die Fahrtvorbereitung – etwa wenn ein Fahrer sich seinem Fahrzeug nähert – umfasst u.a. das Einstellen der präferierten Fahrzeugtemperatur über die Standheizung, die Übernahme eines Fahrtziels sowie das Abfragen von Stauinformationen und ggf. das Vorschlagen von Alternativrouten.

Die Vorbereitung der Ankunft umfasst u.a.:

- Die Übermittlung von Daten an ein Hotel, welche Präferenzen im Hinblick auf den Hotelaufenthalt bzw. an ein Zimmer bestehen (u.a. Fernsehsenderliste/-sortierung, Temperatur, Klimaanlagenfunktion).
- Das Informieren des Fahrers, welches Terminal und Gate bei der Ankunft an einem Flughafen aufzusuchen ist. Weitere Anwendungsmöglichkeiten beziehen sich auf die Ankunft an Bahnhöfen oder in Parkhäusern und können diverse Transportmittel umfassen (Taxi, Zug, Bus, ÖPNV, Car-Sharing etc.).
- Die Fahrtstrecke kann ebenfalls hinsichtlich der persönlichen Präferenzen eines Fahrers gestaltet werden. Etwa könnte für Fahrer mit hohen Ausprägungen im Bereich „Sensation Seeking“, die eine hohe Anfälligkeit für Langeweile haben, eine besonders abwechslungsreiche Strecke gewählt werden, während Fahrer mit einem hohen Wert in Neurotizismus bzw. dessen Subskala Ängstlichkeit eher ruhigere Routen wählen könnten.

3.6 Gebäude-Energiemanagement

Hinsichtlich der Akzeptanz von Gebäude-Energiemanagementsystemen ist die individuelle Anpassung an die jeweiligen Präferenzen und Bedürfnisse der Nutzer eine essentielle Voraussetzung. Durch die digitale Persönlichkeit werden wichtige Informationen über die Bedürfnisse des Nutzers zur Geräte- und Anlagenkoordination im Gebäude bereitgestellt, die automatisiert genutzt werden können.

Die individuellen Anpassungen des Gebäude-Energiemanagements umfassen u.a.:

- Kopplung von Fahrzeugdaten und Gebäude-Energiemanagement: Das Gebäude-Energiemanagement passt z.B. die Koordination thermischer Energieflüsse individuell an die Präferenzen der eintreffenden Personen an.
- Adaptive Benutzerinteraktion: Die Schnittstellen zwischen dem Gebäude-Energiemanagement und dem Nutzer werden an dessen Präferenzen in Bezug auf die Interaktion mit dem Gebäude-Energiemanagement angepasst [BKS12].
- Anpassung der Optimierungsstrategien im Gebäude-Energiemanagement: z.B. wirtschaftlicher Einsatz vs. Maximierung des Anteils erneuerbarer Energien.
- Konfiguration des Zusammenspiels dezentraler Erzeugung, Speicher und flexibler Verbraucher hinsichtlich der Netzautarkie des Gebäudes.

4 Methodik zur Entwicklung der digitalen Persönlichkeit

Im Fokus unseres Ansatzes steht die Perspektive des Nutzers. Das angestrebte Ergebnis der Entwicklung und Realisierung sind Anwendungen und Geräte, die sich unter Zuhilfenahme der mobilen digitalen Persönlichkeit an die Bedürfnisse und an die Verfassung des Nutzers anpassen. Zur Entwicklung der digitalen Persönlichkeit schlagen wir eine zweistufige Methodik vor, die mehrere Disziplinen einbezieht: Die erste Stufe befasst sich mit der Erhebung von Anforderungen. In dieser Stufe sind kommunikationswissenschaftliche und psycho-diagnostische Ansätze maßgeblich. Die zweite Stufe befasst sich mit der technischen Realisierung in Anwendungen. In der zweiten Stufe dominieren ingenieurwissenschaftliche und informationstechnische Ansätze.

Arten der Datenerhebung: Grundlegend für die vorgeschlagene Methodik ist die Unterscheidung der geeigneten, wie auch erforderlichen Arten der Erhebung von Nutzerdaten. Nach unserem Ansatz wird die digitale Persönlichkeit des Nutzers durch zwei Arten der Datenerhebung erfasst und aktualisiert, vgl. dazu auch Abbildung 2.

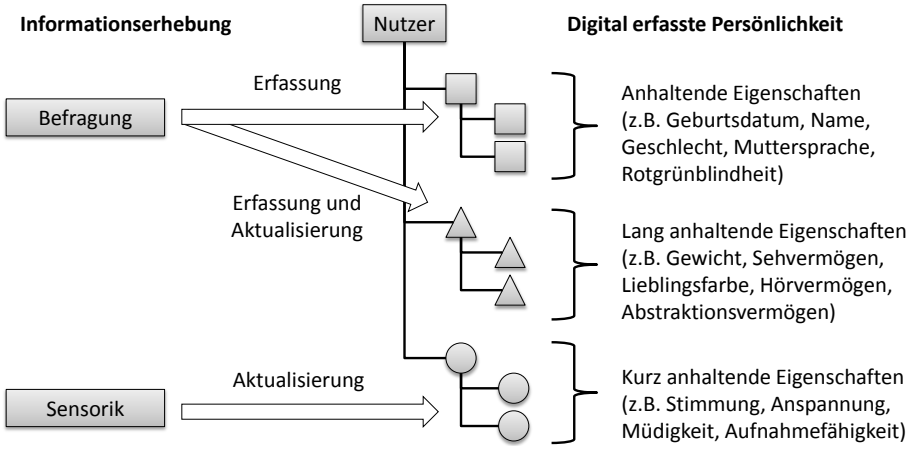


Abbildung 2: Arten der Datenerhebung mit ihrem Bezug zur digitalen Persönlichkeit

(1) Zum einen hinterlegt der Nutzer längerfristige individuelle Eigenschaften und Persönlichkeitsmerkmale, wie z.B. Alter, Sehschwäche, Essgewohnheiten, technische Fertigkeiten, Ängstlichkeit, Sensation Seeking, den kulturellen Hintergrund sowie seine Präferenzen zum Einsatz regenerativer Energiequellen und zum Autarkiegrad des Gebäudes. Aus der Kombination erhobener Daten wird die digitale Persönlichkeit technisch generiert bzw. nach einer Anpassung aktualisiert.

(2) Zum anderen reagiert die mobile digitale Persönlichkeit über Sensoren auf die Verfassung des Nutzers. Beispielsweise könnten durch einen Pulsmesser Hinweise auf den aktuellen Erregungsgrad eines Nutzers erfasst werden. Auch könnte die Spracheingabe eines Nutzers beispielsweise hinsichtlich Stressindikatoren analysiert werden.

4.1 Stufe 1 – Erhebung der Nutzeranforderungen

Die Erfassung der Nutzeranforderungen bzw. des individuellen Profils könnte vorab durch eine interviewgestützte Software geschehen, die mit einer Datenbank hinterlegt ist. Hier können relevante Eigenschaften und Anforderungen eingegeben werden, die im Folgenden skizziert werden. Im Sinne eines adaptiven Systems könnten dann über die Dauer der Nutzung hinweg weitere Daten gesammelt werden, sodass eine immer bessere Vorhersage der Nutzerbedarfe und der situativen Anforderungen eines Nutzers entsteht.

Wir schlagen dazu eine Kombination aus verschiedenen quantitativen und qualitativen Methoden zur Nutzerintegration vor:

- Standardisierte Erfassung des individuellen Persönlichkeitsprofils: Mittels verschiedener Persönlichkeitsfragebogen können unterschiedliche, relevante Persönlichkeitseigenschaften erfasst werden. So könnte Sensation Seeking mittels einer Sensation Seeking Scale [Ho02] erfasst werden. Darüber hinaus bieten Tests zur Erfassung der fünf grundlegenden Persönlichkeitsdimensionen Neurotizismus, Extraversion, Offenheit, Gewissenhaftigkeit und Verträglichkeit sowie deren Subfacetten wie der NEO-PI-R [CM92] eine gute Basis zur Erhebung allgemeiner Verhaltensdispositionen.
- Durchführung von physiologischen und kognitiven Leistungstests: Zur Erfassung der physiologischen und kognitiven Fähigkeiten eines jeweiligen Nutzers könnten Tests zur Erfassung der individuellen Konzentrations- und Reaktionsfähigkeit sowie der kognitiven Leistungsfähigkeit (z.B. Erfassung komplexer Situationen) genutzt werden.
- Interviews mit potentiellen Nutzern: Die mobile digitale Persönlichkeit basiert zudem auf Qualifikationsprofilen. Diese werden zunächst durch leitfadengestützte Interviews mit potentiellen Nutzern erhoben. Mit zunehmender Akzeptanz des Ansatzes liegen eine Standardisierung des Fragenkatalogs und eine stärkere Softwareunterstützung der Interviews nahe. In den Interviews könnten Faktoren erhoben werden wie u.a. individuelle Bedürfnisse eines Nutzers und Angaben zu seinen technischen Kompetenzen. Die Ergebnisse der Erhebung sind eine Basis für das Informationsmodell der digitalen Persönlichkeit.

- Erhebung von Nutzeranforderungen: Die Erhebung von Anforderungen verschiedener Nutzergruppen erfolgt im Hinblick auf (1) die mobile digitale Persönlichkeit und auf (2) Anwendungen und Geräte. Neben Interviews könnten Nutzeranforderungen im Hinblick auf Anwendungen (z.B. Apps) zusätzlich in Usability-Tests erhoben werden: Bestehende Anwendungen (wie Navigation, Gebäude-Energiemanagement) könnten von Nutzern verschiedener Zielgruppen in Usability-Tests überprüft werden. Diese Tests könnten audiovisuell aufgezeichnet, die verbalen Äußerungen der Test-Teilnehmer transkribiert und ausgewertet werden. Die Usability-Tests würden Einblicke in Schwächen und Potentiale bestehender Apps aus der Nutzerperspektive erlauben. Die Ergebnisse würden anschließend evaluiert und in die nutzergerechte Konzeption von Anwendungen einfließen. Die Daten stünden dann zudem für zukünftige Profile bzw. Anwendungen zur Verfügung.
- Evaluation der entwickelten technischen Lösungen durch Nutzer: Anwendungsprobleme und Optimierungsvorschläge bezogen auf die mobile digitale Persönlichkeit und darauf aufbauende Anwendungen werden in Usability-Tests mit Nutzern ermittelt. Testpersonen der Zielgruppen werden aufgefordert, bestimmte Bedienaufgaben auszuführen, zu kommentieren und Verbesserungsideen zu nennen (Methode des lauten Denkens). Die Ergebnisse werden ausgewertet. Auf Basis der Usability-Tests werden die Anwendungen modifiziert und verbessert. Die Einrichtung einer dauerhaften Feedback-Möglichkeit, etwa über ein Web-Formular, ist ebenfalls denkbar.

4.2 Stufe 2 – Technische Realisierung

Zur technischen Realisierung der mobilen digitalen Persönlichkeit sehen wir verschiedene Aktivitäten vor:

- Definition eines digitalen Persönlichkeitsmodells: Ziel ist es, Eigenschaften eines digitalen Persönlichkeitsmodells zu definieren und die Bedürfnisse des Nutzers zu verstehen und vorherzusagen. Zulässige Eigenschaften und Funktionen (z.B. „Lieblingsfarbe“ festlegen, „Lieblingsfarbe“ abfragen, Eigenschaftsblock „Farbempfinden“ abfragen, Eigenschaftsblock „Farbempfinden“ zurücksetzen) werden im sog. digitalen Persönlichkeitsmodell spezifiziert.
- Implementierung eines mobilen Diensts: Implementierung eines mobilen Diensts, der den kontrollierten Zugriff auf die digitale Persönlichkeit ermöglicht. Die zu entwickelnde Sicherheitsarchitektur muss die Sensibilität der erfassten Daten berücksichtigen. Eine Bereitstellung der digitalen Persönlichkeit in der Cloud ist aus diesem Gesichtspunkt kritisch zu betrachten. Die Bereitstellung des Diensts in einem Smartphone soll die kontaktlose aber lokale Verbindung zu nutzeradaptiven Anwendungen und Technik über Kommunikationswege wie Bluetooth, NFC oder WLAN ermöglichen. Im Anschluss an die Bereitstellung erfolgt die Befüllung des Diensts mit Nutzerinformationen. Anwendungen, die auf dem Smartphone ausgeführt werden (z.B. die Benutzerschnittstelle zum Gebäude-Energiemanagement oder eine Navigationsanwendung), sollen unmittelbar auf diesen Dienst zugreifen können. Umfassendere

Funktionalitäten wie „Zugangskontrolle einrichten“ oder „Zugriff gewähren“ werden ebenfalls definiert. Auf diese Weise wird den Anwendungen der Zugriff genehmigt, bestimmte, ggf. zu einem Paket gebündelte Umfänge der digitalen Persönlichkeit zu lesen, anzulegen (z.B. durch Abfragen und Interviews) oder auch zu ändern (mittels Fragebogenbearbeitung, Sensoren oder Klickverhalten).
- Anpassung von Applikationen an die mobile digitale Persönlichkeit: Umsetzung erhobener Anforderungen in Applikationen für Smartphones und Wearables, wie z.B. die Anpassung einer Energiemanagement- und Navigationsanwendung an die mobile digitale Persönlichkeit.

5 Zusammenfassung und Ausblick

Im Fokus des vorgestellten Konzepts der mobilen digitalen Persönlichkeit stehen die Berücksichtigung und Einbeziehung unterschiedlicher Nutzer und ihrer Bedürfnisse. Wir haben vorgeschlagen, die Persönlichkeitsmerkmale eines Nutzers durch eine psychometrische Testbatterie zu erfassen und digital bereitzustellen, damit sich Anwendungen und Geräte an den Nutzer anpassen können. Darüber hinaus bedarf es einer Menge an Diensten im Bereich der Sensorik und Sensordatenfusion, die die kurz anhaltenden Persönlichkeitsparameter, Bedürfnisse und Wünsche des Nutzers erfassen bzw. diese berechnen, antizipieren und bereitstellen. Wir betrachten die informationelle Selbstbestimmung des Nutzers dabei als fundamentale Anforderung. Wir haben Szenarien im Bereich der Mobilität und des Gebäude-Energiemanagements diskutiert, in denen die digitale Persönlichkeit eingesetzt werden könnte. Wir haben schließlich eine 2-stufige Methodik dargestellt, um die Entwicklung eines digitalen Persönlichkeitsmodells zu unterstützen.

Das Ziel des vorgestellten Ansatzes besteht darin, eine einfach zu bedienende und transparente Lösung zu schaffen, welche die Technik dazu befähigt, sich selbst an den Nutzer anzupassen. Zur Erreichung des Ziels sehen wir verschiedene Aufgaben:

- Analyse und Dokumentation von Anforderungen, Wünschen, Vorwissen und Ängsten, die verschiedene Altersgruppen (20+, 30+, 40+ ,50+, 55+, 60+, 65+, 70+) in Bezug auf Technik und deren Anwendungen haben. Dieses Wissen trägt wesentlich dazu bei, die Anforderungen an Akzeptanz und Zugänglichkeit an die Technik und deren Anwendungen zu formulieren sowie auf adaptive Bedienkonzepte unter Einsatz dieser Nutzereigenschaften zu übertragen.
- Die Spezifikation eines Informationsmodells, mit dem individuelle Eigenschaften einer Person, welche für eine nutzergerechte Anpassung von Anwendungen und Technik wesentlich sind, erfasst und digital bereitgestellt werden können.
- Die konkrete Ausgestaltung relevanter Eigenschaften von Personen verschiedener Altersgruppen mit Schwerpunkt auf ausgewählte Anwendungsfälle.
- Die Standardisierung der Abfrageschnittstelle, um das Konzept zu etablieren.
- Die Spezifikation eines Bedien- und Sicherheitskonzepts, das informationelle Selbstbestimmung ermöglicht, Zugriffsberechtigungen handhabbar macht und das Risiko des Datenmissbrauchs entschärft.
- Die prototypische Umsetzung durch ein Datenformat, die Implementierung als Dienst sowie die Evaluation der entstehenden Lösung.

Literaturverzeichnis

[Ar09] Armac, I. et al.: Privacy-friendly Smart Environments. In: Proceedings of the 3rd International Conference on Next Generation Mobile Applications, Services and Technologies, IEEE, 2009.

[As13] Astonsoft: EssentialPIM, 2013.

[AZ09] Arning, K.; Ziefle, M.: Different Perspectives on Technology Acceptance: The Role of Technology Type and Age. In: A. Holzinger and K. Miesenberger (Hrsg.): Human – Computer Interaction for eInclusion. LNCS 5889, 20–41, Springer, 2009.

[BKS12] Becker, B.; Kellerer, A.; Schmeck, H.: User Interaction Interface for Energy Management in Smart Homes. In: Proceedings of the 3rd IEEE PES Conference on Innovative Smart Grid Technologies (ISGT), IEEE, Washington, 2012.

[Ca04] Camp, L. J.: Digital Identity. In: IEEE Technology and Society Magazine, 23(3):34–41, 2004.

[CM92] Costa, P. T. (Jr.); McCrae, R. R. (Hrsg.): Revised NEO Personality Inventory (NEO-PI-R) and NEO Five Factor Inventory (NEO-FFI). Professional Manual. Odessa, Psychological Assessment Resources, 1992.

[Ga08] Garijo Mazario, F. J. et al.: Method, Device and Program for Personal Information Management. Patentnummer US 2008/0263014 A1, 2008.

[Gr04] Greisle, A.: Informations- und Kommunikationstechnologien für flexible Arbeitskonzepte. E-Work Nutzerstudie. Fraunhofer IAO, Stuttgart, 2004.

[He06] Heckmann, D.: Ubiquitous User Modeling. Akademische Verlagsgesellschaft Aka GmbH, Berlin, 2006.

[Ho02] Hoyle, R. H. et al.: Reliability and Validity of a Brief Measure of Sensation Seeking. In: Personality and Individual Differences, 32(3):401–414, 2002.

[HZR10] Holzinger, A.; Ziefle, M.; Röcker, C.: Human-Computer-Interaction & Usability Engineering for Elderly (HCI4Aging) - A Research Agenda. In: K. Miesenberger et al. (Hrsg.): 12th ICCHP - International Conference on Computers Helping People with Special Needs. Part II, LNCS 6180, 556–559, Springer, 2010.

[Ja05] Jakobs, E.-M.: Der mündige Nutzer. Technik für und mit Menschen. In: Computer in der Alltagswelt – Chancen für Deutschland? Acatech Tagungsband, 60–65, 2005.

[JLZ09] Jakobs, E.-M.; Lehnen, K.; Ziefle, M.: Alter und Technik. Eine Studie zur altersbezogenen Wahrnehmung und Gestaltung von Technik. Edition Wissenschaft. Apprimus, 2008.

[JZ11] Jakobs, E.-M.; Ziefle, M.: Mobilität für Ältere: Fahrerassistenzsysteme für ältere Fahrer. In: G. Rudinger & K. Kocherscheid (Hrsg.): Ältere Verkehrsteilnehmer - Gefährdet oder gefährlich? Defizite, Kompensationsmechanismen und Präventionsmöglichkeiten, 181–205, Göttingen, V & R unipress, 2011.

[Ko01] Kobsa, A.: Generic User Modeling Systems. In: User Modeling and User-Adapted Interaction, 11(1-2):49–63, 2001.

[Me13] Mercedes-Benz: Mercedes-Benz Apps, 2013.

[Ou10] Ourega, J. D.: Methods and Systems for Building, Managing and Sharing a Digital Identifty of a User over a Social Network. Patentnummer US 2010/0250618 A1, 2010.

[Te08] Terracciano, A. et al.: Personality Predictors of Longevity: Activity, Emotional Stability, and Conscientiousness. In: Psychosomatic Medicine, 70(6):621–627, 2008.

[TSH11] Thiel, S., Schuller, A.; Hermann, F.: Navigating the Personal Information Sphere. In: Informatik 2011, GI, 2011.

[VU12] VUMS Cluster: Virtual User Modelling. White Paper, 2012.

[Wi05] Windley, P.: Digital Identity. O'Reilly Media, 2005.

[Zi12] Ziefle, M.: Ungewissheit und Unsicherheit bei der Einführung neuer Technologien. Exploring Uncertainty, Springer Gabler, 2013.

TRUMPF

Der wichtigste Rohstoff bei der Herstellung von Hightech-Produkten: Herzblut.

Als weltweit führendes Technologieunternehmen mit Schwerpunkten in der Blechbearbeitung, Lasertechnik und Elektronik glauben wir daran, dass man Gutes immer noch besser machen kann. Dafür brauchen wir zunehmend mehr Softwareentwicklungs-Kompetenz und Informatik-Know-how, denn softwarebasierte Produkte bringen das Unternehmen ebenso voran wie IT-gestützte Prozesse. Innovationen sind für uns essentiell, nicht nur, wenn es um unsere Produkte und Prozesse geht, sondern auch im Hinblick auf Unternehmenskultur, Mitarbeiterförderung und gesellschaftliches Engagement. Für ein Umfeld, in dem neben Innovationen vor allem eines wachsen kann: Begeisterung.

www.trumpf.com

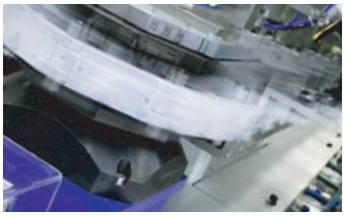
SIEMENS

Kristina Pähr,
Industry Automation, Amberg, Deutschland

Ihr Job ist es, Befehle zu erteilen.

Indem sie Anlagen programmiert.

Kristina Pähr hat bei Siemens den Job gefunden, der zu ihr und ihrer Qualifikation passt. So kann die Software-Entwicklerin beim Planen und Herstellen von automatisierten Anlagen ihre kreative Begabung und Leidenschaft für Technik einbringen. Besonders schön ist es für sie, wenn die Maschinen zum ersten Mal laufen. Entdecken auch Sie Siemens für Ihre Karriere und bewerben Sie sich jetzt online. Gemeinsam machen wir den Unterschied.



Automatisierung von Maschinen und Anlagen.

Siemens Karriere

siemens.de/karriere

Roboterkontrollarchitekturen

(ROBOCONTROL)

Roboterkontrollarchitekturen (ROBOCONTROL)[1]

Abstract: In den letzten Jahren sind vor allem in der Service-Robotik und in der Assistenzrobotik sehr komplexe Systeme entstanden. Bedingt durch bessere Sensorik, neue Methoden der Perzeption, leistungsfähige Verfahren der Lokalisation und der Kartenerstellung sowie kognitiver Systemkomponenten konnten unter anderem diese Fortschritte erzielt werden. Ein weiterer Aspekt, der diesen positiven Trend unterstützt, sind umfangreiche Open Source Robotik-Bibliotheken.So können beispielsweise durch das von Willow Garage entwickelte Robot Operating System – ROS Basissysteme schneller aufgebaut werden. Zum anderen hat es zur besseren Wiederverwendbarkeit und Integrierbarkeit von umfangreichen Sammlungen an Robotik-Methoden in der Community beigetragen. Dennoch stellt die wachsende Komplexität von Robotersystemen immer noch eine große Herausforderung dar. Angefangen von einer geeigneten Embedded Architektur, über die Unterstützung der Steuerungssoftware-Entwicklung durch leistungsfähige Frameworks bis hin zur Steuerungsarchitektur muss ein geeignetes Systemdesign gefunden werden, um einen effizienten Entwicklungsprozess zu ermöglichen und nichtfunktionale Eigenschaften wie Echtzeitfähigkeit, Zuverlässigkeit, Skalierbarkeit oder Sicherheit zu gewährleisten.Der Workshop greift diese Thematik auf. Wissenschaftler, die auf diesen Forschungsfeldern arbeiten, sollen im Rahmen des Workshops ihre neusten wissenschaftlichen Ergebnisse präsentieren. Darüber hinaus sollen wissenschaftliche Fragestellungen herausgearbeitet werden, die bisher nur unzureichend oder nicht gelöst wurden.

Thema und Schwerpunkte

Im Rahmen des Workshops sollen Beiträge insbesondere zu den folgenden Themen präsentiert werden:

- Architekturen eingebetteter Systeme
- Software Frameworks
- Steuerungsarchitekturen (taskbasierte oder verhaltensbasierte Ansätze)

Adressierter Teilnehmerkreis

Wissenschaftler, die im Bereich der Robotik forschen und Robotersysteme entwickeln. Neben Gruppenleitern sollen insbesondere auch wissenschaftliche Mitarbeiter sowohl als Teilnehmer als auch als Vortragende mitwirken.

[1] http://agrosy.informatik.uni-kl.de/informatik2014/

Organisationskomitee: Prof. Dr. Karsten Berns (RRLAB, TU Kaiserslautern), Prof. Dr.-Ing. Rüdiger Dillmann (HIS, Karlsruher Institut für Technologie), Prof. Dr.-Ing. Erik Maehle (Institut für Technische Informatik, Universität zu Lübeck)

High-Level Robot Control with ArmarX

Nikolaus Vahrenkamp, Mirko Wächter, Manfred Kröhnert,
Peter Kaiser, Kai Welke, Tamim Asfour

High Performance Humanoid Technologies (H^2T)
Karlsruhe Institute for Technology (KIT)
Adenauerring 2
76131 Karlsruhe
{vahrenkamp,waechter,kroehnert,kaiser,welke,asfour}@kit.edu

Abstract: The robot development environment (RDE) ArmarX aims at providing an infrastructure for developing a customized robot framework that allows realizing distributed robot software components. This includes communication properties, start-up and error handling, mechanisms for state implementations, interface definitions and concepts for the structured development of robot programs. In addition to this core functionality, we will show in this paper that ArmarX provides customizable building blocks for high level robot control and how these components can be used to build a generic backbone of the robot software architecture. ArmarX provides standard interfaces and ready-to-use implementations of several core components which are needed to setup a distributed robot software framework.

1 Introduction

In the last decade, service and humanoid robot technologies have made enormous progress in terms of integrating sensori-motor capabilities (see Figure 1). The trend towards dynamic and interactive areas of application leads to complex, distributed and asynchronously operating software systems where the robot state is distributed over multiple components. These dedicated distributed components can be developed and tested independent of the whole system, for example in a simulation environment. Nevertheless, software development complexity increases due to synchronizing aspects, concurrent access, data distribution and all challenges that arise in parallel software development.

To support development and integration of all required capabilities of such robots is the goal of robot development environments (RDEs). Different functionalities and capabilities of the robot software should be glued together by such environments in terms of system architecture and communication. In addition, RDEs should provide programming interfaces, allowing programmers to include new capabilities at appropriate levels of abstraction with minimal training effort.

As stated in [WVW+13] several RDEs have been presented during the last decades accompanying the development of robotic platforms. Different levels of abstraction are realized by RDEs, depending on the robot platform and its intended application.

Figure 1: The humanoid robots ARMAR-IIIa [AAV+08] and ARMAR-IV [ASP+13].

Several RDE frameworks put their focus on the control level, such as OpenRTM [ASK08], MatLab/Simulink®, MCA[1], whereas others focus on the implementation of higher level capabilities of the system (e.g. ROS [QCG+09], Yarp [MFN06] and Orocos [BSK03]).

In [WVW+13] we presented the ArmarX RDE framework and its disclosure capabilities. Compared to earlier work, we will focus on the layered architecture of ArmarX in this paper. In addition, we will discuss the generic features of ArmarX that allow creating tailored robot frameworks on top of which application specific robot programs can be implemented.

2 The Robot Development Environment ArmarX

The Robot Development Environment ArmarX provides several components, functionalities, interfaces and generic ready-to-use components for building distributed robot applications. In contrast to traditional message passing protocols, well-defined interface definitions and flexible communication mechanisms are used. ArmarX is organized in several layers as shown in Figure 2. The *Middleware Layer* provides all facilities for implementing distributed applications. It abstracts the communication mechanisms, provides basic building blocks of distributed applications, and facilities for visualization and debugging. Bindings to a wide area of programming languages covering Python, C++, C# and Java are provided. While the *Middleware Layer* comprises a predefined set of components, interfaces and methods, the *Robot Framework Layer* serves several extendable and exchangeable components which can be assembled to an appropriate robot software framework. Currently provided are interfaces and corresponding components for memory, robot and world model, perception, and execution. Since no framework is suitable for all robot systems, this layer can be customized and adapted to the demands of the specific robot platform in order to tailor the robot framework to the needs of the application of the robot. ArmarX comprises several customizable components like memory structures

[1] http://www.mca2.org/

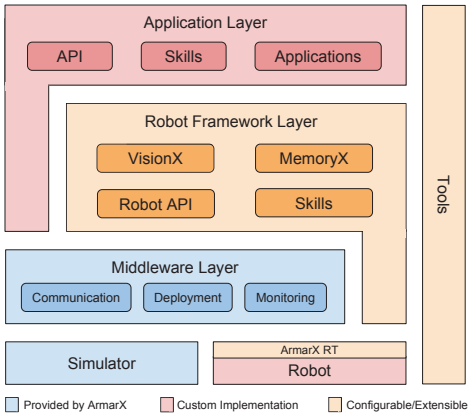


Figure 2: The Middleware Layer of the ArmarX framework provides core functionalities like distributed communication, deployment, start up mechanisms, and network transparent statecharts. The Robot Framework Layer consists of several ready-to-use parameterizable and extendable components regarding perception, memory and generic robot API methods (e.g. kinematics and visualization). Additionally, a set of basic skills such as visual servoing, grasping or navigation are included. Robot programs are developed in the Application Layer supported by the ArmarX toolset. The hardware abstraction layer (bottom of the figure) implements the interfaces of the Robot Framework layer for a specific hardware type or the simulator. Different colors indicate at which points extensions can be realized. Blue layers provide ready-to-use concepts without the possibility for extension. Red layers require user specific implementations. Within orange layers, several ready-to-use components are available which can be parametrized and/or exchanged by user specific implementations.

with interfaces to corresponding update processes compatible with the perception components. Additionally, a generic robot API is provided which includes a network transparent representation of the robot's sensor state, generic methods related to kinematics, collision detection, motion and grasp planning and scene representation. A set of basic skills is provided as configurable statecharts and can be adapted to the needs of the robot framework. The robot application is realized as a network transparent statechart within the *Application Layer*. ArmarX supports developing application specific extensions to APIs and the skill library. Tool support comprises GUI elements, a statechart editor, a physics-based simulation environment and state disclosure mechanisms for online inspection and debugging.

As shown in [WVW+13], the main design principles of ArmarX are distributed processing, interoperability, and disclosure of system state. In this work we will focus on the interoperability features of the framework provided by the ArmarX Middleware Layer, allowing the implementation of ArmarX components in different programming languages and their execution on different operating systems. Another aspect of interoperability is covered by the Robot Framework Layer. The proposed structures allow tailoring the software framework according to application-specific requirements and the capabilities of the robot. To this end, ArmarX provides generic building blocks which can be customized, parameterized, and extended to ease the possibility of embedding ready-to-use components on different robot platforms.

3 ArmarX Middleware: The Core Layer

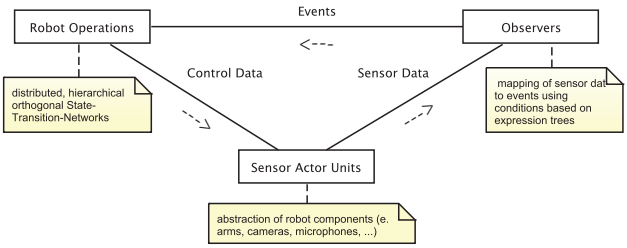


Figure 3: The application programming interface comprises four different elements. *Sensor-Actor Units* serve as abstraction of robot components, *Observers* generate events from continuous sensory data streams resulting in transitions between *Operations*. Operations are organized as hierarchical state-transitions networks. All elements are connected via the *communication* mechanisms (arrows in the figure).

The middleware layer of ArmarX comprises four main building blocks: *inter-object communication*, *Sensor-Actor Units*, *Observers*, and *Operations*. *Inter-object communication* provides convenient mechanisms for communication between objects in the system. These objects can be any of the other buildings blocks. The *Sensor-Actor Units* offer a generic interface on a low level of abstraction for robot components like motor controllers or sensors. ArmarX includes a default set of sensor-actor units for common robot components, e.g. joints, force/torque-sensors and haptic sensors. Implementing additional sensor-actor units for unsupported components is encouraged and straightforward. *Observers* monitor sensor values from sensor-actor units and send application specific events to configured *Operations*. These operations process events and send resulting control commands to the sensor-actor units controling the robot. More details can be found in [WVW+13]. In addition to these building blocks, monitoring facilities for data flow and system state are incorporated into the ArmarX framework.

3.1 Inter-Object Communication

As shown in [WVW+13], ArmarX uses well established communication mechanisms (e.g. Ethernet, TCP/IP, UDP) to provide convenient ways to create distributed applications for robot architectures. To this end, we employ the Internet Communication Engine (Ice) by ZeroC as fundamental communication middleware [Hen04].

Ice implements type-safe distributed communication over Ethernet with a variety of supported platforms and programming languages. It manages the connections between objects and provides convenient communication modes such as synchronous and asynchronous remote procedure calls and publisher-subscriber mechanisms. Communication interfaces are defined in the platform and programming language independent Interface Definition Language (IDL) *Slice*, which is then compiled to the chosen programming language.

3.2 Monitoring

Monitoring is an important concept for any robotic system. ArmarX provides means to retrieve CPU load, network utilization, IO operations, and data update frequency information. Additionally, generated sensor measurements, control flow information as well as more abstract information like condition histories can be retrieved.

Monitoring information can be gathered within a distributed ArmarX robot program for profiling running robot programs and identifying bottlenecks in the distributed application. This information can be used to improve the deployment of ArmarX components and therefore the load balancing in order to provide seamless program execution. Furthermore, resource usage profiles of ArmarX components can be generated by performing offline analysis on monitoring data.

Using the disclosure mechanisms of ArmarX [WVW+13] it is even possible to monitor the execution of robot program statecharts to build an execution model of the robot software components. Combining such execution models with generated resource profiles allows for creating control-flow dependent resource usage diagrams.

4 ArmarX RT: Real-Time Hardware Abstraction

Sensor-actor units provide a sufficient and thin hardware abstraction when real-time requirements do not apply, e.g. when working in simulation or when using inherently stable robots. However, for many robots, especially bipeds, the low-level hardware control software needs to satisfy real-time constraints in order to maintain the robot's stability and safety. To address this issue, ArmarX supports a second layer of hardware abstraction that is intended to run in a real-time context on a Xenomai[2] real-time OS.

The purpose of this hardware backend is to control the attached hardware entirely within a real-time context. Communication between hardware backends and sensor-actor units happens via Xenomai's real-time pipes. Figure 4 depicts the complete hardware abstraction layer including a generic real-time hardware backend for CANopen[3]-based robots that is included in ArmarX. Additionally, adding real-time hardware backends for other hardware as well as operating systems other than Xenomai is supported.

5 ArmarX Robot Framework: Generic Building Blocks for High-Level Robot Control

ArmarX provides a generic robot framework layer that offers standard components and interfaces to ease development of tailored robot frameworks. Several generic and ready-to-use API components can be parameterized and customized in order to build robot specific

[2] http://www.xenomai.org
[3] http://www.can-cia.org

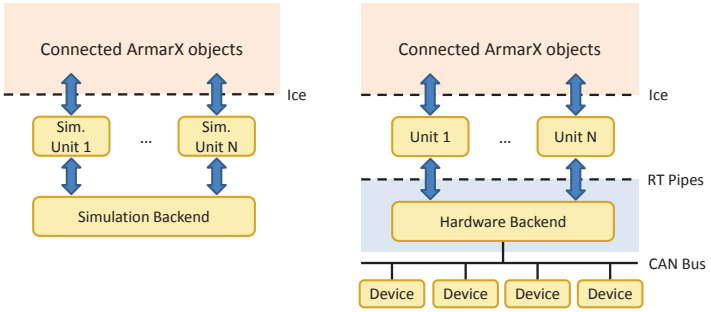


Figure 4: ArmarX hardware abstraction for a simulated robot (left) compared to the real-time capable hardware abstraction for a CANopen based robot (right). Both setups are possible with ArmarX. Switching between a simulation backend and the real hardware is completely transparent to higher-level ArmarX objects.

APIs. Robot model, world model, perception and memory components are available, each providing entry points for custom robot program implementations. These components will be briefly discussed in the following section.

5.1 Robot API and Skills

The robot API comprises several components which can be used off-the-shelf by specifying required parameter settings. If needed, the API can be extended in order to reflect specialized capabilities of the robot.

5.1.1 Kinematics, Robot and World Models

The internal robot model consists of the kinematic structure, sensors, model files and additional parameters such as mass or inertial tensors. The data is consistently specified via XML or COLLADA files compatible to the robot simulation toolbox *Simox* [VKU+12]. This allows making use of all Simox related functionalities, such as inverse kinematics, collision detection, motion and grasp planning or defining reference frames for coordinate transformations. Further, ArmarX provides network transparent access to a shared robot model data structure. This allows all distributed components to query robot internal data such as joint values or to perform coordinate transformations.

5.1.2 Kinematic Units and Robot Skills

The robot API provides reference implementations for simulating kinematic units. Several control modes such as position, velocity or torque can be used out of the box. The robot skill components cover basic functionality like Cartesian manipulator control, zero force control, platform navigation, unified end effector access, and visual servoing. This skill

library can be used by parameterizing related components, by extending them and/or by re-implementing the corresponding interfaces.

5.2 MemoryX

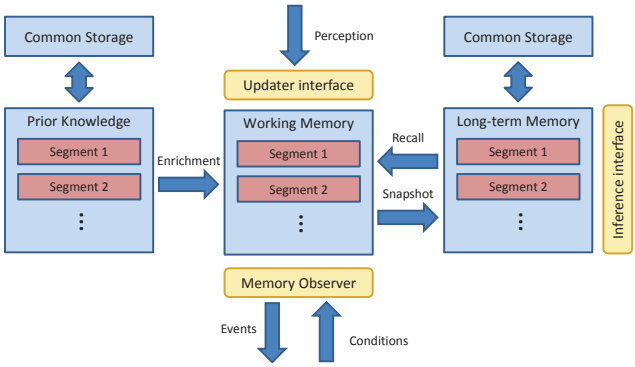


Figure 5: The offered MemoryX architecture consisting of a working memory and a long-term memory. Both memories can be accessed within the distributed application. Appropriate interfaces allow attaching processes to the memory for updating and inference.

MemoryX, the memory layer of ArmarX, includes basic building blocks for memory structures which can either be held in the system's memory or made persistent in a non-relational database [WVW+13, WKK+13]. Based on these building blocks, the memory architecture illustrated in Figure 5 is realized. A key element of MemoryX are network transparent access facilities which allows consistently updating or querying the memory within the distributed application. The architecture consists of different memory types such as working memory (WM), long-term memory (LTM), and prior knowledge (PK). Each memory type is organized in segments which can be individually addressed and used to store arbitrary data types or classes. Each segment may cover data related to classes of known objects, perceived locations of object instances or positions and status of agents in the world.

- The *Working Memory* represents the current context of the robot's internal state. It can be updated by perceptual processes or prior knowledge via an updater interface.
- *Prior Knowledge* contains information which is already known to the robot and which has been predefined by the user or operator. Entities in the PK can be enriched with known data such as 3D models or features for object recognition.
- The *Long-Term Memory* provides long-term storage capabilities and offers an inference interface which allows attaching learning and inference processes. The LTM allows creating snapshots of the current WM state to be used for later re-loading.

Although PK and LTM provide capabilities for long-term data storage, both memories differ in their conceptual role within MemoryX. While the PK holds user generated data and provides enrichment methods for entities of the WM, the LTM is used to create snapshots of the current content of the WM in order to recall them later and/or use them for inference processes.

As depicted in Figure 5, the different memories can be organized in segments. MemoryX provides several types of segments which are listed below:

- The *Object Classes* segment holds information on classes of objects and their relation. Enrichment with 3D models and or localization methods are provided.
- In the *Object Instances* segment information is stored about the positions of actual object instances which have for example been recognized by the perception system. These instance entities link to the corresponding object types within the Object Classes segment.
- The *Agent* segment stores information related to agents (e.g. robots, humans) in the world. The robot itself and any other operators, users, or robots relate to entries in this memory segment.

This memory structure allows the realization of powerful update mechanisms. For example, a perception component localizes a known object and updates the WM. This update creates a new entity of the object with the current location and automatically enriches it with the 3D model from the prior knowledge database. Besides the possibility of directly addressing the WM, an observer exists which allows installing conditions based on the memory content. If the associated content changes the matching events will get generated and processed by the according states of the robot program.

Note, that the memory components are expandable in order to customize the memory structure. However, when a lightweight framework is sufficient, it is possible to use only a subset of the components in order to avoid bloated setups.

5.3 VisionX

With VisionX we provide a perception building block in the robot framework layer which offers facilities for sensor processing in the distributed robot application. Processing currently covers stereo camera images, RGB-D data and offline video streams coming from persistent data files. VisionX consists of interfaces for image providers and image processors as illustrated in Figure 6. The image provider abstracts the imaging hardware and provides the sensor data as a stream either via shared memory or over Ethernet. Different image processors can be implemented that realize tasks related to object perception, human perception, and scene perception. The MemoryX update interface is used to write the processing results to the working memory. Optionally, perception confidence and probabilistic measures can be stored additionally in order to support quality queries and to allow for probabilistic reasoning.

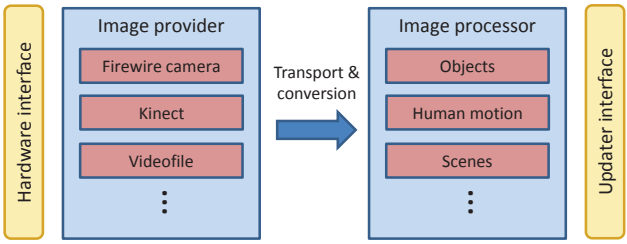


Figure 6: Image processing in the VisionX robot API. The image provider abstracts the hardware and streams the data via shared memory or Ethernet to the image processor. The processing result is written to the working memory.

6 The ArmarX Toolset

ArmarX is equipped with a set of tools designed to ease the development process in the creation and debugging phase. ArmarXPackage, a command line tool to easily create and manage own projects, several GUI viewer tools to inspect the system state on different levels, and a simulator are the key applications of the ArmarX toolset.

6.1 ArmarXPackage

Creating new C++ projects by hand is a tedious and often repetitive task. It is common practice to recycle an existing project structure and adapt it to new demands, often differing in details only. On this account, ArmarX offers the command line tool ArmarXPackage to easily create and manage user projects that depend on ArmarX. With this tool, standard ArmarX components like network accessible objects, statecharts or gui-plugins are created from templates and integrated into the project structure.

6.2 ArmarXGui

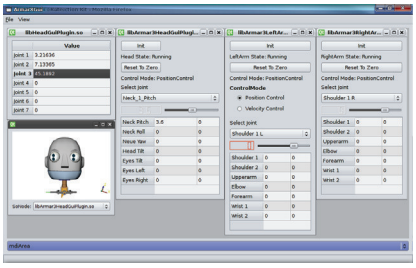
Disclosure of the system state is an important aspect of the ArmarX RDE and allows programmers to access data of many parts of the system they are working on. This data is required for debugging, monitoring and profiling purposes. Since, the amount of available data increases with the size of the developed system an abstraction of the data flow into easy to grasp visualizations is required.

To allow creating visualizations easily, we provide an extensible graphical user interface (GUI) named *ArmarXGui* which is based on the well-established Qt framework. An ArmarXGui instance provides extension points for loading and displaying custom user sup-

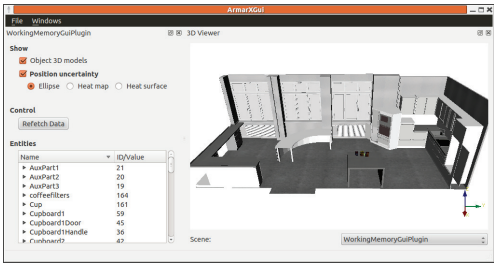
plied plugins. Each plugin has the ability to communicate with the ArmarX framework and offers a visualization which is displayed in the GUI's window. A variety of ready-to-use GUI plugins is already available within ArmarX, such as LogViewer, data plotter, 3D visualization widgets or MemoryX inspection tools.

An exemplary GUI layout is presented in Figure 7(a) containing a 3D visualization and plugins related to several sensor-actor units of the robot. These sensor-actor plugins can be used for emitting control commands as well as for visualizing internal data such as joint values. Below, an exemplary selection of GUI plugins are listed:

- **LogViewer** This plugin allows displaying, grouping, filtering, and searching of all gathered log messages emmitted by ArmarXObjects anywhere in the distributed system. Furthermore, it is possible to inspect extended information such as component state, application IDs and back tracing information.
- **Working Memory** The content of the robot's working memory can be visualized in a 3D view of the current environment. Figure 7(b) depicts an exemplary snapshot of the current content of a robot's working memory.
- **StatechartViewer** Since one statechart design principle has been state disclosure, it is possible to extract and visualize the program logic of an ArmarX robot program at runtime. With the StatechartViewer plugin it is possible to visualize each hierarchy level in the statechart of a robot program. All substates within the hierarchy level, connecting transitions as well as input- and output parameters of each state can be inspected. Both data flow and the currently active state at each hierarchy level are continuously updated.
- **Prior Memory Object Editor** This plugin offers a convenient way to add, inspect and update the object data that is stored in the *Prior Memory*. Properties such as the 3D object model, the recognition method, the motion model or parent object types can be viewed and changed.



(a) Gui with 3D visualization and sensor-actor unit plugins.



(b) Gui plugin for inspecting the current content of the robot's Working Memory.

Figure 7: ArmarXGui provides a variety of ready-to-use widgets and can be extended by a plugin mechanism in order to customize graphical user interaction

6.3 ArmarX Simulator

In order to ease development of robot applications it is of high interest that an RDE provides mechanisms to simulate program flow and robot behavior. Although a simulation environment cannot calculate the fully correct physical interaction, sensor information or execution timing, the structural setup of a robot program, in particular when distributed components are used, and an approximated physical robot behavior can be tested before the application is executed on the real robot. The *ArmarX Simulator* provides such a simulation environment within the ArmarX framework. The simulator comprises simulations of motors, sensors and dynamics and communicates with the ArmarX framework by implementing the robot's *KinematicUnit* interfaces in order to receive motor commands and to serve sensor feedback. This enables developers to run a robot program completely in simulation in order to test and inspect the program logic. An exemplary scene is shown in Figure 8.

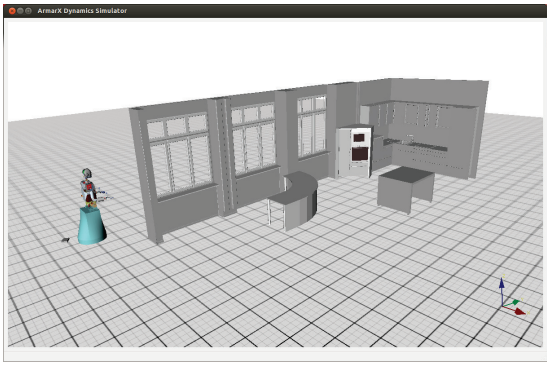


Figure 8: The KIT kitchen scene in the ArmarX Simulator.

7 Conclusion

The robot development environment ArmarX provides infrastructure for building customized robot frameworks to ease the development of distributed robot software. The layered architecture of ArmarX supports lower level functionalities like access to real time components (*ArmarX RT*) and core features like communication, deployment, start-up, and error handling (*ArmarX Middleware Layer*). Additionally, ArmarX offers ready-to-use framework components for perception, memory, and control that can be used with little programming effort but remain open for modification and extension (*ArmarX Robot Framework Layer*). These customizable building blocks for high level robot control can be used to build a robot software framework and to tailor it according to robot capabilities and demands of potential robot applications. Finally, the ArmarX toolset offers a variety of components to support the implementation and debugging of complex robot applications.

8 Acknowledgement

This work was partially supported by the European Union Seventh Framework Programme under grant agreement no. 270273 (Xperience) and the German Research Foundation (DFG) as part of the Transregional Collaborative Research Centre Invasive Computing (SFB/TR 89).

The authors would like to thank all members and students of the Humanoids group at KIT for their various contributions to this work.

References

[AAV+08] T. Asfour, P. Azad, N. Vahrenkamp, K. Regenstein, A. Bierbaum, K. Welke, J. Schröder, and R. Dillmann. Toward Humanoid Manipulation in Human-Centred Environments. *Robotics and Autonomous Systems*, 56:54–65, January 2008. 1

[ASK08] Noriaki Ando, Takashi Suehiro, and Tetsuo Kotoku. A Software Platform for Component Based RT-System Development: OpenRTM-Aist. In *Proceedings of the 1st International Conference on Simulation, Modeling, and Programming for Autonomous Robots*, SIMPAR '08, pages 87–98, Berlin, Heidelberg, 2008. Springer-Verlag. 1

[ASP+13] T. Asfour, J. Schill, H. Peters, C. Klas, J. Bücker, C. Sander, S. Schulz, A. Kargov, T. Werner, and V. Bartenbach. ARMAR-4: A 63 DOF Torque Controlled Humanoid Robot. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2013. 1

[BSK03] Herman Bruyninckx, Peter Soetens, and Bob Koninckx. The Real-Time Motion Control Core of the Orocos Project. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2766–2771, 2003. 1

[Hen04] M. Henning. A new approach to object-oriented middleware. *Internet Computing, IEEE*, 8(1):66–75, 2004. 3.1

[MFN06] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. YARP: Yet Another Robot Platform. *International Journal on Advanced Robotics Systems*, 2006. 43–48. 1

[QCG+09] Morgan Quigley, Ken Conley, Brian P. Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y. Ng. ROS: an open-source Robot Operating System. In *ICRA Workshop on Open Source Software*, 2009. 1

[VKU+12] N. Vahrenkamp, M. Kröhnert, S. Ulbrich, T. Asfour, G. Metta, R. Dillmann, and G. Sandini. Simox: A Robotics Toolbox for Simulation, Motion and Grasp Planning. In *International Conference on Intelligent Autonomous Systems (IAS)*, pages 585–594, 2012. 5.1.1

[WKK+13] K. Welke, P. Kaiser, A. Kozlov, N. Adermann, T. Asfour, M. Lewis, and M. Steedman. Grounded Spatial Symbols for Task Planning Based on Experience. In *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, pages 474–491, 2013. 5.2

[WVW+13] K. Welke, N. Vahrenkamp, M. Wächter, M. Kroehnert, and T. Asfour. The ArmarX Framework - Supporting high level robot programming through state disclosure. In *INFORMATIK Workshop*, 2013. 1, 2, 3, 3.1, 3.2, 5.2

Service Robot Control Architectures for Flexible and Robust Real-World Task Execution: Best Practices and Patterns

Matthias Lutz, Dennis Stampfer, Alex Lotz, Christian Schlegel

Fakultät Informatik
Hochschule Ulm
Prittwitzstrasse 10
89075 Ulm
{lutz, stampfer, lotz, schlegel}@hs-ulm.de

Abstract: Successfully building complex service robotic systems that robustly operate in real-world environments gives a lot of insights into what are valuable patterns in architecting such systems. In this paper, we describe some of our insights with respect to robot control architectures and robotics software systems engineering. The focus of the paper is not on discussing specific functionalities but to explain some best practices that evolved out of our experience in building complex software intensive robotic systems. These behind-the-scenes insights are in particular relevant in moving forward towards a robotics business ecosystem with separation of roles and based on model-driven approaches for handling systems of systems.

1 Introduction and Motivation

Robotic systems are complex, software intensive and heterogeneous composite systems. Robotics as a science of integration depends on structures that guide the overall system design, the system integration and that even support run-time adaptation according to the executed task, the current context and the available resources. Such structures called architectural principles for robotics do not appear out of nowhere but should reflect and explicate practices and techniques developed and matured over many years of experience within all kinds of projects and with involvement of all kinds of stakeholders.

We consider architectural principles good if they provide partitioning schemes that organize different and various views on a robotics system such that one can effectively and efficiently cope with the complexity of the whole lifecycle of such a system (design, composition, deployment, operation, evolution, adaptation etc.).

There are partitioning schemes that are generally considered a good engineering practice and that are independent of the robotics domain. For example, *separation of concerns* [Chr89][Dij76] [Par72][BEH+11][RE96] is one of the most fundamental principles in software engineering and plays a vital role in robotics software systems as well [SSL12b][VKB14]. It aims at identifying orthogonal concerns in order to address com-

plexity by decoupling. It should drive the identification of the right decomposition of a problem. *Separation of roles* [SSL12b] is another partitioning scheme that proved to be a very successful approach for organizing business ecosystems where stakeholders have mutual benefits in collaborating and competing and where they share efforts and risks. Nevertheless, there is still the need to *identify* and to *explicate* behind-the-scenes insights related to achieving *separation of concerns* and of *separation of roles* in robotics.

The focus of the paper is not on discussing specific functionalities but to explain some best practices that evolved out of our experience in building complex software intensive robotic systems [Ulma] [Ulmb] (figure 1).

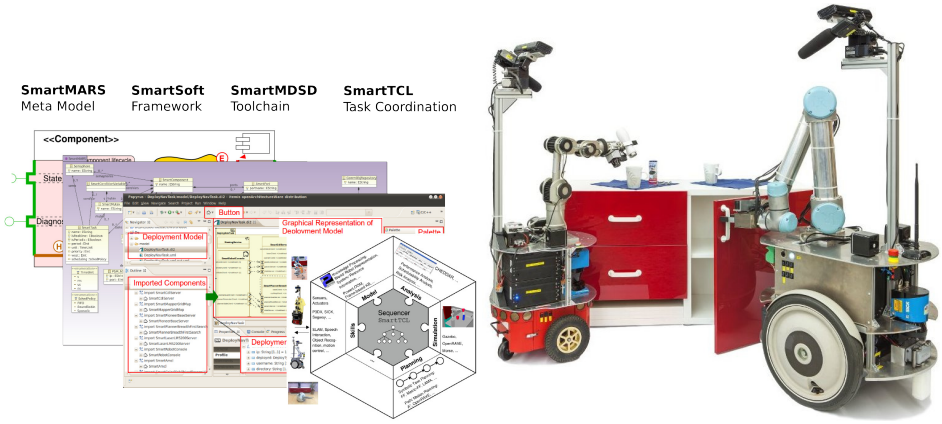


Figure 1: SMARTSOFT toolchain (left), service robots Kate and Larry (right)

2 Freedom from Choice vs. Freedom of Choice

We consider a generic robotics reference architecture that is refined in a top-down approach to end up with more and more specific architectures for dedicated robotic systems as illusive and unrewarding. A specific robotic control architecture always is a trade-off between various conflicting requirements. Thus, different co-existing architectural principles and their implementations represent different trade-offs of the interplay between e.g. modularity, adaptability, performance, dependability, resources etc. For a lawn-mower robot in a high-volume market, it might make perfect sense to loose extensibility and modularity in its software in order to save memory and processing power. However, it might also make sense to preserve some modularity in order to be able to support e.g. customization by user-downloadable apps.

The interesting question is what kind of architectural patterns form the sweet spot in supporting *separation of concerns* and *separation of roles*. That immediately comes along with the question of how much structure is needed for which aspects and is beneficial in which settings and how to support conformance to these structures.

One approach is called *freedom of choice*. One tries to support as many different schemes as possible and then leaves it to the user to decide which one best fits his needs. However, that requires huge expertise and discipline at the user side in order to avoid mixing non-interoperable schemes. Typically, academia tends towards preferring this approach since it seems to be as open and flexible as possible. However, the price to pay is high since there is no guidance with respect to ensuring composability and system level conformance.

Freedom from choice gives clear guidance with respect to selected structures and can ensure composability and system level conformance. However, there is a high responsibility in coming up with the appropriate structures such that they do not block progress and future designs.

An appropriate sweet spot between *freedom of choice* and *freedom from choice* can only be found by first agreeing on paramount objectives that give guidance. In our view, these are the need for *separation of roles* and *separation of concerns*. In line with the subsidiarity principle, architectural patterns should impose structures only as far as these cannot be achieved at a more local level. For example, structures and patterns that ensure *composability* restrict freedom of choice in order to allow *separation of roles* between component developers and system integrators. Typically, industry tends towards *freedom from choice* by agreeing on standards that is the minimum set of structures required to establish a business ecosystem.

An example for *freedom of choice* is ROS. One of the main design rules the ROS founders follow is *"We do not wrap your main."* [CGCG10]. What they mean by this is, among others, that they do not want to enforce any architectural design decisions for developers using ROS. As consequence, with ROS every developer can use his own personally preferred architecture. It then is very likely that the implemented architecture is in conflict with those defined by other parties. As also stated in [CGCG10], this can lead to confusions as this inevitably leads to the need for every ROS user to first understand the architectural decisions of each individual component before being able to reuse them in their own systems. The proposed solution in [CGCG10] for this problem is just to extensively document each ROS component on the ROS web portal. However, this does not circumvent the need to extensively analyse and understand the source code in order to adjust it or to implement workarounds in order to somehow make components compatible and reusable.

It is obvious that ROS (according to its overall design philosophy) does not yet give enough structure in an appropriate format in order to better address *separation of roles* and *separation of concerns* as is needed for a robotics business ecosystem. The minimum needed structures are a sound component model (in order to separate at least the role of component developer and system integrator and still ensure composability). Furthermore, it has to be formalized for use in model-driven tools in order to support *separation of concerns* (e.g. to maintain semantics independent of the OS / middleware mapping) and to assist the different roles in conforming to superordinated structures. The component model and model-driven toolchains of SMARTSOFT are following the approach of *freedom from choice* and explicate those structures[SSL12b].

3 Best Practices in Designing System Architectures

The design of the overall system architecture is an important part of the development process. The challenge is to break down the overall system into smaller units that can be solved by the individual stakeholders / project partners. The goal is to enable them to focus on their individual parts of the system by providing building blocks (software components) which can later be composed and integrated to the final application and target system.

Figure 2 illustrates the workflow showing our best practice in developing service-oriented component based architectures and applications. In a first step, services are defined (1), then they are reused and aggregated to components in order to provide components and groups of components (2). The outcome is a set of building blocks out of which applications can be composed (3).

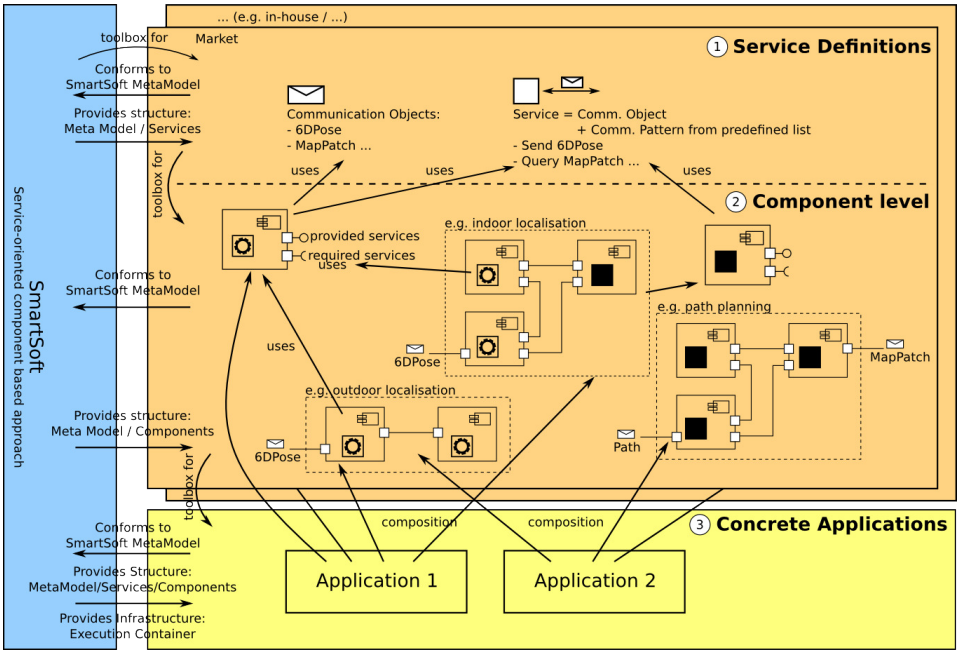


Figure 2: Best practices workflow for designing system architectures

The service-oriented component-based approach SMARTSOFT provides structure, infrastructure and tool support at all levels and steps (Figure 2, left). Amongst others, it includes:

- SmartMARS MetaModel. It defines the structure for communication objects, a set of communication patterns, the structure of services and components, the structure for deployment.
- SMARTSOFT Framework and implementation. In its current state, two exchangeable implementations (ACE- and CORBA-based middlewares). Execution containers for several platforms and operating systems.

- SMARTSOFT MDSD Toolchain. An integrated MDSD toolchain for development that supports the *separation of roles*. The toolchain covers the development process of modelling communication objects, components and systems.

The first level **Service Definitions** (Figure 2: (1)) uses the structure provided by SMARTSOFT to define *Services* that might be provided (made available) or required (relevant) in an application. Services are the basic architectural entities. They guarantee that supplied components can be integrated into concrete applications. At the same time, they keep the architecture and implementation containers flexible due to the service-level and component-level abstractions. Further, they allow to identify white spots in the architecture as early as possible (required services that no-one provides or provided services that no-one requires).

Services consist of communication objects [SSL12a] (data structure for communication: attributes, data types, access methods) and one selected communication pattern [SSL12a] (how a communication object is being communicated). A communication object together with one selected communication pattern (out of a set of communication patterns defined in SMARTSOFT) becomes a service.

For example, at this level the user defines what a "location" for the robotics system actually is, i.e. whether it is represented as full 6D cartesian pose or with geographical position coordinates with or without uncertainty, orientation, etcetera. He can even decide on separate communication objects for both types of representations if required in the domain. It is also defined that e.g. a service providing regular location updates every 0.2s (push timed pattern for location communication object with period 0.2s) is required in the considered domain.

The second level **Component Level** focuses on aggregating services to components or groups of components (Figure 2: (2)). They provide or require the services from (1). There might be competing alternative components providing the same services from different suppliers with different characteristics. Both open source and closed source implementations can be used since one can rely on service descriptions. For example, the common denominator for the core functionality of localisation could have been defined in (1) as the communication object "6DPose" (x, y, z, yaw, pitch, roll) that needs to be pushed to subscribers. Now there might be alternative components or groups of components for different localisation scenarios and accuracy requirements which consist of a different structure of components and are provided by 3rd party component suppliers in a market. These alternatives become alternatives only because they provide the same service ("periodically publish 6D pose") that makes them exchangeable. Developers can choose between alternatives in the design phase and alternatives might even be swapped at runtime. Again, this level conforms to the SMARTSOFT MetaModel which provides structure in the form of components that can provide and require services from (1) and therefore provide further structure to realize / implement them.

Finally, there are **Concrete Applications** in the last step (Figure 2: (3)). When users develop new applications, these applications can rely on service definitions in order to either implement these services or reuse 3rd party implementations of these services. Compatibility is given by the definition of services (1). SMARTSOFT also contributes to this

level by providing an execution container for components. It maps the component hull [SSL12b] to the execution environment (operating system and processor architecture).

From our experience, the best practice is to regard services as the most important building blocks for designing the architecture. Even though components implement functionalities and provide/require these services, the granularity of components is arbitrary and not decisive for the architecture. Component granularity might differ from application to application, whereas the granularity of services can remain the same. For example, the services ”localization” and ”navigationCommand” might be provided by two components from two suppliers or by one single component as long as it provides the defined services. The service-oriented component model naturally supports competing alternatives as building blocks for systems while still ensuring composability. Thus, openness with respect to software service definitions is in no way in conflict with the progress in implementations.

Services as stable entities in the design of systems ensure system level conformance, define responsibilities and allow to identify white spots that have to be covered in a very early stage of the workflow. The strength of this structured workflow is that repeating architectural patterns as well as project or application specific structures can be identified and defined very early and in parallel to the implementation of functionalities. This decoupling from architecture design and implementation has proven successful in the later integration where the system parts (components) just fit together. To an extend, this design is also independent from the hardware architecture: SmartSoft provides abstractions that leave space for integrating hardware by either providing the execution containers to (computational) hardware platforms or by transforming hardware specific sensor values into standardized services.

The goal during service discussions is to converge to the least common denominator of communication objects and services that are relevant in the individual domain. The smaller this set is, the better is the composability that can be achieved with building blocks (software components) that provide and require these services. However, one can at the same time keep specific communication objects and specific services in order to best exploit unique abilities and features of the software components. Finally, it is about finding the right balance between ”too specific” and ”too general” in order to support composability and reuse while not loosing unique characteristics. The illustrated workflow helps in this respect at a very early stage.

Decoupling of services and components brings the decisive flexibility that is a must-have for system composition and establishing a software business ecosystem, for both in-house building blocks and towards a ”market” of building blocks for robotics. At the same it time is a structured approach that ensures system level conformance.

Applying this workflow and discussing the services of a system with project stakeholders at a very early stage also helps each of them: by thinking in detail about what services they require or what services they provide, stakeholders reported that the workflow is very helpful for their own components and expectations.

4 Communication Objects as Self-Contained Entities

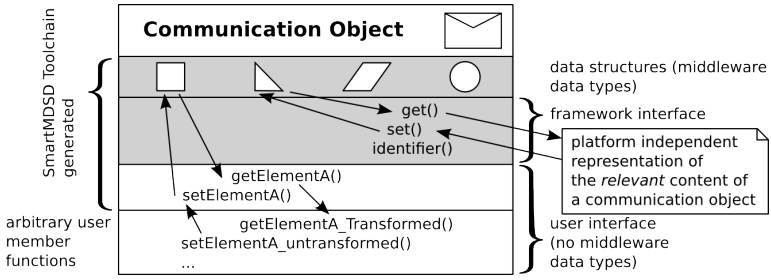


Figure 3: Communication Object with its interfaces for user and framework.

Communication objects [SSL12a], as shown in figure 3, define the data structures (figure 3, grey upper) to be transmitted via a communication pattern between components. Communication objects are C++-like objects decorated with additional member functions (figure 3, grey lower) for internal use by the framework. Besides attributes for communication, communication objects can define user access functions (figure 3, bottom).

It is our best practice to define communication objects as self-contained local entities, communicated *by value* between components via services. All the information needed to process communication objects within a component should be contained within that communication object and should not need additional communication in order to get that information. Communication objects should not only transport data, but also provide a locally extensible interface (user member functions: figure 3, white elements) to access the data, without influencing the overall system.

The self-containment of communication objects avoids fine grained intercomponent communication when processing them, thus helps to achieve one of the most important goals of service-oriented component based approaches: decoupling the sphere of influence between the components (*separation of concerns*). Loosely coupled components enable the *separation of roles*, reuse and easy system coordination.

For example a communication object for sensor data is tagged with a position (itself a communication object) where it was recorded. User member functions of the communication object perform the transformations from the sensor data into different formats, units, coordinate systems, etcetera. This is done without the need of further communication and coordination, keeping the effects limited to the local entity (component), as well as without reimplementing this logic over and over again. Regardless of the point of time when the communication object (e.g. camera image) is used, the information within the object (e.g. capture pose) forms a valid snapshot of the relevant system state in relation to the communicated data.

Communication objects that are not self-contained would require further intercomponent communication that would add to further system and coordination complexity. Other components would have to be activated and configured to get the additional information. Self-

contained communication objects typically also reduce the total amount of required inter-component connections, which again reduces system and coordination complexity. From an architectural point of view this best practice supports an early and expressive design of the components services, as all the relevant information is explicated and collected in the design phase of a system.

User member functions of the communication object provide access to data structures in those data types that are required by the user in the individual components. Commonly used functionality to access the data (e.g. transformation of their representation) can be implemented once and are provided with the communication object. Local extensions of communication objects, extending data types or user member functions, are implemented in locally inherited communication objects and do not spread across the system.

The user access functions are also used to prevent the communication middleware data types from polluting the user space. Using the SmartMDSD Toolchain, the interface functions (figure 3, white upper) to those communication middleware data structures are generated to make the user access function in the communication object independent of the underlying communication middleware, following the principle of *separation of concerns* in the same way as the SMARTSOFT components [SLL+13]. This ensures user code (figure 3, white lower) that is independent of the underlying communication middleware.

5 System Orchestration

System orchestration deals with coordinating the individual parts (components) of the system in such a way that the overall system performs a concrete task. It consists of plenty of challenges and the following will concentrate on best practices that enable *separation of concerns* and roles.

A robot system is separated into multiple layers by distinguishing hierarchical task decomposition and situation-driven task execution from fast reactive control algorithms on the skill layer. The *sequencer* bridges continuous processing and event-driven task execution. It orchestrates the software components in the system and assigns decision spaces to components. The components on the skill layer execute all kinds of algorithms, hardware drivers or control loops needed to perform the task guided by the *sequencer*. Details of the multi-layered system architecture can be found in [IRLVCS12] [SLL+13].

5.1 Coordination of Skill Components

Figure 4 shows an excerpt of the coordination within the collaborative robot butler scenario [Ulmb]. Given such a complex real world scenario (44 active components, open ended environment), the need for systematic coordination, reuse and separation of concerns becomes obvious.

The particularly interesting point is how the interface between *sequencer* and the skill

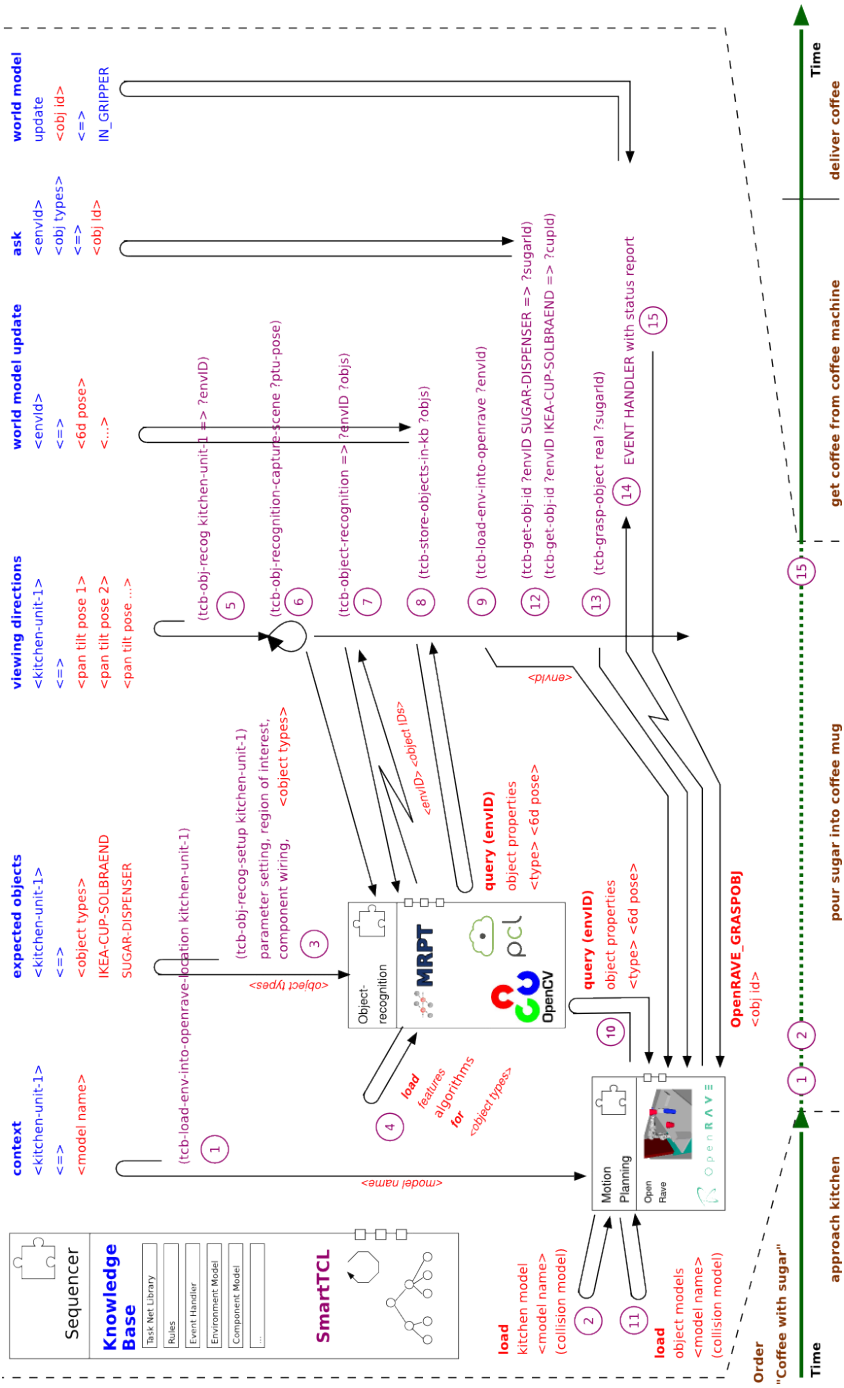


Figure 4: System coordination within the collaborative robot butler scenario. [Ulmb]

components is organized and how it distinguishes from the interfaces between each of the skill components. The interfaces within the skill components mostly consist of data (sub-symbolic) needed for reactive control algorithms such as sensor data, trajectories or sub-goals to approach, for example. The interface between *sequencer*, and skill components is used for coordination and typically includes symbolic data. One reusable pattern that is stable across all our systems is the orchestration cycle of the skill components driven by the *sequencer*: first the configuration of the components, then the execution or trigger of the actions in the components, events send to the *sequencer* driving it and fainally and optionally fetching resulting data from the components.

Given a task and the current context, the *sequencer* configures skill components according to the actions to be performed, using the SMARTSOFT parameter communication pattern. E.g. the object recognition component is configured to search for the object types *SUGAR-DISPENSER* and *IKEA-CUP-SOLBRAEND* (figure 4: (3)). Once configured, the components are activated for continuous unsing state communication patter or single operation using a trigger (parameter communication pattern). In figure 4: (7)), the object recognition component is triggered to perform recognition (once - triggered) given the configuration done before. The components run independent of further interaction with the *sequencer* as long as the skill components are able to operate within the configured decision spaces. If either the configured goal is achieved or components are not able to achieve the goal, an event using the corresponding SMARTSOFT event communication patter is send back from the individual skill component to the *sequencer*. The object recognition component fires an event received by the *sequencer* telling that objects with their corresponding *ids* where found (figure 4: (7)), for example. It is the *sequencer's* responsibility to activate those events from the skill components that are necessary to perform the task given the current context. Dependent on the event, the *sequencer* might then fetch the result of the performed action in symbolic representation from the skill component. The object recognition component is queried for the recognized object using the beforehand via event transmitted *ids* (figure 4: (8)), for example.

The separation of high-level task coordination and low-level control loops as well as the above described interface enforces the *separation of roles* and concerns in the development of complex systems. Explicating the configurations of components as well as their response (events), enables a systematic system orchestration and a hand over from the component developer to the role of the system integrator.

5.2 Distributed (Local) Knowledge Representation

At the *sequencer* level, information is stored and processed almost entirely in symbolic form. The sub-symbolic information and interaction is located at the level of the skill components. For many algorithms at skill level, reference data, models or configuration sets, etc. are required. The object recognition component for example needs all kind of modeled data to feed the different recognition algorithms, while other components such as the motion planning component again needs other views on the models. The individual models are best located close by the components using them (*separation of concerns*).

The coordination of the system however requires a common understanding of objects, locations, rooms etc. among the system parts.

Therefore, it is our best practice to connect the individual local views on the models and data with system-wide symbols, without spreading the local sub-symbolic information of component through the system. Those symbols are defined and linked to the modeled information at the *sequencer* level. Object types, for example are modeled at *sequencer* level and system wide uniquely identified by symbols (e.g. *IKEA-CUP-SOLBRAEND*). The object recognition component needs to get a shape model, whereas for manipulation planning a grasping model is needed. Both components get the reference to the modeled object type during the configuration using the unique symbol (figure 4: (4) and (11)).

This separation reduces the system complexity due to less intercomponent communication and keeps the responsibility for the local data bound to the role of the component developer while still being able of system wide coordination in the role of the system integrator.

6 Conclusions

All the described best practices and architectural principles were found and applied during many years of robotics software systems engineering. The SMARTSOFT approach has been used in various areas, by several users for many years. It has been used for education by students, e.g in the Robocup@Home competition. In our lab, every year a new team of students, without prior robotics knowledge, continues to work with the system of the previous team with no overlap. The black-box view of the components is the only way for the teams to continue and reuse the complex system parts. The SMARTSOFT approach has been used in several research projects, such as the *ZAFH Servicerobotik*, *FIONA* and *iserveU* and builds the stable foundation for many project partners in academia and industry. It is further used in intralogistic scenarios (e.g. [Ulmb]) in projects with partners from industry using the Robotino®3 robot.

Finding generic best practices in robot control architectures is about finding the sweet spot between those structures needed and those providing artificial restrictions. Structure is needed to ensure composability of the system parts which is vital for a robotics business ecosystem. Following the objectives of *separation of roles* and *separation of concerns* allows to guide which structures are needed. Following those objectives we presented some insights in those architectural structures and processes that helped us to organize and build a number of complex real world robotic software systems [Ulmb].

Our next steps are to further exploit model-driven techniques and enhance our model-driven toolchain in order to provide support to all roles of the development process with our best practices.

Acknowledgements

The research leading to these results has been partially funded by BMBF
(iserveU 01IM12008B and FIONA 01IS13017C).

References

[BEH+11] A. Bjorkelund, L. Edstrom, M. Haage, J. Malec, K. Nilsson, P. Nugues, S.G. Robertz, D. Storkle, A. Blomdell, R. Johansson, M. Linderoth, A. Nilsson, A. Robertsson, A. Stolt, and H. Bruyninckx. On the integration of skilled robot motions for productivity in manufacturing. In *Assembly and Manufacturing (ISAM), 2011 IEEE International Symposium on*, pages 1–9, May 2011.

[CGCG10] S. Cousins, B. Gerkey, K. Conley, and W. Garage. Sharing Software with ROS [ROS Topics]. *IEEE Robotics Automation Magazine*, 17(2):12–14, June 2010.

[Chr89] R. Chris. *Elements of functional programming*. Addison-Wesley Longman Publishing Co, Boston, MA., 1989.

[Dij76] E. Dijkstra. *A Discipline of Programming*. Prentice Hall, Englewood Cliffs, NJ., 1976.

[IRLVCS12] Juan F. Inglés-Romero, Alex Lotz, Cristina Vicente-Chicote, and Christian Schlegel. Dealing with Run-Time Variability in Service Robotics: Towards a DSL for Non-Functional Properties. In Emanuele Menegatti, editor, *3rd Int. Workshop on Domain-Specific Languages and Models for Robotic Systems (DSLRob-12), Proc. of SIMPAR 2012 Workshop*, Tsukuba, Japan, November 2012. http://arxiv.org/pdf/1303.4296.

[Par72] D. L. Parnas. On the Criteria to Be Used in Decomposing Systems into Modules. *Commun. ACM*, 15(12):1053–1058, December 1972.

[RE96] M. Radestock and S. Eisenbach. Coordination in evolving systems. In *Trends in Distributed Systems CORBA and Beyond*, volume 1161 of *Lecture Notes in Computer Science*, pages 162–176. Springer Berlin Heidelberg, 1996.

[SLL+13] Christian Schlegel, Alex Lotz, Matthias Lutz, Dennis Stampfer, Juan F. Inglés-Romero, and Cristina Vicente-Chicote. Model-Driven Software Systems Engineering in Robotics: Covering the Complete Life-Cycle of a Robot. In *Informatik 2013, Workshop Roboter-Kontrollarchitekturen*. Springer LNI der GI, Koblenz, September 2013.

[SSL12a] Christian Schlegel, Andreas Steck, and Alex Lotz. Model-Driven Software Development in Robotics: Communication Patterns as Key for a Robotics Component Model. In Daisuke Chugo and Sho Yokota, editors, *Introduction to Modern Robotics*. iConcept Press Ltd., 2012. Available from: http://www.iconceptpress.com/download/paper/101215045543.pdf.

[SSL12b] Christian Schlegel, Andreas Steck, and Alex Lotz. Robotic Software Systems: From Code-Driven to Model-Driven Software Development. In Dr. Ashish Dutta, editor, *Robotic Systems - Applications, Control and Programming*, chapter 23. InTech, 2012. Available from: http://www.intechopen.com/books/robotic-systems-applications-control-and-programming/robotic-software-systems-from-code-driven-to-model-driven-software-development.

[Ulma] Service Robotics Ulm. Website. http://www.servicerobotik-ulm.de.

[Ulmb] Service Robotics Ulm. Youtube Channel. http://youtube.com/user/RoboticsAtHsUlm.

[VKB14] Dominick Vanthienen, Markus Klotzbücher, and Herman Bruyninckx. The 5C-based architectural Composition Pattern: lessons learned from re-developing the iTaSC framework for constraint-based robot programming. *Jorunal of Software Engineering for Robotics (JOSER)*, 5(1):17–35, May 2014.

Seamless Extension of a Robot Control Framework to Bare Metal Embedded Nodes

Steffen Schütz, Max Reichardt, Michael Arndt, Karsten Berns

Robotics Research Lab, Department of Computer Science
University of Kaiserslautern
Gottlieb-Daimler-Straße
67663 Kaiserslautern, Germany
{schuetz, reichardt, arndt, berns}@cs.uni-kl.de

Abstract: Engineering complex robot control systems typically suffers from a major break in development process between different platforms – powerful computers running an operating systems and bare metal embedded computing nodes. This work presents an approach to close this gap by running an existing state-of-the-art software framework with only minor modifications on all platforms. In this context, we discuss relevant design principles of robotic frameworks and validate their importance.

1 Motivation

Robotic software frameworks have major impact on development effort and quality of robot control systems. Complex systems typically involve embedded hardware in order to interact with the magnitude of sensors and actuators. It is often advantageous to shift low level tasks to dedicated hardware – which has the potential to significantly reduce energy consumption, weight, and size. These are critical factors for many robotic systems.

A lot of research on robotic frameworks targets powerful computing platforms in order to achieve the many relevant quality attributes in robotics systems [RFB13a]. Framework support for small embedded nodes is, however, limited and there is typically a big gap in the development process. Developers are often only familiar with one of the two *worlds*.

As current compilers mostly support embedded architectures, the software development process can be more unified. Obviously, there is a break where interaction with dedicated hardware is required. Hence, the presented work targets at extending the FINROC framework[1] [RFB13b] to support *bare metal* (no operating system) embedded processors – in the case of this work a soft core run in an FPGA.

Overall, to better the situation, we pursue to achieve the following goals:

- Provide similar software development process
- Retain as many quality attributes as possible

[1] http://finroc.org/

- Achieve transparent integration – standard FINROC applications as well as tooling
- Light-weight implementation performing with low cycle times (in the range of kHz)
- Allow component reuse across the two *worlds*

In this paper, we present an approach to reach these goals. This is achieved by focusing on efficient HW/SW-codesign. FPGAs provide the necessary flexibility in this respect.

2 Related Work

Overall, we are not aware of any component-based software framework for complex robot control systems that can also be executed *unmodified* bare metal on embedded platforms.

The closest are model-based approaches in solutions such as OpenRTM-aist [ASK08] or SmartSoft [SHLS09]. Platform-independent models of components and applications can be transformed to platform-specific executables. All target platforms include an operating system. Notably, SmartSoft has been used with an 8-bit microcontroller (ATmega128) with only 128 kB of flash memory and 4 kB of SRAM running the FreeRTOS[2] operating system [SHLS09]. Model-based approaches typically require a complex toolchain.

The Player Project is a well-known, discontinued framework for Linux with a "small and fast" implementation [GVH03] – mostly in C++. It also runs on embedded Linux systems. Without device drivers, the binary size is around 64 kB.

Other solutions were specifically designed for embedded platforms. Unity [LSS13] is a notable open source robotics framework for FPGAs (only), including "HW designs, gateware (GW, VHDL) and SW libraries" and model-based code generation for integration with "a high-level software framework" running on a PC via Unity-Link.

Hartmann et al. [HSM13] propose a framework for a CPU with an FPGA coprocessor with a focus on dynamic reconfiguration. This includes automatically distributing *atomic algorithms* "to the CPU and FPGA to ensure the best runtime with the current set of algorithms". Atomic algorithms are implemented inside "Basic control units" – connectible building blocks that run on both architectures. It is still in an early stage of development.

Robot Makers GmbH developed FINROC-lite – "a light-weight control framework specially designed for embedded systems" [HHLF14]. It is an independent implementation in C++ suitable for FPGA soft cores – currently without a component model. It has sophisticated integration with FINROC.

PEIS-ecologies [SB05] envision to create ambient intelligence using connected, highly heterogeneous [SBG+08] devices, ranging from simple radio frequency tags to PCs. The architecture deals with this heterogeneity by keeping the kernel highly portable. According to [SBG+08], the kernel has been tested "on hardware platforms ranging from multi-core 64bit processors down to 8-bit microcontrollers, and on software environments such as Linux, MacOS, Windows, TinyOS, OpenR and embedded Linux" [SBG+08, III.].

[2] `http://www.freertos.org/`

3 Conceptual Approach

3.1 Software Framework – Design Considerations

In order to be suitable for running bare metal on embedded platforms, a software framework must not depend on any functionality provided by an operating system – such as a file system, processes or even threads. Cross-compiling and static linking must be possible. Furthermore, it is highly advantageous if there are no dependencies on external libraries (e.g. boost), as they often depend on operating system functionality and would need to be cross-compiled as well. Resource consumption is critical, too. In particular, computational overhead should be minimal. Memory consumption and the size of compiled binaries is also relevant. The most efficient way of exchanging data between different processing units on an FPGA is typically via some kind of shared memory. Thus, it is advantageous if a framework's ports can be set up to store their current data at specific memory locations. From a broader view, all these requirements are closely related to the *portability* of robotics software – an important quality attribute. A framework suitable for bare metal embedded platforms is arguably highly portable.

A central question when creating a bare metal variant of a software framework is, whether to use the same code base as the full PC version or whether to create an independent (re)implementation. This decision has an impact on development and maintenance effort – as well as efficiency. When using the same code base, it is less effort to keep both variants aligned and behaving in the same way. However, the code will inevitably contain `#ifdef` preprocessor directives and care needs to be taken that code changes do not break other framework configurations. With a reimplementation, arguably more optimized realizations are possible and likely – at least if the PC version is not to be cluttered with `#ifdefs`.

With most frameworks for robotics used on PCs, it is difficult to reach all the requirements for embedded platforms mentioned above. Thus, separate light-weight solutions are typically created. Notably, due to missing C++11 support for the Altera NIOS®II soft core until only recently, FINROC was not suitable for our hardware platform either.

Evaluating all these factors, we opted to adapt the existing FINROC framework [RFB13b] that we use and develop for implementation of complex robot control systems on embedded PCs to also run without an operating system. We believe that FINROC has some key features that make it particularly suitable for this project:

- A highly modular framework implementation: Framework functionality can be tailored to application and platform requirements (see [RFB13b] and Fig. 4).
- Relevant features do not depend on external libraries, middleware, or operating system functionality
- Multiple components can be instantiated in the same executable
- An efficient implementation

Notably, we see FINROC as a proof-of-concept for the proposed approach. Any other framework with these properties should be suitable as well.

3.2 FPGA-based Embedded Nodes

To provide a high degree of flexibility on the embedded side, the targeted embedded nodes are FPGA-based. Using the elaborated toolchains provided by the FPGA-manufacturers, it has become feasible to implement complex systems incorporating multiple soft cores that can be extended by custom IP-cores. Thereby it has become feasible to implement high-performance dedicated embedded systems.

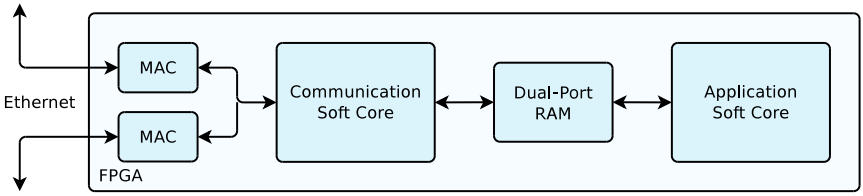


Figure 1: Concept for the FPGA SoC

In order to execute the FINROC framework with good performance, the underlying design approach is to distribute application and communication on two separate soft cores (Fig. 1). The communication soft core interfaces the Ethernet MAC, handles incoming Ethernet frames, and executes the handling of the higher-level communication protocol (see 3.3). For inter-processor communication, the most efficient mechanism is shared memory. Most modern FPGAs feature dual-ported RAM entities allowing simultaneous read/write access by two masters, which is especially suitable for this scenario.

The more detailed design of the system should be guided by two requirements: On *communication side*, to keep bus cycle times low, packets must be handled efficiently. On *application side*, the overhead due to the handling of process data must be kept minimal.

By following a consequent HW/SW co-design approach, a system can be designed that provides the required infrastructure for the performant deployment of FINROC.

3.3 Communication Protocol

The protocol to communicate with the embedded nodes is driven by several requirements:

- Provide a high bandwidth
- Use of standard hardware
- Light-weight regarding communication and computation
- Flexible network topology
- Reliable data exchange
- Tailored to the framework communication

- Support dynamic initialization during run-time
- Programming of embedded nodes via the bus system

The following design decisions emerged. For physical and data link layer of the OSI model, Ethernet is the most suitable solution – hardware is widely available, it is well standardized, and offers bandwidths in the gigabit range. Moreover, using Ethernet on the data link layer allows a flexible network topology. If the embedded nodes feature two Ethernet interfaces and support frame forwarding, any mixture between star and line topologies can be realized by using standard layer two switches. On the network and transport layer of the OSI model, UDP/IP is chosen. Thus, on PC side, the protocol can be executed from user space using UDP sockets. Further on, both – IP as well as UDP – include a checksum to test frame integrity. While TCP guarantees reliable, ordered, and error-checked streams of packets on the transport layer, it causes overhead in the form of additional protocol logic, as well as additional bus load. This makes it unsuitable for systems where computational resources are limited. Package loss, fragmentation, and reordering must be handled at a higher layer – if needed.

An alternative would have been to use one of the industrial Ethernet-based fieldbus systems like EtherCAT[3] or Ethernet Powerlink[4]. Not building on top of a rather strictly defined fieldbus system allows for the protocol to be tailored to the framework communication and the requirements of the typical application scenario. Additional features, such as the dynamic initialization during run-time or the possibility to program the embedded nodes via the bus, are easy to realize. This of course comes at the cost of increased latency and jitter – hard real-time cannot be guaranteed running the bus master on top of a UDP socket on a standard OS.

The *Framework-integrated Embedded Protocol* (FINEMBP) is located at the session layer of the OSI reference model. Inspired by Ethernet Powerlink, FINEMBP separates the communication into cycles. As illustrated in Fig. 2, the cycle is split into an *isochronous*, an *asynchronous*, and an optional *idle* phase.

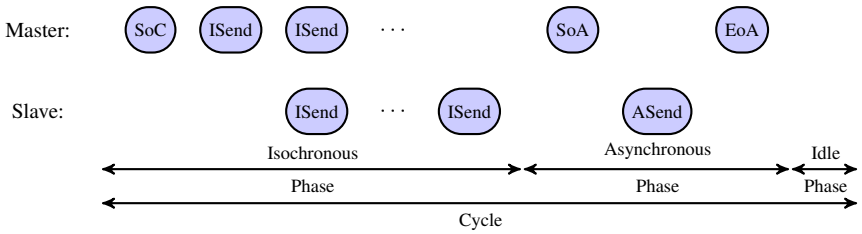


Figure 2: Basic segmentation of a full cycle in FinEmbP.

The *SoC*, *SoA* and *EoA* frames structure the FINEMBP cycle. They have to be received by every node and therefore are sent as a broadcast/multicast. In the isochronous phase,

[3] http://www.ethercat.org/
[4] http://www.ethernet-powerlink.org/

cyclic process data is exchanged by a poll mechanism. The *ISend* packets, sent by the master, contain the process data bound for the corresponding node, while the *ISend* response contains the most recent process data, produced by the node. The master sends an *ISend* packet to each of the nodes containing process data. During the asynchronous phase the master can either send or receive any data that does not require cyclic transmission. The asynchronous phase can be used – e.g. for network initialization and data that does not require cyclic transmission.

The underlying state machine of the FINEMBP protocol is shown in Fig. 3. Each state encapsulates a master/node-specific sub state machine.

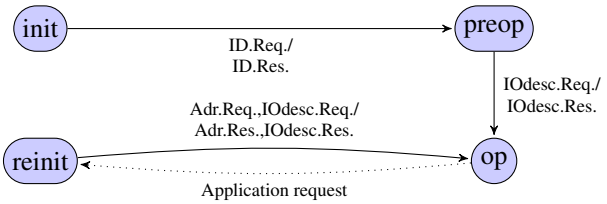


Figure 3: Basic states machine for FINEMBP master and nodes

In the *init* state the master checks the network for FINEMBP nodes. If one or more nodes are available, the master assigns each of them an IP address and thereby manipulates OSI layer 3. Once all available nodes got assigned an IP, the master and all nodes are in the *preop* state. In the *preop* state, the master queries each node for its structure information. The structure information includes all information describing the FINROC application deployed on the node. After the structure information of all nodes has been exchanged, the system is in the *op* state. From now on, cyclic data is exchanged as described above. With the initialization mechanism being allocated in the asynchronous phase of the FINEMBP cycle, additional nodes can be initialized while the network is running.

4 Implementation Details

4.1 FINROC

As introduced in [RFB13b], FINROC is implemented in a highly modular way – composed from slim software entities with a specific concern. By selecting a suitable set of these entities, it can be tailored to application and platform requirements. For the bare metal version of FINROC, we targeted a minimal configuration for executing the common FINROC component types – defined in the *structure* and *ib2c* plugins. The configuration is illustrated in Fig. 4[5].

Although, we generally try to minimize use of the C++ preprocessor for maintenance rea-

[5] RRLIBs are framework-independent libraries, whereas plugins contain optional functionality for the FINROC core. Lines of code were counted using David A. Wheeler's 'SLOCCount'.

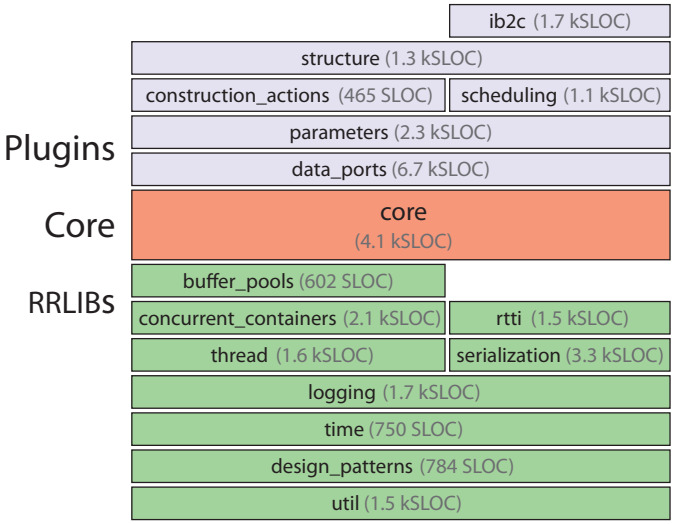


Figure 4: FINROC's highly modular core in the configuration running bare metal on the FPGA soft core

sons, there are futhermore few `#defines` that can be set for specific target platforms – most importantly `RRLIB_SINGLE_THREADED` which will compile FINROC in single-threaded mode with e.g. a major performance increase for data ports. Furthermore, pre-processing directives can check whether certain software entities (RRLIBs and plugins) are present[6].

Without a network protocol, the default FINROC configuration only depends on the platform-independent *libxml2* library[7]. It is mainly used for processing configuration files in the *parameters* plugin and *rrlib_logging*, as well as serializing application structure in the *runtime_construction* plugin. As this is not relevant for FINROC's bare metal version, the dependency is now optional – availability of XML-related functionality depending on availability of *libxml2*.

Notably, there is a single-threaded variant of *rrlib_thread*. This was preferred to removing it from the configuration, as the latter would require cluttering the code with `#ifdefs`. If `RRLIB_SINGLE_THREADED` is defined, many classes such as mutexes and locks are empty, resulting in nops wherever they are used. Thread objects still exist and could be used for basic handling of virtual threads. The RRLIBs *concurrent_containers* and *buffer_pools* employ policy-based design [Ale01] – e.g. using different implementations (policies) depending on the number of producers and consumers. This always includes a single-threaded variant.

Similarly, FINROC's data ports use different implementations depending on the type of data they transfer (more precisely, the type traits of *T*). This is illustrated in Fig. 5. For

[6] FINROC's build system creates respective `#defines`
[7] `http://xmlsoft.org/`

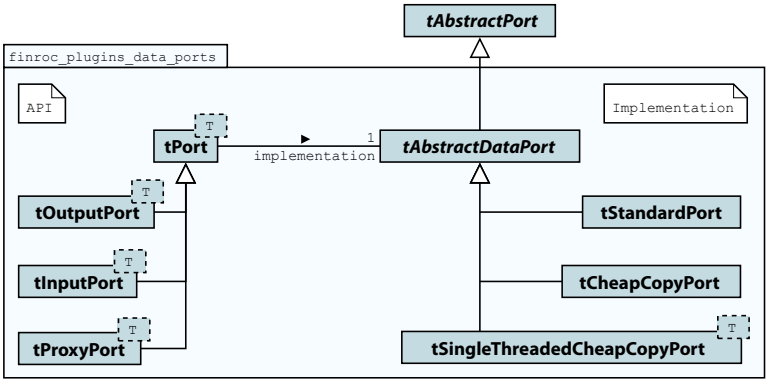


Figure 5: Port classes in FINROC's *data_ports* plugin (simplified)

large data types such as images that are expensive to copy, a zero-copy implementation with buffer pools and reference counting is used [RFB13b] – *tStandardPort*. It is also used for types with internal memory management such as *std::string* – as memory allocation is to be avoided in real-time applications and also relatively expensive. For cheaply copied types such as scalars and vectors, there is an optimized implementation using thread-local buffer pools – *tCheapCopyPort*. In single-threaded mode, however, there is no necessity for buffer pools at all. Instead, a port's current value can simply be stored at a fixed location in memory, as there is no concurrent access. Thus, a simple and very different single-threaded implementation for cheaply-copied types was created. The implementation allows to set the memory location the port data is stored to – possibly shared memory that processing units on an FPGA use to exchange data.

In theory, connected *tStandardPort*s could operate on the same data buffer. However, e.g. the smart pointer classes in the API allow to reference potentially multiple buffers of the same port. As copying is not a viable option for complex types, the standard implementation is used on single-threaded platforms, too.

An issue during compilation was that the *libstdc++* shipped with Altera's gcc 4.7 compiler does not fully support the C++11 standard – e.g. *std::chrono* is missing. Furthermore, some functions defined in headers cannot be found when linking. Thus, we added simple implementations of the relevant missing parts to a small target-platform-specific library. This was added to the target platform's compiler and linker flags. In future compiler versions this will likely be obsolete.

Overall, using a highly modular framework core combined with techniques such as policy-based design, the bare metal version of FINROC could be created adding only a limited amount of `#ifdef`s (no nested ones in particular).

4.2 FPGA HW/SW-System

The system outlined in 3.2 has been implemented on the eSM Board developed and produced by elrest Automationssysteme GmbH. At the core of the board, there is an Altera[8] Cyclone IV FPGA. Due to the external peripherals connected to the FPGA, the board is designed to optimally support the deployment of two soft cores. As it also features two Ethernet interfaces, it is suitable for the use in networks with line topology.

A detailed illustration of the implemented logic system in the FPGA is provided in Fig. 6. Only the components and bus connections that are relevant for the data flow of incoming Ethernet frames are illustrated.

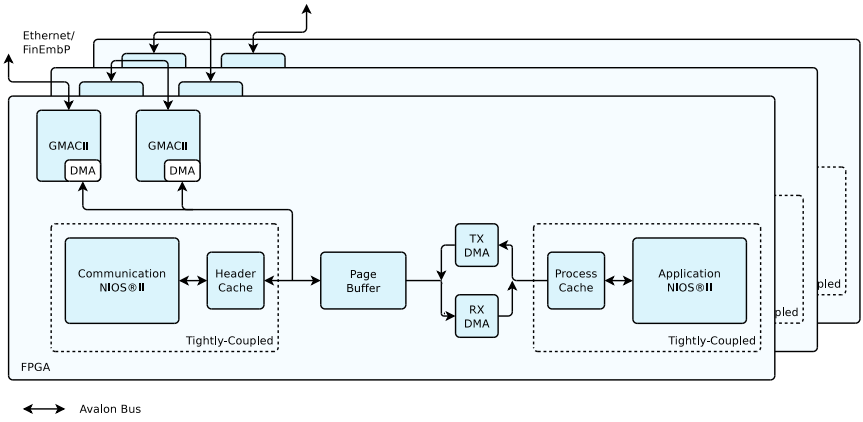


Figure 6: Detailed illustration of the FPGA system.

The implemented soft core is the NIOS®II provided by Altera. Ethernet medium access is managed by two GMACIIs, a high performance Gigabit Ethernet MAC developed by IFI[9], which are interfaced by the *Communication NIOS®II*. A prominent feature of the GMACII is the integrated DMA, which is alignment-aware, byte-exact and provides on-the-fly checksum calculation. Both soft cores have dual-port RAM attached as tightly-coupled memory, which allows for fast one-cycle read/write access.

On communication side, the tightly-coupled memory is mainly used for header caching, while on application side, the process data is located in the memory. The two processors are coupled through a third dual-port RAM, the *Page Buffer*. Besides holding a structure for the inter-processor communication, once the node is initialized, it serves for the exchange of coherent pages of process data. Therefore, the memory is segmented into several areas, buffering the different process data pages. Using two pages for cyclic process data and some page flipping logic, delays in case of concurrent access from application and communication side can be avoided.

On the application side the data is transferred to and from the *Page Buffer* using the

[8] http://www.altera.com/
[9] http://www.ifi-pld.de/

GMACII-integrated DMAs. The application system is extended by two DMAs as well – the *TX DMA* for data transfer from the *Process Cache* to the *Page Buffer*, the *RX DMA* vice versa.

Software-wise, the *Communication NIOS® II* runs a UDP stack handling layer two, three, and four of the OSI model. To allow for Ethernet line topologies, the stack features frame forwarding on OSI layer two. The implementation of the UDP stack makes extensive use of the *GMACII-DMA*s for fast frame handling and checksum calculation. On top of the UDP stack runs the client-FSM handling the FINEMBP protocol and the inter-processor communication. A new Ethernet frame arriving at one of the Ethernet interfaces is, depending on the the frame and the state of the FINEMBP FSM, either forwarded to the second Ethernet interface, copied to the *Page Buffer*, or both. If data is copied to the *Page Buffer*, this will be indicated to the *Application NIOS® II* by setting the according flag. Again, depending on the context, the FINEMBP client-FSM might respond to the incoming frame by sending an adequate frame. In most cases, the response contains process data from the *Page Buffer*. On application side, at the beginning of a FINROC cycle, the *Page Buffer* is checked for new incoming process data. If data is available, it is copied to the *Process Cache* using the *RX-DMA*. As mentioned above, the FINROC ports are mapped to constant memory locations, located in the *Process Cache*. Thus, this two-staged buffer setup relieves the *Application NIOS® II* of the need for any pointer arithmetic. At the end of the FINROC cycle, outgoing data is transferred to the *Page Buffer* via the *TX-DMA*.

By handling data transfers with the DMAs and avoiding pointer arithmetic due to the two-staged buffer setup, the resulting overhead for the *Application NIOS® II* is kept minimal.

4.3 Build System

The build system used to build FINROC needed only minor modifications. So-called *targets* contain definitions for the compiler and compile flags to use. For this work, the NIOS® II compiler provided by Altera is specified and some soft-core-specific compiler/linker flags are added. Two modifications to the build system itself were needed:

Running the code on a bare metal embedded processor requires static linkage of the final binary. This is in contrast to the usual procedure for running with an operating system. There, dynamic linking is highly advantageous to save space. The build system was thus extended to not only produce shared library objects , but also to produce archives to enable static linkage of the binaries, when requested.

Additionally, the build-process requires a so-called *board support package* containing important definitions of the hardware (including a linker script), prior to the linking process. Also, the binaries that result from compiling and linking need some post-processing of the ELF binaries so that they can be run on the NIOS® II soft core. This led to the extension of the build system by two hooks that can be overridden by targets (called *pre-build-hook* and *post-build-hook*). For this work, the *pre-build* hook builds the board support package and the *post-build* hook post-processes the ELF binaries so that they can directly be uploaded to the embedded soft core.

5 Benchmark

The benchmark for evaluating the performance of the proposed system is running a simple PID closed-loop controller within a FINROC module on the soft core. To make the benchmark as transparent as possible, an existing implementation of a PID controller for Atmel AVR microcontrollers has been used instead of implementing a proprietary solution. The documentation and the implementation of this controller are given in Atmel's application note *AVR221: Discrete PID Controller*[10].

This application note provides an implementation in plain C. The FINROC module is just a wrapper around the C implementation. It instantiates a controller and periodically feeds it with sensor data to obtain the control-signal output. The resulting statically linked FINROC binary has a size of about $2.5\,\mathrm{MB}$. Using this setup, it was possible to run the FINROC framework with an update rate of $25\mathrm{kHz}$. Thus also the control loop of the PID controller runs with this frequency, which is more than enough for most typical closed-loop control applications [ASHO+07]. For comparison, the same (single threaded) FINROC-implementation of this controller runs with an update rate of $3.75\,\mathrm{MHz}$ on an Intel Core i7-920 CPU clocked at $2.67\,\mathrm{GHz}$.

6 Conclusion and Outlook

To conclude, an approach was presented to deploy a state-of-the-art component-based robotics framework on an FPGA soft core. Using a highly modular framework core combined with techniques such as policy-based design, a bare metal version of robotic software frameworks can be created with only minor code changes. Results show negligible computational overhead from using a full-featured framework. In fact, a performance level significantly above the initial expectations and goals of the authors was reached. In particular, closed-loop controls can be executed at frequencies above $10\,\mathrm{kHz}$. However, binary size of several megabytes and the dependency on C++11 is a limiting factor regarding portability to some other embedded architectures. Without an operating system and minimal I/O latency, the developed systems seems to be an ideal platform for the development of real-time applications – without the typical gap in the development process. Furthermore, this facilitates the realization of highly distributed robot control systems with dedicated hardware nodes.

Next steps will be to improve integration with FINROC's tooling and to gain experience from use in projects. If required, further optimizations are possible – e.g. reduction of binary size. Notably, we observed that including certain C++ standard headers have a significant impact in this respect already.

[10] `http://www.atmel.com/webdoc/atmel.docs/atmel.docs.33085.19841.html`

References

[Ale01] Andrei Alexandrescu. *Modern C++ design: generic programming and design patterns applied*. Addison-Wesley Longman Publishing Co., Inc., Boston, USA, 2001.

[ASHO+07] Alin Albu-Schäffer, Sami Haddadin, Christian Ott, Andreas Stemmer, Thomas Wimböck, and Gerd Hirzinger. The DLR lightweight robot: design and control concepts for robots in human environments. *Industrial Robot: An International Journal*, 34(5):376–385, 2007.

[ASK08] Noriaki Ando, Takashi Suehiro, and Tetsuo Kotoku. A Software Platform for Component Based RT-System Development: OpenRTM-Aist. In Stefano Carpin, Itsuki Noda, Enrico Pagello, Monica Reggiani, and Oskar von Stryk, editors, *Simulation, Modeling, and Programming for Autonomous Robots*, volume 5325 of *Lecture Notes in Computer Science*, pages 87–98. Springer Berlin / Heidelberg, 2008.

[GVH03] Brian P. Gerkey, Richard T. Vaughan, and Andrew Howard. The Player/Stage Project: Tools for Multi-Robot and Distributed Sensor Systems. In *11th International Conference on Advanced Robotics (ICAR 2003)*, pages 317–323, Coimbra, Portugal, June 30 - July 3 2003.

[HHLF14] Carsten Hillenbrand, Jochen Hirth, Bernd Helge Leroch, and Martin Frank. Closed-Loop Joint Angle Control for a Multi-Axes Hydraulics Arms - Towards Autonomous Construction Machines. In *Commercial Vehicle Technology 2014 - Proceedings of the Commercial Vehicle Technology Symposium (CVT)*, pages 37–46, Kaiserslautern, Germany, March 11-13 2014.

[HSM13] Jan Hartmann, Walter Stechele, and Erik Maehle. Self-reconfigurable Control Architecture for Complex Robots. In *Informatik 2013*, Lecture Notes in Informatics (LNI), Koblenz, Germany, September 16-20 2013. Springer.

[LSS13] Anders B. Lange, Ulrik P. Schultz, and Anders S. Soerensen. Unity: A Unified Software/Hardware Framework for Rapid Prototyping of Experimental Robot Controllers using FPGAs. In Davide Brugali, editor, *Proc. of the 8th full-day Workshop on Software Development and Integration in Robotics*, Karlsruhe, Germany, May 2013.

[RFB13a] Max Reichardt, Tobias Föhst, and Karsten Berns. Design Principles in Robot Control Frameworks. In *Informatik 2013*, Lecture Notes in Informatics (LNI), Koblenz, Germany, September 16-20 2013. Springer.

[RFB13b] Max Reichardt, Tobias Föhst, and Karsten Berns. On Software Quality-motivated Design of a Real-time Framework for Complex Robot Control Systems. *Electronic Communications of the EASST*, Software Quality and Maintainability(60), August 2013. http://journal.ub.tu-berlin.de/eceasst/article/view/855.

[SB05] A. Saffiotti and M. Broxvall. PEIS Ecologies: Ambient Intelligence meets Autonomous Robotics. In *Proc of the Int Conf on Smart Objects and Ambient Intelligence (sOc-EUSAI)*, pages 275–280, Grenoble, France, 2005.

[SBG+08] A. Saffiotti, M. Broxvall, M. Gritti, K. LeBlanc, R. Lundh, J. Rashid, B.S. Seo, and Y.J. Cho. The PEIS-Ecology Project: Vision and Results. In *Proc of the IEEE/RSJ Int Conf on Intelligent Robots and Systems (IROS)*, pages 2329–2335, Nice, France, September 2008.

[SHLS09] Christian Schlegel, Thomas Haßler, Alex Lotz, and Andreas Steck. Robotic Software Systems: From Code-Driven to Model-Driven Designs. In *14th International Conference on Advanced Robotics (ICAR)*, Munich, Germany, June 22-26 2009.

Increasing ROS Reliability and Safety through Advanced Introspection Capabilities

Andreas Bihlmaier, Heinz Wörn

Institute for Anthropomatics and Robotics (IAR)
Intelligent Process Control and Robotics Lab (IPR)
Karlsruhe Institute of Technology (KIT)
Engler-Bunte-Ring 8
76131 Karlsruhe
andreas.bihlmaier@kit.edu
woern@kit.edu

Abstract: The Robot Operating System (ROS) brought together many different robotics research groups by providing a common software ecosystem. One important part of ROS is the publish-subscribe messaging infrastructure that facilitates building modular and reusable distributed robotics systems. However, the inherent complexity – flexibility, diversity and dynamics – of such a system opposes properties essential to robotics: reliability and safety. The paper details how to gain control of a ROS-based robotics system through the extension of ROS by advanced introspection capabilities. Once the introspection based on distributed gathering of metadata is in place, we are able to define a reference state for the whole ROS system. A monitoring node continuously compares the reference to the actual state. In case of a deviation the operator is alerted through a dashboard, in addition appropriate countermeasures can be initiated by the robotics system itself.

1 Introduction

The Robot Operating System (ROS) can be seen as consisting of a middleware, generic supporting tools, robotics specific libraries and a large collection of robotics software based on the former. Our focus is on the first two components: ROS as middleware together with its supporting tools. Without recounting the middleware architecture [QCG+09], it is important to recall the general components. Overall, the ROS middleware is a publish/subscribe architecture [TvS07]. The ROS master provides the naming and registration services for all ROS nodes. If node A publishes a topic T, it registers T with the master. When node B wants to subscribe to T, it queries the master for the network address of T and contacts A at that address. Afterwards the data is directly sent between A and B, without further involvement of the master.

This architecture allows dynamic addition and removal of nodes, flexible communication patterns and modular robotics software design. However, it can also become difficult to track down problems in such a complex system. The complexity is due to the high

diversity and dynamics of the system. There are many different nodes communicating on many different topics with communicating patterns varying during runtime – e.g. involved nodes, numbers of publishers and subscribers, frequency and latency. Fortunately, a utility visualizing the current ROS graph is part of the ROS support tools. It shows which nodes exist, the connections between them and the message type of each topic. Unfortunately, there are two important shortcomings: First, it is not possible to get information about the data, if any, which is sent between the nodes. Second, there is no tool available, which allows the definition a good reference state of the ROS (sub)graph that *should* be the case and check it against the actual state. We will show in the remainder of the paper, how these shortcomings can be overcome and how this can benefit reliability and safety of ROS-based robotic systems.

The following section (2) goes through all the components that are part of our proposal for advanced ROS introspection capabilities. In order to put these capabilities in concrete terms, we present a case study on a robotic surgery research platform in section 3, which is based on ROS. Section 4 concludes the paper with a discussion about advantages provided by advanced introspection and the obstacles that have to be overcome in order to utilize it.

2 Advanced ROS Introspection

Our proposed extension to existing ROS introspection features (see Fig.1) consists of four elements: First, distributed measurement and publication of metadata about a node's communication. Second, a format to define reference states of a ROS subgraph in a manner that preserves runtime flexibility. Third, a monitoring node that collects all metadata, compares it to the reference and runs predefined actions in case of a deviation. Fourth, a rqt GUI to provide clear visual feedback about the robot systems state, utilizing the output of the monitoring node. Each component will be detailed in the remainder of this section.

2.1 Acquiring Metadata

The data transport of ROS, usually TCPROS, already provides all required metadata on receiving or sending a message either in its header fields or through the sockets API. Utilized network bandwidth, message frequency, jitter and even latency can be acquired for many messages with negligible computation. Furthermore, additional metadata can be generated, such as the correlation between incoming and outgoing messages. These correlations support the reasoning about the common ROS pattern of publishing messages after doing computation on a received message. By way of specifying a new ROS msg type, all these measures can be published for each topic using readily available ROS components. Since only a few floating point numbers are required for the metadata of each topic, the additional network traffic comprises only a small fraction of the data traffic.[1]

[1] In future versions of ROS, based on recent work by [FHM14], it will be possible to collect this metadata, as */statistics*, for all C++ and Python nodes without any modification.

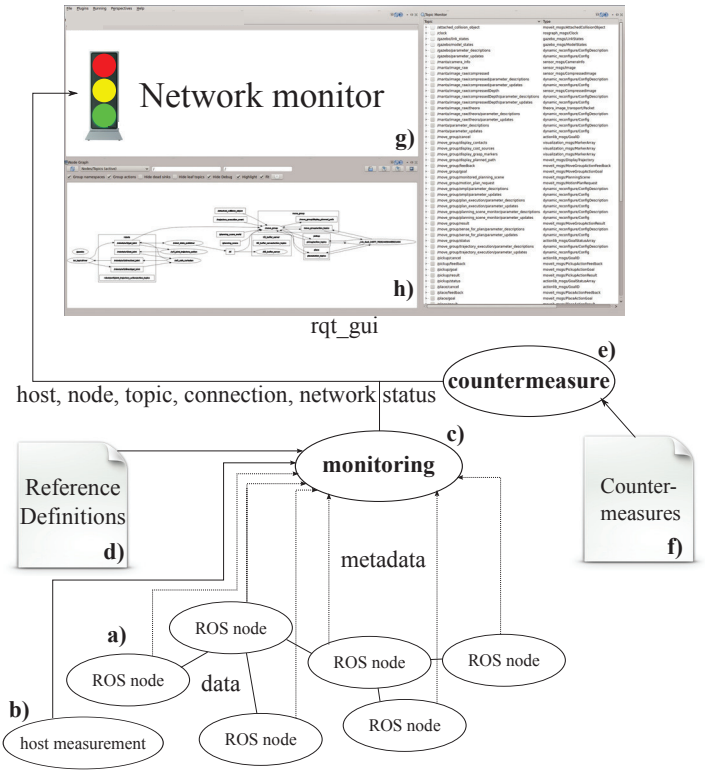


Figure 1: Overview of the advanced introspection capabilities for ROS: a) Gathering and publishing of metadata; b) Nodes to gather host specific data; c) Monitoring node to compare actual metadata against d) reference definitions; e) Node to automatically execute countermeasures defined in f); g) rqt_gui plugin to inspect host, node, topic and connection metadata and to alert about errors; h) rqt_graph extended with metadata information.

In addition to acquiring metadata about message passing traffic, it is also beneficial to measure host and node related features. Such features comprise – per host and per node – CPU, memory and raw network utilization as well as hardware health properties, e.g. CPU and GPU temperatures. The current introspection and monitoring features of ROS for these properties (the rqt process monitor plugin) provide only limited access to them and just for a single host. Given the fact that ROS-based systems are usually distributed across multiple machines, taken together with the flexibility to decide on which host a node runs during startup, operating system level tools regularly prove themselves as insufficient.

2.2 Defining network reference states

The format for reference definitions must allow their definition in a modular and orthogonal way. As a result it is possible to define constraints for different parts of the ROS graph and different – not contradicting – constraints for the same set of nodes. All attributes must support wildcards or value ranges in order to keep flexibility while attaining reliability through monitoring and self-diagnosis. In contrast to other approaches [ZSM+13][MWV14], we propose a simple manual or at most semi-automatic approach. Because automatic approaches are – at least for a research setting – either too sensitive, i.e. have far too many false-positive alarms, or they require too much fine tuning effort and then only work for a very specific setup that does not last long. Therefore we manually define multiple partial constraints that are known to be *necessary* conditions for a working system.

The case study (section 3) will exemplify how the above considerations can be applied to actual systems. Yet, there are also some monitoring rules that can be reasonably applied to most systems. They largely relate to hosts and nodes seen as processes running on a certain host. Often memory and network utilization should stay below some threshold to avoid swapping for the former and package loss for the latter resource. Otherwise both can result in sudden, highly non-linear performance degradation. For CPU (and GPU) utilization, we have to distinguish two cases: On the one hand, there are nodes that consume only a fixed amount of computational resources in the absence of errors. On the other hand, some nodes use up all available computation power to generate as many results per time unit as possible. For the limited consumers, we can define a reference state as usual. For the unlimited case, one can argue that it is ill advised in the first place to have components of unknown performance in a robotic system at all. Nevertheless, it can still make sense to at least define a lower bound for the activity and result frequency of such a node and define falling below this threshold as a system error.

2.3 Monitoring through Metadata

As noted previously, due to the low bandwidth and computation demands, it is possible to collect the *meta*data of all ROS nodes at a central node, even for a large system. This central monitoring node is responsible for comparing the metadata against all constraints given by the reference state. Again, using a custom ROS message type, the result of the comparison is published for each connection, topic, node and host computer. The data can be either used for a robot dashboard (cf. the following subsection) or reacted to by another node in order to increase reliability or at least safety. For example, if the latency to the node for some robot's main environment sensor, such as a RGBD camera, significantly increases or the node completely fails to deliver data, an emergency stop could be initiated instead of the robot crashing into an obstacle. This does not solve the notorious "ROS is not a realtime framework" problem, but at least a higher level of fail-safety can be guaranteed, in case time constraints are violated.

The definition of countermeasures is based on the evaluated metadata (cf. Fig.1). Each metadata property is classified by the monitoring node as either unknown, valid, too-low or too-high. A countermeasure action consists of a logical conditional that is matched against these states and an action that is performed when a match occures.

2.4 Robot status dashboard

Even when running a moderate number of servers, it is good practice to use IT infrastructure monitoring solutions, such as the well-known software Nagios [Jos13]. These software packages enable an administrator to know that a problem has arisen and, more important, what kind of problem it is and where it occurred. Instead of "something is not working right", the monitoring software at least points out "component A, D and X failed" – usually through some kind of dashboard GUI. Our monitoring node allows to transfer this state of affairs from IT infrastructure to distributed robotics systems. Within the ROS ecosystem the adequate implementation of such a dashboard is a rqt[2] GUI plugin. The monitoring plugin provides an overview of the general system state and aggregated access to the metadata. In order to remain usable for larger systems and at the same time provide all details required to track down the source of an error, the system is presented in hierarchical groups. This way multiple hosts, the nodes running on each of these hosts, the various published and subscribed topics of each node and the individual connections making up each topic can be inspected.

3 Case study: Robotic Surgery

The platform to evaluate the advanced ROS Introspection capabilities is a modular robotic system for minimally-invasive surgery (MIS) [NBK+13]. Automated endoscopic camera guidance [BW14] is the particular application targeted in the case study.

3.1 System Overview

The core components for the camera guidance task are as follows (Fig.2): On the perception side we either use an industrial camera with an Ethernet interface or a dedicated endoscope camera through a frame grabber. The system's action is achieved by either – in order of decreasing degrees of freedom – a KUKA LWR4+, a Universal Robots UR5 or a TRUMPF ViKY motorized endoscope holder. However, these are not the only active ROS components in the target scenario, where one robot guides the endoscope and two other robots, equipped with MIS instruments, are telemanipulated. In addition, our robotic surgery platform contains a variety of data-intensive sensors monitoring the OR room.

[2] http://wiki.ros.org/rqt

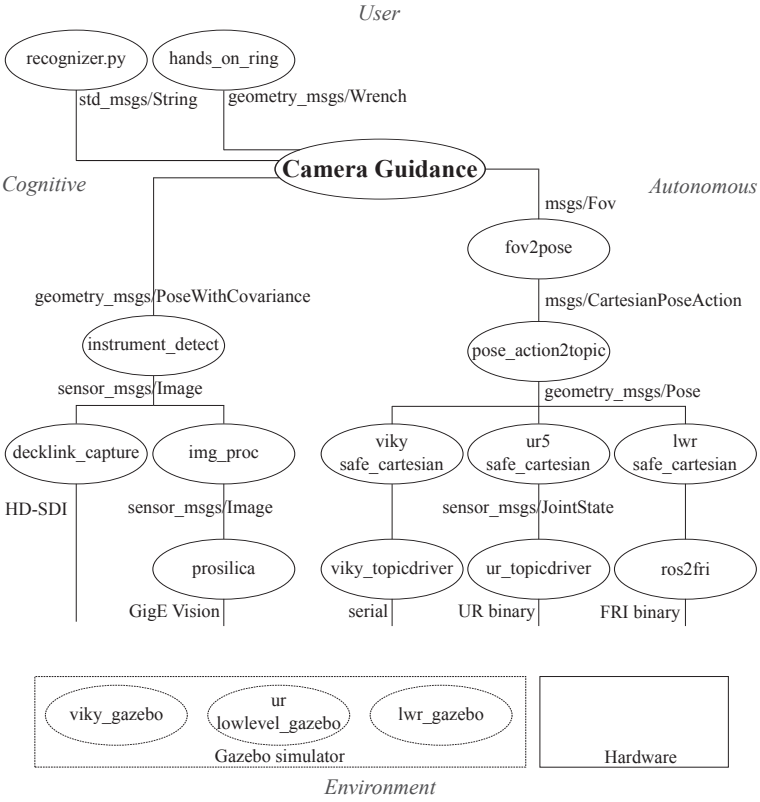


Figure 2: This figure shows part of the ROS graph that makes up the infrastructure for a robotic automated endoscopic camera guidance system (cf. [BW14]).

Given the metadata for all hosts, nodes, topics and connections is published, the first step is to define a number of constraints on the valid state of each ROS component. To illustrate the utilization of the advanced introspection capabilities, we define a reference state for the higher-level camera sensor nodes (see left part of Fig.2) and the lower-level actuator nodes (see right-hand side). In top to bottom direction, we first define that the instrument_detect node should send at least 25 Pose messages per second. Also, we define a maximum delay and jitter value for the connection to the camera guidance node. Going further down in the graph, we define a reference state for the Image carrying topic. Here we have two possibilities: First, to define a different, specific reference for each image source. Second, to define a common, less rigorous reference that is valid for both image sources. In our case one camera provides images at exactly 30 frames per second (FPS) and the other varies between 50 and 60 FPS. Taking into account the big picture of the endoscope guidance system, the important property is to have at least 25 Poses per second. Therefore it is enough to monitor for the same number of images per second. Without going into more detail for the actuator nodes, we want to point out that, based on the same considerations,

we either define a generic rule that holds for all the actuators in use or is actuator specific. By taking advantage of ROS launch files, both cases are easy to deal with. The reference constraints are uploaded from the launch file to the monitoring node via the parameter server.

3.2 Reliability and Safety

Reliability and safety issues do not only occur in surgical robotics, but they become quite obvious when a robot has to control a 30 cm long steel shaft within the patient during surgery. It is, of course, not enough to depend on the ROS systems self-diagnosis in order to guarantee safety. Yet, just as in IT infrastructure monitoring, it is an essential requirement that the operator of the system does not have to scrutinize all parts of the system in order to find out what the problem is. Rather it is commendable that the system points him right at it.

4 Discussion and Outlook

We have detailed why advanced introspection capabilities are important for the reliability and safety of robotics systems based on a distributed architecture such as ROS. Our account then showed how this can be achieved by the generation and analysis of per node metadata, also providing a ROS implementation. One open question concerns the amount of effort required to write down reference definitions that detect anomalies properly, i.e. have a high accuracy, but are general enough not to impair on ROS flexibility. This and further questions related to the real world practicability of self-diagnosis through advanced introspection, are part of our ongoing research into surgical robotics and ROS.

Acknowledgment

This research was carried out with the support of the German Research Foundation (DFG) within project I05, SFB/TRR 125 “Cognition-Guided Surgery”.

References

[BW14] Andreas Bihlmaier and Heinz Wörn. Automated Endoscopic Camera Guidance: A Knowledge-Based System towards Robot Assisted Surgery. In *Proceedings for the Joint Conference of ISR 2014 (45th International Symposium on Robotics) Und ROBOTIK 2014 (8th German Conference on Robotics)*, pages 617–622, 2014.

[FHM14] Dariush Forouher, Jan Hartmann, and Erik Maehle. Data Flow Analysis in ROS. In

Proceedings for the Joint Conference of ISR 2014 (45th International Symposium on Robotics) Und ROBOTIK 2014 (8th German Conference on Robotics), pages 643–648, 2014.

[Jos13] David Josephsen. *Nagios: Building Enterprise-Grade Monitoring Infrastructures for Systems and Networks*. Prentice Hall Press, Upper Saddle River, NJ, USA, 2 edition, 2013.

[MWV14] Valiallah Monajjemi, Jens Wawerla, and Richard Vaughan. Drums: A Middleware-Aware Distributed Robot Monitoring System. In *Computer and Robot Vision (CRV), 2014 Canadian Conference on*, pages 211–218, May 2014.

[NBK+13] P. Nicolai, T. Brennecke, M. Kunze, L. Schreiter, T. Beyl, Y. Zhang, J. Mintenbeck, J. Raczkowsky, and H. Wörn. The OP:Sense surgical robotics platform: first feasibility studies and current research. *International Journal of Computer Assisted Radiology and Surgery*, 1(8):136–137, 2013.

[QCG+09] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source Robot Operating System. In *ICRA Workshop On Open Source Software*, volume 3, 2009.

[TvS07] Andrew S. Tanenbaum and Maarten van Steen. *Distributed systems*. Pearson, Prentice Hall, Upper Saddle River, NJ, 2 edition, 2007.

[ZSM+13] S. Zaman, G. Steinbauer, J. Maurer, P. Lepej, and S. Uran. An integrated model-based diagnosis and repair architecture for ROS-based robot systems. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 482–489, May 2013.

Health Signal Generation in the ORCA (Organic Robot Control Architecture) Framework

Raphael Maas and Erik Maehle

Institute of Computer Engineering
University of Lübeck
Ratzeburger Allee 160
D-23562 Lübeck, Germany
maehle@iti.uni-luebeck.de
maas@iti.uni-luebeck.de

Abstract: This paper evaluates various information-theoretical methods for the generation of health signals in the Organic Robot Control Architecture (ORCA). Also a new method based on the joint probability distribution of signals is proposed. The results show that the new method is superior to the known ones for the studied worst-case artificial and real scenarios. As an example for the application of health signals in ORCA, an adaptable path planner is described which considers the health status of a partially damaged robot.

1 Introduction

Autonomous mobile robots are becoming more and more complex in order to be able to execute more and more challenging tasks. Therefore, they also need sophisticated control systems which are able to master their complexity. With the complexity of a system the probability that it becomes faulty also increases. For mobile robots there are not only internal faults but also interaction faults with its environment. To ensure a high reliability and availability these faults have to be handled by the control system for achieving fault tolerance. Classical fault tolerance techniques which are mostly based on some kind of replication are, however, not suitable, since they require a considerable amount of additional redundant components which mean higher cost and energy consumption. Instead the natural redundancy of the system should be exploited and a fail-soft behaviour achieved.

Natural organisms show a high degree of fault tolerance in order to be able to survive. An example is the immune system which allows to checking the body for infections and reacts by destroying the intruders to restore health. Often self-organisation is used in organisms to allow a flexible response to abnormal situations. Organic computing is an approach to make these properties also available to complex embedded computing systems.

ORCA (Organic Robot Control Architecture) is a software framework developed at the University of Lübeck which supports the implementation of organic computing

principles in particular for robotic systems. In ORCA so-called health signals are used to represent the health status of the robot. If the health signal drops, the control system reacts by organizing itself to continue the robot's mission in the best still possible way.

Thus, the generation of such health signals, which is the main subject of this paper, plays an important role in ORCA. Health signal generation is a very challenging task since it must be able to distinguish between normal healthy behaviour and abnormal faulty behaviour for very many, also unforeseen situations. An explicit fault model usually assumed in classical fault tolerance would become much too complex and is therefore not applicable.

Several concepts have already been proposed for health signal generation. One approach is inspired by the immune system and uses fuzzy rules with adaptable weights to detect abnormal situations [BrR08, JaM08]. Another concept is based on information theory and uses e. g. mutual information of signals for health signal generation [LJE07]. The latter approach will be further elaborated in this paper.

It has been an ongoing scientific interest to rate and measure the degree of information for given data sets in general. Different interpretations and approaches can be found in [HeQ95, PMV03, CoT06]. This includes for example varying normalizations of the mutual information like the ones presented in [LJE07] and [SHH99] that serve as a basis for this work.

The paper is organized as follows. First the ORCA framework is briefly presented in Chapter 2. Then the information-theoretical background used for health signal generation is given in Chapter 3. Besides already known approaches, also a new method based on the joint probability distribution of signals is introduced. To assess the quality of health signal generation, in Chapter 4 artificial worst case as well as real scenarios are evaluated for the proposed methods. Finally, Chapter 5 describes an example of the use of health signals for adaptable path planning for partially damaged robots.

2 Organic Robot Control Architecture (ORCA)

2.1 ORCA Framework

ORCA is a hierarchical and modular architecture which is based on two types of modules. Basic Control Units (BCUs) provide the basic functionality of the control system. The organic behaviour is implemented by separate Organic Control Units (OCUs). They constantly monitor the BCUs. In case OCUs detect an anomaly, they are able to change the behaviour of the BCUs so that the anomaly is handled in the best still possible way. The BCUs alone are sufficient to assure the functionality of the robot in the normal case, OCUs come only into play for abnormal situations.

Fig. 1 shows as simple example a differential drive of a wheeled mobile robot [MMG12]. There is a BCU on the drive layer which provides the interface of the drive to higher layers (not shown) with primitives like FORWARD, TURN LEFT, TURN

RIGHT etc. Its main function is to compute the corresponding speeds for the two motors which are given to separate motor controllers for the right and the left motor. These BCUs alone already implement the full functionality of the drive in the fault-free case. To support fault tolerance two additional OCUs are provided, one for each motor controller. They monitor their BCU for abnormal behaviour. If a motor shows e. g. a slower speed because it is damaged, they can counteract by changing the parameters like the set point of their motor controller.

ORCA has already been applied successfully to the walking robot OSCAR and allows it e.g. to continue walking even with damaged legs or to adapt to difficult terrain [MBG11].

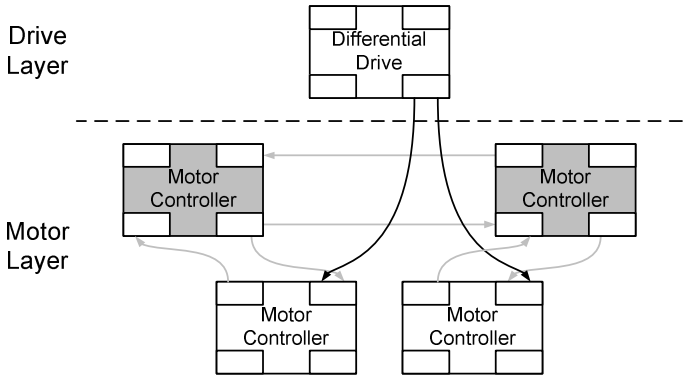


Figure 1: BCUs (white) and OCUs (gray) for a differential drive in ORCA. (Figure taken from [MMG12].)

2.2 Health Signals

An important concept of ORCA are health signals. They give an indication of the health status of a BCU which is monitored by an OCU. In abnormal situations the health signal drops and causes an OCU to intervene similarly to the natural immune system in case of an infection of the body. For the differential drive in Fig. 1 the motor controller BCUs generate a health signal which indicates the health state of their motors respectively. These health signals can be passed to other OCUs and combined to the health state on a higher layer, e. g. in the example in Fig. 1 for the whole drive.

3 Information-Theoretical Background

This chapter gives a brief introduction to the required information-theoretical background in section 3.1 by defining some basic identifiers and measures that are used to describe or rate signals. On this basis, section 3.2 presents special normalizations of the mutual information measure. Finally Section 3.3 introduces the signal classification

$\bar{C}_\omega$ as a measure to evaluate the similarity of two signals by analyzing their joint probability.

3.1 Signal Analysis

Let $S := (s_0, s_1, \dots, s_{t-1})$ be a discrete signal of length t, where s_z with $0 \leq z \leq t-1$ is the current value at time z. For each signal S there is a finite sample space $H_S := (Q_S, P_S)$. In this context $Q_S = \{(q_S)_0, (q_S)_1, \dots, (q_S)_{n-1}\}$ is the set of elementary events in S, where each item $(q_S)_i \in Q_S$ with $0 \leq i \leq n-1$ has a probability of occurrence $p((q_S)_i) \in [0,1]$ resulting in the vector $P_S := (p((q_S)_0), p((q_S)_1, \dots, p((q_S)_{n-1}))$. Hereafter p_i with $0 \leq i \leq n-1$ is used as a short form for $p((q_S)_i)$.

A visual representation of H_S is a histogram. This means that the elementary events in H_S are plotted against their respective probabilities.

Let $H_S = (Q_S, P_S)$ and $H_R = (Q_R, P_R)$ be the sample spaces of two signals S and R. The joint sample space of these signals is then given as $H_{SR} := (Q_S \times Q_R, P_{SR})$. The product set is defined as $Q_S \times Q_R := \left\{ \left((q_S)_i, (q_R)_j\right) \middle| \begin{matrix} i = 0, \dots, n-1 \\ j = 0, \dots, m-1 \end{matrix} \right\}$. Hereafter q_{ij} is used as a short form for $\left((q_S)_i, (q_R)_j\right)$ and p_{ij} stands for the joint probability $p\left(q_{ij}\right)$.

A visual representation of H_{SR} is a two dimensional histogram, where H_S is plotted against H_R and the joint probability of each possible combination is depicted as an intensity value.

The entropy of a given signal S is defined as $E_\alpha(H_S) := -\sum_{p_i \in P_S} p_i \log_\alpha p_i$, where α is used to set the base of the information measure. Common values of α are 2 (bit), e (nat) and 10 (Hartley) [HeQ95]. The joint entropy of two signals S and R is defined accordingly as $E_\alpha(H_{SR}) := -\sum_{p_{ij} \in P_{SR}} p_{ij} \log_\alpha p_{ij}$.

3.2 Mutual Information

Different definitions of the mutual information can be found throughout literature [HeQ95, PMV03, CoT06]. One possible definition focuses on the entropies of the involved signals and their joint entropy. In this context the mutual information is defined as $I_{\alpha(S,R)} := E_\alpha(H_S) + E_\alpha(H_R) - E_\alpha(H_{SR})$ and may reach maximal values if the joint entropy is minimal. It is furthermore possible to use normalized versions of the mutual information like $\hat{I}_\alpha(S,R) := \frac{I_\alpha(S,R)}{E_\alpha(H_S) \cdot E_\alpha(H_R)}$ [LJE07] or $\tilde{I}_\alpha(S,R) := \frac{E_\alpha(H_S) + E_\alpha(H_R)}{E_\alpha(H_{SR})}$ [SHH99].

In the context of health signal generation it is highly undesirable to use unbound measures like I_α and $\hat{I}_\alpha$. Therefore, this paper presents variants of these measures that are normalized with reference to a given signal. The goal of this approach is to achieve codomains that are bound by a [0, 1] interval. The basic idea of the normalization step is

to scale a given mutual information measure with the best possible fit of two signals. This results in the following three definitions:

$$I_\alpha^S(S,R) \coloneqq \frac{I_\alpha(S,R)}{I_\alpha(S,S)}$$

$$\hat{I}_\alpha^S(S,R) \coloneqq \frac{\hat{I}_\alpha(S,R)}{\hat{I}_\alpha(S,S)}$$

$$\tilde{I}_\alpha^S(S,R) \coloneqq \frac{\tilde{I}_\alpha(S,R)}{\tilde{I}_\alpha(S,S)}$$

A detailed derivation can be found in [Maa14].

3.3 Signal Classification

As mentioned in section 3.1 the joint entropy of two signals can be represented as a two-dimensional histogram. The distribution of elementary events within such a histogram depends on the characteristics of the involved signals. If two signals have a perfect match, their elementary events occur only in identical pairs. Therefore the two-dimensional histogram of $H_{RR} \coloneqq (Q_R \times Q_R, P_{RR})$ exhibits a diagonal. The more two signals differ from each other, the more their elementary events migrate away from the diagonal. In the worst case all elementary events are equally spread across the whole histogram.

The basic idea behind the signal classification approach is to rate the deviation of each elementary event q_i from the diagonal $\vec{g}$ in the joint probability distribution. Fig. 2a gives a visual impression of this approach. The diagonal of a perfect match goes from the bottom left corner to the top right corner and is called $\vec{g}$. The spatial boundaries are given by the two points $(x_{\min}, y_{\min})$ and $(x_{\max}, y_{\max})$ which are the minimal and maximal elementary events of two signals X and Y. The deviation of q_i is measured as the perpendicular distance $|\overrightarrow{d_i}|$ of q_i to its foot on $\vec{g}$. This results in the fact that different maximal deviations are possible, depending on the position of the foot point. Fig. 2b illustrates this effect. As can be seen the vectors $\vec{l}_i$ and $\vec{l}_j$ show the relation $|\vec{l}_i| < |\vec{l}_j|$. For an actual rating in the context of health signals such a behaviour is undesirable because each elementary event should be considered with the same weight.

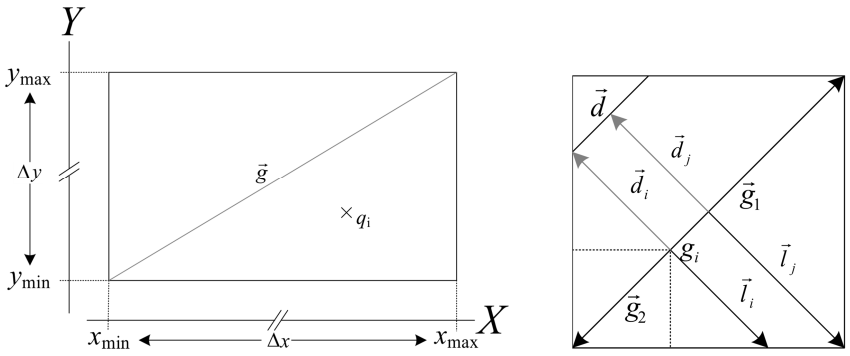
To smooth the weights of varying elementary events, the deviation of q_i is observed on a relative basis as $n_i \coloneqq \frac{|\vec{d}_i|}{|\vec{l}_i|}$. In this way the deviation is observed relative to its potential maximum. It has to be noted that this relation is only correct for square histograms, but is also used for rectangular histograms due to performance aspects.

Based on the previous considerations a precursor signal classification is defined as $C(S,R) \coloneqq \sum_{i=0}^{n-1} n_i p_i$ with $p_i \in P_{SR}$. Furthermore, a normalized version of this function

is defined as $C_\omega(S,R) \coloneqq \frac{C(S,R)}{\omega}$. This allows it to bind the classification to a range of [0, 1]. As a last step the final signal classification is defined as

$$\bar{C}_\omega(S,R) \coloneqq 1 - C_\omega(S,R).$$

A detailed derivation of the signal classification and an appropriate allocation of ω can be found in [Maa14].



(a) Two-dimensional histogram (b) Distance relations

Figure 2: (a) shows a two-dimensional histogram for the two signals X and Y. The diagonal $\vec{g}$ represents a perfect match of X and Y. The elementary event q_i deviates from this match. (b) illustrates different maxima that occur for varying foot points on $\vec{g}$. (Figures taken from [Maa14].)

4 Health Signal Generation

This chapter presents test scenarios for the previously introduced methods for health signal generation. These methods are then evaluated and compared with a reference method for all test scenarios.

4.1 Methods

The basic idea of the information-theoretical approach is to use signal properties like entropy, mutual information or probability distribution to distinguish normal signals from abnormal ones and to compute a health signal based on the extent of the deviation.

A straight-forward approach to measure the similarity of two signals S and R is to simply compute the sum of absolute differences between them

$$\mathrm{SAD}(S,R) \coloneqq \sum_{i=0}^{t-1} |s_i - r_i|,$$

which is in the following taken as a reference.

Furthermore, the three mutual information-based measures defined above

$$I_\alpha^S, \hat{I}_\alpha^S, \tilde{I}_\alpha^S$$

are used for health signals.

Also the new method for signal classification

$$\bar{C}_\omega(S, X)$$

which is based on the joint probability distribution of two signals S and X is evaluated.

4.2 Test Scenarios

In the following the quality of health signal generation for the information-theoretical approach shall be evaluated. For this purpose, nine representative scenarios are defined and the health signal generation according to the five methods described above are compared.

All scenarios assume three different observed signals:

- an optimal signal for the given situation
- a similar fault-free signal
- a faulty signal.

Scenarios 1 to 6 are artificial worst-case scenarios, scenarios 7 to 9 use real measurement data from damaged robots. In the first three scenarios and the real scenario 7 the abnormal signals differ in their levels form the normal ones, in scenarios 4 to 6 and 8 there is a loss of structure in the abnormal signal, e. g. due to noise. All scenarios are described in detail in [Maa14]. As an example scenario 9 is elaborated in the following.

A real mobile robot has been used for health signal generation (Fig. 3). It uses two kinds of sensors for obstacle avoidance: two ultrasonic distance sensors (US) on the left and right front side with the cones shown in Fig. 3 and an infrared distance sensor (IR) pointing in the forward direction with a very narrow cone pictured as a red line. The BCUs of the robot execute a simple reactive obstacle avoidance behaviour which is based on potential fields and vector addition. The resulting robot trajectory is represented by the dotted line. An OCU monitors the signals from the US and IR sensors. For the experiment, the average of the US sensor signals is taken as a fault-free reference signal S. The fault-free signal of the IR sensor is signal R_1 demonstrating a fault-free normal case. In order to inject a fault, the IR sensor is detached from its A/D converter resulting in a low level noisy signal R_2 for this abnormal case. All three signals are smoothed by a median filter to reduce noise and smooth signal inconsistencies that are caused by the layout of the sensors.

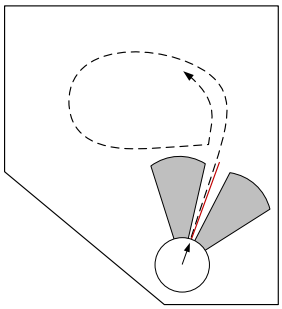


Figure 3: Mobile Robot with US- and IR-Sensors from Scenario 9. (Figure taken from [Maa14].)

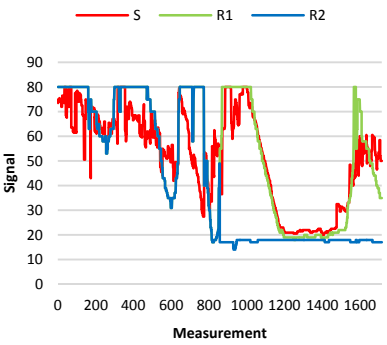
4.3 Results

For the real world scenario 9 described above, the measured signal S for the optimal case, signal R_1 for the similar normal fault-three case and signal R_2 for the abnormal faulty case are shown in Fig. 4a. All three signals have a similar course until the fault is injected at $t = 858$. Then signal R_2 of the IR sensor drops to a low nearly constant level while the normal fault-free signals S and R_1 continue to run very similarly to each other while the robot continues moving.

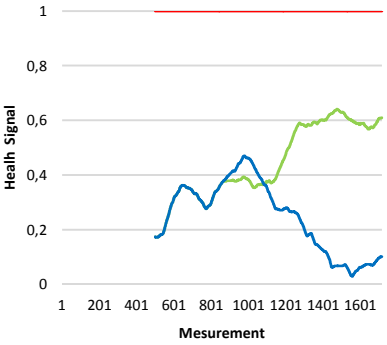
In Fig. 4b to Fig. 4f the corresponding health signals according to the five methods presented in Chapter 4.1 are pictured. As can be seen all health signals react to the fault in R_2 by dropping, however, with different characteristics. For the fault-free sensor signal R_1, all health signals have a lower value than expected.

The two mutual information-based measures in Fig. 4b and 4d look similar and are able to clearly distinguish between the normal and abnormal situation. However, also in the fault-free case they only have a low health value although the system is healthy. This is better for Fig. 4c, but here the difference between the faulty and fault-free situations is not so clear any more. The SAD method in Fig. 4e has a rather high health signal as long as no fault occurs and then drops clearly. However, it later rises to a high level again though the fault is still there. As can be seen from Fig. 4a, this happens because the signal level for the fault-free sensors S is also very low from $t = 1150$ to 1500 resulting in only a small absolute difference to the faulty signal R_2. In summary, the health signal generation is not satisfactory for all these four methods.

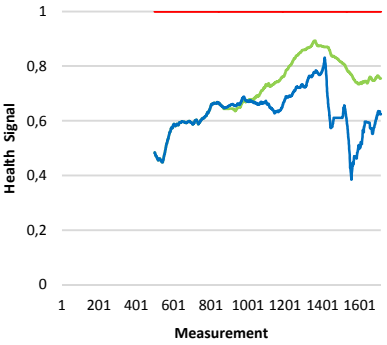
The best results are observed for the new proposed probability distribution-based method in Fig 4f. It generates high health signal values for the normal case and clearly lower ones in case of the anomaly. It can be observed that $\bar{C}_\omega(S, R_2)$ shows a light reaction to the fault at first and then a strong reaction. This effect occurs because the signals were not completely assessed at once, but with a sliding window. The weaker reaction shortly after the fault can be explained by the fact that the window still holds fault-free values that keep the rating at a certain level. The health signal drops significantly, when the sliding window has gathered enough faulty data and the structure of the corresponding two-dimensional histogram changes.



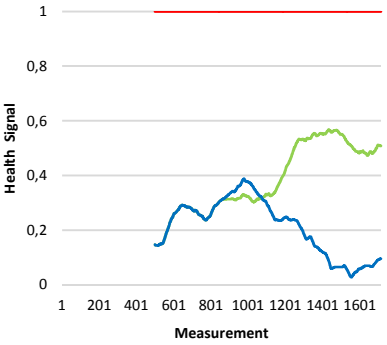
(a) Signals S, R_1 und R_2



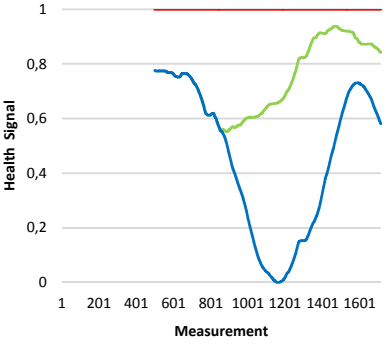
(b) $I_{10}^{S}(S,X)$



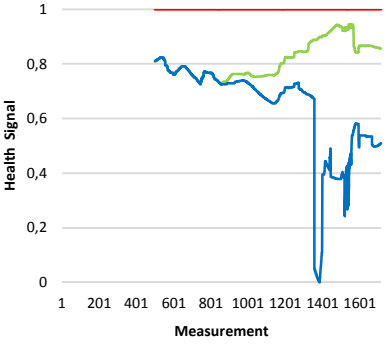
(c) $\hat{I}_{10}^{S}(S,X)$



(d) $\tilde{I}_{10}^{S}(S,X)$



(e) $\mathrm{SAD}_{\omega}(S,X)$



(f) $\bar{\mathrm{C}}_{\omega}(S,X)$

Figure 4: Results for Health Signal Generation in Scenario 9. (Figure based on [Maa14].)

A summary of all studied measures and all scenarios is shown in Tab. 1. Six categories are used to indicate the performance of all methods. If the resulting signals show positive and negative aspects the corresponding method is rated with $\pm$. Light tendencies to either positive or negative results are represented by $\oplus$ or $\ominus$. Strong tendencies are shown as $\oplus\oplus$ or $\ominus\ominus$. If a method is not applicable in a given scenario, it is marked with $\otimes$.

It can be seen that the new method based on joint probability distributions is superior to the three methods based on mutual information and SAD (Sum of Absolute Differences) for the studied scenarios. The measures based on mutual information are even worse than SAD in most cases and turn out to be not suitable for health signal generation.

Table 1: Summary of the Evaluation of Health Signal Generation for Scenario 1 to 9. (Figure based on [Maa14].)

Scenario	I_α^S	$\hat{I}_\alpha^S$	$\tilde{I}_\alpha^S$	SAD_ω	$\bar{C}_\omega$
1	$\otimes$	$\otimes$	$\otimes$	$\oplus$	$\oplus\oplus$
2	$\ominus\ominus$	$\otimes$	$\ominus\ominus$	$\oplus$	$\oplus\oplus$
3	$\otimes$	$\otimes$	$\otimes$	$\oplus$	$\oplus$
4	$\otimes$	$\otimes$	$\otimes$	$\otimes$	$\oplus$
5	$\ominus\ominus$	$\otimes$	$\ominus\ominus$	$\otimes$	$\oplus$
6	$\pm$	$\pm$	$\pm$	$\pm$	$\oplus$
7	$\otimes$	$\ominus\ominus$	$\ominus\ominus$	$\oplus$	$\oplus$
8	$\otimes$	$\ominus\ominus$	$\ominus\ominus$	$\oplus$	$\oplus$
9	$\pm$	$\pm$	$\pm$	$\pm$	$\oplus$

5 Adaptive Path Planning

In this chapter an outlook is given on how health signals can be used for adaptive path planning in difficult terrain for injured mobile robots (e.g. wheeled robots with damaged drives or walking robots with damaged legs [MBG11]). For this purpose, a classical wavefront-based path planner [BLL92, Jar95] is executed by a BCU. In case of a lowered health signal, the OCU modifies the map of the terrain by changing those areas which are too difficult to pass for the current health status of the robot into obstacles. Thus the BCU replans the path around them (see Fig. 5). A detailed description and evaluation of different approaches can be found in [MaM11, Maa14].

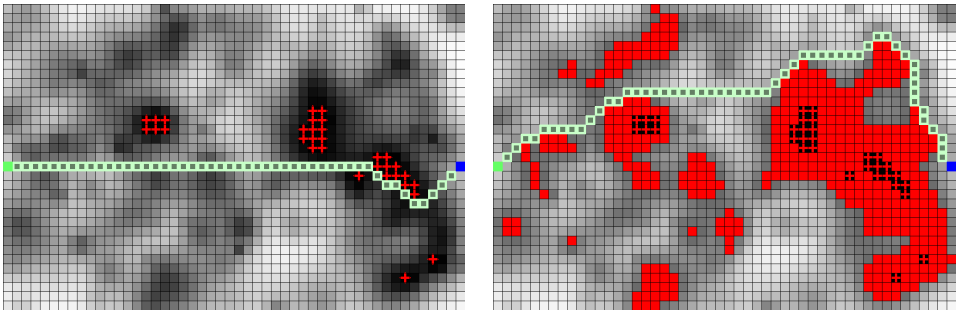


Figure 5: Adaptive Path Planning for a mobile robot based on health signals (left: 100% healthy, right 90% healthy). Difficulties of the terrain are shown as grey levels, obstacles in red. (Figure taken from [Maa14].)

6 Conclusion

This paper has presented an evaluation of the quality of health signal generation for various information-theoretical methods in the ORCA framework. Health signals are used in ORCA by Organic Control Units (OCUs) in order to describe the health status of Basic Control Units (BCUs) monitored by them. It turned out that SAD as well as known mutual information-based measures do not always yield satisfactory results. Only a newly proposed method based on the joint probability distribution of signals showed the desired properties. It was also presented how health signals can be used to plan a suitable path through difficult terrain so that even partly damaged robots can reach their goal.

Future comparative studies should also include other than information-theoretical methods like e.g. fuzzy rule sets. Moreover, it needs to be studied further which information about the health status of a BCU is actually required by higher levels like the path planner. The health signal alone only gives a quantitative measure on how healthy a robot is but does not say what kind of damage it has and what capabilities still remain. Therefore, it can be desirable that the OCUs at higher levels are able to request this information from OCUs at lower levels.

Acknowledgement

This work was funded in part by the German Research Foundation (DFG) within the priority programme Organic Computing (1183) under grant reference MA 1412/8-2.

References

[BrR08] Brockmann, W.; Rosemann, N.: Instantaneous anomaly detection in online learning fuzzy systems. In: Workshop on Genetic and Evolving Systems (GEFS). 2008; p. 23-28.

[JaM08] Jakimovski, B.; Maehle, E.: Artificial Immune System Based Robot Anomaly Detection Engine for Fault Tolerant Robots. In: Autonomic and Trusted Computing, Lecture Notes in Computer Science, 5060, 2008. Springer Berlin, Heidelberg. 2008; p. 177-190.

[LJE07] Larionova, S.; Jakimovski, B.; El Sayed Auf, A.; Litza, M.; Moesch, F.; Maehle, E.; Brockmann, W.: Toward a Fault Tolerant Mobile Robot: Mutual Information for Monitoring of the Robot Health Status. In: International Workshop on Technical Challenges for Dependable Robots in Human Environments. IEEE/RAS, 2007; p. 1-5.

[HeQ95] Heise, W.; Quattrocchi, P.: Informations- und Codierungstheorie, Springer Berlin Heidelberg, 1995.

[PMV03] Pluim, J. P. W.; Maintz, J. B. A.; Viergever, M. A.: Mutual-information-based registration of medical images: a survey. In: Transactions on Medical Imaging, Vol. 22, No. 8. IEEE, 2003; p. 986-1004.

[CoT06] Cover, T. M.; Thomas, J. A.: Elements of Information Theory, John Wiley & Sons, 2006.

[SHH99] Studholme, C.; Hill, D. L. G.; Hawkes, D. J.: An overlap invariant entropy measure of 3D medical image alignment. In: Pattern Recognition, Vol. 32, No. 1. 1999; p. 71-86.

[MMG12] Maas, R.; Maehle, E.; Großpietsch, K. E.: Applying the Organic Robot Control Architecture ORCA to Cyber-Physical Systems. In: EUROMICRO Conf. on Software Engineering and Advanced Applications (SEAA), 2012. IEEE, 2012; p. 250-257.

[MBG11] Maehle, E.; Brockmann, W.; Großpietsch, K.-E., El Sayed Auf, A.; Jakimovski, B.; Krannich, S.; Litza, M.; Maas, R.; Al-Homsy, A.: Application of the Organic Robot Control Architecture ORCA to the Six-legged Walking Robot OSCAR. In: Organic Computing - A Paradigm Shift for Complex Systems. Birkhäuser-Springer, Basel 2011; p. 517-530.

[Maa14] Maas, R.: Adaptive Kontrollstrategien für mobile Roboter basierend auf Organic Computing Prinzipien, University of Lübeck, 2014 (to appear).

[BLL92] Barraquand, J.; Langlois, B.; Latombe, J.: Numerical Potential Field Techniques for Robot Path Planning. In: Trans. on Man and Cybernetics, 22. IEEE, 1992; p. 224-241.

[Jar95] Jarvis, R.: Collision Free Trajectory Planning Using Distance Transforms. In: Mechanical Engineering Transactions, ME10. 1995; p. 187-191.

[MaM11] Maas, R.; Maehle, E.: Fault Tolerant and Adaptive Path Planning for Mobile Robots Based on Health Signals. In: International Conference on Architecture of Computer Systems (ARCS) 2011, Workshop Proceedings. VDE Verlag, 2011; p. 58-63.

Elektronische Wahlen: Unterstützung der Wahlprozesse mittels Technik

(eVoting)

Elektronische Wahlen: Unterstützung der Wahlprozesse mittels Technik

Katharina Bräunlich[1], Stephan Neumann[2], Jurlind Budurushi[2], Rüdiger Grimm[1]

[1]Universität Koblenz-Landau
IT-Risk-Management
Universitätsstr. 1
56070 Koblenz
{braeunlich, grimm}@uni-koblenz.de

[2]Technische Universität Darmstadt
Security, Usability, and Society - SecUSo
Hochschulstr. 10
64289 Darmstadt
{stephan.neumann, jurlind.budurushi}@cased.de

1 Vorwort

Mittels Wahlen werden Ämter oder Aufgaben vergeben. Sie finden in vielen unterschiedlichen Gesellschaftsbereichen statt. Beispiele sind unter anderem Bundestags-, Landtags-, Kommunal-, Sozial-, Betriebsrats- und Vereinswahlen. Ihre politische und/oder gesellschaftliche Bedeutung erfordert die Einhaltung maximaler Sicherheitsanforderungen bezüglich technischer, rechtlicher, organisatorischer und sozialer Aspekte. Die hohen Sicherheitsanforderungen an elektronische Wahlsysteme werden meist mit Hilfe komplexer Wahlsysteme umgesetzt. Auf Grund ihres hohen Komplexitätsgrades bei gleichzeitiger Nutzbarkeit sind elektronische Wahlsysteme als Forschungsgegenstand von besonderem Interesse. Insbesondere die scheinbare Unvereinbarkeit von Verifizierbarkeit (öffentliche Wahl) und Anonymität (geheime Wahl), sowie die Gestaltung benutzbarer, transparenter und vertrauenswürdiger Systeme vor dem Hintergrund der meist komplexen kryptographischen Protokolle zur gleichzeitigen Umsetzung von Verifizierbarkeit und Anonymität gilt es zu diskutieren und zu lösen. Ebenso existieren interdisziplinäre Fragestellungen wie zum Beispiel die Identifizierung und Operationalisierung geeigneter vertrauensschaffender Maßnahmen in Hinblick auf elektronische Wahlsysteme sowie der Ausgestaltung des gesetzlichen Gestaltungsspielraums, d.h. den geforderten Grad der Umsetzung der Wahlrechtsgrundsätze. Diese und andere Forschungsfragen gilt es, fächerübergreifend zu diskutieren und zu lösen, bevor der Einsatz elektronischer Wahlen bei politischen Wahlen in Deutschland denkbar und sinnvoll scheint.

Der Beitrag *„Prozessmodell und -analyse für die Stadtrats und Kommunalwahl 2014 in Koblenz“* [Simic-Draws, Bräunlich] identifiziert und spezifiziert die Wahlprozesse, welche es mittels Technik zu unterstützen gilt.

Ein weiterer Themenschwerpunkt ist die kritische Auseinandersetzung mit Erfahrungen aus dem Ausland (Österreich, Schweiz, Estland, Norwegen, Luxemburg) bei Pilotprojekten und bei rechtlich-bindenden Wahlen, welche gewinnbringend in die Diskussion um sichere und rechtsgemäße elektronische Wahlen eingebracht werden sollten. Hierzu liefern die die folgenden vier Artikel einen Beitrag: *„Iraqi Elections in 2014: A Privacy Requirement Evaluation Based on a Polling Place Experience“* [Al-Shammari, Villafiorita], *„Online-Partizipation in der EU? Erste Erfahrungen mit Europäischen Bürgerinitiativen aus Sicht Österreichs“* [Stein, Wenda], *„Das Zentrale Wählerregister – Ein skalierbares Instrument zur Bürgerbeteiligung mit 1:1-Verifikation“* [Stein, Wenda] und *„Identifying Building Blocks of Internet Voting: Preliminary Findings“* [Krimmer].

Die Gestaltung benutzbarer, transparenter und vertrauenswürdiger Systeme mittels kryptographischer Protokolle wird in den beiden Beiträgen *„A Lightweight Implementation of a Shuffle Proof for Electronic Voting Systems“* [Locher, Haenni] und *„Coercion-Resistant Internet Voting in Practice“* [Feier, Neumann, Volkamer] adressiert.

Die Evaluierung und Zertifizierung elektronischer Wahlsysteme soll das Vertrauen in die Sicherheit elektronischer Wahlprodukten stärken. Die Common Criteria sind ein international anerkannter Kriterienkatalog zur Evaluierung und Zertifizierung von IT-Systemen. Für elektronische Wahlsysteme mit einem Papierausdruck zur Verifizierbarkeit ihrer korrekten Ausführung begründet der Artikel *„Entwicklung eines Common Criteria Schutzprofils für elektronische Wahlgeräte mit Paper Audit Trail“* [Budurushi, Neumann, Shala, Volkamer] entsprechende Sicherheitsanforderungen.

Prozessmodell und -analyse für die Stadtrats- und Kommunalwahl 2014 in Koblenz

Daniela Simic-Draws, Katharina Bräunlich

Universität Koblenz-Landau
Fachbereich Informatik
Universitätsstr. 1
56070 Koblenz
{dsimic, braeunlich}@uni-koblenz.de

Abstract: Im Rahmen dieser Arbeit wurde der Ablauf der Stadtratswahl in Koblenz 2014 als Prozessmodell spezifiziert. Dieses Prozessmodell besteht aus vier Teilprozessen: Der Wahlvorbereitung, der Stimmabgabe, der Stimmauszählung und der Wahlnachbereitung. Die Wahlvorbereitung und Stimmauszählung werden bei der genannten Wahl durch elektronische Hilfsmittel unterstützt. Dies ist in Hinblick auf die Diskussion bezüglich des Einsatzes und Umfangs elektronischer Hilfsmittel und deren Rechtmäßigkeit im Kontext von Wahlen interessant. Das hier spezifizierte Prozessmodell kann ein tieferes Verständnis für die komplexen Abläufe der Wahl schaffen, die Nachvollziehbarkeit sowohl für Wahlorganisatoren als auch die Wahlöffentlichkeit verbessern als auch helfen, sicherheitskritische Aspekte aufzudecken und so die Rechtmäßigkeit der eingesetzten elektronischen Hilfsmittel zu bewerten.

1 Einleitung

Am 25. Mai 2014 findet in Koblenz neben der Europawahl auch die Kommunalwahl statt. Die Durchführung dieser Wahl ist aufgrund ihrer Eigenheiten nicht nur für den Bürger, sondern auch für Wahlorganisatoren eine komplexe Angelegenheit. Aus der Perspektive der Wahlorganisatoren sind über den Zeitraum des eigentlichen Wahlvorgangs hinweg komplexe Abläufe erforderlich. Im Rahmen einer Kooperation mit dem Ordnungsamt Koblenz wurde der Ablauf der Stadtratswahl[1] als Prozess modelliert. Dieses Prozessmodell kann nicht nur einer besseren Verständlichkeit der Abläufe für die Wahlorganisatoren dienen, sondern kann auch dem Bürger öffentlich zugänglich gemacht werden und somit zu einer besseren Nachvollziehbarkeit der Wahl gemäß dem Öffentlichkeitsgrundsatz[2] beitragen. Darüber hinaus erlaubt das Prozessmodell eine Untersuchung der Wahlabläufe in Hinblick auf sicherheitskritische Aspekte [Sim14].

Das in diesem Artikel vorgestellte Prozessmodell spezifiziert die Durchführung der Stadtratswahl für die Stadt Koblenz 2014. Es besteht aus vier Teilprozessen: Der Wahlvorbereitung, der Stimmabgabe, der Stimmauszählung und der Wahlnachbereitung.

[1] Die Stadtratswahl ist Bestandteil der Kommunalwahl.
[2] BVerfG: Art 38 Abs. 1 GG i.V.m. Art. 20 Abs. 1, 2 GG

Die Stimmabgabe kann bei der Europa- sowie Kommunalwahl 2014 in Koblenz in Papierform vor Ort (Urnenwahl) oder per Briefwahl erfolgen. Die Erstellung des Wählerverzeichnisses in der Wahlvorbereitung und die Stimmauszählung erfolgen bei der genannten Wahl unter Zuhilfenahme elektronischer Hilfsmittel. Dies erscheint im Zuge der Diskussionen um die Rechtmäßigkeit elektronischer Wahlen[3] und der Einhaltung der Wahlrechtsgrundsätze interessant. Das hier vorgestellte Prozessmodell kann dazu beitragen, sicherheitskritische Aspekte aufzudecken und die Rechtmäßigkeit der eingesetzten elektronischen Hilfsmittel zu bewerten.

Nach [SS08, S. 62] besteht ein Prozess aus einer „Reihe von Aktivitäten, die aus einem definierten Input ein definiertes Ergebnis (Output) erzeugt". Es lassen sich somit auch nicht-wertschöpfende Aktivitätsabfolgen[4] wie beispielsweise in der öffentlichen Verwaltung mit diesem Prozessbegriff fassen. Zur Erhebung und Dokumentation können die aus dem Geschäftsprozessmanagement bekannten Verfahren z. B. nach [Fi06] verwendet werden. Ein Prozess startet i.d.R. mit einer Anforderung, die als Startereignis aufgefasst wird und endet mit der Erfüllung dieser Anforderung. Das Endereignis ist somit das Erreichen des vordefinierten Resultats. Dieses Ergebnis soll innerhalb der gegebenen Anwendungsumgebung erreicht werden. Ein Prozess beschreibt demnach die Interaktion unterschiedlichster Ressourcen (Menschen, IT-Systeme, Informationen, Güter,…) miteinander unter Berücksichtigung vorgegebener Regelwerke[5] zur Erreichung des vorgegebenen Ergebnisses. Die Prozessmodellierung dient der systematischen Dokumentation von Prozessen, welche als Grundlage der Prozessanalyse und -optimierung dient. Ein Prozessmodell erlaubt somit auch die Sicherheitsanalyse sowohl der Prozesse als auch deren Einbettung in die Einsatzumgebung.

Dieser Artikel ist wie folgt gegliedert: Zunächst erfolgt in Abschnitt 2 ein Überblick über verwandte Arbeiten, welche die Spezifizierung von Sicherheitsanforderungen zum Ziel haben. Dabei werden insbesondere prozessorientierte Methoden berücksichtigt. Darüber hinaus werden solche Arbeiten vorgestellt, welche sich mit der Spezifikation von Sicherheitseigenschaften bei elektronischen Wahlen auseinandersetzen. Die Eigenschaften und Besonderheiten von Kommunalwahlen werden anschließend in Abschnitt 3 vorgestellt. In Abschnitt 4 erfolgt dann die Beschreibung und Darstellung eines der IT-gestützten Teilprozesse, die im Rahmen der diesjährigen Kommunalwahl in Koblenz durchlaufen werden: Die Auszählung der Stimmen. Dieser Teilprozess wird in Abschnitt 5 einer Sicherheitsanalyse unterworfen. Der Artikel schließt mit einer Zusammenfassung der Ergebnisse sowie einem Fazit.

[3] Gemäß [Kr02] bezeichnet „e-Voting im Allgemeinen" die elektronisch unterstützte Durchführung mindestens einer der Prozesse Wähleridentifizierung, Stimmabgabe und/oder Stimmauszählung. D.h., es können sowohl der gesamte Wahlablauf als auch nur einzelne Schritte wie Wählerauthentifizierung, Erfassung, Weiterleitung, Speicherung und/oder Auswertung der Stimmen elektronisch erfolgen.

[4] Eine Abgrenzung zwischen Prozess und Geschäftsprozess ist wichtig, da bei letzterem die Anforderungen des und die zu erbringenden Leistungen für den Kunden – repräsentiert durch die wertschöpfenden Tätigkeiten eines Unternehmens – im Fokus stehen. Im Kontext der öffentlichen Verwaltung findet in der Regel keine Transformation in Güter / Dienstleistungen mit höherem Geldwert statt.

[5] Dies umfasst organisationsinterne Regelwerke wie z. B. Prüfungen, Rückkopplungen, Aufteilen von Verantwortlichkeiten etc.; diese können auch Governance Richtlinien entstammen.

2 Verwandte Arbeiten

Sicherheitsanforderungen können mit Hilfe einer Vielzahl unterschiedlicher Methoden spezifiziert werden: Eine international anerkannte und etablierte Methode zur Spezifikation von Sicherheitsanforderungen an IT-Produkte und die Evaluierung dieser IT-Produkte in Hinblick auf die spezifizierten Sicherheitsanforderungen sind die Common Criteria[6] [CC12]. Diese bieten einen umfassenden Katalog an standardisierten funktionalen Sicherheitsanforderungen, welche von einem spezifischen IT-Produkt oder einer generischen Produktgruppe gefordert werden. Die Common Criteria können daher sowohl im Produktentstehungs- als auch im eigentlichen Produktlebenszyklus angewendet werden. Ein Vorgehen, welches speziell für Softwareentwicklungsprojekte konzeptioniert wurde, ist SQuaRE (Security Quality Requirements Engineering) nach [MHS05]. Diese Methode zielt darauf ab, in neun Schritten kategorisierte und priorisierte Sicherheitsanforderungen zu spezifizieren. Ebenfalls eine Anwendung im frühen Stadium der Software-Entwicklung sieht CLASP (Comprehensive Lightweight Application Security Process) nach [Vie05] vor. CLASP liefert eine Systematik in Form von Prozessbausteinen, welche in den regulären Software-Entwicklungs- bzw. Verbesserungsprozess integriert werden können. Dabei werden basierend auf Rollen, Ressourcen und erwarteten Interaktionen korrespondierende Sicherheitsanforderungen abgeleitet. Eine Vorgehensweise, mit der rechtliche Anforderungen in technische Implementierungsvorschläge konkretisiert werden, liefert [HPR93] mit KORA (Konkretisierung rechtlicher Anforderungen). Technische Sicherheitsanforderungen können aufgrund der Zusammenführung juristischen und informatischen Fachwissens systematisch und für beide Seiten nachvollziehbar von den abstrakten rechtlichen Vorgaben abgeleitet werden. In [Sim+13] erfolgt eine Integration von KORA, den Common Criteria und den BSI-Grundschutzkatalogen [BSI11]. Die Sprachmodelle der Common Criteria und des IT-Grundschutz werden hierbei durch das methodische Vorgehen von KORA angereichert und umgekehrt. Die Berücksichtigung sowohl rechtlicher als auch technischer und organisatorischer Aspekte erlaubt somit eine ganzheitliche Sicht auf IT-Sicherheit.

Eine Integration der Prozess- mit der Sicherheitsperspektive wird bereits in einigen Arbeiten thematisiert: Das SLP-Konzept nach [MBH14] basiert auf etablierten Standards und beschreibt deren Zusammenspiel, um sichere Logistikprozesse entwerfen zu können. Die Erweiterung bestehender Notationen zur Geschäftsprozessmodellierung wird von einigen Arbeiten mit jeweils unterschiedlichen Schwerpunkten thematisiert: eine Menge festgelegter Sicherheitsanforderungen wird in Form grafischer Elemente repräsentiert, mit denen bestehende Prozessmodelle annotiert werden wie z. B. in [DK06], [Ci+11] und [Pa+12]. Eine syntaktische Erweiterung der BPMN, indem neue Notationselemente spezifiziert werden, nehmen z. B. [RFP07a], [RFP07b] und [Lo+05] vor. Das Versehen bestehender Notationselemente mit einer neuen Bedeutung im Sinne einer semantischen Erweiterung nehmen z. B. [AMA12] vor.

Im Anwendungsbereich der elektronischen Wahlen existieren zahlreiche Arbeiten, welche sich mit der Spezifikation von Sicherheitseigenschaften beschäftigen:

[6] Die drei Teile der Common Criteria wurden als DIN ISO/IEC 15408-1...3 veröffentlicht.

Anforderungen an elektronische Wahlen bzw. Internetwahlen im Allgemeinen werden z.B. in [CoE04], [VV08] oder [Br+13] adressiert. Bei der prozessorientierten Anforderungsanalyse für elektronische Wahlen wird oftmals der Fokus auf bestimmte Sicherheitsziele wie z.B. Quittungsfreiheit, Verifizierbarkeit oder Wahlgeheimnis gelegt. Diese Sicherheitsziele werden formal spezifiziert. Anschließend kann für ein ebenfalls formal spezifiziertes Produkt bzw. Protokoll die Erfüllung dieser Sicherheitseigenschaft nachgewiesen werden. Dies erfolgt z.B. in [DKR09], [JP06] in Bezug auf das Wahlgeheimnis für Receipt-Freeness und Privacy, z.B. in [KRS10] für Verifizierbarkeit und z.B. in [KR05], [BHM08] für Gleichheit (Fairness) und Wahlberechtigung (Eligibility).

In keiner der genannten Arbeiten wird der gesamte Prozess einer Wahl über alle Wahlphasen hinweg, von Wahlvorbereitung über Stimmabgabe bis hin zur Ergebnisermittlung, unter Einbezug der beteiligten Menschen und ihrer Interessen betrachtet. Diese Arbeiten erlauben somit keine ganzheitliche Schwachstellenanalyse wie die in dieser Arbeit angewandte Vorgehensweise nach [Sim14], auf welche im Abschnitt 5 näher eingegangen wird.

3 Kommunalwahlen

Kommunalwahlen zählen ebenso wie Bundestags- oder Landtagswahlen zu den parlamentarischen Wahlen und unterliegen somit den verfassungsrechtlichen Wahlrechtsgrundsätzen der allgemeinen, freien, gleichen, geheimen, unmittelbaren und öffentlichen Wahl gemäß Art.38 Abs. 1 Satz 1 GG[7] sowie Art. 38 in Verbindung mit 20 Abs. 1 und Abs. 2 GG. Bei Kommunalwahlen werden beispielsweise Gemeinde-, Stadt- oder Landräte oder in der Direktwahl Bürgermeister gewählt. Im föderalistischen System der Bundesrepublik Deutschland ist das Kommunalwahlrecht Landesrecht. Die genaue Ausgestaltung von Kommunalwahlen kann sich daher von Bundesland zu Bundesland unterscheiden[8]. Da diese Arbeit sich mit der Kommunal- und Stadtratswahl der Stadt Koblenz in 2014 beschäftigt, beziehen sich die nachfolgenden Erläuterungen auf das Kommunalwahlrecht in Rheinland-Pfalz wie es durch das Kommunalwahlgesetz [KWG94] des Landes Rheinland-Pfalz festgeschrieben ist. Die Kommunalwahl in Rheinland-Pfalz ist eine Verhältniswahl mit offenen Listen. Bei der Verhältniswahl erhält eine Partei entsprechend ihrem Stimmanteil Mandate.[9] Die Wahllisten werden in der Regel durch die Parteien aufgestellt. Da die Parteien bei der Zusammenstellung der Wahllisten nicht auf Parteimitglieder beschränkt sind, spricht man von offenen Listen. Die Anzahl der zu vergebenden Stimmen ist hierbei abhängig von der Zahl der zu vergebenden Sitze [KWG94, §32]. Des Weiteren erlaubt das KWG von Rheinland-Pfalz das Kumulieren und Panaschieren (vgl. §32 Sätze 3 und 4). Kumulieren bezeichnet die Möglichkeit, einem Kandidaten mehrere Stimmen geben zu können, um so dessen

[7] Grundgesetz

[8] Unterschiede bestehen z. B. hinsichtlich dem aktiven / passiven Wahlalter (18/18), der Wahlperiode (5 Jahre), dem Sitzzuteilungsverfahren (Hare/Niemeyer), der Listenform (offen) und der Anzahl an Stimmen (Zahl der zu vergebenden Sitze) – vgl. dazu [KWG94].

[9] Im Gegensatz dazu wird das Wahlgebiet bei Mehrheitswahlen entsprechend der zu vergebenen Mandate in Wahlkreise eingeteilt. Das Mandat erhält dann der Kandidat mit den meisten Stimmen in seinem Wahlkreis.

Position innerhalb der Wahlliste zu stärken. Panaschieren beschreibt das Verteilen mehrerer verfügbarer Stimmen durch den Wähler auf Kandidaten von unterschiedlichen Wahllisten [Br+13].
Das in diesem Artikel vorgestellte Prozessmodell spezifiziert die Durchführung der Stadtratswahlen in Koblenz 2014. Die Stadtratswahl folgt den oben aufgeführten Reglementierungen der Kommunalwahl.

4 Prozessmodell für die Stadtratswahl 2014 in Koblenz

Im Rahmen einer Kooperation mit dem Ordnungsamt Koblenz wurde der gesamte Ablauf der Stadtratswahl 2014 für die Stadt Koblenz als Prozess modelliert. Im Folgenden wird ein Auszug aus diesem Prozessmodell vorgestellt. Dabei liegt der Fokus auf einem der beiden IT-gestützten Teilprozesse, nämlich der elektronisch unterstützten Stimmauszählung[10]. Dieser Teilprozess wird nachfolgend hinsichtlich seines Ist-Zustandes beschrieben. Eine Analyse hinsichtlich sicherheitskritischer Aspekte erfolgt in Abschnitt 5.

Das Kumulieren und Panaschieren steht häufig in der Kritik, die Stimmabgabe zu verkomplizieren und so als Fehlerquelle für ungewollt ungültige Stimmzettel bei der Stimmabgabe durch den Wähler sowie als Fehlerquelle bei der Stimmauszählung zu dienen. Bei der Kommunalwahl in Koblenz werden für die 76 Stimmbezirke und Briefwahlbezirke über 900 Mitglieder des Wahlvorstandes an der Auszählung beteiligt sein. Bei diesen handelt es sich überwiegend um juristische Laien. Hierbei besteht z.B. die Gefahr, dass falsch ausgefüllte Stimmzettel nicht gesetzeskonform erfasst oder unterschiedlich interpretiert werden. Zudem müssen die ermittelten Stimmen mit Hilfe eines Sitzzuteilungsverfahrens in die Abgeordnetensitze umgerechnet werden, was manuell recht langwierig ist. Dies könnten Gründe sein, warum in Koblenz (wie in vielen anderen Kommunen auch) die Stimmauszählung unter Zuhilfenahme elektronischer Hilfsmittel durchgeführt wird.

Im Rahmen der Kommunalwahl Koblenz wurde vom Landeswahlleiter Rheinland-Pfalz das Stimmenauszählungs-Programm „PC Wahl“ (genauer: die Module „Heiler“ und „Berechnung der Sitzverteilung“) der Berninger Software GmbH freigegeben[11]. Die Freigabe erfolgt gemäß [KWO83, §55a] und umfasst bspw. die algorithmisch korrekte Implementierung der juristisch vorgegebenen Heilungsvorschriften[12].

[10] Die elektronisch unterstützte Erstellung des Wählerverzeichnisses wird hier nicht weiter betrachtet.
[11] http://www.wahlen.rlp.de/kw/bek/20140212_Bekanntmachung_Freigabe_PCWahl_Staatsanzeiger.pdf (Abrufdatum: 01.04.2014)
[12] Beispiel: Laut KWG können für einen Kandidaten max. drei Stimmen vergeben werden. Vergibt ein Wähler nun sechs Stimmen an einen Kandidaten, so wird der Stimmzettel nicht ungültig, sondern er wird bei der Erfassung „geheilt“. Das heißt, der Stimmzettel wird 1:1 mit den betreffenden sechs Stimmen erfasst, es fließen aber nur drei dieser Stimmen in die Berechnung mit ein (vgl. §37 (3) KWG).

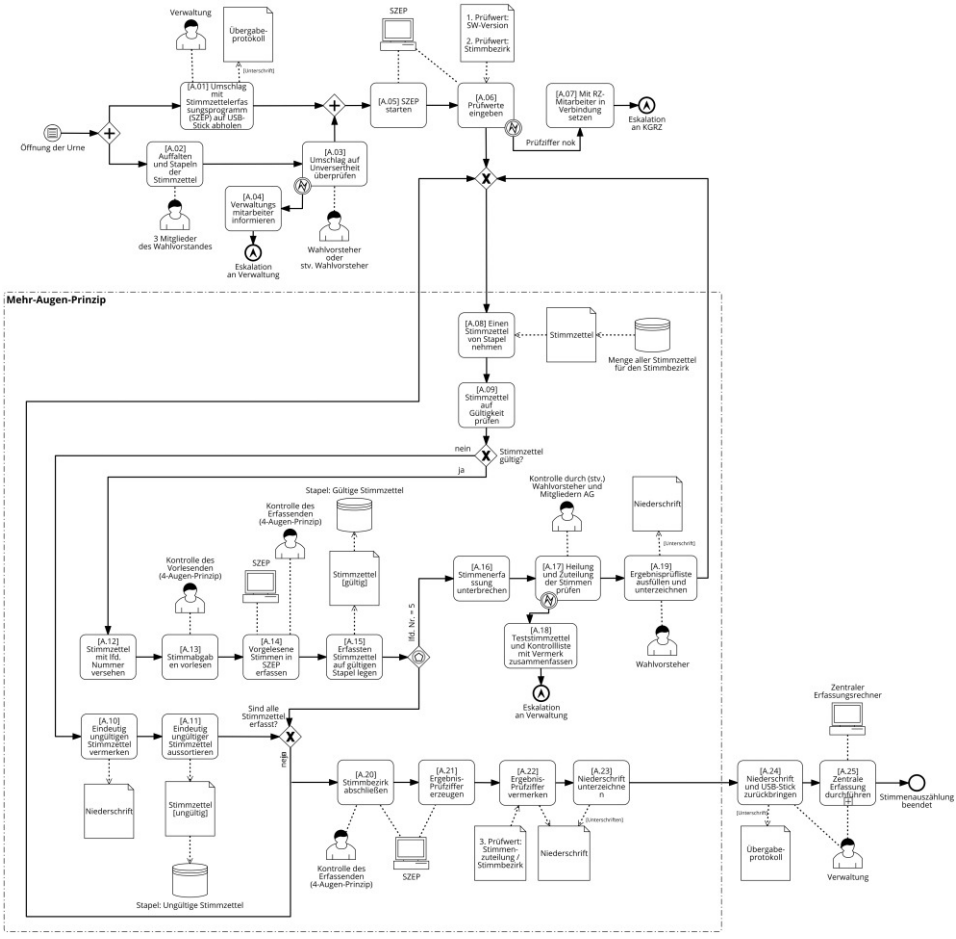


Abbildung 1: Die elektronisch unterstützte Ergebnisermittlung als Prozessverlauf

Die Auszählung[13] beginnt bei dem vorliegenden Teilprozess mit der Öffnung der Urne an dem Tag nach der Wahl durch den Wahlvorsteher[14]. Die Stimmzettel werden von drei Mitgliedern des Wahlvorstandes zunächst aufgefaltet und gestapelt. Ein weiteres Mitglied des Wahlvorstands lässt sich den USB-Stick mit dem Stimmzettelerfassungsprogramm (SZEP) von Mitarbeitern der Verwaltung aushändigen. Dieser Vorgang wird in einem Übergabeprotokoll festgehalten (Name, Funktion der Person). Bevor der Umschlag geöffnet und der Datenträger durch den Wahlvorsteher

[13] Die Ausstattung und Lage der Arbeitsplätze zur Auszählung je Stimmbezirk sind in der „Dienstanweisung für den Einsatz von Personal-Computern“ geregelt.

[14] Die Urnen inklusive der begleitenden Dokumente wurden zuvor durch zwei Mitarbeiter der Verwaltung von den einzelnen Wahllokalen in die Rhein-Mosel-Halle transportiert. Dort wurden die Urnen über Nacht unter Bewachung durch ein privates Sicherheitsunternehmen verwahrt. Die Schlüssel für die Urnen befinden sich währenddessen in Besitz des jeweiligen Wahlvorstehers.

oder dessen Stellvertreter entnommen wird, ist der Umschlag auf seine Unversehrtheit zu überprüfen[15]. Sind Beschädigungen zu erkennen, dann muss die Verwaltung informiert und der Datenträger ausgetauscht werden.

Das Einlesen der Stimmzettel erfolgt gesondert für jeden Wahlbezirk auf Stand-Alone-Rechnern. Das Software-Modul „Heiler" ist wahlbezirksspezifisch und wird direkt vom USB-Stick gestartet. Bevor Stimmen erfasst werden können, müssen zwei Prüfziffern (PZ) eingegeben werden. Die erste PZ garantiert die Verwendung der freigegebenen Programmversion. Die zweite PZ stellt den korrekten Wahlbezirk sicher. Die Bedienung erfolgt durch die sogenannte Arbeitsgruppe, welche sich als Teilmenge aus dem sechsköpfigen Wahlvorstand bildet und gemäß [KWO83, §55b] aus mindestens drei Personen besteht.

Initial wird die korrekte Erfassung der Stimmzettel getestet: Dazu werden die ersten fünf Stimmzettel im Programm erfasst. Anschließend prüft die Arbeitsgruppe im Beisein des Wahlvorstehers, ob das SZEP die Heilung und Zuteilung der Stimmen korrekt durchführt[16]. Das Ergebnis des Testlaufs wird in der Niederschrift dokumentiert und vom Wahlvorsteher unterschrieben. Anschließend erfolgt die Erfassung der restlichen Stimmzettel. Eindeutig ungültige Stimmzettel nach §37 KWG werden von der Arbeitsgruppe aussortiert und im Protokoll vermerkt. Diese werden nicht elektronisch erfasst. Alle anderen Stimmzettel werden von einem Mitglied der Arbeitsgruppe laut vorgelesen und von einem zweiten Mitglied in identischer Form in das SZEP übernommen. Sowohl der „Vorlesende" als auch der „Erfasser" werden durch je eine Person kontrolliert. Sowohl der Papierstimmzettel als auch sein digitales Äquivalent werden mit einer laufenden Nummer versehen. Damit kann jeder im SZEP erfasste Stimmzettel im Nachhinein mit dem Original verglichen werden (Nachvollziehbarkeit). Nach der Erfassung des letzten Stimmzettels erfolgt die abschließende Speicherung der Stimmauszählung. Dieser Vorgang erzeugt die dritte Prüfziffer – die sog. Ergebnisprüfziffer, welche auf der Niederschrift vermerkt wird. Diese Prüfziffer garantiert die Integrität der erfassten Ergebnisse. Die Ergebnisse der jeweiligen Wahlbezirke werden mittels USB-Sticks von dem Wahlvorsteher zur zentralen Erfassung an Mitarbeiter der Verwaltung übergeben. Die zentrale Erfassung erfolgt an einem separaten Rechner, welcher mit dem Internet verbunden ist. Die Mitarbeiter der Verwaltung nehmen die Wahlniederschrift inklusive aller begleitenden Dokumente und den USB-Stick entgegen und protokollieren den Erhalt dieser Dokumente.

[15] Gemäß Bescheid über die Freigabe des Stimmenauswertungsprogramms des Landeswahlleiters, vgl. http://www.wahlen.rlp.de/kw/bek/20140212_Bekanntmachung_Freigabe_PCWahl_Staatsanzeiger.pdf (Abrufdatum: 16.04.2014)

[16] Endet der Testlauf nicht wie erwartet, erfolgt eine Meldung an die Verwaltung. Es werden dann sowohl die Teststimmzettel als auch das Kontrolldatenblatt mit einem entsprechenden Vermerk der Wahlniederschrift beigefügt. Die Verwaltung nimmt anschließend eine Überprüfung in eigener Zuständigkeit vor. Dieser Vorgang wird an den Wahlausschuss berichtet.

5 Identifikation sicherheitskritischer Aspekte

Im Folgenden wird der im vorherigen Abschnitt beschriebene Teilprozess bezüglich der Stimmauszählung hinsichtlich sicherheitskritischer Aspekte nach der in [Sim14] beschriebenen Vorgehensweise analysiert. Die Strukturierung der nachfolgenden Unterabschnitte orientiert sich an den Arbeitsschritten der verwendeten Methodik.

5.1 Zieldefinition

Zunächst erfolgt eine informelle Festlegung der zu betrachtenden Problemstellung. Diese sollte einen Prozess (d.h. eine zu erfüllende Anforderung) adressieren und kann z. B. für Interessenkonflikte oder einzelne Sicherheitsaspekte konkretisiert werden. Für den zu untersuchenden Teilprozess könnte sie lauten: „An welchen Stellen kann der elektronisch unterstützte Auszählungsvorgang durch die am Prozess beteiligten Personen gefährdet sein?“.

5.2 Modellierung des Prozesses

Sobald der Untersuchungsgegenstand festgelegt worden ist, wird der zu betrachtende Prozess z. B. im Rahmen von Prozessworkshops erarbeitet und modelliert. Der dem Teilprozess zugrunde liegende Hauptprozess („Durchführung der Stadtratswahl Koblenz 2014“) wurde in Zusammenarbeit mit dem Ordnungsamt Koblenz auf Basis eines Referenzprozessmodells konkretisiert. Die detaillierte Erfassung des Teilprozesses erfolgte anschließend mit Hilfe von Experteninterviews.

5.3 Identifikation der globalen Interessenlagen

Anschließend werden für den zu untersuchenden Prozess die globalen Interessenlagen[17] der beteiligten Anwender sowie eines potenziellen externen Angreifers erfasst. Stark abstrahiert sind vier Fälle möglich:

- Es existiert kein externer Angreifer und die Prozessbeteiligten agieren regelkonform.
- Die Prozessbeteiligten agieren regelkonform; ein externer Angreifer versucht das Prozessziel zu manipulieren[18].
- Es findet ein Kompromittieren des Prozesses durch einen oder Kooperation mehrerer interner Angreifer statt.
- Es findet eine Kooperation zwischen Innen- und Außentäter statt.

An dem zu betrachtenden Teilprozess (Auszählung der Stimmen) sind die sechs Mitglieder des Wahlvorstands sowie optionale Hilfspersonen beteiligt. Der betrachtete Prozess ist solange nicht gestört, wie die Interessen der Beteiligten in der regelkonformen Auszählung der Stimmen bestehen. Umgekehrt kann der Prozessverlauf

[17] Die „globale Interessenlage“ bezieht sich auf den Hauptprozess; eine „lokale Interessenlage“ ist spezifisch für einen Teilprozess bzw. eine Aktivität. Globale und lokale Interessenlagen unterscheiden sich hinsichtlich ihres Abstraktionsgrades (global = „Manipulation der Wahl“ vs. lokal = „Manipulation bei der Auszählung“).
[18] Für diese Untersuchung ist die Motivation der unterschiedlichen Interessenlagen nicht von Bedeutung.

gestört werden, sobald einer oder mehrere der am Prozess Beteiligten versuchen, ihre eigenen Interessen durchzusetzen. Beispielsweise könnte ein Mitglied des Wahlvorstands bestochen worden sein, Stimmen für einen bestimmten Kandidaten zu unterschlagen. Die auf den Interessenkonflikten von Innentätern basierenden Bedrohungen, welche den Prozessverlauf stören, wurden mittels einer Analyse der Aktivitäten identifiziert.

5.4 Zerlegung des Prozessmodells

Da das Abstraktionsniveau des erfassten Hauptprozesses zu groß ist und daher detaillierte Abläufe noch nicht sichtbar sind, erfolgt zunächst dessen Zerlegung in Teilprozesse. Der im Rahmen dieser Arbeit zu untersuchende Teilprozess wird weiter zerlegt, bis dass die einzelnen durchzuführenden Aufgaben ersichtlich werden. Diese werden als Aktivitäten bezeichnet und stehen für die kleinstmöglichen Elemente von Prozessen. Sie haben den Charakter von Handlungsanweisungen. Beispielsweise ergibt sich aus dem Hauptprozess „Durchführung der Stadtratswahlen Koblenz 2014“ u.a. der Teilprozess „Wahlergebnis ermitteln“. Dort wird die elektronische Erfassung der Stimmzettel als Aktivität modelliert. Bei genauerer Betrachtung kann diese als weiterer Unterprozess mit den entsprechenden Aktivitäten modelliert werden. Was als Aktivität oder Sub-Prozess aufgefasst wird, ist also abhängig von der Zieldefinition. Der im Folgenden zu untersuchende Teilprozess wurde in Abschnitt 4 modelliert und detailliert erläutert.

5.5 Aktivitätsanalyse

Zunächst wird jede Aktivität unter Berücksichtigung der angebundenen Prozessbausteine wie beteiligte Rollen (hier: Wahlvorstand, Verwaltung), verwendete IT-Systeme (hier: SZEP auf USB-Stick, Desktop-PC, netzwerkfähiger zentraler Erfassungsrechner), Datenobjekte (hier: Stimmzettel, Protokolle) etc. beschrieben. Für jede Aktivität wurden mögliche Bedrohungen[19] in abstrahierter Form identifiziert (vgl. Tab. 1). Diese richten sich gegen das Prozessziel (global) bzw. davon abgeleitet auf das Aktivitätsziel (lokal) und basieren auf den in 5.3 identifizierten Interessenkonflikten.

Bedrohung [B.n]	**Aktivität [A.n]**
[B.01]: Umschlag mit USB-Stick (SZEP) wird durch Mitglied des Wahlvorstands oder Verwaltungsmitarbeiter unterschlagen	[A.01], [A.24]
[B.02]: Umschlag mit USB-Stick (SZEP) wird durch Mitglied des Wahlvorstands oder Verwaltungsmitarbeiter manipuliert	[A.01], [A.24]

[19] Eine Abgrenzung der Begriffe Bedrohung, Angriffstechnik und Angriff wird in [Gr+14] vorgenommen.

[B.03]: Unterschlagen von Stimmzetteln („löschen“)	[A.02], [A.08], [A.09], [A.10], [A.11], [A.12], [A.13], [A.14], [A.15], [A.24]
[B.04]: Hinzufügen von Stimmzetteln	[A.02], [A.08], [A.09], [A.10], [A.13], [A.14], [A.15], [A.24]
[B.05]: Manipulation bestehender Stimmzettel (Kreuz hinzufügen, Streichung vornehmen, ungültig markieren, falsche lfd. Nr., etc.)	[A.02], [A.08], [A.12], [A.13], [A.14], [A.15], [A.24]
[B.06]: Bewusste Falschmeldung vornehmen (Anwender)[20]	[A.03], [A.17], [A.19], [A.20], [A.22], [A.23]
[B.07]: Unterschlagen einer erforderlichen Meldung (Anwender)[21]	[A,03], [A.17], [A.19], [A.20], [A.22], [A.23]
[B.08]: Nicht freigegebene Software wird verwendet	[A.05], [A.25]
[B.09]: Das System gibt fälschlicherweise eine Fehlermeldung aus	[A.06], [A.14], [A.20], [A.21], [A.25]
[B.10]: Das System gibt fälschlicherweise keine Fehlermeldung aus	[A.06], [A.14], [A.20], [A.21], [A.25]
[B.11]: SZEP erzeugt falsche Prüfziffer	[A.21]

Tabelle 1: Bedrohungen (abstrahiert), welche die Auszählung beeinflussen können; ohne Berücksichtigung eingesetzter Sicherheitsmaßnahmen

Die von den Prozessbeteiligten durch ihre Interessenkonflikte verursachten Bedrohungen und zu deren Vermeidung getroffene Maßnahmen sollen nachfolgend betrachtet werden:

[B.01] Umschlag mit USB-Stick (SZEP) wird durch Mitglied des Wahlvorstands oder Verwaltungsmitarbeiter unterschlagen: Die Abholung des Umschlags mit dem SZEP erfolgt durch ein Mitglied des Wahlvorstands. Dieses Mitglied hat die Möglichkeit, den Umschlag verschwinden zu lassen bzw. am Auszählungsplatz dessen Erhalt zu bestreiten. Ein Erfolg dieses Angriffs ist nicht gegeben, da bei Bedarf jederzeit ein neuer stimmbezirksspezifischer USB-Stick erstellt werden kann. Ferner wird die Abholung durch die Mitarbeiter der Verwaltung protokolliert, so dass ein Verlust durch den Abholenden ihm zugerechnet werden kann. Bei einem Unterschlagen durch die ausgebende Stelle ist das Ziel fraglich. In diesem Fall würde der abholende Wahlhelfer auf Ausgabe des Umschlages beharren und gegebenenfalls existierende Möglichkeiten

[20] Resultiert in der Verzögerung des Prozessablaufs z. B. indem fälschlicherweise behauptet wird, dass der ausgegebene Umschlag beschädigt gewesen sei oder indem eine falsche Unterschrift vorgenommen wird.
[21] Entspricht dem Gegenteil von [B.07], da hier der Prozessverlauf bewusst kompromittiert wird, z. B. aufgrund von Mitwisserschaft oder auch, um den Prozess zu verzögern (fehlende Unterschrift).

der Eskalation wahrnehmen. Auch eine Kooperation beider wäre nicht zielführend, da in diesem Falle die anderen Mitglieder der Arbeitsgruppe intervenieren könnten.

[B.02] Umschlag mit USB-Stick (SZEP) wird durch Mitglied des Wahlvorstands oder Verwaltungsmitarbeiter manipuliert: Der Austausch des Umschlags ist durch den Verwaltungsmitarbeiter oder durch den Abholenden auf dem Weg von der Ausgabestelle bis zum Auszählungsplatz möglich. Ein Austausch impliziert einen vorbereiteten Angriff; d.h. es wurde ein nicht authentisches oder manipuliertes SZEP bereits im Vorfeld auf einem USB-Stick installiert. Dieser sollte von seiner äußeren Beschaffenheit so sein, dass er einer visuellen Prüfung standhält. Gleiches gilt für den Umschlag, damit die Integritätsprüfung durch den Wahlvorstand erfolgreich ist. Dieser Angriff schlägt fehl, sobald der Abgleich mit der öffentlich bekanntgegebenen Prüfziffer am Arbeitsplatz erfolgt. Auch eine Kooperation beider Akteure ist nicht zielführend, so lange der Abgleich der Prüfziffer innerhalb der Arbeitsgruppe korrekt erfolgt.

[B.03] Unterschlagen von Stimmzetteln („löschen"): Das Löschen von Stimmzetteln betrifft sowohl Papier- als auch elektronische Stimmzettel. Papierstimmzettel könnten z. B. für ungültig erklärt werden, obwohl sie gültig sind. Es ist auch denkbar, dass die Erfassung bewusst unterbleibt. Ebenso ist es nach der Auszählung und dem Abschließen des Stimmbezirks möglich, Stimmzettel zu löschen. Die genannten Angriffe werden mit Ausnahme der systemseitigen Löschung durch das Mehr-Augen-Prinzip unterbunden. Die Manipulation der elektronischen Auszählungsergebnisse wird bei der zentralen Konsolidierung aufgedeckt: Nach Abschluss des Stimmbezirks wird eine Prüfziffer generiert, die auf der Wahlniederschrift erfasst wird. Eine nachträgliche Änderung der Ergebnisse ist möglich; diese resultiert aber in einer neuen Prüfziffer. Da bei der zentralen Erfassung die Prüfziffern abgeglichen werden, wird an dieser Stelle eine Manipulation auffallen.

[B.04] Hinzufügen von Stimmzetteln: Das Hinzufügen von Papierstimmzetteln ist dann möglich, wenn am Ort der Auszählung nicht verwendete Stimmzettel verfügbar sind. Der Angriff kann als besonders erfolgreich gewertet werden, wenn der Angreifer zusätzlich Zugang zu dem Auszug des Wählerverzeichnisses hat und er zu nachträglich ausgefüllten Stimmzetteln einen entsprechenden Vermerk machen kann. Das Hinzufügen elektronischer Stimmzettel ist ebenfalls möglich; bei nachträglicher Kontrolle wird aber das Fehlen der korrespondierenden Papierstimmzettel auffallen. Beide Angriffe sind aufgrund des Mehr-Augen-Prinzips zu vernachlässigen. Eine Nacherfassung ist aufgrund der unter [B.03] genannten Aspekte nicht durchsetzbar.

[B.05] Manipulation bestehender Stimmzettel: Im Gegensatz zu [B.04] gestaltet es sich wesentlich einfacher, bestehende Stimmzettel ungültig zu markieren bzw. nicht vergebene Stimmen nachträglich zu verteilen. Dies gilt sowohl für Papier- als auch für elektronische Stimmzettel. Auch hier unterbinden das Mehr-Augen-Prinzip sowie die Kontrolle der Prüfziffern bei der Erfassung den Erfolg dieses Angriffs.

[B.06] Bewusste Falschmeldung vornehmen (Anwender): Diese Bedrohung ist nicht sicherheitsrelevant, da sie lediglich in einer Verzögerung des Prozesses resultiert. Würde z. B. fälschlicherweise ein beschädigter Umschlag gemeldet werden, so würde der

Prozess solange stoppen, bis dass ein Ersatz-USB-Stick generiert und ausgegeben wird. Gleiches gilt für falsche Unterschriften – sofern eine solche auffällt bzw. gemeldet wird, liegt es im Ermessen der Verwaltung, wie mit einem solchen Fall umgegangen wird. Eine Neuzählung ist in solchen Fällen nicht zu erwarten.

[B.07] Unterschlagen einer erforderlichen Meldung (Anwender): Im Gegensatz zu [B.06] ist diese Bedrohung sicherheitsrelevant. Z. B. wird im Rahmen der Integritätsprüfung festgestellt, dass der ausgegebene Umschlag beschädigt ist und dieser Umstand würde nicht eskaliert. Dann würde die Verwendung des manipulierten SZEP bei dem Abgleich der Prüfziffern auffallen. Auch ist es denkbar, dass die Aufzeichnung der Prüfziffer auf der Niederschrift nach Abschluss des Stimmbezirks bewusst unterbleibt, um eine Nacherfassung von Stimmen zu forcieren. Zur Vermeidung solcher bewussten und unbewussten Unterlassungen dient das Mehr-Augen-Prinzip.

[B.08] Nicht freigegebene Software wird verwendet: Bei dieser Bedrohung für [A.05] und [A.25] relevanten Bedrohung müssen zwei Fälle unterschieden werden. Im ersten Fall handelt es sich um eine versehentliche Herausgabe einer nicht freigegebenen Software-Version bzw. den Versuch, das freigegebene SZEP („Heiler"-Modul) durch ein anderes Programm oder eine nicht-freigegebene Version zu ersetzen, um so die Stimmerfassung zu manipulieren (vgl. [B.02]). Durch den Abgleich der Prüfziffer werden beide Ereignisse bei der Arbeitsgruppe auffallen. Weiterhin kann es sich bei dieser Bedrohung um den Versuch handeln, das Software-Modul „Berechnung der Sitzverteilung" zu manipulieren bzw. durch ein anderes Programm zu verwenden. Dieser Fall ist als kritisch anzusehen. Dieses Software-Modul läuft auf einem zentralen Netzwerkrechner, und dient der Zusammenführung der Ergebnisse aus den einzelnen Wahlbezirken. Diese Teilergebnisse werden mittels USB-Stick vom Wahlvorstand des jeweiligen Wahlbezirks an Mitarbeiter der Verwaltung übergeben und vom USB-Stick auf dem zentralen Rechner eingelesen und zur Ermittlung der finalen Sitzverteilung mittels Software ausgewertet. Ein manipuliertes Programm würde es somit ermöglichen, das Wahlergebnis zu verändern. Sollte dies durch keine weiteren Sicherheitsmaßnahmen abgesichert sein (z.B. Prüfziffer für freigegebene Programmversion, Testlauf zur Überprüfung der korrekten Funktionsweise der Berechnung), bliebe diese Manipulation unbemerkt.

Die Funktionalität der verwendeten Software wird von den Bedrohungen **[B.09] – [B.11]** adressiert. Diese Bedrohungen müssen im Vorfeld durch eingehende und fachmännische Analyse der zu verwendenden Software ausgeschlossen werden, so dass diese Bedrohungen für die freigegebene Softwareversion ausgeschlossen werden können.

6 Fazit

Im Rahmen einer Kooperation mit der Stadt Koblenz wurde die Stadtratswahl 2104 modelliert und analysiert. Dieses Prozessmodell spezifiziert vier Teilprozesse: Wahlvorbereitung, Stimmabgabe, Stimmauszählung und Wahlnachbereitung. Von diesen vier Teilprozessen werden die Prozesse der Wahlvorbereitung und der Stimmauszählung mittels elektronischer Hilfsmittel unterstützt. Hier wurde exemplarisch

der Teilprozess der Stimmauszählung detailliert spezifiziert und hinsichtlich sicherheitskritischer Aspekte analysiert. Dabei wurde methodisch in Anlehnung an [Sim14] vorgegangen.

Die in Abschnitt 5 beschriebene Prozessanalyse identifiziert eine Menge von möglichen Bedrohungen. Es zeigt sich jedoch, dass diese Bedrohungen durch organisatorische Sicherheitsmaßnahmen wie dem Mehr-Augen-Prinzip weitestgehend ausgeschlossen bzw. deren Risiko hinreichend minimiert werden können. Voraussetzung ist hier jedoch, dass innerhalb der Mehr-Augen-Gruppe mindestens ein Mitglied integer ist. Die Eingabe der Stimmzettel erfolgt beispielsweise so, dass eine Person den Stimmzettel vorliest und dabei durch eine weitere Person kontrolliert wird. Um die Integrität des Ergebnisses garantieren zu können, muss demnach entweder der Vorlesende oder der diesen Kontrollierende integer sein, besser noch beide. Dieser Zusammenhang lässt sich formal mittels des sogenannten k-Resilience Term ausdrücken [VG09]. Die Formulierung des k-Resilience Terms anhand des Prozessmodells ist ein Anknüpfungspunkt für zukünftige Arbeiten.

Als sicherheitskritisch konnte die Berechnung der Sitzverteilung auf dem zentralen Netzwerkrechner identifiziert werden (Aktivität [A.25] in Abb.1). Hier ist unklar, ob von Seiten der Wahlorganisatoren weitere Sicherheitsmaßnahmen ergriffen werden, um die ausschließliche Verwendung der freigegebenen Software-Version sicherzustellen, und ob Test- und/oder Kontrollroutinen existieren, um deren korrekte Funktionsweise zu gewährleisten. Hier besteht nach Meinung der Autoren sowohl wegen der Anbindung des Rechners an das Internet und der damit verbundenen Bedrohung durch einen externen Netzwerkangreifer sowie des Black-Box-Charakters der Berechnung das größte Manipulationspotenzial.

Weiterhin erfolgte die Prozessanalyse im Rahmen dieses Artikels nur auszugsweise für den Teilprozess der Stimmauszählung. Im Rahmen der Kooperation mit der Stadt Koblenz wurde jedoch der gesamte Ablauf der Stadtratswahl modelliert und analysiert. Hier zeigt sich, dass über den gesamten Prozess gesehen die öffentliche Nachvollziehbarkeit nicht durchgängig gewährleistet werden kann. Die Stimmabgabe sowie die Stimmauszählung sind in vollem Maße öffentlich nachvollziehbar. Die verschlossenen Urnen und die in der Urne enthaltene Wahlniederschrift werden von zwei Verwaltungsmitarbeitern gegen 18:00 Uhr am Wahltag vom Wahllokal in die Rhein-Mosel-Halle transportiert. Dort werden die Urnen zentral unter Bewachung durch eine private Sicherheitsfirma bis zur Auszählung am darauffolgenden Tag verwahrt. In diesem Zeitfenster ist keine öffentliche Kontrolle möglich. Das Öffnen der Urnen und der Austausch des Inhalts (abgegebene Stimmen und/oder Wahlniederschrift) wären demnach nicht erkennbar, insofern die Urne vor Abtransport aus dem Wahllokal nicht erneut versiegelt würde. Weiterhin könnte ein derartiger Angriff den Ausgang der gesamten Wahl beeinflussen. Dies wäre ein Angriff der unabhängig von dem Einsatz elektronischer Hilfsmittel rein auf Grund der Ablauforganisation erfolgen könnte, allerdings wie bei vollständig elektronisch durchgeführten Wahlen auf das gesamte Wahlergebnis skalieren. Hier wären weitere organisatorische Sicherheitsmaßnahmen oder eine Umstrukturierung innerhalb des Prozessablaufes zu diskutieren.

Abschließend lässt sich feststellen, dass das hier auszugsweise vorgestellte Prozessmodell das Verständnis für die komplexen Abläufe der Wahl vertieft und dazu beiträgt, sicherheitskritische Punkte zu identifizieren.

Literaturverzeichnis

[AMA12] Altuhova, O.; Matulevičius, R.; Ahmed, N.: Towards Definition of Secure Business Processes. LNBIR: Workshop on Information Systems Security Engineering, 1-15, 2012.

[BHM08] Backes, M.; Hritcu, C.; Maffei, M.: Automated verification of remote electronic voting protocols in the applied pi-calculus. CSF, pages 195-209, 2008.

[Br+13] Bräunlich, K.; Grimm, R.; Richter, P., Roßnagel, A.: Sichere Internetwahlen - Ein rechtswissenschaftlich- informationstechnisches Modell, Nomos-Verlag, 2013.

[BSI11] Bundesamt für Sicherheit in der Informationstechnik: IT-Grundschutzkataloge. Bundesanzeiger-Verlag, Köln, 12. Ergänzungslieferung, 2011.

[CC12] Common Criteria and Common Evaluation Methodology, Version 3.1, 2012.

[Ci+11] Ciuciu, I.; Zhao, G.; Mülle, J.; von Stackelberg, S.; Vasquez, C.; Haberecht, T.; Meersman, R.; Böhm, K.: Semantic Support for Security-Annotated Business Process models. LNBIP, Vol. 81, 284-298, 2011.

[CoE04] Council of Europe: Legal, Operational and Technical Standards for e-voting, recommendation rec(2004)11 adopted by the Committee of Ministers of the Council of Europe and Explanatory Memorandum. Strasbourg, 2004.

[DK06] Dzhendova, G.; Kalmring, D.: Modellierung von IT-Sicherheit. Analyse und Synthese. Proc. DACH Security, 2006.

[DKR09] Delaune, S., Kremer, S., & Ryan, M.: Verifying privacy-type properties of electronic voting protocols. Journal of Computer Security, 17(4):435-487, 2009.

[Fi06] Fischermanns, G.: Praxishandbuch Prozessmanagement. 6. Auflage, Verlag Dr. Götz Schmidt, Gießen, 2006.

[Gr+14] Grimm, R.; Simić-Draws, D.; Bräunlich, K.; Kasten, A.; Meletiadou, A.: Referenzmodell für ein Vorgehen bei der IT-Sicherheitsanalyse. Zur Veröffentlichung angenommen in: Informatik Spektrum 2014.

[HPR93] Hammer, V.; Pordesch, U.; Roßnagel, A.: KORA. Eine Methode zur Konkretisierung rechtlicher Anforderungen zu technischen Gestaltungsvorschlägen für Informations- und Kommunikationssysteme. Infotech 1/1993, S. 21ff, 1993.

[JP06] Jonker, H. L.; Pieters, W.: Receipt-freeness as a special case of anonymity in epistemic logic. Proc. IAVoSS Workshop On Trustworthy Elections, 2006

[Kr02] Krimmer, R.: e-Voting.at: Elektronische Demokratie am Beispiel der österreichischen Hochschülerschaftswahlen. Working Paper 05/2002 des Instituts für Informationsverarbeitung und -wirtschaft, Wirtschaftsuniversität Wien, 2002.

[KR05] Kremer, S.; Ryan, M.: Analysis of an electronic voting protocol in the applied pi calculus. ESOP, pages 186-200, 2005.

[KRS10] Kremer, S., Ryan, M.; Smyth, B.: Election verifiability in electronic voting protocols. ESORICS, pages 389-404, 2010.

[KWG94] Kommunalwahlgesetz online verfügbar unter http://landesrecht.rlp.de/jportal/?quelle=jlink&query=KomWG+RP&psml=bsrlpprod.psml [zuletzt eingesehen am 01.04.2014].

[KWO83] Kommunalwahlordnung, online verfügbar unter http://landesrecht.rlp.de/jportal/?quelle=jlink&query=KomWO+RP&psml=bsrlpprod.psml [zuletzt eingesehen am 21.03.2014].

[Lo+05] Lopez, J.; Montenegro, J. A.; Vivas, J. L.; Okamoto, E.; Dawson, E.: Specification and Design of Advanced Authentication and Authorization Services. Computer Standards & Interfaces 27(5), 467-478, 2005.

[MBH14] Middelhoff, M.; Böhle, C.; Hellingrath, B.: Modeling and Analyzing Information Security in Secure Logistics Business Processes. Proc. MKWI 2014, 1925-1926, 2014.

[MHS05] Mead, N. R.; Hough, E. D.; Stehney II, T. R.: Security Quality Requirement Engineering (SQUARE) Methodology. Technical Report CMU/SEI-2005-TR-009. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA, 2005.

[Pa+12] Paja, E.; Giorgini, P.; Paul, S.; Meland, P. H.: Security Requirements Engineering Support for Security-Annotated Business Processes. LNBIP, 77-89, 2012.

[RFP07a] Rodríguez, A.; Fernández-Medina, E.; Piattini, M.: A BPMN Extension for the Modeling of Security Requirements in Business Processes. Proc. IEICE Trans. Inf. & Syst., Vol. E90-D, No. 4, 2007.

[RFP07b] Rodríguez, A.; Fernández-Medina, E.; Piattini, M.: M-BPSec: A Method for Security Requirement Elicitation from a UML 2.0 Business Process Specification. LNI Vol. 4802, 106-115, 2007.

[Ru09] Rupp, C.: Requirements Engineering und Management. Hanser, München, 2009.

[Sim+13] Simić-Draws, D., Neumann, S.; Kahlert, A.; Richter, P.; Grimm, R.; Volkamer, M.; Roßnagel, A.: Holistic and Law Compatible IT Security Evaluation: Integration of Common Criteria, ISO 27001/IT-Grundschutz and KORA. In: IJISP, 7(3), 16-35, 2013.

[Sim14] Simić-Draws, D.: Ein Vorgehensmodell zur Durchführung einer prozessorientierten Sicherheitsanalyse. Proc. MKWI 2014, 1889–1896, 2014.

[SS08] Schmelzer, H. J.; Sesselmann, W.: Geschäftsprozessmanagement in der Praxis. Hanser, München, 2008.

[VG09] Volkamer, M.; Grimm, R.: Determine the Resilience of Evaluated Internet Voting Systems, In: IEEE CS Digital Library, doi: 10.1109/RE-VOTE.2009.2, First International Workshop on Requirements Engineering for E-Voting Systems, p. 47-54, 2009

[Vie05] Viega, J.: Building security requirements with CLASP. Proc. Workshop on software engineering for secure systems (SESS), 1-7, 2005.

[VV08] Volkamer, M., Vogt, R.: Common criteria protection profile for basic set of security requirements for online voting products. BSI-CC-PP-0037, Version 1.0, retrieved April 18, 2008, from http://www.bsi.bund.de/

Iraqi Elections in 2014: a Privacy Requirement Evaluation Based on a Polling Place Experience

Ali Fawzi Najm Al-Shammari[1] and Adolfo Villafiorita [2]

[1]Computer Science Dept., University of Kerbala, Kerbala, Iraq
[1]Computer Science Dept., Free University of Bolzano, Bolzano, Italy
[1,2]ICT4G, Fondazione Bruno Kessler, Trento, Italy
[1]alifawzi@uokerbala.edu.iq
[2]adolfo.villafiorita@fbk.eu

Abstract: Democracy was not a well known term for the Iraqi people till 2003, when the US and coalition forces begun a military action against the Iraqi regime at that time. After that, the new constitution in Iraq allows the citizen to elect the government, the local governments, and the parliament, every 4 years. For this reason, the Independent High Electoral Commission (IHEC) was established to manage, control, and run the elections. In the first two elections, the voting system was a "regular" paper-based system. Recently, in 2014, the process has been improved by making the voter authorization process electronic. In this paper, we will describe the latest voting system used in the Iraqi election and make some evaluation about the system fulfillments of privacy requirements, based on a polling station experience.

1 Introduction

The voting system used in the parliament Iraqi elections in 2005 and 2009 was a paper-based voting system. This system has the advantage of being usable, since it has a simple voting procedure that allows most of the voters to vote in autonomy. Public verifiability is supported, since the cast votes are tallied publicly in each polling station after closing the poll. Authorised public entities, i.e., observers, monitor this process and can keep a copy of the tallied station result.

In the recent election, in 2014, the IHEC introduced a new component, implemented by Andra [Ind14], to partially automate the voting procedure. More specifically, they added a component that can manage the voter identification process. This component uses a smart card reader and a biometric fingerprint scanner to identify voters. Each voter, therefore, is provided with a smart identification card (SID) in the pre-election phase to be used in the election day as an identifier. The IHEC claims that this improvement can mitigate the risk of double voting, vote changing and vote stuffing, since it allows them to verify that each tallied vote belongs to single eligible voter. This is what we call the "eligibility verifiability" [KRS10].

However, based on the current available resources on this system [IHE14, II14], we were

able to analyse the election process in a polling station. This analysis shows that the system is vulnerable to a set of risks that can compromise the privacy requirements, namely: anonymity, receipt freeness, and coercion resistance. More in details [LSVB09], by compromising *anonymity* we mean that it is possible to know which voter has cast a particular vote. By compromising *receipt freeness* we mean that it is possible for a voter to present evidence about his/her selection. Finally, by compromising *coercion resistance* we mean that there is a way for a third party to control a voter's selection. In this paper, we also define the term *risk* as an exposure to a failure in a system functionality, caused by an attack.

Such risks might cause violations of the Iraqi legislation. For instance, Article 5 of the Iraqi constitution states that "The people are the source of authority and legitimacy, which they shall exercise in a direct, general, secret ballot and through their constitutional institutions" [Com05]. Also, Article 5-3 of the IHEC election law number 13 on 2013 states that a voter must cast a vote secretly [oC13].

In this paper, we describe the voting procedures adopted in the recent elections in Iraq. We start by describing the election components in the poll, the stakeholders, and the voting procedures. After that, we highlight the possible risk scenarios and the assumptions under which these are made possible. Finally, we provide our recommendation to mitigate such risks by proposing a slight modification to procedures and to components.

2 Polling Station Experience

The election in Iraq divides the precincts at a provincial level. Each province has a number of polling places, which are voting centers that contain one or more polling station. The polling station is where voters can cast their votes. In this section, we describe the election procedure in a polling station. First, we describe the station materials and stakeholder, and second the voting procedures.

2.1 Station Materials

There are two types of materials in a polling station: sensitive and non-sensitive. The sensitive materials include those materials that could be used to influence the election result, namely: paper ballot, ballot stamp, voters list, voting ink, station forms, smart card reader device, and supervisor smart card. The non-sensitive materials includes the logistic materials, for instance, the voting cabin, the empty ballot box, pens, and guideline materials. Beside these materials, we mention the voter smart card, since it is used also inside a polling station. More in details:

- **Paper Ballot Pack:** It contains 50 ballots that are serialised with a unique serial number. This number is written on the ballot in two formats. The first in clear text, and the second encoded in a QR code. A voter can vote either by selecting a party

or by selecting a party and a candidate. An example of the ballot is shown Figure 1.

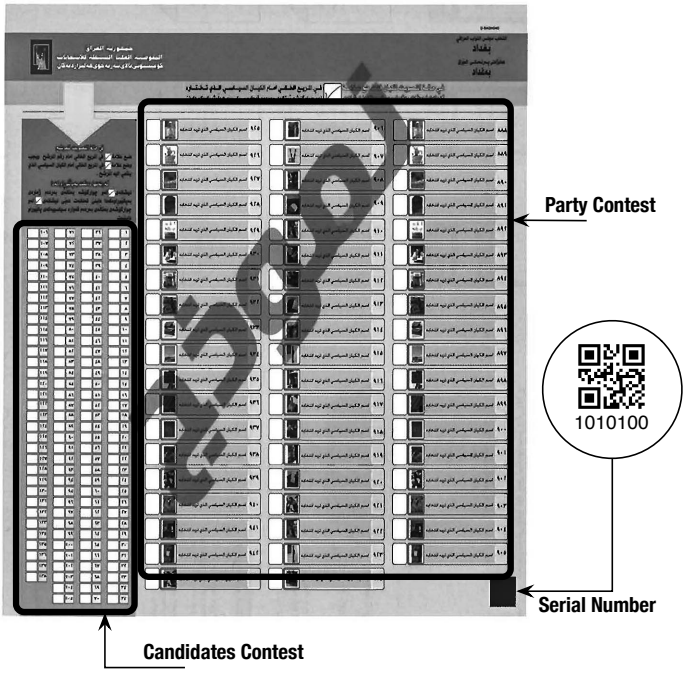


Figure 1: Sample ballot [IHE14].

- **Ballot Stamp:** a manual stamp used to mark the back of a ballot issued to a voter.
- **Voters List:** A hard copy list that contains the names of the eligible voter of a specific station. Each station has a different voters list, since each voter must cast a vote in a dedicated station, that is fixed in the pre-election phase, i.e., during the voter registration process.
- **Voting Ink:** It is used to mark the index finger of any voter who cast a vote. The ink lasts for a few days. Therefore, it considered as a voting evidence that is used to mitigate any malicious try of multiple voting.
- **Station Forms:** hard copy forms used to record voting process output in a station by a station manager. More details about the recorded date is provided in Section 2.3.
- **Voter Smart Identification Card (SID):** a smart card that stores a voter's information, namely: name, date of birth, address, biometric fingerprint, polling place name, and polling station number. This card is used to identify a voter in the election day.

- **Smart Card Reader System (SCRS):** An electronic system that is used to identify eligible voters in a polling station, and also to collect the serial numbers of cast votes to support verifiability. It contains an offline database of the eligible voters. More details about its functionality are provided in Section 2.3.

  Technically, this system comprises three main hardware components, as shown in Figure 2, namely: a tablet provided with a front camera (BQ Maxwell Plus), a thermal printer with a smart card reader (DATECS DPP-250 2"), a fingerprint reader (Futronic FS80). All of these components are gathered in a single box that could be sealed to avoid any unauthorised access. The software of the SCRS is an application installed in Android 4.1 Jelly Bean system.

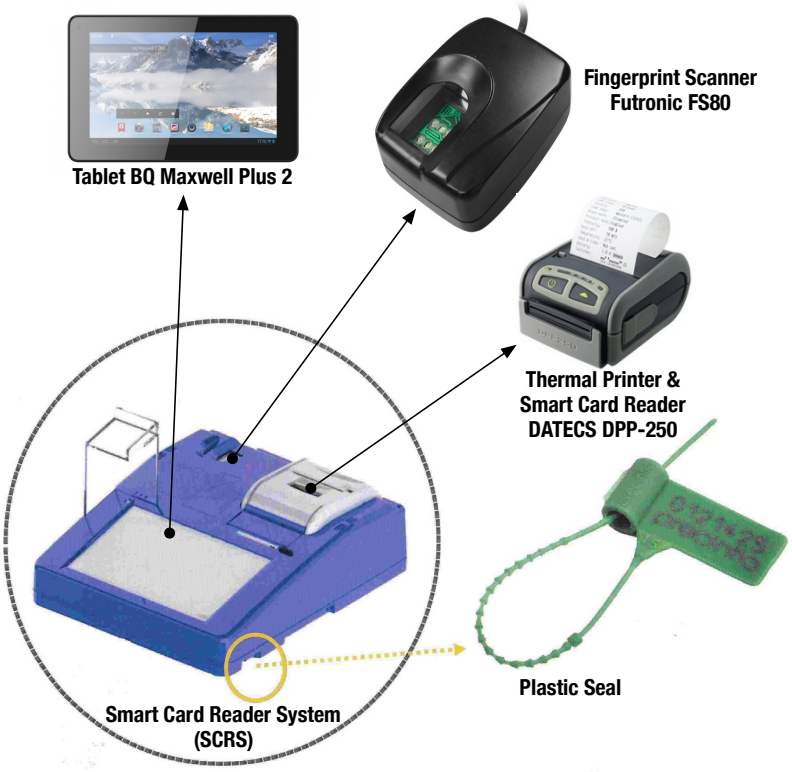


Figure 2: The components of the Smart Card Reader System (SCRS) [II14].

- **Supervisor Smart Card (SSC):** A smart card used to activate the SCRS, and to collect the election data from SCRS after closing the poll.
- **Security Seal:** Plastic seal provided with a unique serial number, used to secure the ballot box, and the SCRS.
- **Ballot Box:** Transparent plastic box where the voter can cast his/her vote.

- **Voting Cabin:** Portable voting cabin where a voter can fill a ballot privately.
- **Secure Plastic Bag:** Special plastic bags used to secure the sensitive materials, e.g., forms, SSC, after closing the poll. They are secure since each one has a unique number and it is not possible to take any material outside the bag without ripping it.

2.2 Stakeholders

Various stakeholders can and need to access a polling station during an election. All of them can positively or negatively impact the security of an election. Understanding who they are and their role is therefore important to reason about the security of an election.

More in details, for the Iraqi election, we distinguish:

- **Polling Place Manager (PPM):** an employee that is responsible for administrating and controlling the general processes in a polling place.
- **Station Manager (SM):** an employee that is responsible for receiving the station materials (sensitive and non-sensitive) in the pre-election phase, installing the station materials before starting the election, administrating voting process during the election, opening and closing the poll, administrating the tallying process after closing the poll, delivering the polling station output, e.g., tallying result, and station materials to the election officials at the end of the election day.
- **Authorization Officer (AO):** an employee that is responsible for verifying voters' eligibility for voting during the election.
- **Ballot Issuer (BI):** an employee that is responsible for issuing paper ballots to eligible voters.
- **Ballot Box Observer (BBO):** an employee that is responsible for observing the ballot box during election, and verifying that each voter marks his/her finger in the permanent voting ink before casting a vote in the ballot box.
- **Technician:** an employee that is responsible for troubleshooting and maintaining the technical components of the SCRSs in a polling place.
- **Election Observer (EO):** an authorised public entity or candidate's agent that is responsible for observing the election process in a polling station during the election day. One or more observers can access a polling station with a maximum limit.

Other authorised people can access a polling station, namely: observer's translators, the press, election officials, and security.

2.3 Election Procedures

We divide the description into three different phases, namely: before opening the station, during the election, and after closing the station.

Before opening the polling station. The day starts when the SM receives a set of ballot packs from the PPM and records the range of their serial numbers in the station form. After that, the SM seals the ballot box using four plastic seals, one for each side, and records their serial numbers. Before opening the poll, the voting equipment is placed in the polling station in a specific configuration: the SCRS and the ballot packs placed on the first side of the room, the voting cabin on the second side, the ballot box and the voting ink on the third side, and finally, the fourth side contains the chairs where the election observers can sit to observe the election.

To open the poll, the AO turns on and starts the SCRS using the SSC. Note that the SCRS is programmed to start only on a specific day and at a specific time.

During the election. A voter arrives at the authorisation desk, where the SCRS is installed, holding his/her SID card and other identification document, e.g., national ID or passport. The AO inserts voter's SID to the SCRS, and scans voter's fingerprint in order to verify his/her eligibility to vote. Voter's data is verified with the database. If the voter is eligible, then the SCRS does the following: displays the voter is authorised to cast a vote, saves the time of the voter's access, blocks the voter's SID (so it cannot be used again in this election), and update the voter's status in the database.

After that, the AO passes the voter's document to the BI, who holds the ballot packs and the voters list. The BI checks the voter's name in the voter's list and issue the paper ballot. Before giving the issued ballot to the voter, the AO take it and scans its QR code with the SCRS. This is done in order to store the issued ballot serial number in the SCRC. After that, the voter takes the ballot to the voting cabin to vote privately, and then cast the vote in the ballot box. The ballot box observer controls that each voter marks his/her finger in the voting ink before casting the ballot.

There is a special case to be discussed here, namely what happens if the SCRS fails in reading the SID of a voter, or if the data in the SID does not correspond to any record in the system database. In this case, the SM checks the existence of the voter's name in the voters list. If it exists, then the voter can vote. In this case, the SM takes the voter's SID, puts it in an envelope, and writes on the envelope the name of the voter, the card serial number, and the reason for collecting the SID. Then, the BI releases a ballot for the voter without being scanned by the SCRS. Also, the voter signs in the voter's list, and the AO signs the back of the ballot with a statement "Smart card was not readable". This is to indicate that the serial number of this ballot has not been scanned by the system.

After closing the station. At the end of the election day, the SCRS stops accepting any voter's SID. At this point, the SM closes the poll and starts the tallying process. In the first

step, the SM secures the ballot box by locking the top hole of the box using a plastic seal, and record its serial number. Then, the station manager asks the AO to store the SCRS data into the SSC and prints the SCRS report. This report includes the polling place name, the polling station name, the total number of eligible voters, the total number of voters who accessed the station, the total number of scanned fingerprints, the total number of the scanned QR codes, the time and date of opening and closing the poll, and finally, the list of scanned ballots serial numbers. The SM updates the station forms with the report data.

After that, the polling officer collects the SSC and the SCRS printed report and puts them in a secure plastic bag. The secure bag serial number is also recorded in the station form. All other sensitive information, e.g., the remaining blank ballot and voting ink, are also secured in plastic bags.

After that, the election employees in the station perform the tallying process publicly. The SM records the count result in the station form. Note that the public observers can keep a copy of this form for verification purposes. However, all of the sensitive materials are secured and sent to the central election office, where the election officials tally and audit the final election result. More specifically, they count the cast votes, compare the result with the station forms, and verify that the serial numbers of the cast votes are the same as the serial numbers of the scanned blank ballots by the SCRS, this to verify that there were no votes stuffed or changed.

Note that, at any time during the election day, the technician of the polling place could be called to solve any technical issue if it occurs.

3 The System Privacy Issues

The IHEC claims that the current system mitigates the chance of compromising the cast votes in the after-election phase. For instance, changing or stuffing a vote requires to hack the data collected from the SCRS (that is, the data in the SSCs and the printed reports) in order to perform a non-detectable attack.

However, in elections there is a trade-off between ensuring the correctness of elections and ensuring secrecy of votes. This is evident, for instance, in the use of ink to mark the finger of voters. While it helps ensure no one can vote twice, at the same time it makes evident to anyone who voted in an election. A similar reasoning can be performed on the new electoral procedures and, based on our evaluation, we found that this system is vulnerable to a set of risks against privacy. More specifically, there are different possible scenarios allow to compromise the anonymity, coercion resistance, and receipt freeness.

Before describing the vulnerabilities, Table 1 recaps the different access privileges stakeholders' have to electoral assets. This is done to understand what actor could compromise or access assets to alter assets or disclose private information. For instance, the table shows that the BI and the AO can read (access) the serial number of each issued ballot, and the name of each eligible voter. This could compromise the privacy, as we will describe in the risks scenarios later.

Stakeholder	Access during the election	Access after closing a station	Access in after election phase
Voter	-Ballot number	-	-
Station Manager	-Eligible Voter's information that has failed in SID verification	-Cast vote (Station)	-
Ballot Issuer	-Ballot number -Voter's information -Voters list data -Eligible Voter's information that has failed in SID verification	-Cast vote (Station)	-
Authorization Officer	-Ballot number -Voter's information - Eligible Voter's information that has failed in SID verification	-Cast vote (Station)	-
Election Official	-SCRS component (Precinct)	-	-Cast votes (Precinct) -SCRS data (Precinct) -Voters list data -Eligible Voter's information that has failed in SID verification
Component Firm	-SCRS component (Precinct)	-	-SCRS component (Precinct)
Technician	-SCRS component (Precinct & Station)	-	-SCRS component (Precinct & Station)

Table 1: The different access abilities for the election stakeholders. Note that the (Station) refers to a station access level, and the (Precinct) refers to a precinct access level.

In the following we show how the different access privileges allows one or more actors to compromise the secrecy of elections. All risks requires one or more assumptions to be in place: therefore, the probability of enacting them varies. Still, it is non-zero and understanding under what conditions these attacks can be performed is a first step to devise procedures which enforce stricter privacy requirements while not compromising any of the traceability requirements that help ensure elections are fair.

Scenario 1. A third party could know how the voter voted. More specifically, a malicious third party could ask a voter to write down the ballot's serial number on a piece of paper during the vote casting process, i.e., in the voting cabin, and provide it as an evidence about how he/she voted. The malicious third party could collaborate with a malicious election official, who has a sufficient access to the ballot box, to reveal the vote of a given ballot serial number.

Scenario 2. A malicious AO could reveal the link between vote and voter. More specifically, when a voter arrives, the AO asks him/her for the SID to be verified by the SCRS. At this time, the AO write down the voter's name on a piece of paper, or memorise it. After that, the AO asks the BI to issue a blank ballot. When the AO scans the ballot QR code using the SCRS, he/she can read the ballot serial number. This means that the AO could also write the ballot number in front of the voter's name in the same paper, or memorise it. There are now different ways and occasions during which the voters' votes can

be revealed. The first is during the vote tallying: the AO could consult his/her list of serial numbers-voter's name to disclose all votes; a systematic approach, however, would be easy to detect. The second is after the tallying process, if an actor has access to the ballots and the list **after** the tallying process. This attack would be more difficult to detect, but also more difficult to enact, since it requires the cooperation of more actors. Notice that this attack could be performed on a subset of voters (and thus become easier to perform), for instance to make vote selling easier to perform.

Scenario 3. A malicious SM, AO, BI, and election official could reveal the vote of voters whose SID failed verification, but who are still eligible to vote. More specifically, if the card fails, he/she asks the BI to check the name of the voter in the voters' list. If the name exists, then the voter is eligible to vote. At this point, a malicious AO and BI could write down or memorise, the name of the voter. After that, the SM collects the failed SID in a secure envelope. At this stage, the SM knows the name of the voter. After the voting process, the voter signs the voters' list, and the BI signs the back of the blank ballot before giving it to the voter. This means that the vote, in this special case, will be cast on a ballot signed on the back and, thus, easily recognizable. This allows any stakeholder who has a sufficient access to the cast votes to recognise the specific ballot.

The SM, AO, and BI can then easily associate vote and voters. Regarding the election officials, since they have the access to the voters' list, the collected failed SID, the SCRD report and the cast votes. This allows them to link the special case vote with its voter, if it happened just one time per a single station. For instance, linking the marked vote with voter's signature on the voters list.

As an additional special case, the voters selling their votes could compromise their SID's on purpose, to ensure their votes get cast in ballots signed by the election officials. This could ensure all the vote sold are easily recognizable, although an easier scenario is just taking a picture of one's vote in the cabin.

Scenario 4. A malicious or hacked SCRS could save data that links the voters' identity and their ballot serial numbers. The malicious data could be saved either in the internal memory of the SCRS tablet component, or in the tablet removable memory. In this case, a malicious election official who has a sufficient access to the malicious data could collaborate with another malicious election official who has a sufficient access to the cast votes, both of them could be the same person, in order link between vote, serial number and voter.

Scenario 5. A malicious or hacked SCRS could use the tablet wireless interface, e.g., the WiFi or the Bluetooth, to broadcast malicious data that links a voter with ballot serial number during the election day. More specifically, when the SCRS reads both of a voter data and ballot QR code, it broadcasts the linking data to a malicious device, e.g., a smart phone with a malicious application, that could be handled by a malicious person nearby. For instance, malicious observer, employee, and security. The secret could be revealed by

providing the collected data to a malicious election official who has a sufficient access to the cast votes.

In our analysis, the root failure causes of the scenarios mentioned above can be traced to two main issues. The first issue is that the SCRS is used both for voter authentication and for ballot issuing. While this simplifies traceability (one voter-one vote), it also gives the possibility for a malicious SCRS to link a voter and his/her ballot, i.e., by linking the voter identification data and ballot serial number. The second issue is that the ballot number is not protected, i.e., it is written in clear text. This makes it possible for a malicious employee and voter to keep this number in order to break the anonymity later by collaborating with a malicious stakeholder who has a sufficient access to the cast votes.

4 Improvement Recommendations

We propose two approaches to mitigate the possible system vulnerabilities. The first is to change the procedure, and the second makes a slight change in the ballot.

Considering the procedures, the scanning of the ballot QR must not be performed during the voter identification process. We recommend that this done after closing the election, during the vote tallying process. This can avoid the possibility of a malicious SCRS to link a vote by ballot, since the votes' serial numbers will be scanned randomly. Furthermore, such approach would not compromise the integrity, since it is not possible to stuff a vote to the box during the election, while the entire day process is observed by the public observers and by the station employees.

The second recommendation is to hide the ballot serial number, in a way that prevents anyone from reading it during the election, and allows the SM to reveal it later, during the tallying process. Hiding the ballot serial number could be performed by using different techniques, for instance scratch to reveal the QR code, similar mechanism used in this system [AR06], or by using invisible ink and marking pen as used by [CCC+08]. Furthermore, we recommend that the ballot number should be randomised instead of being serialised. Since, with a serialised ballot pack, revealing the serial number of a single ballot means revealing the numbers of all ballots in the pack.

Considering our recommendations, the voting process could be described as the following: the voter identifies him/herself using SCRS, receives the ballot with the anonymous code, fills the ballot privately, cast the filled ballot in the ballot box. The design of the ballot must allow the box observer to check that the ballot code is maintained hidden (not revealed) while the ballot folded in voter's hand (i.e., if the voter maliciously revealed the ballot number, then it must be avoided). After closing the election and during the tallying process, the SM reveals the ballots code, using a sufficient tool, and scans them using the SCRS. This would mitigate the possibility of performing the mentioned risk scenarios, while keeping the same level of traceability of the current implementation.

5 Conclusion

Iraqi has recently transitioned to a democracy and it is defining its procedures to ensure elections are fair and all voters are given the possibility of voting in secrecy. Simplifying quite a lot on the research in the area, there is, traditionally, a trade-off between these conflicting requirements: to ensure elections are fair, more traceability on the process is required; to ensure voters' votes are secret, less traceability is usually called upon.

In new democracies, where also the apparatus of the Public Administration needs to be reinforced (consider, for instance, the introduction of identification cards or the maintenance of voters' lists), the Iraqi electoral commission has introduced various novelties to make the improve traceability and fairness of elections.

This is an important step to make elections more credible. However, it also comes with new risks and this paper envisages that the Iraqi voting system has a set of vulnerabilities that could compromise the privacy requirements. The vulnerabilities are caused by two main issues, which are the mix of two critical processes to be performed by the same component, and the use of serialised ballots with a readable serial number. We proposed some system modification to avoid these issues with the consideration of minimal changes. It includes changing the voting procedure slightly, changing the physical ballot (the serial code only), and modifying the component software slightly. These changes will add an extra cost caused by the change in ballot production depending on the used code hiding technique, plus the cost of the software modification.

References

[AR06] Ben Adida and Ronald L. Rivest. Scratch & Vote: Self-contained Paper-based Cryptographic Voting. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*, WPES '06, pages 29–40. ACM, 2006.

[CCC+08] David Chaum, Richard Carback, Jeremy Clark, Aleksander Essex, Stefan Popoveniuc, Ronald L. Rivest, Peter Y. A. Ryan, Emily Shen, and Alan T. Sherman. Scantegrity II: End-to-end Verifiability for Optical Scan Election Systems Using Invisible Ink Confirmation Codes. In *Proceedings of the Conference on Electronic Voting Technology*, EVT'08, pages 14:1–14:13. USENIX Association, 2008.

[Com05] Iraqi Constitutional Committee. Iraqi Constitution. `http://www.iraqinationality.gov.iq/attach/iraqi_constitution.pdf`, 2005.

[IHE14] IHEC. Elections and tallying Procedures for Iraqi Parliament Election 2014 - Training Kit Manual . Iraqi Independent High Electoral Commission, 2014.

[II14] IHEC and Indra. Election Components - Training Kit Manual . Iraqi Independent High Electoral Commission, 2014.

[Ind14] Indra. ELECTORAL PROCESSES INDRA. `http://www.indracompany.com/en/soluciones-y-servicios/solucion/electoral-processes/1361/offering`, 2014.

[KRS10] Steve Kremer, Mark Ryan, and Ben Smyth. Election Verifiability in Electronic Voting Protocols. In *Proceedings of the 15th European Conference on Research in Computer Security*, ESORICS'10, pages 389–404. Springer-Verlag, 2010.

[LSVB09] Lucie Langer, Axel Schmidt, Melanie Volkamer, and Johannes Buchmann. Classifying Privacy and Verifiability Requirements for Electronic Voting, 2009.

[oC13] Iraqi Board of Commissioners. IHEC election law number 13. `http://www.ihec.iq/ihecftp/2014/sys-2014/sys-13.pdf`, 2013.

Online-Partizipation in der EU? - Erste Erfahrungen mit Europäischen Bürgerinitiativen aus Sicht Österreichs

Mag. Robert Stein / Mag. Gregor Wenda, MBA
Abteilung für Wahlangelegenheiten
Bundesministerium für Inneres, Republik Österreich
Herrengasse 7, 1014 Wien, Österreich
robert.stein@bmi.gv.at
gregor.wenda@bmi.gv.at

Abstract: Im Mai 2012 wurden bei der Europäischen Kommission die ersten Europäischen Bürgerinitiativen (EBI) registriert. Gerade die Möglichkeit, Unterstützungsbekundungen für dieses neue partizipative Instrument der Europäischen Union nicht nur auf Papier, sondern auch online zu sammeln, erregte in der Zivilgesellschaft großen Zuspruch. Der Start des „Online-Sammelns" verlief durch technische Pannen jedoch verzögert, die praktische Anwendung der Verordnung über die Bürgerinitiative brachte Regelungslücken ans Licht. Inzwischen liegen in Österreich nach Prüfung von fünf Europäischen Bürgerinitiativen genügend Erfahrungswerte vor, um einen ersten Rückblick zu unternehmen und im Lichte der EU-Evaluierung 2015 Verbesserungsvorschläge zu formulieren.

1 Rückblick

Zeitgleich mit der „Verordnung (EU) Nr. 211/2011 des Europäischen Parlaments und des Rates vom 16. Februar 2011 über die Bürgerinitiative"[1], die ab 1. April 2012 Geltung entfaltete, traten viele innerstaatliche Regelungen in den damals 27 EU-Mitgliedstaaten[2] in Kraft, darunter auch das Europäische-Bürgerinitiative-Gesetz (EBIG)[3] in Österreich. Laut Verordnung ist vorgesehen, dass Unterstützungsbekundungen für durch die Kommission zugelassene Europäische Bürgerinitiativen (EBI) innerhalb eines Zeitfensters von einem Jahr auf vorgegebenen Formularen in Papierform oder via Internet mit einem „Online-Sammelsystem" (*online collection system*) gesammelt werden können. Die „moderne Technologie als Instrument der partizipatorischen Demokratie" sollte, so die Verordnung, „sinnvoll" eingesetzt werden.[4] Dieses Online-Sammeln war ein Novum, das in der Zivilgesellschaft von Anbeginn großen Zuspruch erregte und auch Hoffnungen schürte. Am Rande sei erwähnt, dass die Abgabe einer Unterstützungsbekundung mittels digitaler Signatur[5] in

[1] ABl L 2011/65, 1.
[2] Kroatien kam als 28. Mitgliedstaat am 1. Juli 2013 zur Europäischen Union.
[3] Bundesgesetz über die Durchführung von Europäischen Bürgerinitiativen (Europäische-Bürgerinitiative-Gesetz – EBIG), BGBl I 2012/12 idF BGBl I 2013/115.
[4] Vgl. EBI-Verordnung, Präambel, Rz 14.
[5] Die EBI-Verordnung verlangt lediglich eine fortgeschrittene, keine qualifizierte digitale Signatur.

der Verordnung – zumindest theoretisch – ebenfalls normiert ist.[6] Damit eine Europäische Bürgerinitiative „erfolgreich", also der Verordnung entsprechend, von der Kommission und vom Europäischen Parlament weiterbehandelt wird, sind Mindestquoten an Unterstützungsbekundungen erforderlich: Insgesamt müssen eine Million gültige Unterstützungsbekundungen erzielt werden, wobei als Mindesterfordernis in zumindest sieben Mitgliedstaaten die jeweiligen Mindestquoten (z.B. Österreich 13.500, Deutschland 74.250) überschritten werden muss.

In der Praxis waren die ersten Europäischen Bürgerinitiativen Mitte Mai 2012 registriert, sodass die Organisatorinnen und Organisatoren mit dem Sammeln von Unterstützungsbekundungen beginnen konnten. Da Systeme zur Online-Sammlung „angemessene Sicherheitsmerkmale" aufweisen sollten, „um unter anderem zu gewährleisten, dass die Daten sicher gesammelt und gespeichert werden"[7], wurden zuvor noch detaillierte technische Spezifikationen für Online-Sammelsysteme festgelegt.[8] Entsprechend der Verordnung hatte die Kommission auch eine Open-Source-Software[9] für Online-Sammelsysteme entwickelt[10] und bot diese den Mitgliedstaaten zur Unterstützung der Sammlung auf elektronischem Weg an. Die Software stellt jedoch nur einen Teil des in Verordnung verlangten Online-Sammelsystems dar. Ein solches System muss für jede Europäische Bürgerinitiative angepasst und stets einem der Mitgliedstaaten zur Zertifizierung vorgelegt werden.[11] Die Implementierung solcher Online-Sammelsysteme auf eigenen EDV-Plattformen stellte alle Organisatorinnen und Organisatoren von Beginn an vor nicht unerhebliche Probleme. Abgesehen davon, dass die von der Kommission bereitgestellte Software, wie unter anderem ein in Österreich durchgeführter Test zeigte, erhebliche Mängel aufwies, war es den Organisatorinnen und Organisatoren offensichtlich unmöglich, zertifizierungstaugliche Systeme ins Leben zu rufen. Hierzu hätte es grundlegender und wohl kostenintensiver Fachkenntnisse aus dem IT-Bereich bedurft. Die Kommission kam daher den Organisatoren der einzelnen registrierten Europäischen Bürgerinitiativen insofern entgegen, dass für den Anfang eine „zentrale" Lösung angeboten wurde. Dabei kam die Zertifizierungsstelle des Mitgliedstaates Luxemburg für die jeweilige Implementierung zu Hilfe.[12]

Die Unzukömmlichkeiten bzw. Verzögerungen beim Start der Online-Sammlungen führten rasch zum Unmut von Organisatorinnen und Organisatoren, die unter anderem Sorge äußerten, durch das in der Verordnung vorgegebene Fristengefüge benachteiligt zu werden und von den zwölf vorgegebenen Monaten gleich mehrere für das Sammeln

[6] Die konkrete Handhabung bleibt in der gesamten Verordnung im Dunkeln.
[7] Vgl. EBI-Verordnung, Präambel, Rz 14.
[8] Durchführungsverordnung (EU) Nr. 1179/2011 der Kommission vom 17. November 2011 zur Festlegung der technischen Spezifikationen für Online-Sammelsysteme gemäß der Verordnung (EU) Nr. 211/2011 des Europäischen Parlaments und des Rates über die Bürgerinitiative.
[9] Vgl. EBI-Verordnung, Präambel, Rz 16.
[10] Laut Verordnung sollte die Software ab 1. Jänner 2012 zur Verfügung stehen, alsbald stellte sich jedoch heraus, dass die zu diesem Zeitpunkt bereit gestellte Version noch im Entwicklungsstadium steckte.
[11] Die Zertifizierung ist sodann in allen Mitgliedstaaten gültig.
[12] Obwohl sich viele Mitgliedstaaten auf einen Antrag zur Zertifizierung von Online-Sammelsystemen gründlich – und auch mit Kosten verbunden – vorbereitet haben, wurden nach Wissensstand der Autoren bislang neben Luxemburg nur in Deutschland Online-Sammel-Systeme zertifiziert.

von Unterstützungsbekundungen via Internet zu verlieren.[13] Die Kommission begegnete der Kritik, indem sie – einigermaßen überraschend – die Frist für die Vorlage von Unterstützungsbekundungen bei den einzelnen Mitgliedstaaten bis zum 1. November 2013 erstreckte.[14] Nach Ansicht des österreichischen Bundesministeriums für Inneres erfolgte dieser Schritt ohne Rechtsgrundlage und führte zu einem nicht unbeträchtlichen Spannungsfeld zwischen der neuen Vorgabe und der österreichischen innerstaatlichen Rechtslage, die (wie die Verordnung) keine solche Verlängerung erlaubte. Die Probleme haben sich allerdings, wie sich später gezeigt hat, zumindest in Österreich „in Luft aufgelöst", weil keine der bis Ende 2013 zur Überprüfung vorgelegten Europäischen Bürgerinitiativen das ursprüngliche Zeitfenster der Verordnung nach österreichischer Interpretation (Näheres siehe unten) durchbrachen.

2 Erfahrungen bei der Überprüfung von Unterstützungsbekundungen in Österreich

Die Verordnung räumt den Mitgliedstaaten ein, nicht sämtliche Unterstützungsbekundungen einer Europäischen Bürgerinitiative zu überprüfen, sondern sich auch Stichproben, also eines *random sampling*, zu bedienen.[15] Die Verordnung enthält keine näheren Regelungen über die Ausgestaltung dieses *random sampling*, insbesondere sind auch keine Konfidenzintervalle vorgegeben. Allerdings hat die Kommission im April 2013 den Mitgliedstaaten, basierend auf Erfahrungen beim Bundesverwaltungsamt in Köln, Vorschläge unterbreitet, wie eine Überprüfung unter Zuhilfenahme von Stichproben aussehen könnte.[16] Die Möglichkeit des *random sampling* erstreckt sich übrigens sowohl auf die Überprüfung von Unterstützungsbekundungen in Papierform, als auch auf elektronisch (mit einem Online-Sammelsystem) eingebrachte Unterstützungsbekundungen. Österreich hat sich dessen ungeachtet durch innerstaatliche Regelungen dazu entschieden, Unterstützungsbekundungen systematisch einer Überprüfung zu unterziehen und sich keiner Stichproben-Ziehung zu bedienen.[17]

Es stellt sich die Frage, was unter der „Überprüfung" einer Unterstützungsbekundung zu verstehen ist. Klare Vorgabe der Verordnung ist, dass darunter nicht die Überprüfung einer Unterschrift fällt; auch eine treffsichere Überprüfung der mit einer elektronisch abgegebenen Unterstützungsbekundung erfolgten Willensbekundung ist nicht vorgesehen. Dass die Qualität der Überprüfung in den einzelnen Mitgliedstaaten unterschiedlich gesehen wird, lässt sich schon aus dem durch Annex III zur Verordnung

[13] Vgl. etwa den Website-Eintrag des MEP Gerhard Häfner vom 6. Juni 2012 zur Kritik von Organisatoren an der Software: „Vor allem bemängeln sie, dass diese grundsätzliche Web-Standards nicht erfüllt und die Vorschriften zur Datensicherheit für die meisten Initiativen unvertretbar hohe Kosten bedeuten." (http://www.geraldhaefner.de/2012/06/softwareprobleme-duerfen-nicht-zu-lasten-der-europaeischen-buergerinitiative-gehen/).

[14] Vgl. Mitteilung von Juraj Nociar, Head of Cabinet of European Commission Vice-President Šefčovič, vom 30. Oktober 2012, kommuniziert durch eine E-Mail-Aussendung von Head of Unit Mario Tenreiro am 31. Oktober 2012 an alle Mitgliedstaaten.

[15] Vgl. EBI-Verordnung, Präambel, Rz 18.

[16] „Guidelines and Recommendations for Practical Implementation" der EK vom 15. April 2013.

[17] Vgl. § 3 Abs. 6 EBIG: „Hat die Bundeswahlbehörde sämtliche Unterstützungsbekundungen überprüft …".

vorgegebenen Katalog der in den einzelnen Mitgliedstaaten anzugebenden Daten ableiten. Während es in neun Mitgliedstaaten, darunter der Bundesrepublik Deutschland, nicht erforderlich ist, eine „ID-Nummer“ anzugeben (darunter kann eine Passnummer zu verstehen sein, aber etwa auch die Nummer einer Sozialversicherungskarte oder eines anderen Dokumentes), ist in anderen Mitgliedstaaten – allerdings in sehr unterschiedlicher Qualität – die Angabe einer solchen Ausweisnummer gesetzlich verankert.

In Österreich ist durch das EBIG festgeschrieben, dass eine unterstützungswillige Person die Nummer ihres Passes oder ihres Personalausweises anzugeben hat.[18] Für den österreichischen Gesetzgeber stand bei der Vorgabe, eine der beiden Nummern einzutragen, die Überprüfbarkeit der Existenz dieser Person im Vordergrund, ebenso jedoch die Implementierung eines „Sicherheitsmerkmals“, mit dem Datendiebstahl stark eingeschränkt, wenn vielleicht auch nicht gänzlich verhindert werden kann. Die Beschränkung der Zulässigkeit auf eine dieser zwei Ausweisnummern ebnete die Möglichkeit, ein kostengünstige und, wie sich herausgestellt hat, effiziente EDV-Applikation zu schaffen, mit der die Unterstützungsbekundungen in hoher Treffsicherheit und mit einer denkbar geringen Anzahl an „Zweifelsfällen“ überprüft werden können. Die im Bundesministerium für Inneres *in-house* entwickelte Applikation beruht auf einer systematischen Abfrage der Ausweisdaten im österreichischen Identitätsdokumentenregister (IDR).[19] Über dieses Register sind die Nummern sämtlicher gültiger Reisepässe und Personalausweise abrufbar.

Grundsätzlich wird bei der Überprüfung von Unterstützungsbekundungen in Papierform und der Überprüfung von Online-Unterstützungsbekundungen in gleicher Weise vorgegangen. Es wird von der Bundeswahlbehörde durch das Bundesministerium für Inneres abgefragt, ob die angegebene Passnummer oder Personalausweisnummer den im Register erfassten Daten zuordenbar ist. Im Fall eines Treffers wird der Datensatz einer unterstützungswilligen Person in einer eigenen Datenbank gespeichert, um mit einem systematischen Abgleich mit dieser Datenbank auch doppelte Unterstützungsbekundungen auszuschließen. Klar ist, dass für einen gültigen Treffer Toleranzvorgaben in einigen Bereichen vorzugeben waren, insbesondere was die Schreibweise von diakritischen Zeichen, die Schreibweise eines Vornamens oder einer Anschrift betrifft. Keine Toleranz wurde bei der Implementierung der Applikation hinsichtlich des Geburtsdatums verankert, handelt es sich bei diesem Datum doch sowohl um eine Vorgabe der Verordnung, als auch des innerstaatlichen Rechts.

Schematisch gesehen kann eine Abfrage einen von folgenden Ausgängen haben:

- Die Abfrage bewirkt einen Treffer und somit eine gültige Unterstützungsbekundung.
- Die Abfrage bewirkt einen Treffer, der jedoch wegen eines bereits vorangegangen Treffers, auf die gleiche Person lautend, keine Gültigkeit der Unterstützungsbekundung nach sich zieht.

[18] Vgl. § 3 Abs. 5 EBIG.
[19] Zentrale Evidenz gemäß § 22b des Paßgesetzes 1992.

- Die Abfrage erzielt keinen Treffer, weil die angegebene Nummer keinem vorhandenen Datensatz zugeordnet werden kann.
- Die Abfrage bewirkt einen Treffer, der jedoch deshalb zu keiner gültigen Unterstützungserklärung auslöst, weil mit der (offensichtlich unzutreffenden) Angabe einer Passnummer oder Personalausweisnummer der Datensatz einer anderen Person „erraten" worden ist.

Zweifelsfälle werden in der Applikation elektronisch ausgewiesen, Sachbearbeiterinnen und Sachbearbeiter entscheiden danach „manuell", ob es sich bei einem Treffer um eine gültige Unterstützungsbekundung handelt oder nicht.

Bei der Überprüfung von Unterstützungsbekundungen in Papierform wird so vorgegangen, dass die Sachbearbeiterin oder der Sachbearbeiter lediglich die angegebene Nummer eingeben muss, die unmittelbar zu einer Abfrage eines Datensatzes im Identitätsdokumentenregister führt. Aufgrund des *response* auf das *request,* das ohne Zeitverzögerung erfolgt, kann die Sachbearbeiterin oder der Sachbearbeiter mit wenigen Mausklicks oder Tastenbefehlen eine Entscheidung treffen, wie mit dem Datensatz umzugehen ist. Die Möglichkeit der Entscheidung beinhaltet auch die Möglichkeit der Generierung eines „Zweifelfalls", über den später, allenfalls im Kollegium oder durch Vorgesetzte, entschieden werden kann.

Die Anwendung der im Bundesministerium für Inneres entwickelten Applikation hat sich im zweiten Halbjahr 2013 und im ersten Halbjahr 2014 überraschend gut bewährt. Insgesamt wurden damit bislang fünf in Österreich vorgelegte Europäische Bürgerinitiativen behandelt.[20] Sieht man von ganz geringfügigen Anfangsschwierigkeiten ab, war die Applikation rasch einsatzbereit und hat ermöglicht, dass selbst Bürgerinitiativen bei denen „zigtausend"[21] Unterstützungsbekundungen in Papierform vorgelegt worden sind, durch eine angemessene Zahl an Sachbearbeiterinnen und Sachbearbeiterinnen rechtzeitig bewältigt wurden. Bei dieser Abarbeitung wurde erkennbar, dass die Qualität der Unterstützungsbekundungen bei den einzelnen Europäischen Bürgerinitiativen sehr unterschiedlich ist. Grundsätzlich ist die Datenqualität – das wird nicht überraschen – bei Unterstützungsbekundungen aus einem Online-Sammelsystem besser, als bei Unterstützungsbekundungen, die in Papierform abgegeben worden sind. Bei manchen Europäischen Bürgerinitiativen war insbesondere bei den Unterstützungserklärungen in Papierform eine relativ große Zahl[22] ungültig, zumeist, weil nicht existente Ausweisnummern abgegeben worden sind oder weil unterstützungswillige Personen versucht haben, eine europäische Bürgerinitiative zweimal oder öfter zu unterstützen.

[20] Bislang wurden der Bundeswahlbehörde Unterstützungsbekundungen österreichischer Wahlberechtigter für folgende EBIs vorgelegt: „Wasser und sanitäre Grundversorgung sind ein Menschenrecht! Wasser ist ein öffentliches Gut und keine Handelsware!" (kurz „Right2Water"), „Einer von uns", „Stop vivisection", „Bedingungsloses Grundeinkommen" (BGE) - Erforschung eines Weges zu emanzipatorischen sozialstaatlichen Rahmenbedingungen in der EU" und „30km/h – macht die Straßen lebenswert!".

[21] Die meisten Unterstützungsbekundungen (64.843, davon 57.643 gültige) wurden bislang von der EBI „Right2Water" eingereicht.

[22] In einem Fall waren 12.860 von insgesamt 37.833 Unterstützungsbekundungen ungültig.

3 Bilanz zwei Jahre nach Inkrafttreten der Verordnung

In einer am 15. April 2014 abgehaltenen Konferenz[23] in Brüssel, die als „Europäische-Bürgerinitiative-Tag 2014“ bezeichnet wurde und bei der vor allem Vertreterinnen und Vertreter von zugelassenen und nicht zugelassenen Europäischen Bürgerinitiativen präsent waren, konnten mehrere Kernprobleme herausgearbeitet werden, die keineswegs überraschen:

- Zulassung oder Nichtzulassung von Europäischen Bürgerinitiativen.

  Die Frage der Registrierung oder Nicht-Registrierung von Europäischen Bürgerinitiativen geht an der Verwaltung in den einzelnen Mitgliedstaaten vorbei, da es sich hierbei ausschließlich um eine Zuständigkeit der Europäischen Kommission handelt. Die Mitgliedstaaten sehen nur auf der Homepage der Kommission, ob und wann eine Europäische Bürgerinitiative zugelassen oder nicht zugelassen wurde. Auch Zurückziehungen von Bürgerinitiativen sind dort ersichtlich. Vertreterinnen und Vertreter von Europäischen Bürgerinitiativen beklagten, dass die Entscheidung der Kommission zum Teil nicht nachvollziehbar sei und dass es an Begründungen für die Nichtzulassung mangle. Aus österreichischer Sicht kann dazu angemerkt werden, dass für die Zulassung oder Nichtzulassung neben klaren gesetzlichen Regelungen gegebenenfalls, insbesondere bei Versagen einer Registrierung, eine ausführliche Begründung wichtig wäre – vor allem mit Blick auf den Umstand, dass Organisatoren mit dieser Begründung besser abwägen könnten, ob es sich lohnen würde, nachfolgenden Rechtschutz in Anspruch zu nehmen.

- Unterschiedlichen Zugangsbedingungen bei der Abgabe von Unterstützungsbekundungen gemäß Annex III der Verordnung

  Scharfe Kritik wurde von Organisatorinnen und Organisatoren daran geübt, dass manche Mitgliedstaaten bezüglich der Zulässigkeit der Abgabe einer Unterstützungsbekundung auf den Wohnsitz, andere auf die Staatsangehörigkeit abstellten und dass nur ganz wenige Mitgliedstaaten Unterstützungsbekundungen aufgrund des Vorhandenseins eines der beiden Merkmale zuließen. Das Problem ist schon bei Inkrafttreten der Verordnung bekannt und wurde unter anderem in der Literatur ausführlich dokumentiert.[24]

[23] „European Citizens' Initiative Day 2014: Building up success” (http://www.eesc.europa.eu/?i=portal.en.events-and-activities-eci-day-2014).

[24] Vgl Müller-Török/Stein, The Assignment of European Citizens to Member States in the Regulation on the European Citizens’ Initiative 53; Müller-Török/Stein, Die Rechtsakte zur Europäischen Bürgerinitiative – Offene Fragen und Risiken für Organisatoren, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Transformation juristischer Sprachen, Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012 (2012) 237.

Selbst wenn es keine klaren Analysen, sondern nur Schätzungen gibt, gilt als sicher, dass wohl Millionen von Unionsbürgerinnen und Unionsbürgern die Möglichkeit hätten, Unterstützungsbekundungen – wenn auch rechtswidrig – in zwei Mitgliedstaaten abzugeben. Wesentlich problematischer erscheint es aber, dass einer anderen Gruppe von Unionsbürgerinnen und Unionsbürgern (es wird sich ebenso um Millionen handeln) die Möglichkeit der Abgabe einer Unterstützungsbekundung versagt bleibt. Betroffen sind jene Unionsbürgerinnen und Unionsbürger, die aus einem Herkunftsmitgliedstaat stammen, in dem bei der Abgabe einer gültigen Unterstützungsbekundung auf den Wohnsitz abstellt wird, und die in einem Mitgliedstaat wohnen, der auf die Staatsbürgerschaft abstellt (Beispiel: Bürger des Vereinigten Königreichs in Österreich). Eine rasche Lösung dieses Problems – auch in der Evaluierungsphase nach drei Jahren – ist nach Prognose der Autoren nicht in Sicht, haben doch die jeweiligen Mitgliedstaaten Gründe, warum sie sich für eine bestimmte Regelung ausgesprochen haben. Eine Harmonisierung würde bedeuten, dass sehr viele Mitgliedstaaten Änderungen in ihren Rechtsordnungen verankern müssten, die nur – wenn überhaupt – mit großer Vorlaufzeit und häufig mit großen Kosten umsetzbar wären.

- Wegfall des Erfordernisses der Angabe von ID-Nummern

  Von der Zivilgesellschaft und Politikerinnen und Politikern wird immer wieder verlangt, dass alle Mitgliedstaaten auf das Erfordernis der Angabe einer ID-Nummer verzichten sollten. Dies würde, rein formal, bedeuten, dass bei elektronischen Unterstützungsbekundungen nur noch die Existenz der elektronisch erfassten Bürgerinnen und Bürger bewiesen werden könnte, es aber auf die tatsächliche Willensäußerung durch den Bürger oder die Bürgerin keinerlei Hinweis mehr gäbe. Allem Anschein nach wird die Gefahr eines Datendiebstahles[25], der ohne das Erfordernis der Angabe einer ID-Nummer um ein vielfaches vereinfacht würde, bei Weitem unterschätzt. Dies umso mehr, als eine stichprobenartige Rückfrage bei unterstützungswilligen Bürgerinnen und Bürgern[26] durch die Verordnung nicht vorgesehen und somit auch innerstaatlich gesetzlich nicht gedeckt wäre.

4 Evaluierung 2015 – offene Fragen aus Österreich

Neben den oben geschilderten Fragestellungen bedarf es für Österreich bei der Evaluierung der Verordnung über die Bürgerinitiative, die für das Jahr 2015 vorgesehen

[25] Zum Risiko eines Datendiebstahls siehe u.a. Stein/Wenda, EBI vor dem Start: Legistische Maßnahmen zur Implementierung der Europäischen Bürgerinitiative in Österreich, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Transformation juristischer Sprachen, Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012 (2012) 229.

[26] Zum Beispiel in der Form: „Trifft es zu, dass Sie die Europäische Bürgerinitiative XY unterstützt haben?“

ist[27], hinsichtlich weiterer Fragen dringend einer Evaluierung. Einige seien wie folgt angeführt:

- Digitale Signatur – wie?

  Wie anfangs erwähnt, ist neben dem Sammeln von Unterstützungsbekundung mit einem Online-Sammelsystem und in Papierform auch eine Unterstützung der Europäischen Bürgerinitiative mit Hilfe einer fortgeschrittenen elektronischen Signatur[28] vorgesehen. Bemerkenswert ist, dass diese Signatur im Gefüge der Verordnung dort angesiedelt ist, wo von der Unterstützung in Papierform die Rede ist. Man kann mit Fug und Recht sagen, dass auf Ebene der Kommission bezüglich deren tatsächlicher Verwendung bis zum Jahr 2012 kaum tiefergreifende Erörterungen stattgefunden haben. Aus österreichischer Sicht müsste diese Frage aber unbedingt geklärt werden. Naheliegend wäre es, dass die Abgabe einer digitalen Signatur in technischer Hinsicht im Rahmen des Online-Sammelsystems angesiedelt wäre, wobei den unterstützungswilligen Bürgerinnen und Bürgern mit Sicherheit zugestanden werden müsste, dass die Abgabe einer digitalen Signatur die Angabe einer ID-Nummer ersetzt.

- Zeitfenster – ja/nein?

  Ursprünglich schien die Kommission davon auszugehen, dass für eine Europäische Bürgerinitiative bereits spätestens ein Jahr nach dem Datum der Registrierung keine Unterstützungsbekundungen mehr gesammelt werden dürfen. Aufgrund der großen Probleme bei der Implementierung der Online-Sammelsysteme hat sich diese absolute Ansicht aber allem Anschein nach. Die von Österreich wiederholt postulierte Interpretation[29], wonach das in der Verordnung genannte Zeitfenster von insgesamt zwölf Monaten mit Blick auf das Gesamtfristengefüge um – zumindest – drei Monate nach hinten geschoben werden sollte[30], wurde zuletzt von Seiten der Kommission nicht mehr in Frage gestellt. Rechtssicherheit würde freilich nur eine Präzisierung der Verordnung bringen.

- Verschlüsselung

[27] Vgl. Art. 22 der EBI-Verordnung.
[28] Im Sinne der Richtlinie 1999/93/EG des Europäischen Parlaments und des Rates vom 13. Dezember 1999 über gemeinschaftliche Rahmenbedingungen für elektronische Signaturen (vgl. Art. 5 Abs. 3 der EBI-Verordnung).
[29] Vgl. u.a. Stein/Wenda, EBI vor dem Start: Legistische Massnahmen zur Implementierung der Europäischen Bürgerinitiative in Österreich, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Transformation juristischer Sprachen, Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012 (2012) 229.
[30] Vgl. Art. 5 Abs. 5 iVm Art. 12 Abs. 3 der EBI-Verordnung.

Das von der Europäischen Kommission angebotene Online-Sammel-System beinhaltet eine – in der Verordnung gar nicht vorgesehene – Verschlüsselung, wobei eine Software zu verwenden ist, die z.B. auf dem Rechner des Bundesministeriums für Inneres wegen der dortigen Sicherheitsvorgaben gar nicht erst installiert werden darf. Österreich, und dem Anschein nach auch andere Mitgliedstaaten, haben sich damit beholfen, die Entschlüsselungssoftware auf „Stand-Alone-Geräten" zu installieren. Zielführender wäre es jedoch, bei der Vollziehung der Europäischen Bürgerinitiative auf nicht genutzte Verschlüsselungsinfrastrukturen der Kommission zurückzugreifen.

- Form der Einbringung der Europäischen Bürgerinitiative

Auf Grund der vielfältigen, in der Verordnung vorgesehenen Möglichkeiten werden Europäische Bürgerinitiativen auf höchst unterschiedliche Weise eingebracht. Typisch war bislang, dass die Unterstützungsbekundungen in Papierform persönlich überbracht wurden und jene, die jene, die online eingegangen waren, auf einem zusätzlichen Datenträger gespeichert wurden.[31] Beim Lesen des Datenträgers kam es dann mit Blick auf die Verschlüsselung wiederholt zu Problemen, die erst durch eingehende Kontakte mit den Organisatorinnen und Organisatoren gelöst werden konnten. In zwei Fällen wurden alle Unterstützungsbekundungen ausschließlich im Postweg – und ohne vorherige Ankündigung – übermittelt; die weitere Kommunikation war dadurch nur verzögert zu bewerkstelligen. Pro futuro wäre eine allgemein gültige Regelung für die Überbringung der Daten der Unterstützungsbekundungen anzustreben.

5 Resümee

Gemessen an den Schwierigkeiten, die schon in der legistischen Ausgestaltungsphase der Verordnung zwischen 2010 und 2012 zu verzeichnen waren[32], hat die praktische Durchführung der Europäischen Bürgerinitiativen in Österreich bislang erstaunlich gut geklappt. Nunmehr wären baldige Vorarbeiten für die Evaluierung, die im Jahr 2015 stattzufinden hat, anzugehen. Nur mit einer entsprechenden Vorlaufzeit besteht die Chance, eine größere Anzahl der angeführten Probleme nachhaltig zu lösen, zumal die unterschiedlichen innerstaatlichen Gegebenheiten in den einzelnen Mitgliedstaaten prägende Faktoren sind. Für Österreich erscheint es angeraten, bei vielen Fragen Flexibilität zu zeigen, bei bestimmten Punkten – insbesondere der obligaten Angabe einer Ausweisnummer – aber am bewährten Kurs festzuhalten.

[31] Im Endeffekt waren die Dateien stets im XML-Format vorhanden.
[32] Vgl. u.a. Stein/Wenda, ECI Ready to Roll? – On the Challenge to implement the Citizens' Initiative in Member States, in Prosser/Golob/Leitner/Šimić (Hrsg), Easter European eGov Days 2011 (2011) 63.

Literaturverzeichnis

[MTS11] Müller-Török, R.; Stein, R.: The Assignment of European Citizens to Member States in the Regulation on the European Citizens' Initiative 53, 2011.

[MTS12] Müller-Török, R.; Stein, R.: Die Rechtsakte zur Europäischen Bürgerinitiative – Offene Fragen und Risiken für Organisatoren, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Transformation juristischer Sprachen, Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012 (2012) 237.

[SW11a] Stein, R.; Wenda, G.: ECI and the Member States: The Austrian Perspective, in Pichler/Kaufmann (Hrsg), Modern Transnational Democracy – How the 2012 launch of the European Citizens' Initiative can change the world, Schriften zur Rechtspolitik 33 (2011) 91.

[SW11b] Stein, R.; Wenda, G.: ECI Ready to Roll? – On the Challenge to implement the Citizens' Initiative in Member States, in Prosser/Golob/Leitner/Šimić (Hrsg), Easter European eGov Days 2011 (2011) 63.

[SW11c] Stein, R.; Wenda, G.: Implementing the ECI: Challenges for the Member States, in Prosser (Hrsg), EDEM 2011, Proceedings of the 5th International Conference on E-Democracy (2011) 45.

[SW12a] Stein, R.; Wenda, G.: Europäische Bürgerinitiative in Kraft – Mitpartizipieren auch online, eGovernment Review 10/2012, 30.

[SW12b] Stein, R.; Wenda, G.: EBI vor dem Start: Legistische Massnahmen zur Implementierung der Europäischen Bürgerinitiative in Österreich, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Transformation juristischer Sprachen, Tagungsband des 15. Internationalen Rechtsinformatik Symposions IRIS 2012 (2012) 229.

Identifying Building Blocks of Internet Voting: Preliminary Findings

Robert Krimmer[1]

Tallinn University of Technology
Ragnar Nurkse School of Innovation and Governance
Akadeemia tee 3
12618 Tallinn, Estonia
robert.krimmer@ttu.ee

Abstract: In this paper we present preliminary findings of a case study analysing the implementation experience with Internet voting in Austrian academic tests and one election. The project identified a first set of six building blocks which seem important for the conduct of Internet voting, which include legal basis, identification, secrecy, transparency, control, and evaluation.

1 Introduction

Recognizing each other as equal is the central precondition for a society to develop democratic institutions [Dahl1998]. Over time, societies have developed various forms of decision making, which are often also ways that culture is expressed. In one way or another, all democracies use elections to make legitimate decisions. As Kofi Annan [KA2000] once put it: 'Democracy must be more than free elections, [but] it is also true […] that it cannot be less.' As such elections are one expression of the culture and identity of a society; however, they are also processes in continuous evolution that differ from society to society.

An election is a formal process by which members of a society make a choice on a topical question or candidate for a public office. Since the end of the Second World War, many efforts have been undertaken to establish a common understanding of what constitutes a democratic election. The United Nations' first set out standards [UN1966] for the main qualitative factors that an election has to satisfy: holding genuine periodic elections, extending the right to all citizens to stand and to cast a (secret) vote, and allowing free expression of the will of the vote. But these standards say little about procedures, leaving unanswered the questions of how and with which methods elections are to be conducted.

[1] Research for this article was supported in part by Estonian Research Council's institutional grant IUT19-13.

Initially, elections were held in controlled environments using a variety of technologies including vocal communication, show of hands, stones, wax tablets, or fragments of clay. Today paper ballots are most commonly used. With the increasing need to accommodate citizens' mobility, paper ballots are also used to allow for voting in uncontrolled environments, i.e. to allow the voter to make choice either inside or outside of a dedicated polling station, using the postal service to receive and return the ballot.

As soon as electricity became available, the question arose, whether it could be used to allow for electrified and later electronic voting. One of the earliest efforts to develop electronic vote casting machines date back to 1849 when Martin de Brettes developed an early model. Werner von Siemens and Thomas Edison followed with similar proposals All of them were designed for the use by parliamentarians. Despite public interest, none of these inventions were ever actually used [Krim2012].

It took over 100 years for the use of electronic voting machines in polling stations to spread happening in the 1970s in the U.S. [Toka2005] and the Netherlands [Leye2010], and in the late 1990s in Germany [BWGV]. At about the same time, research started to make the visionaries' dreams come closer to reality. It was David Chaum [Chau1982], on the basis of the advances in cryptography by Diffie & Hellman [DH1976] with the use of asymmetric keys as well as ist implementation by Rivest, Shamir & Adleman [RSA1978], who made casting secret votes via the Internet possible for the first time.

It took a little longer for this experiment to reach real elections. When the Internet became available to the public in the 1990s, a transformative process gained momentum. It started by enabling online purchasing (e-commerce), continued by processing government transactions (e-government) and finally also reached democratic procedures, bringing with it remote electronic voting that promised to transform democracy.

Unlike the storm that shocked the traditional business world, developments in election process take much longer, mostly because introducing remote electronic voting involves many more questions than the basic ones of who is able to offer a product cheapest and fastest.

Nevertheless a number of experiments with remote electronic voting via the Internet. The first legally binding elections via the Internet took place in the United States with the Arizona primaries (for an analysis of this trial in 2000 see Solop in [Solo2004]) and in the student parliament elections in Osnabrück/Germany [Stege2000] in early 2000. In Switzerland discussions started in 1998; since 2004 the Swiss have been gaining experience in a step-by-step approach, starting with three pilot cantons Geneva, Neuchâtel, and Zurich [Brau2004], and expanding in recent years to thirteen [Burn2011]. The small Baltic republic of Estonia was the first country in the world to introduce remote electronic voting for all elections in 2005. Since then it has held six nation-wide elections with an electronic remote channel [VVK2014]. Australia experimented in 2007 with remote electronic voting, also investigating how vision-impaired persons can benefit from such a voting channel [AEC2007]. In 2011 and 2013, Norway piloted a remote electronic voting channel in local government elections that allows the possibility for the voter to verify if the vote was cast as intended [MLGR2011, 2013].

Overall, the number of users of electronic voting has increased over the past few years. But the trend has not been uniform, and has spurred intensive discussions amongst election administrators, researchers, vendors, media, observers, and last but not least voters. While the decision whether or not to introduce remote electronic voting remains a question to be answered by responsible decision makers, the implementation of Internet voting presents challenges of its own.

In this paper we present preliminary findings of a project which tries to identify the building blocks of Internet voting. The project deals with the case of Austria, where after several experiments, including two academic tests in 2003 [PKKU2003] and 2004 [PKKU2004], a highly contested Internet voting project was implemented in 2009 [KET2010]. These three elections where examined in detail for the identification of building blocks which seemed important for the conduct of Internet voting.

2 Methodology

The aim of the study is to identify building blocks of Internet voting. Such an approach, to conduct an in-depth study of a contemporary phenomenon using multiple sources of evidence in its real-life context, is the typical application of case-study methodology [Yin2003]. This is an area where traditionally there has been little prior research in the IS field [BGM1987]. For this case study the following sources of information were used for its compilation: (i) project reports of the academic tests and the election, (ii) academic papers, (iii) legal documents, (v) media reports on the issue, (vi) as well as further documents that help understand the case study context and content.

The case study analysis was organized in three phases: (i) define and design; (ii) prepare, collect, and analyze; and (iii) analyze and conclude, following Yin [2003]. For the analysis of the data, a reference framework was used [Krim2012] to identify emerging patterns throughout the implementations [Klug2000].

3 Identifying Building Blocks

The preliminary findings of the project hint that a first six areas have to be paid attention to when designing, building and finally deploying a remote electronic voting channel via the Internet:

1. The adaptation of the legal basis;
2. The selection of technical means to solve the main paradox of unequivocally identifying the eligible voter; and at the same time
3. Ensuring the secrecy of the vote;
4. Evaluating that the software works as required;
5. Giving the election commission control over the process; and
6. Enabling overall transparency for the process.

Figure 1 gives an overview of these topics and in the following these building blocks are described in more detail.

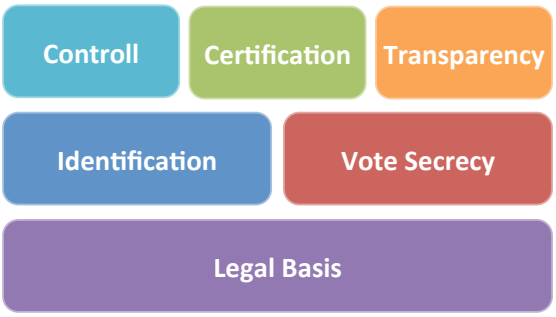


Figure 1: Building Blocks of Internet Voting

3.1 Legal Basis

When designing the legal basis for electronic voting, the first question is whether or not it is in line with international commitments. Here the Venice Commission of the Council of Europe commissioned a study [Grab2004a, Grab2004b] which found general compatibility. In addition, the Council of Europe also passed a recommendation on how electronic voting systems should be designed [CoE2004]. At the third meeting of reviewing the recommendation, it was amended by two documents to reflect recent developments in transparency and certification [CoE2011a, CoE2011b]. Most publications on national legislation regarding remote electronic voting concentrated on whether it is in line with the constitutional requirements of the respective country [Zisk2004; Karp2005]. In the end, project experience showed that most importantly the implementation of the law's principles into an ordinance needs to include technical detail [GoWe2012].

3.2 Identification

The identification of voters is an essential part of the whole voting process, and closely linked to the available online identity management infrastructure in the country where the election is being conducted. Some countries have begun to equip their citizens with smart cards [Maat2004] and in the case of Austria the smart cards are linked to the existing population registers [LHP2002]. For others such solutions were (i) too costly, (ii) delayed due to data protection concerns [RRM2005], or (iii) delayed for a long time due to lack of national certification providers [CH2007]. While in such cases one-time passwords (transaction authorization numbers – TAN) were used, which resulted in high costs for printing and distribution of voting cards for each election [Brau2004]. However the increased level of security came at the price of usability. While TAN are easily used by voters, smart cards can require high level of transaction costs to issue them as well as prove a high barrier to participation which is hard to overcome.

3.3 Vote Secrecy

Ensuring the freedom of the voter to cast a ballot of his/her choice requires that it remain impossible to link a voter and his/her vote, both at the time of casting the vote as well as in the future. Many algorithms have been proposed in the past 30 years, all of which hide either the vote or the voter by cryptographic and/or organizational means. For an overview of different available algorithms see [SP2006; Paul2011].

Most of this research however does not include experience in real-world elections. This can be assumed to be the reason why most algorithms used in practice are of less sophisticated nature than those considered state-of-the-art in research.

The most successful form of supporting free cast of a vote is allowing the voter to cast a vote more than once, while ensuring that only the last vote counts [VG2006]. However, this requires changing legal regulations which in the Austrian case was not possible.

3.4 Control by the Electoral Committee

Traditional voting processes are organized by an election committee. Often election administrators have a legal background and only limited technical experience. They often consult with technically experienced personnel or companies to conduct the electronic voting processes. Still, election committees should remain in full control of the conduct of the election. This becomes in particular challenging when there is need to allow the election commission to start, stop or interrupt the process. Most algorithmic solutions propose no technical means for this and therefore require organizational measures through regulation, such as detailled contractual relations with the vendor helping to implement this control element.

3.5 Certification

In addition to the necessity of overall trust- and transparency-enhancing measures, the correct functioning of the electronic voting software is in doubt if it is not checked before its actual use. Before evaluations can be performed, one has to translate legal requirements into functional and organizational requirements [LSBV2010]. Here, the technical part of the Council of Europe recommendation [CoE2004, CoE2011a/b] has made a fundamental contribution to the development of generally accepted technical requirements. Ahead of such international technical standards that can be used for certification, it was necessary to develop national approach in the case of Austria [ASIT2009]. Re-use of these techniques by others is limited, as they are either designed for specific existing systems, tied to national (electoral) legislation, or too generic [Volk2009].

3.6 Transparency

Paper-based voting processes are easy to understand and to follow. The use of electronic means presents the inherent problem that electronic bits and bytes cannot be seen. This in effect results in a process that requires access to documentation of the actual proceeding of the operation of the electronic voting system, as well as advanced mathematical and technical knowledge to understand the overall logic behind it. While early efforts introduced confirmation numbers that would allow voters to verify that their confirmation number is included in a public bulletin board, recent research proposes the use of end-to-end verifiability approaches [Ryan2009], which would allow the voter to check whether his/her vote was cast as intended, recorded as cast, as well as counted as recorded. The proposals use (mathematical) proofs to allow these checks. Practical experience with end-to-end verifiable systems is limited (only available in Norway and Estonia) and trust and transparency in the conduct of the electronic election remain pre-requisites, but can possibly be enhanced through efforts like this, in particular when considering universal approaches which could support observation efforts [KrVo2006; NVT2013].

4 Summary

The preliminary analysis of this project shows insights in the development process that the field of remote electronic voting has taken. Clearly the experiences with implementing such new election technologies have shown that what once seemed a technical problem became much more. At the beginning the technical solutions for the main problem of verifying the eligibility of voters and keeping their vote secret was at the center of attention. Later more sophisticated algorithms were developed and functionalities like trying to give more control to election commissions were added. However, the experiences showed, that accurate legal regulations are needed, which not only show the interaction with the constitutional legal texts but also on how to give accountability to a remote electronic voting channel through legal means. International standards were a first step, but much more necessary are regulations based on actual experience which show how remote electronic voting channels can be realized and where it is needed in order to avoid problems identified in pilot implementations. Further this practical knowledge also shows that sophisticated algorithms are not always the key to success. Much rather, several key implementations make use of very basic technical means to realize the tasks given by law. One should not forget about the voters. They not only need to use such systems but also need to understand the processes in order to build up trust. It requires trust in the election administration; otherwise suspicion will arise with more technology in an election process.

Last but not least it also shows that remote electronic voting is one of the most challenging IT projects. Not only does the requirement for secrecy of a vote rule out many approaches towards IT-security in the Internet, but also elections themselves are special projects: they have a fixed date when they have to take place – if the system is ready or not.

References

[AEC2007] Australian Election Commission: *Federal Electronic Voting Trials*. Canberra 2007. Available from http://www.aec.gov.au/Voting/electronic-voting.htm, retrieved 03-11-2013.

[ASIT2009] A-SIT: *Bescheinigungen nach § 34 Abs. 6 HSG 1998*. Available from http://www.a-sit.at/de/bestaetigungsstelle/bescheinigungen_hsg/index.php, retrieved 16-04-2014.

[Brau2004] Braun, Nadja: E-Voting: Switzerland's Projects and their Legal Framework. Paper read at ESF TED Workshop on Electronic Voting in Europe, GI LNI, Bonn 2004.

[BGM1987] Benbasat, I., Goldstein, D. K., & Mead, M.: The Case Research Strategy in Studies of Information Systems. *MIS Quarterly, 11*(3), 369-386, 1987.

[Burn2011] Burnand, Frédéric. 2011. *E-Voting 2011 auf Sparflamme*. Swissinfo 2011, retrieved on 11-08-2013.

[BWGV1975] *Bundeswahlgeräteverordnung of 3 September 1975 (BGBl. I p. 2459), last changed on 20 April 1999 (BGBl. I p. 749)*. Available from http://www.gesetze-im-internet.de/bundesrecht/bwahlgv/gesamt.pdf, retrieved 11-02-2014.

[CH2007] Schweizer Bundesrat: *Anfrage 07.1031 zu 'Digitale Signatur',* 2007. Available from http://www.parlament.ch/d/suche/seiten/geschaefte.aspx?gesch_id=20071031 retrieved on 03-05-2014.

[Chau1982] Chaum, David. Blind Signatures for Untraceable Payments. Paper read at Advances in Cryptology: Proceedings of CRYPTO '82, Santa Barbara 1982.

[CoE2004] Council of Europe: *Legal, operational and technical standards for e-voting. Recommundation Rec(2004)11 and explanatory memorandum*. Council of Europe, Strassbourg 2004.

[CoE2011a] Council of Europe: *Guidelines of the Committee of Ministers of the CoE on Certification of E-voting Systems (2011)*. Council of Europe, Strassbourg 2011. Available from http://www.coe.int/t/dgap/democracy/activities/ggis/E-voting/Evoting%202010/Biennial_Nov_meeting/Guidelines_certification_EN.pdf, retrieved on 20-02-2012.

[CoE2011b] Council of Europe: *Guidelines of the Committee of Ministers of the CoE on Transparency of E-enabled Elections (2011)*. Council of Europe, Strassbourg 2011. Available from http://www.coe.int/t/dgap/democracy/activities/ggis/E-voting/E-voting%202010/Biennial_Nov_meeting/Guidelines_transparency_EN.pdf, retrieved on 20-02-2012.

[Dahl1998] Dahl, Robert A.: *On Democracy*. New Haven, London: Yale University 1998.

[DH1976] Diffie, Whitfield, Hellman, Martin E.: New Directions in Cryptography. *IEEE Transactions on Information Theory* 1976, 22 (6): 644–654.

[Grab2004a] Grabenwarter, Christoph: Briefwahl und E-Voting. *Journal für Rechtspolitik* 12 (2): 70-77, 2004.

[Grab2004b] Grabenwarter, Christoph: *Report on the compatibility of remote voting and electronic voting with the standards of the Council of Europe, 2004*. Available from http://www.venice.coe.int/docs/2004/CDL-AD(2004)012-e.pdf, retrieved on 12-12-2008.

[GoWe2012] Goby, Barbara, Weichsel, Herbert: Das E-Voting-Erkenntnis des VfGH: gesetzwidrige Ausgestaltung der ÖH-Wahlordnung. Zeitschrift für Hochschulrecht, Hochschulmanagement und Hochschulpolitik (Zfhr) 2012(11)3: 118-125.

[KA2000] Annan, Kofi: *UN Secretary General Kofi Annan's Closing Remarks to the Ministerial on June 27th 2000*. Available from http://www.demcoalition.org/pdf/un_secertary_gen_kofi_annan.pdf, retrieved 02-02-2014.

[Karp2005] Karpen, Ulrich: Elektronische Wahlen? Nomos Verlag, Baden-Baden 2005.

[KET2010] Krimmer, Robert, Ehringfeld, Andreas, & Traxl, Markus : Evaluierungsbericht: E-Voting bei den Hochschülerinnen- und Hochschülerschaftswahlen 2009. Bundesministerium für Wissenschaft und Forschung (BMWF), Vienna 2010.

[Klug2000] Kluge, Susanne: Empirically Grounded Construction of Types and Typologies in Qualitative Social Research. *Forum Qualitative Sozialforschung,* 1(1), Art. 14, 2000.

[Krim2012] Krimmer, Robert: The Evolution of E-voting: Why Voting Technology is Used and How it Affects Democracy. Dissertation, Tallinn University of Technology, Tallinn 2012, Available from http://digi.lib.ttu.ee/i/?798, retrieved 02-06-2014.

[KrVo2006] Krimmer, Robert, Volkamer, Melanie: Observing Threats to Voter's Anonymity: Election Observation of Electronic Voting, Electronic Government EGOV, Trauner Verlag, 43-52, Linz 2006.

[NVT2013] OSCE/ODIHR: Handbook for the Observation of New Voting Technologies, Warsaw, 2013.

[Leye2010] Leyenaar, Monique: *Electronic Voting: A Case Study from the Netherlands. Presentation at CiO Seminar on 'Present State, and Future Prospects of the Application of Electronic Voting in the OSCE Participating States' in Vienna on 16 and 17 September 2010*. Available from http://www.osce.org/cio/71355, retrieved on 02-04-2014.

[LHP2002] Leitold, Herbert, Hollosi, Arno & Posch, Reinhard: Security architecture of the Austrian citizen card concept, Vienna 2002.

[LSBV2010] Langer, Lucie, Schmidt, Axel, Buchmann, Johannes, Volkamer, Melanie: A taxonomy refining the security requirements for electronic voting: analyzing helios as a proof of concept. ARES, IEEE 2010.

[Maat2004] Maaten, Epp: Towards Remote E-Voting: Estonian Case. In: TED Workshop on Electronic Voting in Europe, GI LNI, Bonn 2004.

[MLGR2011] Ministry of Local Government and Regional Development: *The e-vote project,* Oslo 2011. Available from http://www.regjeringen.no/en/dep/krd/prosjekter/e-vote-2011-project.html?id=597658, retrieved on 11-03-2012.

[MLGR2013] Ministry of Local Government and Regional Development: *The e-vote 2013 project*, Oslo 2013. Available from http://www.regjeringen.no/nb/dep/kmd/kampanjer/valg/nytt-om-valg/2014/Rapport-fra-Cartersenteret-om-forsok-med-internettstemmegivning.html?id=756400, retrieved on 05-05-2014.

[Paul2011] Paulsen, Christian: *Sicherheit von Internetwahlen*. Hamburg: BoD, Books on Demand, 2011.

[PKKU2003] Prosser, A., Kofler, R., Krimmer, R., & Unger, M. K.: Die erste Internet-Wahl Österreichs: Ein Erfahrungsbericht von e-Voting.at. The first Internet-Election in Austria: The Findings by e-Voting.at, Institute of Information Processing and Information Management, Vienna 2003.

[PKKU2004] Prosser, A., Kofler, R., Krimmer, R., & Unger, M. K.: E-Voting Election Test to the Austrian Federal Presidency Election 2004. Institute of Information Processing and Information Management, Vienna 2004.

[RRM2005] Reichl, H., A. Roßnagel, and G. Müller. 2005. *Digitaler Personalausweis: Eine Machbarkeitsstudie*: Duv.

[RSA1978] Rivest, Ronald L., Adi Shamir, and Leonard M. Adleman.. A Method for Obtaining Digital Signatures and Public-Key Cryptosystems. *Communications of the ACM* 21 (2):120-126, 1978.

[Ryan2009] Ryan, Peter Y.A., David Bismark, James Heather, Steven Schneider, and Zhe Xia: Prêt à voter: a voter-verifiable voting system. *IEEE Transactions on Information Forensics and Security* 4 (4), 2009.

[Solo2004] Solop, Frederic I.: Electronic Voting in the United States: At the Leading Edge or Lagging Behind? In *Electronic Voting and Democracy: A Comparative Analysis*, edited by N. Kersting and H. Baldersheim. London: Palgrave, 2004.

[SP2006] Sampigethaya, Krishna, and Radha Poovendra. A framework and taxonomy for comparison of electronic voting schemes. *Computers & Security* 25 (2):137-153, 2006.

[Stege2000] Stegers, Fiete: Virtuelle Wahl: Ab 0:00 Uhr kann geklickt werden. *SPIEGEL Online,* 2000.

[Toka2005] Tokaji, Dan P.: The Paperless Chase: Electronic Voting and Democratic Values. *Fordham Law Review* 57, 2005.

[UN1966] United Nations: *International Covenant on Civil and Political Rights* 1966. Available from http://www2.ohchr.org/english/law/ccpr.htm, retrieved on 20-04-2013.

[VG2006] Volkamer, M., & Grimm, R.: Multiple Casts in Online Voting: Analyzing Chances. In R. Krimmer (Ed.), *Electronic Voting 2006* (Vol. 86, pp. 97-106). Bregenz: GI, 2006.

[Volk2009] Volkamer, Melanie: *Evaluation of electronic voting: requirements and evaluation procedures to support responsible election authorities*. Vol. 30. Berlin: Springer LNBIP, 2009.

[VVK2014] Vabariigi Valimiskomisjon: *Internet Voting - Voting Methods in Estonia - Estonian National Electoral Committee* 2014. Available from http://www.vvk.ee/voting-methods-in-estonia/engindex, visited 06-05-2014.

[Yin2003] Yin, R. K.: *Case Study Research. Design and Methods* (3 ed.). Thousand Oaks, London, New Delhi, Sage Publications, 2003.

[Zisk2004] Ziska, Bernd: *Verfassungsrechtliche Rahmenbedingungen für E-Voting* 2004. Available from http://www.rechtsprobleme.at/doks/ziska-e-voting.pdf, retrieved 03-02-2014.

A Lightweight Implementation of a Shuffle Proof for Electronic Voting Systems

Philipp Locher[1,2] and Rolf Haenni[1]

[1] Research Institute for Security in the Information Society
Bern University of Applied Sciences, CH-2501 Biel, Switzerland
{philipp.locher, rolf.haenni}@bfh.ch
[2] Department of Informatics
University of Fribourg, CH-1700 Fribourg, Switzerland
philipp.locher@unifr.ch

Abstract: In the usual setting of a verifiable mix-net, an input batch of ciphertexts is shuffled through a series of mixers into an output batch of ciphertexts while hiding the mapping between input and output and preserving the plaintexts. Unlike shuffling, proving the correctness of a shuffle is relatively expensive and difficult to implement. In this paper, we present a new implementation of a shuffle proof based on the proof system proposed by Wikström and Terelius. The implementation offers a clean and intuitive application programming interface and can be used as a lightweight cryptographic component in applications of verifiable mix-nets. Verifiable electronic voting is the most prominent target application area.[1]

1 Introduction

Verifiable mix-nets are important building blocks in electronic voting protocols. They are used to provide vote secrecy by anonymizing the voting channel from the voter into the final tally. Some protocols use re-encryption mix-nets to shuffle the list of encrypted votes [BGP11, RBH$^+$09, RBH$^+$09], while other protocols require mix-nets to shuffle the voters' credentials [Nef01, JCJ05, HS11]. In both cases, the shuffling is performed through a series of mixers. To demonstrate the correctness of a mix-net shuffle, mixers provide individual zero-knowledge proofs—called *shuffle proofs*—to certify each step of the shuffling process. The link between the input and output of a mix-net remains hidden, as long as at least one trustworthy mixer is involved in the shuffling.

Shuffle proofs can be constructed in various ways. The first efficient shuffle proofs were proposed independently by Neff [Nef01] and Furukawa and Sako [FS01]. Neff's approach, which is based on the invariance of polynomials under the permutation of their roots, has later been improved by Groth, Ishai, and Bayer [Gro10, GI08, BG12]. The Furukawa and Sako approach is based on a commitment to a permutation matrix. Later, Wikström showed how to split the shuffle proof in an offline and online phase [Wik09]. Together with

[1] This work is supported by the Swiss National Science foundation, under the grant 200021L-140650/1.

Terelius, Wikström presented an improved and generalized proof, that allows choosing the permutation from a restricted subset of all permutations [TW10].

To the best of our knowledge, Wikström's *Verificatum* is currently the only off-the-shelf mix-net implementation that offers a complete shuffle proof [Wik13]. Verificatum has been used in the 2013 parliamentary election in Norway and for University elections in Israel. The complete Java source code is publicly available under a research license. Aside from Verificatum, multiple prototype implementations of shuffle proofs with corresponding performance tests have been mentioned in the literature [FMM+02, FMS10, BGP11, BG12], but none of them is available as a stand-alone library. The most recent performance analysis in [BG12] reports slightly better running times for Groth's approach compared to Verificatum, but this can be explained by the chosen programming languages (C++ vs. Java) and different levels of code optimization.

Contribution. We present a new implementation of a shuffle proof based on the proof system of Wikström and Terelius [Wik09, TW10]. Our implementation differs from Verificatum in multiple ways. First, we have embedded the shuffle proof in a cryptographic library with a clean and intuitive application programming interface. This greatly simplifies the integration of a shuffle proof in applications such as a mix-net. In other words, while Verificatum *is* a full-featured mix-net, we provide the necessary toolbox for building one. Second, as part of a cryptographic library, our implementation offers enhanced flexibility with respect to the homomorphic encryption system in use or the underlying algebraic group. It also supports proofs for shuffles that are not based on re-encryption, for example the one required in [HS11] for mixing voter credentials. Our shuffle proof implementation is therefore applicable in many different scenarios and is not even restricted to the context of verifiable mix-nets [Joa14]. The whole library comes as a lightweight Java component, which can be ported to any device—even to a notebook or smartphone—running a Java Virtual Machine. The source code is publicly available and free for non-commercial use.

Paper Overview. This paper gives an introduction and overview of our shuffle proof implementation. Section 2 presents a summary of the necessary technical background to understand Wikström's shuffle proof as implemented. Section 3 first gives some details about the design of the whole library and its components, and then presents a complete example of usage from a programmer's perspective. Section 4 concludes the paper with an outlook and an overview of ongoing work.

2 Shuffle Proof

We give a short introduction to Wikström's shuffle proof as presented in [Wik09, TW10]. To accentuate its essence, we fade out some technical details in our summary of the proof. In particular, we describe the proof in terms of two homomorphic one-way functions, from which respective preimage proofs are derived. We believe that the compactness of this representation is very instructive and simplifies the understanding of the proof.

2.1 Cryptographic Preliminaries

We denote by G_q a cyclic group of prime order q, for which the decisional Diffie-Hellman assumption is believed to hold. For simplicity, we write G_q always multiplicatively and assume that independent generators $g, h_1, \ldots, h_N \in G_q$ are publicly known (for $N = 1$ we write $h = h_1$). For an arbitrary group H and any pair of vectors $\bar{u} \in H^N$ and $\bar{e} \in \mathbb{Z}^N$, we use the notation $\langle \bar{u}, \bar{e} \rangle$ for both $\sum_{i=1}^{N} e_i u_i$ and $\prod_{i=1}^{N} u_i^{e_i}$, depending on whether H is written additively or multiplicatively. For an arbitrary finite set S, we write $r \in_R S$ for picking the value r uniformly at random from S.

Generalized Pedersen Commitments. We use Pedersen commitments $Com(m, r) = g^r h^m$ over G_q to commit to an integer $m \in \mathbb{Z}_q$ with randomization $r \in_R \mathbb{Z}_q$. To commit to a vector $\bar{m} = (m_1, \ldots, m_N) \in \mathbb{Z}_q^N$ of integers, we use generalized Pedersen commitments $Com(\bar{m}, r) = g^r \prod_{i=1}^{N} h_i^{m_i}$. To commit to an $N \times N$-matrix $M \in \mathbb{Z}_q^{N \times N}$, we compute generalized Pedersen commitments column-wise by

$$Com(M, \bar{r}) = \big(Com(\bar{m}_1, r_1), \ldots, Com(\bar{m}_N, r_N)\big),$$

where $\bar{m}_j$ denotes the j-th column of M and $\bar{r} = (r_1, \ldots, r_N) \in_R \mathbb{Z}_q^N$ the corresponding randomization vector. In case M is a permutation matrix relative to a permutation π of size N, then committing to M in this way allows computing a commitment to a permuted vector based on the matrix commitment (i.e., without knowing M or π). This is a consequence of the fact that generalized Pedersen commitments are additively homomorphic. More precisely, if $\bar{e}' \in \mathbb{Z}^N$ denotes the vector of integers obtained from $\bar{e} \in \mathbb{Z}^N$ by permuting its values according to π, then

$$\langle Com(M, \bar{r}), \bar{e} \rangle = Com(M\bar{e}, \langle \bar{r}, \bar{e} \rangle) = Com(\bar{e}', r)$$

is a commitment of $\bar{e}'$ with randomization $r = \langle \bar{r}, \bar{e} \rangle$.

Homomorphic Encryptions. For a randomized asymmetric encryption scheme, such as for example ElGamal or Paillier, we write $u = Enc_{pk}(m, r)$ for encrypting a plaintext $m \in \mathcal{M}$ with randomization $r \in \mathcal{R}$ and public key pk into a ciphertext $u \in \mathcal{C}$. Let $\odot$, $\oplus$, and $\otimes$ be respective group operations on $\mathcal{M}$, $\mathcal{R}$, and $\mathcal{C}$. An encryption scheme is homomorphic, if $Enc_{pk}(m_1, r_1) \otimes Enc_{pk}(m_2, r_2) = Enc_{pk}(m_1 \odot m_2, r_1 \oplus r_2)$ for all $m_1, m_2 \in \mathcal{M}$ and $r_1, r_2 \in \mathcal{R}$. A homomorphic encryption scheme allows a ciphertext $u = Enc_{pk}(m, r)$ to be re-encrypted with a new randomization r'. We write $ReEnc_{pk}(u, r') = u \otimes Enc_{pk}(1, r') = Enc_{pk}(m, r \oplus r')$, where $1 \in \mathcal{M}$ denotes the identity element of the plaintext space.

Zero-Knowledge Proofs. A zero-knowledge proof of knowledge is an interactive protocol in which a prover P convinces a verifier V that P knows a value (private input) satisfying a certain predicate (public input) without revealing any information about the value. Σ-proofs are zero-knowledge proofs of knowledge based on a three-message protocol: P passes a commitment t to V, V replies with a randomly chosen challenge c, and

P sends a response s back to V. The triple (t, c, s) is called proof transcript, which V either accepts or rejects. A large class of Σ-proofs results from any group homomorphism $\phi : G \to H$. Let $\odot$ and $\otimes$ be respective operators of G and H. If $x \in G$ is the private input known to P and $y = \phi(x) \in H$ the public input known to P and V, we write

$$\Sigma\text{-proof}\Big[\ x \mid\ y = \phi(x)\Big]$$

for the Σ-proof of knowing the preimage x. It can be constructed by the following standard procedure: P picks $r \in_R G$ and sends $t = \phi(r)$ to V, V replies with $c \in_R C \subseteq \mathbb{Z}$, and P sends the response $s = r \odot x^c$ back to V. V accepts the proof, if and only if $\phi(s) = t \otimes y^c$. This general construction of preimage Σ-proofs covers many known proofs of knowledge as special cases [Mau09]. It can be turned into a non-interactive proof by obtaining c from a random oracle using (t, y) as query [FS86].[2]

2.2 Proof of a Shuffle

The shuffle proof according to Wikström and Terelius consists of an offline and an online phase [Wik09, TW10]. In the offline phase, the mixer commits to a permutation matrix and proves under zero knowledge that the commitment contains indeed a permutation matrix. An upper bound for the size of the shuffle is required to conduct this phase prior to the actual shuffling. Later, in the online phase, the mixer performs the shuffle according to the committed permutation matrix and proves the correctness of the shuffle under zero knowledge.

Offline Phase. Let π be a permutation of size N and $\bar{c}_\pi = Com(M, \bar{s}) \in G_q^N$ a commitment to the corresponding permutation matrix M. If $\bar{x} = (x_1, \ldots, x_N)$ is a vector of N independent variables, then an $N \times N$-matrix M over $\mathbb{Z}_q$ is a permutation matrix if and only if $\prod_{i=1}^N \langle \bar{m}_i, \bar{x} \rangle = \prod_{i=1}^N x_i$ and $M\bar{1} = \bar{1}$. These properties allow to prove that $\bar{c}_\pi$ is a commitment to a permutation matrix [TW10]:

$$\Sigma\text{-proof}\left[\begin{array}{c} v, w \in \mathbb{Z}_q \\ \bar{e}' \in \mathbb{Z}_q^N \end{array} \middle|\ Com(\bar{1}, v) = \langle \bar{c}_\pi, \bar{1} \rangle \ \wedge\ Com(\bar{e}', w) = \langle \bar{c}_\pi, \bar{e} \rangle \ \wedge\ \prod_{i=1}^N e'_i = \prod_{i=1}^N e_i \right],$$

where $v = \langle \bar{s}, \bar{1} \rangle$, $w = \langle \bar{s}, \bar{e} \rangle$, and $\bar{e}' = (e'_1, \ldots, e'_N) = (e_{\pi(1)}, \ldots, e_{\pi(N)})$ are the private inputs. The vector $\bar{e} = (e_1, \ldots, e_N) \in \mathbb{Z}_q^N$ is a public input selected and communicated beforehand by the verifier.[3] The last part of the proof, which consists in showing the equality $\prod_{i=1}^N e'_i = \prod_{i=1}^N e_i$, can be achieved using a recursive commitment structure $c_1, \ldots, c_N$ with base case $c_0 = h$ [Wik12]. This leads to a slightly different representation

[2] To establish a binding between prover and proof, the prover's identity is sometimes adjoined to the random oracle query.

[3] Usually, the values e_i are selected from a subset $[0, 2^{k_e} - 1]^N \subseteq \mathbb{Z}_q$, where k_e is a security parameter.

of the above proof:

$$\Sigma\text{-proof}\left[\begin{array}{c|l} v,w,d \in \mathbb{Z}_q & Com(\bar{1},v) = \langle \bar{c}_\pi, \bar{1}\rangle \wedge Com(\bar{e}',w) = \langle \bar{c}_\pi, \bar{e}\rangle \wedge \\ \bar{t},\bar{e}' \in \mathbb{Z}_q^N & \bigwedge_{i=1}^N (c_i = g^{t_i} c_{i-1}^{e'_i}) \wedge Com(0,d) = c_N / h^{\prod_{i=1}^N e_i} \end{array}\right],$$

where $\bar{t} = (t_1, \ldots, t_N) \in_R \mathbb{Z}_q^N$ and $d = d_N$ are additional private inputs for $d_0 = 0$ and $d_i = t_i + e'_i\, d_{i-1}$ for $i > 0$. This leads directly a homomorphic one-way function,

$$\phi_{offline}(v,w,\bar{t},d,\bar{e}') = \Big(Com(\bar{1},v), Com(\bar{e}',w), g^{t_1} c_0^{e'_1}, \ldots, g^{t_N} c_{N-1}^{e'_N}, Com(0,d)\Big),$$

which can be used for constructing a preimage proof by the standard procedure.

Online Phase. Let $\mathcal{C}$ denote the ciphertext space of the given homomorphic encryption scheme. We assume that the largest cyclic subgroup of $\mathcal{C}$ is of the same order q as the cyclic group used for the Pedersen commitments.[4] The input list of ciphertexts is denoted by $\bar{u} = (u_1, \ldots, u_N) \in \mathcal{C}^N$ and the corresponding shuffled list of permuted and re-encrypted ciphertexts by $\bar{u}' = (u'_1, \ldots, u'_N) \in \mathcal{C}^N$. Again, a public vector $\bar{e} \in \mathbb{Z}_q^N$ is selected and communicated beforehand by the verifier. A proof that $\bar{u}'$ has been formed correctly by shuffling $\bar{u}$ according to the committed permutation matrix $\bar{c}_\pi = Com(M, \bar{s})$ can be constructed as follows [Wik09]:

$$\Sigma\text{-proof}\left[\begin{array}{c|l} r,w \in \mathbb{Z}_q & Com(\bar{e}',w) = \langle \bar{c}_\pi, \bar{e}\rangle \ \wedge \ \prod_{i=1}^N (u'_i)^{e'_i} = \prod_{i=1}^N (u_i)^{e_i}\, Enc(1,r) \\ \bar{e}' \in \mathbb{Z}_q^N & \end{array}\right],$$

where $r = \langle \bar{r}, \bar{e}\rangle$, $w = \langle \bar{s}, \bar{e}\rangle$, and $\bar{e}' = (e'_1, \ldots, e'_N) = (e_{\pi(1)}, \ldots, e_{\pi(N)})$ are the private inputs for $\bar{r} = (r_1, \ldots, r_N) \in_R \mathbb{Z}_q^N$ and $u'_i = ReEnc(u_{\pi(i)}, r_{\pi(i)})$. Again, this implies a homomorphic one-way function,

$$\phi_{online}(r,w,\bar{e}') = \Big(Com(\bar{e}',w), \textstyle\prod_{i=1}^N (u'_i)^{e'_i}\, Enc(1,-r)\Big),$$

for which a preimage proof can be constructed by the standard procedure.

3 Design and Implementation

Some selected details of our shuffle proof implementation are the focus of this section. The goal is to make the reader familiar with the design and some basic concepts of our implementation. For this, we give first some background information about *UniCrypt*, the cryptographic library into which our shuffle proof is embedded. Then we discuss a complete example of usage to illustrate the provided interface from a programmer's perspective. The chosen example covers the full process of generating and mixing some ElGamal encryptions and constructing and verifying the offline and online proofs.

[4] For simplicity, we expect the same group order for the encryptions and commitments, but this is not a necessary requirement for the proof.

3.1 UniCrypt

UniCrypt is a Java library developed for the purpose of simplifying the implementation of cryptographic voting protocols.[5] It consists of two layers, one for the mathematical fundament and one for the cryptographic primitives. The mathematical layer deals with all sorts of algebraic structures, corresponding elements, and functions. Its purpose is to provide strict type safety on a mathematical level: elements always "know" the algebraic structure to which they belong, i.e., applying a group operator is only allowed for elements of the group and evaluating a function is is only allowed for elements of the domain.

The cryptographic layer provides interfaces and implementations of various cryptographic schemes: symmetric and asymmetric encryption, secret sharing, commitments, digital signatures, zero-knowledge proofs, mix-nets, and more. It also contains a package for generating pseudo-random numbers, common reference strings, or random oracles. In the remaining paragraphs, we provide additional information to some of the cryptographic components, which appear in the example of the following subsection. Corresponding top-level Java interfaces exist in the UniCrypt library.

Mixer. This interface specifies the functionality of a pure cryptographic shuffle: an input list of values is shuffled into an output list of values without proving its correctness. Currently, two implementations of this interface are available: a *re-encryption mixer*, which permutes and re-encrypts a list of ciphertexts of a homomorphic encryption scheme, and a so-called *identity mixer*, which shuffles credentials according to the method described in [HS11].

Proof System. This is the general interface for many different types of zero-knowledge proofs. It specifies two principal methods, one for generating a proof for given private and public inputs, and one for verifying such proofs. Multiple standard zero-knowledge proofs of knowledge are implemented in a generic way. There are implementations for basic preimage proofs, for conjunctive or disjunctive compositions of preimage proofs, and for equality and inequality proofs. They can all be applied to any homomorphic function. There are also implementations for validity proofs, which can deal with ElGamal encryptions and Pedersen commitments. The online part of the shuffle proof is implemented in two different ways, one for a re-encryption mixer and one for an identity mixer.

Challenge Generator. The proof system implementation in UniCrypt is very flexible about constructing proofs in an interactive or non-interactive manner. The process of creating the challenge is abstracted in our concept of a *challenge generator*. During the construction or verification of a proof, the proof system simply calls the challenge generator to get a suitable challenge. The details of selecting the challenge in a concrete implementation are hidden behind the interface. The default implementations are non-interactive, which obtain the challenge from calling a random oracle.

[5] The source code of the UniCrypt library is publicly available on GitHub under a dual AGPLv3/commercial licence, see https://github.com/bfh-evg/unicrypt.

3.2 Example of Usage

To present our shuffle proof implementation from a programmer's point of view, we present a complete example of a re-encryption shuffle including proving and verifying the correctness of the shuffle. To keep the code as tight as possible, the example operates often in a default manner. For example, the independent generators used in the generalized Pedersen commitment are implicitly derived from the default common reference string, and non-interactive challenge generators are created automatically by the proof systems.

Setup. We first create a list of ElGamal ciphertexts, which in a normal application is given by the context. For this, we randomly select a cyclic group $G_q \subseteq \mathbb{Z}_p^*$ such that $p = 2q + 1$ is a safe prime of a specified bit length. Then the ElGamal encryption scheme is instantiated based on the default generator of the selected cyclic group, and a public key pk is chosen at random (we don't decrypt in this example, so no private key is needed). Finally, the list $\bar{u}$ of input ciphertexts is created based on random messages.

```
// Create cyclic group for random safe prime (1024 bits)
CyclicGroup group = GStarModSafePrime.getRandomInstance(1024);

// Create ElGamal encryption scheme and select random public key pk
ElGamalEncryptionScheme elGamal =
        ElGamalEncryptionScheme.getInstance(group.getDefaultGenerator());
Element pk = group.getRandomElement();

// Set shuffle size and create random ElGamal ciphertexts
int n = 100;
Tuple ciphertexts = Tuple.getInstance();
for (int i = 0; i < n; i++) {
    Pair c = elGamal.encrypt(pk, group.getRandomElement());
    ciphertexts = ciphertexts.add(c);
}
```

Listing 1: Setup

Shuffle. To shuffle the list of input ciphertexts $\bar{u}$, a re-encryption mixer is instantiated based on the ElGamal encryption scheme, the public key pk, and the shuffle size N. Then permutation π and the re-encryption randomizations $\bar{r}$ are selected at random. To be able to proof the correctness of the shuffle later, it is important to create these values explicitly. Finally, calling the shuffle method of the mixer outputs a list of ciphertexts $\bar{u}'$.

```
// Create mixer, a random permutation pi, and randomizations r
ReEncryptionMixer mixer = ReEncryptionMixer.getInstance(elGamal, pk, n);
PermutationElement pi = mixer.getPermutationGroup().getRandomElement();
Tuple r = mixer.generateRandomizations();

// Shuffle ciphertexts using pi and r
Tuple shuffledCiphertexts = mixer.shuffle(ciphertexts, pi, r);
```

Listing 2: Shuffle

Offline Phase. The first step in the offline phase of the shuffle proof is to generate the permutation matrix commitment $\bar{c}_\pi$ relative to π. For this, we instantiate a permutation commitment scheme and select the commitment randomizations $\bar{s}$ explicitly at random. To prove that $\bar{c}_\pi$ is a commitment to a permutation, we instantiate a permutation commitment proof system, which allows creating the proof using $(\pi, \bar{s})$ as private and $\bar{c}_\pi$ as public input. Note that the code of Listing 3 could be executed before the ciphertexts are shuffled in the last line of Listing 2.

```
// Create permutation commitment c_pi based on pi and randomizations s
PermutationCommitmentScheme pcs =
        PermutationCommitmentScheme.getInstance(group, n);
Tuple s = pcs.getRandomizationSpace().getRandomElement();
Tuple c_pi = pcs.commit(pi, s);

// Create permutation commitment proof system
PermutationCommitmentProofSystem pcps =
        PermutationCommitmentProofSystem.getInstance(group, n);

// Define private and public inputs
Pair offlinePrivateInput = Pair.getInstance(pi, s);
Element offlinePublicInput = c_pi;

// Generate permutation commitment proof
Pair offlineProof =
        pcps.generate(offlinePrivateInput, offlinePublicInput);
```

Listing 3: Online Phase (Proof of Knowledge of Permutation Matrix)

Online Phase. Finally, the shuffle proof can be generated with the help of a re-encryption shuffle proof system. The triples $(\pi, \bar{s}, \bar{r})$ and $(\bar{c}_\pi, \bar{u}, \bar{u}')$ are the private and public inputs, respectively.

```
// Create shuffle proof system
ReEncryptionShuffleProofSystem rsps =
        ReEncryptionShuffleProofSystem.getInstance(group, n, elGamal, pk);

// Define private and public inputs
Triple onlinePrivateInput = Triple.getInstance(pi, s, r);
Triple onlinePublicInput =
        Triple.getInstance(c_pi, ciphertexts, shuffledCiphertexts);

// Generate shuffle proof
Triple onlineProof = rsps.generate(onlinePrivateInput, onlinePublicInput);
```

Listing 4: Online Phase (Commitment Consistent Proof of a Shuffle)

Verification. The verification of the overall proof is straightforward: just call the verification methods of the proof systems with the corresponding proof and the public input as arguments, and make sure that the same permutation commitment is included in both public inputs. The proof systems are either given by the context or can be created based on common values.

```
// Verify permutation commitment proof
boolean v1 = pcps.verify(offlineProof, offlinePublicInput);

// Verify shuffle proof
boolean v2 = rsps.verify(onlineProof, onlinePublicInput);

// Verify equality of permutation commitments
boolean v3 =
        offlinePublicInput.isEquivalent(onlinePublicInput.getFirst());

if (v1 && v2 && v3) success();
```

Listing 5: Proof Verification

4 Conclusion

In this paper, we presented a short summary of the shuffle proof of Wikström and Terelius and presented an overview of a new implementation embedded in a lightweight Java library. With our example of shuffling a list of ElGamal ciphertexts, we illustrated the construction of a shuffle proof from a programming perspective. It turns out that the using the library is straightforward and intuitive. As a standalone library, it can be easily integrated in any mix-net based electronic voting system.

The results of preliminary performance tests are comparable to the results reported in the literature for other shuffle proof implementations (100,000 ElGamal ciphertexts within a few minutes, for a residue class of 160/1024 bits and on a standard notebook). We expect performance improvements by further optimizing the implementation (multi-exponentiation, pre-computations, caching, etc.). The flexibility of our library with respect to working with different groups—for example by using elliptic curves—can further speed up the proof generation and verification without code modifications.

References

[BG12] S. Bayer and J. Groth. Efficient Zero-Knowledge Argument for Correctness of a Shuffle. In D. Pointcheval and T. Johansson, editors, *EUROCRYPT'12, 31st Annual International Conference on Theory and Applications of Cryptographic Techniques*, LNCS 7237, pages 263–280, Cambridge, UK, 2012.

[BGP11] P. Bulens, D. Giry, and O. Pereira. Running Mixnet-Based Elections with Helios. In H. Shacham and V. Teague, editors, *EVT/WOTE'11, Electronic Voting Technology Workshop/Workshop on Trustworthy Elections*, San Francisco, USA, 2011.

[FMM+02] J. Furukawa, H. Miyauchi, K. Mori, S. Obana, and K. Sako. An Implementation of a Universally Verifiable Electronic Voting Scheme based on Shuffling. In M. Blaze, editor, *FC'02, 6th International Conference on Financial Cryptography*, LNCS 2357, pages 16–30, Southampton, Bermuda, 2002.

[FMS10] J. Furukawa, K. Mori, and K. Sako. An Implementation of a Mix-Net Based Network Voting Scheme and Its Use in a Private Organization. In D. Chaum, M. Jakobsson, R. Rivest, P. Y. A. Ryan, J. Benaloh, M. Kutylowski, and B. Adida, editors, *Towards Trustworthy Elections*, LNCS 6000, pages 141–154. Springer, 2010.

[FS86] A. Fiat and A. Shamir. How to Prove Yourself: Practical Solutions to Identification and Signature Problems. In A. M. Odlyzko, editor, *CRYPTO'86, 6th Annual International Cryptology Conference on Advances in Cryptology*, pages 186–194, 1986.

[FS01] J. Furukawa and K. Sako. An Efficient Scheme for Proving a Shuffle. In J. Kilian, editor, *CRYPTO'01, 21st Annual International Cryptology Conference on Advances in Cryptology*, LNCS 2139, pages 368–387, Santa Barbara, USA, 2001.

[GI08] J. Groth and Y. Ishai. Sub-Linear Zero-Knowledge Argument for Correctness of a Shuffle. In N. Smart, editor, *EUROCRYPT'08, 27th International Conference on the Theory and Applications of Cryptographic Techniques*, LNCS 4965, pages 379–396, Istanbul, Turkey, 2008.

[Gro10] J. Groth. A Verifiable Secret Shuffle of Homomorphic Encryptions. *Journal of Cryptology*, 23(4):546–579, 2010.

[HS11] R. Haenni and O. Spycher. Secure Internet Voting on Limited Devices with Anonymized DSA Public Keys. In H. Shacham and V. Teague, editors, *EVT/WOTE'11, Electronic Voting Technology Workshop/Workshop on Trustworthy Elections*, 2011.

[JCJ05] A. Juels, D. Catalano, and M. Jakobsson. Coercion-Resistant Electronic Elections. In V. Atluri, S. De Capitani di Vimercati, and R. Dingledine, editors, *WPES'05, 4th ACM Workshop on Privacy in the Electronic Society*, pages 61–70, Alexandria, USA, 2005.

[Joa14] R. Joaquim. How to Prove the Validity of a Complex Ballot Encryption to the Voter and the Public. *Journal of Information Security and Applications*, accepted, 2014.

[Mau09] U. Maurer. Unifying Zero-Knowledge Proofs of Knowledge. In B. Preneel, editor, *AFRICACRYPT'09, 2nd International Conference on Cryptology in Africa*, volume 5580 of *LNCS 5580*, pages 272–286, Gammarth, Tunisia, 2009.

[Nef01] C. A. Neff. A Verifiable Secret Shuffle and its Application to E-Voting. In P. Samarati, editor, *CCS'01, 8th ACM Conference on Computer and Communications Security*, pages 116–125, Philadelphia, USA, 2001.

[RBH+09] P. Y. A. Ryan, D. Bismark, J. Heather, S. Schneider, and X. Zhe. Prêt à Voter: a Voter-Verifiable Voting System. *IEEE Transactions on Information Forensics and Security*, 4(4):662–673, 2009.

[TW10] B. Terelius and D. Wikström. Proofs of Restricted Shuffles. In D. J. Bernstein and T. Lange, editors, *AFRICACRYPT'10, 3rd International Conference on Cryptology in Africa*, LNCS 6055, pages 100–113, Stellenbosch, South Africa, 2010.

[Wik09] D. Wikström. A Commitment-Consistent Proof of a Shuffle. In C. Boyd and J. González Nieto, editors, *ACISP'09, 14th Australasian Conference on Information Security and Privacy*, LNCS 5594, pages 407–421, Brisbane, Australia, 2009.

[Wik12] D. Wikström. *How to Implement a Stand-alone Verifier for the Verificatum Mix-Net*. Verificatum AB, Stockholm, Sweden, 2012.

[Wik13] D. Wikström. *User Manual for the Verificatum Mix-Net Version 1.2.0*. Verificatum AB, Stockholm, Sweden, 2013.

Coercion-Resistant Internet Voting in Practice

Christian Feier, Stephan Neumann, Melanie Volkamer

Technische Universität Darmstadt / CASED
Hochschulstrae 10
64289 Darmstadt
name.surname@cased.de

Abstract: Internet voting continues to raise interest both among research and society. Throughout the last decades, many Internet voting schemes have been developed, each one providing particular properties such as receipt-freeness or end-to-end verifiability. One attractive scheme is the JCJ / Civitas scheme due to its property of making coercion attacks ineffective. Neumann and Volkamer [NV12] analyzed the scheme and identified significant usability issues. To overcome these drawbacks, the authors extended the original work by integrating smart cards. In a follow-up work, Neumann et al. [NFVK13] conducted a theoretical performance analysis for this extension and improved the extension towards its applicability in real-world elections. Their analysis left several real-world considerations open for future work. The present work addresses these gaps: We present a prototype implementation of the revised extension and assess its real-world performance. Based on this contribution, we are able to conclude that the revised extension is feasible to be used in real-world elections.

1 Introduction

Internet voting continues to attract interest both among research and society. Starting with Chaum's seminal work [Cha81], throughout the last decades, many Internet voting schemes have been developed. e.g. FOO [AF92], Helios [Adi08], and the Estonian Internet voting scheme [MM06]. One scheme of particular relevance is the JCJ / Civitas scheme [JCJ05, CCM08] due to its promised coercion-resistance property. Coercion is one of the crucial problems that accompanies the conduction of remote elections such as Internet voting [KV06]. While the JCJ / Civitas scheme addresses the highly valuable security requirements of coercion-resistance, the scheme lacks usability due to sophisticated credential handling as outlined by Neumann and Volkamer [NV12]. To overcome this shortcoming, Neumann and Volkamer revised the scheme. Their extension builds upon the integration of smart cards to simplify the credential handling to improve the usability. While their proposal certainly improves the JCJ / Civitas scheme from a usability perspective, it turns out that their extension does not take performance into account. As a consequence thereof, it includes several routines which cannot be implemented sufficiently performant on smart cards, e.g. the proof of well-formedness. In a follow-up work, Neumann et al. [NFVK13] conducted an improvement of the extension towards its applicability in real-world elections by modifying or removing routines which do not run

performant on smart cards. As a second part they made a theoretical performance analysis for the revised extension. Their work did not consider varying variables like network- and transmission time or smart card operations like copy data from non-persistent to persistent storage.

To address these shortcomings, the present work takes up the [NFVK13] scheme. As a contribution it analyzes the need of all changes made from [NFVK13] and provides some alternative ways to improve its performance. In addition to that, the present work provides ideas how to handle certificates and precises the calculations with respect to network traffic for 2048 bit encryption. As a last contribution a prototype including all made modifications will be presented.

The remainder of this work is structured as follows: In Section 2, we provide a short overview about the [NFVK13] scheme. In Section 3, we explain the made design decisions in terms of implementation details of the prototype followed by a short system design explanation in Section 4. In Section 5, we reanalyze the performance with respect to 2048 bit encryption blocks and the suggested certificate handling. Furthermore we make a practical test using our prototype and compare these results with the theoretical results. Later on in Section 6 we present the interfaces of our prototype and guide through an election. Section 7 concludes this work and possible directions for future research are outlined.

2 The NFVK13 Scheme Overview

The [NFVK13] scheme builds upon the extension by Neumann and Volkamer [NV12]. The scheme proofed the feasibility for real-world elections by optimizing the registration- and voting phase by changing and removing some unnecessary subroutines. In this section we summarize the [NFVK13] scheme.

Setup Phase. The supervisor initialises the election and publishes all details on the bulletin board. The *tabulation tellers* generate a distributed ElGamal election key pk_{EK} and publish the corresponding public key on the bulletin board.

Registration Phase. As a first step a voter *v* personally consults the *supervised registration authority* (SRA). The authority checks that the voter is not under direct influence of any coercers. The voter is then requested to insert her smart card into the card reader and the SRA authenticates to the smart card as the SRA. In the next step she is asked to set her personal owner PIN. Afterwards the SRA stores the private supervised credential share c^v_{SRA} on the voter's smart card which is a normal credential share as used by the registration tellers. The smart card will then calculate the coefficient $c_p = \frac{c^v_{SRA}}{PIN}$ and store it. This is the only value which is stored during this consultation. The voter leaves the SRA afterwards.
To a later moment in time the voter connects to the election website to finish the registration process. She is asked to choose her preferred registration tellers out of all available registration tellers. After this all preferred registration tellers, from now on called

trusted registration tellers (TRT(v)), will be saved on the smart card. Afterwards a connection to each TRT will be established and the private credential share c^v_{RT} along with an encryption $S'_i = \{c^v_{RT}\}_{pk_{EK}}$ and the used random r will be directly forwarded to the smart card. At the same time the registration teller will publish the public credential share $S_i = \{c^v_{RT}\}_{pk_{EK}}$, which is a reencryption of c^v_{RT}, on the bulletin board. The smart card then verifies the DVRP proving that S'_i is a reencrypion of S_i from the bulletin board. After obtaining all private credential shares the smart cards calculates the credential factor

$$c_{factor} = c_p \cdot \prod_{i \in TRT(v)} c_{RT_i}$$

Voting Phase. After finishing the registration phase the voter can start the voting process. To do this she visits the election website. On this website she can select the candidates of her choice. After the selection is finalized the selection is forwarded to the smart card which then encrypts the selection using the election key pk_{EK} and asks the voter to enter her owner PIN. The card then calculates the private credential as follows

$$c = PIN \cdot c_{factor}$$

If she entered the real PIN it will result in the voter's real credential to be calculated, if not a fake credential is generated. This credential will be encrypted using pk_{EK} and calculates the proof of knowledge (PKCV) which shows that the voter knows both, the credential and the vote. Furthermore the smart card computes $h(\{vote\}_{pk_{EK}}||\{c\}_{pk_{EK}}||PKCV)$ as a verification code and displays this code on the card reader's display. The ballot will then be sent to all available ballot boxes.

Tallying Phase. In the tallying phase, all tabulation tellers retrieve the ballots from all ballot boxes and the public credentials stored on the bulletin board. Zero-knowledge proofs are verified, duplicates (due to vote-updating) and unauthorized votes (due to the use of fake credentials) are eliminated. Finally, encrypted credentials of remaining ballots are discarded and the respective encrypted votes are distributively decrypted. Each step of the tabulation tellers is publicly verifiable based on a set of zero-knowledge proofs.

3 Design Decisions

[NFVK13] made some recommendations we did not keep. In this section we explain the made design decisions.

Multos vs. Java Card. In [NFVK13] the calculations are on the basis of Java Cards and the timings from [BCGS09]. As proposed in [NFVK13] Java cards only offer some high level methods like RSA encryption, but do not offer any methods to directly calculate modular operations like addition, subtraction, multiplication and division on the crypto co-processor. Due to this for example the modular multiplication needs to be calculated using a mapping on RSA CRT decryption. This means one modular multiplication on

Java cards needs as long as one RSA CRT decryption which produces a big overhead. Multos instead offers direct crypto co-processor access using MEL assembler commands. Therefore on Multos cards it is possible to calculate modular operations directly on the crypto co-processor which results in better timings.
Due to the advantages over Java cards we switched for our implementation from Java- to Multos cards.

Protocol Revision. Neither [CCM08] nor [NV12] define protocols for the communication between registration teller and the bulletin board and communication between the smart card and the bulletin board. Furthermore at the beginning of the implementation the Multos C-API did not offer a mapping on AES. Therefore, the protocol has been changed to use RSA encryption and signatures to authenticate and exchange data with the registration teller. [CCM08] uses Needham-Schroeder-Loewe protocol (NSL) to authenticate to the registration teller and to exchange an AES key for further communication. The NSL protocol needs two RSA encryptions to authenticate and exchange the AES key. The proposed protocol change needs three RSA encryptions to authenticate and exchange all data. Assuming that one RSA encryption is at most as slow as the data communication using AES, the feasibility of the [NFVK13] protocol is shown if the feasibility with the modified protocol can be proven. In the protocol we use the following notation: The entities are the smart card (S), Registration Teller (RT) and BulletinBoard (BB). The used keys are the public (pk_X) and private (sk_X) RSA key of entity *X* and the public ElGamal election key (pk_{EK}). $\{T\}_K$ describes the encryption of *T* using the public key *K* and $Sig_K(T)$ is the signature of *T* signed with the private key *K*. Furthermore N_X denotes the nonce generated by entity *X*, ID_X is the ID of entity *X* and s_i denotes the credential share of registration teller i and S_i is the ElGamal encrypted credential share using the public election key. A protocol step is denoted by X $\rightarrow$ Y: T which means that entity *X* sends the message *T* to entity *Y*. The modified protocol is described as follows:

1. S $\rightarrow$ BB: request registration teller certificate
2. BB $\rightarrow$ S: registration tellers certificate
3. S $\rightarrow$ RT: N_V, ID_V, trusted RTs, $Sig_{sk_V}(N_V||ID_V||\text{trusted RTs})$
4. RT $\rightarrow$ BB: $\{N_{RT}, ID_V\}_{pk_{BB}}$
5. BB $\rightarrow$ RT: pk_V, pk_{EK}, $Sig_{sk_{BB}}(pk_V, pk_{EK}, N_{RT}, ID_V)$
6. RT $\rightarrow$ BB: N'_{RT}, ID_{RT}, ID_V, trusted RTs, S_i, $Sig_{sk_{RT}}(N'_{RT}||ID_{RT}||ID_V\text{trusted RTs}||S_i)$, $Sig_{sk_{RT}}(\text{trusted RTs}, S_i, ID_{RT})$
7. RT $\rightarrow$ S: $\{(s_i, r)\}_{pk_V}, S_i, Sig_{sk_{RT}}(s_i||r||S_i)$
8. S $\rightarrow$ BB: S_i, $\{N'_V, ID_V, \text{trusted RTs}\}_{pk_{BB}}$ $Sig_{sk_V}(S_i||\{N'_V, ID_V, \text{trusted RTs}\}_{pk_{BB}})$
9. BB $\rightarrow$ S: $\{True/False, True/False, N'_V\}_{pk_V}$

In step 1 and 2, the specific registration teller certificate will be requested and loaded onto the card. In step 3 the card authenticates to the registration teller due to its knowledge

of sk_V. In step 4 the registration teller requests the public key of V which the bulletin board responds in step 5. The registration teller can verify the identity of the card. If the verification is successful it will send the trusted registration tellers and the encrypted credential share of voter V to the bulletin board and sends the same encrypted credential share and the credential share to the voter in step 7. To save performance only the private credential share s_i and used randomness r are send encrypted. The public credential share is sent unencrypted. To ensure its integrity, the data is signed. Due to the fact the the resources of the smart card are limited, the signature of $s_i||r||S_i$ is generated so that the card does not need to store $\{(s_i, r)\}_{pk_V}$. After the card received the credential share and verified that the encrypted credential share is a reencryption it sends a request to the bulletin board in step 8. This request contains the trusted registration tellers and and the encrypted credential share. The bulletin board will respond with if the encrypted credential is on the bulletin board and if the trusted tion tellers are correct. The used nonce N'_V which was sent encrypted in step 8 is used as an integrity check. This nonce N'_V will be resent by the bulletin board in step 9. The smart card will check whether the nonce is the same and only accept the credential share of registration teller i if the check passed successfully.

PIN Handling. To avoid side channel attacks to spy out the PIN code during the voting process [NFVK13] recommends to calculate coefficients c_p and c_{factor} so that the smart card does not need to know the real PIN but be able to calculate the credential depending on the entered PIN. As a drawback this way costs some performance. Through the security model in [NFVK13] the smart card needs to be trusted. Therefore it is not necessary that the card does not know the correct PIN as long as there is no way that the PIN leaves the card. Therefore the implemented prototype is saving the real PIN and compares it with the entered PIN. It will always compare the full PIN and will not stop at the first wrong digit. If the entered PIN is correct the PIN will be hashed using SHA-1 and the real credential c will be encrypted. If the entered PIN is wrong the PIN will be hashed and instead of the real credential the hash value of the wrong PIN will be encrypted. Using this way the card will always perform the same steps independent of the entered PIN which will prevent side channel attacks. Additionally it is not necessary to calculate the coefficients c_p and c_{factor} which will result in a slightly better performance in the registration phase.

Certificate Handling. To ensure authenticity of messages from the authorities, namely the registration tellers and bulletin board, an authentic RSA key is necessary. A card should be used for several elections but due to the lack of space on the smart card only a small number of certificates could be stored. To solve this problem only one certificate called *Root Certificate* will be stored on the card. This Root Certificate could be set during the production process of the card and can not be changed after it is set once.
To initialise a secure communication to an authority the, authority's certificate needs to be loaded on the card at first. The card contains a slim DER parser[1] which extracts the necessary information from the certificate. To be authentic a certificate needs to be signed from the *Root Certificate*, must be valid and must match a subjectDN pattern e.g. the common name CN must start with `RT` if the certificate belongs to a registration teller.

[1] Distinguished Encoding Rules parser after RFC 5280 http://www.ietf.org/rfc/rfc5280.txt

Thereby, a trust chain will be built and only one certificate needs to be stored permanently on the card and there is no restriction on how many elections can be held with the card. As a drawback one should mention that the card needs to verify a certificate at every reload which may lower the performance.

4 System Architecture

In Figure 1, the system design is depicted. The responsibility of each component will be explained.

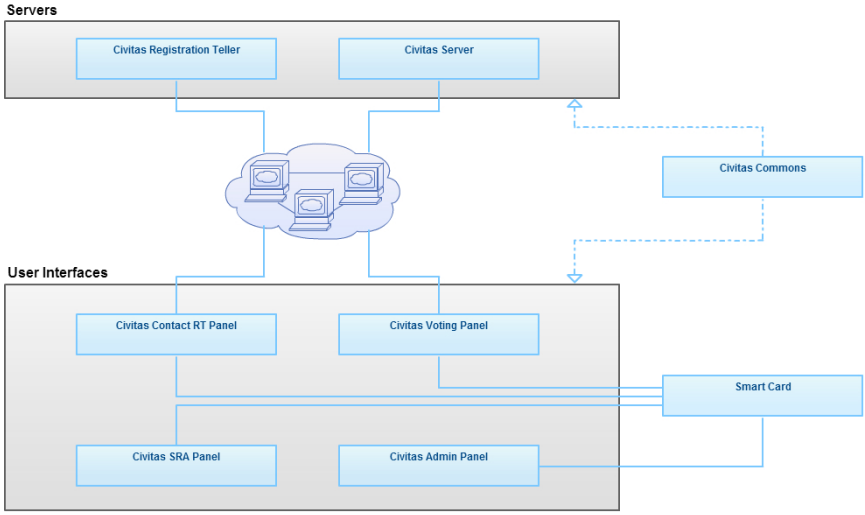


Figure 1: System Design

Civitas Server: The Civitas Server is a slim implementation of a real Civitas Server. It contains the bulletin board and offers an interface to communicate with the bulletin board. Furthermore this component includes the tallying module.
Civitas Registration Teller: The registration teller component is contacted by the voter to receive her private credential share. Several instances of registration teller components can run at different locations. The registration teller will be contacted remotely by the contact RT panel component.
Civitas Contact RT Panel: This component is a JApplet which offers a graphical user interface for the voter to contact a registration teller component and get her private credential share. This JApplet can directly communicate with the smart card.
Civitas Voting Panel: The voting panel is a JApplet which offers a graphical user interface for the voter to make her vote. It can directly communicate with the smart card.
Civitas SRA Panel: This component offers a graphical user interface for the supervised registration authority to authenticate to a smart card, allow the voter to set her owner PIN

and set the supervised credential share to the smart card. Furthermore it transmits the current election parameters to the smart card. The component can directly communicate to the smart card.
Civitas Admin Panel: The administration panel offers a graphical user interface to initialize the smart card. It offers methods to set the UUID, set the root certificate and the signed card certificate. The admin panel is an optional panel and is not necessary to run an election. It can directly communicate with the card.
Smart Card: The smart card component is responsible to manage the credential shares, to generate the ballot which will be send to the bulletin board and to generate the verification code which will be shown on the card reader's display.
Civitas Commons: The commons package is used by all components. It includes commonly used classes like communication messages, offers an interface to communicate with the smart card and contains the whole cryptography methods.

5 Performance Analysis

In this section we concretize the formulas from [NFVK13] with respect to the changes made to the protocol described in section 3 and present results for a practical run from our prototype.

5.1 Smart Card Timings

The modular operations are *addition* ($a + b \bmod p$), *subtraction* ($a - b \bmod p$), *multiplication* ($a \cdot b \bmod p$), *modular exponentiation* ($a^b \bmod p$) and modular division $\frac{a}{b} \bmod p = a \cdot b^{p-2} \bmod p$, which corresponds to 1 subtraction, 1 multiplication, and 1 exponentiation. The measured timings for the operations are found in 1.

Card / Operation	Addition	Subtraction	Multiplication	Expo.	Division
Java Card 1536 bit	0.082	0.082	0.517	0.430	1.029
Multos 1536 bit	0.038	0.038	0.047	0.421	0.506
Multos 2048 bit	0.056	0.058	0.066	1.272	1.396

Table 1: Java and Multos Timing comparison

5.2 Concretize Formulas

Registration Phase. The formulas proposed in [NFVK13] only considered asymmetric cryptography operations like RSA and ElGamal encryption but did not consider the symmetric AES encryption which is used to secure the communication. The modified protocol described in section 3 only uses asymmetric encryption schemes to ensure a secure com-

munication. Furthermore as stated in section 3, for a communication to an authority the certificate of the authority needs to be loaded on the card and then verified. To verify a certificate, a signature needs to be verified using `SHA1WithRSAEncryption`. The whole process to transmit a 2048 bit RSA key certificate with `SHA1WithRSAEncryption` signature, parse it and verify it takes approximately 4.4 seconds. For further reference, we define the whole certificate verification process as $t_{verify} = 4.403$ seconds.
We consider 2048 bit encryption which means that up to 255 bytes can be encrypted at once using RSA. Also we only analyse all steps in which the smart card is involved, namely step 3, 7, 8 and 9. To initiate the communication, the registration teller certificate needs to be loaded. To generate step 3 the card needs to generate one signature which results in one modular exponentiation and one certificate verification. Assuming that the credential share and the used randomness needs up to 255 byte, one RSA encryption and one signature generation is necessary to transmit the data. After decryption the reencrypted credential share needs to be validated by performing one ElGamal encryption. Summed up step 7 needs four modular exponentiations and one modular multiplication.
To perform step 8, communication to the bulletin board is required which means a new certificate needs to be loaded and verified. Furthermore, one RSA decryption and one signature verification is necessary which results in two modular exponentiations and one certificate verification for step 8. For step 9, one RSA decryption which means one modular exponentiation is required. After all registration tellers have been contacted, the credential c can be generated by multiplying all credential shares s_i. In summary, the time needed throughout the registration phase is given as follows:

$$\begin{aligned} t_{registration} &= (1 \cdot t_{exp} + 1 \cdot t_{verify} + 4 \cdot t_{exp} + 1 \cdot t_{mul} + 2 \cdot t_{exp} + 1 \cdot t_{verify} + 1 \cdot t_{exp}) \\ &\quad \cdot |TRT| + t_{mul} \cdot |TRT| \\ &= (8 \cdot t_{exp} + 1 \cdot t_{mul} + 2 \cdot t_{verify}) \cdot |TRT| + t_{mul} \cdot |TRT| \end{aligned}$$

Voting Phase. In order to vote the credential and the vote needs to be encrypted and one proof of knowledge (PKCV) needs to be performed which results in four modular exponentiations and two multiplications for the two ElGamal encryptions and two modular exponentiations, two modular multiplications and two modular subtractions for the PKCV proof. Hence, the needed time is given as follows:

$$\begin{aligned} t_{voting} &= 4 \cdot t_{exp} + 2 \cdot t_{mul} + 2 \cdot t_{exp} + 2 \cdot t_{mul} + 2 \cdot t_{sub} \\ &= 6 \cdot t_{exp} + 4 \cdot t_{mul} + 2 \cdot t_{sub} \end{aligned}$$

5.3 Practical Results

The results from section 5.2 do not consider network traffic and transmission time between the client computer and the smart card. In this section we present our results from a live test of our prototype. The test system consists of four computers (one voter client, two registration tellers and one bulletin board) connected via a 100 Mbit LAN connection. The used smart card is a Multos ML3 and the used card reader is an Omnikey 3821. The needed time will be measured after all user interaction is done, means for the registration phase

after the trusted registration tellers are chosen and the 'proceed' button is pressed. For the voting phase it means after all candidates are chosen and the PIN is entered. To contact two registration tellers 46.782 seconds were needed, which means 23.391 per registration teller and the voting phase took 12.979 seconds. The theoretical solution took with respect to the timings from table 1 and $t_{verify} = 4.403$ results in a registration time of 38.228 (i.e. 19.114 seconds per registration teller) and a voting time of 7.809 seconds. The theoretical analysis differs from the practical results. This is reasonable due to the fact that the theoretical analysis did not consider network and transmission traffic. Furthermore it did only consider modular operations like modular addition and the time needed to verify a certificate. Other processes on the card like hashing, copying data from non-persistent to persistent storage, reset data and many other operations could not be considered. We think that 30 seconds per registration teller and 13 seconds to vote are still feasible.

6 Interface

After the system has been implemented from a functionality perspective, user interfaces have been developed. These have been developed in cooperation with interface design experts and designers. In this section we present the first prototypes to guide a voter through the voting process. The interface is placed in a webbrowser using HTML as a markup language and a JApplet component to communicate with the card reader on the clients computer. Using this way the voter just needs to connect to a website and does not need to download additional software.

First, a voter connects to the election website (Figure 2). On this website the voter finds all election related information, namely the electoral register, the candidates, the authorities, the public authority keys, and the public election key. Furthermore, the voter can find all received ballots and encrypted credential shares on the bulletin board. There are several independent institutes which offer the voting software. The voter can decide over which institute she casts her vote.

On the welcome site (Figure 3) of the voting institute the voter is asked whether she already contacted her supervised registration authority. If not, she has to do this first and come back to this site. Furthermore she can choose if she already contacted her trusted registration tellers (for example in case of revoting). If she did not contact them before, she will be directed to the trusted registration tellers contact site (Figure 4). On this site a JApplet will be loaded. The voter will be instructed to insert her smart card into the card reader and choose her trusted registration tellers by simply clicking on them in the JApplet. After the voter made her choice, she confirms by the 'contact' button. Thereafter, the smart card contacts all chosen registration tellers and eventually returns a notifications once the registration process has been finished. The voter can start the voting process by clicking 'proceed'.

On the voting site (Figure 5) another JApplet will be loaded. The voter is requested to insert her smart card into the card reader if not already done and to decide which candidate/s she wants to vote for. After pressing 'vote' she will be asked to enter the candidates ID into

the card reader's PIN pad and submit it via the the card reader's submit button. The card asks for the owner's PIN and correspondingly calculates the ballot as described in section 2. After the calculation process is done, the card will send the ballot to the bulletin board. The corresponding verification code will now be displayed on the card reader's display.

After casting her vote, the voter is directed to the last site (Figure 6). This site informs the voter that her ballot is cast successfully. Furthermore it informs her that she can verify that her vote by visiting the bulletin board and check if her verification appears.

7 Conclusion and Future Work

Throughout the last decades, many Internet voting schemes have been proposed in the literature. One of these schemes is the coercion-resistant JCJ / Civitas [JCJ05, CCM08] scheme. As shown by Neumann and Volkamer [NV12], the scheme builds upon a number of trust assumptions and lacks in usability. Correspondingly, Neumann and Volkamer presented a proposal [NV12] to overcome these shortcomings by incorporating smart cards into the JCJ / Civitas scheme. Nevertheless, their work did not consider performance. In 2013, Neumann et al. [NFVK13] therefore analyzed the previously named proposal with regard to its feasibility. After slight protocol modifications, the authors were able to theoretically conclude the feasibility of the JCJ / Civitas extension.

As a final step of the previously outlined history, the present work proves the feasibility in a practical way. In the first part, we gave a short overview about the [NFVK13] scheme and afterwards outlined further design decisions made throughout the implementation of the scheme as well as the system design. In the second part, we refined the previous theoretical performance analysis from [NFVK13] and made a practical analysis. Third, we presented interface prototypes and provided a corresponding walktrough through the voting process. Summing up these contributions, we are able to conclude that that [NFVK13] is feasible to be applied in real-world elections.

In the future we plan to optimize the performance of the scheme. This can be done by exchanging space for performance e.g. by caching the bulletin board certificate. This would save one verification (currently 4 seconds) per registration teller in the registration phase. Furthermore optimizations for specific card types are imaginable e.g. by using the full non-persistent memory. This would save a few copy routines and would speed up the whole routine.

Acknowledgment. This work has been developed within the project ComVote, which is funded by the Center for Advanced Security Research Darmstadt (CASED), Germany. We also thank the reviewers for their valuable comments that helped to considerably improve the quality of this work.

References

[Adi08] Ben Adida. Helios: Web-based Open-Audit Voting. In Paul C. van Oorschot, editor, *USENIX Security Symposium*, pages 335–348. USENIX Association, 2008.

[AF92] K Ohta A. Fujioka, T. Okamoto. A practical secret voting scheme for large scale elections. *J. Seberry and Y. Zheng, editors, Advances in Cryptology AUSCRYPT'92*, page 244251, 1992.

[BCGS09] Patrik Bichsel, Jan Camenisch, Thomas Groß, and Victor Shoup. Anonymous credentials on a standard java card. In *ACM Conference on Computer and Communications Security*, CCS '09, pages 600–610. ACM, 2009.

[CCM08] Michael R. Clarkson, Stephen Chong, and Andrew C. Myers. Civitas: Toward a Secure Voting System. In *IEEE Symposium on Security and Privacy*, pages 354–368. IEEE Computer Society, 2008.

[Cha81] David L. Chaum. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM*, 24(2):84–90, 1981.

[JCJ05] Ari Juels, Dario Catalano, and Markus Jakobsson. Coercion-resistant electronic elections. In *ACM Workshop on Privacy in the Electronic Society*, pages 61–70. ACM, 2005.

[KV06] Robert Krimmer and Melanie Volkamer. *Observing Threats to Voters Anonymity: Election Observation of Electronic Voting*. na, 2006.

[MM06] lle Madise and Tarvi Martens. E-voting in Estonia 2005. The first Practice of Countrywide binding Internet Voting in the World. In Robert Krimmer, editor, *Electronic Voting*, volume 86 of *LNI*, pages 15–26. GI, 2006.

[NFVK13] Stephan Neumann, Christian Feier, Melanie Volkamer, and Reto Koenig. Towards A Practical JCJ / Civitas Implementation. Cryptology ePrint Archive, Report 2013/464, 2013. `http://eprint.iacr.org/`.

[NV12] Stephan Neumann and Melanie Volkamer. Civitas and the Real World: Problems and Solutions from a Practical Point of View. In *International Conference on Availability, Reliability and Security (AReS 2012)*, pages 180–185. IEEE Computer Society, 2012.

A Interface Mockups

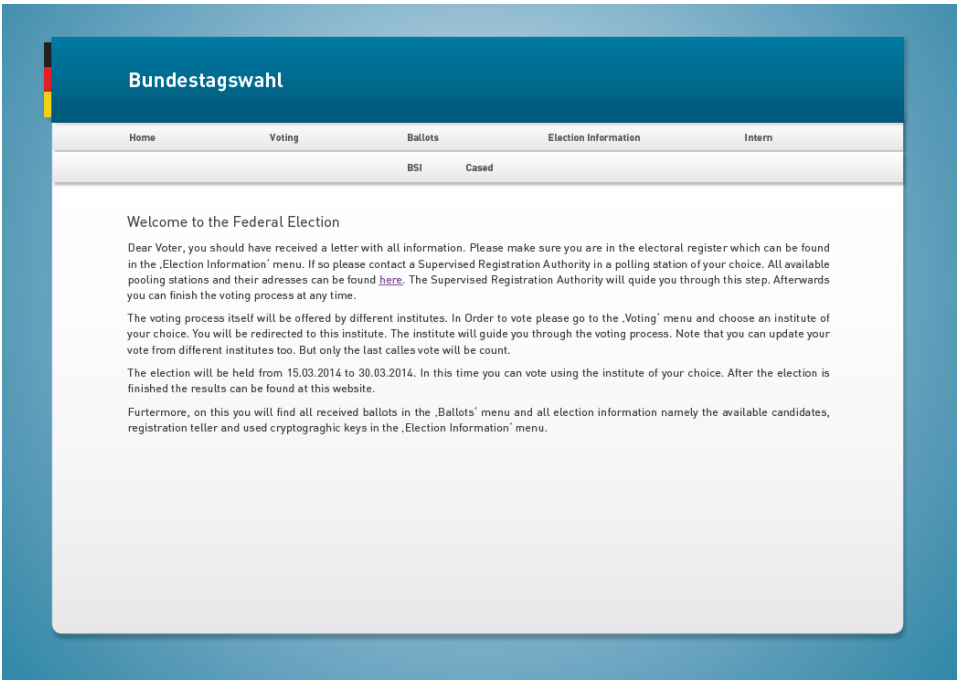


Figure 2: Bulletin Board

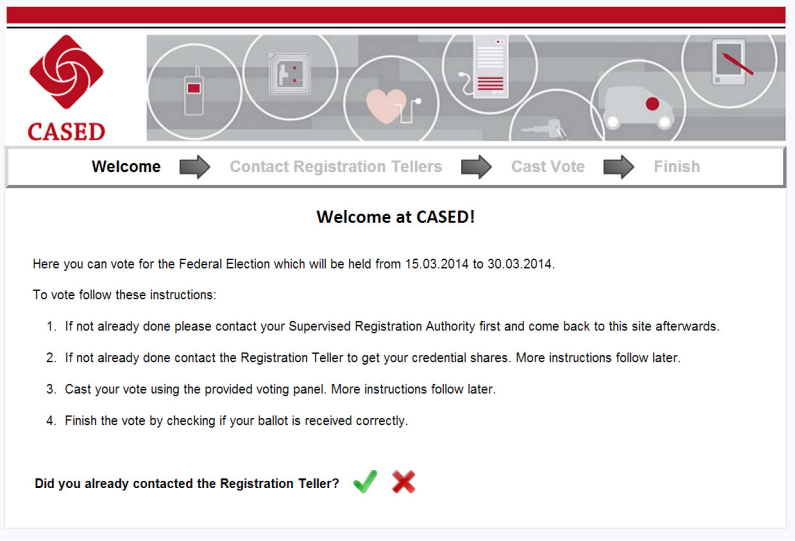


Figure 3: Vote Welcome Site

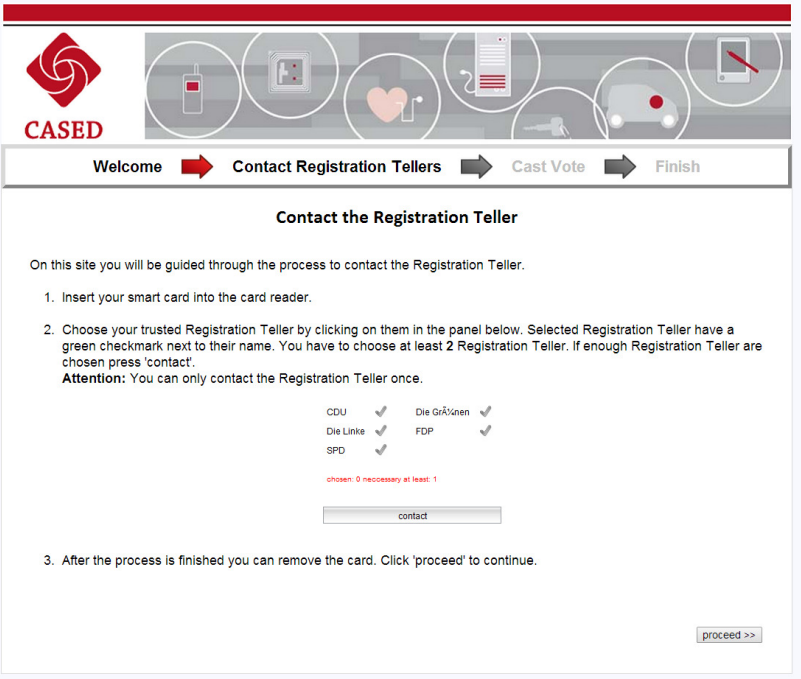


Figure 4: Contact Registration Teller Site

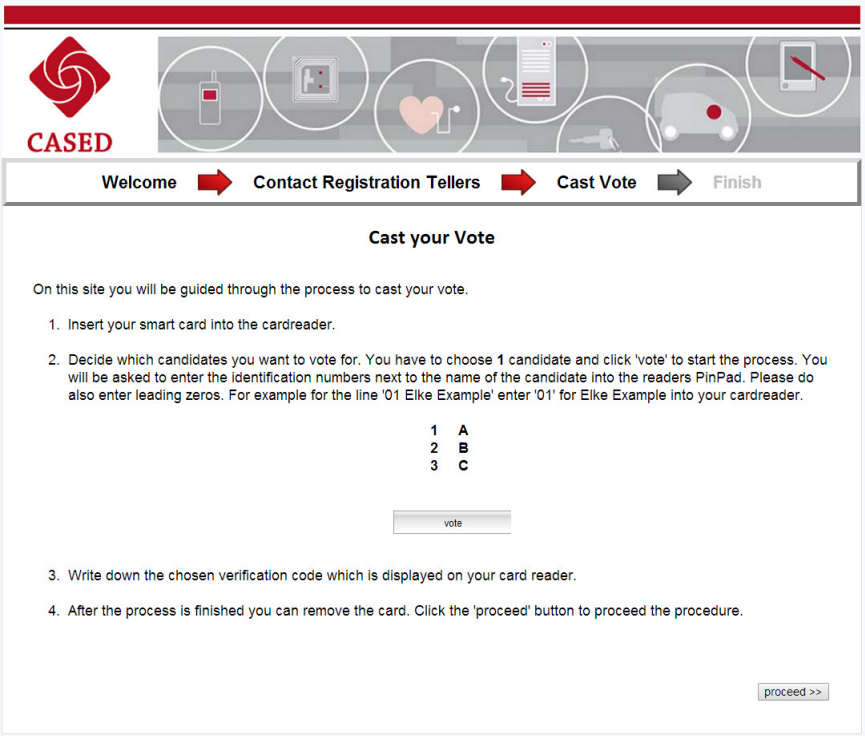


Figure 5: Cast Ballot Site

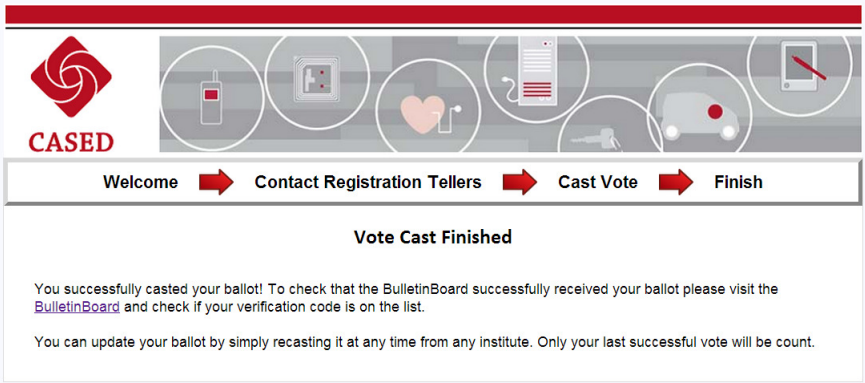


Figure 6: Vote Cast Finish Site

Entwicklung eines Common Criteria Schutzprofils für elektronische Wahlgeräte mit Paper Audit Trail

Jurlind Budurushi, Stephan Neumann, Genc Shala, Melanie Volkamer

Security, Usability and Society
Technische Universität Darmstadt
Hochschulstraße 10
64289 Darmstadt
name.surname@cased.de

Abstract: Mit dem Urteil vom 3. März 2009 hat das Bundesverfassungsgericht die bislang in der Bundesrepublik Deutschland eingesetzten Wahlgeräte für verfassungswidrig erklärt. Grund für dieses Urteil war die fehlende Umsetzung des Prinzips der *Öffentlichkeit der Wahl*. Mit dem Urteil erklärte das Gericht jedoch nicht grundsätzlich den Einsatz elektronischer Wahlgeräte für verfassungswidrig. Im Rahmen des von der DFG geförderten Projekts 'VerKonWa' wurde das EasyVote System entwickelt, welches den Öffentlichkeitsgrundsatz durch sogenante Paper Audit Trails umsetzt. Im Rahmen dieser Arbeit berichten wir über die Erfahrung bei der Entwicklung eines Common Criteria Schutzprofils für elektronische Wahlgeräte mit Paper Audit Trails.

1 Einführung

In Deutschland waren bis zum Jahr 2005 elektronische Wahlgeräte im Einsatz. Diese sowie die entsprechende Bundeswahlgeräteverordnung wurden mit dem Urteil des Bundesverfassungsgerichts vom 3. März 2009 [Bun09] für verfassungswidrig erklärt, weil weder die Geräte noch die Verordnung mit dem Prinzip der *Öffentlichkeit der Wahl* vereinbar sind. Dieses Urteil untersagt nicht grundsätzlich den Einsatz elektronischer Wahlgeräte, solange diese mit der *Öffentlichkeit der Wahl* vereinbar sind, d.h. jeder Wähler die wesentlichen Schritte der Wahlhandlung und Ergebnisermittlung zuverlässig und ohne besondere Sachkenntnisse überprüfen kann [HVB12]. Diese Überprüfbarkeit wird in der technischen Literatur als Verifizierbarkeit bezeichnet.

Im Rahmen des von der DFG geförderten Projekts 'VerKonWa' wurde das EasyVote System entwickelt, welches den insbesondere geforderten Öffentlichkeitsgrundsatz durch sogenante Paper Audit Trails umsetzt. Für diesen Ansatz sollte im Rahmen dieser Arbeit ein Common Criteria Schutzprofil entwickelt werden, welches sicherstellt, dass die rechtlichen Vorgaben - u.a. die Öffentlichkeit der Wahl - von zukünftigen Produkten, die nach diesem Schutzprofil ervaluiert werden, umgesetzt werden. Dazu sollten als Input existierende Schutzprofile verwendet werden und eine Methode verwendet werden, die zur systemantischen Entwicklung von Schutzrprofilen aus rechtlichen Vorgaben vorgestellt wur-

de [SDNK+13]. In dieser Arbeit berichten wid über die unterschiedlichen Quellen und die Erfahrung bei der Erstellung eines Schutzprofils.

2 Related Work: Existierende Schutzprofile

Es existieren bereits eine Reihe von Common Criteria (CC) Schutzprofilen im Kontext von "elektronischen Wahlen", die wir als Grundlage nehmen. Diese werden im Folgenden vorgestellt:

Schutzprofil: Digitales Wahlstift-System (BSI-PP-0031). Das Schutzprofil [Bun07] zielt insbesondere auf den Einsatz *Digitalen Wahlstift-Systems* [AMBS07] für politische Wahlen ab. Der Evaluierungsgegenstand besteht aus folgenden Komponenten: 1) Der "Digitale Wahlstift" und die zugehörige "Dockingstation", 2) Firmware zur Aufzeichnung der Stimme, 3) Datenbank zur Speicherung der elektronischen Stimmen (elektronische Urne) und 4) Software zur Kontrolle der Abläufe der Wahlhandlung, Bewertung, Auszählung und Feststellung des Ergebnisses. Das Schutzprofil entspricht den Vorgaben der CC-Standard Version 2.3 und schlägt die Vertrauenswürdigkeitsstufe EAL3+ vor.

Schutzprofil: Basissatz von Sicherheitsanforderungen an Online-Wahlprodukte (BSI-PP-0037). Dieses Schutzprofil [Bun08a] definiert die Sicherheitsanforderungen, welche von Online-Wahlsysteme erfüllt werden müssen. Der Evaluierungsgegenstand ist ein verteiltes Client-Server-System, welches aus folgenden Komponenten besteht: 1) Dienstnutzer (Client), d.h. die Anwendersoftware des Wählers und 2) Dienstleister (Server), d.h. die Serversoftware. Die client-seitige Komponente ermöglicht den Wählern ihre Stimme online abzugeben. Die server-seitige Komponente verwaltet die Wahlberechtigungsliste und die Urne. Das Schutzprofil entspricht den Vorgaben der CC-Standard Version 3.1 Revision 2 und schlägt die Vertrauenswürdigkeitsstufe EAL2+ vor.

Schutzprofil: IEEE P1583. Das "IEEE P1583" Schutzprofil definiert die minimalen Sicherheitsanforderungen, die elektronische Wahlgeräte umsetzen müssen, um mit den Richtlinien 2002 FEC (Federal Election Commission Guidelines) und HAVA (Help America Vote Act) konform zu sein. Bei diesem Schutzprofil wurde keine Methodik zur Herleitung technischer Anforderungen aus den spezifischen rechtlichen Anforderungen verwendet. Der Evaluierungsgegenstand besteht aus der Software und Hardware des Wahlgeräts. Das Schutzprofil entspricht den Vorgaben der CC-Standard Version 2.3 und schlägt die Vertrauenswürdigkeitsstufe EAL2 vor.

Schutzprofil: PP-CIVIS. Ähnlich zu dem vorherigen Schutzprofil definiert das CIVIS-Schutzprofil [Sec06] von Wahlgeräten zu erfüllende Sicherheitsanforderungen. Der Evaluierungsgegenstand ist ebenfalls die Software und Hardware des Wahlgeräts. Das Schutz-

profil entspricht den Vorgaben der CC-Standard Version 3.0 und schlägt die Vertrauenswürdigkeitsstufe EAL2+ vor.

Schutzprofil: Kontrollierte elektronische Wahlgeräte. Dieses Schutzprofil [KKY12] definiert Sicherheitsanforderungen, die von Wahlgeräten erfüllt werden sollten. In diesem Schutzprofil wird die Bedeutung der Verifizierbarkeit hervorgehoben, aber ähnlich zu allen vorherigen Schutzprofilen werden keine Anforderungen bezüglich Verifizierbarkeit definiert. Der Evaluierungsgegenstand wird nicht genau spezifiziert. Das Schutzprofil entspricht den Vorgaben der CC-Standard Version 3.1 und schlägt die Vertrauenswürdigkeitsstufe EAL3+ vor.

Schutzprofil: Sichere elektronische Wahlgeräte. Im Gegensatz zu den IEEE P1583 und PP-CIVIS Schutzprofilen behandelt dieses Schutzprofil [LLWK10] Kiosk-Wahlgeräte. Kiosk-Wahlgeräte sind über ein Netzwerk verbundene Wahlgeräte, über die Wähler am Wahltag ihre Stimme an öffentlichen Plätzen abgeben können. Die Besonderheit dieses Schutzprofiles besteht darin, dass es das einzige Schutzprofil ist, welches Verifizierbarkeit einbezieht und es als Sicherheitsziel des Evaluierungsgegenstandes definiert. Der Evaluierungsgegenstand besteht aus der Software und Hardware des Wahlgeräts. Das Schutzprofil entspricht den Vorgaben des CC-Standard Version 3.1 und schlägt die Vertrauenswürdigkeitsstufe EAL4+ vor.

Vergleich existierende Schutzprofile. Tabelle 1 stellt einen Vergleich aller existierenden Schutzprofile, die für elektronische Wahlgeräte entworfen wurden, dar. Der Vergleich zeigt, dass alle Schutzprofile außer "Sichere elektronische Wahlgeräte" Verifizierbarkeit nicht einbeziehen und somit nicht mit dem Prinzip der *Öffentlichkeit der Wahl* konform sind. Wie bereits von Buchmann et al. [BNV14] festgestellt wurde, deckt jedoch das Schutzprofil "Sichere elektronische Wahlgeräte" Verifizierbarkeit nicht adäquat ab: Einerseits wird Verifizierbarkeit als Sicherheitsziel des Evaluierungsgegenstandes definiert, jedoch wird keine Beziehung (rationale) zwischen Sicherheitszielen und Sicherheitsanforderungen angegeben. Andererseits führt die Erweiterung der Sicherheitsanforderungen zu einem Schutzprofil, das außerhalb des CC-Standards. Letztlich werden die Sicherheitsanforderungen nicht systematisch aus rechtlichen Anforderungen heraus abgeleitet, was die Rechtskonformität der danach evaluierten Produkte in Frage stellt.

Verifizierbarkeit und Common Criteria In diesem Zusammenhang ist auch die Diskussion von Buchmann et al. [BNV14] zum Thema Schutzprofilen im Kontext von elektronischen Wahlen von Interesse. Einerseits wird dort festgestellt, dass es der Common Criteria an Flexibilität fehlt, den gesetzlichen Gestaltungsspielraum [NKH+13] umzusetzen. Andererseits wird festgestellt, dass die Berücksichtigung der Verifizierbarkeit in Schutzprofilen als besondere Herausforderung gilt, da entsprechende CC Komponenten nicht vorhanden sind. Gemäß ihrer Erkenntnisse, schlagen Buchmann et al. vor, auf Annahmen an die Vertrauenswürdigkeit des Wahlsystems zu verzichten, um die Verifizierbarkeit so implizit in das Schutzprofil zu integrieren.

Schutzprofil	Version	EAL	EVG	Verifizierbarkeit
BSI-PP-0031	2.3	3+	Docking Station Firmware, E-Urne, Software	Nein
BSI-PP-0037	3.1	2+	Client Server	Nein
IEEE P1583	2.3	2	Software, Hardware	Nein
PP-CIVIS	3.0	2+	Software, Hardware	Nein
Kontrollierte ele. Wahlgeräte	3.1	3+	-	Nein
Sichere ele. Wahlgeräte	3.1	4+	Software, Hardware	Ja

Tabelle 1: Vergleich existierender Schutzprofile.

3 Methode zur Entwicklung von Schutzprofilen

Dieser Abschnitt widmet sich den von Simić-Draws et al. [SDNK+13] vorgeschlagenen Schnittstellen zwischen den Methodiken und Standards *Konkretisierung Rechtlicher Anforderungen* (KORA) [Roß08], *Common Criteria* (CC) [Int09] und *IT-Grundschutz* [Bun08b] Standard. Die Schnittestellen erlauben u.a. Schutzprofile auf Basis rechtlicher Vorgaben zu entwickeln. Entsprechend ist die Idee, diesen Ansatz zu verwenden, um ein Schutzprofil für elektronische Wahlgeräte mit Paper Audit Trails zu entwickeln. Zunächst werden die einzelnen Methodiken und Standards beschrieben, dann die Schnittstellen.

KORA - Konkretisierung Rechtlicher Anforderungen Die Methodik *Konkretisierung Rechtlicher Anforderungen* (KORA) wurde von der *Projektgruppe verfassungsverträgliche Technikgestaltung* (provet) an der Universität Kassel entwickelt. Ziel der Methodik ist die Entwicklung technischer Konzepte juristisch zu begleiten und somit die Verfassungsverträglichkeit der Technik zu sichern. KORA sieht dazu folgenden Prozess vor: Nachdem man die rechtlichen Vorgaben, die für die neue Technik relevant sind, identifiziert hat, werden folgende Schritte ausgeführt: 1) *Konkretisierung rechtlicher Anforderungen*: Risiken und Chancen werden beschrieben, welche durch den Einsatz der Technik einen Einfluss auf die rechtlichen Vorgaben haben; 2) *Konkretisierung rechtlicher Kriterien*: Einzelne Aspekte der Anforderungen werden betont, welche die Technik erfüllen soll; 3) *Ableitung technischer Ziele*: Funktionen werden beschrieben, die den rechtlichen Kriterien entsprechen sollen. Hier findet ein Übergang von der juristischen in die technische Sprache statt; 4) *Ableitung technischer Gestaltungsvorschläge*: Konkrete technische Gestaltungsvorschläge werden aus den technischen Zielen abgeleitet.

CC - Common Criteria Der Common Criteria (CC)-Standard wurde entwickelt, um Produkte der Informationssicherheit zu evaluieren. Der CC-Standard umfasst zwei wesentliche Grundkonzepte: Schutzprofile und Security Targets. Der Fokus dieser Arbeit liegt auf Schutzprofilen, welche im Gegensatz zu Security Targets allgemein und unabhängig von einem konkreten Produkt definiert werden. Ein Schutzprofil besteht aus folgenden Teilen: 1) Definition des Sicherheitsproblems, 2) Sicherheitsziele, 3) funktionale Sicherheitsanforderungen und Vertrauenswürdigkeitsanforderungen an das Produkt (die spezifische Technik). Alle Anforderungen werden in zwei Katalogen bereitgestellt: 1) Komponenten der funktionalen Sicherheit; 2) Komponenten der vertrauenswürdigen Sicherheit. Diese Komponenten werden weiter in Klassen und Familien unterteilt, welche zur Überführung und Konkretisierung von Sicherheitszielen in Anforderungen verwendet werden. Durch die Bereitstellung von Anforderungskatalogen ist ein einheitliches Vokabular und somit die Vergleichbarkeit von Produkten gegeben. Schutzprofile werden darüber hinaus mit einer Vertrauenswürdigkeitsstufe (*engl.*: Evaluation Assurance Level (EAL)) versehen, die eine Prüftiefe für das Produkt gegen das Schutzprofil vorgibt. So kann die Prüftiefe zwischen EAL 1 (funktionell getestet) und EAL 7 (formal verifizierter Entwurf und getestet) liegen.

IT-Grundschutz Der IT-Grundschutz wurde vom Bundesamt für Sicherheit in der Informationstechnik (BSI) entwickelt. Ziel des IT Grundschutzes ist es eine einfache Methode bereit zu stellen, die Sicherheitsmaßnahmen mit einer an die Bedürfnisse angepassten Sicherheitsstufe für Standard-Schutzanforderungen zum Schutz von Geschäftsprozessen, Anwendungen und IT-Systemen liefert. Dazu werden organisatorische, personelle, infrastrukturelle und technische Maßnahmen vorgeschlagen. Der IT-Grundschutz besteht auf folgenden Katalogen: Bausteinen, Gefährdungen und Schutzmaßnahmen. Der Bausteine-Katalog ist in fünf Stufen unterteilt, nach denen sich Verantwortlichkeiten trennen lassen. Diese Stufen sind: Übergreifende Aspekte, Infrastruktur, IT-Systeme, Netze, Anwendungen. Für jeden konkreten Baustein (z.B. B 5.21 Webanwendungen) werden konkrete Gefährdungen aus dem Schichtenmodell des Gefährdungskatalogs identifiziert, sowie entsprechende Schutzmaßnahmen aus dem Schichtenmodell des Schutzmaßnahmenkatalogs genannt, die einen Grundschutz liefern.

Schnittstellen zwischen KORA, CC und IT-Grundschutz Simić-Draws et al. [SDNK+13] identifizieren Schwachstellen der jeweiligen Methodiken und Standards. So bietet KORA zwar eine rechtlich fundierte Grundlage zur Ableitung technischer Anforderungen, jedoch genügt diese Ableitung nicht zur vollständigen Evaluation von IT-Systemen. Während die Common Criteria ein etablierter IT Evaluationsstandard ist, werden rechtliche Anforderungen nicht betrachtet und eine rechtskonforme Evaluation von Produkten kann nicht ohne weiteres stattfinden. Schließlich bietet der IT-Grundschutz eine umfassende Sicherheitssicht, während auch der IT-Grundschutz einerseits rechtliche Aspekte nicht betrachtet, andererseits eine Produktevaluation nicht auf dem Detailgrad der Common Criteria stattfindet. Auf Grundlage dieser Erkenntnis erarbeiten Simić-Draws et al. eine Methodik, die die Schwachstellen der einzelnen Methodiken und Standards durch eine geeignete Komposition eliminiert. Diese Komposition ist in Abbildung 1 dargestellt.

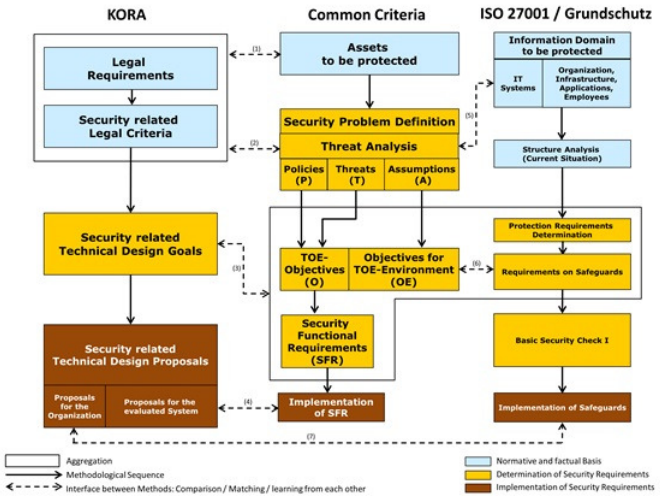


Abbildung 1: Schnittstellen zwischen KORA, CC und IT-Grundschutz [SDNK+13].

4 Operationalisierung der Methode

Bei dem Versuch, die Methode anzuwenden, hat sich gezeigt, dass weitere Konkretisierungen notwendig sind, um diese in Bezug auf die Entwicklung von Schutzprofilen operationalisieren zu können. Die im Rahmen dieser Arbeit vorgeschlagenen Konkretisierungen sind in Abbildung 2 fett umrandet.

Im Vordergrund stehen dabei zunächst die zu schützenden Werte, die als Basis für alle weiteren Schritte bei der Entwicklung eines Schutzprofils dienen. Wir schlagen vor, die zu schützenden Werte wie folgt aus den mittels KORA ermittelten rechtlichen Kriterien abzuleiten: In einem ersten Schritt werden die Kriterien semantisch analysiert und in Teilkriterien unterteilt. In einem zweiten Schritt sollen in einem interdisziplinären Diskurs zwischen juristischen und technischen Experten die sicherheitsrelevanten Teilkriterien identifiziert werden. Im nächsten Schritt können dann auf Basis der sicherheitsrelevanten Teilkriterien die zu schützenden Werte identifiziert werden.

Die Grundlage für die Beschreibung der Einsatzumgebung (Bedrohungen, Annahmen und Sicherheitspolitiken) bildet klassich eine Bedrohungsanalyse. Um diese auf die zuvor identfizierten zu schützenden Werte anwenden zu können, wird folgender Prozess vorgeschlagen: Basierend auf den kryptographischen Schutzzielen Authentizität, Integrität und Vertraulichkeit soll im interdiziplinären Dialog geklärt werden, welches die relevanten Schutzziele für den jeweiligen zu schützenden Wert sind. Die Bedrohungsanalyse kann dann anhand von Bedrohungsbäumen durchgeführt werden.

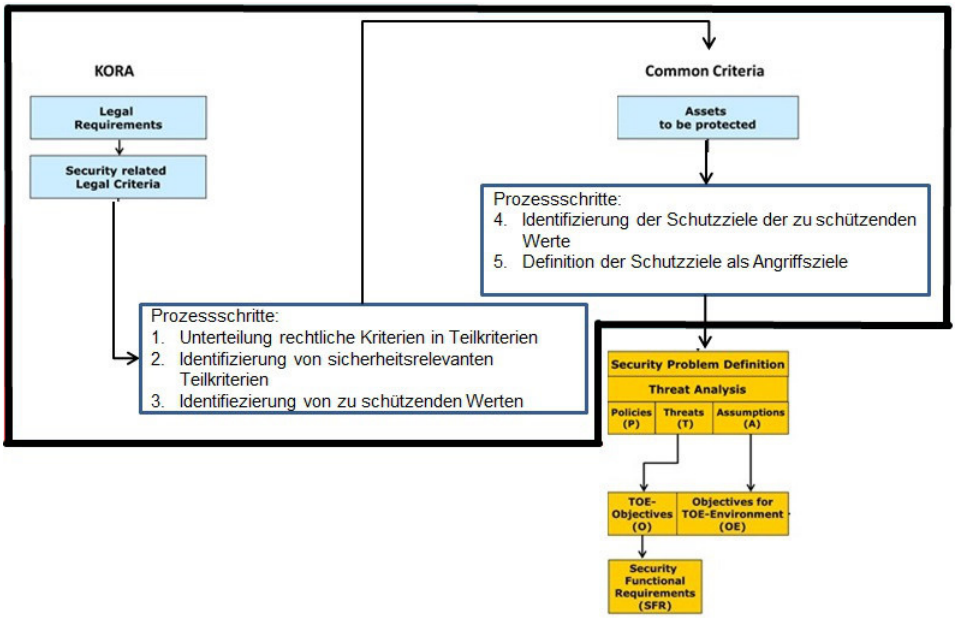


Abbildung 2: Fokus und Verfeinerung der Evaluationsmethodik [SDNK+13].

5 Entwurf des Schutzprofils

Aufgrund der Tatsache, dass ein vollständiges Schutzprofil den Rahmen dieser Arbeit sprengen würde, beschränken wir uns an dieser Stelle auf folgende Bestandteile des Schutzprofils: 1) Beschreibung des Evaluierungsgegenstands 2) Liste der zu schützenden Werte (an ausgewählten Beispielen), 3) Bedrohungsgraph für die zu schützenden Werte als Basis für die Definition der Einsatzumgebung.

5.1 Der Evaluierungsgegenstand

Der betrachtete Evaluierungsgegenstand (EVG) ist ein elektronisches Wahlsystem, das die Kandidatenauswahl an einem elektronischen Wahlgerät sowie die Auszählung an einem anderen elektronischen Wahlgerät vorsieht. Der EVG besteht aus einem hybriden Wahlgerät. Ähnlich zur papier-basierten Wahl identifiziert sich der Wähler zunächst gegenüber dem Wahlvorstand. Danach betritt er die Wahlkabine und benutzt das elektronische Wahlgerät um durch das Auswählen der Kandidaten seinen Stimmzettel zu erstellen. Sobald der Wähler seine Auswahl bestätigt wird der Stimmzettel ausgedruckt. Der ausgedruckte Stimmzettel enthält einen menschlich- und einen maschinen-lesbaren Teil. Beide Teile ent-

halten dieselbe Information, nämlich die Auswahl des Wählers.[1] Der menschlich-lesbare Teil dient dem Wähler zur Überprüfung, dass der gedruckte Stimmzettel der am Wahlgerät abgegebenen Stimme entspricht. Nach der Überprüfung durch den Wähler faltet dieser den Stimmzettel, verlässt die Wahlkabine und wirft den Stimmzettel in die Wahlurne. Der maschinen-lesbare Teil des Stimmzettels ermöglicht ein automatisiertes Auszählen der gedruckten Stimmzettel. Es gibt bereits einige Wahlsysteme (Wahlgeräte), welche diesem Evaluierungsgegenstand entsprechen, zum Beispiel [BNFL+12], [VBD11], [Veg12], and [Vot11].

Die Komponenten des Evaluierungsgegenstand sind in Abbildung 3 dargestellt. Der Evaluierungsgegenstand beinhaltet die Hardware und Software dieser Komponenten.

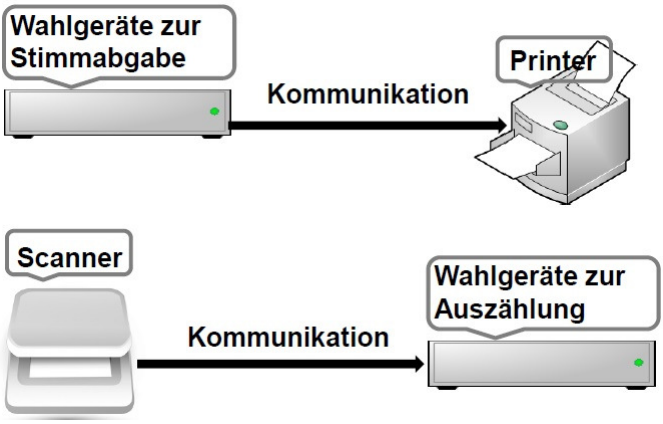


Abbildung 3: Komponentente des Evaluierungsgegenstandes.

5.2 Identifizierung der zu schützenden Werte

In diesem Abschnitt wenden wir die in dieser Arbeit eingeführte Verfeinerung der Evaluationsmethodik von Simić-Draws et al. [SDNK+13] an und führen die Identifizierung von zu schützenden Werten beispielhaft durch. Als Beispiel verwenden wir das rechtliche Kriterium der Manipulationssicherheit. Dieses Kriterium steht indirekt im Bezug zur Verifizierbarkeit, weil es voraussetzt, dass Manipulationen nicht unentdeckt bleiben können, d.h. es fordert einen Mechanismus zur Verifikation.

Aus dem rechtlichen Kriterium der Manipulationssicherheit[2], werden folgende sicherheitsrelevante Teilkriterien abgeleitet:

1. Eine Manipulation von amtlich zugelassenen und genehmigten Geräten darf nicht

[1] Im Falle einer Abweichung zwischen dem menschlich- und dem maschinen-lesbaren Teil, zählt ausschließlich der menschlich-lesbare Teil des Stimmzettels.

[2] Die vollständige Definition dieses rechtliche Kriterium befindet sich in [Hen14].

unentdeckt bleiben.

2. Eine nachträgliche Änderung der Software bzw. Hardware des EVGs muss unmöglich sein oder darf zumindest nicht unentdeckt bleiben.

5.3 Identifizierung von Bedrohungen

Die beiden zuvor identifizierten sicherheitsrelevanten Teilkriterien sind über den zu schützenden Wert *Geräte* verbunden. Das einzige relevante Schutzziel dieses zu schützenden Wertes ist die Integrität und das daraus entstehende Angriffsziel ist *Verletzung der Integrität der Geräte*. Der Bedrohungsbaum zu diesem Angriffsziel ist in Abbildung 4 dargestellt. Einige der möglichen Bedrohungen, welche in diesem Bedrohungsbaum identifiziert wurden, wurden aus vorherigen Arbeiten entnommen, siehe Prosser et al. [PKKU04] and Krimmer and Volkamer [KV08].

Aus den Blättern dieses Bedrohungsbaumes können die entsprechenden *Bedrohungen* und *Annahmen* identifiziert werden. In Anlehnung an Buchmann et al. [BNV14] kann nur eine Kombination an zulässigen Bedrohungen und Annahmen in einem Schutzprofil abgebildet werden, nicht aber alle laut Gestaltungsspielraum möglichen. Welche Kombination gewählt wird, sollte im interdisziplinären Dialog entschieden werden.

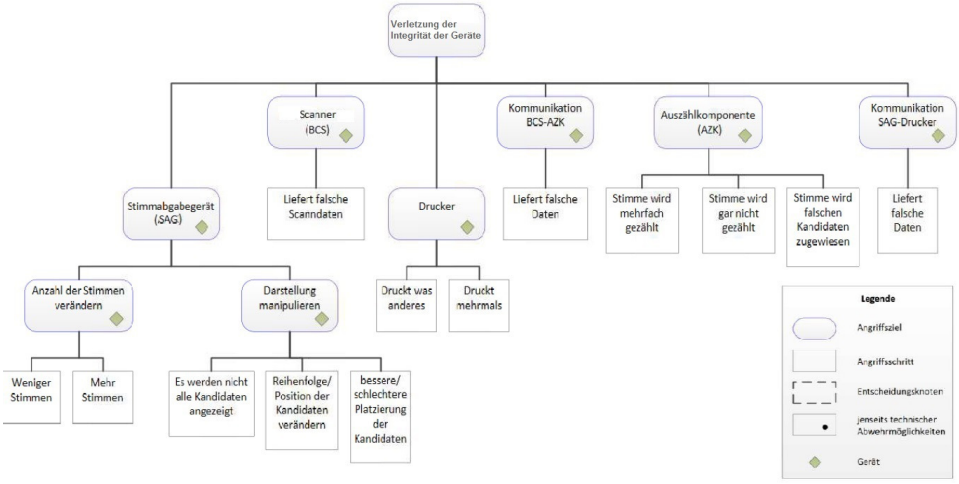


Abbildung 4: Bedrohungsbaum “Manipulierte Geräte”.

6 Zusammenfassung

Mit dem stetigen Fortschreiten der Entwicklung, dringen Technologien in immer neue Bereiche unseres Lebens vor. War die Durchführung von Wahlen über Jahrhunderte ein relativ simpler Prozess, so eröffnen elektronische Wahlsysteme neue Möglichkeiten zur Schaffung und Festigung von Demokratien. Obwohl elektronische Wahlen in Deutschland bereits praktisch durchgeführt wurden, hat das Bundesverfassungsgericht mit dem Urteil vom 3. März 2009 die Durchführung bisheriger elektronischer Wahlen für verfassungswidrig erklärt. Grund dieses Urteils war die Tatsache, dass die eingesetzten Wahlgeräte nicht dem Prinzip der *Öffentlichkeit der Wahl* entsprachen. Die Durchführung rechtskonformer elektronischer Wahlen ist eine Herausforderung von besonderem öffentlichen Interesse und bedarf interdisziplinärer Forschungsarbeit. Ziel dieser Arbeit war es einen ersten Schritt in Richtung Entwurf eines Schutzprofils zur Evaluierung von elektronischen Wahlgeräten mit Paper Audit Trails auf Basis rechtlicher Vorgaben zu entwickeln. Dazu wurde in dieser Arbeit eine der Schnittstellen zwischen KORA und dem CC-Standard der Evaluationsmethodik nach Simić-Draws et al. [SDNK+13] bezüglich einer systematischen Methodik zur Identifizierung von zu schützenden Werten verfeinert.

Danksagung

Wir danken den anonymen Gutachtern mit deren Hilfe die Qualität dieser Arbeit deutlich verbessert werden konnte. Diese Arbeit wurde im Rahmen des durch die DFG geförderten Projektes 'VerkonWa' - Verfassungskonforme Umsetzung von elektronischen Wahlen sowie durch das vom Center for Advanced Security Research Darmstadt (CASED) geförderte Projekt ComVote entwickelt.

Literatur

[AMBS07] J. Arzt-Mergemeier, W. Beiss und T. Steffens. The Digital Voting Pen at the Hamburg Elections 2008: Electronic Voting Closest to Conventional Voting. In A. Alkassar und M. Volkamer, Hrsg., *E-Voting and Identity*, Jgg. 4896 of *Lecture Notes in Computer Science*, Seiten 88–98. Springer Berlin Heidelberg, 2007.

[BNFL+12] J. Ben-Nun, N. Fahri, M. Llewellyn, B. Riva, A. Rosen, A. Ta-Shma und D. Wikström. A New Implementation of a Dual (Paper and Cryptographic) Voting System. In *Electronic Voting*, Jgg. 205 of *LNI*, Seiten 315–329, 2012.

[BNV14] J. Buchmann, S. Neumann und M. Volkamer. Tauglichkeit von Common Criteria-Schutzprofilen für Internetwahlen in Deutschland. *Datenschutz und Datensicherheit - DuD*, 38(2):98–102, 2014.

[Bun07] Bundesamt für Sicherheit in der Informationstechnik. Schutzprofil Digitales Wahlstift-System (Version 1.0.1), März 2007.

[Bun08a] Bundesamt für Sicherheit in der Informationstechnik. Basissatz von Sicherheitsanforderungen an Online-Wahlprodukte (Version 1.0), April 2008. https://www.commoncriteriaportal.org/files/ppfiles/pp0037b.pdf.

[Bun08b] Bundesamt für Sicherheit in der Informationstechnik. IT-Grundschutz Methodology, BSI Standard 100-2 (Version 2.0), May 2008.

[Bun09] Bundesverfassungsgerichts. Entscheidungen des Bundesverfassungsgerichts. BVerfGE 123, 39 - 81, March 2009. http://www.bverfg.de/entscheidungen/cs20090303$_2bvc$000307.$html$.

[Hen14] M. Henning. Transparenz und Sicherheit in der Demokratie - Verfassungsverträglichkeit elektronischer Wahlgeräte. Diss., Universität Kassel, 2014.

[HVB12] M. Henning, M. Volkamer und J. Budurushi. Elektronische Kandidatenauswahl und automatisierte Stimmermittlung am Beispiel hessischer Kommunalwahlen. *Die Öffentliche Verwaltung (DÖV)*, (20), October 2012.

[Int09] International Organization for Standardization. ISO/IEC 15408: Common Criteria for Information Technology Security Evaluation, Part 1–3 (Version 3.1, Revision 3), July 2009.

[KKY12] G. R. Karokola, S. Kowalski und L. Yngström. Secure e-Government services: Protection Profile for Electronic Voting - A Case of Tanzania. In *IST-Africa 2012 Conference Proceedings*, 2012.

[KV08] R. Krimmer und M. Volkamer. Observing Threats to Voter's Anonymity: Election Observatio of Electronic Voting. In S. J. Krishna und N. K. Agarwal, Hrsg., *E-Voting - Perspectives and Experiences*, The Icfai University Press, 2008.

[LLWK10] K. Lee, Y. Lee, D. Won und S. Kim. Protection Profile for Secure E-Voting Systems. In J. Kwak, R. Deng, Y. Won und G. Wang, Hrsg., *Information Security, Practice and Experience*, Jgg. 6047 of *Lecture Notes in Computer Science*, Seiten 386–397. Springer Berlin Heidelberg, 2010.

[NKH+13] Stephan Neumann, Anna Kahlert, Maria Henning, Philipp Richter, Hugo Jonker und Melanie Volkamer. Modeling the German Legal Latitude Principles. In *Electronic Participation*, Seiten 49–56. Springer, 2013.

[PKKU04] A. Prosser, R. Kofler, R. Krimmer und M. K. Unger. Security assets in e-voting. In *The International Workshop on Electronic Voting in Europe*, 2004.

[Roß08] A. Roßnagel. Rechtswissenschaftliche Gestaltung von Informationstechnik. In H. Kortzfleisch und O. Bohl, Hrsg., *Wissen, Vernetzung, Virtualisierung*, Seiten 381–390. Köln, Germany: Josef Eul Verlag, 2008.

[SDNK+13] D. Simić-Draws, S. Neumann, A. Kahlert, P. Richter, R. Grimm, M. Volkamer und A. Roßnagel. Holistic and Law compatible IT Security Evaluation: Integration of Common Criteria, ISO 27001/IT-Grundschutz and KORA. *International Journal of Information Security and Privacy (IJISP)*, 7(3):16–35, 2013.

[Sec06] Secretariat general de la defense nationale. Protection Profile: Machine a voter, PP-CIVIS (Version 1.0), Juni 2006. https://www.commoncriteriaportal.org/files/ppfiles/pp0037b.pdf.

[VBD11] M. Volkamer, J. Budurushi und D. Demirel. Vote casting device with VV-SV-PAT for elections with complicated ballot papers. In *International Workshop on Requirements Engineering for Electronic Voting Systems (REVOTE'11)*, Seiten 1–8. IEEE, August 2011.

[Veg12] C. Vegas. The New Belgian E-voting System. In *Electronic Voting*, Jgg. 205 of *LNI*, Seiten 199–211, 2012.

[Vot11] 2011. Vot.ar, http://www.vot-ar.com.ar/en/system-votation/.

Das Zentrale Wählerregister – Ein skalierbares Instrument zur Bürgerbeteiligung mit 1:1-Verifikation

Mag. Robert Stein / Mag. Gregor Wenda, MBA
Abteilung für Wahlangelegenheiten
Bundesministerium für Inneres, Republik Österreich
Herrengasse 7, 1014 Wien, Österreich
robert.stein@bmi.gv.at
gregor.wenda@bmi.gv.at

Abstract: Im Rahmen einer geplanten Reform der „Direkten Demokratie" wurde 2012 im österreichischen Parlament von Abgeordneten der Regierungsparteien ein Initiativantrag eingebracht, der unter anderem die Schaffung eines neuen Zentralen Wählerregisters vorsah. Ein solches Register könnte neben einer erstmaligen Vernetzung der Wählerdaten in Österreich auch eine Erneuerung des direktdemokratischen Instruments der „Volksbegehren" nach sich ziehen, denn der Gesetzesentwurf würde die Möglichkeit einer Online-Teilnahme vorsehen. Die seit 2013 laufende neue Legislaturperiode wird zeigen, ob die Umsetzung eines Zentralen Wählerregisters weiterhin realistisch ist. Im Fall seiner Einrichtung könnte sogar die Basis für eine allfällige zukünftige Einführung der elektronischen Stimmabgabe gelegt werden.

1 Vorgeschichte

Anders als in anderen europäischen Staaten[1] gibt es in Österreich derzeit keine zentrale Evidenz der Wahlberechtigten. Bis ins Jahr 2003 wurde das bestehende System von dezentral bei den Gemeinden geführten Karteien oder Datenbanken mit regelmäßig durchzuführenden Überprüfungen, so genannten Zuzugsmitteilungen, und einem Reklamationsverfahren[2] aufrechterhalten. Eine Person wurde eingetragen, wenn sie sich in einer Gemeinde angemeldet hatte; aufgrund der „Zuzugsmitteilung" wurde sie von der „alten" Gemeinde gelöscht. Zweifelsfälle in jeglicher Richtung wurden aufgrund von Reklamationen bereinigt. Das System war keineswegs „wasserdicht", vor allem ab dem Zeitpunkt, zu dem Auslandsösterreicherinnen und Auslandsösterreicher in die Wählerevidenz eines „Anknüpfungspunkts" (zumeist dem Ort des letzten

[1] So besteht etwa in Albanien, Armenien, Finnland oder Litauen eine entsprechende zentrale Erfassung (vgl. die Zusammenfassung der OSZE/ODIHR-Veranstaltung „Seminar on Synergies between Population and Voter Registration: Implications for the Accuracy of Voter Lists" am 31. August und 1. September 2011 in Warschau). Allgemein zur Errichtung Zentraler Wählerregister in Europa s. auch das Handbuch „Electoral Law" des Europarates (2008) 131 f.

[2] Dies bedeutet, dass Wahlberechtigte bei den in den Gemeinden geführten Listen hinein und hinaus „reklamiert" werden konnten – durch jedermann in jegliche Richtung. Dieses „Reklamationsrecht" besteht dem Grunde nach bis heute. Vgl. § 4ff des Wählerevidenzgesetzes 1973.

Hauptwohnsitzes im Inland) einzutragen waren.[3] Die Datenqualität muss rückblickend als schlecht bezeichnet werden.

Im Jahr 2003 gab es insofern eine Veränderung, als eine DV-technische Bindung der örtlichen Wählerevidenzen an das Zentrale Melderegister[4] implementiert wurde. Seit diesem Zeitpunkt kann davon ausgegangen werden, dass die örtlichen Wählerevidenzen eine 1:1-Abbildung der Datensätze jener Personen darstellen, die in der jeweiligen Gemeinde ihren Hauptwohnsitz haben. Auslandsösterreicherinnen und Auslandsösterreicher werden – von dieser Akzessorietät unabhängig – weiterhin aufgrund von Anträgen „manuell" eingetragen bzw. auch aufgrund von Fristabläufen (Zehn-Jahres-Frist) „manuell" gelöscht. Die korrekte Erfassung von Auslandsösterreicherinnen und Auslandsösterreichern musste und muss aber weiterhin als Schwachstelle dieses Systems angesehen werden. Als weitere Schwachstelle ist im Jahr 2011 die Erfassung von Häftlingen hinzugekommen[5], die seitdem in Österreich in der Regel auch bei mehrjährigen Haftstrafen nicht vom Wahlrecht ausgeschlossen sind, allerdings nicht in der Gemeinde des Sitzes der Justizanstalt, sondern der letzten Gemeinde des Hauptwohnsitz in Österreich zu erfassen sind.[6]

2 Volksbegehren

Unter Zuhilfenahme der örtlichen Wählerevidenzen wird in Österreich seit Jahrzehnten ein populäres Instrument der direkten Demokratie, das „Volksbegehren"[7], administriert. Seit 1963 sind 37 solche Begehren abgehalten worden.[8] Bei einem Volksbegehren handelt es sich um eine kraft Gesetzes zweiphasig implementierte Initiative, die bei Erfüllung bestimmter Voraussetzungen zur obligaten Behandlung eines Ansinnens im österreichischen Parlament führen kann. Neben den inhaltlichen Voraussetzungen („durch Bundesgesetz zu regelnde Angelegenheit") muss in der ersten Phase vor der Einbringung eines Einleitungsantrages eine Mindestzahl an Unterstützungserklärungen (8.401) gesammelt werden. Proponentinnen und Proponenten eines Volksbegehrens haben hierzu das „angefangene" Jahr und das darauffolgende Kalenderjahr Zeit. Unterstützungserklärungen bedürfen jedenfalls der Beglaubigung der Unterschrift, sei es, indem die Unterschrift vor dem Organwalter oder der Organwalterin einer Gemeinde geleistet wurde, sei es dass sie bei einem Gericht oder vor einem Notar getätigt wurde. Dazu kommt die Bestätigung durch die Gemeinde, dass die jeweilige unterstützungswillige Person in der Wählerevidenz dieser Gemeinde eingetragen ist. Ein

[3] Aufgrund eines Erkenntnisses des Verfassungsgerichtshofes vom 16. März 1989 (VfSlg 12.023/1989) fand mit dem Wahlrechtsänderungsgesetz 1990 (BGBl 1990/148) erstmals eine Norm Eingang in die Rechtsordnung, die eine Stimmabgabe aus dem Ausland ermöglichte (vgl. Stein/Vogl/Wenda NRWO4 [2013] Anm 1 zu § 60).
[4] Novelle zum Meldegesetz, BGBl I Nr. 28/2001.
[5] Der Novellierung der Wahlausschließungsgründe im Jahr 2011 (BGBl I 2011/43) ging das von 2004 bis 2010 beim EGMR laufende Verfahren Frodl gegen Österreich (ÖJZ 2010, 734) voraus.
[6] Vgl. Art. 6 Abs. 4 B-VG (diese verfassungsrechtliche Grundlage wurde zuletzt mit BGBl I 2013/115 präzisiert).
[7] Vgl. Art 41 B-VG.
[8] Für eine komplette Liste vgl.
http://www.bmi.gv.at/cms/BMI_wahlen/volksbegehren/vb_xx_periode/start.aspx

anlässlich der Erteilung der Bestätigung in der örtlichen Wählerevidenz angebrachter Sperrvermerk verhindert, dass unterstützungswillige Personen ein Volksbegehren mehrfach unterstützen. Kraft Bundesverfassung[9] ist die Abgabe einer Unterstützungserklärung – wie auch die Abgabe einer Unterschrift im Eintragungsverfahren – im Moment auf Österreicherinnen und Österreicher mit Hauptwohnsitz im Bundesgebiet beschränkt.

Haben Proponentinnen und Proponenten eines Volksbegehrens eine ausreichende Anzahl an Unterstützungserklärungen gesammelt, so können sie der Bundesministerin für Inneres einen Einleitungsantrag vorlegen, der – nach Überprüfung der Unterstützungserklärungen durch das Bundesministerium für Inneres – zur Festlegung eines Eintragungszeitraumes von acht Tagen führt. Auch das Eintragungsverfahren erfolgt unter Zugrundelegung der örtlichen Wählerevidenzen, aus denen für den Eintragungszeitraum Stimmlisten 1:1 gewonnen werden. Personen, die ein Volksbegehren bereits vor dem Einleitungsverfahren mittels Unterstützungserklärung unterstützt haben, sind von einer Unterschrift im Eintragungsverfahren ausgeschlossen. Die Zahl der Unterstützungserklärungen wird jedoch auf die erzielten Unterschriften angerechnet. Der zu überschreitende Schwellwert für die Behandlung eines Volksbegehrens im Parlament liegt bei 100.000 Unterschriften (inklusive den bereits vorgelegten Unterstützungserklärungen).[10]

Beim Volksbegehren nach geltender Rechtslage sollte zwar zu 100 Prozent sichergestellt sein, dass Personen ein solches nur einmal unterstützen können, das Prozedere muss aber als aufwändig und im EDV-Zeitalter als unzeitgemäß angesehen werden. Hinzu kommt ein weiterer „bremsender" Faktor: Grundsätzlich können unterstützungswillige Personen ein Volksbegehren nur durch Aufsuchen des Gemeindeamts oder einer dafür vorgesehenen Stelle an ihrem Hauptwohnsitz unterstützen bzw. dafür unterschreiben. Da es sich bei der Unterstützung eines Volksbegehrens nicht um eine geheime Stimmabgabe im Rahmen einer Wahl handelt, wird bei einer Unterschriftsleistung in der „Heimatgemeinde" unvermeidlich eine bestimmte politische Präferenz kundgetan. Dieser Umstand hat in der Vergangenheit dem Anschein nach immer wieder Bürgerinnen und Bürger davon abgehalten, an einem Volksbegehren teilzunehmen. Die Angst, Nachteile durch einen Bürgermeister oder eine Bürgermeisterin zu erleiden, der die Teilnahme an einem Volksbegehren entgegen der eigenen politischen Ausrichtung nicht gutheißen würde, scheint ein – subjektiver – „Hemmschuh" zu sein. Auch von der Möglichkeit der Beantragung einer Stimmkarte, mit der für ein Volksbegehren anderswo im Inland eine Eintragung getätigt werden kann, wird in der Regel nur selten Gebrauch gemacht.[11] Um zumindest den Kreis der Teilnehmenden weiter zu erhöhen, besteht seit kurzer Zeit die Möglichkeit, dass bettlägerige Personen von einem Organwalter oder einer Organwalterin der Gemeinde am Krankenbett aufgesucht werden.

[9] Vgl. Art. 41 Abs. 2 B-VG.
[10] Vgl. Anm 7.
[11] Auch in diesem Fall müsste in der Gemeinde des Hauptwohnsitzes die Ausstellung einer Stimmkarte beantragt werden, woraus sich für Bedienstete der Gemeinde der Wille zur Unterstützung eines konkreten Volksbegehrens schließen ließe.

3 Initiative der österreichischen Bundesregierung

Im September 2012 nahm die österreichische Bundesregierung in Aussicht, für die Schaffung eines „Zentralen Wählerregisters" („ZeWaeR") einzutreten.[12] Folgt man der Begründung zum Initiativantrag 2177/A vom 30. Jänner 2013[13], so sind dort eine Reihe von Vorteilen in Aussicht genommen, die die Schaffung eines ZeWaeR mit sich bringen würde. Losgelöst von einem „Volksbegehren neu" (Näheres siehe unten) sind vorab zu nennen:[14]

- Möglichkeit des „Clearings" aller Wahlberechtigten, insbesondere in Hinblick auf eine allfällige Doppelregistrierung von vermeintlichen Auslandsösterreicherinnen und Auslandsösterreichern;
- Wegfall des die Gemeinden treffenden Erfordernisses der Datenübermittlung an das Bundesministerium für Inneres für die bestehende Zentrale Wählerevidenz (diese dient ausschließlich zur Weitergabe an die im Nationalrat vertretenen Parteien) und für die Zentrale Europa-Wählerevidenz (diese dient zur Beschickung der anderen Mitgliedstaaten im Zusammenhang mit dem Datenaustausch nach der Richtlinie 93/109/EG);
- verbesserte Datenqualität bei der Weitergabe der Wählerevidenzdaten an die im Nationalrat vertretenen Parteien;
- zielsichere Zuordnung von Häftlingen zur Wählerevidenz der für sie zuständigen Gemeinde während der Haft;
- stark vereinfachte Beauskunftung anderer EU-Mitgliedstaaten im Sinne der oben angeführten Richtlinie bezüglich wahlberechtigter Unionsbürgerinnen und Unionsbürger bei Europawahlen;
- wesentliche Vereinfachung für Gemeinden bei Amtshandlungen im Zusammenhang mit der Wählerevidenz und der Europa-Wählerevidenz.

Besondere Vorteile eines neuen ZeWaeR werden daneben bei einer völlig neugestalteten Administration des direkt-demokratischen Instruments der „Volksbegehren" (auch als „Volksbegehren neu" bezeichnet) gesehen. Nach dem genannten Gesetzesentwurf könnten Volksbegehren von den Bürgerinnen und Bürgern durch Aufsuchen jedes beliebigen österreichischen Gemeindeamtes unterstützt oder im Eintragungsverfahren unterschrieben werden. Die Unterstützung eines Volksbegehrens im Weg des Internet soll ebenfalls möglich sein, allerdings nur für Personen, die über die Möglichkeit einer qualifizierten digitalen Signatur verfügen.[15] Mittels qualifizierter digitaler Signatur

[12] Vgl. OTS von BM Spindelegger vom 11. September 2012: „Mehr Bürgernähe durch zentrales Wählerregister".

[13] Antrag gemäß § 26 GOG-NR der Abgeordneten Dr. Josef Cap, Karlheinz Kopf, Kolleginnen und Kollegen betreffend ein Bundesgesetz, mit dem das Bundes-Verfassungsgesetz, das Bundesgesetz über die Geschäftsordnung des Nationalrates, die Nationalrats-Wahlordnung 1992, das Bundespräsidentenwahlgesetz 1971, die Europawahlordnung, das Europa-Wählerevidenzgesetz, das Volksabstimmungsgesetz 1972 und das Volksbefragungsgesetz 1989 geändert, das Volksbegehrengesetz 2013 und das Wählerevidenzgesetz 2013 erlassen sowie das Volksbegehrengesetz 1973 und das Wählerevidenzgesetz 1973 aufgehoben werden.

[14] Vgl. IA 2177/A 24. GP, Begründung zu Art. 4.

[15] Funktion der „Bürgerkarte", entweder physisch auf einer Karte oder mittels „Handy-Signatur" (vgl. www.buergerkarte.at).

könnten auch Auslandösterreicherinnen und Auslandsösterreicher erstmals die Möglichkeit bekommen, österreichische Volksbegehren zu unterstützen. Ihnen ist, wie oben erwähnt, die Teilnahme an Volksbegehren in Ermangelung eines Hauptwohnsitzes in Österreich bislang – sogar durch die österreichische Bundesverfassung „abgesichert“[16] – verschlossen. Das aufwendige Prozedere für die Ermittlung des Ergebnisses von Volksbegehren, das nach derzeitiger Rechtslage die Gemeinden sowie die Bezirkswahlbehörden betrifft, könnte ebenfalls auf ein Minimum reduziert werden. Alles würde in kürzester Zeit und praktisch vollautomatisch von Statten gehen.

Schon beim ersten Gesetzesentwurf aus dem Jänner 2013 waren weitere Elemente der Bürgerbeteiligung vorgesehen, die sich des ZeWaeR bedienen sollten.[17] Dazu zählten die „Parlamentarische Bürgeranfrage“ sowie die „Parlamentarische Bürgerinitiative“. Bei der Weiterentwicklung des Entwurfs wurde das erstgenannte Element wieder fallen gelassen.[18] Gleichzeitig wurde in den Entwurf eine Aufwertung des Volksbegehrens in Richtung einer obligaten anschließenden Abhaltung einer Volksbefragung[19] aufgenommen. Als Bedingung hierfür sollten Volksbegehren gewisse Voraussetzungen (Form der Einbringung, Erfüllung einer Mindestquote, Beschränkung auf bestimmte Anliegen) erfüllen. Im Gegensatz zur Durchführung von Volksbegehren sollte sich bei der Durchführung von Volksbefragungen – folgt man dem Entwurf in seiner letzten Fassung – nichts ändern, da ein solcher Schritt einen Übergang zu einem E-Voting-System darstellen würde[20] und erstmals eine „elektronische Stimmabgabe“ bei Wahlen[21] ermöglichen könnte.[22] Derzeit ist E-Voting in Österreich allerdings kein politisches Ziel.[23]

4 „Volksbegehren neu“ – ein durchdachtes System

Schon bei Konzipierung des Entwurfes wurde auf technische Details geachtet, die im Fall einer Umsetzung von „Volksbegehren neu“ zu beachten wären. Im Zuge der parlamentarischen Behandlung des Entwurfs in der vergangenen (24.) Gesetzgebungsperiode wurden die Details in Absprache mit Expertinnen und Experten der Klubs der Parlamentsfraktionen präzisiert. Die Möglichkeit der Online-

[16] Vgl. Anm. 7.
[17] Vgl. IA 2177/A 24. GP.
[18] Vgl. Stein, Kaum mehr überschaubar – Änderungen im Wahlrecht, RFG 3/2013, 114.
[19] Vgl. Art. 49b B-VG.
[20] Eine Volksbefragung ist in ihrem Ablauf einer Wahl vergleichbar. Unter anderem gilt das Prinzip der geheimen Stimmabgabe.
[21] Vgl. „Der Standard“ vom 21. Juli 2012: Ostermayer zu E-Voting: „Warum nicht?“ http://derstandard.at/1342139760040/Ostermayer-zu-E-Voting-Warum-nicht
[22] Schon die Arbeitsgruppe zu E-Voting im BM.I empfahl 2004 die Einrichtung einer neuen „Zentralen Wählerevidenz“, vgl. Bericht der Unterarbeitsgruppe 2 „Technik“ vom 3.11.2004, 23: „Es bietet sich an, auf Basis der in Österreich elektronisch geführten Register, wie zentrales Melderegister (ZMR) oder Ergänzungsregister, eine Zentrale Wählerevidenz (ZWE) aufzubauen.“ (http://www.a-sit.at/pdfs/Endbericht_E-Voting_UAG2-Technik.pdf).
[23] Im aktuellen Regierungsprogramm wurde die Materie der elektronischen Stimmabgabe, im Gegensatz zu früheren Regierungsabkommen, in keiner Weise aufgenommen. Eine besonders explizite Formulierung enthielt nur das Regierungsprogramm für die 23. GP mit dem Auftrag einer „Prüfung der elektronischen Stimmabgabe“ ab Jänner 2007.

Unterstützung würde eine Registrierung des Volksbegehrens im Vorhinein unerlässlich machen[24], zumal der Bund eine Infrastruktur für die Unterstützung des Volksbegehrens zur Verfügung zu stellen hätte.[25] Nach erfolgter Registrierung könnten Unterstützungserklärungen für ein Volksbegehren gesammelt werden. Diese Sammlung könnte – als Novum – erstmals online erfolgen, oder weiterhin vor einem Organwalter (einer Organwalterin). Dabei könnte, losgelöst vom Hauptwohnsitz, eine beliebige Gemeinde aufgesucht werden. In beiden Fällen würde die Tätigung einer Unterstützungserklärung zu einer – für ZeWaeR-Benutzungsberechtigte in anderen Gemeinden grundsätzlich nicht sichtbaren – Vormerkung (eines so genannten „Flag“ in der Datenbank) führen. Für die Tätigung einer Online-Unterstützungserklärung wäre, wie erwähnt, die Verwendung einer qualifizierten digitalen Signatur zwingend erforderlich. Im Fall der Unterstützung des Volksbegehrens in einer Gemeinde wären – nachdem der Organwalter (die Organwalterin) die Identität der unterstützungswilligen Person festgestellt hat und die Applikation keine Ablehnung wegen einer bereits vorhandenen Vormerkung angezeigt hat – durch die ZeWaeR-Applikation zwei Ausdrucke zu generieren, nämlich die Unterstützungserklärung und eine Bestätigung hierüber. Der (die) Unterstützungswillige hätte eine Unterschrift zu leisten (die Bestätigung dient ledig lediglich zu Beweiszwecken und verbleibt bei der Gemeinde) und der (die) Unterstützungswillige würde eine vom Organwalter (von der Organwalterin) unterfertigte Bestätigung (ebenfalls lediglich zu Beweiszwecken) erhalten. Die Proponentinnen und Proponenten – wie auch das Bundesministerium für Inneres – könnten sich über die Zahl der bislang getätigten Unterstützungserklärungen jederzeit (online mit entsprechenden Zugangsberechtigungen) informieren.

Bei dem im Entwurfstext gewählten Begriff „Anwendung“[26] handelt es sich um einen technologieneutralen Begriff, mit dem die unterschiedlichsten technischen Umsetzungsformen (z.B. Formular, eigene Webseite, "App.", sonstige technische Umsetzung) abgedeckt werden sollten. Dass es sich um eine elektronische Form der Unterstützung handeln muss, ergibt sich daraus, dass die im Gesetzestext normierte Anwendung in der Lage sein muss, eine elektronische Signatur zu verarbeiten. Mit dem Wortlaut der Regelung laut Entwurf wäre ausgeschlossen, dass ein Volksbegehren via E-Mail oder im Weg einer nicht seitens der Behörde bereitgestellten Applikation unterstützt werden könnte.

Wäre eine ausreichende Zahl an Unterstützungserklärungen getätigt, so könnten die Proponentinnen und Proponenten eines Volksbegehrens jederzeit einen Einleitungsantrag einbringen. Danach wäre die Möglichkeit, weitere Unterstützungserklärungen zu tätigen, automatisch gesperrt. Wie nach geltendem Recht hätte die Bundesministerin für Inneres einen Eintragungszeitraum festzusetzen. Das bewährte Fristengefüge würde sich somit nicht ändern; mit Blick auf die Möglichkeit,

[24] Nach derzeit geltender Rechtslage sind Proponentinnen und Proponenten bei der Sammlung von Unterstützungserklärungen noch völlig auf sich allein gestellt; eine Registrierung vor dem Sammlungsstart ist nicht vorgesehen.

[25] Als „Übereilungsschutz“ ist im genannten Entwurf für die Registrierung eines Volksbegehrens eine Gebühr von € 500 vorgesehen. Um diesen Betrag reduziert sich der zu entrichtende Druckkostenbeitrag, der bei der Einbringung des Einleitungsantrags fällig wird. Die Einbringung eines Volksbegehrens würde sich gegenüber der geltenden Rechtslage also nicht verteuern.

[26] Vgl. § 5 des Entwurfes zu einem „Volksbegehrengesetz 2013 – VoBeG“ (Art. 3, 2177/A 24. GP).

für das Volksbegehren online eine Unterschrift zu leisten, wurde im Entwurf lediglich vorgeschlagen, auf das obligate Offenhalten des Eintragungslokales am Sonntag zu verzichten. Während des Eintragungszeitraums könnten Unterschriften in gleicher Weise geleistet werden, wie dies für Unterstützungserklärungen oben detailliert beschrieben wurde. Auch das Abfragen der Zahl der geleisteten Unterschriften wäre in gleicher Weise jederzeit möglich. Unmittelbar nach Ende des Eintragungszeitraums könnte die vorläufige Zahl der Eintragungen „auf Knopfdruck" durch das Bundesministerium für Inneres ermittelt werden. Den Bezirkswahlbehörden käme bei der Ergebnisfeststellung keine Rolle mehr zu. Die Feststellung des amtlichen Ergebnisses hätte wie bisher die Bundeswahlbehörde vorzunehmen, etwa drei Wochen nach Ende des Eintragungszeitraums.

Im Zug der parlamentarischen Behandlung der Entwürfe wurde bereits über viele Details zum Modell eines „Volksbegehren neu" nachgedacht.[27] So wäre z.B. vorgesehen, dass die Online-Unterstützung, wie in solchen Anwendungsfällen durchaus üblich, in zwei Phasen zu erfolgen hätte. Nachdem sich eine unterstützungswillige Person über den Text des Anliegens informiert und durch Mausklick ihre Absicht bekundet hat, das Volksbegehren unterstützen zu wollen oder eine Unterschrift im Eintragungsverfahren zu leisten, hätte sie sich – ähnlich wie beim Einloggen bei einem österreichischen Sozialversicherungsträger – zunächst zu identifizieren. Nach der Identifizierung würde sie in einem weiteren Schritt, gleichsam als Übereilungsschutz, gefragt, ob sie mit einer qualifizierten digitalen Signatur die entsprechende Willenserklärung tätigen möchte.

Im Falle einer Unterstützung vor einem Organwalter oder einer Organwalterin einer – beliebigen – Gemeinde wäre dieses Prozedere sehr ähnlich. Die unterstützungswillige Person wäre zunächst durch Abgleich der Daten in den Ausweispapieren mit jenen im ZeWaeR zu identifizieren. Dabei könnten die Gemeinde-Bediensteten noch nicht sehen, ob die unterstützungswillige Person allenfalls schon woanders oder online eine entsprechende Willenserklärung abgegeben hätte. Erst anlässlich des zweiten Schritts, nämlich beim Anklicken der mündlich übermittelten Willenserklärung, könnte indirekt klar werden, dass bereits eine Unterstützung getätigt worden wäre. Geplant wäre allerdings, dass selbst in diesem Fall nicht mit einer Meldung wie „Sie haben bereits unterstützt oder unterschrieben" vorgegangen wird, sondern dass die beteiligten Personen eingeladen werden, sich an eine Clearingstelle zu wenden. Mit der beschriebenen Lösung würde einem wesentlichen Kritikpunkt an der Administration des Volksbegehrens nach geltendem Recht Rechnung getragen, nämlich der notwendigen Offenlegung bestimmter politischer Interesse bei der Gemeinde des Hauptwohnsitzes.

Zwar kann das Prozedere – sieht man vom Identifizierungsvorgang ab – nicht mit einem Anonymisierungsvorgang, wie ihn ein E-Voting-Verfahren zwingend voraussetzen würde, verglichen werden. In der Praxis wäre für eine unterstützungswillige Person jedoch auch beim geplanten „Volksbegehren neu" garantiert, dass von der Unterstützung vor einer – beliebigen – Gemeinde und seinem Ansinnen nur der Organwalter (die Organwalterin) Kenntnis erlangt, im Fall einer Online-Unterstützung diese de facto

[27] Vgl. Stein, Volksbegehren goes online, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Abstraktion und Applikation, Tagungsband des 16. Internationalen Rechtsinformatik-Symposions IRIS 2013 (2013) 303.

völlig anonym bleibt und allenfalls im Fall eines Verfahrens vor dem Verfassungsgerichtshof dem mit der Entscheidung befassten Personenkreis bekannt wird. Der auf den ersten Blick im Raum stehende Umkehrschluss, nämlich, dass auch aus einer Nicht-Unterstützung einer bestimmten Initiative seitens der Entscheidungsträgerinnen und Entscheidungsträger in einer Gemeinde Rückschlüsse auf eine bestimmte Person gezogen werden könnten, kann in der Praxis bei näherer Betrachtung der Gegebenheiten in Österreich weitestgehend vernachlässigt werden. Abgesehen davon, dass Volksbegehren in der Vergangenheit von maximal 15 Prozent der Bevölkerung, zumeist aber von einem viel geringeren Anteil, unterstützt worden sind, könnte sich ein Bürger (oder eine Bürgerin) im Fall eines mentalen Druckes, ein Volksbegehren „nachweislich" zu unterstützen, bei der geplanten Rechtlage darauf berufen, diesen Schritt bereits gesetzt zu haben, ohne dass dies durch die Druck ausübende Person nachgeprüft werden könnte.

In der (bisher) letzten Entwicklungsstufe des Entwurfs wurde auch einem Hauptkritikpunkt aus dem Bereich des Datenschutzes begegnet.[28] Ursprünglich war nicht in hinreichendem Maß klargestellt, welche Stellen „Eigentümer" der Daten sind. Auch in Hinkunft wären nun die Gemeinden als „Eigentümer" der Daten die einzigen Stellen, die individuelle Berichtigungen an Datensätzen vornehmen können. Eine Ausnahme von Verbot eines Zugriffs durch „Nicht-Eigentümer" würde lediglich die Zulässigkeit des systematischen Löschens von „Flags" darstellen, z.B. in den Fällen, in denen ein Volksbegehren zurückgezogen würde oder die entsprechenden gesetzlichen Fristen abgelaufen wären. Die für das ZeWaeR gewählte Rechtskonstruktion würde es somit unmöglich machen, dass die Bundesministerin ohne Gesetzesbruch die Namen jener Personen, die ein Volksbegehren unterstützt oder unterschrieben haben, aus der Datenbank ausfiltern und in der Folge deren Daten weitergeben könnte.[29] Es ist mit Sicherheit davon auszugehen, dass im Fall einer Implementierung des ZeWaeR mit dem „Volksbegehren neu" diesem Umstand durch EDV-technische Maßnahmen Rechnung getragen würde, wie dies bei vergleichbaren Datenbank-Applikationen[30] in Österreich Standard ist.

Schon bei der Konzipierung des Gesetzesentwurfs wurde ein großes Augenmerk auf die Skalierbarkeit der Applikation gelegt (Näheres hierzu siehe auch Punkt 5). Auch wenn die Einführung von E-Voting in Österreich derzeit nicht im Raum steht, so könnten wesentliche Elemente aus dem Bereich „Volksbegehren neu" – insbesondere, was die Vorgangsweise bei der Identifizierung der wahlberechtigten Personen betrifft – in ein allfälliges zukünftiges E-Voting-Modell übernommen werden, sofern dieses, wie insbesondere aus Estland bekannt, die Identifizierung mittels qualifizierter digitaler Signatur beinhalten würde. Nach einer erfolgreichen Implementierung des ZeWaeR-Modells ließe sich die erforderliche Vorbereitungszeit für die Einführung von E-Voting

[28] Zu den Kritikpunkten in der parlamentarischen Vorbereitung vgl. u.a. http://www.parlament.gv.at/PAKT/PR/JAHR_2013/PK0295/.

[29] Diese Bedenken waren ursprünglich von Kritikern ins Treffen geführt worden, vgl. etwa „Wer sich online engagiert, wird vermerkt", Futurezone, 9.4.2013 (http://futurezone.at/netzpolitik/wer-sich-online-engagiert-wird-vermerkt/24.595.284).

[30] Etwa beim Zentralen Melderegister (ZMR) oder beim Zentralen Vereinsregister (ZVR).

deutlich verkürzen, weil auf eine E-Voting-taugliche Datenbank bereits zurückgegriffen werden könnte.

5 Wie geht es weiter mit „ZeWaeR“?

Bis zum Ende der 24. Gesetzgebungsperiode fand der Entwurf zum Demokratiepaket, der als Kernbereich die Schaffung des ZeWaeR enthielt, nicht die erforderliche parlamentarische Verfassungsmehrheit (Zwei-Drittel-Mehrheit im Nationalrat). Die Regierungsparteien – sie sind in der 24. und in der 25. Gesetzgebungsperiode identisch – haben den erwähnten Entwurf jedoch in das Regierungsübereinkommen für die neue Gesetzgebungsperiode als weiterhin anzustrebendes Ziel übernommen und angekündigt, offene Fragen zum Entwurf im Rahmen von Enquete-Kommissionen zu klären.[31] In der Ressortstrategie „Innen.Sicher“ des Bundesministeriums für Inneres ist das Zentrale Wählerregister explizit als Schwerpunkt des Innenressorts für 2014 genannt.[32] Zum gegenwärtigen Zeitpunkt kann nicht gesagt werden, ob der Entwurf in dieser Gesetzgebungsperiode eine Mehrheit finden wird. Die Gründe für ein möglicherweise neuerliches Scheitern würden wohl nicht am Konzept des ZeWaeR liegen, sondern an unter Umständen stattfindenden Junktimen betreffend andere Elemente des Entwurfs, insbesondere was den beschriebenen Automatismus von „Volksbegehren – Volksbefragung“ betrifft.

Immer wieder laut gewordener Kritik[33] gegen das geplante ZeWaeR in seinem Kernbereich, wonach sich die Bundesministerin für Inneres aufgrund der vorhandenen „Flags“ jederzeit einen Überblick über die Einstellung der Bürgerinnen und Bürger, die eine Unterstützung oder Unterschrift für ein Volksbegehren geleistet haben, verschaffen könnte, ist entgegenzuhalten, dass der Entwurf umfangreiche datenschutzrechtliche Vorkehrungen enthält, die im Lauf der parlamentarischen Behandlung in der 24. Gesetzgebungsperiode noch wesentlich verbessert worden sind. Nicht außer Acht gelassen werden darf dabei die geltende Sachlage: Rein theoretisch werden die auf die Abgabe von Unterstützungserklärungen zu Volksbegehren vorgenommenen Sperrvermerke, wie auch die Vormerkungen über die Unterschriften im Einleitungsverfahren, bei den Gemeinden dezentral verwaltet. In der Praxis gibt es, weil im Rahmen der Privatwirtschaftsverwaltung stattfindend, ohne diesbezügliche gesetzliche Regelung zudem auch jetzt schon große Datenbanken, die Daten von Unterstützungserklärungen oder allenfalls von Unterschriften zu Volksbegehren enthalten. Einerseits beruht die Existenz dieser Daten auf EDV-Lösungen in größeren österreichischen Städten, andererseits bedient eine ziemlich kleine Anzahl an Internetprovidern zahlreiche Gemeinden, so dass die entsprechenden Daten gar nicht in

[31] Vgl. Arbeitsprogramm der österreichischen Bundesregierung 2013-2018, Politische Partizipation und Grundrechte (2013) 90 (http://www.bka.gv.at/DocView.axd?CobId=53264).

[32] Vgl.“SI 19 – Zentrales Wählerregister“ (http://www.innensicher.at/files/INNEN.SICHER.2014_Langversion.pdf).

[33] Vgl Stellungnahmen im parlamentarischen Begutachtungsverfahren, abrufbar unter http://www.parlament.gv.at/PAKT/VHG/XXIV/A/A_02177/index.shtml

den Rechnern der Gemeinden, sondern in den Festplatten weniger Unternehmen gespeichert sind.

Zusammenfassend kann gesagt werden, dass ein ZeWaeR mit der Aussicht auf „Volksbegehren neu“ und mit der Möglichkeit der Skalierung auf andere Bürgerbeteiligungsverfahren einen wesentlichen Schritt in Richtung einer Bürgerbeteiligung unter Zuhilfenahme moderner Informationstechnologie darstellen würde. Bei Änderung der politischen Willenslage in Österreich lägen sogar technische Grundpfeiler vor, die einen ersten Schritt zur Implementierung einer E-Voting-Lösung darstellen würden. Es bleibt daher zu hoffen, dass der Einführung des ZeWaeR letztendlich doch noch Erfolg beschieden ist – und nicht ein Scheitern durch politische Implikationen, die außerhalb dieses Konzepts liegen.

Literaturverzeichnis

[CoE08] Council of Europe, Electoral Law. 2008.

[Lei13] Leiss, W.: „Dateneinsicht elektronisch“, Demokratiepaket umsetzen, Verwaltungsaufwand reduzieren, Kommunal 6/2013, 12.

[Ste13a] Stein, R.: „Ein ehrgeiziges Projekt“, Zentrales Wählerregister, Kommunal 3/2013, 18.

[Ste13b] Stein, R.: Kaum mehr überschaubar – Änderungen im Wahlrecht, RFG 3/2013, 114.

[Ste13c] Stein, R.: Volksbegehren goes online, in Schweighofer/Kummer/Hötzendorfer (Hrsg), Abstraktion und Applikation, Tagungsband des 16. Internationalen Rechtsinformatik-Symposions IRIS 2013 (2013) 303.

[SVW13] Stein, R.; Vogl, G.; Wenda, G.: Nationalratswahlordnung (NRWO), 4. Auflage (2013).

Informationssysteme mit Open Source

(ISOS 2014)

Informationssysteme mit Open Source (ISOS 2014)

11. Workshop des GI-Arbeitskreis Informationssysteme mit Open Source

Chairs

Prof. Dr. Olaf Herden (DHBW Stuttgart/Horb)
Prof. Dr. Holger Hinrichs (FH Lübeck)

Programmkomitee

Dr. D. Boles (Uni Oldenburg)
Prof. Dr. R. Bruns (Hochschule Hannover)
Prof. Dr. H. Eirund (Hochschule Bremen)
Prof. Dr. O. Herden (DHBW Stuttgart)
Prof. Dr. H. Hinrichs (FH Lübeck)
Prof. Dr. T. Kessel (DHBW Stuttgart)
Prof. Dr. B. Müller (Ostfalia)
Prof. Dr. M. Nüttgens (Uni Hamburg)
Prof. Dr. N. Preiß (DHBW Stuttgart)
Prof. Dr. F. Rump (Hochschule Emden/Leer)
Dr. H. Smolinski (IBM Deutschland)
Prof. Dr. H. Tapken (Hochschule Osnabrück)
Prof. Dr. T. Teschke (Hochschule Bremen)
Prof. Dr. O. Zukunft (HAW Hamburg)

Die Nutzung von Open Source Software zur Entwicklung von Informationssystemen ist in Industrie, Verwaltung und Wissenschaft ein beliebter Ansatz. Seit über 10 Jahren widmet sich der jährliche Workshop "Information Systems mit Open Source" (ISOS) diesem Trend. Der Workshop behandelt alle Aspekte der Open Source Welt – von Lizensierungsfragen bis Projektmanagement, von Standardisierungen bis Frameworks, von Softwarearchitekturen bis Fallstudien. Sowohl Praktiker als auch Wissenschaftler sind aufgerufen der ISOS-Community beizutreten und ihr Wissen mit uns zu teilen.

In den diesjährigen Beiträgen werden im Workshop die Themen Konfigurations-management-Frameworks, Kommunikationsserver, Server-Push-Mechanismen und asynchrone Servertechnologien behandelt.

Potenziale von Server Push-Mechanismen im Kontext von Informationssystemen

Herbert Neuendorf

Fakultät Wirtschaft – Studiengang Wirtschaftsinformatik
Duale Hochschule Baden-Württemberg Mosbach
Lohrtalweg 10
74821 Mosbach
neuendorf@dhbw-mosbach.de

Abstract: Ein Aspekt der Gestaltung von Informationssystemen ist die effektive Präsentation und Aktualisierung der Informationsinhalte in Echtzeit. In diesem Beitrag werden innovative Möglichkeiten von Server Push-Technologien beleuchtet, die im Rahmen standardisierter Protokolle und plattformunabhängiger Oberflächentechnologien zur Verfügung stehen. Der Beitrag versteht sich als Hinweis auf neuartige Kommunikations-Muster im Rahmen serverbasierter Anwendungen und umreißt ein offenes Forschungsfeld als work in progress.

1 Motivation und Überblick

Neuartige Technologien und offene Standards im Bereich Server Push unterstützen serverbasierte Anwendungen. Ohne vertraute serverbasierte Programmiermodelle aufgeben zu müssen, können asynchrone, event-getriebene Echtzeit-Benachrichtigungen und –Aktualisierungen der involvierten Clients dargestellt werden. Dabei handelt es sich primär um die Standards Server Sent Events (SSE) und WebSockets (WS), durch die eine polling-freie Kommunikation vom Server zum Client und dadurch eine Echtzeit-Aktualisierung des Clients (plugin-frei) erreicht wird.

Die genannten Standards stellen eine wichtige Komponente zur Realisation von Informationssystemen dar, und ermöglichen die Realisation latenzfreier Kommunikations- und Geschäftsprozesse, ohne auf proprietäre Oberflächentechnologien und Deployment-Verfahren zurückgreifen zu müssen. Dabei ist eine zunehmende Unterstützung der Standards durch gängige Browser zu beobachten.

Indem das reine Request-Response-Modell von HTTP erweitert wird um server-getriebene Benachrichtigungen, lassen sich auch im Rahmen browserbasierter Frontends und serverbasierter Inside-Out-Programmiermodelle Echtzeit-Anwendungen darstellen. Technologische Innovationen im Bereich browserbasierter Clients unterstützen somit auch serverbasierte Informationssysteme.

1.1 Server-Sent Events und Websockets im Überblick

Die Standards SSE und WS sind Teil der HTML5-Spezifikation [BER14] und betreffen die Kommunikation zwischen Server und Client (Browser), wobei die beiden Protokolle auf unterschiedlicher Protokollebene agieren (s.u.). In beiden Fällen handelt es sich um die Möglichkeit, vom Server auf Basis einer persistenten Verbindung wiederholt Daten an den Client zu senden (Server Push), ohne erneute Einzel-Requests der Clients voraus-zusetzen (wie XHR-Polling und -Long Polling). Dabei beruht SSE auf Half-Duplex-Kommunikation, während WS Full-Duplex-Kommunikation unterstützt. Durch den Aufbau einer stehenden Verbindung zwischen Client und Server entfällt die Belastung des Netzwerks durch ständigen aufwendigen Neuaufbau von TCP-Verbindungen. Dadurch reduziert sich auch der Stromverbrauch mobiler Geräte [GRI13].

Innerhalb einer Client-Anwendung können - getrieben durch vom Server ausgelöste Events und gesendete Daten - jederzeit Aktualisierungen verzugslos vorgenommen werden. Speziell SSE lassen sich serverseitig direkt implementieren, ohne dass der Server das Protokoll nativ unterstützen müsste. Somit lässt sich auch mit nicht WS-fähigen Servern ein Push-Modell zur Client-Benachrichtigung implementieren.

Einem serverbasierten Informationssystem ist es somit möglich, einen stets aktuellen Systemzustand in Quasi-Echtzeit auf dem Client abzubilden.

1.2 Anwendungsfelder Server Push

Durch Server Push-Technologien erschließen sich Anwendungsfelder innerhalb lose gekoppelter Systeme, die bislang proprietären Technologien vorbehalten waren. Im folgenden werden einige repräsentative anwendungsbezogene und anwendungs-übergreifende Szenarien genannt.

a) Anwendungsbezogene Szenarien:

- Im betrieblichen Umfeld können dispositive betriebswirtschaftliche Daten zum Monitoring innerhalb von Management-Dashboards ständig aktualisiert werden - relevant z.B. im Bereich des Transportwesens und der mobilen Logistik sowie bei mobilen Dienstleistungen (z.B. im Bereich Facilty Mangement).

- Im Bereich Business Intelligence (BI) erfordern Berechnungen auf großen Datenmengen auch weiterhin eine batch-orientierte Arbeitsweise. Bei Vorliegen berechneter Kennzahlen können browserbasierte BI-Cockpits aktualisiert werden.

- Clients können jederzeit auf Events reagieren, die im Rahmen von ereignis-orientierten Geschäftsprozessen erzeugt werden. Der Status ablaufender Prozesse kann Prozessbeteiligten gemeldet werden. Dies kann zur Unterstützung von Kollaborationsplattformen und Echtzeit-Handelsplattformen eingesetzt werden, sowie mobiles Arbeiten unterstützen.

- Innerhalb von Konferenzsystemen, SocialMedia- und Gaming-Anwendungen ist Instant Messaging realisierbar.

- Im Bereich Mobile Commerce ließen sich Location Based Services (für ortsbezogene aktuelle Informationen) innerhalb von Märkten realisieren, ohne dass Kunden zur Installation von Anwendungen genötigt wären.

b) Anwendungsübergreifende Szenarien:

- Ein Server kann proaktiv und verzugslos Statusmeldungen an Clients streamen, bzw. Probleme oder Schwellwertüberschreitungen melden.

- Messaging-Architekturen können dargestellt werden, da hier größere Datenmengen vom Message Broker (Server) zu einem oder mehreren Clients fließen und Forderungen nach Echtzeitdatenversorgung geringer Latenz bestehen. Messaging-Architekturen können eingesetzt werden, um verschiedene Teile von Informationssystemen lose zu koppeln und zu integrieren. Grundsätzlich lassen sich auch Peer-to-Peer-Ansätze unterstützen, bei denen der Server als Relay-Instanz fungiert. So sind über einen intermediären Websocket-Server 1:n Broadcast-Mechanismen zu zahlreichen Clients realisierbar.

- Innerhalb von Browser-Frontends profitiert die Kombination mehrerer Applikationen zu Meta-Applikationen (Mashups) speziell durch die Möglichkeiten der WS-Technologie.

- Auf Basis von Websockets sind Virtual Network Computing und Remote Desktop Applikationen darstellbar.

- Aktuelle Anwendungspotenziale ergeben sich im Bereich der Machine-to-Machine-Kommunikation. Die Echtzeit-Nachrichtenübermittlung von zu trackenden Objekten (Smart Meter, Fahrzeuge, Schließanlagen, etc.) könnte unterstützt werden, ohne zu propriätären Protokollen greifen zu müssen.

Alle genannten Szenarien setzen eine verzugslose oder sogar kontinuierliche Statusaktualisierung und Synchronisation des Clients voraus. Trotz der Zustands- und Session-Freiheit von HTTP können stateless arbeitende Server mittels Server Push den Benutzerkontext initiativ sofort aktualisieren, ohne diesen selbst längerfristig bis zur nächsten Client-Anfrage vorhalten zu müssen. Insofern stellt Server Push eine grundsätzliche technische und semantische Ergänzung des zustandslosen HTTP-Protokolls dar. Eine zusätzliche Last auf dem Server entsteht nur durch das Offenhalten der Verbindung zum Client. Jedoch wird das Latenzproblem gelöst, das entsteht, wenn der Client Statusaktualisierungen stets aktiv anfordern muss.

In allen Fällen lassen sich browserbasierte und auch mobile Lösungen latenzfrei realisieren, ohne auf spezielle Programmiermodelle angewiesen zu sein, die einem serverbasierten InSide-Out-System einen Outside-In-Programmieransatz aufzwingen würden. Der Vorteil browserbasierter Frontends im betrieblichen und mobilen Umfeld beruht auf deren installationsfreier Aktualität. Die Notwendigkeit eines

Endgerätemanagements (BYOD-Problematik) wird (bis auf die Installation eines einheitlichen Browsertyps) deutlich entschärft. Durch den Verzicht auf proprietäre Plug-In-Technologien erhöht sich die Zukunfts- und Investitionssicherheit von Anwendungen.

2 Server-Sent Events

Server-Sent Events (SSE) arbeiten auf Basis von HTTP [HIC12b]. Sie werden bereits von gängigen Browsern (Firefox, Chrome, Safari, Opera – nicht jedoch Internet Explorer) unterstützt. SSE beruhen auf einer Half-Duplex-Technologie, d.h. es ist nur vorgesehen, textuelle Daten vom Server zum Client (Browser) zu senden. Voraussetzung ist ein vorausgegangener Verbindungsaufbau des Clients zum Server. Im Anschluss kann dieser Server jederzeit Benachrichtigungen (push notifications) und Informationsaktualisierungen in Echtzeit über eine langlebige offene HTTP-Verbindung zum Client senden, ohne dass dazu periodisches clientseitiges Polling erforderlich wäre.

Dies stellt gegenüber der XHR-Technologie (Ajax) [KES14b] auf Basis von periodischem Client-Polling bzw. Long-Polling („hanging GET“ oder „Comet“) (mit entsprechender Bandbreitenbelastung bzw. verzögerter Benachrichtigung) einen deutlichen Effizienzgewinn dar. Mittels SSE sind im Gegensatz zu XHR auch effiziente Streaming-Lösungen darstellbar, während die geplante Erweiterung des XHR-Protokolls durch eine Stream API [MOU13] momentan noch nicht verfügbar ist [GRI13].

SSE nutzen HTTP und setzen (im Gegensatz zu Websockets) nicht die serverseitige Unterstützung eines weiteren Protokolls voraus. Vielmehr kann die HTTP-Response des Servers in einem konventionellen Servlet ohne direkte serverseitige Protokoll-Unterstützung implementiert werden. Auf Seiten des Clients übernimmt der Browser das Verbindungsmanagement und die Details der Datenextraktion, so dass die Applikation sich auf die Geschäftslogik fokussieren kann.

SSE unterliegen (wie XHR) grundsätzlich der Same-Origin-Policy (SOP) [BAR10], so dass Cross-Origin Requests (ursprungsübergreifende Kommunikation) nicht zugelassen sind. Somit können Benachrichtigungen nur von dem Server an den Client gesendet werden, an den der eröffnende Request des Clients adressiert war. Inzwischen wurde die Spezifikation jedoch zugunsten größerer Flexibilität unter Berücksichtigung möglichst hoher Sicherheit erweitert, so dass Cross-Origin Resource Sharing (CORS) [KES14a] erlaubt wird, um Webclients Cross-Origin Requests gegen fremde Server zu ermöglichen. Allerdings wird diese SSE-Erweiterung erst von wenigen Browsern unterstützt. Sicherheit wird also je nach Browsertyp durch SOP oder die CORS-API gewährleistet.

2.1 Verbindungsaufbau und Datentransfer

Clientseitig beruht die SSE-API ausschließlich auf dem `EventSource`-Objekt. Zum Verbindungsaufbau zum Server genügt es, ein `EventSource`-Objekt mittels URL zu erstellen. In beschränktem Maße kann der Client dabei Daten zum Server senden, indem

der URL ein Query-String angefügt wird. Daraufhin baut der Browser eine Verbindung zum Server auf und sendet einen HTTP-Get-Request.

Im Client sind Handler-Funktionen für zu behandelnde Events zu implementieren. Diese werden asynchron aufgerufen und aktualisierte Daten übergeben, wenn vom Server eine Benachrichtigung gesendet und ein entsprechendes Event im Client ausgelöst wird – ohne jegliches (potenziell verzögertes) Polling durch den Client. Im Client stehen die Events `open`, `close`, `error` und `message` zur Verfügung.

Auf Seiten des Servers sind bezüglich des EventStream-Datenformats nur wenige Anforderungen zu beachten: Der Content-Type der Response muss explizit text/event-stream lauten und die textuelle Response in UTF-8 kodiert sein. Die eigentlichen Nutzdaten sind Werte eines oder mehrerer aufeinander folgender `data:`-Felder. Auch das Versenden im JSON-Format ist möglich. Für syntaktische Details verweisen wir auf die Spezifikation [HIC12b].

Nachrichten können mit einem frei wählbaren Eventnamen versehen werden. Serverseitig wird dazu der Benachrichtigung eine Zeile mit `event:<name>` hinzugefügt. Auf diese Weise können innerhalb eines einzigen Server Push mehrere verschiedene benamte Nachrichten gleichzeitig an den Client gesendet und von diesem separat behandelt werden. Benamte Nachrichten lösen entsprechend benamte Events aus, die durch spezielle Event Handler des Clients separat behandelt werden können. Auf diese Weise lassen sich Verhaltensweisen des Clients feingranular durch serverseitige Benachrichtigungen auslösen – z.B. für unterschiedliche Statusmeldungen des Servers.

Zur Realisation von Streaming-Lösungen muss die Verbindung permanent offen gehalten werden. Deshalb sendet das `EventSource`-Objekt des Clients automatisch einen neuerlichen Get-Request zum Server, sobald die Verbindung erfolgreich beendet oder durch einen Verbindungsfehler unterbrochen wurde. Konfigurierbar ist das zeitliche Client-Verhalten durch die `retry`-Option der Server-Response. Die client-seitige API abstrahiert von den Details des Aufbaus der Verbindung, deren automatischer Aufrechterhaltung und Wiederherstellung, was die Implementierung clientseitiger Anwendungslogik erleichtert.

Der Neuaufbau einer geschlossenen Verbindung durch den Client kann im Client durch Aufruf der Methode `close()` des `EventSource`-Objekts beendet werden. Dieser Aufruf schließt sofort die aktuelle Verbindung zum Server. Auch servergesteuert lässt sich die `EventSource` stoppen, indem eine HTTP-Response ohne Content-Type Header oder mit Status-Code ungleich 200 gesendet wird. Letzteres ist bei Streaming-Szenarien nicht möglich, da Header-Informationen bereits zu Streaming-Beginn zum Client gesendet wurden. In diesem Fall muss der Server ein speziell benamtes Event senden, auf das der Client mit Schließen des `EventSource`-Objekts reagiert.

Mittels des Nachrichtenelements `id` kann der Server jeder Nachricht einen beliebigen Identifikations-String zuweisen, der im Attribut `lastEventID` des `event`-Objekts abgelegt wird. Auf diese Weise kann der Client überprüfen, ob er alle Nachrichten erhalten hat. Wird die Verbindung zum Server beendet oder unterbrochen, so sendet der

Client beim nächsten Request den zusätzlichen HTTP-Header `Last-Event-ID` zum Server, der den letzten vom Client erhaltenen `id`-Wert enthält. Der Server kann den Header abfragen und entsprechend reagieren - z.B. alle Nachrichten seit dieser ID zum Client senden oder den Verlust ignorieren.

Zur Erhöhung der Sicherheit steht auf dem `event`-Objekt das Attribut `origin` zur Verfügung. Somit kann der Ursprung jeder Servernachricht im Client überprüft werden.

2.2 Bewertung

SSE sind das Mittel der Wahl, wenn textuelle Echtzeitdaten vom Server geliefert werden sollen, ohne dass der User einen Rückmeldungskanal zum Server benötigt. Markante Eigenschaften des SSE-Protokolls sind seine einfache und flexible Handhabbarkeit. Es wird HTTP als Trägerprotokoll genutzt, das von CORS unterstützt werden kann. Da der Rahmen des HTTP-Protokolls nicht verlassen wird, treten bei Proxy-Servern keine Akzeptanzprobleme beim Verbindungsaufbau auf.

Durch Verwendung von HTTP können SSE serverseitig im Rahmen jedes Webframeworks dargestellt werden, dass eine Manipulation der HTTP-Response erlaubt. Eine native Unterstützung durch spezielle Server oder Webframeworks ist nicht erforderlich. Browserseitig sind SSE relativ einfach durch Javascript-Bibliotheken (Polyfills) abbildbar (wie im Falle IE9), falls sie nicht nativ unterstützt werden.

Somit sollten SSE zur Realisation textueller Server Push-Szenarien geringer Latenz in Betracht gezogen werden, ohne im Vornherein auf die Websocket-Technologie zu fokussieren.

3 Websockets

Das Websocket-Protokoll wird spezifiziert durch die IEFT (RFC6455) [FET11], die Websocket-API durch das W3C [HIC12a]. Websockets ermöglichen eine bidirektionale Full-Duplex-Kommunikation textueller oder binärer Daten zwischen einem beliebigen Client (u.a. Browser) und einem Server auf Basis des WS-Protokolls. Somit werden (im Gegensatz zu SSE) WS-fähige Server zur Realisation benötigt. Eine beispielhafte Java-basierte Server-API ist spezifiziert in [COW13].

Das WS-Protokoll wurde speziell zur bidirektionalen Kommunikation entworfen, statt (wie bei SSE) eine zusätzliche Protokollschicht oberhalb von HTTP zu definieren. Dadurch entfällt zusätzlicher HTTP-Overhead nach Aufbau der Verbindung.

3.1 Verbindungsaufbau

Das WS-Protokoll [FET11] verwendet einen initialen HTTP-Request über HTTP-Standard-Ports und den HTTP-Upgrade-Mechanismus, um beim Opening-Handshake eine HTTP-Verbindung zu einer WS-Verbindung aufzuwerten. Das WS-Protokoll wird

mittlerweile von allen gängigen Browsern unterstützt. Neben dem WS-Protokoll existiert auch das WSS-Schema, das zur Verschlüsselung auf TLS (SSL) aufsetzt.

Bislang unterstützen WS noch kein Verbindungs-Multiplexing - d.h. jede WS-Verbindung basiert auf einer dedizierten TCP-Verbindung. Dies kann problematisch sein angesichts der begrenzten Zahl offener Browser-Verbindungen. Aktuell ist eine Multiplex-Erweiterung für WS in Entwicklung [TAM12], durch die eine physische TCP-Verbindung als Träger mehrerer separater logischer WS-Verbindungen fungieren kann, indem jede einzelne WS-Verbindung durch eine eigene Kanal-ID ausgezeichnet wird. Mehrere WS-Verbindungen (channels) können somit über dieselbe geteilte TCP-Verbindung dargestellt werden.

Grundsätzlich geht die Verbindungsaufnahme stets vom Client aus. Webclients können keine WS-Verbindungen akzeptieren, die sie nicht selbst initiiert haben. Somit sind Clients vor ungewollten Verbindungsaufnahmen geschützt.

Eine wichtige Neuerung ist die Cross-Domain-Fähigkeit von WS, die nicht den Beschränkungen der Same-Origin-Policy (wie SSE) unterworfen sind, so dass eine Kommunikation zwischen verschiednen Domänen möglich ist. Innerhalb einer WS-Applikation können somit z.B. separate Server für die Kern-Applikation einerseits und Streaming-Aufgaben andererseits eingesetzt werden (client side aggregation). Somit entfällt die Notwendigkeit, die gesamte Applikation auf einem zentralen Server zu hosten (server side aggregation) und der damit verbundene Engpass. Eine Verbindung kann zu jedem beliebigen Server aufgebaut werden, sofern das Handshake-Protokoll korrekt ausgeführt wird.

Dem Schutz des Servers vor DenialOfService-Attacken dient das Header-Feld `Origin` innerhalb der WS-Client-Anfrage. Durch dieses werden Schema, Host und Port des anfragenden Clients spezifiziert. Der `Origin`-Header ermöglicht es dem Server, WS-Verbindungen zu akzeptieren oder abzulehnen. Der Browser ist verpflichtet den `Origin`-Header beim Handshake mitzusenden. Der Server kann anhand des Wertes entscheiden, ob er den Verbindungsaufbau bestätigt oder durch eine Fehlermeldung ablehnt. Serverseitig kann vorgegeben werden, ob die Serverdienste Clients jeglicher Domäne oder nur definierter Domänen zugänglich sind.

3.2 Unterstützung von Subprotokollen

Das clientseitige `Websocket`-Objekts wird durch Angabe der Server-URL im Client erzeugt [HIC12a]. Als optionaler Konstruktor-Parameter können Namen von Subprotokollen (als String-Array) an den Server übermittelt werden, die auf Basis der WS-Kommunikation verwendet werden sollen. Das WS-Protokoll kennzeichnet die übermittelten Daten als textuell oder binär, macht selbst aber keine weiteren Vorgaben über deren Format und das ihrer Nutzung zugrunde liegende Protokoll. Der Server wählt ein von ihm unterstütztes Protokoll aus, so dass Client und Server mittels des ausgehandelten Zusatzprotokolls interagieren können. Die Mitteilung der erfolgreichen Protokollauswahl durch den Server und der somit erreichten Protokollvereinbarung

zwischen Client und Server geschieht über den WS-Header `Sec-WebSocket-Protocol` der Server-Antwort und das gefüllte Attribut `protocol` des `WebSocket`-Objekts. Kann der Server das vom Client vorgeschlagene Protokoll nicht unterstützen, so schließt er die Verbindung und das `error`-Event wird im Client ausgelöst. Bei den Protokollen kann es sich um offiziell registrierte WS-Subprotokolle handeln, weit verbreitete standardisierte Protokolle oder um selbst definierte Protokolle, z.B. auch Kompressionsverfahren bei textuellem Datenverkehr [GRI13].

Vorteil der WS-Technologie ist somit, dass Protokollerweiterungen und zusätzliche Protokolle spezieller Semantik und Funktionalität auf der WS-Protokollschicht aufbauen können. WS kann als Transportschicht für höhere Protokolle dienen. WS baut auf TCP/TLS auf, wird aber selbst auch als Transportprotokoll betrachtet, da Protokolle der Applikationsschicht auf WS fußen können. Typische Beispiele sind das eXtensible Messaging and Presence Protocol (XMPP) und das Simple Text Oriented Messaging Protocol (STOMP) [WAN13]. Im Rahmen einer WS-Anwendung wird HTTP jedoch weiterhin genutzt werden, um statische Ressourcen (HTML, CSS, JS) zu laden. Insgesamt ergibt sich somit der Protokollstapel gemäß Abbildung 1.

<table>
<tr><td colspan="3">Applikation</td></tr>
<tr><td>SSE</td><td>XMPP</td><td>STOMP</td></tr>
<tr><td>HTTP</td><td colspan="2">WS</td></tr>
<tr><td colspan="3">TLS</td></tr>
<tr><td colspan="3">TCP</td></tr>
</table>

Abbildung1: Protokollstapel im Kontext der Protokolle SSE und WS.

3.3 Kommunikation und Datenaustausch

Das Kommunikationsmodell folgt clientseitig dem Eventhandler-Muster; es können Behandler (asynchrone Callback-Funktionen) für die Events `open`, `message` und `close` definiert werden. Die WS-API arbeitet rein eventbasiert, jegliches Polling entfällt. Events und Benachrichtigungen werden asynchron empfangen und verarbeitet.

Meldet der Server einen Fehler, so wird auf dem Client das `error`-Event ausgelöst und die Verbindung geschlossen. Ein automatischer Neuaufbau der Verbindung durch den Client (wie bei SSE) findet nicht statt. Der Client kann durch Aufruf der `close()`-Methode des `Websocket`-Objekts die Verbindung zum Server schließen.

Sobald die Verbindung zum Server etabliert wurde (sichtbar durch Auslösen des `open`-Events im Client), kann der Client Daten an den Server mittels der Methode `send()` des `Connection`-Objekts senden. Ebenso kann der Server jederzeit Daten an den Client senden, wodurch in diesem das `message`-Event ausgelöst wird und die Daten im

`data`-Attribut des `event`-Objekts zur Verfügung stehen. In beiden Fällen kann es sich um textuelle Daten oder um Binärdaten handeln. Es besteht für den Server die Möglichkeit, alle verbundenen Clients zu adressieren und Aktualisierungen an diese gemeinsam zu versenden [COW13].

Eine aktive WS-Verbindung kann vor serverseitiger Überlastung geschützt werden, indem der Client die Datenübertragungsrate anpasst (throttling). Das `Websocket`-Attribut `bufferedAmount` legt die Zahl der Bytes fest, die auf dem Client vor der Versendung zum Server zwischengepuffert werden.

3.4 Proxy-Problematik

WS-Verbindungen können längere Zeit geöffnet bleiben. Dies kann zum Erreichen von Timeouts in zwischengeschalteten intermediären (forward oder reverse) Proxy-Servern oder Routern führen, durch die die (scheinbar untätige TCP- und somit auch WS-) Verbindung geschlossen wird. Um dies zu vermeiden, unterstützt das WS-Protokoll browser- und serverseitig initiierte `ping`- und `pong`-Aufrufe in frei wählbaren Intervallen, um die Verbindung offen zu halten.

Ebenso wie HTTP kann auch WS den Schutz durch Tunneling via TLS nutzen; dies entspricht dem Schema WSS (Websocket Secure) und kommt durch einen HTTPS-Handshake-Upgrade von HTTPS zu WSS zustande. Neben reinen Sicherheitsaspekten gestaltet sich dadurch die Durchleitung des WS-Verkehrs durch intermediäre Proxies unproblematisch, da verschlüsselter Datenverkehr durch Proxies nicht inspiziert und somit durchgeleitet wird. Ferner sehen Proxies für Tunnel typischerweise großzügigere Timeout-Einstellungen vor.

3.5 Bewertung

Durch WS wird ein neues symmetrisches Kommunikationspattern für serverbasierte Anwendungen eröffnet und das HTTP-Request-Response-Dogma durchbrochen. Dazu müssen WS-Server jedoch zahlreiche gleichzeitig geöffnete WS-Verbindungen bewältigen. Dies setzt eine performante und hochgradig nebenläufige Serverarchitektur voraus, typischerweise auf Basis von Thread-Pooling und nichtblockierenden Prozessen.

Insgesamt ermöglichen WS erstmals eine bidirektionale textuelle und binäre asynchrone Echtzeit-Kommunikation zwischen Client und Server mit minimaler Benachrichtigungsverzögerung.

Besondere Stärke des Websocket-Protokolls ist seine grundsätzliche Fähigkeit, als Träger beliebiger bidirektionaler Subprotokolle zwischen Client und Server zu fungieren, die innerhalb des Eröffnungs-Handshakes zwischen Client und Server vereinbart werden können.

4 Praktische Umsetzungen

Praktische Erfahrungen mit den Standards SSE und WS konnte im Rahmen von studentischen Projekten des Studiengangs Wirtschaftsinformatik der DHBW Mosbach gesammelt werden. Hierbei wurde eine hochschulbezogene Standort-WebApp realisiert, die aktuelle Informationen zu Themen des studentischen Alltags browserbasiert darstellt. Im Rahmen der Anwendung wurde u.a. bereits eine Chat-Möglichkeit auf Basis von WS realisiert sowie ein News-Ticker auf Basis von SSE. Dabei wurden SSE auf aktuellen Versionen der Browser Chrome, Firefox, Opera und Chrome für Android betrieben, WS auch auf Internet Explorer. Als Server kamen Tomcat und Jetty zum Einsatz

Momentan wird eine Firmenbefragung vorbereitet, um potenzielle betriebliche Nutzungsmöglichkeiten der Standards SSE und WS zu evaluieren und im Rahmen kooperativer Forschungsprojekte zu pilotieren.

5 Fazit

In dem Maße, in dem sich das Internet wandelt von einer dokumentenorientierten zu einer ereignisorientierten Nutzungs- und Wahrnehmungsweise (augmented web), werden auch standardisierte Server Push-Technologien an Bedeutung gewinnen. Speziell durch SSE und WS erschließen sich serverbasierten Informationssystemen weitere Kommunikationskanäle und Kommunikationsmuster. In Tabelle 1 werden charakteristische Protokolleigenschaften vergleichend gegenüber gestellt.

Eigenschaft	**TCP**	**UDP**	**HTTP**	**XHR**	**SSE**	**WS**
Adressierung	IP:Port	IP:Port	URL	URL	URL	URL
Übertragung	Full Duplex	Datagramms	Half Duplex	Half Duplex	Half Duplex	Full Duplex
Inhalt	Bytestream	Bytestream	Mimes	Mimes	Text	Text + Binär
Basisprotokoll	IP	IP	TCP	HTTP	HTTP	TCP
Verbindungs-orientiert	Ja	Nein	Nein	Nein	Ja	Ja
Verlässlichkeit	Ja	Nein	Ja	Ja	Ja	Ja
Cross Domain Messaging	--	--	(Nein) (CORS)	(Nein) (CORS)	(Nein) (CORS)	Ja

Tabelle 1: Eigenschaften verschiedener Web-Protokolle im Umfeld von Server Push.

Server Push-Technologien unterstützen und stärken vertraute serverbasierte Architekturen und Inside-Out-Programmiermodelle, durch die Geschäfts- und Prozesslogik auf dem Server zentralisiert, stabil und einheitlich bereitgestellt werden.

Hinz und Kessel [HIN13] weisen darauf hin, dass webbasierte Frontendsysteme zunehmend auf dedizierte Usergruppen und native Frontendtechnologien zugeschnitten wurden, so dass durch parallele Mehrfachentwicklung ähnlicher Abläufe siloartige, monolithische Infrastrukturen mit hohen Kosten der Neuerstellung und Wartung entstanden. Die konsequente Nutzung neuartiger, standardisierter Kommunikationskanäle kann dazu beitragen, einer architekturellen Zersplitterung und Schaffung redundanter Zugriffskanäle und Systemstrukturen vorzubeugen.

Literaturverzeichnis

[BAR10] Barth, A. (IETF) (2010): The Web Origin Concept (Draft), November 2010. http://tools.ietf.org/html/draft-abarth-origin-09 (Abgriff 22.04.2014)

[BER14] Berjon, R. et.al. (W3C) (2014): HTML5 – A vocabulary and associated APIs for HTML and XHTML. W3C Candidate Recommendation, 04 February 2014. www.w3.org/TR/html5/ (Abgriff 22.04.2014)

[COW13] Coward, D. (2013): Java API for WebSocket. JSR 356, Version 1.0, 26 April 2013, https://jcp.org/en/jsr/detail?id=356 (Abgriff 22.04.2014)

[FET11] Fette, I.; Melnikov, A. (IETF) (2011): RFC 6455 - The WebSocket Protokoll, December 2011. http://tools.ietf.org/html/rfc6455.txt (Abgriff 22.04.2014)

[GRI13] Grigorik, I. (2013): High Performance Browser Networking. O'Reilly Media. Sebastopol.

[HIC12a] Hickson, I. (W3C) (2012): The WebSocket API. W3C Candidate Recommendation, 20 September 2012. www.w3.org/TR/websockets/ (Abgriff 22.04.2014)

[HIC12b] Hickson, I. (W3C) (2012): Server-Sent Events. W3C Candidate Recommendation, 11 December 2012. www.w3.org/TR/eventsource/ (Abgriff 22.04.2014)

[HIT13] Hitz, M.; Kessel, K. (2013): Einfluss der Nutzung und Auswahl von Open Source Software beim Entwurf einer Multikanal-Architektur. Workshop ISOS 2013. GI-Tagung 2013.

[KES14a] van Kesteren, A. (W3C) (2014): Cross-Origin Resource Sharing. W3C Recommendation, 16 January 2014. www.w3.org/TR/cors/ (Abgriff 22.04.2014)

[KES14b] van Kesteren, A. (W3C) (2014): XMLHttpRequest Level 1. W3C Working Draft, 30 January 2014. www.w3.org/TR/XMLHttpRequest/ (Abgriff 22.04.2014)

[MOU13] Moussa, F; Yoshino, T (2013): Streams API. W3C Working Draft, 05 November 2013. www.w3.org/TR/streams-api/ (Abgriff 22.04.2014)

[TAM12] Tamplin, J.; Yoshino, T. (IETF) (2012): A Multiplexing Extension for WebSockets (Internet-Draft), 28 February 2012. https://tools.ietf.org/html/draft-tamplin-hybi-google-mux-03 (Abgriff 22.04.2014)

[WAN13] Wang,W.; Salim,F.; Moskovits,P. (2013): The Definitive Guide to HTML5 WebSockets. Apress. New York.

Systemarchitektur eines verteilen, ereignisgetriebenen Konfigurationsmanagement-Frameworks auf Basis von Open-Source-Komponenten

Martin Helmich

Hochschule Osnabrück
martin.helmich@hs-osnabrueck.de

Mittwald CM Service GmbH & Co. KG
m.helmich@mittwald.de

Abstract: Moderne Open-Source-Konfigurationsmanagementsysteme wie CFEngine, Puppet oder Chef arbeiten als zentralisierte Systeme, in denen der Soll-Zustand eines IT-Systems in deklarativer Art beschrieben wird. Diese Definitionen werden von auf den Zielsystemen laufenden Agenten abgerufen und mit dem Ist-Zustand abgeglichen. Dieser Ansatz ist nur bedingt geeignet für Systeme mit extrem häufigen Konfigurationsänderungen, die zudem unter Einhaltung bestimmter Zeitbedingungen umgesetzt werden müssen. Darüber hinaus ist es in vielen Lösungen nicht möglich, hostübergreifende Abhängigkeiten von Systemzuständen zu definieren.

Dieser Artikel beschreibt eine mögliche Architektur für ein Softwaresystem, das als Teilkomponente eines komplexen Konfigurationsmanagementsystems genutzt werden kann. Insbesondere wird darauf eingegangen, wie aus frei verfügbaren Open-Source-Komponenten eine hochskalierbare Message-Broker-basierte Architektur aufgebaut werden kann, die in der Lage ist, beliebig programmierbare Konfigurationsänderungen mit sehr geringer Latenz auf einer großen Anzahl heterogen konfigurierter Hosts durchzuführen.

Im Unterschied zu den oben genannten Lösungen arbeitet das hier vorgestellte System nicht mit einer deklarativen Beschreibung des Soll-Zustandes, sondern nimmt Aufträge entgegen, die in Form imperativer Skripte auf den Zielsystemen ausgeführt werden. Ferner ist es imstande, Abhängigkeiten einzelner Konfigurationsänderungen untereinander zu berücksichtigen und somit auch komplexe, aufeinander aufbauende, hostübergreifende Folgen von Aufträgen zu automatisieren.

1 Einleitung und Motivation

Zur effizienten Administration größerer IT-Systeme werden heutzutage Konfigurationsmanagementlösungen wie CFEngine [CFE14a], Chef [Che14] oder Puppet [Pup14b] eingesetzt [DJV10]. In bestimmten Fällen stoßen diese jedoch an ihre Grenzen; sei es in der Anzahl der zu verwaltenden Systeme, der Häufigkeit, mit der Änderungen an verwaltete Systeme propagiert werden müssen, oder hinsichtlich der Zugänglichkeit durch den Endanwender.

Ein spezieller Anwendungsfall besteht bei Hosting- oder Cloud-Dienstleistern, die bestimmte Bereiche der von ihnen zur Verfügung gestellten Infrastruktur ihren Endkunden selbst zur freien Konfiguration überlassen möchten. Besondere Herausforderungen in solchen Systemen bestehen darin, Konfigurationsänderungen schnell umzusetzen und ein besonders hohes Maß an Stabilität sicherzustellen. Aufgrund der Vielzahl unprivilegierter Nutzer muss zudem ein hohes Maß an Sicherheit garantiert werden. Nicht zuletzt muss dem Nutzer eine gut bedienbare Nutzerschnittstelle zur Verfügung gestellt werden.

Ein konkreter Fall eines solchen Systems ist das Serverautomatisierungssystem des Webhosting-Unternehmens MITTWALD CM SERVICE [Mit14]. Dieses betreibt eine Infrastruktur aus derzeit ca. 8.000 Servern (Tendenz steigend), von denen einige dediziert einzelnen Kunden zur Verfügung stehen und andere von mehreren Kunden gemeinsam genutzt werden (*shared hosting*). Das System bietet dem Kunden eine grafische Oberfläche an, über die gebuchte Webhosting-Instanzen konfiguriert werden können (z.B. Domains, DNS, E-Mail).

Ziel dieses Artikels ist es, eine auf Open-Source-Komponenten basierende Architektur für eine Teilkomponente eines Konfigurationsmanagementsystems vorzustellen, die eine effiziente Ausführung von beliebig programmierbaren Aufgaben auf einzelnen Hosts oder Host-Gruppen eines großen Netzwerks ermöglicht und gleichzeitig die Skalierbarkeit des Gesamtsystems sicherstellt. Insbesondere wird dabei auf die Implementierung einer Publish-Subscribe-Architektur [TvS07] mittels des Advanced Message Queuing Protocol (AMQP) [OAS12] eingegangen. Diese Architektur soll ein ähnliches System ablösen, welches im Unternehmen bereits im Einsatz ist. Dieses bearbeitet derzeit bereits mehrere 10.000 Konfigurationsänderungen pro Tag, wird jedoch in mittelfristiger Zukunft voraussichtlich an die Grenzen seiner Skalierbarkeit stoßen.

2 Stand der Forschung

Die Grundlage moderner Konfigurationsmanagementsysteme bildet [Bur95], in welchem das Werkzeug CFEngine erstmalig vorgestellt wurde. Eine formalere Beschreibung der theoretischen Grundlagen solcher deklarativen Konfigurationsmanagementwerkzeuge erfolgt in [CHIK03] und [BC06]. In [CHIK03] wird insbesondere der Begriff der Konvergenz definiert. Dieser Begriff beschreibt die Eigenschaft eines Systems, stets einen zentral definierten Soll-Zustand anzustreben. In [DJV10] werden aktuelle Werkzeuge (sowohl als Open-Source-Software (OSS) als auch kommerziell) anhand verschiedener Kriterien untersucht. Als relevante OSS-Werkzeuge werden dabei Chef [Che14], Puppet [Pup14b] und das bereits genannte CFEngine [CFE14a] identifziert.

Neuere Ansätze hinsichtlich des Konfigurationsmanagements sind Salt [Sal14] und Ansible [Ans14]. Die Besonderheit von Salt ist der Nachrichtenaustausch mithilfe von Message Queues gegenüber den von den übrigen Lösungen genutzten REST-APIs [ZGW+13]. Die in diesem Artikel vorgestellte Architektur arbeitet ebenfalls mit Message Queues, setzt jedoch mit einem AMQP-Broker auf eine dedizierte Middleware, während Salt mit ZeroMQ auf eine eingebettete Queue setzt [Hos13].

In [VJ13] wird ein Ansatz vorgestellt, mit welchem auch komplexere Konfigurationsaufgaben automatisiert werden können. In Ergänzung zu dem üblichen deklarativen Ansatz erlaubt [VJ13] auch imperative Programmierung der System-Konfigurationen und kann Systemzustände hostübergreifend modellieren.

Im Bereich der Message-Broker hat sich in den letzten Jahren das Advanced Message Queueing Protocol (AMQP) als Standard etabliert. AMQP wurde erstmals vorgestellt in [Vin06] und später endgültig standardisiert in [OAS12]. Eine Beschreibung des AMQP zugrunde liegenden Modells liefert [KD14] (vgl. auch Abb. 1). [YZF+11] untersuchen den Aufbau skalierbarer Architekturen in Cloud-Computing-Umgebungen unter Einsatz von AMQP-Queues, hier insbesondere RabbitMQ [Piv14b]. Zielsetzung ist hier jedoch primär die effiziente Ausführung rechenintensiver Aufgaben. In [FLRU13] wird die Leistungsfähigkeit von AMQP-basierten Services (auch hier insbesondere RabbitMQ) mit RESTful Webservices verglichen.

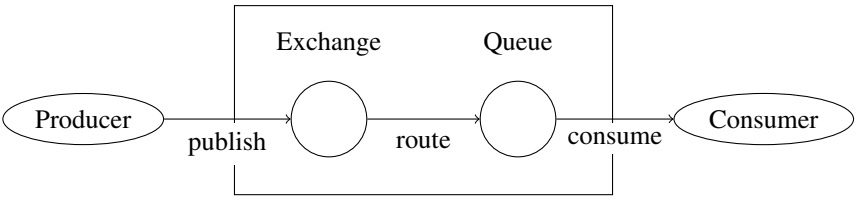


Abbildung 1: Das AMQP zugrunde liegende Modell nach [KD14].

3 Anforderungen an das System

3.1 Funktionale Anforderungen

Im Unterschied zu ganzheitlichen Lösungen wie [Che14], [Pup14b] oder [CFE14a] liegt der Fokus der hier vorgestellten Architektur ähnlich wie bei [Pup14a] ausschließlich auf der effizienten Ausführung von Konfigurationsaufträgen auf den Hosts eines Netzwerks. Diese Lösung ist dazu vorgesehen, in ein Gesamtsystem integriert zu werden und soll entsprechende Schnittstellen für Drittapplikationen anbieten, die ihrerseits den Sollzustand des Systems modellieren und entsprechende Änderungen an die Zielsysteme propagieren.

Primär muss das System eine Programmierschnittstelle anbieten (beispielsweise per SOAP [GHM+07] oder RESTful-HTTP [Fie00]), über die Drittanwendungen Aufträge an beliebige Hosts verteilen können. Diese werden dann in Form vordefinierter und parametrisiert aufgerufener Skripte auf den Hosts ausgeführt, um einen bestimmten Konfigurationszustand herzustellen.

Die auf den Zielsystemen auszuführenden Skripte sollen versioniert in einem zentralen Repository verwaltet werden können. Dabei muss eine heterogene Systemzusammensetzung berücksichtigt werden; je nach Konfiguration des Hosts (Betriebssystem, Distribution, Version, ...) sind zur Bearbeitung desselben Auftrags möglicherweise unterschiedliche Skripte notwendig.

3.2 Qualitative Anforderungen

Eine zentrale Anforderung liegt in der Skalierbarkeit des Systems. So stellen [DJV10] fest, dass zahlreiche der OSS-Werkzeuge nur in Netzwerken bis 10.000 Hosts skalieren. CFEngine ist nach Herstellerangaben [CFE14b] in Standardkonfigurationen für ca. 2.000 Hosts pro Management-Server geeignet; zur weiteren Skalierung sind sehr spezielle Anpassungen erforderlich. Eine weitere Einschränkung ergibt sich hinsichtlich der Anzahl der pro Host konfigurierbaren Ressourcen [CFE14b]. In der vorliegenden Systemarchitektur ergibt es sich, dass auf einzelnen Hosts teils sehr viele Ressourcen verwaltet werden müssen (etwa mehrere 100.000 Firewall-Regeln oder mehrere Millionen DNS-Einträge). Zudem bestehen strenge Anforderungen hinsichtlich der Zeit, in der Aufträge an die Hosts zugestellt und dort bearbeitet werden müssen. Die Zustell- und Bearbeitungszeiten sollen mit zunehmender Skalierung des Systems konstant bleiben.

Da das System in einem Netzwerk aus heterogen konfigurierten Hosts eingesetzt werden soll, besteht ebenfalls die Anforderung der Portabilität. Die Systeme umfassen diverse Linux-Distributionen wie etwa Debian, Suse, CentOS und Embedded-Varianten (z.B. auf Netzwerkhardware); auch die Nutzung von Windows-Hosts ist zukünftig nicht ausgeschlossen.

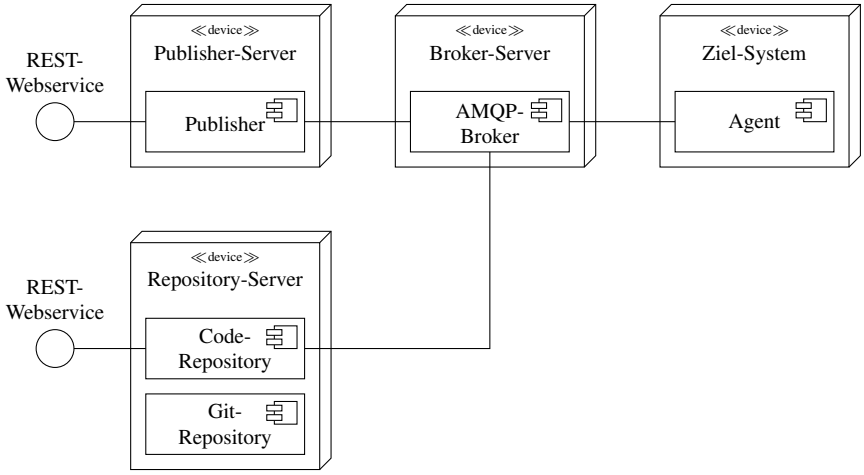


Abbildung 2: Ein exemplarisches Deployment-Diagramm der vorgestellten Systemarchitektur.

4 Systemarchitektur

Die einzelnen Systemkomponenten bestehen aus einem Publisher, der Drittapplikationen Schnittstellen anbietet, um beliebige Konfigurationsänderungen auf Hosts im Netzwerk durchzuführen. Der Publisher veröffentlicht zugestellte Konfigurationsaufträge in einem Message Broker, der diese an auf den Hosts laufende Software-Agenten zustellt. Ein Code-Repository hält in verschiedenen Skriptsprachen frei programmierbare Vorlagen für auf den Hosts auszuführende Skripte bereit, die von den Software-Agenten bei Bedarf abgerufen werden. Abbildung 2 zeigt eine exemplarische Verteilung der hier vorgestellten Architektur.

4.1 Publisher-Komponente

Ein Publisher ist jeweils dafür zuständig, Aufträge für eine von ihm verwaltete Gruppe von Hosts entgegen zu nehmen. Hierzu wird eine REST-Schnittstelle [Fie00] angeboten, die von Dritt-Applikationen angesprochen werden kann. Um trotz der grundsätzlich asynchronen Bearbeitung von Aufträgen einen Rückkanal anzubieten (etwa, um Applikationen über abgeschlossene Aufträge zu informieren), kann mit HTTP-Callbacks [Til11] gearbeitet werden. Listing 1 zeigt eine exemplarische REST-Anfrage, mit welcher eine Liste von DNS-Einträgen auf einem Host-Cluster erstellt werden kann; wurden die Aufträge bearbeitet, sendet der Publisher die Ergebnisse an die im `X-TXE-CallbackURI`-Header spezifizierte URI.

Listing 1: Beispiel eines REST-Aufrufs zum Erstellen eines neuen Auftrags

```
POST /tasks HTTP/1.1
Host: publisher.local
Authentication: ...
X-TXE-CallbackURI: https://client.local/finished/dns
X-TXE-HMAC: ...

[{
  "prodecure": "dns_create_records",
  "clusters": ["dns.external"],
  "hosts": ["dns01.internal", "dns02.internal"],
  "data": {
    "records": [
      {"type": "A", "host": "example.com", value: "1.2.3.4"},
      {"type": "AAAA", "host": "example.com", value="::abcd"}
    ]
  }
}]
```

4.2 AMQP-Broker

Die Zustellung der Aufträge an die Host-Systeme erfolgt durch eine nachrichtenorientierte Middleware [DEF+08]. Die bekanntesten Open-Source-Implementierungen des Standard-Protokolls AMQP [OAS14] sind Apache QPID [Apa14] und RabbitMQ [Piv14b]. Abbildung 3 zeigt eine exemplarische Konfiguration eines AMQP-Brokers, um Aufgaben an einzelne Hosts zu verteilen. Die Knoten N_2 und N_3 sind zudem in einem Cluster zusammengefasst und Nachrichten mit dem Routing Key *task.c1* werden im Round-Robin-Verfahren an jeweils einen Knoten des Clusters zugestellt.

Diese Konfiguration ist protokollgemäß jedoch weder statisch noch zentral vorgegeben, sondern wird zur Laufzeit von den einzelnen Clients konfiguriert [KD14]. Auf diese Weise ist es möglich, dass Agenten gegenüber dem Broker selbst spezifizieren, welche Nachrichten benötigt werden. Zudem können so auf einfache Weise Gruppen von Hosts verwaltet werden. Je nach Konfiguration ist es möglich, dass jeweils ein Host einer Gruppe eine Nachricht erhält oder alle Hosts jede Nachricht erhalten.

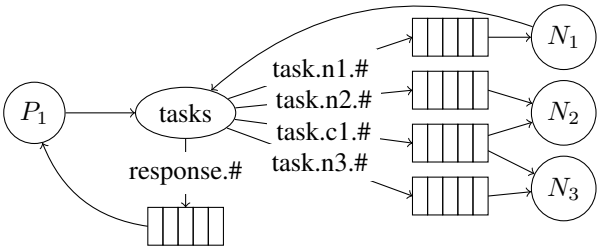


Abbildung 3: Eine exemplarische AMQP-Architektur mit einem Publisher und drei Knoten, die Tasks bearbeiten.

4.3 Subscriber-Komponente

Auf den zu verwaltenden Endsystemen arbeitet ein Software-Agent, dem vom AMQP-Broker neue Aufträge zugesandt werden. Hierzu wird eine Vorlage für ein auszuführendes Skript benötigt, die der Agent bei Bedarf aus dem Code-Repository nachlädt [ZGW+13]. Die Zustellung dieser Vorlagen kann ebenfalls über AMQP erfolgen (Abbildung 6). Auf diese Weise kann auch erreicht werden, dass ein Agent, nachdem einmalig eine bestimmte Skript-Vorlage angefordert wurde, automatisch über Änderungen an dieser Vorlage benachrichtigt wird. Abbildung 6 sowie Listings 2 und 3 zeigen die entsprechende AMQP-Konfiguration sowie exemplarische Anfragen und Antworten zur Anforderung von Vorlagen.

Zudem müssen die Software-Agenten auf den Zielsystemen in der Lage sein, komplexe Ketten voneinander abhängiger Aufträge host-übergreifend zu bearbeiten. Abbildung 4 zeigt eine solche Kette von Aufträgen, die sich über fünf verschiedene Hosts erstreckt: Zum Anlegen einer neuen Webhosting-Instanz mit vorinstallierter Anwendungssoftware müssen neue DNS-Records erstellt (t_D), Datenbank-Benutzer (t_U) und Datenbanken (t_B) angelegt, eine neue virtuelle Maschine erstellt und provisioniert (t_V), ein virtueller Host konfiguriert (t_H), Softwarepakete und deren Abhängigkeiten installiert (t_A) und zeitgesteuerte Dienste eingerichtet (t_C) werden. In diesem Beispiel kommt hinzu, dass einer dieser Hosts erst während dieser Auftragskette erstellt wird; dementsprechend muss der Agent in der Lage sein, sich selbst auf weiteren Host-Systemen zu installieren.

Für die Ausführung solcher Auftragsketten wären mehrere Lösungen denkbar. Im einfachsten Falle könnten die Aufgaben vom Publisher topologisch sortiert und sequentiell abgearbeitet werden. Für eine Host-übergreifende Lösung könnte der Publisher sicherstellen, dass die Aufträge erst dann zugestellt werden, wenn die zuvor benötigten Aufträge bearbeitet wurden. Eine solche sequentielle Bearbeitung von Auftragsketten wäre jedoch potentiell ineffizient, da (etwa wie in Abbildung 4) durchaus Teilaufgaben parallel bearbeitet werden können. Für eine effiziente Ausführung muss der Publisher also sicherstellen, dass nach Bearbeitung eines Auftrages stets alle weiteren abhängigen Aufträge (sofern keine weiteren Abhängigkeiten existieren) an die Zielsysteme zugestellt werden.

4.4 Code-Repository

Das Code-Repository verwaltet die Vorlagen für auf den Hosts auszuführende Skripte. Da das System eine große Anzahl heterogen konfigurierter Hosts verwalten muss, ist es sinnvoll, für jedes Skript Zielsystem-spezifische Facetten eines Skripts zu verwalten. So muss beispielsweise zur Installation eines Webservers auf einem Debian Linux-Host anders vorgegangen werden als auf einem CentOS-Host (Abbildung 5). Um eine Versionierung der Vorlagen sicherzustellen, werden diese im Hintergrund mithilfe eines SCM-Systems, wie etwa Git [Git14], verwaltet.

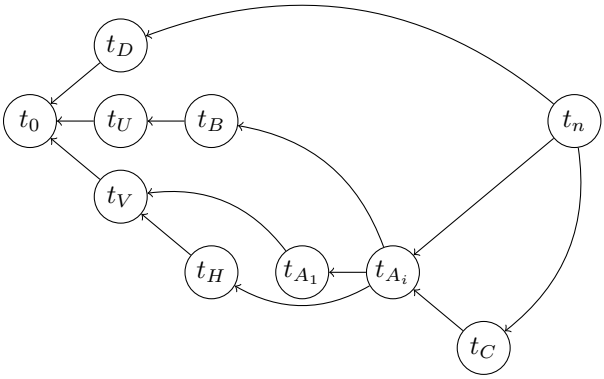


Abbildung 4: Beispiel einer komplexen Kette auf Aufgaben, die sich über fünf verschiedene Hosts verteilt (von denen einer dynamisch während dessen erstellt wird).

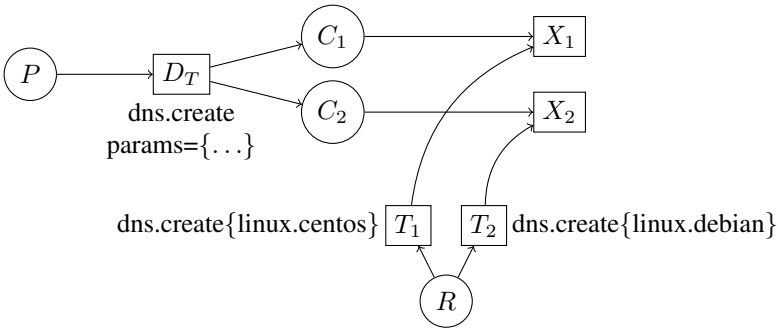


Abbildung 5: Ablauf, nach welchem die Software-Agenten C ausführbare Skripte X auf Grundlage der vom Publisher übermittelten Eingabeparameter D_T und einer von der Konfiguration des Host-Systems abhängigen Vorlage T erstellen.

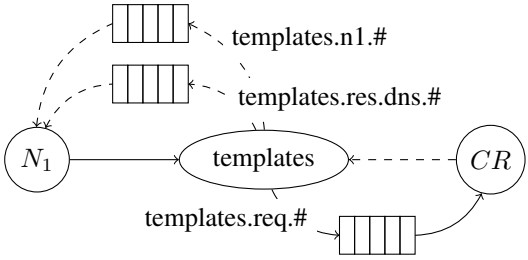


Abbildung 6: Nutzung von AMQP zum Zustellen von Skript-Vorlagen an den Software-Agenten.

Listing 2: Beispiel-AMQP-Nachricht eines Agenten zur Anforderungen einer bestimmten Vorlage

```
Routing-Key:
  templates.req.dns.create

{
  "template": "dns.create",
  "target_queue": "templates.
      node100",
  "host_name": "n1",
  "host_properties": {
    "os": "linux",
    "dist": "rhel",
    "version": "6.4",
    "service": "bind"
  }
}
```

Listing 3: Antwort auf die Anfrage nach einem bestimmten Template (Listing 2).

```
Routing-Key:
  templates.n1

{
  "template": "dns.create.linux.
      rhel.any.bind",
  "language": "bash",
  "content": "..."
}
```

4.5 Skalierungsszenarien

Da die Systemarchitektur auf einen AMQP-Broker als zentrale Komponente angewiesen ist, stellt sich hier die Frage nach der Skalierbarkeit des Systems. Diese Frage wurde auch in [MSPP10] behandelt. Hier werden verschiedene Möglichkeiten untersucht, föderierte AMQP-Broker mit Apache QPID zu realisieren. Auch RabbitMQ bietet mit *federation* und *clustering* sogar mehrere Mechanismen zur Verteilung und Skalierung von Brokern [Piv14a].

Alle übrigen Systemkomponenten sind aufgrund der Publish-Subscribe-Architektur von AMQP automatisch skalierbar. So können theoretisch beliebig viele Publisher Aufträge annehmen und über den Broker veröffentlichen und ebenso können beliebig viele Code-Repositories neue Skript-Vorlagen veröffentlichen. Eine Synchronisation mehrerer Code-Repositories untereinander kann über die eingebauten Synchronisierungsmechanismen von Git erfolgen.

5 Zusammenfassung und Ausblick

Abbildung 7 zeigt abschließend noch einmal die Kommunikationsbeziehungen der einzelnen Komponenten untereinander beim Bearbeiten eines einzelnen Auftrags durch das System.

Im Unterschied zu Lösungen wie [Che14], [Pup14b] oder [CFE14a] versucht die in diesem Artikel vorgestellte Architektur nicht, das Problem des Konfigurationsmanagements ganzheitlich zu lösen. Stattdessen konzentriert sich der hier vorgestellte Vorschlag bewusst

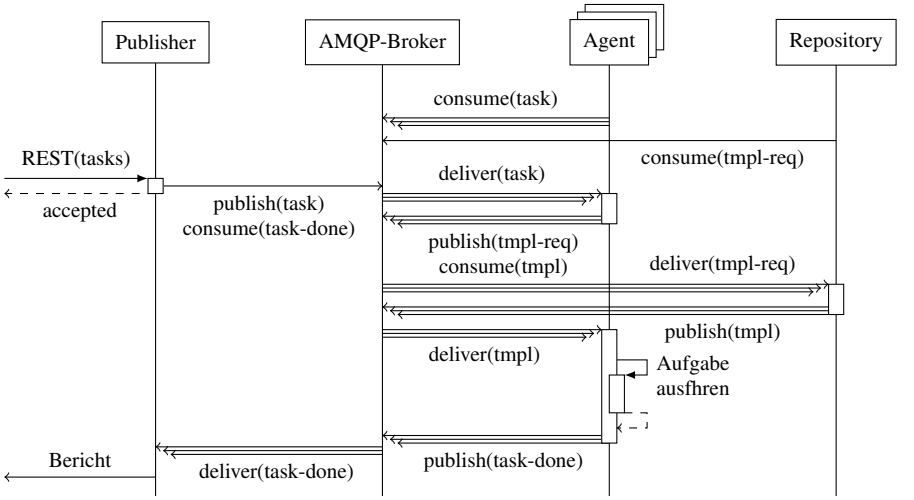


Abbildung 7: Ein Sequenzdiagramm, welches die Ausführung eines einzelnen Auftrags auf drei Zielsystemen verdeutlicht. Die Nachrichten von und zum AMQP-Broker entsprechen denen in [OAS14] definierten Methoden des AMP-Protokolls.

darauf, das Teilproblem der effizienten Ausführung von Skripten auf einer Vielzahl von heterogen konfigurierten Hosts (10.000 und mehr) unter Einhaltung strenger Zeitbedingungen von wenigen Minuten zu lösen.

Das vorgestellte System bietet eine zentrale Schnittstelle an, um beliebige Aufträge auf entfernten Systemen durchzuführen. Um ein ganzheitliches Konfigurationsmanagement zu erreichen, werden weitere Systemkomponenten benötigt, die den erwünschten Zustand des Systems modellieren und Änderungen an diesem Zustand über entsprechende Schnittstellen publizieren.

Im Unterschied zu vergleichbaren Lösungen abstrahiert die hier vorgestellte Architektur die Kommunikation mit den einzelnen Hosts durch den Einsatz einer nachrichtenorientierten Middleware. Weiterhin wird auf den Endsystemen auszuführender Code zentral vorgehalten, und es können komplexe, hostübergreifende Auftragsketten orchestriert werden. Die zugrunde liegende AMQP-Middleware ermöglicht es, alle Komponenten redundant und somit hochverfügbar und skalierbar zu halten.

Literatur

[Ans14] Ansible, Inc. Ansible is Simple IT Automation. http://www.ansible.com/home, 2014.

[Apa14] Apache Software Foundation. Apache Qpid: Messaging built on AMQP. http://qpid.apache.org/, 2014.

[BC06] Mark Burgess und Alva L. Couch. Modeling Next Generation Configuration Management Tools. In *LISA*, Seiten 131–147, 2006.

[Bur95] Mark Burgess. Cfengine: a site configuration engine. In *USENIX Computing systems*, Jgg. 8, Seiten 309–337, 1995.

[CFE14a] CFEngine AS. CFEngine: Configuration management software. http://cfengine.com/index, 2014.

[CFE14b] CFEngine AS. CFEngine: Scale and Scalability. https://cfengine.com/archive/manuals/reader, 2014.

[Che14] Chef Software, Inc. Chef: IT automation for speed and awesomness. http://www.getchef.com/, 2014.

[CHIK03] Alva L. Couch, John Hart, Elizabeth G. Idhaw und Dominic Kallas. Seeking Closure in an Open World: A Behavioral Agent Approach to Configuration Management. *Proceedings of the 17th Large Installations Systems Administration (LISA) conference (San Diego, CA, USA, 10/2003), Usenix Association*, 3:125–148, 2003.

[DEF+08] Jürgen Dunkel, Andreas Eberhart, Stefan Fischer, Carsten Kleiner und Arne Koschel. *Systemarchitekturen für verteilte Anwendungen: Client-Server, Multi-Tier, SOA, Event Driven Architecture, P2P, Grid, Web 2.0*. Carl Hanser-Verlag, München, 1. Auflage, September 2008.

[DJV10] Thomas Delaet, Wouter Joosen und Bart Vanbrabant. A survey of system configuration tools. *Proceedings of the 24th Large Installations Systems Administration (LISA) conference (San Jose, CA, USA, 11/2010), Usenix Association*, 2010.

[Fie00] Roy Thomas Fielding. *Architectural styles and the design of network-based software architectures*. Dissertation, University of California, Irvine, 2000.

[FLRU13] J.L. Fernandes, I.C. Lopes, J.J.P.C. Rodrigues und S. Ullah. Performance evaluation of RESTful web services and AMQP protocol. In *2013 Fifth International Conference on Ubiquitous and Future Networks (ICUFN)*, Seiten 810–815, Juli 2013.

[GHM+07] Martin Gudgin, Marc Hadley, Noah Mendelsohn, Jean-Jacques Moreau, Henrik Frystyk Nielsen, Anish Karmarkar und Yves Lafon. SOAP Version 1.2 Part 1: Messaging Framework, April 2007.

[Git14] Git-SCM. Git. http://git-scm.com/, 2014.

[Hos13] Ben Hosmer. Getting Started with Salt Stack: The Other Configuration Management System Built with Python. *Linux Journal*, Januar 2013.

[KD14] Michael Klishin und Chris Duncan. The AMQP 0-9-1 Model Explained, 2014.

[Mit14] Mittwald CM Service GmbH & Co. KG. Reseller Webhosting für Agenturen & Freelancer. https://www.mittwald.de/, 2014.

[MSPP10] Gregory Marsh, Ajay P Sampat, Sreeram Potluri und Dhabaleswar K Panda. Scaling Advanced Message Queuing Protocol (AMQP) Architecture with Broker Federation and InfiniBand. Bericht, 2010.

[OAS12] OASIS. OASIS Advanced Message Queuing Protocol (AMQP) Version 1.0. http://docs.oasis-open.org/amqp/core/v1.0/os/amqp-core-overview-v1.0-os.html#toc, Oktober 2012. OASIS-Standard.

[OAS14] OASIS. AMQP Architecture. https://www.amqp.org/product/architecture, 2014.

[Piv14a] Pivotal Software, Inc. RabbitMQ: Distributed RabbitMQ brokers. http://www.rabbitmq.com/distributed.html, 2014.

[Piv14b] Pivotal Software, Inc. RabbitMQ: Messaging that just works, 2014.

[Pup14a] Puppet Labs, Inc. MCollective: Programmatic Execution of Systems Administrations Actions. http://puppetlabs.com/mcollective, 2014.

[Pup14b] Puppet Labs, Inc. Puppet Enterprise: IT Automation Software for System Administrators. http://puppetlabs.com/puppet/puppet-enterprise, 2014.

[Sal14] SaltStack, Inc. Salt Documentation, Release 2014.1.1. https://media.readthedocs.org/pdf/salt/latest/salt.pdf, März 2014.

[Til11] Stefan Tilkov. *REST und HTTP: Einsatz der Architektur des Web für Integrationsszenarien*. dpunkt.verlag, Heidelberg, 2. Auflage, 2011.

[TvS07] Andrew S. Tanenbaum und Maarten van Steen. *Verteilte Systeme: Prinzipien und Paradigmen*. Pearson, 2. Auflage, November 2007.

[Vin06] S. Vinoski. Advanced Message Queuing Protocol. *IEEE Internet Computing*, 10(6):87–89, November 2006.

[VJ13] Bart Vanbrabant und Wouter Joosen. A framework for integrated configuration management tools. In *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*, Seiten 534–540. IEEE, 2013.

[YZF+11] Jiayuan Yue, Zhuo Zhang, Junhong Fu, Shengqing Lu, Xinmin Li und Yunfu Shen. Extensible architecture for high-throughput task processing based on hybrid cloud infrastructure. In *2011 International Conference on Electronics, Communications and Control (ICECC)*, Seiten 1452–1455, September 2011.

[ZGW+13] Sai Zeng, Shang Guo, Fred Wu, Constantin Adam, Long Wang, Cashchakanithara Venugopal, Rajeev Puri und Ramesh Palakodeti. Universal Script Wrapper – An innovative solution to manage endpoints in large and heterogeneous environment. In *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium on*, Seiten 916–919. IEEE, 2013.

Erfahrungsbericht über den Einsatz des open source Kommunikationsservers Mirth Connect am Universitätsklinikum Bonn

P. Köppen, S. Langenberg
uk-it Köln Bonn
Universitätsklinikum Bonn
Sigmund-Freud-Str. 25
53105 Bonn
Philipp.Koeppen@ukb.uni-bonn.de
Stefan.Langenberg@ukb.uni-bonn.de

Abstract: Kommunikationsserver spielen im Krankenhaus zur Verknüpfung heterogener medizinischer Anwendungssysteme eine zentrale Rolle. Seit 2013 wird am Universitätsklinikum Bonn der open source Kommunikationsserver Mirth Connect eingeführt. Mit zusätzlichen, nicht einer open source Lizenz unterliegenden Erweiterungen ist das System für den Produktiveinsatz im Krankenhaus geeignet.

1 Einleitung

Derzeit gibt es kein Informationssystem das alle Funktionen, die in einem Krankenhaus benötigt werden, bereitstellen kann. [KG01] In vielen Einrichtungen existieren neben einem eher administrativ ausgerichteten Patientenmanagementsystem und klinischen Dokumentationssystem spezialisierte Subsysteme, z.B. für das Labor oder die Radiologie. Diese müssen über Schnittstellen miteinander kommunizieren, um effektive anwendungsübergreifende Arbeitsprozesse zu ermöglichen. Am UKB wurde dazu ein Prozess orientiertes Integrationskonzept in einer Nachrichten basierten Architektur gewählt [CHKT06]. Hierzu hat sich der Kommunikationsstandard HL7 im Krankenhaus etabliert [HBD99].

Mittels HL7 lassen sich zwar grundsätzlich Kommunikationsbeziehungen zwischen den einzelnen Anwendungen direkt realisieren, wie dies beispielsweise am Universitätsklinikum Freiburg im Sinne einer HL7-basierten serviceorientierten Architektur realisiert wurde. Um jedoch die Komplexität der Kommunikationsbeziehungen zu reduzieren, kommen Kommunikationsserver zum Einsatz [Bra10]. Sie haben folgende Funktion: „Ein Kommunikationsserver überwacht alle Verbindungen zwischen den Applikationen in einer Organisation. Eine Verbindung wird zwischen jedem System und dem Kommunikationsserver hergestellt und die Nachrichten zwischen diesen bidirektionalen Verbindungen können übersetzt, gefiltert, archiviert, modifiziert und zu ihrem endgültigen Ziel weitergeleitet werden.“ [Spr07]. Treffender ist daher die englische Bezeichnung *integration engine*.

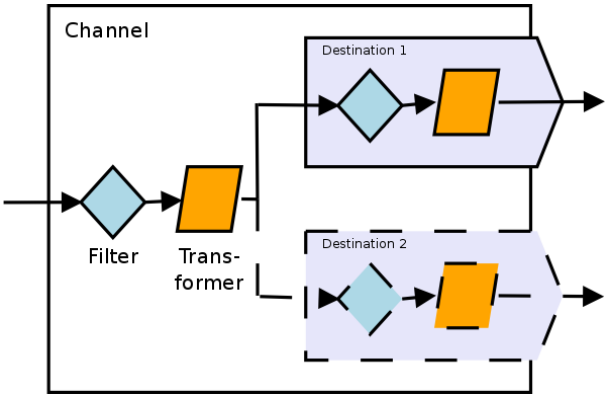


Abbildung 1: Channel mit Quellkonnektor, allen möglichen Filtern und Transformatoren sowie zwei Zielkonnektoren

Der Kommunikationsserver Mirth Connect wurde von der Mirth Corporation, die von Quality Systems, Inc. im Jahre 2013 übernommen wurde entwickelt [QSI]. Der Java-Quellcode wurde unter der Mozilla Public License (MPL) 1.1 veröffentlicht.

Kommunikationsbeziehungen werden mit Mirth Connect mittels sog. Channel verwirklicht [HW03]. Mindestens einen Channel benötigt man, um Daten von A nach B zu transportieren. Jeder Channel läuft als Thread innerhalb eines Mirth Connect Prozesses. Jeder Channel hat einen Eingang (Quellkonnektor), über den Daten von einem externen System entgegengenommen werden können. Weiterhin kann jeder Channel auch Daten von einem oder mehrerer anderer Channel entgegennehmen, wodurch sich auch komplexere Kommunikationsbeziehungen umsetzen lassen. Für die verarbeiteten Daten gibt es einen oder mehrere Zielkonnektoren (destinations).

Jeweils auf der Eingangs- wie Ausgangsseite befinden sich Filter und Transformatoren zur internen Verarbeitung der Nachrichten, siehe Abb. 1. Intern werden die Nachrichten nach XML konvertiert. Empfangene und gesendete Nachrichten werden für einen einstellbaren Zeitraum in einer Datenbank gespeichert, wobei die Datenbanken Apache Derby, MySQL, PostgreSQL und MS SQL unterstützt werden. Mittels des in Java geschriebenen JavaScript Interpreters Rhino [Rhi] sind eigene Programmentwicklungen innerhalb von Mirth Connect ausführbar. Die Administration des Systems erfolgt über eine per Java Web Start aufrufbare graphische Benutzerschnittstelle, siehe Abb 2.

2 Ergebnisse und Erfahrungen

Seit Mitte 2013 verfügt das Universitätsklinikum Bonn über eine Produktions- und Entwicklungsumgebung einer Mirth Connect 2.0.5 Installation mit von der Firma OSM-GmbH entwickelten Erweiterungen (Mirth Connect powered by OSM), diese Erweite-

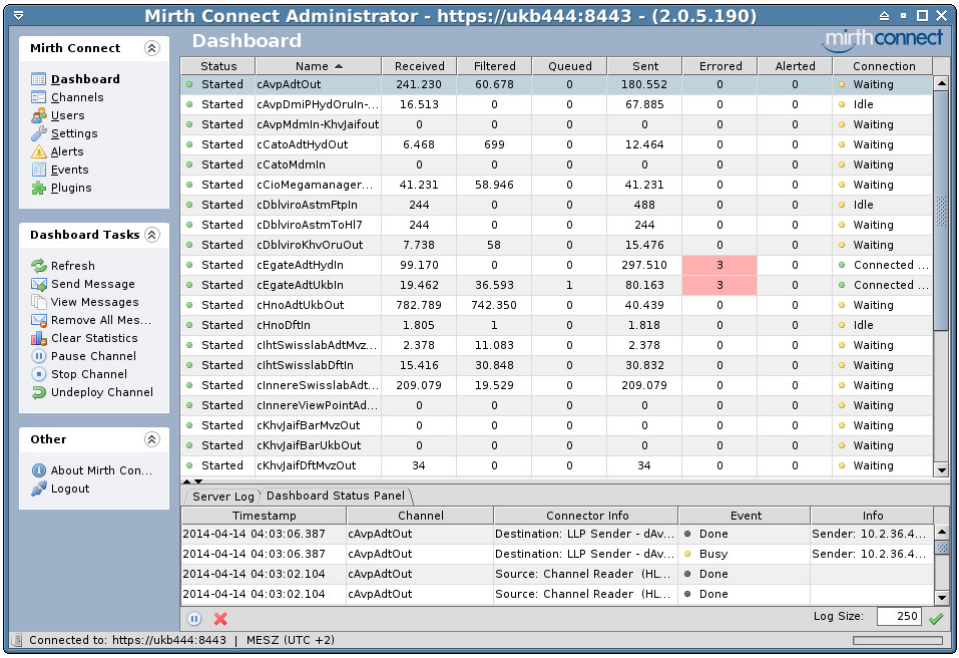


Abbildung 2: Mirth Connect Dashboard mit Ansicht aller Channel

rungen unterliegenden jedoch nicht einer open source Lizenz. Darüber hinaus verwenden mindestens 4 klinische Subsysteme Mirth Connect als Schnittstelle zu den zentralen Kommunikationsservern. Bisher wurde der Kommunikationsserver eGate 5-SRE verwendet. Da die Weiterentwicklung dieses Produktes seitens der Firma Oracle allerdings ungewiss ist, haben wir und auf Anraten der OSM-GmbH für den Umstieg auf Mirth Connect entschieden. Da der Support für eGate 5-SRE noch bis 2017 gesichert ist, besteht für den Umstieg kein Zeitdruck. Zunächst implementieren wir alle neue Schnittstelle unter Mirth Connect, nebenher stellen wir sukzessive die bestehenden Schnittstellen auf Mirth Connect um.

Die zentralen Mirth Connect Installationen werden unter Windows 2008 64bit mit MySQL als Datenbank betrieben.

Ein systematischer Vergleich von Mirth Connect mit dem bisher verwendeten eGate 5-SRE sowie dem Cloverleaf Kommunikationsserver ist in Tab. 1 dargestellt.

2.1 Protokollunterstützung

Mirth Connect verwendet XML als internes Datenformat zur Nachrichtenspeicherung. Andere Formate wie CSV, HL7, DICOM werden in den Quellkonnektoren nach XML gewan-

Tabelle 1: Vergleich von Mirth Connect mit zwei kommerziellen Kommunikationsservern mit hohem Marktanteil im Krankenhaus

	Mirth Connect	**eGate 5-SRE**	**Cloverleaf 5.8**
Konnektoren	Dateisystem, CIFS, (S)FTP, TCP/IP-Socket, HTTP, Mail, Web-Service (SOAP), JMS, DICOM, Datenbank, selbstentwickelte JavaScript Konnektoren	Dateisystem, FTP, TCP/IP-Socket, weitere als Zusatzmodul (z.B. SAP-BAPI, SAP-HCM)	Dateisystem, FTP, SFTP als Zusatzmodul, TCP/IP-Socket, HTTP, Mail, DICOM, selbstentwickelte Java oder Tcl Konnektoren
Nachrichtenprotokolle	XML, HL7 2.x, HL7 3.x, CSV, UN/EDIFACT, SAP-HCM (OSM Erweiterung)	XML, HL7 2.x, CSV, UN/EDIFACT, Frei definierbare Formate, SAP-HCM (OSM Erweiterung)	XML, HL7 2.x, HL7 3.x, HRL (Zusammenstellung aus VRL/FRL) , VRL (Variable Formate), FRL (Fixe Formate)
Nachrichtentransformation	JavaScript, XSLT	Monk (Scheme), XSLT, Java	Tcl, Java
Transaktionssicherheit	bedingt vorhanden	vorhanden	vorhanden
Alerting	z.T. vorhanden, sonst über internes JavaScript	über externe shell Skripte	Networkmonitor, Alertkonfigurator
Nachrichtenarchivierung	Datenbank	Dateisystem	Dateisystem
Dokumentation	Online Zugang nur bei Supportvertrag mit QSI	vollständig	vollständig

delt. Die XML-Strukturen werden für jede Nachricht einzeln dynamisch generiert, so dass auch nicht Standard konforme Nachrichten verarbeitet werden können.

Das für das Patientenmanagement verwendete SAP Modul IS-H verwendet den SAP eigenen Nachrichtenstandard HCM [SAP99]. Mit einer Programmerweiterung von OSM kann auch dieses verarbeitet werden.

Für von Mirth Connect nicht primär unterstützte Nachrichtenformate muss man sich mittels JavaScript einen Parser schreiben, um das Format nach XML zu konvertieren. So konnte mit Mirth Connect auch eine verbesserte Integration eines Virologie Laborsystems in einem ASTM-ähnlichen Befundformat realisiert werden.

Auf der Transportebene unterstützt Mirth Connect TCP/IP MLLP (meist verwendet im Zusammenhang mit HL7), dateibasierte Protokolle (File, SMB, FTP, SCP). Das Umbenennen von Dateien nach der Übermittlung wird von Mirth Connect erst ab Version 3 im Standard unterstützt. Mit diesem Mechanismus möchte man verhindern, dass eine Datei bereits vom Zielsystem verarbeitet wird, während sie noch geschrieben wird. Semaphore Dateien werden lediglich in der OSM-Erweiterung unterstützt.

Für den Import und Export werden weiterhin Web Services (SOAP, WSDL) sowie HTTP unterstützt.

Der direkte Zugriff auf Datenbanken ist ebenfalls möglich. So wurde die Übermittlung von Behandlungsdokumenten aus dem zentralen Dokumentenmanagementsystem (DMS) an ein Einweiserportal entwickelt. Mirth Connect fragt regelmäßig neue Eingänge von Dokumenten ab und übermittelt diese im HL7 Standard an ein Einweiserportal. Die Einbindung von Datenbanken in Kommunikationsprozesse ist mit Mirth sehr anwenderfreundlich möglich, da die notwendigen Datenbanktreiber bereits in die Software integriert sind. Eine SQL Abfrage, die in einem JavaScript eingebunden ist kann auch Abfrageparameter enthalten, die erst zur Laufzeit ermittelt werden.

```
// Get timestamp of last SQL execution
var strFile = "LastTimestamp.seq";
var contents = FileUtil.read(strFile);
var strLetzterAbfrageZeitpunkt = contents; // z.B.: "01032014100000";
// Get Data from Oracle SQL Database
var dbConn = DatabaseConnectionFactory.createDatabaseConnection( ... );
//Create a SQL Statement to be executed
var sqlStr =
"select ... FROM ... " +
"where ..." +
"BPD.DOKUMENTENDATUM >= TO_TIMESTAMP ('" + strLetzterAbfrageZeitpunkt
 + "', 'DDMMYYYYHH24MISS')";
// Execute SQL Statement and save them as results
var result = dbConn.executeCachedQuery(sqlStr);
dbConn.close();
return result;
```

Neben der Echtzeitübermittlung von HL7 Nachrichten, wird Mirth Connect auch zur Extraktion und Bereitstellung von Datenpools genutzt, die für Anwender aus den administrativen Bereichen (z.B. Medizincontrolling) als Basis für Auswertungen verwendet werden. Insbesondere SQL Abfragen mit längeren Antwortzeiten, werden mit Mirth Connect zu einem festgelegten Zeitpunkt in der Nacht ausgeführt und die gewonnenen Daten per FTP oder SMB auf Abteilungslaufwerke abgelegt. Somit werden die Abfragezeiten außerhalb der Arbeitszeit verlegt.

2.2 Nachrichtenfilterung und Übersetzung

Einfache Filter und Transformationen lassen sich mit der graphischen Benutzeroberfläche erstellen. Besser und flexibler ist dies jedoch mit dem integrierten JavaScript möglich, Beispiel siehe Abb. 3.

Eine MDM-T01 Nachricht mit base64-kodiertem eingebettetem Dokumenteninhalt lässt sich beispielsweise aus übermittelnden Dokumenten (PDF, TIFF, JPEG) in einem Transformationsprozess mittels JavaScript erzeugen.

```
var uncpath = "\\server\share\" + file;
var contents = FileUtil.readBytes(uncpath); // (PDF-)Datei einlesen per UNC-Pfad:

// Dokument base64 kodieren.
var encData  = FileUtil.encode(contents);

//remove the newlines to produce valid Base64
var index = encData.indexOf("\r\n");
    while(index != -1){
        encData = encData.replace("\r\n","");
        index = encData.indexOf("\r\n");
    }

// Base64 kodiertes (PDF-)Dokument in OBX.5.1 Feld schreiben.
tmp['OBX']['OBX.5']['OBX.5.1'] = encData;
```

Hl7-Nachricht:

```
MSH|^~\&|ORBIS_ADT|MVZ|MEDOS|RAD|201403241134||ADT^A34|1408963|P|2.3|||AL|NE|D||DE
EVN|A34|201403051010|201403051010||UKB17525ADM||
PID|1|6643112|06643112||Test^MVZ^^^^||19901010|M||
MRG|06655915|
```

Interne XML-Darstellung der HL7-Nachricht (Ausschnitt):

```
<HL7Message>
    <MSH>
        ...
        <MSH.4>
            <MSH.4.1>MVZ</MSH.4.1>
        </MSH.4>
        <MSH.5>
            <MSH.5.1>MEDOS</MSH.5.1>
        </MSH.5>
        <MSH.6>
            <MSH.6.1>RAD</MSH.6.1>
        </MSH.6>
        <MSH.7>
            <MSH.7.1>201403241134</MSH.7.1>
        </MSH.7>
        <MSH.8/>
        <MSH.9>
            <MSH.9.1>ADT</MSH.9.1>
            <MSH.9.2>A34</MSH.9.2>
        </MSH.9>
...
```

Zugriff auf Nachrichtenelemente mit JavaScript am Beispiel eines Filters:

```
var valid_Events = new
RegExp("A01|A02|A03|A04|A05|A06|A07|A08|A11|A12|A13");
var valid_Ambulance = new RegExp("M1551|M1552");
var valid_Institute = new RegExp("M15IS");

var process = false;

if (msg['EVN']['EVN.1']['EVN.1.1'].toString() == "A34") process = true;
else if (valid_Events.test(msg['EVN']['EVN.1']['EVN.1.1'].toString()))
{
if (valid_Ambulance.test(msg['PV1']['PV1.3']['PV1.3.4'].toString())
||

valid_Institute.test(msg['PV1']['PV1.3']['PV1.3.5'].toString()))  process =
true
}

return process;
```

Abbildung 3: Beispiel einer einfachen HL7-Nachricht und deren interne XML-Darstellung, sowie den Zugriff auf Feldinhalte in einer JavaScript Filterregel

2.3 Transaktionssicherheit

Verbesserungen bei der Transaktionssicherheit gibt es mit der Version 3. Bei Version 2 war diese lediglich für TCP/IP-Socketkommunikation gegeben. Wenn auf ein Zielsystem mit einem dateibasierten Transportprotokoll wie beispielsweise FTP eine Nachricht übertragen wird und das Zielsystem nicht verfügbar ist, so läuft die Verarbeitung der Nachricht auf einen Fehler. Seit Version 3, sowie in der speziellen OSM-Version, die wir verwenden, wird die Nachricht in einer Queue gepuffert und nach Wiederverfügbarkeit des Zielsystems automatisch nachkommuniziert.

Weitere Probleme bei der Transaktionssicherheit treten bei der Übertragung sehr großer Dateien auf. Bei erschöpftem Java heap space wird die Quelldatei zwar gelöscht, die Zieldatei aufgrund des Fehlers aber nicht geschrieben.

2.4 Monitoring und Alarme

Alle Fehlermeldungen werden in einer Log-Datei protokolliert. Die Zuordnung zu einem bestimmten Channel ist dabei meist schwierig. Pro Channel werden die letzten 200 (einstellbar) Log-Einträge gespeichert und über die graphische Benutzeroberfläche angezeigt. Solange eine Nachricht noch nicht nach der Verarbeitung aus der Datenbank gelöscht wurde, lässt sich über die Benutzeroberfläche der Erfolg der Verarbeitung kontrollieren.

Bei Fehlerzuständen lassen sich Email-Alerts konfigurieren. Bei dem o.g. Beispiel des nicht verfügbaren Zielsystems sind allerdings solche Alerts wenig hilfreich, wird doch bei jedem Versuch einer fehlerhaften Nachrichtenübermittlung eine Email verschickt. An dieser Stellen kann man allerdings sich mit JavaScript behelfen: Innerhalb eines Channels kann per JavaScript die Anzahl der Nachrichten in der Queue ermittelt werden, die nicht zu einem Zielsystem kommuniziert werden konnten. Erst wenn diese Anzahl einen bestimmten Schwellwert überschreitet, wird eine Fehlermeldung versandt. Nach Behebung des Fehlers beim Zielsystem kann dann nach Unterschreitung des Schwellwerts per Email der Alarm wieder aufgehoben werden. Ein solcher Mechanismus ist allerdings nur anwendbar, wenn für jedes Zielsystem ein eigener Channel zuständig ist, siehe Abb 4.

2.5 Archivierung

Für einen gewissen Zeitraum lassen sich die bereits kommunizierten Nachrichten in der Datenbank von Mirth Connect aufbewahren, z.B. für 7 Tage. Für die Langzeitarchivierung ist dieses Verfahren aufgrund des benötigten Speicherbedarfs nicht geeignet. Man kann das Problem wie folgt lösen: Die Verarbeitung, die eigentlich über einen Channel abgewickelt werden könnte, wird in einen Quell- und Zielchannel aufgespalten. Wir senden daher aus jedem Quellchannel über einen Archivkonnektor alle Nachrichten eines Tages in eine Archivdatei. Bei Bedarf lassen sich einzelne Archivdateien oder Teile davon über einen Resend-Channel an den gewünschten Zielchannel nachkommunizieren, siehe Abb. 4. Ein

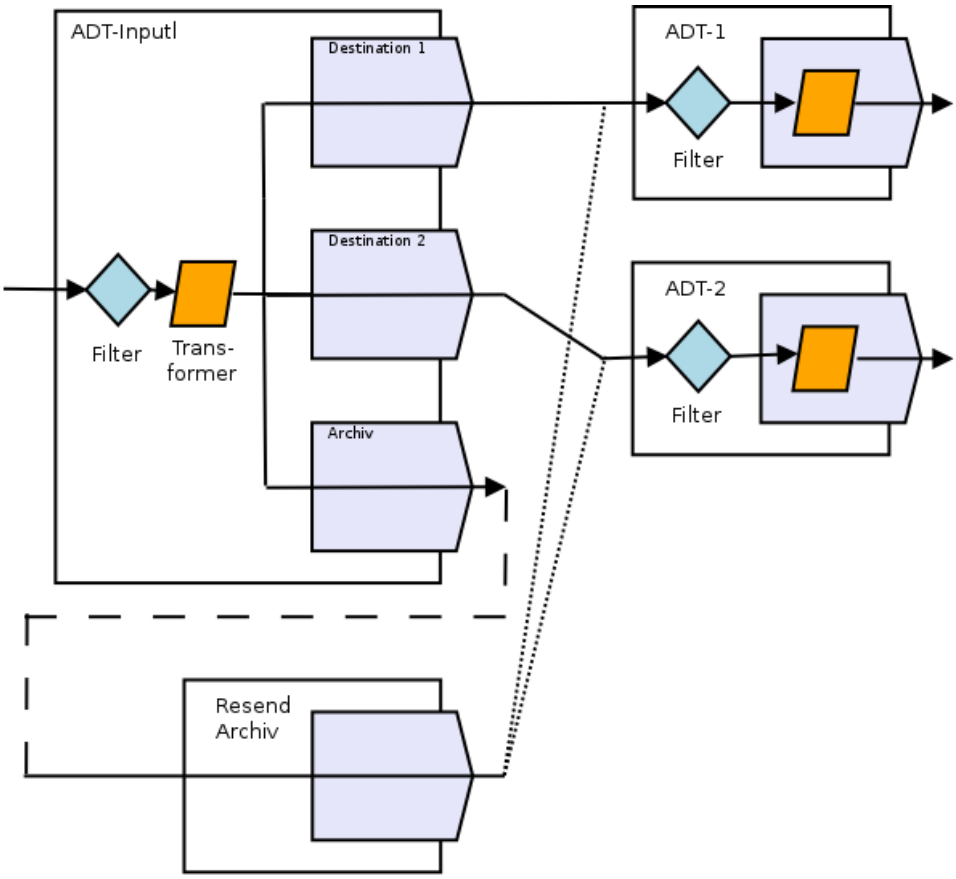


Abbildung 4: Nachrichtenverarbeitung mit mehreren Channeln am Beispiel des Transfers von Patientenstamm- und Bewegungsdaten von einem zentralen Patientenmanagementsystem an mehrere Subsysteme (ADT-1 und ADT-2). Im Quellchannel werden die Nachrichten zusätzlich über einen Archivkonnektor tageweise gebündelt als Archivdatei abgelegt. Bei Bedarf können diese über einen Resend-Channel an den gewünschten Zielchannel nachkommuniziert werden. Dabei durchlaufen sie in dem Zielsystemchannel die gleichen Filter- und Transformationsschritte wie die ursprüngliche Nachricht.

vergleichbarer Mechanismus für jeden Channel wird mit Version 3 im Standard bereitgestellt.

2.6 Dokumentation

Das grobe Fundament für den fachlichen Einstieg in die Mirth Connect Umgebung wurde durch die Basisschulung der OSM-GmbH gelegt. Wesentlich für den Aufbau von Wissen zu Schnittstellenentwicklung ist jedoch die sogenannte Community-Plattform des Herstellers in den USA. Es besteht die Möglichkeit sich kostenfrei an der Plattform anzumelden und somit Einsicht in zahlreiche Anfragen und Problemlösungen zu erhalten. Der Umfang der Forenbeiträge ist sehr umfangreich und zu den meisten Fragestellungen konnte eine Lösung gefunden werden, die ggf. leicht angepasst werden musste. Die Beträge sind durchweg in englischer Sprache verfasst. Je nach Problem gestaltet sich die Suche nach einem passenden Beitrag mit Lösungsansatz recht langwierig. Mirth Connect verwendet intern die Skriptsprache JavaScript, deren Programmierkonzepte und Syntax sehr gut dokumentiert ist. Daher konnte die Umsetzung von Anforderungen zu Formatanpassungen bei Schnittstellenprojekten in Programmcode schnell umgesetzt werden. Mirth Connect bietet die Möglichkeit die konfigurierte Schnittstellenlogik in Gesamtheit oder einzeln in einer XML Konfigurationsdatei abzuspeichern. Diese Datei kann von einem anderen Mirth Connect-System importiert werden und sofort ist die gleiche Schnittstelle implementiert. Dies erleichtert den Austausch mit anderen Nutzern erheblich: Realisierte Schnittstellenlösungen können als konkrete Implementierungsvorlage schnell und bequem bereitgestellt werden; offene Implementierungsprobleme werden durch Konfigurationsentwürfe konkretisiert. Dieser Nutzen ist jedoch stark abhängig von der Partizipationsbereitschaft und dem allgemeinen Wissenstand innerhalb der Nutzergemeinde.

3 Fazit

Mirth Connect v2 mit den von OSM bereitgestellten Erweiterungen kann unseren bisherigen Kommunikationsserver eGate 5 weitestgehend ersetzten. Mit Version 3 sind viele dieser Erweiterungen bereits im Standard enthalten.

Mit Mirth Connect lassen sich Schnittstellen mit deutlich geringerem Aufwand als mit eGate realisieren. Erste produktiv verwendete Schnittstellen konnten von eGate auf Mirth Connect umgestellt werden.

Durch den open source Charakter der Software konnten Lizenzkosten für die Beschaffung eines neuen Kommunikationsservers eingespart werden. Durch den kostenpflichtigen Support eines Drittanbieters ist das Betriebsrisiko nicht höher als bei der Verwendung eines kommerziellen Produktes. Die rasche Verbreitung, die Mirth Connect zwischenzeitlich international gefunden hat, lässt eine dauerhafte und dynamischen Weiterentwicklung erwarten.

Um den Austausch zwischen den Mirth Connect Benutzern im deutschsprachigen Raum

zu verbessern, haben wir eine deutschsprachige Internetplattform zu Mirth Connect ins Leben gerufen [Lan].

Literatur

[Bra10] E.M. Brauer. Haben Kommunikationsserver eine Zukunft? - Künftige Anforderungen und Kriterien. *Krankenhaus-IT Journal*, 4:53–54, 2010.

[CHKT06] S. Conrad, W. Hasselbring, A. Koschel und R. Tritsch. *Enterprise Application Integration*. Elsevier, 1. Auflage, 2006.

[HBD99] K.U. Heitmann, B. Blobel und J. Dudeck. *HL7-Kommunikationsstandard in der Medizin*. Mönch Verlag, 1. Auflage, 1999.

[HW03] G. Hohpe und B. Woolf. *Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions*. Addison-Wesley, 1. Auflage, 2003.

[KG01] K.A. Kuhn und D.A. Giuse. From Hospital Information Systems to Health Information Systems. *Method. Inform. Med.*, 4:275–287, 2001.

[Lan] Mirth Connect – Der universelle open source Kommunikationsserver, http://mirthconnect.wordpress.com/.

[QSI] Offizielle Seite von Quality Systems, Inc. (QSI) zu Mirth Connect: http://www.mirthcorp.com/products/mirth-connect.

[Rhi] Rhino, https://developer.mozilla.org/en/docs/Rhino.

[SAP99] SAP AG. *IS-HCM Guide for External System Partners*, 1999.

[Spr07] R. Spronk. Die Rolle eines Kommunikationsservers: Ein Überblick für das Management, http://www.ringholm.de/docs/00100_de.htm. Whitepaper, 2007.

Erfolg innovativer Open-Source-Ansätze am Beispiel asynchroner Servertechnologien

Michael Hitz, Thomas Kessel
DHBW-Stuttgart
Paulinenstraße 50
70178 Stuttgart

Abstract: Ist Open Source innovativ oder doch nur eine ”copy cat”? Viele Projekte im Open Source (OS) Umfeld haben zum Ziel, für bestehende Ansätze oder offene Standards eine Lösung ”nachzubauen” und damit eine in der Anschaffung kostengünstige Alternative zu kommerziellen Produkten und durch Beteiligung der Community gereifte Lösung anzubieten. Open-Source-Projekte können aber durchaus auch neue, innovative Ansätze hervorbringen und damit zum Innovationstreiber werden. Im Rahmen dieses Papiers betrachten wir die Konzepte des Projekts **Vert.x** im Bereich asynchroner Servertechnologien genauer und untersuchen aus technischer Sicht an einer konkreten Problemstellung, ob die dort eingeführten Ansätze für Alltagsprobleme im Unternehmenskontext nutzbringend eingesetzt und damit als Innovationserfolg gewertet werden können.

1 Einleitung

Das Open-Source-Konzept [OSI14] hat sich in den meisten Anwendungsbereichen etabliert und bietet Lösungen an, die kommerziellen Produkten ebenbürtig sind. Das aktuelle Portfolio geht von Betriebssystemen (z.B. Linux, Android), Anwendungssoftware (z.B. Typo3, Activiti, mySQL), bis hin zu Werkzeugen (z.B. Eclipse, Mantis) und Frameworks für die Softwareentwicklung (z.B. Spring, Hibernate). Der aktuelle Stand der wissenschaftlichen Forschung zu Open Source ist in [CWHW12] dokumentiert.

Viele Projekte haben dabei zum Ziel, bekannte und etablierte Technologien, Spezifikationen oder Standards als Open-Source-Lösung zu implementieren (z.B. JBoss für JEE, VLC media player für Multimedia-Player, Libre Office für Bürosoftware) und damit eine kostengünstige Alternative zu kommerziellen Produkten anzubieten. In einigen Fällen (z.B. Open Office) haben die Lösungen sogar ihren Ursprung im kommerziellen Umfeld und wurden zur weiteren Entwicklung der Open-Source-Community übergeben. Im Gegensatz zu kommerzieller Software ist die eingebundene Entwickler-Community für die nachhaltige Entwicklung der Open-Source-Software entscheidend [BGH+12].

Insbesondere Open-Source-Projekte im Bereich der Softwareentwicklung besitzen häufig aber auch innovativen Charakter. Die Innovation reicht dabei von Verbesserungen beste-

hender Ansätze bis hin zu gänzlich neuen Konzepten, die in kommerziellen Produkten noch nicht verfügbar sind. Ein prominentes Beispiel sind die Entwicklungen im Bereich der NoSQL-Datenbanken, in welchem verschiedenste Ansätze in den letzten Jahren entstanden [BBLW13].

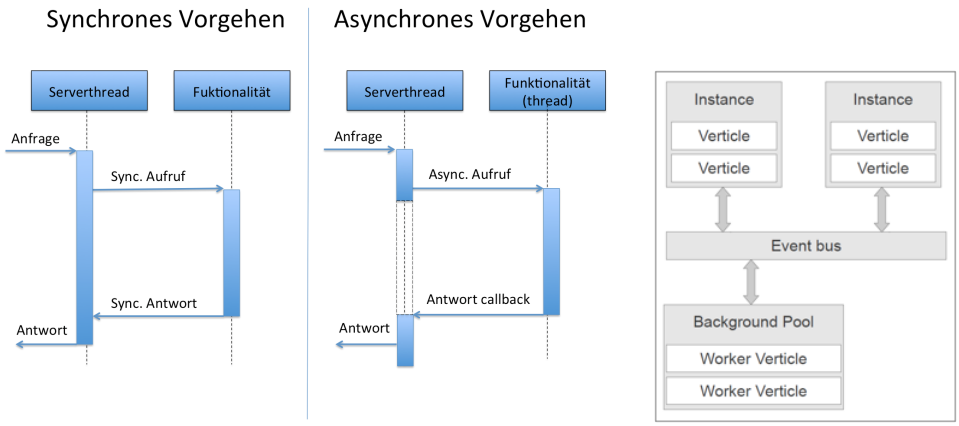
Während etablierte Open-Source-Produkte häufig durch Unternehmen oder Dachorganisationen wie Apache oder die Eclipse Foundation betreut werden und damit auf eine breite Community zählen können, beginnen innovative Projekte meist als kleine Projekte auf Open-Source-Portalen, z.B. http://sourceforge.net/ oder http://www.javaforge.com.

Ein innovativer Ansatz, der seinen Ursprung im Open-Source-Umfeld hat, sind asynchrone Frameworks zur Anwendungserstellung für webbasierte Anwendungen und Servertechnologien, die ebenfalls auf dem synchronen, event-getriebenen Vorgehen aufbauen. Die bekanntesten Produkte, die mittlerweile einen Einzug in die Praxis gefunden haben ([MHB14], [Wol13],[Gro13]), sind *Node.js* [Nod14] und *Vert.x* [Ver14b]. Der asynchrone Ansatz zur Abarbeitung von Anfragen verspricht eine optimalere Ausnutzung von Serverressourcen und kann so durch eine optimierte Nutzung der Serverressourcen zu Kosteneinsparungen beim Betrieb von Webanwendungen führen.

Im Rahmen des Projekts "Kompetenzzentrum Open Source" (KOS) an der DHBW Stuttgart wird für konkrete Open-Source-Produkte untersucht, inwiefern sie einen Nutzen im Unternehmensumfeld erbringen können. In diesem Kontext wurde **Vert.x** näher analysiert ([BBSW14]). Die Implementierung Vert.x wurde ursprünglich von Entwicklern der Firma VMWare initiiert und im August 2013 an die Eclipse Foundation übergeben. Obwohl der Verbreitungsgrad des Produkts noch nicht so hoch wie der von Node.js ist, verspricht es aber durch seine größere Sprachunabhängigkeit einen flexibleren Einsatz als Node.js, das auf JavaScript beschränkt ist.

In diesem Papier wird Vert.x als Vertreter einer innovativen Technologie technisch genauer betrachtet, für die es – nach unserem Kenntnisstand - keine vergleichbaren kommerziellen Produkte gibt. Es werden Optimierungsmöglichkeiten aufgezeigt, die mit dem neuen Ansatz erreicht werden können. Insbesondere steht die in der Praxis häufig auftretende Problemstellung einer zeitweiligen Blockierung bzw. hohen Auslastung von Webservern durch langlaufende Operationen im Vordergrund - die Lösung dieser Problemstellung stellt einen hohen praktischen Nutzen dar.

Der Einsatz innovativer Open-Source-Software in einem Unternehmen birgt eine Reihe von Risiken [HK13] - insbesondere in Hinsicht auf den Investitionsschutz. Die Kosten für die Einführung eines Open-Source-Produkts mit hohem Innovationsgrad erscheinen auf den ersten Blick gering, jedoch können sich die Folgekosten bei der Integration in die eigene Produktentwicklung als hoch erweisen, wenn man auf auf "das falsche Pferd" setzt und die Entwicklung des Open-Source-Produkts eingestellt wird. Mittels Open Source Maturity Models versucht man solche Risiken einzuschätzen. Fokus des vorliegenden Papiers sind jedoch die technischen Aspekte. Eine Betrachtung der Reife und Nachhaltigkeit kann anhand [GHJ12], [GKPP12] vorgenommen werden.



(a) Synchrones vs. asynchrones Vorgehen

(b) Kernelemente von Vert.x

Abbildung 1: Synchron vs. asynchron und Vert.x Kernelemente

2 Asynchrone Servertechnologien

Synchron vs. Asynchron

Die gängigsten, heute eingesetzten Web-/Applicationserver arbeiten nach einem synchronen Prinzip: wenn ein Request an den Server eingeht, wird dieser an einen Thread gebunden, der für die Abarbeitung zuständig ist. Während der gesamten Lebensdauer der Anfrage, ist dieser Thread somit belegt. Sobald die Anfrage bearbeitet ist, wird der Thread wieder freigegeben und kann weitere Anfragen bearbeiten.

Asynchrone Ansätze haben zum Ziel, diese Kopplung von Request und Thread, also deren Bindung bis zur vollständigen Abarbeitung auch aufwändiger Requests, aufzulösen. Dies wird durch eine ereignisgesteuerte Programmierung erreicht: es existiert ein Hauptthread, der Anfragen entgegennimmt. Operationen, die ausgeführt werden sollen, werden asynchron bearbeitet: das bedeutet, ihre Ausführung wird in eigenen Threads angestoßen, es wird aber nicht im Hauptthread auf ein Ergebnis gewartet. Es wird lediglich eine *callback*-Methode angegeben, die bei Beendigung der Operation gerufen wird. Dieses Ereignis kann dann die Weiterbearbeitung steuern (z.B. die Aufbereitung der Antwort an den Rufer). Indem auf diese Weise ein- oder ausgabeintensive Operationen von sog. Eventhandlern in nebenläufigen Threads bearbeitet werden, wird die Eventschleife nicht blockiert. Sie kann also weiterhin ihre Hauptfunktionen erfüllen: die Abfrage neuer Events und Aufruf des passenden Eventhandlers.

Abbildung 1(a) zeigt die unterschiedlichen Herangehensweisen in einem Ablaufdiagramm: während beim synchronen Vorgehen der Thread während der Abarbeitung des Requests ständig aktiv ist und auf die Antwort einer synchron gerufenen Funktionalität wartet (also faktisch blockiert ist), ist der Serverthread im Fall des asynchronen Aufrufs der Funktiona-

lität (und dessen Abarbeitung in einem eigenen Thread) frei für die Abarbeitung anderer Aufgaben und somit nicht blockiert. Erst wenn die gerufene Funktionalität abgearbeitet ist, wird der Serverthread wieder beansprucht.

Vert.x als konkrete Implementierung

Vert.x ist eine Implementierung des obigen asynchronen Konzepts, und basiert auf Java-Technologien. Es wird in einer Java Virtual Machine (JVM) ausgeführt und ein Vert.x Webserver kann damit mehrere Prozessorkerne nutzen, ohne dass der Benutzer sich um die Prozessverwaltung oder die Kommunikation zwischen den Prozessen kümmern muss [Ver14a]. Vert.x bezeichnet sich selbst als *polyglott*. Dies bedeutet, dass Anwendungen in mehreren Sprachen erstellt und auf dem Server ausgeführt werden können, sofern diese auf einer JVM umgesetzt werden können (z. Zt. sind dies Ruby, Groovy, Python, Java, JavaScript, Clojure) [Ver14b]. Diese größere Sprachenunabhängigkeit war entscheidend für die Auswahl von Vert.x für das Papier [BBSW14], auf dem die folgenden Untersuchungen basieren.

Abbildung 1(b) (entnommen aus [BBS+13]) illustriert das Kernkonzept von Vert.x: die Kernelemente von Vert.x sind sog. *Verticles*, die zu Modulen gepackt werden können und welche die Anwendungslogik implementieren. Mehrere Verticles können innerhalb einer *Vert.x-Instanz* ausgeführt werden, die wiederum in einer eigenen JVM-Instanz bzw. einem Prozessor ausgeführt wird. Vert.x ordnet so jedes Verticle genau einer Eventschleife zu und verhindert damit, dass eine Verticle-Instanz zur gleichen Zeit in mehr als einem Thread ausgeführt wird.

Für den Fall, dass ein Verticle ein- oder ausgabeintensiven (d.h. blockierenden) Code enthält, kann es als sogenanntes "Worker Verticle" deklariert werden. Es ist dann nicht der Vert.x-Eventschleife zugeordnet, sondern wird beim Aufruf in einem speziellen *Workerthread* abgearbeitet. Dadurch wird eine Blockade der Eventschleife verhindert [Ver14a].

Anwendungsbereiche asynchroner Ansätze

Die Implementierungen von Node.js und Vert.x basieren darauf, dass es nur *einen* Serverthread gibt, der alle eingehenden Anfragen beantwortet. Dieser erhält die gesamte Rechenleistung des Prozessors. Dies beschränkt zwar auf den ersten Blick die Nutzbarkeit des Ansatzes für Unternehmensanwendungen, die z.T. rechenintensive Operationen ausführen müssen - zumindest, wenn diese im Hauptthread ausgeführt werden müssten.

Aber die Voraussetzung für das asynchrone Vorgehen ist, dass die im Hauptthread gerufenen Operationen asynchron arbeiten können. Hierzu bieten die asynchronen Serverimplementierungen, insbesondere für die - aus der Laufzeitsicht - teuren Ein-/Ausgabeoperationen, bereits asynchrone Implementierungen an. Die Anwendungsbereiche, die für den Einsatz asynchroner Server darum häufig als optimal angegeben werden, beziehen sich meist auf die von den Implementierungen gelieferten Funktionalitäten: Ein-/Ausgabeoperationen, kurzlaufende Requests von Webanwendungen, Entgegennahme und Bereit-

stellung von großen Datenmengen (z.B. für Mediastreaming, Lieferung von Daten bzw. Dateien aus Objekt-Datenbanken) [Cap13], [BBS+13].

Im Unternehmensumfeld geht es allerdings häufig darum, bestehende Funktionalitäten aufzurufen, die nur auf den Backend-IT-Systemen des Unternehmens verfügbar sind, und diese Operationen sind z.T. (sehr) zeitaufwändig und somit langlaufend. In Artikeln und Literatur (z.B. [Rod12], [Cap13]) wird deshalb häufig das asynchrone Konzept als nicht geeignet für solche Anwendungsfälle bezeichnet (z.B. rechenintensive Anwendungen auf dem Server, Operationen auf relationalen Datenbanken). Dies gilt dann, wenn diese Aktionen innerhalb des Serverthreads ausgeführt werden sollen.

Die im vorangegangenen Abschnitt beschriebenen *Worker-Verticles*, die in Vert.x vorgesehen sind, gestatten es jedoch, langlaufende Operationen in externe Threads auszulagern und so diesen scheinbaren Nachteil auszugleichen.

So kann eine konkrete Problemstellung untersucht werden, die in der Praxis in Unternehmensportalen häufig auftritt - die temporäre Blockade von Webservern durch langlaufende Anfragen.

3 Nutzen in der Praxis

Auf Webservern, die Anfragen in großen Unternehmensportalen bearbeiten, treffen zahlreiche Requests zur gleichen Zeit ein. Üblicherweise sind die meisten gerufenen Aktionen dabei kurzlaufend (z.B. die Auslieferung einer angeforderten Seite oder das Absenden eines Mailformulars). Es gibt jedoch auch eine Reihe von Anfragen, bei denen die Antwort auf den Request des Nutzers einige Zeit in Anspruch nimmt. Meist sind hier Zugriffe auf das Backend der IT-Infrastruktur involviert, die eine längere Wartezeit erfordern.

Die aktuell verbreitet eingesetzten Webservertechnologien (z.B. Tomcat, JBoss, Apache Webserver) bearbeiten einen eingehenden Request synchron - in einem eigenen Thread. Hierzu steht ein Thread-Pool zur Verfügung, aus der - bei einer Anfrage - ein Thread für die Abarbeitung entnommen und danach wieder für andere Anfragen freigegeben wird.

Bei langlaufenden Anfragen kann diese Vorgehensweise dazu führen, dass ein Server weitgehend blockiert wird durch die vielen langlaufende Anfragen, denn der Server ist dann vorwiegend damit beschäftigt, auf die Antworten des Backends der IT-Infrastruktur zu warten. In dieser Zeit können somit keine weiteren Anfragen beantworten werden. Dies führt in ungünstigen Last-Situationen dazu, dass der Server alle weiteren Anfragen ablehnt und die Website nicht mehr erreicht werden kann, bis wieder neue Threads freigegeben werden können.

Das betrachtete Szenario sollen Self-Service-Anwendungen sein, wie sie in Versicherungs- oder Banking-Portalen häufig vorkommen. Hier existieren sowohl lang- als auch kurzlau-

fende Operationen (z.B. Informationen zu Produkten, Kontoübersicht, Durchführung einer Überweisung, Adressänderung). Folgende Rahmenbedingungen wurden hierfür aufgrund der Praxiserfahrungen zugrunde gelegt:

- Es existieren viele kurzlaufende Anfragen (z.B. die Anzeige von Produktinformation, die Adressänderung)
- Die Zahl der langlaufenden Anfragen ist gering, aber die Abarbeitungsdauer ist hoch, da zeitaufwändige Backendaufrufe notwendig sind (z.B. Durchführen einer Überweisung und gleichzeitige Vertragsanzeige)
- Langlaufende Operationen dürfen nicht zur Blockade des Servers führen.
- kurze Operationen haben Priorität. Bei Häufung langlaufender Operationen dürfen diese länger dauern. Kurze Operationen sollen aber nicht blockiert werden.

Für diese Konstellation wurde ein Lösungsansatz entworfen, der die asynchronen Features von Vert.x ausnutzt.

Lösungsansatz

Das in unseren Untersuchungen ([BBSW14]) zugrunde gelegte Prinzip nutzt die asynchrone Herangehensweise aus und vermeidet die Probleme, die zu einer Blockierung der Hauptschleife führen können.

Die Operationen wurden als Verticles in Java implementiert und als REST-ful Serviceaufrufe vom Server zur Verfügung gestellt.

Kurzlaufende Operationen wurden auf dem Server als *reguläre Verticles* konfiguriert und somit werden die von Vert.x bereitgestellten Mittel genutzt (z.B. optimierte I/O für statische Inhalte), jedoch nicht in Worker-Verticles ausgelagert. Zeitintensive Operationen werden als *Worker-Verticles* konfiguriert, die bei einem Aufruf in einem gesonderten Thread ausgeführt werden. Sie belasten so nicht den Serverthread, sondern sie werden zur Abarbeitung für einen eigenen Workerthread vorgemerkt, der dem Threadpool entnommen wird.

Diese Anordnung verspricht theoretisch, dass auf diese Weise in der Gesamtsicht keine Blockade des Hauptthreads auftreten kann (oder erst dann, wenn die Last auf den Server extrem hoch wird). Die kurzlaufenden Operationen werden unmittelbar ausgeführt, die Abarbeitung der langlaufenden Operationen wird jedoch nach außen verlagert.

Für die langlaufenden Operationen tritt dennoch dasselbe Problem auf, wie bei der Abarbeitung durch einen synchronen Server mit einer beschränkten Threadpool-Größe: bei Erreichung der Threadpool-Grenze können keine weiteren Anfragen abgearbeitet werden. Sie müssen warten, bis wieder ein Thread frei wird. Der Unterschied zum synchronen Verhalten liegt aber darin, dass der Server weiterhin kurzlaufende Anfragen entgegennehmen kann - aus Nutzersicht – ist er also nicht blockiert und somit weiterhin verfügbar.

Dies ist aus Sicht des Betriebs von webbasierten Anwendungen ein großer Vorteil, der ohne diesen Ansatz nicht erreicht werden könnte.

Erfolgsmessung: Vergleich des asynchronen und synchronen Ansatzes

In diesem Abschnitt werden die Ergebnisse der Untersuchungen aus [BBSW14] zusammengefasst. Eine detaillierte Betrachtung findet sich dort.

Zur Messung des Verhaltens von Webservern wurden für die Untersuchung zwei vergleichbare Infrastrukturen aufgebaut. Die synchrone Infrastruktur bestand aus einem Apache Tomcat-Server, welcher die Funktionalität in Form von REST-ful Services zur Verfügung stellte. Die asynchrone Implementierung wurde mit Vert.x umgesetzt, welche dieselbe Schnittstelle anbot. Um eine Vergleichbarkeit hinsichtlich der Threadzahl zu erreichen, wurde dem Apache Tomcat-Server eine feste Threadzahl zugewiesen. Dieselbe Anzahl an Threads wurde für den Threadpool für die langlaufenden Worker-Verticles verwendet.

Die Untersuchungen wurde auf einer relativ schwachen Hardwareausstattung durchgeführt; dies entspricht zwar nicht der Realität von Rechenzentren eines Unternehmens, sie zeigt aber deutlich die Skalierungseffekte, da der Sättigungsbereich des Servers bei Anfragen schneller erreicht wird. Um den Versuchsaufbau möglichst einfach zu halten, wurde aus diesem Grund zudem den Serverinstanzen nur jeweils ein einziger Prozessorkern zugewiesen. Da sowohl Tomcat als auch Vert.x mehrere Prozessorkerne unterstützen können, wird hiermit keiner Technologie ein potentieller Vorteil gegeben.

Die Beispielfunktionen wurden aus dem Bankenbereich gewählt. Abbildung 3 illustriert die gewählten UseCases. Für die Aktionen wurden feste Laufzeiten gewählt und so ein Satz von kurz- bzw. langlaufenden Aktionen bestimmt. Es wurde zudem ein Test-Client (s. Abbildung 2) entwickelt, mit dem die Anzahl der Requests und der Anteil der jeweiligen Operationen bestimmt und auf den Server geschickt werden können. So können bei den Simulationen Änderungen hinsichtlich des Verhältnisses von kurz- zu langlaufenden Operationen durchgeführt werden.

Die Messungen wurden für verschiedene Konfigurationen der Vert.x-Verticles vorgenommen (Verteilung von Verticles auf die Workerthreads) um die Effekte nachweisen zu können. Wir beschränken uns hier nur auf einen Teil der Szenarien aus [BBSW14].

Gemessen wurde die Durchlaufzeit der einzelnen Requestarten und damit das Antwortverhalten des Gesamtsystems für die einzelnen UseCases.

- **Konfiguration 1:** Alle Requestarten werden von Worker-Verticles verarbeitet. Die Zuteilung der Threads erfolgt nach Eingang der Anfragen.
- **Konfiguration 2:** Alle Requestarten werden von Worker-Verticles bearbeitet. Die Zuteilung der Threads erfolgt jedoch explizit. Hierbei wird jedem Verticle eine Anzahl von Threads aus dem Pool zur Verfügung gestellt. Langlaufende Operationen erhalten eine höhere Zahl von Threads. Die Tabelle in Abbildung 4 zeigt eine solche

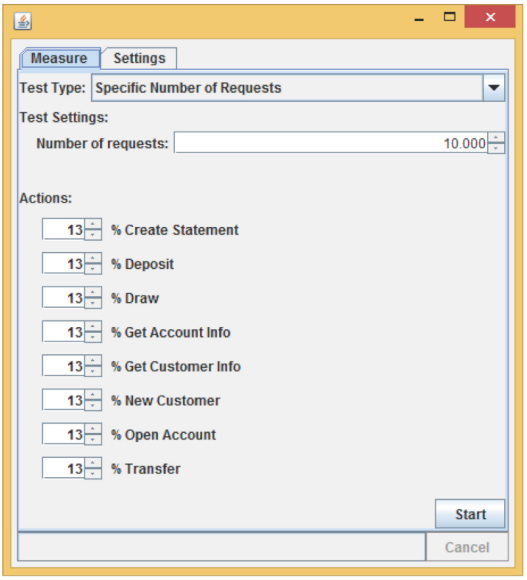


Abbildung 2: Benchmarktool zur Simulation von Anfragen

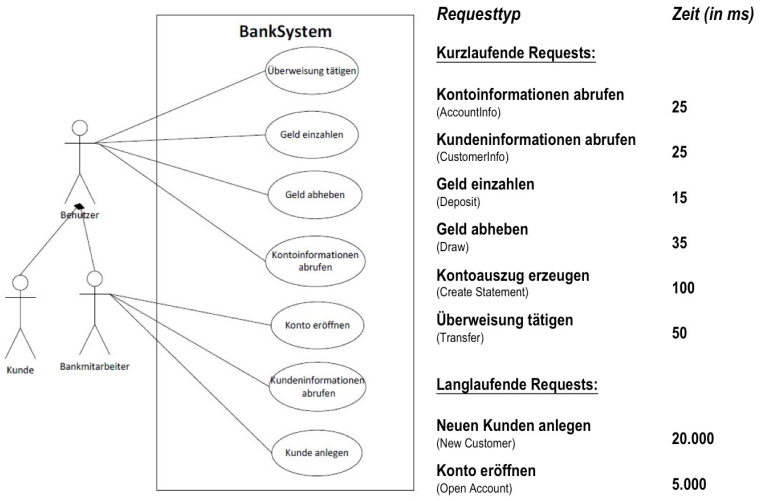


Abbildung 3: Gewählte UseCases und Laufzeiten in der Benchmarkanwendung

Zuordnung beispielhaft. Hier wird aufgrund der Abarbeitungszeit und des prozentualen Anteils an der Summe der Aufrufe eine Gewichtung vergeben, die zu einer Zuordnung von Threads für die Requestart führt.

Berechnung der Zuweisung der Workerthreads							
Requestart	**Ausarbeitungszeit**	**Anteile**	**"Gewicht"**	**Anteil am Gesamtgewicht**	**1 Thread je kurzem Typ**	**Aufteilung des Rests auf die langen Typen**	**Verteilung**
DepositWorker	15	15	225	0,03	1		**1**
AccountWorker	25	15	375	0,06	1		**1**
CustomerWorker	25	15	375	0,06	1		**1**
DrawWorker	35	15	525	0,08	1		**1**
TransferWorker	50	15	750	0,12	1		**1**
StatementWorker	100	15	1500	0,23	1		**1**
OpenAccountWorker	5000	5	25000	3,88	14	2,8	**3**
NewCustomerWorker	20000	5	100000	15,53		11,2	**11**
			128750				

Abbildung 4: Explizite Zuordnung von Workerthreads zu Transaktionen

- **Konfiguration 3:** Abarbeitung einiger kurzlaufenden Requestarten im Serverthread. Die verbleibenden Requests sind analog zu Szenario 2. Hierbei soll jedoch der Effekt untersucht werden, wenn der Hauptthread Aufgaben übernimmt, ohne dass hierzu eine Threadverwaltung eingeschaltet werden muss. Die könnte den Threadpool für sehr kurze Operationen entlasten.

Die Konfiguration 1 entspricht im Grundsatz der synchronen Vorgehensweise. Wie erwartet wurden hier keine Vorteile bei der Abarbeitung erzielt, da das Erreichen der Threadpoolgröße zum Warten auf einen freiwerdenden Thread für alle Requests führt. Allerdings begann der synchrone Server frühzeitig mit der Ablehnung von Anfragen und blockierte damit.

Die Ergebnisse der Konfiguration 2 zeigten signifikante Veränderungen in den Antwortzeiten, im Vergleich zu den Ergebnissen des synchronen Servers. Abbildung 5 stellt die Differenz der Antwortzeiten der asynchronen und synchronen Lösung für unterschiedliche Requestarten dar. Die *Deposit*-Operation ist explizit als ein Repräsentant für eine kurzlaufende Operation aufgeführt. Bei einer steigenden Anzahl von Anfragen verbesserte sich die Antwortzeit für kurzlaufenden Anfragen signifikant. Die langlaufenden Anfragen hingegen wurden langsamer abgearbeitet. Dieser Effekt kann gedämpft werden, wenn den langlaufenden Operationen anteilig mehr Threads zugewiesen werden. Auch hier begann der synchrone Server mit der (frühzeitigen) Ablehnung von Anfragen.

Die Konfiguration 3 brachte auf den ersten Blick sehr verblüffende Ergebnisse (vgl. Laufzeitdifferenzen in Abbildung 6). In dieser Konfiguration wurde lediglich die *Deposit* Anfrage als normales Verticle festgelegt und damit nicht mehr durch einen Workerthread abgearbeitet. Die sich ergebenden Verbesserungen waren signifikant - insbesondere auch für die langlaufenden Operationen. Die *Deposit*-Operation wurde erheblich schneller (da das Threadmanagement entfiel). Da diese Operation 15% der Anfragen ausmachte (vgl. Abbildung 4), wurde demnach in gleichem Maße Pool-Kapazität frei - was der Abarbeitung der Langläufer zugute kam.

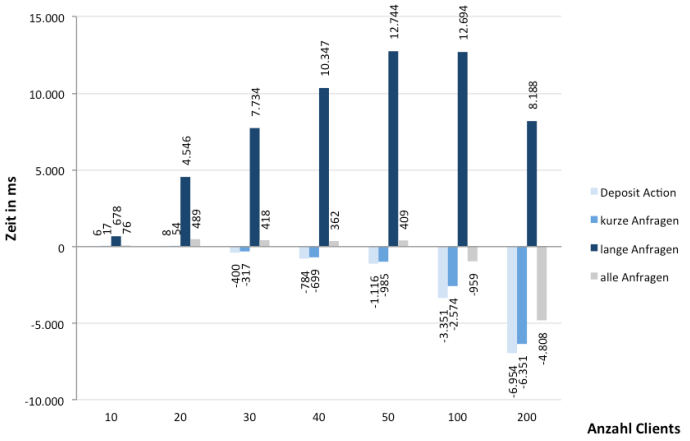


Abbildung 5: Vergleich Konfiguration 2 mit synchroner Verarbeitung

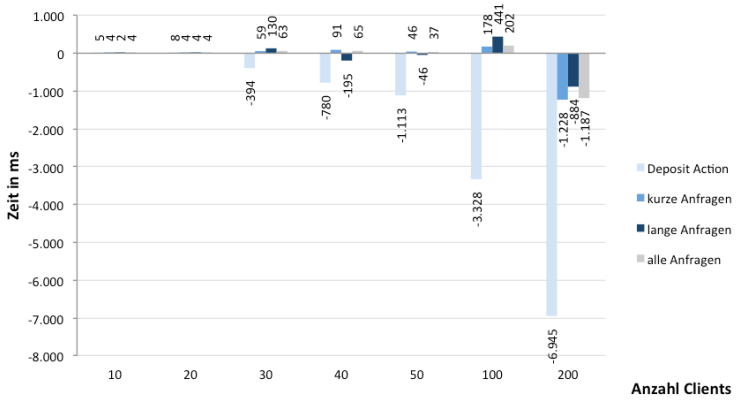


Abbildung 6: Vergleich Konfiguration 3 mit synchroner Verarbeitung

Die Ergebnisse unserer Untersuchungen geben Hinweise auf eine deutliche Verbesserung der Serverauslastung bei durchdachtem Einsatz von Vert.x. Durch die Konfiguration und die Verteilung von Aufgaben (in Form von Verticles) auf dedizierte Worker konnten schon Verbesserungen erreicht werden. Die deutlichste Verbesserung entstand aber in der Konfiguration, die kurzlaufende Operationen nicht dem Threadpool zuwies, da hier die Aufwände für die Threadverwaltung entfallen. Unsere formulierten Rahmenbedingungen konnten erfüllt werden und die Vermutung, dass sich durch den Einsatz eines asynchronen Vorgehens Vorteile ergeben, hat sich in der Versuchsreihe bestätigt. Um eine optimale Lösung für einen konkreten Anwendungsfall zu erreichen, bedarf es allerdings einer de-

taillierten Analyse der Gegebenheiten in einer konkreten Anwendung. Die Entscheidung, welche Aktivitäten in einen Threadpool ausgelagert werden sollen und wie viel Last der Serverthread verarbeiten kann, muss im Einzelfall ermittelt und eine entsprechende Konfiguration gefunden werden.

4 Zusammenfassung und Bewertung

Die Ergebnisse unserer technischen Untersuchungen geben einen deutlichen Hinweis, dass die im Open-Source-Umfeld entstandenen asynchronen Ansätze für die Umsetzung eines Servers eine Lösungsalternative zu den bestehenden synchronen Ansätzen im Webumfeld bieten. Auch wenn die Vorteile nur nach einer detaillierten Analyse des Anwendungsfalls gehoben werden können, so ermöglicht der Einsatz eines asynchronen Vorgehens eine erhebliche Verbesserung der Servernutzung und kann die Servicequalität eines Portals aus Nutzersicht erhöhen - der Ansatz stellt damit einen hohen Nutzen dar.

Im Rahmen dieses Papiers wurde lediglich auf den Nutzen des asynchronen *Ansatzes* aus technischer Sicht eingegangen und keine Bewertung des *Projekts Vert.x* hinsichtlich seiner Beständigkeit vorgenommen. Um die Einsetzbarkeit des konkreten Produktes im Unternehmensumfeld zu bewerten, kann eine Nutzwertanalyse des Projekts und ein Vergleich der Ergebnisse mit anderen Projekten angewendet werden ([GHJ12], [GKPP12]).

Die asynchronen Konzepte haben ihr Fundament im Open-Source-Umfeld. Wenn auch teilweise von kommerziellen Unternehmen initiiert, wurden die wichtigen Schritte als Open Source Projekt durchgeführt - was sicherlich auch zur Verbreitung und damit Stabilisierung der Ansätze geführt hat. Der Nutzen des Ansatzes ist im Unternehmenskontext hoch - und kann damit als echter Innovationserfolg der Open-Source-Gemeinde gewertet werden.

Literatur

[BBLW13] Teresa Bogolowski, Tana Brunner, Sophie Lingelbach und Christin Wattler. NoSQL Datenbanken - Typisierung, Seminararbeit DHBW-Stuttgart, 2013. In *Projektergebnisse des Kompetenzzentrums Open Source*, 2013.

[BBS+13] Can Paul Bineytioglu, Joe André Boden, Rocco Schulz, Max Vökler und Robert Wawrzyniak. Evaluation of Asynchronous Server Technologies, Seminararbeit DHBW-Stuttgart, 2013. In *Projektergebnisse des Kompetenzzentrums Open Source*, 2013.

[BBSW14] Christian Brummer, Michael Busam, Thomas Scherer und Matthias Welz. Asynchrone Servertechnologien zur optimalen Auslastung von Webservern, Seminararbeit DHBW-

Stuttgart, 2014. In *Projektergebnisse des Kompetenzzentrums Open Source*. Duale Hochschule Baden-Württemberg, Stuttgart, 2014.

[BGH+12] Robert Bruchhardt, Anne Golembowska, Maximilian Heinemeyer, Felix Kugler, Stephen Said und Jennifer Zohar. Beständigkeit eines Open Source Projektes - Analyse von Anerkennungssystemen in Open Source Projekten, Semnararbeit DHBW-Stuttgart, 2012. In *Projektergebnisse des Kompetenzzentrums Open Source*, 2012.

[Cap13] Tomislav Capan. Why The Hell Would I Use Node.js? A Case-by-Case Introduction - http://www.toptal.com/nodejs/why-the-hell-would-i-use-node-js, 2013.

[CWHW12] Kevin Crowston, Kangning Wei, James Howison und Andrea Wiggins. Free/Libre open-source software development. *ACM Computing Surveys*, 44(2):1–35, Februar 2012.

[GHJ12] Franziska Gorhan, Juliana Hettinger und Maren Wolter Juliane Schulz. Development of a Model Evaluating the Maturity of Open Source Software, Seminararbeit DHBW-Stuttgart, 2012. In *Projektergebnisse des Kompetenzzentrums Open Source*, 2012.

[GKPP12] Anne Golembowska, Madeline Klink, Jana Petrovic und Daniel Prescher. Entwicklung eines Modells zur Bewertung von Open Source Produkten hinsichtlich eines produktiven Einsatzes, Seminararbeit DHBW-Stuttgart, 2012. In *Projektergebnisse des Kompetenzzentrums Open Source*, 2012.

[Gro13] Christian Gross. Node.js - hinter dem Hype. Sollte man sich als Java-Entwickler mit Node.js beschäftigen? *JavaMagazin*, 13(4), 2013.

[HK13] Michael Hitz und Thomas Kessel. Einfluss der Nutzung und Auswahl von Open Source Software beim Entwurf einer Multikanal-Architektur. In *GI-Jahrestagung 2013*, Seiten 1324–1338, 2013.

[MHB14] Sascha Möllering, Mariam Hakobyan und Björn Stahl. Vert.x im Unternehmenseinsatz - Entwicklung und Betrieb von asynchronen Applikationen mit Vert.x in der Praxis. *JavaMagazin*, 14(4), 2014.

[Nod14] Nodejs. Node.js Website - http://nodejs.org, 2014.

[OSI14] OSI (Open Source Initiative). The Open Source Definition – http://www.opensource.org/osd, 2014.

[Rod12] Golo Roden. *Node.js & Co - Skalierbare, hochperformante und echtzeitfähige Webanwendungen professionell in JavaScript entwickeln*. dpunkt.verlag, 2012.

[Ver14a] Vert.x. Vert.x Manual - http://vertx.io/manual.html, 2014.

[Ver14b] Vert.x. Vert.x Website - http://vertx.io, 2014.

[Wol13] Eberhard Wolff. vert.x – alles wird alles: Polyglott – asynchron – modular. *JavaMagazin*, 13(4), 2013.

IT-Governance und Strategisches Informationsmanagement

(ITG-SIM)

IT-Governance und Strategisches Informationsmanagement (ITG-SIM)[1]

Daniel F. Abawi, Matthias Goeken, André Miede, Konrad Walser

1 Vorwort und Einleitung

Ansätze und Strategien zur IT-Governance sind für viele Unternehmen heutzutage nahezu unverzichtbar geworden. Dazu trägt die Notwendigkeit bei, die IT des Unternehmens an den aktuellen Trends der Gestaltung und Entwicklung komplexer IT-Systeme auszurichten und dabei auch neu aufkommende Technologien und Anwendung sowie möglicherweise unter Berücksichtigung von Branchenbesonderheiten die IT auf die Unternehmensziele hin auszurichten.

In diesem Umfeld führt insbesondere die Planung, Steuerung und Kontrolle der Systemlandschaft und deren Zusammenspiel mit der fachlichen Seite zu neuen Herausforderungen, welche durch ein ganzheitliches IT-Management und ein strategisches Informationsmanagement abgedeckt werden müssen. Darüber hinaus verlangen auch externe Faktoren wie die wachsende Anzahl staatlicher Regulatorien (z.B. MiFID, Basel II & III, Sarbanes-Oxley Act, Legal Unbundling in der Energieversorgung) die Einführung einer effektiven und effizienten IT-Governance, die die Einhaltung von gesetzlichen, technischen und internen Regelwerken (auch als Compliance bezeichnet) gewährleistet.

Eine Hauptherausforderung an die IT-Governance stellt die Notwendigkeit einer Steuerung der gesamten Systemlandschaft dar. Parallel hierzu entstehen neue Anforderungen an die Unternehmensstruktur bzw. -organisation, Rollen und Verantwortlichkeiten, Softwarelebenszyklen oder -standards, die allesamt wichtige Einflussfaktoren darstellen.

Um diesen Herausforderungen ganzheitlich begegnen zu können, ist die Entwicklung zeitgemäßer Governance-Ansätze und Methoden sowie unterstützender Applikationen notwendig.

Im Rahmen des Workshops sollen aktuelle Forschungsarbeiten einschließlich „Work in Progress“ zum Themenkomplex IT-Governance sowie dem strategischen Informationsmanagement vorgestellt und diskutiert werden. Der Workshop soll auch Industrieteilnehmern die Möglichkeit geben, ihre Erfahrungen aus der Praxis der IT-Governance zu berichten. Dabei sind unter anderem die folgenden Themen von Interesse:

[1] http://www.itg-sim.de/

- Governance-Frameworks sowie deren Anwendbarkeit in unterschiedlichen Szenarien
- Standardisierungen im Governance-Umfeld
- Richtlinienmodelle für die IT-Governance
- Best Practices für die IT-Governance
- Governance aus Sicht von Geschäftsprozessen
- Data Governance
- Strategisches Informationsmanagement
- Softwarelebenszyklusmanagement
- Monitoring in Verteilten Systemen
- SOA-Governance und Cloud-Governance
- Service-Management und Risk-Management
- Security, Privacy & Trust
- Service-Engineering

In diesem Jahr konnten wir 14 Einreichungen für den Workshop gewinnen, aus denen 8 für den Tagungsband und zur Präsentation ausgewählt wurden. Bedanken möchten sich die Organisatoren bei den Gutachtern sowie den Autoren für ihren Beitrag.

2 Organisatoren und Programmkomitee

Organisatoren: Prof. Dr. Daniel F. Abawi (htw saar), Prof. Dr. Matthias Goeken (Hochschule der Deutschen Bundesbank), Prof. Dr.-Ing. André Miede (htw saar)

Programmkomitee: Prof. Dr. Carsten Felden (Technische Universität Bergakademie Freiberg), Prof. Dr. Horst Hellbrück (Fachhochschule Lübeck), Prof. Dr. Volker Herwig (Fachhochschule Erfurt), Dr. Wolfgang Johannsen (It's Okay - Governance in Practice), Prof. Dr.-Ing. Arne Koschel (Hochschule Hannover), Dr.-Ing. Ulrich Lampe (TU Darmstadt), Dr. Marc Lohmann (msg systems AG), Prof. Dr. Giselher Pankratz (Hochschule der Deutschen Bundesbank), Dr.-Ing. Stefan Schulte (Technische Universität Wien), Marco Tröbs (BearingPoint GmbH), Dr. Thomas Widjaja (TU Darmstadt & CASED), Prof. Dr. Till Winkler (Copenhagen Business School), Prof. Dr. Alfred Zimmermann (Hochschule Reutlingen)

Ein semantisches Netz des Informationsmanagements im Krankenhaus

Franziska Jahn, Michael Schaaf, Barbara Paech, Alfred Winter
Institut für Medizinische Informatik, Statistik und Epidemiologie
Universität Leipzig
Härtelstr. 16-18
04107 Leipzig
franziska.jahn@imise.uni-leipzig.de
michael.schaaf@imise.uni-leipzig.de
paech@informatik.uni-heidelberg.de
alfred.winter@imise.uni-leipzig.de

Abstract: Die Literatur zum Informationsmanagement ist durch eine Vielzahl unterschiedlicher theoretischer Ansätze und praxisorientierter Frameworks geprägt. Zusammenhänge zwischen Aspekten des strategischen Informationsmanagements, der IT Governance, des IT Service Managements und dem Informationssystems sind nicht ausreichend formalisiert, um auf deren Grundlage entscheidungsunterstützende Werkzeuge für CIOs und Geschäftsleitung zu konstruieren. Insbesondere in Krankenhäusern ist der strategische Einsatz von Komponenten des Informationssystems bei geringen IT-Budgets eine herausfordernde Aufgabe. Im Rahmen des Requirements Engineering eines entscheidungsunterstützenden Werkzeugs für den CIO soll ein Semantisches Netz des Informationsmanagements im Krankenhaus (SNIK) konstruiert werden. Diese Arbeit skizziert die Methodik zur Konstruktion des SNIK und stellt Ergebnisse der ersten beiden von sechs Schritten auf dem Weg zu SNIK vor. Um sowohl die Werkzeugentwicklung für den CIO auf Basis von SNIK unterstützen zu können als auch SNIK in der Lehre des Informationsmanagements später einsetzen zu können, müssen jegliche Designentscheidungen in iterativen Zyklen überdacht und angepasst werden.

1 Einleitung und Fragestellung

CIOs in Krankenhäusern und die Krankenhausleitung benötigen ein tiefes Verständnis dafür, wie das Informationssystem (IS) zum Erreichen der strategischen Krankenhausziele eingesetzt werden kann. Kennzahlen zu Kosten und Effizienz des Informationssystems bilden nicht alle entscheidungsrelevanten Informationen für das strategische Informationsmanagement ab. Vielmehr sind CIO und Krankenhausleitung auf vernetzte Informationen über das Projektgeschehen im taktischen Informationsmanagement (IM), den Betrieb des IS im operativen IM sowie die langfristige Weiterentwicklung im strategischen IM angewiesen. Die Informationen über

das IM müssen wiederum mit Informationen über das IS, den von ihm unterstützten Geschäftsprozessen und der Unternehmensstrategie vernetzt sein. Sowohl CIO als auch Krankenhausleitung benötigen Unterstützung bei der Navigation durch diese vernetzten Informationen, wozu ein Werkzeug in Form einer Business-Analyse-Software für das vernetzte Informationsmanagement hilfreich wäre.

Vor der Konstruktion einer solchen Software sieht man sich im Requirements Engineering mit den Begriffswelten der unterschiedlichen problemorientierten, aufgabenorientierten und prozessorientierten Ansätze und Frameworks für das Informationsmanagement [Kr10] konfrontiert. Problemorientierte Ansätze fokussieren auf das strategische IM und das Strategic Alignment [BP85]. Hilfsmittel wie Balanced Scorecards [IKB02] oder der GQM+Strategies [Ba10] helfen, den Beitrag des IS zu den Krankenhauszielen qualitativ und quantitativ zu beschreiben. Aufgabenorientierte Ansätze des IM (z.B. [HS09]), sammeln die vielfältigen Aufgaben vom Projektmanagement über das operative Management bis zur Strategieformulierung und stellen Zusammenhänge zwischen den Aufgaben her. Prozessorientierte Ansätze (z. B. [ÖBH92]) detaillieren die Aufgaben des IM. Frameworks wie COBIT (Control Objectives for Information and Related Technology, [IS12]) oder ITIL (Information Technology Infrastructure Library, z. B. [CC05]) gliedern sich hier ebenfalls ein, da sie Prozessbeschreibungen für unterschiedliche Aspekte des strategischen und operativen Informationsmanagements liefern. COBIT und ITIL stellen darüber hinaus Beziehungen zwischen Informationsmanagement und IT Governance her.

Die Begriffsvielfalt aufgrund unterschiedlicher Ansätze und Frameworks innerhalb des Informationsmanagements geht mit der Verbreitung von Synonymen, Homonymen, Vergröberungen und Begriffsüberschneidungen einher. Ebenso sind die methodischen Ansätze nicht hinreichend formalisiert, weshalb auch die Zusammenhänge zwischen strategischem, taktischem und operativem Informationsmanagement sowie IT Service Management und IT Governance nicht ausreichend beschrieben sind. Es besteht die Herausforderung, vorhandene Werkzeuge für das IM im Krankenhaus, die auf unterschiedlichen Ansätzen und Frameworks basieren und somit unterschiedliche Begriffe verwenden, semantisch zu integrieren, um vernetztes IM unterstützen zu können. Darüber hinaus werden relativ aufwändig im Unternehmen zu etablierende Frameworks wie ITIL oder COBIT in den mit vergleichsweise geringen IT-Budgets [Pr08] ausgestatteten Informationsmanagementabteilungen in Krankenhäusern seltener umgesetzt.

Langfristig soll deshalb ein semantisches Netz des Informationsmanagements im Krankenhaus (SNIK) konstruiert werden, das die komplexen Zusammenhänge der verschiedenen Ansätze des IM beschreibt. Abbildung 1 skizziert einen möglichen, vereinfachten Ausschnitt aus SNIK. Es werden beispielhaft einige Zusammenhänge zwischen Begriffen des strategischen Informationsmanagements („Krankenhausziel“ und „IT-Ziel“), einem Begriff des taktischen Informationsmanagements („IT-Projekt“), einer Komponente des Informationssystems („Anwendungssystem“) und Personalressourcen („IT-Mitarbeiter“) dargestellt.

Auf Basis des SNIK sollen zukünftig entscheidungsunterstützende Werkzeuge für CIOs von Krankenhäusern entwickelt werden können. Derartige Werkzeuge integrieren Informationen aus dem Krankenhaus und dessen Informationssystem für die vielfältigen Teilaufgaben des Informationsmanagements und unterstützen die Kommunikation zwischen CIO und Krankenhausleitung.

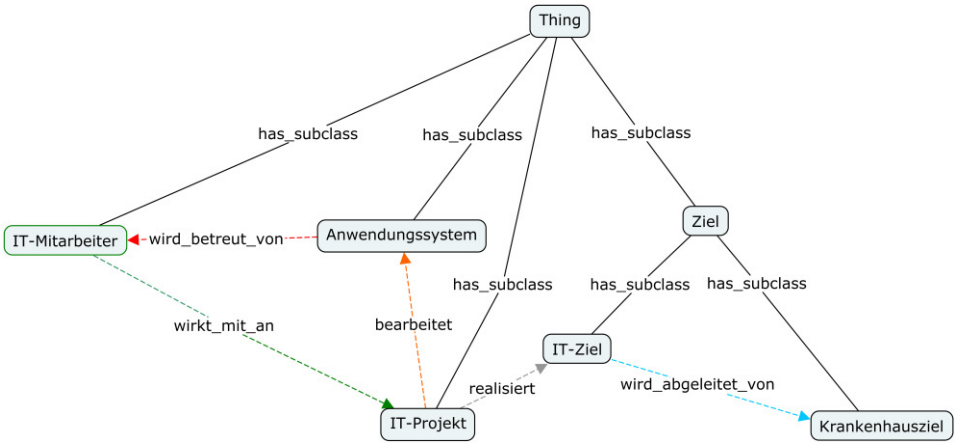


Abbildung 1: Möglicher, stark vereinfachter Ausschnitt aus dem Semantischen Netz des Informationsmanagements im Krankenhaus (erstellt mit CmapTools).

Der gesamte Entwicklungsprozess von SNIK wird sechs Schritte umfassen und leitet sich aus etablierten Methoden zur Entwicklung von Ontologien ab (siehe Kapitel 2). Die vorliegende Arbeit konzentriert sich auf die ersten beiden Schritte und trifft im Sinne eines Ontologieentwicklungsprozesses Festlegungen zur

1) Spezifikation sowie zur

2) Bewertung des Integrationspotentials existierender semantischer Netze und Ontologien.

(siehe Kapitel 3.1).

2 Methoden

Die Entwicklung des semantischen Netzes kombiniert Teilschritte aus den Vorgehensmodellen „Ontology Development 101“ [NM01] und „Methontology“ [FGJ97] für die Ontologieentwicklung. Tabelle 1 stellt die sechs definierten Schritte zur Entwicklung von SNIK sowie konkrete Aktivitäten zu deren Umsetzung dar.

Tabelle 1. Kombination von Ontology Development 101 und Methontology zur Entwicklung des SNIK

	Entwicklungsschritt	Aktivitäten
1	Spezifikation	Festlegen von Randbedingungen (z. B. Ein- und Ausschluss von Konzepten, Granularität)
2	Bewertung des Integrationspotentials existierender Semantischer Netze und Ontologien	Recherche und Analyse zu bestehenden Top-Level-Ontologien und Domänenontologien hinsichtlich ihrer Wiederverwendbarkeit für SNIK
3	Identifikation von Begriffen	Literaturreview von etablierter und aktueller Literatur zum Informationsmanagement, Nutzung von Methoden des Natural Language Processing zur Literaturaufarbeitung
4	Konzeptualisierung von SNIK	Zuordnung der in Schritt 3 gefundenen Begriffe zu Konzepten, Instanzen, Relationen und Eigenschaften, Übertragung ins Modellierungswerkzeug
5	Spezifikation von Eigenschaftsrestriktionen	Festlegung von Restriktionen in Form von Kardinalitäten, Einheiten und Wertebereichen
6	Entwicklung der Instanzenebene des SNIK	Identifikation von Instanzen durch eine Umfrage unter CIOs und Krankenhausleitern

In dieser Arbeit sollen Schritt 1 und 2 im Fokus stehen. Zunächst werden in der Spezifikation die Randbedingungen für das Semantische Netz festgelegt. Anschließend wird eine Recherche zu verfügbaren Top-Level-Ontologien und Domänenontologien durchgeführt, welche hinsichtlich ihrer Wiederverwendbarkeit nach einem Kriterienkatalog untersucht werden. Auch wird überprüft, inwieweit die methodischen Grundlagen des Unified Medical Language System (UMLS, [U.14]), d.h. die Verwendung eines Metathesaurus, eines Semantic Networks und eines Specialist Lexicon auf SNIK übertragen werden können.

3 Ergebnisse

Die Ergebnisse sich auf die in Tabelle 1 genannten ersten zwei Teilschritte der Entwicklung des SNIK.

3.1 Spezifikation von SNIK

Für die Spezifikation des semantischen Netzes wurden folgende Festlegungen getroffen:

- Zweck: Neben der Konstruktion entscheidungsunterstützender Werkzeuge für den CIO im Krankenhaus soll SNIK auch die Lehre und Forschung zum Informationsmanagement unterstützen. Durch die Strukturierung und Systematisierung der unterschiedlichen Ansätze des Informationsmanagements können die Begriffe des Informationsmanagements besser vermittelt werden. Auf Grundlage von SNIK können bestehende Ansätze auch systematischer weiterentwickelt und branchenspezifisch angepasst werden.

- Formalitätsgrad: Da semantische Netze nicht nur für Computer interpretierbar, sondern auch für menschliche Nutzer verständlich sind, ist im Sinne der genannten Verwendungszwecke die Entwicklung eines semantischen Netzes gegenüber einer Ontologie, die dem Prädikatenkalkül erster Stufe genügen muss, angestrebt.

- Ein- und Ausschlusskriterien: Ein umfangreicher Fragenkatalog, an dem sich der Ein- und Ausschluss von Begriffen orientieren soll, fasst die Fragen zusammen, die mit Hilfe des semantischen Netzes bzw. darauf aufbauenden Werkzeugen für den CIO beantwortet werden sollen, z. B. „Welche IT-Ziele ergeben sich aus einem bestimmten Krankenhausziel?“, „Für welche Aufgaben des strategischen Informationsmanagements hat der CIO Entscheidungsbefugnisse?“, „Welcher (IT-)Mitarbeiter ist Projektleiter eines Projekts, das sich mit einem bestimmten medizinischen Anwendungssystem befasst?“

3.2 Bewertung des Integrationspotentials existierender Semantischer Netze und Ontologien

Sowohl die Wiederverwendung vorhandener Top-Level-Ontologien auf Metaebene des SNIK als auch die Verwendung von Begriffen aus Domänenontologien, die sich mit dem Unternehmensmanagement, dem Informationsmanagement und Informationssystemarchitekturen beschäftigen, erscheinen plausibel für die Entwicklung von SNIK.

Als relevante Top-Level-Ontologien wurden die General Formal Ontology, [He10], die Basic Formal Ontology [GSG04] und Upper Cyc [Cy14] identifiziert. Ebenfalls auf Metaebene erscheint auch der Einsatz eines Metathesaurus, eines Semantic Networks und eines Specialist Lexicon im Sinne der UMLS sinnvoll, wobei die Inhalte der UMLS in Form der medizinischen Begriffe nicht für die Beschreibung des Informationsmanagements geeignet sind. Im Metathesaurus des SNIK wäre das gesamte Vokabular mit eindeutigen Identifikationsnummern enthalten. Ein Begriff und dessen unterschiedliche Bezeichnungen wären über Relationen miteinander verbunden. Ein Semantic Network legt semantische Typen fest, denen alle Begriffe des Metathesaurus zugeordnet werden können. Für das Informationsmanagement wären z. B. semantische

Typen wie „Informationssystemkomponente“, „Rolle“, „Aufgabe des Informationsmanagements“, „Unternehmensprozess“ denkbar. Die in UMLS vorhandenen Relationen wie „Zeitlich in Beziehung stehend“, „Funktionell in Beziehung stehend“ könnten übernommen bzw. erweitert werden. Ein Specialist Lexicon stellt zu jedem Begriff bzw. jeder Bezeichnung in SNIK syntaktische, morphologische und graphemische Informationen zur Verfügung. Die Metakonzepte der UMLS entstanden 1989 und müssen deshalb auf die heute gängigen Methoden zur Erstellung von Ontologien abgebildet werden.

Einzelne Begriffe aus Domänenontologien wie der Business Model Ontology [Os04], e³value [GA01] und Resource-Event-Agent Enterprise Ontology [GM02] kommen für die Wiederverwendung in Frage, wobei diese Ontologien auf Wertschöpfungsketten in Unternehmen fokussieren und nur einige Aspekte von Informationssystemen und Informationsmanagement abbilden. Für Begriffe aus dem Bereich der Informationssysteme können z. B. die Konzepte von ARIS [Sc01] oder 3LGM² [Wi11] wiederverwendet werden.

4 Diskussion

Das zu konstruierende semantische Netz soll laut 3.1 sowohl die akademische Lehre des Informationsmanagements als auch die Konstruktion entscheidungsunterstützender Werkzeuge für das Informationsmanagements unterstützen. Um diesen zwei unterschiedlichen Verwendungszwecken gerecht zu werden, müssen die anfänglich getroffenen Festlegungen für das semantische Netz in regelmäßigen Abständen auf ihre Plausibilität geprüft und ggf. angepasst werden. Z. B. könnte es vonnöten sein, weitere Ein- und Ausschlusskriterien zu definieren. Auch ist der notwendige Formalitätsgrad von SNIK für die angestrebte Softwareentwicklung höher und somit ausschlaggebend, als er für die alleinige Entwicklung einer Lehreinheit zu Konzepten des Informationsmanagements wäre.

Bei späteren konkreten Designentscheidungen wird ebenfalls zwischen Alternativen zu wählen sein, die in unterschiedlichem Maße den beiden Verwendungszwecken zuträglich sind. Auch können sich die wiederverwendeten Top-Level- oder Domänenontologien aus 3.2 bei der Hinzunahme neuer Begriffe für SNIK als ungeeignet erweisen und eine Rekalibrierung erfordern. Der iterative Charakter ist inhärenter Bestandteil der Ontologieentwicklung [NM01] und somit auch der Entwicklung von semantischen Netzen.

Obwohl SNIK krankenhausspezifisch ausgerichtet ist, wird eine Übertragbarkeit auf andere Branchen angestrebt. Dies soll erreicht werden, indem krankenhausspezifische Ausprägungen allein auf Instanzenebene des SNIK formuliert werden. Dementsprechend kann die Konzeptebene von SNIK auf das Informationsmanagement in jeglichen Unternehmen übertragen werden.

5 Literaturverzeichnis

[Ba10] Basili, V. R. et al.: Linking Software Development and Business Strategy Through Measurement. In Computer, 2010, 43; S. 57–65.

[BP85] Benson R.J.; Parker M. M.: Enterprise-wide Information Management - An Introduction to the Concepts, Los Angeles, 1985.

[CC05] CCTA: Service delivery. ITIL - The key to managing IT services. TSO (The Stationery Office), London, 2005.

[Cy14] Cycorp(R): ResearchCyc. http://www.cyc.com/platform/researchcyc, 06.05.2014.

[FGJ97] Fernandez, M.; Gomez-Perez, A.; Juristo N.: Methontology: From Ontological Art Towards Ontological Engineering. Technical Report AAAI Report No.: SS-97-06, 1997.

[GA01] Gordijn, J.; Akkermans, H.: Design and Evaluation of e-Business Models. In IEEE Intelligent Systems, 2001, 16; S. 11–17.

[GM02] Geerts, G.; McCarthy, W.: An ontological analysis of the economic primitives of the extended-REA enterprise information architecture. In International Journal of Accounting Information Systems, 2002, 3; S. 1–16.

[GSG04] Grenon, P.; Smith, B.; Goldberg, L.: Biodynamic Ontology: Applying BFO in the Biomedical Domain. In (Pisanelli, D. M. Hrsg.): Ontologies in medicine. IOS Press, Amsterdam, Burke, VA, 2004; S. 20–38.

[He10] Herre, H.: General Formal Ontology (GFO): A Foundational Ontology for Conceptual Modelling. In (Poli, R.; Healy, M.; Kameas, A. Hrsg.): Theory and applications of ontology. Springer, Dordrecht, London, 2010; S. 297–345.

[HS09] Heinrich, L. J.; Stelzer, D.: Informationsmanagement. Grundlagen, Aufgaben, Methoden. Oldenbourg, München, Wien, 2009.

[IKB02] Inamdar, N.; Kaplan R. S.; Bower, M.: Applying the balanced scorecard in healthcare provider organizations. In J. Healthc Manag, 2002, 47; S. 179–195.

[IS12] ISACA: COBIT 5. Rahmenwerk für Governance und Management der Unternehmens-IT. ISACA, Rolling Meadows, IL, 2012.

[Kr10] Krcmar, H.: Informationsmanagement. Springer, Berlin, Heidelberg, 2010.

[NM01] Noy, N. F.; McGuinness D. L.: Ontology Development 101: A. Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[ÖBH92] Österle, H.; Brenner, W.; Hilbers, K.: Unternehmensführung und Informationssystem: Der Ansatz des St. Galler Informationssystem-Managements. Teubner, Stuttgart, 1992.

[Os04] Osterwalder, A.: The Business Model Ontology – A Proposition in a design science approach. Dissertationsschrift, Lausanne, 2004.

[Pr08] PricewaterhouseCoopers AG WPG: Der Wertbeitrag der IT zum Unternehmenserfolg, Stuttgart, 2008.

[Sc01] Scheer, A.-W.: ARIS - Modellierungsmethoden, Metamodelle, Anwendungen. Springer, Berlin [u.a.], 2001.

[U.14] U.S. National Library of Medicine: Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/, 06.05.2014.

[Wi11] Winter, A. et al.: Health Information Systems. Architectures and Strategies. Springer-Verlag London Limited, London, 2011.

Federal Information Management – Context and Effects

Norbert Ahrend[1], Petra Wolf[2], Michael Räckers[3], Alexander Dentschev[1], Marlen Jurisch[4]

[1]AIOS GmbH
Schönhauser Allee 10-11
10119 Berlin
norbert.ahrend@aios.de
alexander.dentschev@aios.de

[2]fortiss, An-Institut TU München
Guerickestr. 25
80805 München
wolf@fortiss.org

[3]Westfälische Wilhelms-Universität Münster
European Research Center for Information Systems
Leonardo-Campus 3
48149 Münster
michael.raeckers@ercis.uni-muenster.de

[4]Technische Universität München
Lehrstuhl für Wirtschaftsinformatik – I17
Boltzmannstr. 3
85748 Garching bei München
marlen.jurisch@in.tum.de

Abstract: In spite of their common legal basis public sector organizations have developed a plethora of diverse forms and service descriptions for similar or even identical citizen services. As a consequence for citizens as well as for businesses it is difficult to understand the wording of forms and the procedures necessary for citizen and business services. Public administration cannot profit from synergy effects which could be gained by a shared effort for developing these information – although a holistic approach for an integrated information management should be the goal of strategy. The aim of this paper is to combine and integrate data and information from the information domains of services, processes and forms from the public sector domain and provide a holistic approach to make them usable for information managers in the field. To this end, we identify the essential relationships and effects of these information domains on the administrative action. We develop a framework based on the information needs outside of the administration (citizens and private sector) and within the administration, which represents content and interaction. We also examine to which extent such a structure can be further defined on a more detailed level and integrated.

1. Introduction

In spite of their common legal basis public sector organizations have developed a plethora of diverse forms and service descriptions for similar or even identical citizen and business services. As a consequence for citizens as well as for businesses it is difficult to understand the wording of forms and the procedures necessary for citizen and business services. Besides public administration cannot profit from synergy effects which could be gained by a shared effort for developing and maintaining these information.

If the information on citizen and business services needs to be updated as a consequence of changes in the legal basis the quality of the information provided is often heterogeneous in terms of completeness and usability caused by scarce personal resources. This is especially critical for public administrations since that information has to be usable and reliable for citizens and for businesses. In order to meet this challenge several initiatives are aiming to conceptualize specific (public sector) information domains and provide standards concerning structure and content [AWL13], [Bm11], [Le05]. Yet, the federal structure of the organizational units concerned with developing and maintaining citizen service information is still a barrier for leveraging synergy effects by concentrating that effort.

In order to provide the basis for improving the quality of citizen service information as well as the effectiveness of the provisioning process itself common information quality standards and organizational structures for information editing are necessary and from strategic importance for the public sector. Thus harmonization of and shared organizational structures for providing citizen service information are the key elements of the solution proposed in our paper.

There are three major information elements describing a citizen (or business) service: (1) The textual description of the services including the respective legal basis, the responsible part of the public administration and documents needed for processing the service. (2) The corresponding form(s) usually delivers information for the public administration or more general the receiver. Forms can be input documents or output documents. An example for the former one are application forms. Examples for the latter one are notices or output documents of other services such as passports. Forms trigger the (3) process of generating the service. Service descriptions, forms and processes represent individual information domains, which serve different information needs. The different requirements in the three information domains can be attributed to different conditions in the administrations. Different goals also affect information needs and their granularity.

Thus, information about and around services and processes in the public sector can only be harmonized by means of an overarching architecture for a federal information management (FIM) between the federal level, the federal state level and the municipal level. Additionally, the definition of editorial processes plays a central role in order to ensure the quality and availability of relevant information as well as the harmonization of the structural basis for deployment and retrieval of information. The three major information domains of public administration need to be merged and linked in a holistic approach by mapping contexts and their provision of information management services. The corresponding approaches in Switzerland are equally transferable to Germany [Le05], [DJS10], [Wi11], [Sc09], [LSS10].

The focus of the research presented in this article is the representation of the interaction of these three information domains in a holistic FIM. The aim of this paper is therefore to show how the approach of a FIM links the different information needs and can effectively provide adequate information.

In order to leverage synergy effects and enforce the harmonized use of service information, it is suitable to impose or recommend standards [Di08], which have been elaborated top-down across the different administrative levels and to define quasi-standards on the implementation level.

From a methodological perspective, this research focuses on the design science principle. Based on a problem statement and the investigation of the related work, requirements can be defined and can be used as a basis for the development of a concept to support and address this gap. The construction of such a concept is followed by an epistemological aspect, namely the evaluation of the developed concept. Although we will focus on the construction and development of a concept to solve the problem and only partly will look on the evaluation in this article, this research is based on the paradigm of design science, which has recently been predominantly discussed based on Hevner et al.[He04], [Pe07].

Based on a requirements engineering and formulation of a problem statement, we finally collected four questions that are guiding us through the design science process: (1) Which information do the public sector authorities, the citizens and the businesses require? (2) Which relationships and dependencies exist between the different components of information? (3) How can such a structure be gradually refined and integrated within each other? (4) What are, based on this concept, the benefits for the public administrations, citizens and businesses?

To answer these questions, this paper is structured as follows: In chapter 2 related work is discussed. In chapter 3 we present the overall framework of FIM and its components, based on our problem statement and open questions and the aspects we could conclude from the related work. Chapter 4 describes which standardized information (in which information domains) is provided and which relationships exist between the various domains of information, whereas chapter 5 summarizes the effects (and benefits) such integrated approach has. Finally, in chapter 6 we discuss the limitations and findings of our research and give an outlook on future work.

2. Literature Review

The interaction of information or data and business processes is the subject of an extensive research tradition. In IS research in the field of *Business Process Modeling* [Me95] or *Reference Modeling* [Sc98] many scientists examined how the flow of business processes and the required information and data can adequately be represented and designed.

These research efforts concerning (business) process modeling resulted in various concepts and modeling techniques to represent processes and information objects and to map them to each other [Sc91], [DP97]. Many of these modeling techniques allow representing different views to differentiate between the specific information needs of different user roles. A better understanding of this specific information needs leads to an improved *information logistics* [Di08].

The concept of information logistics is often discussed in the field of Master Data Management [SGZ12]. It includes components and processes for distribution and replication of master data elements. For the concept of FIM the challenge is to make the master information available in a central repository and in case of updates or corrections to ensure the distribution of the updated version of the master information. Master information is centrally prepared, but may be adjusted as needed for a particular application. Organizational measures such as version control must be taken to identify master data and their possible adjustment. At the same time, the knowledge about the multiple use of master information can support the standardization of data definitions and thus improve the efficiency of information processing. In particular, the design of high quality process structures requires the optimization across all information objects and their standardization [BH06], [RS97]. Therefore, this standardization should be promoted [BK04]. However, only those processes and systems should be centralized which fulfill the same tasks on the same execution level. The underlying IT support needs binding and uniform exchange formats, data structures and interfaces [KW05].

Reference Modeling provides a concept for this project, which structurally supports harmonization and interoperability by showing general examples of entire industries and requirements for data models, process models, organizational models and other models [RS97], [BS96]. The goal of these reference models is to meet the various requirements of different stakeholders and harmonize and integrate all these information requirements into one model.

The findings from Business Process Modeling can be transferred to the field of public administration. In both fields the process-oriented view becomes more important than the function-oriented view. Carrier of information objects such as forms or notices can be assigned to a particular administration process by merging the process view and the object view [BAF07].

3. Description of the Information Domains

The main principle of a federal information management is to harmonize and integrate the three information domains services, forms and processes into one entity (Figure 1). The different requirements in the three information domains can be attributed to different preconditions in the administrations. In order to take into account the different levels of details of the information needs, the FIM information domains also have to work in an integrated way even if not every information domain is available in the same level of detail.

From a citizen and business perspective the following requirements towards a FIM concept can be formulated: Citizens and especially businesses corresponding with different public administrations at different sites need unambiguous and reliable information on citizen or business services. Ideally that information on a specific service should not differ depending on the municipality or federal state it is requested from. In addition both citizens and business could benefit from the reuse of information: Information on name and address of a citizen or a business is requested at the beginning

of nearly all forms of public administration. If that and other frequently used information was put in a standardized form and structure citizens and even more so businesses could reuse that information in diverse administration contexts.

From a public administration perspective the reuse and especially the interoperability of information is an important requirement. Interoperability refers to sharing information within subject domains like public safety across federal levels as well as to sharing information across subject domains like urban development and transportation planning. In order to improve effectiveness and flexibility in the provision of online citizen services public administrations especially on the municipal level should be able to use pools of shared information on services, forms and processes which can be adapted to local requirements. Changes in the legal basis of citizen services result in high efforts to adjust service descriptions and forms or even processes. These efforts could be reduced if that information and the necessary updates – as a reference basis – were provided centrally. Many forms like the individual ID card are used in the course of diverse administration services. Systematic links between forms, services and respective processes can help to standardize interfaces and improve interoperability.

The overall concept of FIM consists of different concepts which have to be conceptualized and specified in an integrated manner. In doing so, the **FIM Logic** forms the core concept for harmonization. This logic specifies the FIM information domains, their linkage and rules to define the interaction among the FIM further modules of the overall architecture. The **FIM Tools** consist of concepts, methods and standards used to create, maintain and use FIM content. So the FIM tools will include methods for business process modelling, such as BPMN, and also will specify the mechanisms and rules to create service descriptions and reference forms for FIM. The **FIM Building Set** comprises all structural elements needed to create FIM content – by using the FIM Tools. The building set consists for example of wiki modules, form fields, process elements as well as linkage rules. The FIM Building Set provides the basic FIM content ready to be combined in as many ways as possible and as easily as possible reusable. For example, the FIM modular component “forms” consists of LeiKa[1] modules, form fields, form field groups and rules.

For the FIM component "forms" that are, for example, the modular elements of fields, field groups and rules. Further, an unlimited number of fields and field groups can be included in the master form.

Content for the **FIM Library** can be created by means of the elements of the FIM Building Set and by using the methodologies and principles included in the FIM Tools. These are for instance reference forms, which consist of form fields, form field groups and linking rules based on federal law or federal state law. The **FIM Editorial Concept** consists of editorial processes and defines user roles for FIM, which elements have to be maintained and can be used and how to manage the creation, maintenance, access or the quality assurance.

[1] LeiKa = Leistungskatalog

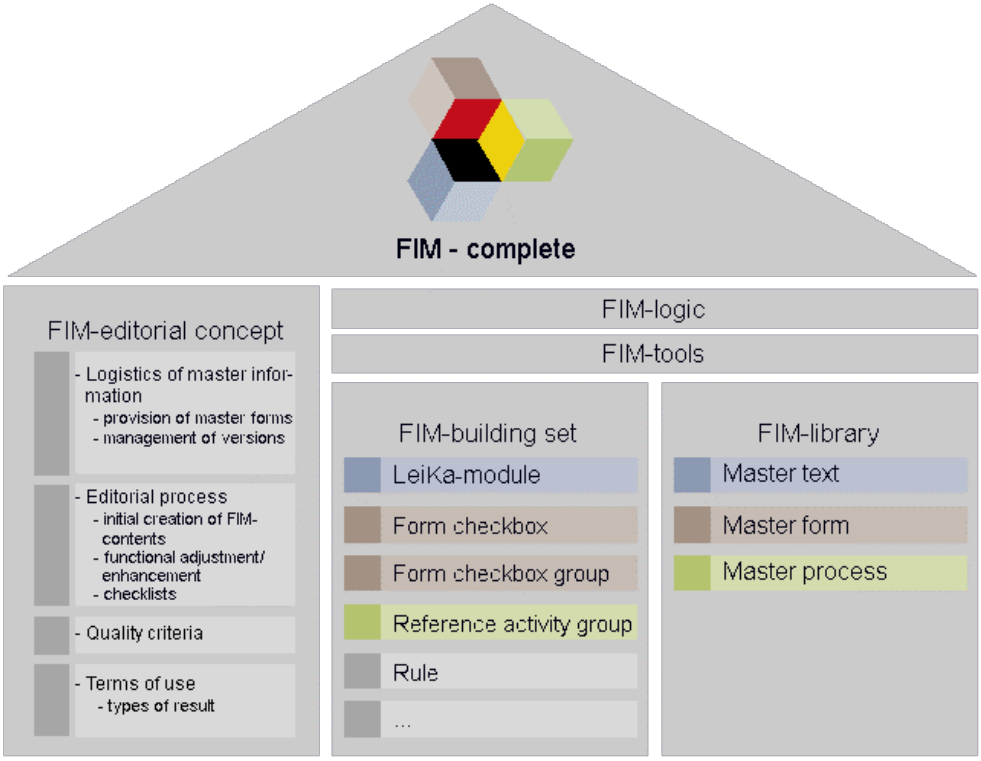


Figure 1: FIM House

4. Interaction of the modular components

One of the major goals of FIM is to not only provide information but also to link this information across the three information domains. The level of detail of these linkages will be continuously refined in three stages during the implementation of the concept in practice. When doing this, the three information domains will mellow in different pace. It is essential that in every stage the interplay of the information domains is possible even if they are not on the same level of detail.

In the first stage, the modular elements are mostly isolated from each other (Figure 2). Forms and processes are black boxes for each other. Only the services are standardized which allows an explicit matching of "Forms" and "Processes" within the area of service descriptions. The element "Services" (LeiKa) includes the description module "Necessary Documents" and "Forms" which lists all documents and forms. The description of services for the application for issuing a weapon, for instance, includes in the first stage a reference to all necessary documents (application form, etc.) in form of an ID form instead of text.

Similarly, the service specification contains a reference to the corresponding process by means of a process ID. Further, the process references the respective service by means of the LeiKa ID.

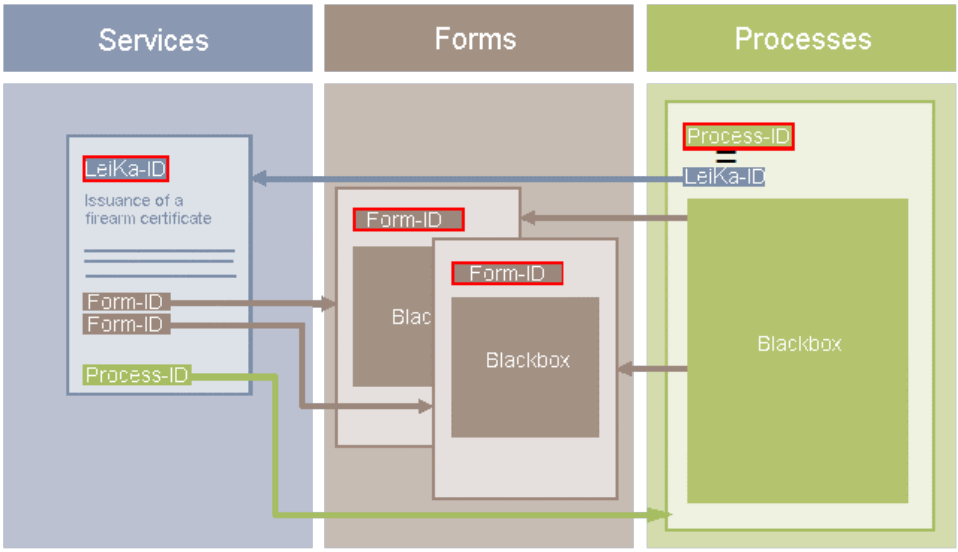


Figure 2: Interaction of the FIM modular elements during the first stage

In the second stage, the details of the black boxes are shown and the linkages between the components are specified in more detail (field group IDs, activity IDs; Figure 3).

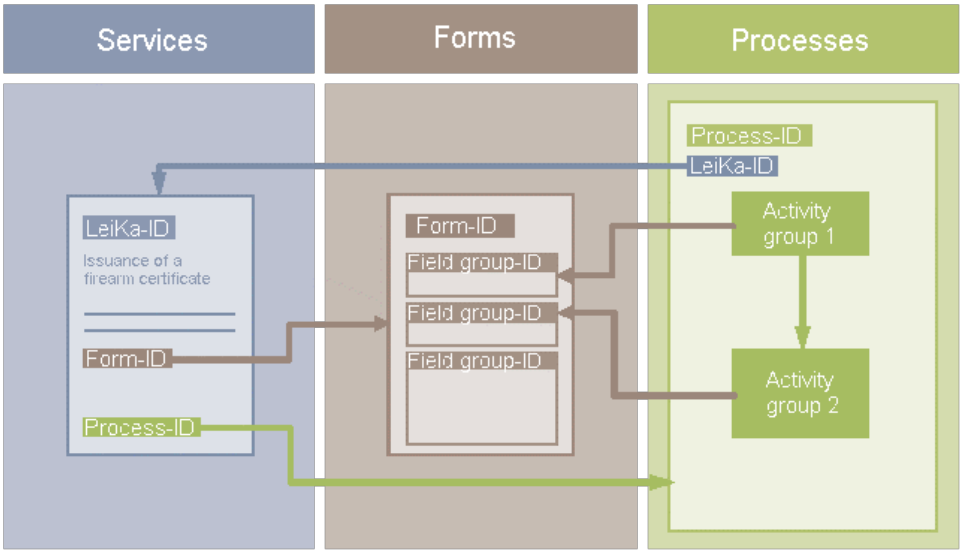


Figure 3: Interaction of the FIM elements in the second stage

In this stage, the recurring combinations of field groups (e.g., the address, consisting of name, street, postal code and city) and activities (e.g., the processing of incoming mail,

consisting of opening, scanning, detecting the metadata, and forwarding to the person in charge) have to be associated with the service description. At the same time, it is a requirement in the process view that the form fields needed for specific activity bundles can clearly be identified. This means that for the processing of incoming mail the field group IDs and their corresponding field groups are stored in the metadata of an activity bundle.

In the third stage, the linkages between the components are also specified in the field level wherever necessary and appropriate (Figure 4). Across all stages of development the linkages between processes, services and LeiKa IDs are stored in the Process Characteristics.

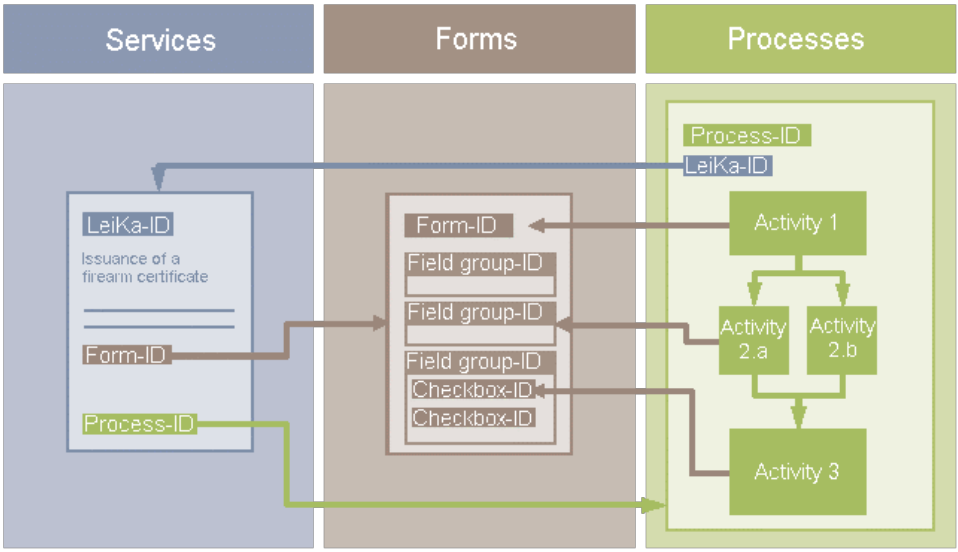


Figure 4: Interaction of the FIM elements in the third stage

For the implementation of this concept, a data model was developed, which reflects the described relationships between the FIM components and represents the core of the FIM logic. Using the data model (i.e., FIM Meta Model), one can link the single service descriptions and their corresponding forms and documents as well as the corresponding process. On a more detailed level, each activity bundle respectively specific activities of a process will be matched to the corresponding form field groups and form fields. One example is the matching of a responsible person to the last name and address of an applicant. The matching is realized by means of IDs, which are stored in the metadata of the FIM components.

The meta model matches one service to one or more forms and documents.

The term **Object Document** designates administration external documents such as blueprints or pictures. They are also used as an information carrier in the context of administrative services.

The matching between service and form or document is realized by means of the IDs: the LeiKa ID, form ID and document ID[2]. Similarly, a service or the respective forms and documents are matched to the corresponding process by means the process ID. The relationships shown below apply to both the main level of information as well as the instantiations at the execution level.

The data model has continuously been refined according to the three stages of development. In the third stage, the relationships between the elements are described (form field and activity (= process step), cf. Figure 5). In a specific form, the content of form field and form field groups can be linked, which are described by rules (e.g. field selection rules, linkage rules, plausibility and consistency rules, structural rules, building regulations or subject-specific rules).

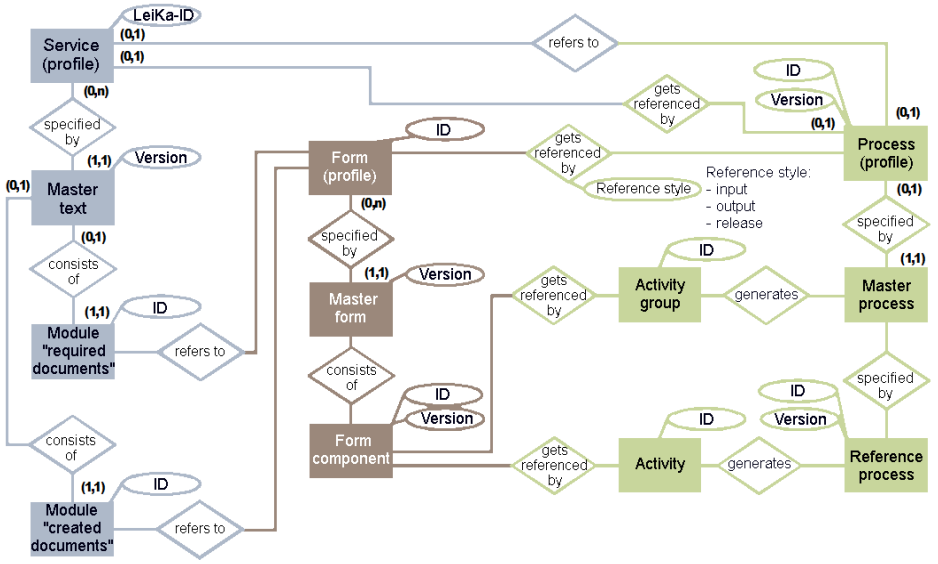


Figure 5: FIM data model in the third stage

Several activities form an activity bundle. Form fields and activities are identified by and associated with IDs. Similar to the matching between activity bundles and form field groups, the matching between activities and form fields take place in the context of a specific form.

5. Effects of a Federal Information Management

Based on the data model and concept presented several effects on the enhancement of a federal information management can be identified and described. These effects are not only based on the structural framework presented above. When digging into more detail and also thinking about harmonization on the content within this structure we will come

[2] For clarity, the illustration of the object document is omitted in the graphical representation.

up with useful and measureable effects for a holistic federal information management. The result of the integrated view is a comprehensive connecting and harmonized interaction of the three information domains. These effects are summarized in Figure 6, we will explain in more detail:

Coming from the beginning, the starting point is the provision of services of the public administration (including administrative services): outside the administration for citizens and companies, as well as within the administration to support or accompany services. The information needs now can structurally be documented with the concept of service descriptions. The service descriptions consist of different modules which form the master text and provide the citizen or enterprises as well as the officials within the public administrations guidelines and support in what the service is about, who is responsible, which information is to be handed in, what are the possible results and further more. Two of the modules from the service descriptions support especially the interplay with further information domains: required and created documents. With this, a link to the domain of forms is structurally given. Furthermore, this structure guides the creation of service descriptions and fosters the user of the concept to document comparable information in a comparable way – and integrate the links to the further information domains in a structured, harmonized way.

An application form or a comparable form usually delivers information for the public administration – or more general the receiver – and triggers further activities. The forms mainly are a structured list of fields and groups of fields, which describe the information the applicant has to hand in for a certain service request. Additionally, the forms also cover further forms in a wider sense like certificates or other documents, which the applicants have to hand in. Thirdly, the domain of forms describe the output information, which is given back to the applicant. These forms or documents like official notifications of verdicts are the results of the process of generating services. The winning effect of this structure is given when it is used to create a set of harmonized and reusable form fields and form field groups. Such set of elements will support the creation of harmonized master forms in which same information is requested with same elements and same concepts.

The processes are the activities running within the public administrations taking the input forms and creating the result forms. When transforming the information into results, several process steps will take place. Again, the effect of the concept presented so far will be apparent when defining a common set of activities and activity bundles and have them prepared for reuse for modelling the processes. Then it will help to harmonize the processes, identify same steps in the processes as same or similar and help the users of the concept to prepare usable process descriptions.

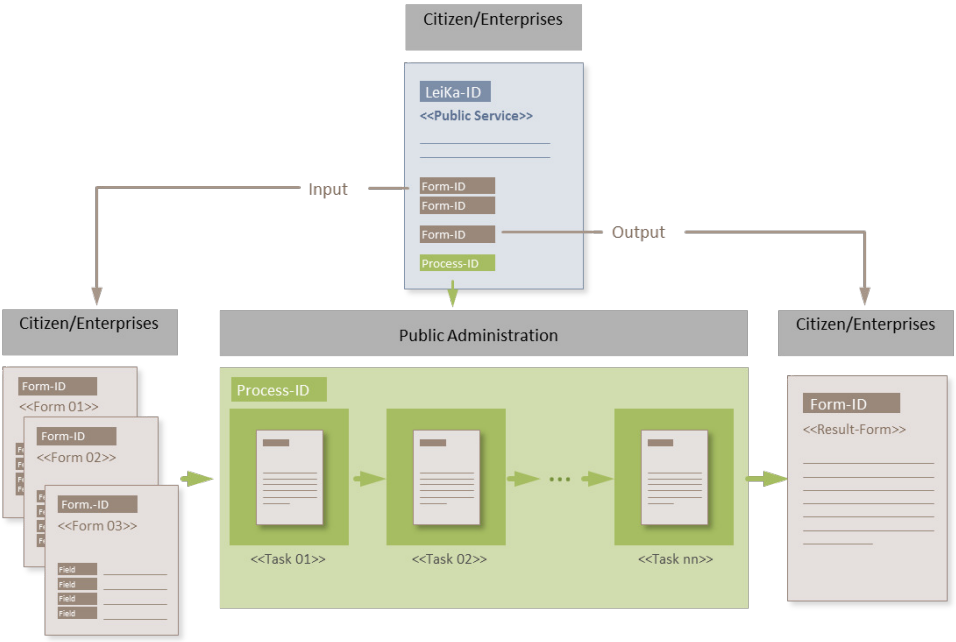


Figure 6: Effects of the integrated view of the information domains

6. Concluding remarks

The aim of this paper is to present the content and especially the relationship between the components of a FIM which contributed to the strategy of an efficient and effective information management for the public sector. Our work was guided by four questions described at the beginning. When analyzing the public sector domain, we identified three information domains relevant for the structuring of the field: service descriptions, business processes and data/information, conceptualized through forms. This contribution shows that the interaction between data or information and business processes is not only relevant for the private sector but also for the public sector [Be12], [Sa11]. While concepts for business process management in the private sector are largely agreed upon, these concepts for public administrations are just about to emerge and not yet tested (Question 1).

We structured the different views identified and developed a concept and data model to describe the relationships and interdependencies of these concepts of information (Question 2). Next, we developed a three step approach of refinement to address the different maturity of the three information domains and the different pace of development of each information domain (Question 3). Although there are appropriate techniques to describe processes and information objects and to associate them to each other [Sc91], their (across federal level and the federal state level) use in the public sector is still in its infancy [Be12]. It is shown that the understanding of the specific

information requirements in certain parts of the process, contributes to improve information logistics.

Finally, we described the effects, this integrated structure will have on the public sector domain (Question 4). Until now, this approach has not been sufficiently anchored in the practice of public administration. Although the potential benefits are recognized, the specifics of the federal administrative organizations inhibit their materialization. Thus, for the better answering of Question 4, intensive evaluation of the concept is a next step in the overall approach and object of further research, which is actually due. However, the multiple use of different types of information on the various administrative levels and in different functional contexts makes the effect of a standardization of data definitions evident. This effect is enhanced when the relationships between the components of a FIM are taken into account. The identified third stage corresponds to the maturity level of the private sector for entire industries [BS96]: universally valid examples are defined, guidelines for data models, process models, organizational models and other models are given.

Bibliography

[AWL13] N. Ahrend, K. Walser, and H. Leopold, “Case Study of the Implementation of Business Process Management in Public Administration in Germany, Switzerland and Austria.” 2013.

[Bm11] BMI, “Abschlussbericht Pilotbetrieb D115,” 2011.

[Le05] M. Lenk, “IMAGI-das Gremium des Bundes zur Koordinierung des Geoinformationswesens in der Bundesverwaltung ,” Bayern , 2005.

[DJS10] F.-D. Dorloff, V. Jahns, and V. Schmitz, “Organisationsübergreifender Austausch öffentlicher Dienstleistungsinformationen: Konzept, Architektur und Erkenntnisse aus dem Piloten Amt@ Direkt,” Essen, 2010.

[Wi11] M. Wind, “Die „neue Zentralität “im E-Government.” 2011.

[Sc09] G. Schweiz, “E-government strategie schweiz,” Bern , 2009.

[LSS10] K. Lenk, T. Schuppan, and M. Schaffroth, “Vernetzte Verwaltung - Organisationskonzept für ein föderales E-Government Schweiz,” 2010.

[Di08] B. Dinter, R. Winter, P. Chamoni, N. Gronau, and K. Turowski, *Synergien durch Integration und Informationslogistik*. Bonn, 2008.

[He04] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design Science in Information Systems Research ,” *MIS Q.* , vol. 28, no. 1, pp. 75–105, 2004.

[Pe07] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A Design Science Research Methodology for Information Systems Research,” *J. Manag. Inf. Syst.*, vol. 24, no. 3, pp. 45–77, 2007.

[Me95] P. Mertens, F. Bodendorf, W. König, A. Picot, and M. Schumann, *Grundzüge der Wirtschaftsinformatik* . Berlin : Springer , 1995.

[Sc98] A.-W. Scheer, *Wirtschaftsinformatik: Referenzmodelle für industrielle Geschäftsprozesse*. Berlin : Springer, 1998.

[Sc91] A.-W. Scheer, *Architektur integrierter Informationssysteme: Grundlagen der Unternehmensmodellierung*. Berlin : Springer, 1991.

[DP97] T. H. Davenport and L. Prusak, *Information Ecology: Mastering the Information and Knowledge Environment*. Oxford University Press, 1997.

[Sc12] R. Scheuch, T. Gansor, and C. Ziller, *Master Data Management: Strategie, Organisation, Architektur*. dpunkt.verlag, 2012.

[BH06] J. Bauer and E. Hayessen, *Prozesscontrolling*. Springer, 2006.

[RS97] M. Rosemann and R. Schütte, “Grundsätze ordnungsmäßiger Referenzmodellierung. Entwicklungsstand und Entwicklungsperspektiven der Referenzmodellierung,” 1997.

[BK04] T. Böhmann and H. Krcmar, “Grundlagen und Entwicklungstrends im IT-Servicemanagement,” *HMD-Praxis der Wirtschaftsinformatik*, pp. 7–21, 2004.

[KW05] H. Kubicek and M. Wind, “Kompatibler Zoo,” *Kommune 21*, vol. 2, p. 21, 2005.

[BS96] J. Becker and R. Schütte, *Handelsinformationssysteme*. Landsberg/Lech , 1996.

[BAF07] J. Becker, L. Algermissen, and T. Falk, *Prozessorientierte Verwaltungsmodernisierung: Prozessmanagement im Zeitalter von E-Government und New Public Management*. Berlin Heidelberg: Springer, 2007.

[Be12] J. Becker, S. Hofmann, M. Jurisch, R. Knackstedt, H. Krcmar, M. Räckers, I. Thome, and P. Wolf, “Prozessorientierte Verwaltung - Status quo und Forschungslücken,” *Lect. Notes Informatics* , pp. 61–72, 2012.

[Sa11] A. F. L. Santana, C. F. Alves, H. R. M. Santos, and A. d. L. C. Felix, “BPM Governance: An Exploratory Study in Public Organizations,” in *Enterprise, Business-Process and Information Systems Modeling*, Springer, 2011, pp. 46–60.

Reifegradbestimmung der IT-Governance: Eine Fallstudie zur Anwendbarkeit des COBIT 5 PAM in der öffentlichen Verwaltung

Tobias Grönert[1], Jens Pöppelbuß[2], Andreas Breiter[3]

[1]Institut für Informationsmanagement Bremen (ifib)
Am Fallturm 1, 28359 Bremen, tgroenert@ifib.de
[2]Juniorprofessur für Industrienahe Dienstleistungen, Universität Bremen,
Wilhelm-Herbst-Str. 5, 28359 Bremen, jepo@is.uni-bremen.de
[3]Institut für Informationsmanagement Bremen (ifib) und Universität Bremen,
Am Fallturm 1, 28359 Bremen, abreiter@ifib.de

Abstract: Der Reifegrad der IT-Governance in öffentlichen Verwaltungen wird im Allgemeinen als gering eingeschätzt. Um öffentlichen Verwaltungen eine Möglichkeit aufzuzeigen, ihren eigenen Reifegrad zu bestimmen, wurden in einer Fallstudie bei der Landeshauptstadt München relevante IT-Governance-Prozesse identifiziert und ein etabliertes Reifegradinstrument angepasst. Als Basis diente das mit COBIT 5 neu eingeführte Process Assessment Model (PAM). Der Beitrag macht deutlich, wie dieses Modell im Kontext öffentlicher Verwaltungen genutzt werden kann und welche Faktoren bei der Übertragung berücksichtigt werden müssen.

1 Einleitung

In Bund, Ländern und kommunalen Körperschaften hat die IT-Unterstützung von Verwaltungsprozessen längst eine existenzielle Bedeutung erreicht. Hiermit einher geht die Notwendigkeit, klare Regeln für das strategische Management der IT zu definieren. Der Reifegrad der IT-Governance in öffentlichen Verwaltungen in Deutschland wird jedoch im Allgemeinen als gering eingeschätzt und dementsprechender Nachholbedarf aufgezeigt [Jo12, We13]. Wissenschaftliche Studien zeigen, dass ansonsten etablierte Referenzmodelle wie COBIT nur selten von der öffentlichen Hand eingesetzt werden [Jo12]. Es finden sich auch nur vereinzelt Fallstudien, die über den konkreten Einsatz entsprechender Frameworks im öffentlichen Sektor berichten [BF11, OBP12]. Insbesondere IT-Governance in deutschen Kommunalverwaltungen ist in der Forschung bislang unterrepräsentiert.

Mangelnde Erfahrungs- und Vergleichswerte erschweren es, die Ist-Situation der IT-Governance in einer öffentlichen Verwaltung angemessen einschätzen zu können. Mit dieser Herausforderung sehen sich vor allem die öffentlichen Verwaltungen konfrontiert, die bereits gezielt die Etablierung von IT-Governance-Strukturen vorantreiben und nach weiteren Entwicklungsmöglichkeiten suchen. So wurde bspw. bei der Landeshauptstadt

München (LHM) sehr früh die wachsende Bedeutung und die Notwendigkeit für übergreifende Strukturen im Bereich der IT-Governance erkannt und in einem umfassenden Restrukturierungsprojekt entsprechende organisatorische Anpassungen realisiert. Eine wesentliche Maßnahme war die Gründung der Hauptabteilung IT-Strategie, IT-Steuerung und IT-Controlling (STRAC), die für alle strategischen IT-Prozesse und für die IT-Governance der LHM zuständig ist.

Im Kontext dieser Reorganisationsmaßnahmen begann auch eine Auseinandersetzung mit Möglichkeiten zur Bewertung des Entwicklungsstandes für die neu definierten IT-Governance-Prozesse. Eine in anderen Bereichen bereits bewährte Möglichkeit zur Bewertung der Ist-Situation von Organisationen und Prozessen sind Reifegrad-Assessments. Zu Reifegradanalysen im Bereich der IT-Governance in öffentlichen Verwaltungen liegen bislang aber nur wenige Erkenntnisse vor [JGW06, LR05]. Es ist unklar, inwieweit sich die aus der Privatwirtschaft bekannten Assessment-Modelle auch in der öffentlichen Verwaltung anwenden lassen. Ziel des Forschungsprojektes war daher die Beantwortung der folgenden Fragestellung:

Wie lässt sich der Reifegrad von IT-Governance-Prozessen am Beispiel der LHM bestimmen?

Um dieses Ziel zu erreichen, wurden die jüngst eingeführten IT-Governance-Strukturen und -Prozesse der LHM analysiert und ein Reifegradinstrument angepasst und erprobt, das Aufschlüsse über den Entwicklungsstand der IT-Governance bei der LHM ermöglichen soll. Als Ausgangspunkt diente das COBIT 5 Process Assessment Model (PAM), um von den Vorteilen eines etablierten IT-Governance-Frameworks in Kombination mit einem standardisierten Assessment-Ansatz (ISO/IEC-Standard 15504 [Is03]) für die Erhebung von Prozessreifegraden profitieren zu können. Der vorliegende Beitrag stellt die Ergebnisse der Fallstudie vor und bietet somit auch für andere öffentlichen Verwaltungen einen Impuls, sich mit Hilfe von Reifegrad-Assessments mit dem Entwicklungsstand ihrer IT-Governance-Prozesse auseinanderzusetzen.

Nach einer Erläuterung der besonderen Stellung der IT-Governance in der öffentlichen Verwaltung folgt die Vorstellung der konkreten Situation der LHM. Anschließend wird das methodische Vorgehen des Forschungsprojekts vorgestellt. Nach einer kurzen Einführung zu Reifegradmodellen im Allgemeinen wird das COBIT 5 PAM, das als Basis für das Reifegradinstrument verwendet wurde, näher betrachtet. Im Anschluss werden die vorgenommenen Anpassungen und die Ergebnisse der Erprobung des Reifegrad-instruments bei der LHM beschrieben. Das Fazit fasst schließlich die wichtigsten Ergebnisse zusammen.

2 IT-Governance in der öffentlichen Verwaltung

2.1 IT-Governance

Für IT-Governance mangelt es an einer einheitlichen Begriffsdefinition [SK13]. Eine populäre Definiton liefern Weill und Ross, die IT-Governance als *„Specifying the decision rights and accountability framework to encourage desirable behavior in the use of IT"* definieren [WR04, S. 8]. Der Standard ISO/IEC 38500, der auch im etablierten IT-Governance-Framework COBIT in seiner aktuellen Version 5 aufgegriffen wird, definiert IT-Governance als *„the system by which the current and future use of IT is directed and controlled"* [Is08, S. 3]. Der Standard identifiziert zudem drei Hauptaufgaben der IT-Governance. Die erste Aufgabe „Evaluate" beschreibt die Auseinandersetzung mit der aktuellen und der zukünftigen Bedeutung und Nutzung der IT im Unternehmen. Die Aufgabe „Direct" umfasst Anweisungen zur Vorbereitung und Implementierung von Strategien, Plänen und Richtlinien, um sicherzustellen, dass die Benutzung der IT den Geschäftszielen entspricht. Die Aufgabe „Monitor" umfasst die Sicherstellung der Einhaltung von Richtlinien und geltender Vorgaben sowie die Überwachung der Leistungsfähigkeit, die wiederum der Aufgabe „Evaluate" als Eingangsgröße dient.

Die IT-Governance leitet sich in der Regel aus der Corporate Governance der betrachteten Organisation ab bzw. ist ein Teilbereich von ihr. Die Governance von Unternehmen und öffentlichen Verwaltungen unterscheiden sich. Öffentliche Betriebe streben nicht nach Gewinn, sondern arbeiten grundsätzlich nach dem Kostendeckungs<-prinzip [WD10, S. 29-30]. Die typische Dreiteilung mit Bundes-, Landes- und Kommunalebene ist konstitutiv für die Governance öffentlicher Verwaltungen in Deutschland. Die Governance von öffentlichen Verwaltungen basiert auf gesellschaftlichen, im Rahmen von Gesetzen verhandelten, Verordnungen. Dementsprechend ist auch die strategische IT-Planung nahezu ausschließlich auf die Erfüllung gesetzlicher Vorgaben ausgerichtet [Jo12]. Sie verfolgen in einem verstärkten Maße das Ziel, Gesetze und Regeln zum Datenschutz und zur Informationssicherheit vorbildlich umzusetzen. Entscheidungen über IT-Projekte sind außerdem als Budget-Entscheidungen Gegenstand politischer Auseinandersetzungen innerhalb und insbesondere zwischen Ressorts. Budget und Termintreue bei der Abwicklung von IT-Projekten spielen wie in der Privatwirtschaft eine bedeutende Rolle, jedoch kann der Projektumfang aufgrund von gesetzlichen Vorgaben nicht so flexibel gehandhabt werden wie es in Unternehmen möglich wäre [Jo12]. Kubicek identifiziert in diesem Kontext eine „Organisationslücke" [Ku92] zwischen IT und Geschäft, d.h. Organisationsstruktur, Prozesse und IT sind in öffentlichen Verwaltungen häufig nicht passgenau aufeinander abgestimmt (für weitergehende Beispiele siehe [Ba90, BW11]).

2.2 Die IT-Organisation der Landeshauptstadt München

Die Landeshauptstadt München (LHM) startete im Jahr 2007 ein stadtweites strategisches Umstrukturierungsprojekt „Münchner IT – Konkrete Umsetzung und Top Priorities" (MIT-KonkreT) zur strategischen Neuausrichtung ihrer gesamten IT. Die durch das Projekt neu entwickelte Organisationsstruktur ist seit Beginn des Jahres 2012 umgesetzt.

Die städtische IT wurde in drei Organisationseinheiten, sogenannte Häuser, aufgeteilt, die jeweils entsprechend ihrer Kernkompetenzen unterschiedliche Aufgaben erfüllen. Die im Folgenden genauer betrachtete Hauptabteilung IT-Strategie, IT-Steuerung und IT-Controlling (STRAC) ist im Direktorium der LHM angesiedelt und häuserübergreifend für alle strategischen IT-Prozesse und die IT-Governance zuständig. Darüber hinaus wurde analog zum Bund die Rolle des IT-Beauftragten eingeführt, die vom Leiter des Direktoriums übernommen wird. In dieser Rolle ist er auch Sprecher des obersten IT-Gremiums der LHM, das IT-Beirat genannt wird.

Im Rahmen der Umstrukturierung wurden auch die IT-Governance-Prozesse neu definiert, die sich zum Teil noch in der Einführungsphase befinden. Die neuen Prozesse sind in einer Prozesslandkarte zusammengefasst. Die Prozesslandkarte wird stetig weiterentwickelt und stellt auch die zuständigen Prozessverantwortlichen und Prozesseigner dar. Darüber hinaus liegen für die meisten Prozesse bereits detailliertere Prozessmodelle und Prozesssteckbriefe vor, in denen Aspekte wie die Prozessziele, Schritte und Rollen für den jeweiligen Prozess festgehalten sind.

3 Methodisches Vorgehen

In dieser Fallstudie wurde ein Instrument zur Erhebung der Reifegrade der IT-Governance-Prozesse basierend auf COBIT 5 PAM angepasst und anschließend bei der LHM angewendet. Das Vorgehen wurde in die folgenden Phasen gegliedert:

1. Vorbereitung und Planung der Entwicklung eines Instruments zur Ermittlung des Reifegrads strategischer Prozesse bei der STRAC.

2. Entwicklung eines Prototyps des Reifegradinstruments.

3. Erprobung bei der STRAC im Rahmen von zwei Workshop-Wellen und Verfeinerung des Instruments.

4. Auswertung und Reflexion der Ergebnisse der Erprobung des Reifegradinstruments.

Das Vorgehen orientiert sich an den Grundsätzen der Aktionsforschung mit dem Ziel, eine praxisnahe Lösung für die konkrete Fragestellung und damit einhergehend aktiv zu Veränderungen in der Organisation beizutragen [Fr98, WH07]. Damit wird bewusst deutlich gemacht, dass die Forschenden aufgrund ihrer Beteiligung am untersuchten Phänomen Stellung beziehen und die Position eines passiven und neutralen Beobachters aufgeben.

4 Reifegrad-Assessments mit dem COBIT 5 PAM

4.1 Reifegradmodelle

Reifegradmodelle sind Instrumente zur Bewertung der Leistungsfähigkeit von Organisationen. Sie definieren Fähigkeitsstufen anhand einer Rangliste, die zum Beispiel von einer Stufe „initial“ über verschiedene Stufen hin zu einer Stufe „optimiert“ reichen kann. Die Stufen werden für Reifegrad-Assessments genutzt, bei denen die Ist-Reife ermittelt und ein gewünschter, in der Zukunft zu erreichender, Zielzustand abgeleitet werden [BKP09, Go09]. Im IT-Management dienen Reifegrad-Assessments insbesondere der Motivation zu einem kontinuierlichen Verbesserungsprozess und der Unterstützung des organisationsinternen Lernprozesses [BF11]. Reifegradmodelle wurden ursprünglich durch die verbreitete Anwendung des Capability Maturity Model (CMM) im Bereich der Softwareentwicklung populär [Eh10, HK11]. Inzwischen gibt es für viele weitere Fähigkeitsbereiche eigene Reifegradmodelle [Po11].

Die Anwendung von Reifegradmodellen ist mit Herausforderungen verbunden [MR09, Po11, SG10]. Zum einen wurde in den vergangenen Jahren eine Vielzahl von Reifegradmodellen in Praxis und Wissenschaft vorgeschlagen. Hieraus ergibt sich die Problematik, ein für den speziellen Anwendungsfall geeignetes Reifegradmodell zu identifizieren und ggf. unter verschiedenen geeigneten Modellen auszuwählen. Zum anderen sind Reifegradmodelle zu einem gewissen Grad generisch gehalten, so dass sie in möglichst vielen Organisationen anwendbar sind. Sie berücksichtigen nicht die Besonderheiten einer einzelnen Organisation. Organisationen sehen daher häufig die Notwendigkeit, die Modelle und zugehörige Assessment-Materialien (wie Formulare und Richtlinien) auf ihre eigenen Bedürfnisse anzupassen [PPG13]. Reifegraderhebungen sollten daher immer mit einem klaren Ziel und Fokus eingesetzt werden, die bei möglichen Anpassungen des zugrundeliegenden Modells entsprechend berücksichtigt werden können.

4.2 COBIT 5 Process Assessment Model

Das weltweit anerkannte IT Governance Framework COBIT wird schon seit über 15 Jahren von der Information Systems Audit and Control Association (ISACA) herausgegeben und weiterentwickelt [Ha06, OBP12]. Im Jahr 2012 wurde die Version 5 des Frameworks vorgestellt. COBIT ist durch seinen generischen Aufbau grundsätzlich auch für öffentliche Verwaltungen geeignet. Teil des COBIT 5 Frameworks ist das neue Process Assessment Model (PAM), das für Reifegrad-Assessments der 37 in COBIT 5 beschriebenen Prozesse verwendet werden kann [Is13a]. Die in den vorherigen Versionen von COBIT enthaltenen prozessspezifischen Reifegradmodelle wurden durch das COBIT 5 PAM abgelöst. Das neue Modell basiert auf dem ISO/IEC-Standard 15504 „Information Technology – Process Assessment“ [Is03], der auch unter dem Namen „SPICE“ bekannt ist.

Das COBIT 5 PAM beinhaltet einen vollständig ausgestalteten Assessment-Ansatz und Hilfestellungen, die auf dem im ISO/IEC 15504 vorgestellten generischen Assessment-

Vorgehen basieren. Begleitend werden in mehreren Publikationen unterschiedliche Klassen von Assessments, umfangreiche Beschreibungen für einzelne Vorgehensschritte des Assessments, sowie weitergehende Informationen zu notwendigen Rollen vorgestellt, die bei der praktischen Durchführung von Assessments unterstützen sollen [Is13b, Is13c]. Das Vorgehen beschreibt die gesamte Durchführung eines Assessment-Projekts von der initialen Planung bis hin zur Erstellung eines Assessment-Reports. Im ersten Schritt, der Initiierung, ist bspw. die Assessment-Klasse auszuwählen. Die Klassen repräsentieren drei Typen von Assessments mit unterschiedlichen Zielsetzungen, Assessoren- und Nachweis-Anforderungen, die von der Überwachung des Fortschritts von kontinuierlichen Verbesserungsprogrammen (Klasse 3), über zuverlässige interne Assessments für internes Reporting (Klasse 2), bis zum Vergleich mit anderen Abteilungen oder Unternehmen reichen (Klasse 1). Darüber hinaus steht mit dem Self-Assessment Guide eine vierte Assessment-Methode zur Verfügung, die aber nur für erste Einschätzungen, nicht aber für belastbare Ergebnisse, gedacht ist.

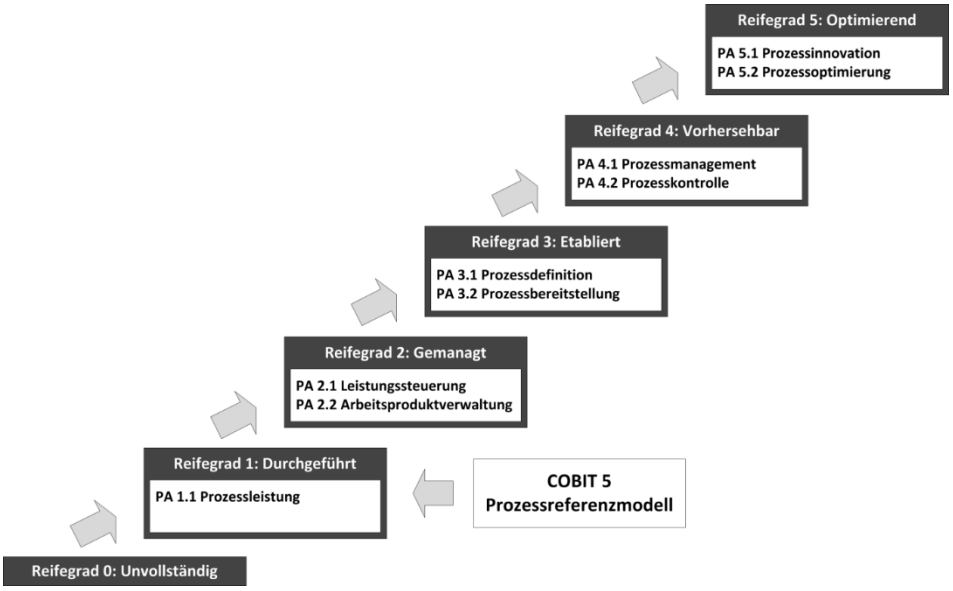


Abbildung 1: COBIT 5 PAM Reifegradmodell

Das Reifegradmodell, das auch Prozessbefähigungsmodell genannt wird, besteht aus sechs Reifegradstufen, denen insgesamt neun Prozessattribute (PA) zugeordnet sind (Abbildung 1). Die Besonderheit des Reifegradmodells ist, dass es neben der Reifegraddimension eine zweite sogenannte Prozessdimension beinhaltet. Die Prozessdimension fügt dem orignär generischen Modell des ISO-Standards 15504 prozessspezifische Anforderungen hinzu, die aus den COBIT Prozessbeschreibungen entnommen werden und zur Erreichung des Reifegrades 1 erfüllt sein müssen. Die Reifegrad-Assessments finden deshalb auf Basis einzelner Prozesse aus dem COBIT 5 Prozessreferenzmodell statt, für die im Vorfeld bei Bedarf eine Zuordnung zu den Prozessen der betrachteten Organisation vorzunehmen ist. Die Bewertungen der

Reifegrade 2 bis 5 basieren ausschließlich auf generischen Prozessattributen und somit für jeden Prozess mit den gleichen Attributen durchgeführt. Die Prozessattribute werden mit Hilfe von Prozessindikatoren beurteilt, die die für die Umsetzung des jeweiligen Prozessattributs notwendigen Maßnahmen beschreiben.

5 Adaption und Anwendung des COBIT 5 PAM

5.1 Adaption

Um die Effektivität der neuen IT-Governance-Prozesse bei der LHM beurteilen sowie weiter verbessern zu können, wurde das COBIT 5 PAM adaptiert. Hierzu wurden die spezifischen Rahmenbedingungen und Anforderungen der STRAC an das Reifegradinstrument berücksichtigt. Insbesondere die Prozessergebnisqualität sollte bei der Messung der Reife berücksichtigt werden. Darüber hinaus bestand die Anforderung, dass durch das Reifegradinstrument konkrete Vorschläge zur Erhöhung der Reife ableitbar sein sollten, um einen kontinuierlichen Verbesserungsprozess etablieren zu können. Vor der Anwendung des COBIT 5 PAM wurden daher folgende Anpassungen des Vorgehens und des Modells an die spezifische Situation der LHM vorgenommen:

- *Anpassung der Assessment-Vorlage:* Das Assessment-Vorgehen wurde an die Assessment-Klasse 3 des COBIT 5 PAM angelehnt. Die angepasste Assessment-Vorlage basiert auf der Self-Assessment-Vorlage aus dem PAM Toolkit. Sie wurde so konzipiert, dass sie bei den Assessments durchgängig ohne Medienbruch angewendet und gleichzeitig zur Protokollierung der Ergebnisse verwendet werden kann. Hierzu wurden ergänzende Felder für Dokumentationen und Nachweise integriert.
- *Detaillierung von Reifegrad 1:* Aufgrund der geringen Zeit, in der die STRAC-Prozesse bis dato implementiert waren, wurde die Wahrscheinlichkeit als hoch eingeschätzt, dass in den ersten Assessments nur geringe Reifegrade erreicht werden. Hierauf wurde mit einer Detaillierung der Bewertungskriterien für die erste Reifgeradstufe reagiert. Dem Prozessattribut „PA 1.1 Prozessleistung“ wurden hierzu insgesamt drei Teilprozessattribute hinzugefügt, die jeweils einzeln bewertet werden: Grundlegende Praktiken, Arbeitsprodukte und Prozessziele. Diese zusätzliche Detaillierung hilft, auch bei einer nicht vollständigen Erreichung des ersten Reifegrads differenzierte Aussagen zum aktuellen Entwicklungsstand machen zu können.
- *Ergänzung von generischen Praktiken:* Zur Unterstützung der Bewertung der Prozessattribute und der untergeordneten Prozessindikatoren der Reifegrade 2 bis 5 wurden so genannte generische Praktiken in die begleitende Dokumentation der Assessment-Vorlage aufgenommen. Die generischen Praktiken beschreiben für jeden der Prozessindikatoren notwendige Einzelschritte der Umsetzung und tragen somit zur Bewertbarkeit der Prozessindikatoren bei. Für die LHM wurden die generischen Praktiken aus der deutschsprachigen Version des „Automotive SPICE“ [Vd07] verwendet, dem ebenso wie dem COBIT 5 PAM der ISO/IEC Standard 15504 zugrunde liegt. Die Übernahme war möglich, da die Attribute

und Indikatoren der Reifegrade 2 bis 5 nur generische, d.h. domänenunabhängige Prozessaspekte adressieren.

- *Erweiterung der Ergebnisdarstellung:* Bei der Ergebnisdarstellung wird die Bewertung der einzelnen Prozessattribute in den Vordergrund gestellt. Aufgrund des Umstrukturierungsprojekts wurden die STRAC-Prozesse vor ihrer ersten Durchführung sehr detailliert geplant und Prozessmodelle erstellt, während das COBIT 5 PAM implizit davon ausgeht, dass derartige Modelle erst aus den bereits gelebten Prozessen hervorgehen. Prozessattribute höherer Reifegrade werden daher evtl. früher erreicht als die niedrigerer Reifegrade. Eine auf vollständig erreichte Reifegrade reduzierte Ergebnisdarstellung würde möglicherweise die bereits erzielten Erfolge unterschlagen.

Die LHM-spezifischen Anpassungen bringen auch Einschränkungen mit sich. So ist die Durchführung von Assessments gemäß der Assessment-Klassen 2 und 1 (bspw. im Sinne eines Benchmarking) unter Umständen nicht mehr bzw. nur noch mit Einschränkungen möglich.

5.2 Anwendung

Das angepasste Reifegradinstrument wurde in Workshops erprobt. Zu Beginn erfolgte eine Auswahl von IT-Governance-Prozessen mit Hilfe der Prozesslandkarte. Anschließend wurde eine Zuordnung der STRAC-Prozesse auf das COBIT 5 Prozessreferenzmodell vorgenommen (Tabelle 1), da die LHM sich nicht strikt an COBIT ausrichtet [Is12].

COBIT 5 Prozesse	**Zugeordnete STRAC Prozesse**
APO02 – Managen der Strategie	IT-Strategieprozess
APO05 – Managen des Portfolios	IT-Vorhabensplanung
APO13 – Managen der Sicherheit	IT-Sicherheitsmanagement
BAI02 – Managen der Definition von Anforderungen	Anforderungsmanagement
MEA01 – Überwachen, Evaluieren und Beurteilen von Leistung und Konformität	Bericht zur IT-Card, Kenngrößendefinition

Tabelle 1: Zuordnung der Prozesse

Die Ergebnisse der Workshops wurden anschließend in einem Assessment Report zusammengeführt, der sich an den von COBIT vorgeschlagenen Inhalten und Strukturen orientiert [Is13b]. Die Reifegradbewertungen wurden nach einem einheitlichen Schema vorgenommen, welches die Anpassung der Reifegradstufe 1, sowie die Darstellung anhand der Erreichung der Prozessattribute, berücksichtigt. In Abbildung 2 wird das verwendete Ergebnisschema am Beispiel des COBIT Prozesses „MEA01 - Überwachen, Evaluieren und Beurteilen von Leistung und Konformität“ gezeigt.

Assessment-Ergebnis: MEA01					
Reifegrad	**Prozessattribut**	**N**	**P**	**L**	**F**
0	-	-	-	-	-
1	*PA 1.1.1 – Grundlegende Praktiken*				F
	PA 1.1.2 – Arbeitsprodukte				F
	PA 1.1.3 – Prozessziele				F
	PA 1.1 – Prozessleistung				F
2	PA 2.1 – Leistungssteuerung			L	
	PA 2.2 – Arbeitsproduktverwaltung				F
3	PA 3.1 – Prozessdefinition				F
	PA 3.2 – Prozessbereitstellung				F
4	PA 4.1 – Prozessmanagement	N			
	PA 4.2 – Prozesskontrolle				
5	PA 5.1 – Prozessinnovation				
	PA 5.2 – Prozessoptimierung				
Erreichter Reifegrad: 2					

Legende:	N	P	L	F
	0-15%	15-50%	50-85%	85-100%

Abbildung 2: Assessment-Ergebnis für den Prozess MEA01

Prozessattribute (PA) können „nicht erreicht“ (N: not achieved, 0-15%), „teilweise erreicht“ (P: partially achieved, 15-50%), „größtenteils erreicht“ (L: largely achieved, 50-85%) oder „vollständig erreicht“ (F: fully achieved, 85-100%) sein. Ein Prozess kann einen Reifegrad nur dann erreichen, wenn alle dem Reifegrad zugeordneten Prozessattribute mindestens „Largely achieved“ und alle Attribute der vorherigen Stufen mit „Fully achieved“ bewertet wurden [Is13a]. Für den betrachteten Prozess ergibt sich daher der Reifegrad 2, auch wenn die PA auf Reifegrad 3 vollständig erreicht sind, da PA 2.1 nicht vollständig erreicht wurde.

Eine Übersicht der Ergebnisse gibt Abbildung 3. Die Prozess-Assessments wurden in begrenzten Zeiträumen durchgeführt, so dass nicht immer eine vollständige Durchführung möglich war. Die Leerräume in der Ergebnisübersicht bedeuten somit, dass diese Prozessattribute nicht bewertet wurden und sollten nicht automatisch als Nichterreichung verstanden werden. Die dargestellten Ergebnisse entsprechen einer Momentaufnahme zum Durchführungszeitpunkt der Prozess-Assessments Mitte des Jahres 2013.

COBIT-Prozess	Reifegrad	0	0				2		3		4		5	
		-	PA 1.1				PA 2.1	PA 2.2	PA 3.1	PA 3.2	PA 4.1	PA 4.2	PA 5.1	PA 5.2
			PA 1.1.1	PA 1.1.2	PA 1.1.3	PA 1.1								
APO02	1	-	L	F	L	L								
APO05	1	-	L	L	L	L	F	F	F	F	N			
APO13	0	-	L	L	P	P	L							
BAI02	3	-	F	F	F	F	F	F	L	L				
MEA01	2	-	F	F	F	F	L	F	F	F	N			

Legende:		N	P	L	F
	N/A	0-15%	15-50%	50-85%	85-100%

Abbildung 3: Übersicht der Assessment-Ergebnisse

Die betrachteten COBIT-Prozesse bzw. die zugeordneten Prozesse weisen im Verhältnis zur Dauer des Bestehens der STRAC und den umfangreichen Folgen der Neustrukturierung für die Organisation bereits eine hohe Reife auf. Insbesondere die Erreichung des ersten Reifegrads durch vier Prozesse ist bereits eine wichtige Errungenschaft, da dies bedeutet, dass diese Prozesse bereits durchgeführt werden können und die Organisation die angestrebten Ergebnisse erzielen kann [Is13b]. Insgesamt bemerkenswert war die große Übereinstimmung der Prozesse mit den Inhalten des COBIT 5 Prozessreferenzmodells, obwohl aktuell keine konsequente Ausrichtung an COBIT vorgesehen ist. Die inhaltlich sehr umfangreichen Prozesse mit hohen Durchlaufzeiten wie der IT-Strategieprozess (APO02) und das IT-Sicherheitsmanagement (APO13) weisen aktuell noch niedrige Reifegrade auf, die aber mit der Komplexität ihrer Entwicklung und Umsetzung erklärbar sind. Die Prozesse der STRAC sind durch die Prozesslandkarte, die Prozessmodelle und die Prozess-Steckbriefe insgesamt gut definiert und beschrieben. Bei konsequenter Weiterführung der umfangreichen Bemühungen im Prozessmanagement ist in absehbarer Zeit eine Erhöhung der Reife der Prozesse der STRAC zu erwarten.

Trotz des Anspruches von COBIT, auch für die öffentliche Verwaltung geeignet zu sein, bestanden inhaltliche Herausforderungen bei der Anwendung des Reifegradinstruments. Dies betraf bei der LHM nicht eindeutig übertragbare Fachbegriffe wie „Business Case“ sowie geforderte Prozessschritte, die mitunter im gegebenen Anwendungsfall nicht direkt umsetzbar oder irrelevant waren. Es konnte in den meisten Fällen bei der Bewertung eine zufriedenstellende Lösung gefunden werden, die notwendigen Klärungen erhöhten jedoch den Aufwand für die Assessments. Ein konkretes Beispiel ist die bei der Erstellung der IT-Strategie (APO02) von COBIT geforderte Berücksichtigung der Unternehmensstrategie. Anders als in privatwirtschaftlichen Unternehmen steht diese nicht zur Verfügung. Aus diesem Grund erfolgt innerhalb des Prozesses zur Erstellung der IT-Strategie eine Aufarbeitung der Strategie der LHM durch Auswertung von Dokumenten der Politik, bestehenden Regelungen und Vorgaben, Erwartungen von Anspruchsgruppen sowie weiteren Quellen. In diesem Kontext wäre auch der Einsatz der COBIT-5-Zielkaskade vorstellbar, die bei der Übersetzung der Organisationsziele in IT-Ziele unterstützen könnte.

Auffallend ist, dass bei der Zuordnung der Prozesse keiner der ausgewählten STRAC-Prozesse einem Prozess aus der in COBIT 5 neu eingeführten Prozessdomäne „Evaluate, Direct and Monitor (EDM)“ zugeordnet wurde. Diese Domäne beinhaltet eigentlich die Hauptaufgaben der IT-Governance, die aber bei der STRAC in die Prozesse aus den

Management-Prozessdomänen integriert wurden. Dieser Umstand liegt mit darin begründet, dass die Zuständigkeit für die IT-Governance bei der LHM innerhalb der IT-Organisation angesiedelt ist. Eine dedizierte Betrachtung in separaten Prozessen, wie sie von COBIT vorgesehen ist, wird durch die LHM als zu aufwändig angesehen. Ein Beispiel hierfür ist der Prozess zur IT Balanced Score Card (IT-Card), ein zentrales Instrument des IT-Controllings in der LHM. Während der Prozess im Wesentlichen dem COBIT-Prozess MEA01 entspricht, werden auch große Teile des COBIT-Prozesses EDM05 aus der Governance-Prozessdomäne durch ihn umgesetzt.

6 Fazit

Das COBIT 5 PAM wurde für die spezifische Situation und Anforderungen der STRAC in der LHM angepasst und angewendet. Es konnte beispielhaft gezeigt werden, dass das COBIT 5 PAM zur Ermittlung des Reifegrads der IT-Governance-Prozesse auch in der öffentlichen Verwaltung eingesetzt werden kann. Die vorgenommenen Anpassungen ermöglichten es, hierbei die spezifischen Anforderungen der LHM als kommunale Verwaltung zu berücksichtigen.

Ein wesentliches Merkmal des Reifegradinstruments ist die Verwendung des COBIT 5 Prozessreferenzmodells zur Bewertung des Reifegrads 1 und somit der inhaltliche Abgleich der STRAC-Prozesse mit anerkannten Good Practices. Dieser Abgleich wurde von den Prozess-Assessment-Teilnehmenden bei der LHM als sehr hilfreich beschrieben. Neben diesen Ergebnissen hat die Adaption und die Anwendung des Reifegradinstruments Prozesswissen innerhalb der STRAC generiert, aus der unter anderem eine Veränderung der STRAC-Prozesslandkarte hervorging. Darüber hinaus sehen sich die Mitarbeiter des IT-Controllings mit Hilfe der erstellten Dokumentation dazu in die Lage, das erstellte Reifegradinstrument in Zukunft eigenständig anzuwenden.

Der Reifegrad der IT-Governance-Prozesse der STRAC ist im Verhältnis zum kurzen Zeitraum des Bestehens der Abteilung bemerkenswert hoch. Auch die inhaltliche Übereinstimmung der Prozesse mit dem COBIT-5-Prozessreferenzmodell waren größer als erwartet. Ein Vergleich der Assessment-Ergebnisse mit anderen öffentlichen Verwaltungen war nicht möglich, da bisher keine entsprechenden vergleichbaren Untersuchungen bekannt und öffentlich verfügbar sind. Mit den vollzogenen Anpassungen halten wir das COBIT 5 PAM grundsätzlich auch im Kontext anderer öffentlichen Verwaltungen für anwendbar. Allerdings sollten bereits Bemühungen im Bereich der IT-Governance und des Prozessmanagements existieren sowie ausreichend Expertise zur Anwendung des COBIT 5 PAM vorhanden sein oder hinzugeholt werden.

Schlussbemerkungen

Wir bedanken uns bei der Landeshauptstadt München für ihre Unterstützung. Dieser Beitrag ist eine erweiterte Version eines Artikels, der in der Fachzeitschrift *IT-Governance* veröffentlicht wurde [GPB14].

Literaturverzeichnis

[Ba90] Barley, S.R.: The alignment of technology and structure through roles and networks. In: Administrative science quarterly Vol. 35, Nr. 1, 1990.

[BF11] Breiter, A.; Fischer, A.: Implementierung von IT Service-Management. Erfolgsfaktoren aus nationalen und internationalen Fallstudien, Xpert.press. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[BKP09] Becker, J.; Knackstedt, R.; Pöppelbuß, J.: Entwicklung von Reifegradmodellen für das IT-Management: Vorgehensmodell und praktische Anwendung. In: Wirtschaftsinformatik Vol. 51, Nr. 3, 2009; S. 249-260.

[BW11] Breiter, A.; Wind, M.: Informationstechnik und ihre Organisationslücken: soziale, politische und rechtliche Dimensionen aus der Sicht von Wissenschaft und Praxis. LIT, Münster, 2011.

[Eh10] Ehsan, N.; Perwaiz, A.; Arif, J.; Mirza, E.; Ishaque, A.: CMMI/SPICE based process improvement. In: Management of Innovation and Technology (ICMIT), 2010 IEEE International Conference on : IEEE, 2010, S. 859-862.

[Fr98] Frank, U.; Klein, S.; Krcmar, H.; Teubner, A.: Aktionsforschung in der WI-Einsatzpotentiale und Einsatzprobleme. In: Wirtschaftsinformatik und Wissenschaftstheorie. Grundpositionen und Theoriekerne. Arbeitsberichte des Instituts für Produktion und Industrielles Informationsmanagement, Nr. 4, 1998; S. 71-90.

[Go09] Gottschalk, P.: Maturity levels for interoperability in digital government. In: Government Information Quarterly Vol. 26, Nr. 1, 2009; S. 75–81.

[GPB14] Grönert, T.; Pöppelbuß, J.; Breiter, A.: Adaption und Anwendung des COBIT 5 Process Assessment Model in der öffentlichen Verwaltung. In: IT-Governance Vol. 8, Nr. 18, 2014, S. 15-21.

[Ha06] Hardy, G.: Using IT governance and COBIT to deliver value with IT and respond to legal, regulatory and compliance challenges. In: Information Security Technical Report Vol. 11, Nr. 1, 2006; S. 55-61.

[HK11] Hertneck, C.; Kneuper, R.: Prozesse verbessern mit CMMI for Services: ein Praxisleitfaden mit Fallstudien. dpunkt.verlag, Heidelberg, 2011.

[Is03] ISO/IEC: ISO/IEC 15504-2:2003. Information technology - Process assessment - Part 2: Performing an assessment. International Organization for Standardization, Switzerland, 2003.

[Is08] ISO/IEC: ISO/IEC 38500:2008. Corporate governance of information technology. International Organization for Standardization, Switzerland, 2008.

[Is12] ISACA: COBIT 5 - Enabling Processes. Information Systems Audit and Control Association, 2012.

[Is13a] ISACA: COBIT 5 - Process Assessment Model (PAM): Using COBIT 5. Information Systems Audit and Control Association, 2013.

[Is13b] ISACA: COBIT 5 - Assessor Guide: Using COBIT 5. Information Systems Audit and Control Association, 2013.

[Is13c] ISACA: COBIT 5 - Self-assessment Guide: Using COBIT 5. Information Systems Audit and Control Association, 2013.

[JGW06] Janssen, M.; Gortmaker, J.; Wagenaar, R.W.: Web Service Orchestration in Public Administration: Challenges, Roles, and Growth Stages. In: Information systems management Vol. 23, Nr. 2, 2006; S. 44-55.

[Jo12] Johannsen, W.: Status der IT-Governance in der öffentlichen Verwaltung. Technische Universität Darmstadt, Darmstadt, 2012. Abgerufen am 06.05.2014 von http://isprat.net/fileadmin/downloads/pdfs/Status_IT-Governance_Referenzmodelle_Abschlussbericht_TUD.pdf.

[Ku92] Kubicek, H.: Die Organisationslücke beim elektronischen Austausch von Geschäftsdokumenten (EDI) zwischen Organisationen. Universität Bremen, Bremen, 1992.

[LR05] Liu, Q.; Ridley, G.: IT Control in the Australian public sector: an international comparison. In: Proceedings of the 13th European Conference of Information Systems, 2005.

[MR09] Mettler, T.; Rohner, P.: Situational Maturity Models as Instrumental Artifacts for Organizational Design. In: Proceedings of the International Conference on Design Science Research in Information Systems and Technology (DESRIST). Philadelphia, 2009.

[OBP12] Al Omari, L.; Barnes, P.H.; Pitman, G.: Optimising COBIT 5 for IT governance: examples from the public sector. In: Proceedings of the ATISR 2012: 2nd International Conference on Applied and Theoretical Information Systems Research (2nd. ATISR2012) : Academy of Taiwan Information Systems Research, 2012.

[OL12] Oliver, D.; Lainhart, J.: COBIT 5: Adding Value Through Effective Geit. In: EDPACS Vol. 46, Nr. 3, 2012; S. 1-12.

[Po11] Pöppelbuß, J.; Niehaves, B.; Simons, A.; Becker, J.: Maturity Models in Information Systems Research: Literature Search and Analysis. In: Communications of the Association for Information Systems Vol. 29, 2011.

[PPG13] Patas, J.; Pöppelbuß, J.; Goeken, M.: Cherry picking with meta-models: a systematic approach for the organization-specific configuration of maturity models. In: Proceedings of the International Conference on Design Science Research in Information Systems and Technology (DESRIST). Helsinki, Finnland, 2013, S. 353-368.

[SK13] Schwertsik, A.R.; Krcmar, H. (Hrsg.): IT-Governance als Teil der organisationalen Governance. Ausgestaltung der IT-Entscheidungsrechte am Beispiel der öffentlichen Verwaltung. Springer Gabler, Wiesbaden, 2013.

[SG10] Solli-Sæther, H.; Gottschalk, P.: The Modeling Process for Stage Model. In: Journal of Organizational Computing and Electronic Commerce Vol. 20, Nr. 3, 2010; S. 279-293.

[Vd07] VDA: Qualitätsmanagement in der Automobilindustrie. Automotive SPICE Prozessassment, Verband der Automobilindustrie e.V., 2007.

[WD10] Wöhe, G.; Döring, U.: Einführung in die allgemeine Betriebswirtschaftslehre. Vahlen, München, 2010.

[We13] Westerfeld, H.: Öffentliche Hand hat IT-Governance-Bedarf. Computerwoche. Abgerufen am 06.05.2014 von http://www.computerwoche.de/a/oeffentliche-hand-hat-it-governance-bedarf,2550151.

[WH07] Wilde, T.; Hess, T.: Research methods in „Wirtschaftsinformatik“ - An empirical study. In: Wirtschaftsinformatik Vol. 49, Nr. 4, 2007; S. 280-288.

[WR04] Weill, P.; Ross, J.W.: IT Governance: How Top Performers Manage IT Decision Rights for Superior Results. Harvard Business Press, Boston, Massachusetts, 2004.

StratIT – Ein Rahmenwerk zu den Inhalten von IT-Strategien. Grundlegung und methodische Fundierung

Konrad Walser, Matthias Goeken, Klaus Schoo, Can Adam Albayrak, Jörg Puchan, Donat Sterren, Jan Thielscher, Raymond Tischendorf, Matthias Knoll, Hans-Jürgen Ley, Jens Bartenschlager, Fritz Pörtig

GI – Fachbereich 5 WI – Fachgruppe Strategisches Informationsmanagement SIM
Leitung der Fachgruppe
Konrad Walser / Matthias Goeken
Berner Fachhochschule / Hochschule der Deutschen
E-Government-Institut / Bundesbank
Morgartenstrasse 2a / Schloss Hachenburg
CH-3000 Bern 22 / D-57627 Hachenburg
konrad.walser@bfh.ch / Matthias.Goeken@bundesbank.de

Abstract: Bis heute ist kein Konsens über die Inhalte von IT-Strategien vorhanden. Daraus folgt eine gewisse Unsicherheit, wenn es gilt, die Themenbereiche einer IT-Strategie festzulegen. Dieser Missstand soll durch den Entwurf eines Rahmenwerks/Frameworks behoben werden. Die Fachgruppe Strategisches Informationsmanagement der Gesellschaft für Informatik (GI-Fachgruppe SIM) hat sich zum Ziel gesetzt, ein entsprechendes Rahmenwerk zu entwickeln. Dieses dient dazu, die Erstellung von IT-Strategien zu unterstützen. Der vorliegende Beitrag beschreibt den aktuellen Stand der Arbeit und zeigt im Sinne einer Diskussionsgrundlage auf, in welche thematischen Perspektiven, Domänen und Elemente sich IT-Strategien gliedern lassen. Darüber hinaus werden die methodische und wissenschaftstheoretische Fundierung sowie der „modus operandi" der Arbeitsgruppe beschrieben. Das Rahmenwerk „StratIT" entsteht in Zusammenarbeit von Forschern und Praktikern.

Vorbemerkung: Die Fachgruppe Strategisches Informationsmanagement der Gesellschaft für Informatik (GI-Fachgruppe SIM) stellt eine Expertengruppe aus Praktikern, Beratern und Wissenschaftlern mit langjähriger Erfahrung im Themenbereich dar. Die Beteiligten entwickeln oder entwickelten IT-Strategien bzw. wurden in entsprechende Entwicklungen einbezogen. Die Mischung, die sich aus unterschiedlichen Rollen und der Herkunft der Beteiligten ergibt, trägt zur Sicherstellung der Betrachtung aus verschiedenen Perspektiven im Sinne einer Triangulation bei. Die Fachgruppe entwickelt StratIT als ein Referenzmodell für Strategieinhalte im Sinne einer Orientierungshilfe zur Vereinfachung der IT-Strategieerstellung. Sie trifft sich dazu viertel- bis halbjährlich.

1 Einleitung

1.1 Problemstellung und Hintergrund

Die Arbeit eines Chief Information Officers (CIO) oder IT-Leiters ist in den letzten Jahren komplexer geworden [HHH08; Ki04; OR12; WA03]. Die Bedeutung der IT hat, etwa gegenüber anderen Leistungsbereichen von Unternehmen, in verschiedenen Branchen überproportional zugenommen. Die Globalisierung verstärkt diesen Trend und die resultierende Komplexität zusätzlich [Bo11]. CIOs oder IT-Leiter haben es in ihrem Alltag mit einer immer größeren Fülle neuer Technologien, Frameworks und Anforderungen zu tun. Überdies ergeben sich daraus auch immer komplexere Sourcing-Konstellationen [GO99], die aus einer strategischen Sicht adäquat zu managen sind. Somit beschäftigen sich CIOs vermehrt mit der Frage, wie sie die Entwicklung der IT im Rahmen der größer werdenden Komplexität so steuern und managen können, dass die IT dem Geschäft den größtmöglichen Nutzen bringt und gleichzeitig regulatorischen Vorgaben nachkommt (Compliance).

Ein zentrales Element der Steuerung der IT stellt neben Unternehmensarchitektur, GRC (Governance, Risk und Compliance) sowie Bebauungsplanung die Entwicklung, Implementierung und Pflege von IT-Strategien als langfristige „Wegweiser“ dar. In der Vergangenheit hat vor allem die praktische Entwicklung und Implementierung von IT-Strategien Aufmerksamkeit erhalten (vgl. [BG09; BG10; Ba11] sowie die dort zitierte Literatur), nicht jedoch die theoretische Untersuchung und Fundierung von deren Inhalten [TPM12]. In der Vergangenheit sind z. B. für die IT-Governance und das IT-Servicemanagements umfangreiche Rahmenwerke entwickelt worden, welche als Best Practices verstanden werden können. Teilweise sind diese von Praktikern und in Berufsverbänden oder unabhängigen Institutionen entstanden (IT Governance Institut, ISACA, OGC, ISO/IEC-Standardisierungsgruppen etc.). Für die IT-Strategie und ihre Inhalte existieren keine vergleichbaren Rahmenwerke und Hilfestellungen. An dieser Stelle ist somit ein Orientierungswunsch der CIOs bezüglich der Inhalte von IT-Strategien auszumachen. Zwar wird IT-Strategien in der Praxis eine große Bedeutung zugemessen [GMS94; WKG97; MPC05; LBZ11] und IT-Strategien sind seit Langem ein Topthema auf der Agenda von CIOs, IT-Leitern und Top-Managern [LBZ11]. Gleichwohl besteht häufig ein hohes Maß an Unklarheit hinsichtlich des Auftrags, wenn eine IT-Strategie zu entwickeln ist.

Die GI-Fachgruppe SIM hat sich zum Ziel gesetzt, die Lücke zu schließen. Zwar gab es in der Vergangenheit Versuche, die Inhalte einer IT-Strategie zu konkretisieren. [Ea96] bspw. unterscheidet Anwendungssystem-, Infrastruktur- und Informationssystemmanagementstrategien und konkretisiert diese weiter in Elemente. Allerdings sind diese Ergebnisse aus heutiger Sicht nicht mehr befriedigend. Aus Forschungssicht stellt die IT-Strategie und deren Inhalte einen „weißen Flecken“ auf der Forschungslandkarte der Wirtschaftsinformatik dar [TPM12; Te13]. Das Thema ist heute zwar von großer praktischer Relevanz. Es genießt jedoch in der akademischen Diskussion eher geringes Interesse. Mit StratIT soll ein Beitrag zur Schließung dieser Lücke geleistet werden. In diesem Zusammenhang stellen sich methodische Fragen etwa folgender Art: Wie methodisch fundiert sind die von Unternehmen und Beratern einge-

setzten Verfahren, insbesondere in Bezug auf thematische Domänen und darin enthaltene Elemente einer IT-Strategie? Welche Rahmenbedingungen sind bei der Konkretisierung einer IT-Strategie und ihrer Inhalte relevant? Wie kann für die IT-Strategie der ideale Themenmix gefunden werden? Wie und woraus leitet sich dieser ab? Lassen sich generische Themen identifizieren, die dann mit Blick auf eine spezifische Unternehmenssituation angepasst werden können (Tailoring)? Welche Umweltfaktoren beeinflussen die Inhalte einer IT-Strategie? Entsprechend ergibt sich die Anforderung, methodisch fundiert die inhaltlichen Elemente einer IT-Strategie zu definieren sowie ein Tailoring des Umfangs der IT-Strategie aus verschiedenen Perspektiven und entsprechend resultierenden Rahmenbedingungen zu unterstützen. Dadurch soll die IT-Strategieerstellung organisations-spezifisch situativ angepasst werden können.

1.2 Motivation

Die Mitglieder der GI-Fachgruppe SIM beabsichtigen, Ziele für StratIT zu formulieren, indem sie die Frage beantworteten: Wie sind die IT und das Business-IT-Alignment durch StratIT bis ins Jahr 2016 verbessert geworden? Die für und mittels StratIT angestrebten Ziele lauten wie folgt:

- StratIT ist eine in Wirtschaft und Wissenschaft anerkannte Hilfestellung zur Entwicklung von IT-Strategien.
- Die IT ist (durch den Einsatz von StratIT) reifer geworden und wird als verlässlicher Partner (an)erkannt und wahrgenommen.
- Die Vorgaben für die IT (bspw. durch eine Business-Strategie) sind vorhanden und hinreichend konkret (Steigerung des Reifegrades des strategischen Denkens in allen funktionalen Bereichen der Unternehmen).
- Durch eine gute IT-Strategie wird die IT-Funktion in Unternehmen nachhaltiger, moderner und professioneller geführt und wahrgenommen.
- Es gibt mehr Unternehmenserfolg durch IT(-Strategie). IT-Strategien werden zu Dokumenten, durch welche die Transformation der IT an-, eingeleitet und sichergestellt wird.

1.3 Zielsetzung

Die Ziele der Fachgruppenarbeit, über die der vorliegende Beitrag berichtet, lauten wie folgt:

- Entwicklung eines Rahmenwerks, das themenorientierte IT-Strategie-Domänen und darin granulare Elemente einer IT-Strategie umfasst; dieses zielt auf eine theoretische Stützung der Praxis
- Theoretische Fundierung der Entwicklung des StratIT-Rahmenwerks und damit verbunden die konzeptionelle Verortung und Fundierung der Entwicklung selbst.

2 Methodische Fundierung der StratIT-Entwicklung

Im Folgenden wird zunächst die methodische Fundierung konkretisiert. Ansätze, die zur Fundierung herangezogen werden, sind der Design-Science-Ansatz und das Method Engineering [HMP04; PTR07] sowie die Begründung mittels Wahrheitstheorien [Fr07; La07].

2.1 Wahrheitstheorien und Begründungsverfahren

Zunächst stellt sich die Frage, wie die Bildung und Entwicklung eines Domänen- und Elementenrasters für IT-Strategiedokumente begründet werden kann. Eine Position in Anlehnung an [Fr07] lautet wie folgt: Konstruktionsorientierte Forschung sollte eine überzeugende wissenschaftstheoretische Fundierung vorweisen, was die Betonung der zentralen Merkmale wissenschaftlicher Forschung – Abstraktion, Originalität und Begründung – empfiehlt. Die Begründung von Forschungsergebnissen stellt dabei eine besondere Herausforderung dar. Für das vorliegende Vorhaben erscheint eine umfassende Begründung durch einen formalen Beweis oder durch eine empirisch bestätigte gehaltvolle Theorie nicht als Option. Daher bleibt nur die Möglichkeit, die Annahmen, die dem Entwurf der Konstruktion zugrunde liegen, explizit zu machen (Transparenz). Dazu gehören Annahmen, die Anforderungen, Entwurfsentscheidungen wie auch die Evaluation betreffen. Für jede Annahme ist die bestmöglich nachvollziehbare Begründung anzustreben. Hier bieten sich die folgenden drei Ansätze an [Fr07; La07]:

- Ausgehend von der Korrespondenztheorie lassen sich folgende Kriterien für die Rechtfertigung der Gültigkeit anführen: Übereinstimmung mit der (wahrgenommenen) Realität bzw. Erfahrung, empirische Bestätigung, Nachweis der erfolgreichen Umsetzung oder Implementierung in der Praxis z. B. anhand von Fallstudien. Für StratIT kann eine Übereinstimmung mit der Realität im ersten Schritt dadurch sichergestellt werden, dass sich die GI-Fachgruppe SIM aus Personen mit praktischen Erfahrungshintergründen zusammensetzt (IT-Strategie-Forschungs-, -Schulungs-, -Beratungs- oder -Umsetzungserfahrung). Diese Erfahrungen fließen in die Entwicklung von StratIT ein und werden darin im Hinblick auf die IT-Strategie-Artefaktentwicklung gekoppelt. Eine weiterführende empirische Bestätigung ergibt sich durch die angestrebte Validierung des Rahmenwerks ausgehend von weiteren Fallstudien, CIO-Interviews, Experten-Interviews, studentischen Arbeiten etc. Der Nachweis der Anwendbarkeit und Umsetzung oder Implementierung von StratIT in der Praxis kann zwar nicht über Konferenzen abgeholt werden, jedoch können über Peers wesentliche Hinweise hinsichtlich Verbesserungen und Einordnung erwartet werden. In Anlehnung an den Design-Science-Ansatz sind Ergebnisse aus dem Problemlösungsprozess zu publizieren, damit sie der Knowledge Base hinzugefügt werden können.
- Ausgehend von der Kohärenztheorie lassen sich folgende Kriterien für die Rechtfertigung anführen: Erforderlich ist die Stimmigkeit der thematischen Domänen und Elemente von IT-Strategien im Kontext bestehender Konzepte oder Theorien. Eine bestehende Theorie, aus der sich die Domänen und Elemente ableiten ließen,

ist nicht bekannt. Es ist aber mindestens eine widerspruchsfreie Einbettung in anerkanntes Wissen zu IT-Strategien und deren Entwicklung herzustellen. Darüber hinaus ist eine Einbettung in die laufende Forschung zu IT-Strategien im Bereich Informationsmanagement in der Wirtschaftsinformatik anzustreben.

– Ausgehend von der Konsenstheorie lassen sich folgende Kriterien für die Rechtfertigung der Gültigkeit anführen: Ein Nachweis der breiten Zustimmung für das aufgestellte Artefakt in der jeweiligen Forschungs- und Anwendungs-Community ist erforderlich: Im vorliegenden Fall ist dies die GI-Fachgruppe SIM sowie erweitert die Wirtschaftsinformatik-Community. Die verschiedenen Versionen des StratIT-Rahmenwerkes werden in einem Delphi-Verfahren innerhalb der Fachgruppe verfeinert und abgestimmt. Durch den Konsens in der Fachgruppe und darüber hinaus wird darüber hinaus Praktikabilität angestrebt und sichergestellt, die im Sinne eines erweiterten Konsenses durch die Validierung durch Praktiker des IT-Strategiemanagements vertieft wird.

2.2 Orientierung am Design-Science-Ansatz

Bei der Entwicklung des StratIT-Rahmenwerkes wird auf der Basis der Design-Science-Research-Methodology (DSRM) von [PTR07] vorgegangen. Deren methodisches Vorgehen sieht sechs Schritte sowie Iterationen vor, um dadurch (mehr) Rigorosität in den Entwicklungsprozess von Artefakten zu bringen. Der vorliegende Beitrag geht nach diesen sechs Schritten vor, die wie folgt lauten:

1. Identify problem and motivation;
2. Define objectives of a solution, what would a better artifact accomplish?;
3. Design and develop artifact;
4. Demonstration, find suitable context, use artifact to solve problem;
5. Evaluation, observe how efficient and effective the design is; iterate back;
6. Communication, scholarly publications, professional publication.

Die ersten beiden Schritte wurden bereits umgesetzt. Die Ergebnisse dazu sind in den vorangegangenen Abschnitten kurz skizziert. Schritt 3 ist in Arbeit und in diesem Beitrag partiell dokumentiert. Die Schritte 4 bis 6 stellen weitere mögliche Arbeitsschritte im Fachgruppenkontext dar.

3 StratIT-Domänen und Elementenraster-Entwicklung

3.1 Literaturrecherche und Erhebung des „State-of-the-Art“

Eine starke Entwicklung der IT-Strategieforschung erfolgte in den 90-er-Jahren des letzten Jahrhunderts insbesondere im angelsächsischen Raum (vgl. u.a. [Ea89; Ga91; HV93; WG98; WP02]). Schwerpunkte hierbei waren die Entwicklung und der Entwicklungsprozess von IT-Strategien, Methoden der strategischen IT-Planung sowie die Abstimmung von IT- und Geschäftsstrategie (Strategic Alignment, u. a. in Anlehnung an [HV93; AJP04; BG09; Al08]). Dazu meinen [TPM12]: “Academic research

so far has focused on the process of strategy formation rather than on the outcome, the resulting IT strategy, and its contents […] only 26% of articles are concerned with the content while 84% address the process of formation". Vergleichbar zeigt die Studie von [Te13], dass insgesamt nur 13 % der untersuchten Publikationen der Frage nachgehen, was unter einer IS-Strategie zu verstehen ist und welche Inhalte sie umfasst oder umfassen soll. Dieser Befund deckt sich u.a. mit den Einschätzungen etwa von [Go99], der feststellt, dass die Forschung sich überwiegend mit dem Prozess der Planung und bislang wenig mit deren Ergebnis, der IS-Strategie [selbst], beschäftigt hat [Te13].

Die folgenden „Strategieinhalte" konnten laut [Te13] identifiziert werden: Wettbewerbsvorteile durch IT, Informationsressourcen, Anwendungssysteme, IT-Infrastruktur, IT-Architektur, IT-Sicherheit, IT-Personal, IT-Organisation sowie IT-Outsourcing. Dabei messen Wissenschaft und Praxis den genannten Inhalten unterschiedliches Gewicht bei. Darüber hinaus lassen sich nach [Te13] zwei Herangehensweisen zu den Inhalten von IT-Strategien unterscheiden: Themenlisten (d. h. die Nennung zu bearbeitender Fragestellungen und Probleme) und Strategiemodelle (d. h. h. Versuche, den Gegenstandsbereich von IS-Strategien sachlogisch-deduktiv zu fassen ([Ea89]; [Ea96]; [Ea00]; vgl. zur Zusammenfassung der Ansätze nach Earl Tabelle 1).

Fragen nach Earl	**Domänen der IT-Strategie**	**Fokus**	**Inhalte**	**Weitere Charakterisierung**
Welche betrieblichen Aufgaben sollen unterstützt werden?	Systemstrategie	Was!	Befasst sich mit den betrieblichen Anwendungsfeldern, in denen IT eingesetzt werden soll, Anwendungsportfolio, durchzuführende Projekte	Geschäftsbezogen und nachfrageorientiert
Wie sollen diese Aufgaben durch IT unterstützt werden?	Technologiestrategie	Wie!	Befasst sich mit der Implementierung der für die Anwendungen einzusetzenden Technologien und die bei der Implementierung zu beachtenden Grundsätze	Implementierungsorientiert und technologiefokussiert
Wer ist für die IT-Unterstützung verantwortlich (einschließlich In-/Outsourcing)?	Managementstrategie	Wofür! Wer!	Befasst sich schließlich mit der Rolle und den Verantwortlichkeiten der IT-Funktion im Unternehmen	Klärt betriebliche Zwecke des IT-Einsatzes und legt benötigte Personalressourcen und Zuständigkeiten für die Aufgabenerfüllung fest.

Tabelle 1: Die drei Strategiearten nach Earl ([Ea89]; [Ea96]; [Ea00]).

[Te13] präsentiert überdies einen Vergleich dreier Beispiele von IT-Strategie-Themenlisten. Dabei werden Listen von [DZW91] zur Literaturanalyse, [CAM92] zur Fallstudienanalyse sowie [SL96] zur Methodenanalyse herangezogen. Für diesen Beitrag erfolgt eine abstrahierte Verdichtung der Themen aus den Themenlisten in Tabelle 2. Dabei wird von spezifischen Konkretisierungen so abstrahiert, dass sie aus heutiger Sicht zweckmäßiger erscheinen [Te13].

Obermenge	Ausrichtung der IT-Strategie	Ressourcen und Werkzeuge	Technologie	Wertschöpfung
Elemente	Ziele Industrie- und Branchenentwicklung Strategie-Definitions- und -Evaluations-zyklus Organisatorischer Wandel Planrealisierungs-forderung Unternehmens-Architektur-Management	Ressourcen Entwicklungsprozeduren Personalausstattung(s-planung) Ausstattungs-Planung Finanz-Ausstattungs-Planung Ausbildung(splanung) Stärken/Schwächen der Ressourcen	Technologie-Architektur Hard- und Software(-Planung) Anwendungs-Architektur	Organisation(splanung) Beschaffung Systementwicklung(s-planung) Implementierung(s-planung) Management von Kosten, Nutzen, Risiken IT-Performance-Management (Organisationaler Wandel) Geschäftsarchitektur

Tabelle 2: IT-Strategie-Themenlisten nach [Te13], basierend auf [DZW91], [CAM92] und [SL96].

3.2 Design und Entwicklung des StratIT-Rahmenwerks

Mit den genannten Themenlisten aus der Literatur und den Themen, die weiter oben erwähnt wurden, liegt bereits eine umfangreiche Liste möglicher Inhalte einer IT-Strategie vor. In der Fachgruppe wurde die Entscheidung getroffen, zwischen thematischen Domänen und Elementen zu unterscheiden. Die Sammlung der Domänen bildet einen groben thematischen Bezugsrahmen aus einer Makroperspektive. In einer detaillierteren Betrachtung, der Mikroperspektive, werden thematische Elemente, die in den Domänen platziert werden, definiert. Zur Verbesserung der Übersichtlichkeit und Orientierung werden zum Teil Domänen in sogenannte Perspektiven gruppiert. Die Elemente benennen Inhalte, zu welchen aus Sicht der IT Entscheidungen definiert und in einer IT-Strategie festgelegt sowie kommuniziert werden müssen. Damit ergibt sich folgender Entwurf des StratIT-Rahmenwerks (Abbildung 1).

Im Folgenden werden (die Domänen präsentiert, sowie) die entsprechenden Elemente (darin) erwähnt.

- *Strategische Prinzipien und Guidelines*: Die strategischen Prinzipien und Guidelines finden sich in den von der Fachgruppe untersuchten Fallstudien. Die detaillierteren strategischen Initiativen und Projekte hängen von der grundlegenden strategischen Stoßrichtung, dem Tailoring sowie den IT-Management-Paradigmen ab, die mit der IT-Strategie unterstützt werden sollen. Aktuelle Elemente, welche den Domänen zugeordnet sind, lauten wie folgt: Relation zu Geschäftsstrategie, Relation IT- versus Business-Geschäftsmodell, Innovation, Strategische IT-Ziele, Vision, Mission, Selbstverständnis, Kultur, Business/IT-Alignment, Governance, Risk, Compliance, IT-Security, Planung, Steuerung/Controlling, Kontinuierliches IT-Strategiemanagement.
- *IT-Wertschöpfungsorientierung:* In dieser Perspektive finden sich die Domänen zum Sourcing (Lieferanten), zur IT-Organisation und -Serviceproduktion (intern und extern) sowie zum Kunden(beziehungs)management. Aktuelle Elemente, wel-

che den Domänen zugeordnet sind, lauten wie folgt: Lieferanten-Management sowie Management von IT-Lieferketten (Sourcing, Lieferantenzahl und -art, In- versus Out-Sourcing, Wertschöpfungstiefe der IT), IT-Organisation (IT-Aufbau- versus -Ablauforganisation), Kunden der IT (interne versus externe Kunden der IT).

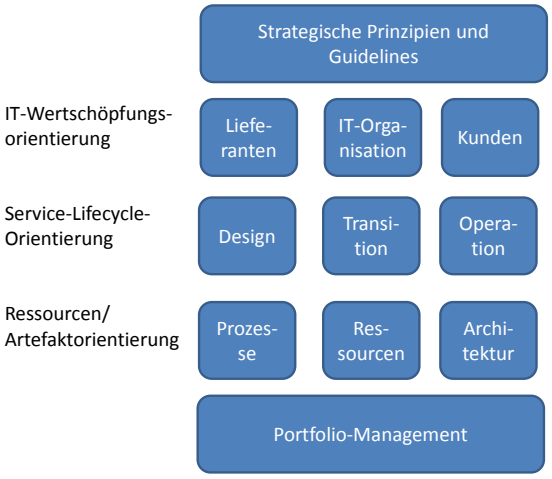


Abbildung 1: Domänen des StratIT-Rahmenwerks.

- *Service-Lifecycle-Perspektive*: Basierend auf dem klassischen Plan-Build-Run, in Abbildung 1 in Anlehnung an die Information Technology Infrastructure Library (ITIL) mit Design, Transition, Operation bezeichnet. Hier können z.B. die aus situativer oder strategischer Sicht relevanten IT-Servicemanagement-Prozesse ausgeführt werden, die in der Strategieperiode eingeführt, verbessert, geprüft werden sollen.
- *Ressourcen/Artefaktorientierung*: Mit u.a. Prozessen, Ressourcen, Architektur. Innerhalb dieser Perspektive kann weiter unterschieden werden in Architektur, nach Ebenen des Architekturmanagements (u.a. Geschäfts-, Anwendungs- und Technologie-Architektur) oder nach Aussagen zum Prozess des Architekturmanagements. Aktuelle Elemente, welche der Domäne Ressource zugeordnet sind, lauten wie folgt: Personal, Finanzen, Infrastruktur, Know-how und Kompetenzen, Information, Daten, Kommunikation.
- *Portfolio-Management*: Besonders in IT-Strategien für (kleinere Unternehmen oder) einzelne Geschäftseinheiten finden sich konkrete Pläne und Budgets zu strategischen Programmen und Projekten. Aktuelle Elemente, welche der Domäne Programme und Elemente zugeordnet sind, lauten wie folgt: Programm-, Projekt-, Service-, Ressourcen-Portfolio, -Prinzipien, -Entscheide, -Planung sowie die Bebauungsplanung.

Die Aufzählung der Perspektiven, Domänen und Elemente erhebt keinen Anspruch auf Vollständigkeit und kann erweitert werden, u. a. um ein Tailoring der Entwicklung des Strategiedokumentes zu unterstützen. Es stellt sich die Frage, wie die verschiedenen Domänen mit den in der Literatur und der Fachgruppe zusammen getragenen Elementen gekoppelt werden können. Geplant ist, die Elemente und ihre Einordnung in das Raster in der nahen Zukunft mittels einer Delphi-Studie zu verifi-

zieren. Dies erfolgt iterativ in mehreren Runden. Zwischen jeder Iteration wird über die Ergebnisse der vorigen Iteration – d.h. Änderungen oder Erweiterungen am StratIT-Rahmenwerk – informiert. Die Iterationen enden, sobald es keine gravierenden Änderungen oder Erweiterungen pro Iteration mehr gibt, also ein weitreichender Konsens erreicht ist.

3.3 Elementenbeschreibung von StratIT

Um eine vergleichbare und konsistente Beschreibung verschiedener Strategieelemente sicherzustellen, ist die Definition einer für das gesamte Rahmenwerk gültigen Struktur vorgesehen, die sieben Beschreibungsdimensionen eines Elementes definiert (Vgl. Abbildung 2).

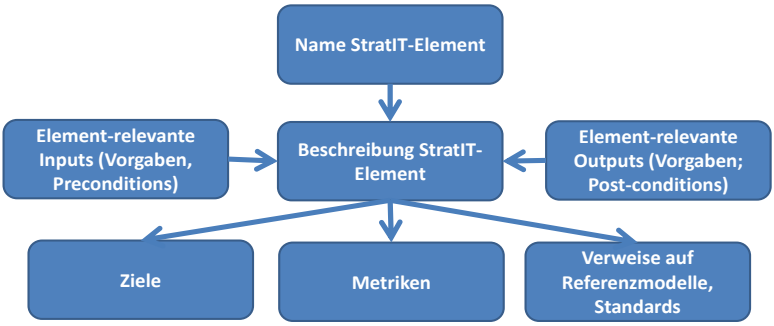


Abbildung 2: Beschreibungsbereiche für das StratIT-Metamodell.

In Anlehnung an das COBIT-Rahmenwerk [ISA12a; ISA12b] werden die folgenden Beschreibungsdimensionen für StratIT-Elemente vorgesehen (StratIT-Metamodell): Name des StratIT-Elements, textliche Beschreibung des Elements, Ziele und Metriken je Element, Inputs im Sinne von Vorgaben (Preconditions), Outputs im Sinne von Ergebnissen (Post-Conditions), sowie Verweise auf Referenzmodelle. Die Beschreibung der einzelnen Elemente von StratIT erfolgt nach diesem Metamodell und ist derzeit in Arbeit.

3.4 Demonstration der Nützlichkeit der Problemlösung und Evaluation

Diese der Entwicklung nachgelagerten Schritte stehen zurzeit noch aus. Der Nutzen wird u. a. anhand der Ziele zu messen sein, ebenfalls ausgehend von der Praktikabilität des Rahmenwerks und dessen Zusatzdokumenten. Ein wesentlicher Beitrag zur Nützlichkeit von StratIT ist die Anpassbarkeit an die Rahmenbedingungen und Ziele, für welche die IT-Strategie entwickelt wird. Die Notwendigkeit dafür wurde in der Fachgruppe betont. Darüber hinaus ist – im Sinne der Konsenstheorie (siehe Kapitel 2.1) – durch die Anwendung des Delphi-Verfahrens beabsichtigt, mittels Rückkopplungen und mehreren Iterationen die Verlässlichkeit des erstellten Artefakts zu erhöhen. Des Weiteren ist eine konzeptionelle Evaluation geplant, bei der das Entwicklungsergebnis den in der Literaturrecherche identifizierten Strategiegegenständen

(siehe Kapitel 3.1) gegenüber gestellt wird. Überdies ist beabsichtigt, StratIT in Fallstudien und studentischen Abschlussarbeiten zur Grundlage zu machen, um seine Anwendbarkeit und Nützlichkeit besser beurteilen zu können.

4 Vergleich, Zusammenfassung und Ausblick

Der vorliegende Beitrag ist als Arbeitsbericht und erster Beitrag zur Konkretisierung des StratIT-Rahmenwerks zu verstehen. Er berichtet im Sinne eines Research-in-Progress-Papers von der Entwicklung eines Rahmenwerks mit thematischen Domänen und granularen Elementen für IT-Strategien. Dieses Entwicklungsvorhaben der GI-Fachgruppe SIM wird im vorliegenden Beitrag mit einem methodologischen und wissenschaftstheoretischen Fundament versehen, das auf dem Design-Science- sowie dem Method-Engineering-Ansatz und Wahrheitstheorien basiert. Ausgehend davon wird eine Entwurfsversion präsentiert. Diese besteht aus Perspektiven, welche Domänen motivieren. In diese werden die IT-Strategieelemente eingeordnet. Darüber hinaus wird das Forschungsvorhaben motiviert und es erfolgt eine Problembeschreibung sowie eine Beschreibung der Ausgangssituation.

StratIT unterscheidet die folgenden Domänenbereiche: Wertschöpfungsorientierung, Service-Lifecycle-Orientierung, Funktionale oder Objekt-orientierte Ableitung von thematischen Elementen (Prozesse, Ressourcen, etc.). Hinzu kommen strategische Leitlinien, IT-GRC (Governance, Risk, Compliance) sowie Portfolio-Management. Ebenfalls neu ist die Idee eines Tailorings von StratIT vor der Erstellung einer konkreten IT-Strategie mit Blick auf unternehmensspezifische Gegeben- und Besonderheiten. Dies erfolgt zum Beispiel vor dem Hintergrund von Unterschieden in der Größe und Rolle der IT in Unternehmen und Unterschieden in deren Wertschöpfungskettenabdeckung.

Im Sinne eines Ausblicks ist weitere Forschungsarbeit zu folgenden Themenbereichen erforderlich: Verifizierung des Nutzens des StratIT-Rahmenwerks für die Praxis, was durch Fallstudien, Validierung durch empirische Forschung sowohl qualitativ wie quantitativ erfolgt, z.B. zur Gruppierung von Elementen im Rahmenwerk (Delphi-Studie), sowie die Methodenentwicklung für ein (situatives) Tailoring von StratIT.

Die folgenden Schritte sind geplant und in Arbeit: Validierung des Rahmenwerks durch existierende IT-Strategien aus der Praxis, Validierung und Erweiterung des StratIT-Frameworks durch weitere Forschungsarbeiten, Überarbeitung und wo nötig Erweiterung des bestehenden Domänen- und Elementenrasters basierend auf weiteren Vergleichen der theoretischen und praktischen Ergebnisse aus Sicht des StratIT-Rahmenwerks.

Literaturverzeichnis

[AJP04] Avison, D.; Jones, J.; Powell, P.; Wilson, D.: Using and Validating the Strategic Alignment Model, Journal of Strategic Information Systems, 13 (2004) 3, S. 223-246.

[Al08] Albayrak, C.A. (2008): IT Strategien planen und umsetzen. Der IT Manager, Management Circle, Eschborn.

[Ba11] Bartenschlager, J. (2011): Implementing IT strategy – Laying a foundation. In: Tagungsband der INFORMATIK 2011 – Lecture Notes in Informatics, Band P192, Bonn.

[BG09] Bartenschlager, J.; Goeken, M. (2009): Designing Artifacts of IT-Strategy for Achieving Business/IT Alignment. In: AMCIS 2009 Proceedings. Paper 494.

[BG10] Bartenschlager, J.; Goeken, M. (2010): IT Strategy Implementation Framework – Bridging Enterprise Architecture and IT Governance. In: Online-Proceedings der AMCIS 2010, Paper 400. Auf: http://aisel.aisnet.org/amcis2010/400 (Aufruf per 2014-03-30).

[BBK05] Boddy, D.; Boonstra, A.; Kennedy, G. (2005): Managing information systems: strategy and organization, Pearson Education, Harlow.

[Bo11] Borhoff, T. (2011): The Role of ICT in the Globalization of Firms. Auf: http://www.davidpublishing.com/davidpublishing/Upfile/2/28/2012/20120228755157 61.pdf (Aufruf per 2014-02-16).

[CAM92] Conrath, D.W.; Ang, J.K,; Mattay, S.: Strategic planning for information systems: a survey of Canadian organizations. Informatica 30 (1992) 4, S. 364-378.

[DZW91] Das, S.R.; Zahra, S.A.; Warkentin, M.E. (1991): Integrating the content and process of strategicMIS planning with competitive strategy. In: Decision Sciences 22 (1991) 5, S. 953-984.

[Ea89] Earl, M.J. (1989): Management strategies for information technology. Prentice Hall, Essex

[Ea96] Earl, M.J. (1996) Integrating IS and the organization. In: Earl MJ (Hrsg.) Information management: the organizational dimension. Oxford University Press, Oxford, S 485-502

[Ea00] Earl, M.J. (2000): Every business is an information business. In: Marchand D.A.; Davenport, T.H. (Hrsg.): Mastering information management. Financial Times Prentice Hall, London, S 16-22

[Fr07] Frank, U. (2007): Ein Vorschlag zur Konfiguration von Forschungsmethoden in der Wirtschaftsinformatik. In: Lehner F.; Zelewski, S. (Hrsg.): Wissenschaftstheoretische Fundierung und wissenschaftliche Orientierung der Wirtschaftsinformatik. GITO, Berlin, S. 158-185.

[Ga91] Galliers, R.D.: Strategic information systems planning: myths, reality and guidelines for successful implementation. In. European Journal of Information Systems 1 (1991) 1, S. 55-64.

[Ga11] Galliers, R.D. (2011): Further developments in information systems strategizing: unpacking the concept. In: Galliers, R.D.; Currie, W. (Hrsg.): The Oxford handbook of management information systems: critical perspectives and new directions. University Press Oxford, Oxford, S. 229-245

[GMS94] Galliers, R.D.; Merali, Y.; Spearing, L.: Coping with information technology? How British executives perceive the key information systems management issues in the mid 1990s. In: Journal of Information Technology 9 (1994) 3, S. 223-238.

[GO99] Gallivan, M.J.; Oh, W. (1999): Analyzing IT Outsourcing Relationships as Alliances among Multiple Clients and Vendors. In: Proceedings der 32. Hawaii International Conference on Systems Science, Hawaii.

[Go99] Gottschalk, P.: Implementation predictors of strategic information systems plans. In: Information and Management 36 (1999) 2, S. 77-91.

[HHH08] Holtschke, B.; Heier, H.; Hummel, T. (2008): Quo vadis CIO?. Springer, Berlin.

[HMP04] Hevner, A. R.; March, S. T.; Park, J.: Design Science in Information Systems Research. In: MIS Quarterly 28 (2004) 1/ March, S. 75-105.

[HV93] Henderson, J.C.; Venkatraman, N. (1993): Strategic alignment: leveraging information technology for transforming organizations. In: IBM Systems Journal 32 (1993) 1, S. 4-16.
[ISA12a] COBIT 5 – A Business Framework for the Governance and Management of Enterprise IT, ISACA, Rolling Meadows.
[ISA12b] COBIT 5 – Enabling Processes, ISACA, Rolling Meadows.
[Ki04] Kirchmann, E.M.W. (2004): Die gewandelte Rolle des CIO. Auf: http://home.arcor.de/edgar.kirchmann/Attachments/Rolle%20des%20CIO.pdf (Aufruf per 2014-02-16)
[La07] Lange, C. (2007): Ein Bezugsrahmen zur Beschreibung von Forschungsgegenständen und -methoden in Wirtschaftsinformatik und Information Systems – ICB Research Report. Auf: www.wi-inf.uni-due.de/FGFrank/documents/Arbeitsberichte_ICB/No1 .pdf (Aufruf per 2013-03-31).
[LBZ11] Luftman, J.; Ben-Zvi, T.: Key issues for IT executives 2011: cautious optimism in uncertain economic times. MIS Quarterly Executive 10 (2011) 4, S. 203-212.
[Le93] Lehner, F. (1993): Informatik-Strategien: Entwicklung, Einsatz und Erfahrungen. Hanser, München.
[MPC05] McGee, K.; Plummer, D.C.; Comport, J.; Tully, J.; Hafner, B.; Mahoney, J.; Fenn, J.; Morello, D.; Mc-Donald, M.P.; Prentice, S.; Kutnik, D. (2005) The gartner scenario 2005: IT leaders' next big decisions. Gartner Inc.
[OR12] Otterbach, N.; Rose, D. (2012): CIO der Zukunft, Themen, Aufgaben, Qualifikation. Auf: http://www.hdm-stuttgart.de/~no012/resources/CIO-der-Zukunft.pdf (Aufruf per 2014-02-16).
[PTR07] Peffers, K.; Tuunanen, T.; Rothenberger, M.S.; Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. In: Journal of Management Information System 24 (2007/2008) 3, S. 45-77.
[SL96] Lederer, A.L.; Salmela, H.: Toward a theory of strategic information systems planning. In: Journal of Strategic Information Systems 5 (1996) 3, S. 237-253.
[SvR97] Smits, M.T.; van der Poel, K.G.; Ribbers, P.M.A.: Assessment of information strategies in insurance companies in the Netherlands. Journal of Strategic Information Systems 6 (1997) 2, S. 129-148.
[Te13] Teubner, A.: Informationssystem-Strategie – Theorie, Praxis und Herausforderungen an die Forschung. In: Wirtschaftsinformatik (2013) 4, S. 239-255.
[TPM12] Teubner, R.A.; Pellengahr, A.R.; Mocker, M. (2012): The IT strategy divide – Professional Practice and Academic Debate, Working Paper No. 12, European Research Center for Information Systems. Becker, J. et al. (Hrsg.), Universität Münster, Münster.
[WA03] Weiss, J.W.; Anderson, D. (2003): CIOs and IT Professionals as Change Agents, Risk and Stakeholder Managers: A Field Study. In: Proceedings 36. Hawaii International Conference on System Sciences (HICSS'03).
[WG98] Ward, J.; Griffiths, P. (1998) Strategic Planning for Information Systems. Wiley, Chichester.
[WKG97] Watson, R.T.; Kelly, G.G.; Galliers, R.D.; Brancheau, J.C.: Key issues in Information systems management: an international perspective. Journal of Management Information Systems 13 (1997) 4, S. 91-115.
[WP02] Ward, J.; Peppard, J. (2002): Strategic Planning for Information Systems. Wiley, Chichester.

Der kostenoptimale Grad funktionaler Redundanz in IT-Anwendungslandschaften

Dr. Thomas Widjaja

Technische Universität Darmstadt
Hochschulstraße 1
64289 Darmstadt
widjaja@is.tu-darmstadt.de

Abstract: Für die Entstehung von unerwünschter funktionaler Redundanz in IT-Anwendungslandschaften existieren zahlreiche Gründe: So können beispielsweise dezentrale Entscheidungsstrukturen oder Unternehmenszusammenschlüsse dazu führen, dass unterschiedliche IT-Anwendungen eine identische fachliche Funktion anbieten. Anders als bei geplanter funktionaler Redundanz – wie beispielsweise bei einem Backup-System – verursacht diese unerwünschte Redundanz Mehrkosten ohne Nutzen zu stiften. Die Reduktion dieser Redundanzen ist jedoch auch mit Kosten für die Konsolidierung der redundanten Anwendungen verbunden. In diesem Artikel wird ein mathematisches Modell vorgestellt, dass IT-Architekten dabei unterstützt, den kostenoptimalen Grad an funktionaler Redundanz in der IT-Anwendungslandschaft zu bestimmen. Das Modell wird im Rahmen eines realen Anwendungsbeispiels evaluiert und dient als Grundlage für praxisrelevante Handlungsempfehlungen und theoretische Implikationen.

1 Einleitung

IT-Anwendungslandschaften sind komplexe Systeme: Eine Vielzahl von Komponenten ist durch unterschiedliche Relationen (Schnittstellen) untereinander verbunden. In einem solchen System existieren oftmals funktionale Redundanzen, d. h. eine identische fachliche Aufgabe wird durch mehr als eine Komponente abgebildet. Diese funktionale Redundanz ist beispielsweise die Folge von unkontrolliertem und unkoordiniertem historischen Wachstum der IT-Landschaft oder entsteht durch externe Einflüsse wie Unternehmenszusammenschlüsse.

Aus der Literatur ist bekannt, dass unnötige funktionale Redundanz[1] ein wesentlicher Treiber für die Betriebskosten der IT-Landschaft ist und, dass dieses Problem durch Standardisierungsmaßnahmen angegangen werden kann. [BY06] Einige Quellen legen sogar den Schluss nahe, dass in der Erkennung und dem Management dieser unerwünschten funktionalen Redundanzen ein wesentlicher Wert des kompletten Enterprise Architecture Managments (EAM) liegt. [z. B. EBD11, Ta11, Za87] In der Praxis zeigt

[1] In diesem Artikel wird ausschließlich die unerwünschte funktionale Redundanz behandelt. Dabei bleibt unberührt, dass es auch gewünschte funktionale Redundanz innerhalb einer IT-Landschaft geben kann. Beispiele für solche gewollte Redundanz sind Dopplungen von Komponenten, um die Ausfallsicherheit zu erhöhen.

sich jedoch, dass die vollständige Reduktion der identifizierten unnötigen Redundanz weder sinnvoll noch möglich ist. Ein Grund dafür ist, dass die Reduktion von bereits existierender funktionaler Redundanz mit erheblichen Kosten verbunden ist. Da Kostenmanagement eine wesentliche Anforderung an CIOs und IT-Architekten ist [GP12, Li06], stellt sich insbesondere für sie die Frage, wie mit existierender Redundanz in den IT-Landschaften umzugehen ist und entsprechend, wie hoch der kostenoptimale Grad an Redundanz in der Zielarchitektur [Ai09] sein sollte.

In der aktuellen Literatur fehlt zum einen die detaillierte Konzeptualisierung des zugrundeliegenden Trade-offs und zum anderen wird bisher kein Entscheidungsunterstützungsmodell vorgeschlagen, das die gegenläufigen Kosteneffekte (insb. „unnötige" Betriebskosten redundanter Anwendungen vs. Kosten der Redundanzreduktion - siehe auch Abbildung 1) zu einer Lösung integriert. Bisher ist es oftmals eine in der Praxis akzeptierte Lösung, in der Ziel-Architektur [Ai09] die identifizierten unerwünschten funktionalen Redundanzen so weit wie möglich zu reduzieren – bis zu welchem Grad dies jedoch ökonomisch sinnvoll ist, bleibt dem „Bauchgefühl" der IT-Architekten überlassen. Die in diesem Artikel vorgeschlagene Konzeptualisierung des zugrundeliegenden Trade-Offs ist damit sowohl für Praxis als auch Wissenschaft relevant, da zum einen ein wesentlicher Mechanismus der Wertgenerierung von EAM detaillierter beleuchtet wird und zum anderen die Grundlagen für ein praxisrelevantes Entscheidungsunterstützungssystem erarbeitet werden.

In Summe zielt der vorliegende Artikel darauf ab, folgende Forschungsfrage zu beantworten: Wie können IT-Architekten bei der Ermittlung des kostenoptimalen Grads an funktionaler Redundanz in einer Ziel-Architektur unter Berücksichtigung der gegenläufigen Kosteneffekte unterstützt werden? Der Artikel ist wie folgt strukturiert: Im nächsten Kapitel wird das beschriebene Problem theoretisch eingeordnet und die bereits existierende Forschung zum Thema aufgearbeitet. In Kapitel 3 wird die zugrundeliegende Problemklasse definiert. Auf dieser Basis wird in Kapitel 4 eine Formalisierung des grundlegenden Trade-offs als Optimierungsproblem vorgeschlagen. Die vorgestellte Modellformulierung wird in Kapitel 5 im Rahmen eines realen Anwendungsfalls in Kooperation mit einem Praxisunternehmen evaluiert. Nachdem die praktische Nutzbarkeit geprüft ist, werden anschließend in Kapitel 6 die Implikationen sowohl aus praktischer als auch theoretischer Sicht diskutiert.

2 Theoretischer Hintergrund

2.1 IT-Architektur aus Systemtheoretischer Perspektive

In diesem Artikel wird [IEEE00] folgend, eine IT-Architektur als System verstanden, das sich aus Komponenten und deren Relationen untereinander zusammensetzt. Diese Sichtweise entspricht damit der klassischen Systemtheorie [HF69]. Dieses hierarchische System [Si62] kann in verschiedene untereinander verbundene Sub-Systeme untergliedert werden: Im EAM-Kontext werden insbesondere Anwendungs-Architektur, Daten-Architektur und Infrastruktur-Architektur unterschieden (z. B. [RWR06]). Im Folgenden

fokussiert der Artikel auf die Anwendungs-Architektur eines Unternehmens. Damit sind die Komponenten des betrachteten Systems IT-Anwendungen und die entsprechenden Relationen repräsentieren Schnittstellen zum Datenaustausch.

2.2 Der Redundanzbegriff im Kontext von IT-Architekturen

Redundanz, von lateinisch redundantia „Überfülle“, beschreibt das Vorhandensein von eigentlich überflüssigen, identischen oder vergleichbaren Ressourcen. Herbert Simon definiert „If a complex structure is completly **un**redundant [Markierung hinzugefügt]- if no aspect of its structure can be inferred from any other - then it is its own simplest description.“ [Si96, S. 209] Im Kontext von IT-Architekturen wird die Förderung von Wiederverwendung und damit die Vermeidung von Redundanzen in zahlreichen Quellen als einer der wesentlichen Nutzen der EAM-Funktion gesehen. [z. B. EBD11, Ta11, Za87] Im Zusammenhang von IT-Systemen sind insbesondere die in diesem Artikel betrachtete funktionale Redundanz und die Datenredundanz von Bedeutung. Datenredundanz bezeichnet eine Situation, in der ein und dieselbe Information an mehr als einem Ort gespeichert ist. Funktionale Redundanz liegt vor, wenn mehr als eine Komponente innerhalb des Systems (siehe Kapitel 2.1) die gleiche Funktion erfüllt. Da die in diesem Artikel betrachteten Komponenten des Systems IT-Anwendungen sind, kann der Redundanzbegriff konkretisiert werden: Da die Kernaufgabe einer IT-Anwendung die Informationstransformation ist, sind im Sinne dieses Artikels zwei Anwendungen funktional redundant, wenn sie für alle Eingabe-Informationen die gleiche Ausgabe-Informationen liefern. Diese Definition von Redundanz ist angelehnt an die Definition von [BB03] aus dem Prozesskontext: „redundancy-that is, a determination of whether there is more than one way to connect an input to an output.“ (S. 339) Dieses Redundanzverständnis lässt neben der beschriebenen binären Klassifizierung, d. h. „vollständiger Redundanz“ (jede Eingabe-Information wird zu einer identischen Ausgabe-Information transformiert), auch eine graduelle Variante zu, d. h. X % der Eingaben führen zu identischen Ausgaben. Zur Vereinfachung wird die folgende Betrachtung, auf das binäre Verständnis fokussiert – die entsprechenden Limitationen dieses Ansatzes werden jedoch in Kapitel 6 wieder aufgegriffen. Dieses Redundanzverständnis ist nicht kommutativ, d. h. aus „Anwendung A ist redundant zu Anwendung B“ folgt *nicht*, dass auch B redundant zu A ist. So kann beispielsweise die vollständige Funktion einer Anwendung A gleichzeitig von Anwendung B angeboten werden - Anwendung B jedoch darüber hinaus wesentlich mehr Funktionen zur Verfügung stellen. Daher kann zwar A durch B aber nicht B durch A ersetzt werden. Der in diesem Artikel verwendete Redundanzbegriff ist jedoch transitiv - denn aus „A ist redundant zu B“ und „B ist redundant zu C“ folgt, dass A redundant zu C ist.

Im Kontext von IT-Architekturen kann erforderliche und unnötige funktionale Redundanz unterschieden werden. Beispiele für erforderliche Redundanz – deren Ziel meist die Erhöhung der Ausfallsicherheit des Systems ist – sind Backup-Systeme, die bei Versagen der Hauptkomponenten die wesentlichen Funktionalitäten übernehmen. Unnötige funktionale Redundanz – also Redundanz bei der die entstehenden Kosten keinem angemessenen Nutzen gegenüberstehen – kann beispielsweise durch unkoordiniertes Wachstum der IT-Landschaft oder Unternehmenszusammenschlüsse entstehen. [z. B.

LC07] Im Folgenden wird auf diese unnötige funktionale Redundanz als Ziel möglicher Konsolidierungsmaßnahmen fokussiert.

2.3 Reduktion von unerwünschter funktionaler Redundanz

Während Standards die „Entstehung“ von Redundanz verhindern können [BY06], zielt Konsolidierung auf die Reduktion von existierender Redundanz ab. Der Begriff „Konsolidierung“ bezeichnet das Zusammenfassen von Elementen zu einem Ganzen. Entsprechend wird in diesem Artikel unter (Anwendungs-)Konsolidierung das Zusammenfassen von zwei funktional redundanten Anwendungen verstanden. Hierbei ist das Ziel, eine der beiden Anwendungen nicht weiter zu betreiben. Dieses Verständnis des Konsolidierungsbegriffes wird im Kontext des IT-Architekturmanagements als Sonderform der IT-Integration verstanden (siehe z. B. [KWS05]). In der Terminologie von [LC07] entspricht der oben beschriebene Konsolidierungsbegriff der „Choose one“-Integrationsstrategie und [BB03] beschreiben dieses Vorgehen als „Synthesis“-Ansatz.

[CLT12] leiten aus ihrem aktuellen und umfassenden Literaturüberblick ab, dass aktuell eine Theorie der “Integration” in der Wirtschaftsinformatik fehlt. Ähnlich argumentieren [LC07], die in einer Serie von Fallstudien Industrieerfahrungen zum Management von funktionaler Redundanz von Software aggregieren. Da im vorliegenden Artikel ein zentraler Trade-off bei der Konsolidierung von IT-Anwendungen (d. h. einem wesentlichen Teilbereich der IT-Integration) konzeptualisiert und formalisiert wird, bietet der vorliegende Artikel einen ersten Ansatzpunkt in diese Richtung.

3 Die untersuchte Problemklasse: Ermittlung des kostenoptimalen Grades von unerwünschter Redundanz in einer IT-Anwendungslandschaft

Existiert in einer IT-Landschaft unerwünschte funktionale Redundanz, gilt es zu prüfen bis zu welchem Grad es ökonomisch sinnvoll ist, diese zu reduzieren. Die Bestimmung des kostenoptimalen Grades von unnötiger funktionaler Redundanz unterliegt folgendem fundamentalen Trade-Off (siehe Abbildung 1): Auf der einen Seite verursachen die „unnötigen“ Anwendungen und die entsprechenden notwendigen Schnittstellen zusätzliche Betriebskosten und Kommunikationskosten.[2] Auf der anderen Seite ist die Reduktion dieser unerwünschten funktionalen Redundanz mit Konsolidierungskosten[3] und ggf. den Kosten für das Entwickeln neuer Schnittstellen und den entsprechenden Kommunikati-

[2] Betriebskosten sind laufende Kosten und umfassen nach der Klassifikation von [Bu01, S. 26ff] folgende Oberkategorien: Systemkosten, Personalkosten, Fremdleistungskosten und sonstige Kosten. Kommunikationskosten sind in Anlehnung an [Wi11] als laufende Kosten für die Informationsübermittlung zwischen zwei Anwendungen definiert.

[3] Konsolidierungskosten sind einmalige Kosten und sind in Anlehnung an [Bu01S. 26ff] als Anschaffungs- und Herstellkosten, Personalkosten, Fremdleistungskosten, Installations- und Implementierungskosten und sonstige Kosten für die Überführung einer redundanten Anwendung A in Anwendung B sowie die Abschaltung von Anwendung A definiert. Siehe [Bu01S. 26ff] für eine detaillierte Aufschlüsselung der aufgelisteten Oberkategorien. Analog sind auch die einmaligen Entwicklungskosten für neue Schnittstellen definiert.

onskosten verbunden. Das Entwickeln neuer Schnittstellen kann nötig sein, da es nach der Konsolidierung der redundanten Anwendung A in die Anwendung B zu Veränderungen des Informationsbedarfs von Anwendung B kommen kann (da Anwendung B in der Zielarchitektur auch mit den ehemaligen Kommunikationspartnern von A Informationen austauschen muss).

Da es sich bei den betrachteten Größen sowohl um laufende Kosten (z. B. Betriebskosten für Anwendungen) als auch um einmalige Kosten (z. B. Kosten für die Konsolidierung der entsprechenden Anwendungen) handelt, ist es notwendig die laufenden Kosten über einen definierten Betrachtungszeitraum zu diskontieren.

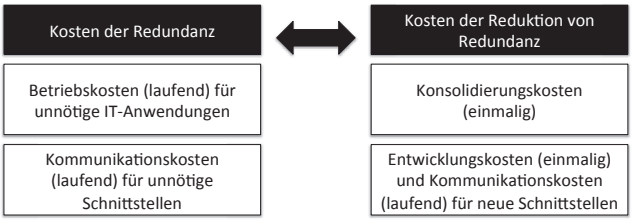


Abbildung 1: Zugrundeliegender Trade-Off

Die Schwierigkeit den kostenoptimalen Konsolidierungsplan zu identifizieren, steigt mit der Komplexität der Redundanz- und den Kommunikationsbeziehungen zwischen den Anwendungen. Eine hohe Komplexität dieser Beziehungen ist beispielsweise zu erwarten, wenn nach einem Merger zweier Finanzdienstleister zwei unterschiedliche technische Umsetzungen zur Unterstützung eines Kreditvergabeprozess existieren. Ein solcher Ausschnitt aus der IT-Anwendungslandschaft ist stark vereinfacht auf der linken Seite von Abbildung 2 als dreigliedriger Prozess (Management der Kundenanfrage, Bonitätsprüfung, Verbuchung) dargestellt. Durch die gerichtete Modellierung der Redundanzbeziehung (gestrichelte Pfeile in Abbildung 2) kann folgende Situation berücksichtigt werden: Während Anwendung A (zum Beispiel das Management der Kundenanfragen in Unternehmen 1) redundant zu Anwendung A' und vice versa ist, gilt dies nicht für Anwendung B und Anwendung B' (die in diesem Beispiel beide die Bonitätsprüfung unterstützen). Zwar ist Anwendung B redundant zu Anwendung B', jedoch nicht umgekehrt (beispielsweise könnten in Anwendung B' wesentlich umfassendere Prüfalgorithmen implementiert sein). Dieses stark vereinfachte Beispiel verdeutlicht zudem, dass eine mögliche Konsolidierung von redundanten Anwendungen neben der Reduktion der Anzahl von Anwendungen auch zur Veränderung der Kommunikationsbeziehungen (ungerichtete durchgezogene Kanten in Abbildung 2) zwischen den Anwendungen führen kann: Für den Fall, dass in der Zielarchitektur Anwendung A' in Anwendung A konsolidiert wird (wie in der Lösung I in Abbildung 2 dargestellt), fallen zum einen Kosten für die Konsolidierung (Zusammenfassen von A und A') und zum anderen Entwicklungs- und Kommunikationskosten für eine neue Schnittstelle zwischen Anwendung A und Anwendung B' an. Demgegenüber können die Betriebskosten für Anwendung A' sowie die Kommunikationskosten für die Schnittstelle zwischen A' und B' eingespart werden. Eine alternative Lösung könnte darin bestehen, die IT-Anwendungen von Unternehmen 1 soweit wie möglich in die IT-Landschaft von Unternehmen 2 aufgehen zu lassen, wie in der Lösung II in Abbildung 2 dargestellt.

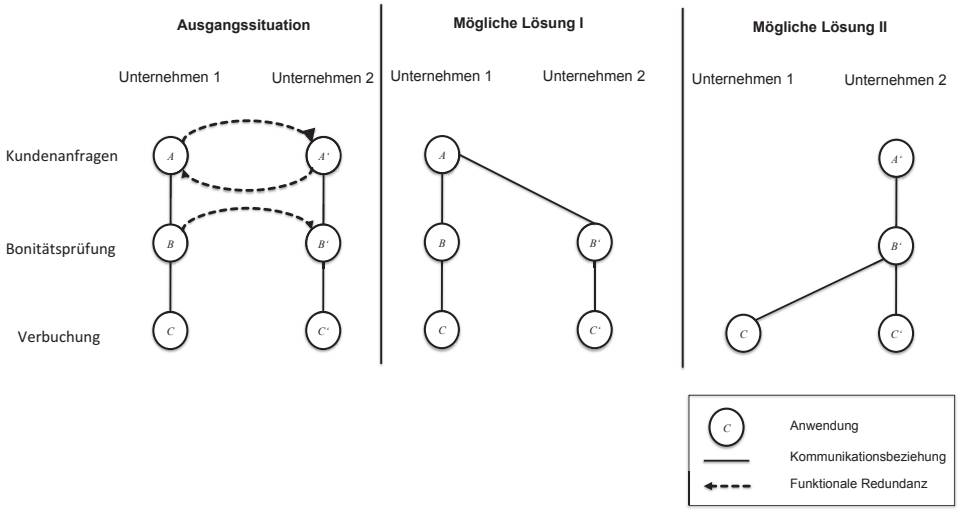


Abbildung 2: Beispielhafte Darstellung der Anwendungslandschaft nach dem Merger von zwei Banken

IT-Architekten wurden bisher bei dem komplexen Entscheidungsproblem, den optimalen Konsolidierungsplan zu identifizieren, nicht durch ein theoretisch fundiertes und formalisiertes Modell unterstützt. Diese Lücke wird im folgenden Kapitel geschlossen.

4 Bestimmung des kostenoptimalen Grads an funktionaler Redundanz als Optimierungsproblem

Zur Formalisierung des Entscheidungsproblems wird die IT-Architektur (d. h., das betrachtete System[4]) als Graph modelliert (siehe Abbildung 3). Die Konten $V = \{1,2,\ldots,n\}$ entsprechen dabei den IT-Anwendungen und es werden zwei Arten von Relationen zwischen den Anwendungen unterschieden: Kommunikationsbeziehungen $E_{Com} = V \times V$ (ausgefüllte Kanten) und Redundanzbeziehungen $E_{Red} = V \times V$ (gestrichelte Pfeile).

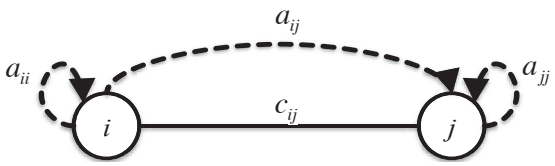


Abbildung 3: Dem Modell zugrundeliegender Graph

[4] Hierbei ist zu beachten, dass die Anwendungslandschaft als System bezüglich verschiedener Abstraktionsebenen betrachtet werden kann [Si62]: Eine Kernbankenanwendung kann zum Beispiel auch selbst als System verstanden werden, das aus unterschiedlichen Sub-Systemen (Modulen) besteht. Für die sinnvolle Anwendung des Modells ist durch den IT-Architekten eine geeignete Abstraktionsebene zu definieren.

Redundanzbeziehungen geben an, dass zwei Anwendungen zusammengefasst werden können und sind mit den Kosten für die Konsolidierung $a_{ij} \in \mathbb{R}_+ \forall (i,j) \in E_{Red}$ gewichtet (siehe Abbildung 3). Als Sonderfall – um eine einfache Modellierung zu ermöglichen - existieren „virtuelle" funktionale Redundanzen, die von jedem Knoten auf sich selbst gerichtet sind (sog. Schlingen) und mit den jeweiligen diskontierten Betriebskosten der Anwendung gewichtet werden.[5] Der Parameterwert $a_{ij} \in \mathbb{R}_+$ repräsentiert also entweder die Betriebskosten der Anwendung *i* (falls $i = j$) oder die Konsolidierungskosten (falls $i = j$).

$$a_{ij} = \begin{cases} Betriebskosten\ von\ Anwendung\ i & | i = j \\ Konsolidierungskosten & | i\ redundant\ zu\ j\ und\ i = j \end{cases} \quad [1]$$

Die ungerichteten Kommunikationsbeziehungen zwischen den Anwendungen repräsentieren Schnittstellen und sind mit diskontierten Kommunikationskosten $c_{ij} \in \mathbb{R}_+ \forall i,j \in V\ und\ i < j$ gewichtet.[6] Zur einfacheren Modellierung wird zudem die Menge $Com_i = \{j | (i,j)\ oder\ (j,i) \in E_{Com}\} \forall i \in V$ definiert, die alle direkten Kommunikationspartner von Anwendung *i* enthält. Zudem ist es möglich, dass Anwendung *i* im Rahmen von Konsolidierungsmaßnahmen neue Kommunikationspartner erhält und entsprechend neue Schnittstellen entwickelt werden müssen – in diesen Fällen müssen neben den Kommunikationskosten für die neue Schnittstelle auch einmalige Entwicklungskosten berücksichtigt werden (in Gleichung [2] als „$Entw. = und\ Kom. = kosten$" bezeichnet). Die Menge $PotCom_i$ enthält alle potenziellen Kommunikationspartner von $i \in V$ in einer Zielarchitektur. Es werden auch diejenigen Anwendungen berücksichtigt, die mit einem aktuellen Kommunikationspartner funktional redundant sind bzw. deren Kommunikationspartner. Damit kann $c_{ij} \in \mathbb{R}_+$ definiert werden als:

$$c_{ij} = \begin{cases} Kommunikationskosten\ zw.\ i\ und\ j & | j \in Com_i\ und\ i < j \\ Entw. = und\ Kom. = kosten\ zw.\ i\ und\ j & | j \in PotCom_i \setminus Com_i\ und\ i < j \\ 2 & | sonst \end{cases} \quad [2]$$

Zur Formulierung des Optimierungsproblems sind zwei binäre Entscheidungsvariablen notwendig. Die Variable $x_{ij} \in \{2,1\}\ \forall (i,j) \in E_{Red}$ nimmt genau dann den Wert 1 an, wenn in der Zielarchitektur die Anwendung *i* in Anwendung *j* konsolidiert wird. Es ist explizit möglich, dass $i = j$ gilt – genau dann, wenn Anwendung *i* „weiterbetrieben" wird.

$$x_{ij} = \begin{cases} 1 & | Anwendung\ i\ in\ Anwendung\ j\ konsolidiert \\ 2 & | sonst \end{cases} \quad [3]$$

Die zweite Entscheidungsvariable $y_{ij} \in \{2,1\}\ \forall i,j \in V$ nimmt genau dann den Wert 1 an, wenn in der Zielarchitektur eine Kommunikationsbeziehung zwischen Anwendung *i* und Anwendung *j* existiert.

[5] Konzeptionell besteht zwischen *i* und *i* also eine funktionale Redundanz.

[6] Da die Kommunikationsbeziehungen ungerichtet sind, ist es ausreichend nur eine Hälfte der Matrix von $c_{ij} \in \mathbb{R}_+$ zu spezifizieren (also für $i < j$).

$$y_{ij} = \begin{cases} 1 & |Anwendung\ i\ tauscht\ mit\ Anwendung\ j\ Informationen\ aus \\ 2 & |sonst \end{cases} \quad [4]$$

In Tabelle 1 sind die verwendeten Symbole sowie deren Bedeutung zusammengefasst.

Tabelle 1: Verwendete Symbole

Symbol	Bedeutung
$\boldsymbol{V = \{1, 2, \dots, n\}}$	Anwendungen
$\boldsymbol{E_{Com} = V \times V}$	Kommunikationsbeziehungen
$\boldsymbol{Com_i = \{j \mid (i,j) oder\ (j,i) \in E_{Com}\} \forall i \in V}$	Kommunikationspartner von *i*
$\boldsymbol{PotCom_i = V}$	Mögliche Kommunikationspartner von *i* in der Zielarchitektur
$\boldsymbol{E_{Red} = V \times V}$	Redundanzbeziehungen
$\boldsymbol{a_{ij} \in \mathbb{R}_+ \forall (i,j) \in E_{Red}}$	Konsolidierungs- bzw. Betriebskosten
$\boldsymbol{c_{ij} \in \mathbb{R}_+ \forall i,j \in V, i < j}$	Entwicklungs- + Kommunikationskosten
$\boldsymbol{x_{ij} \in \{0,1\} \forall (i,j) \in E_{Red}}$	Entscheidungsvariable, die genau dann den Wert 1 annimmt, wenn *i* in *j* betrieben wird. Hinweis: $x_{ii} = 1$ entspricht dem Weiterbetrieb von *i*.
$\boldsymbol{y_{ij} \in \{0,1\}\ \forall i,j \in V}$	Entscheidungsvariable, die genau dann den Wert 1 annimmt, wenn *i* mit *j* Informationen austauscht.
$\boldsymbol{F(x,y) \in \mathbb{R}_+}$	Zielfunktion (Gesamtkosten)
$\boldsymbol{i, i', j, k, m, l \in V}$	Indizes (die jeweils Anwendungen repräsentieren)

Mithilfe dieser Notation kann folgendes Optimierungsproblem formuliert werden:

$$Min!\ F(\boldsymbol{x},\boldsymbol{y}) \coloneqq \underbrace{\sum_{\substack{i,j\ mit \\ (i,j) \in E_{Red}}} x_{ij} \cdot a_{ij}}_{\substack{Konsolidierungs+ \\ und\ Betriebskosten}} = \underbrace{\sum_{\substack{i,j \in V \\ i<j}} y_{ij} \cdot c_{ij}}_{\substack{Entwicklungs+und \\ Kommunikations+ \\ kosten}} \quad [5]$$

Unter den Nebenbedingungen

$$\sum_{\substack{j\ mit \\ (i,j) \in E_{Red}}} x_{ij} = 1 | \forall\ i \in V \quad [6]$$

$$x_{ij} < x_{jj} | \forall\ (i,j) \in E_{Red} \quad [7]$$

$$x_{ij} < y_{jk} | \forall\ (i,j) \in E_{Red}\ und\ k \in \{l | x_{ml} = 1 \wedge m \in Com_i\} \quad [8]$$

$$y_{ij} = y_{ij} | \forall\ i,j \in V \quad [9]$$

$$x_{ij}, y_{kl} \in \{2,1\} | \forall\ (i,j) \in E_{Red}\ und\ k,l \in V \quad [10]$$

Die Zielfunktion berücksichtigt Betriebskosten (Summe 1 für $i = j$), Konsolidierungskosten (Summe 1 für $i = j$) sowie Entwicklungs- und Kommunikationskosten (Summe 2) in der Zielarchitektur. Die erste Nebenbedingungsgruppe (Gleichung 6) garantiert,

dass jede Anwendung entweder weiterbetrieben wird ($i = j$) oder mit einer anderen j Anwendung (zu der i funktional redundant ist) konsolidiert wird. Die zweite Gruppe von Nebenbedingung (Gleichung 7) garantiert, dass sobald eine Anwendung j das „Ziel" einer Konsolidierung ist (d. h. es existiert ein $x_{ij} = 1$), Anwendung j tatsächlich betrieben wird. Die dritte Gruppe von Nebenbedingungen (Gleichung 8) garantiert, dass die Anwendung, die Ziel der Migration ist, alle Kommunikationsbeziehungen der migrierten Anwendung „erbt". Wird also Anwendung i mit Anwendung j konsolidiert, (d. h. $x_{ij} = 1$) muss gelten $y_{jk} = 1$ (wobei der Index k Anwendungen beschreibt, zu denen ehemalige Kommunikationspartner von i (d. h. $m \in Com_i$) konsolidiert wurden). Hierbei zeigt sich u. a. auch, warum die Modellierung der Betriebskosten durch Schlingen im Graph (z. B. $x_{ii} = 1$) nützlich ist. Denn für die Variablen $x_{ml} = 1$ in Gleichung 8 kann somit gelten $m = l$ (genau dann, wenn die Kommunikationspartner von i nicht in anderen Anwendungen konsolidiert wurden). Für den Fall $i = j$ (d. h. die Anwendung i selbst wird weiterbetrieben) garantiert Gleichung 8 zudem, dass $y_{ik} = 1$ gilt, also Anwendung i weiterhin mit den bisherigen Kommunikationspartnern oder deren Konsolidierungszielen kommuniziert.[7] Die vierte Gruppe von Nebenbedingungen (Gleichung 9) ist notwendig, da nur ungerichtete Kommunikationsbeziehungen für $i < j$ betrachtet werden, die Entscheidungsvariable jedoch für $\forall\, i, j \in V$ definiert ist (Gleichung 10). So ist es möglich, die Formulierung der vierten Nebenbedingungsgruppe (Gleichung 8) zu vereinfachen. Die fünfte Gruppe von Nebenbedingungen (Gleichung 10) definiert die binären Wertebereiche der Entscheidungsvariablen.

5 Anwendung des Modells bei einem Dienstleistungsunternehmen

Das in Kapitel 3 konzeptualisierte und in Kapitel 4 formalisierte Entscheidungsproblem wurde für ein internationales Dienstleistungsunternehmen mit mehr als 100.000 Mitarbeitern und für einen Teil der dort historisch gewachsenen IT-Landschaft instanziiert. Insgesamt betreibt das Unternehmen mehr als 200 Anwendungen. Gemeinsam mit zwei Experten wurde der in diesem Artikel präsentierte illustrative Ausschnitt der IT-Architektur ausgewählt. Hierbei wurden 21 Anwendungen, mit 19 Schnittstellen und 11 entsprechende Konsolidierungsmaßnahmen (die in acht Konsolidierungsprojekten gruppiert sind: P01a bis P08)[8] ausgewählt. Der betrachtete Ausschnitt der IT-Landschaft umfasst die wesentlichen IT-Systeme für den Vertrieb des Unternehmens und umfasst mehrere Vertriebskanäle (Filiale, Internet, Großkunden) von mehreren Geschäftsdivisionen inkl. der angrenzenden Hilfssysteme wie Stammdatenmanagement, Kundenbetreuung, Reklamationsmanagement und Customer-Relationship-Management. Ebenfalls betrachtet wurden die Planungssysteme für Personal, Finanzen und Produktion, die in direkter Abhängigkeit der vertriebsunterstützenden Geschäftsprozesse stehen, mitberücksichtigt.

[7] Generell gilt für alle $y_{ij} = 2 | \forall\, i, j \in V$ die nicht in Gleichung 8 oder Gleichung 9 auf den Wert 1 „gezwungen" werden, da $c_{ij} \in \mathbb{R}_+ \forall i, j \in V$ nur positive Werte annehmen kann.

[8] Aufgrund einer besonderen Anforderung in diesem Anwendungsfall wurde über zusätzliche Nebenbedingungen sichergestellt, dass Konsolidierungsprojekte mit einer identischen Projektnummer nur gemeinsam durchgeführt werden (daher die verwendete Nummerierung). Beispielsweise kann „P01a" nur durchgeführt werden, wenn auch „P01b" durchgeführt wird (und umgekehrt).

Für diesen Ausschnitt der IT-Landschaft sind vor der Konsolidierung 32 Mio. € Betriebskosten und 5 Mio. € Kommunikationskosten angefallen. Anschließend wurden auf Grundlage von existierenden Dokumenten zur Kostenschätzungen und in enger Zusammenarbeit mit zwei Experten und einem weiteren Wissenschaftler die für die Modellinstanziierung notwendigen Parameterwerte geschätzt.[9]

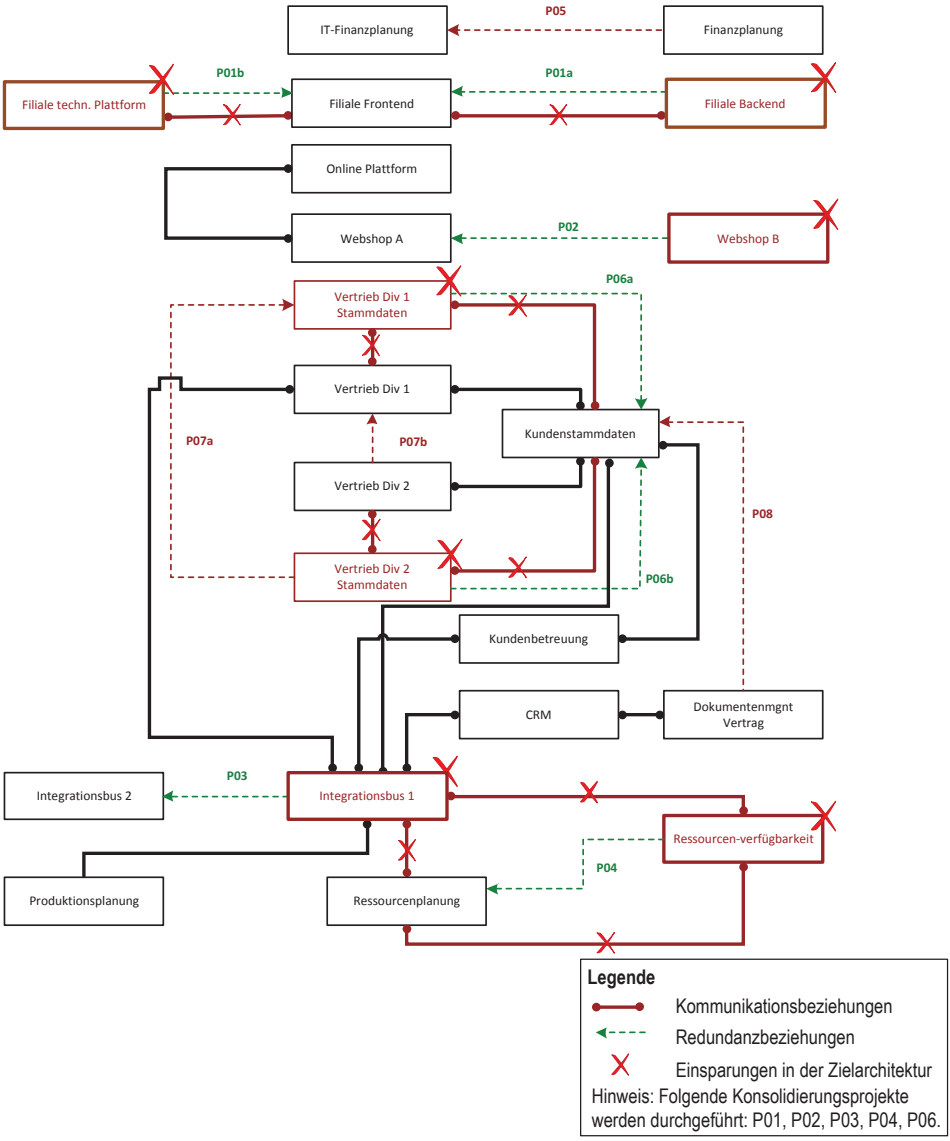


Abbildung 4: Betrachteter Ausschnitt der IT-Landschaft und Darstellung der Zielarchitektur (für eine Darstellung der Soll-Architektur siehe Abbildung 5 im Anhang)

[9] Die ermittelten Parameterwerte werden auf Anfrage an widjaja@is.tu-darmstadt.de zur Verfügung gestellt.

Nach der Modellanwendung umfasst der kostenoptimale Konsolidierungsplan sieben (in grün dargestellt) der elf möglichen Konsolidierungsmaßnahmen. Insgesamt können so sieben funktional redundante Anwendungen (in rot dargestellt) und neun entsprechende Kommunikationsbeziehungen (in rot dargestellt) eingespart werden. Fünf Kommunikationsbeziehungen bleiben unberührt und sechs weitere Kommunikationsbeziehungen müssen aufgrund von Konsolidierungsprojekten neu entwickelt werden. Damit konnten die Gesamtkosten von ca. 37 Mio. € (32 Mio. € Betriebskosten und 5 Mio. € Kommunikationskosten) um ca. 23 % auf ca. 28,5 Mio. € reduziert werden. Während der Evaluation standen insbesondere die Kriterien Validität und Nützlichkeit im Vordergrund. [GH13] Die Ergebnisse des Optimierungsmodells stimmten mit den Erfahrungen der Experten überein und sie betonten die Nützlichkeit des Modells für vergleichbare Konsolidierungsprojekte. Daneben zeigte die Evaluation zudem interessante Ansatzpunkte für weitere Forschung, die im nächsten Kapitel diskutiert werden.

6 Diskussion

Theoretische Implikationen

Die zwei wesentlichen Ergebnisse dieses Artikels sind: 1) Identifikation und Konzeptualisierung des beschriebenen Trade-offs zur Ermittlung des kostenoptimalen Grads an unerwünschter funktionaler Redundanz in einer IT-Architektur. 2) Formalisierung und Lösung dieses Trade-offs mithilfe eines mathematischen Modells. Die hier erarbeiteten Ergebnisse können in die Literatur zu EAM einfließen, da die Identifikation und das Management funktionaler Redundanzen einen großen Wert für Unternehmen stiften kann. [RWR06, Ta11] Nach [LC07], die in einer Serie von Fallstudien Industrieerfahrungen zum Management von funktionaler Redundanz von Software aggregieren, gilt jedoch: „There is no established theory or methods that specifically address this situation." (p. 420) Daher ist die Identifikation und Formalisierung des beschriebenen Trade-offs ein wichtiger erster Schritt in diese Richtung und unterstützt CIOs und IT-Architekten beim Kostenmanagement. [GP12, Li06] Zudem kann die Identifikation des in diesem Artikel beschriebenen Trade-offs die Diskussion um Komplexität von IT-Architekturen bereichern. [z. B. Sc97, SM03, SWK13] Bisher wird argumentiert, dass unnötige funktionale Redundanz zur Steigerung der Anzahl der Systemelemente und damit zur Komplexität beiträgt. Der Artikel zeigt jedoch auf, dass die Komplexitätsreduktion durch Abbau von funktionaler Redundanz nur bis zu einem gewissen Grad ökonomisch sinnvoll ist.

Praktische Implikationen

Der vorgestellte Ansatz ist ein erster Schritt, um das für IT-Architekten relevante Problem der Ermittlung der kostenoptimalen funktionalen Redundanz in der Zielarchitektur zu untersuchen und Lösungswege aufzuzeigen. Das vorgeschlagene mathematische Modell ist die Grundlage für ein Entscheidungsunterstützungssystem, dessen Prototyp aktuell entwickelt wird. Der Prototyp soll primär die Entscheidungsträger bei der Parametereingabe und Ergebnisinterpretation unterstützen und so die praktische Anwendbarkeit des hier vorgestellten mathematischen Modells vereinfachen. Aus unternehmerischer Sicht gilt es jeweils zu prüfen ob der Mehraufwand der Anwendung des vorgestell-

ten Modells für das zu lösende Problem gerechtfertigt ist – so ist das vorgestellte Optimierungsmodell beispielsweise erst ab einer gewissen Problemgröße und Komplexität der Redundanz- und Kommunikationsbeziehungen sinnvoll. Zudem müssen die Kosten für die präzise Datenerhebung vor dem Hintergrund der möglichen Einsparungen gesehen werden.

Limitationen und weiterer Forschungsbedarf

Da hier der aktuelle Stand eines laufenden Forschungsprojekts berichtet wird, bieten sich zahlreiche interessante Möglichkeiten der Weiterentwicklung. Insbesondere wurde das mathematische Modell noch nicht vollständig in einen Prototypen umgesetzt (die Implementierung läuft jedoch bereits). Zudem wurde der vorgestellte Ansatz bisher nur im Rahmen von einem (jedoch umfangreichen) Projekt evaluiert - hier gilt es weitere Fälle zu betrachten. Der im Aritkel beschriebene Trade-Off fokussiert sich auf den kostenoptimalen Grad an funktionaler Redundanz. Aus der Literatur können jedoch neben dem Kostenaspekt auch weitere, oft strategische und damit schwerer quantifizierbare Vorteile reduzierter Redundanz abgeleitet werden: Eine redundanzfreie IT-Architektur kann zum Beispiel leichter an geplante und ungeplante Änderungen der Geschäftsanforderungen angepasst werden. Daher bietet das vorgestellte Modell lediglich eine Entscheidungsperspektive, die im konkreten Fall durch weitere Betrachtungen zu ergänzen ist. Eine sinnvolle Erweiterung wäre zudem die Berücksichtigung von Unsicherheit bei den Eingabeparametern und eine entsprechend robuste stochastische Optimierung. Auch die statische Betrachtung des Entscheidungsproblems stellt eine Vereinfachung dar. Durch die enge Kooperation mit den Praxispartnern bei der Modellevaluation wurde deutlich, dass insbesondere für Unternehmen in denen jahrelange Konsolidierungsprogramme ablaufen, eine dynamische Betrachtung des Entscheidungsproblems hoch interessant ist - d. h. das Ergebnis der Optimierung sollte neben einer Zielarchitektur auch einen Konsolidierungsplan liefern, der die Reihenfolge der Konsolidierungsschritte festlegt. Diese dynamische Betrachtung ist Gegenstand aktueller Modellerweiterungen. Zudem geht der Artikel von einem binären Redundanzbegriff aus - wie bereits in Kapitel 2.2 diskutiert, sind darüber hinaus in der Realität oftmals auch graduelle funktionale Redundanzen zu finden (z. B. Anwendung A kann 80 % der Funktionen von Anwendung B anbieten). Dies gilt es in entsprechende Modellerweiterungen zu integrieren. In diesem Artikel wurde auf funktionale Redundanz auf Anwendungsebene fokussiert. Gerade da weitere theoretische Implikationen für die Forschung zu EA-Komplexität zu erwarten sind, sollten weitere Typen von Redundanzen (z. B. Datenredundanz) und andere Sub-Systeme der EA (wie zum Beispiel Datenarchitektur und Infrastrukturarchitektur) betrachtet und die entsprechenden Wechselwirkungen in das Modell integriert werden.

Dank

Herzlichen Dank an Markus Weckesser und Dennis Schmidt für die Unterstützung bei der Evaluierung des Modells und ihr hilfreiches Feedback zur Modellformulierung. Marcus Purzer und Dr. Ulf Brusch von der Senacor Technologies AG danke ich für die tatkräftige Unterstützung bei der Instanziierung und die zahlreichen Hinweise zur Praxisrelevanz. Zudem bedanke ich mich bei den zwei anonymen Reviewern für die hilfreichen Hinweise.

Literaturverzeichnis

[Ai09] Aier, S.; Kurpjuweit, S.; Saat, J.; Winter, R.: Enterprise Architecture Design as an Engineering Discipline. In: AIS Transactions on Enterprise Systems, 2009. **1**(1); S. 36-43.

[BB03] Basu, A.; Blanning, R.W.: Synthesis and Decomposition of Processes in Organizations. In: Information Systems Research, 2003. **14**(4); S. 337-355.

[Bu01] Buxmann, P.: Informationsmanagement in vernetzten Unternehmen: Wirtschaftlichkeit, Organisationsänderungen und der Erfolgsfaktor Zeit. Deutscher Universitätsverlag, Wiesbaden, 2001.

[BY06] Boh, W.F.; Yellin, D.: Using Enterprise Architecture Standards in Managing Information Technology. In: Journal of Management Information Systems, 2006. **23**(3); S. 163–207.

[CLT12] Chowanetz, M.; Legner, C.; Thiesse, F.: Integration: An Omitted Variable in Information Systems Research. In: *ECIS 2012 Proceedings.* 2012: Barcelona, Spain.

[EBD11] Espinosa, J.A.; Boh, W.F.; DeLone, W.: The organizational impact of enterprise architecture: a research framework. In: 44th Hawaii International Conference on System Sciences (HICSS), 2011. IEEE.

[GH13] Gregor, S.; Hevner, A.: Positioning and Presenting Design Science Research for Maximum Impact. In: MIS Quarterly, 2013. **37**(2); S. 337-355.

[GP12] Guillemette, M.G.; Paré, G.: Toward a New Theory of the Contribution of the IT Function in Organizations. In: MIS Quarterly, 2012. **36**(2); S. 529-551.

[HF69] Hall, A.D.; Fagen, R.E.: Definition of System. Wiley, New York, 1969.

[IEEE00] IEEE: IEEE Recommended Practice for Architectural Description of Software-Intensive Systems. In: IEEE Std 1471-2000, 2000; S. 1-23.

[KWS05] Klesse, M.; Wortmann, F.; Schelp, J.: Erfolgsfaktoren der Applikationsintegration. In: Wirtschaftsinformatik, 2005. **47**(4); S. 259-267.

[LC07] Land, R.; Crnkovic, I.: Software Systems In-House Integration: Architecture, Process Practices, and Strategy Selection. In: Information & Software Technology, 2007. **49**(5); S. 419-444.

[Li06] Lindström, Å.; Johnson, P.; Johansson, E.; Ekstedt, M.; Simonsson, M.: A Survey on CIO Concerns - Do Enterprise Architecture Frameworks Support Them? In: Information Systems Frontiers, 2006. **8**(2); S. 81-90.

[RWR06] Ross, J.W.; Weill, P.; Robertson, D.: Enterprise Architecture as Strategy: Creating a Foundation for Business Execution. Harvard Business Press, Boston, MA, 2006.

[Sc97] Schneberger, S.L.: Distributed computing environments: effects on software maintenance difficulty. In: Journal of Systems and Software, 1997. **37**(2); S. 101-116.

[Si62] Simon, H.A.: The architecture of complexity. In: Proceedings of the American philosophical society, 1962. **106**(6); S. 467-482.

[Si96] Simon, H.A.: The Sciences of the Artificial. MIT Press, Cambridge, MA, 1996.

[SM03] Schneberger, S.L.; McLean, E.R.: The complexity cross: implications for practice. In: Communications of the ACM, 2003. **46**(9); S. 216-225.

[SWK13] Schütz, A.; Widjaja, T.; Kaiser, J.: Complexity in Enterprise Architectures - Conceptualization and Introduction of a Measure from a System Theoretic Perspective. In: *European Conference on Information Systems (ECIS)*. 2013: Utrecht, Netherlands.

[Ta11] Tamm, T.; Seddon, P.B.; Shanks, G.; Reynolds, P.: How Does Enterprise Architecture Add Value to Organisations? In: Communications of the Association for Information Systems, 2011. **28**(1); S. 141-168.

[Wi11] Widjaja, T.: Standardisierungsentscheidungen in mehrschichtigen Systemen: Untersuchung am Beispiel serviceorientierter Architekturen. Gabler, Wiesbaden, 2011.

[Za87] Zachman, J.A.: A Framework for Information Systems Architecture. In: IBM Systems Journal, 1987. **26**(3); S. 276-292.

Anhang

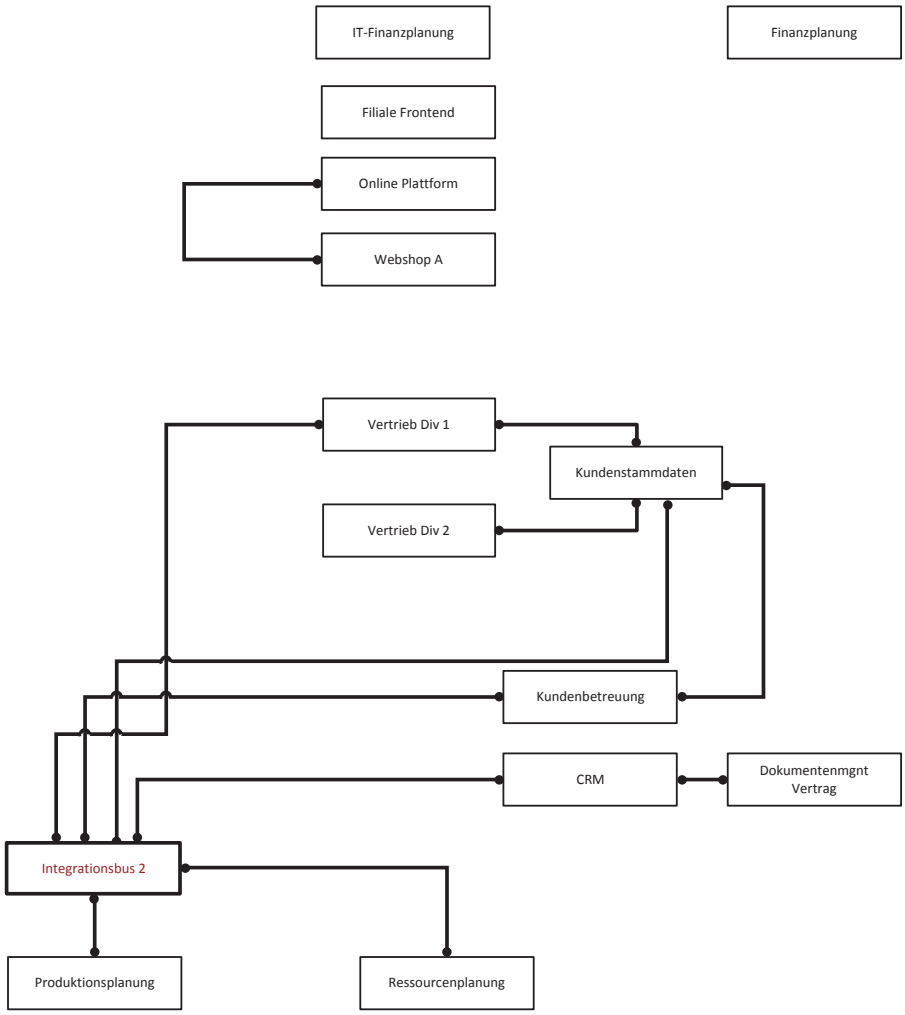


Abbildung 5: Schematische Darstellung der Soll-Architektur

Softwarealterung aus Sicht des IT-Managements – Ergebnisse einer qualitativ-empirischen Analyse in der Finanzindustrie

Wiebke Kappenberg, Paul Drews

Fachbereich Informatik
Universität Hamburg
Vogt-Kölln-Str. 30
22527 Hamburg
wiebke.kappenberg@gmail.com
drews@informatik.uni-hamburg.de

Abstract: Seit 20 Jahren wird das Phänomen der Softwarealterung in der Forschung überwiegend aus der Perspektive der Softwareentwicklung untersucht. Aus Sicht des IT-Managements ist dies unzureichend, da für das Management komplexer Anwendungslandschaften auch organisatorische, finanzielle und strategische Faktoren zu berücksichtigen sind. In Banken und Versicherungen setzen sich die Anwendungslandschaften aus vielfältigen Anwendungen unterschiedlicher Technologien und Generationen zusammen. Es stellt sich in diesem Zusammenhang die Frage, wie und anhand welcher Indikatoren das IT-Management in der Finanzindustrie das Softwarealter bewerten kann und welche Maßnahmen gegen die Softwarealterung ergriffen werden können. Um diese Fragen zu beantworten, wurden in einer qualitativ-empirischen Studie Experteninterviews mit 28 mit Vertretern des IT-Managements der Finanzindustrie durchgeführt. Die Ergebnisse zeigen, dass ein erweiterter Ansatz zur Bewertung von Softwarealterung erforderlich ist.

1 Einleitung

Seit Mitte der 90er Jahre, nachdem Parnas mit der Metapher Softwarealterung (Software Aging) die Aufmerksamkeit der Forschung auf dieses Phänomen lenkte [Pa94], wurden zahlreiche Methoden zur Bewertung des Softwarealters, zum Refactoring und zur Verjüngung von Software (Software Rejuvenation) [Hu95] entwickelt. Diese betrachten jedoch primär technische Eigenschaften, wie beispielsweise die zunehmende Anzahl an Fehlern, den steigenden Ressourcenverbrauch sowie den damit einhergehenden Performanceverlust [BF09] [Co11]. Diese technische Sichtweise ist geeignet, um Softwarealterung aus der Sicht der Softwareentwicklung oder des Produktmanagements zu bewerten. Aus der Sicht des IT-Managements bedarf es jedoch anderer Methoden, da die Bewertung für eine Vielzahl von Anwendungen vorgenommen werden muss. Dabei sind auch nicht-technische Indikatoren zu berücksichtigen. Hinzu kommt, dass für viele Anwendungen keine Möglichkeit besteht, den Quellcode zu analysieren oder zu verändern. Ziel dieses Beitrags ist es daher, eine erweiterte Sichtweise auf das Phänomen der Softwarealterung zu entwickeln, die auch die besonderen Interessen und

Herausforderungen des IT-Managements berücksichtigt. Diese erweiterte Sichtweise wird als das Ergebnis der Analysephase eines gestaltungsorientierten Vorgehens [He04] verstanden, in dessen weiterem Verlauf ein Artefakt entwickelt werden soll, das ein systematisches Management des Softwarealters unterstützt.

Im Anschluss an diese Einleitung ist der Artikel in vier Abschnitte untergliedert. Im folgenden Abschnitt werden die Ergebnisse einer systematischen Literaturanalyse zu Indikatoren für Softwarealterung beschrieben. Im dritten Abschnitt wird das Vorgehen bei der Durchführung einer qualitativ-empirischen Studie zur Softwarealterung aus Sicht des IT-Managements in der Finanzindustrie beschrieben. Die empirische Untersuchung hat das Ziel, Indikatoren zur Bewertung des Softwarealters sowie genutzte Maßnahmen zum Umgang mit diesem Phänomen zu identifizieren. Die Ergebnisse der empirischen Untersuchung werden im vierten Abschnitt beschrieben. Der Artikel schließt mit einer Zusammenfassung und einem Ausblick im fünften Abschnitt.

2 Bisheriger Stand der Forschung zur Softwarealterung

Seit Mitte der 90er Jahre werden Methoden zur Identifikation von Softwarealterung und Maßnahmen zum Umgang mit diesem Phänomen erforscht. Um die bisherigen Ergebnisse systematisch zu konsolidieren, wurde eine Literaturrecherche nach Brink [Br13] durchgeführt, die in drei Schritte untergliedert ist: (1) offene Literaturrecherche in elektronischen Volltextdatenbanken, Fachliteratur und Lehrbüchern, anschließend: Definition und Abgrenzung des Begriffs Softwarealterung (2) Erstellung einer Suchwortliste, diese umfasste zunächst die Suchterme „Software Aging“, „Software Age“, „IT-Management“, „Criteria“ und „Method“, später wurde die Liste um die folgenden Begriffe erweitert: „Code Deterioration“, „Software Risks“, „Software Quality“, „Software Rejuvenation“, „Software Maintenance“, „Software Refactoring“, „Software Health“, „Software Fitness“, „Software Lifetime Expectancy“ (3) Verwendung der Suchterme (kombiniert und einzeln) für eine ausführliche Recherche in sieben relevanten Datenbanken der Wirtschaftsinformatik [KW06], acht Journalen aus dem *„*Senior Scholars*' Basket of Journals“* der AIS sowie in den Publikationen von vier Konferenzen der Wirtschaftsinformatik. Die Literaturrecherche wurde dokumentiert und die Literatur auf Relevanz geprüft. Diese war gegeben, wenn die Abstracts Hinweise auf Methoden und Indikatoren für die Bewertung von Softwarealterung enthielten. Es verblieben 85 Quellen für die nähere Analyse. Im Folgenden werden die Ergebnisse der Literaturanalyse kurz zusammengefasst.

David L. Parnas legte mit seinem Artikel von 1994 den Grundstein für viele anschließende Forschungsarbeiten zum Thema Softwarealterung. Er beschreibt das Phänomen in seinem Artikel wie folgt: *„Programs, like people, get old. We can 't* prevent aging, but we can understand its causes, take steps to limits its effects, temporarily reverse some of the damage it has caused, and prepare for the day when the software is no longer viable.*“* [Pa94] Der Alterungsprozess von Software kann nur bedingt aufgehalten werden und endet stets mit der Abschaltung der veralteten Software oder der Ablösung durch eine neue. Faktoren, die den Alterungsprozess beeinflussen können, sieht Parnas insbesondere in der fehlenden Aktualisierung der Software sowie in

häufigen Änderungen durch unterschiedliche Mitarbeiter [Pa94]. Letztere können dazu führen, dass der Code und die Architektur strukturell zerstört werden. Eine schlechte Wartbarkeit führt zu weiteren Problemen wie abnehmender Leistungsfähigkeit, steigender Fehlerzahl und reduzierter Verfügbarkeit. Parnas geht nur kurz auf die Folgen ein, zu denen er steigende Kosten für Wartung und Weiterentwicklung sowie eine abnehmende Wettbewerbsfähigkeit zählt. Wie das Softwarealter durch das IT-Management konkret bewertet werden kann, wird in seinem Artikel nicht beschrieben. In der Literatur besteht Einigkeit darüber, dass das Phänomen der Softwarealterung nicht unmittelbar mit dem tatsächlichen Alter der Software in Jahren in Beziehung steht. Selbst wenn das tatsächliche Softwarealter (in Jahren) bekannt ist, so liefert dies keine qualitative Aussage über deren Zustand. Daher thematisiert die Forschung überwiegend die Bestimmung des optimalen Verjüngungszeitpunktes der Software, sodass sie ohne spürbare Beeinträchtigungen lange einsatzfähig bleibt. Die weiteren Artikel und deren Definitionen zur Softwarealterung, die im Rahmen der Literaturrecherche identifiziert wurden, stimmen mit den von Parnas genannten Kerneigenschaften überein. In erster Linie wird die Anhäufung von Fehlern mit Softwarealterung in Verbindung gebracht, weil alte Software aufgrund von häufigen und zum Teil inkonsistenten Änderungen schlechter wartbar wird [Ga98]. Insgesamt konnten in der Literatur 54 Indikatoren für Softwarealterung identifiziert werden, von denen die sechs meist genannten in Tabelle 1 aufgeführt sind.

Kategorie	**Beschreibung**	**Anzahl Quellen**	**Exemplarische Quellen**
Ressourcen-verbrauch	Bei alter Software nimmt der Ressourcenverbrauch (insbesondere der Speicherverbrauch) zu.	30	[Ca01] [EDB08] [WHG09]
Verfügbar-keit	Die Ausfallwahrscheinlichkeit nimmt auf Grund von Fehlern zu und die Verfügbarkeit der Software nimmt ab.	28	[An11] [RD07] [ZS08]
Fehler	Im Laufe der Zeit entstehen durch häufige Änderungen mehr Fehler, wie z. B. arithmetische Rundungsfehler oder Systemfehler.	22	[SLM10] [TV02] [VT05]
Performance	Die Performance nimmt im Laufe der Zeit ab, weil z. B. der Speicher nicht korrekt freigegeben wird.	15	[CT02] [GBB02] [VG03]
Antwortzeit	Aufgrund der zunehmenden Programmkomplexität können Antworten länger dauern und fehlerhaft sein.	10	[Bo11] [Ji08] [LXJ10]
Software-qualität	Die Softwarequalität (Funktionalität, Zuverlässigkeit, Effizienz, Wartbarkeit, Übertragbarkeit) nimmt im Laufe der Zeit durch häufige und inkonsistente Änderungen ab.	10	[BF09] [Jo07] [SD00]

Tabelle 1: Die 6 meistgenannten Softwarealterungskriterien aus der Literatur

Anhand dieser Indikatoren wird deutlich, dass Softwarealterung bisher primär aus der Perspektive der Softwareentwicklung erforscht wurde. Diese eng gefasste Sichtweise ist für das IT-Management nicht ausreichend und sollte erweitert werden. Lediglich Bombardieri und Fontana versuchen, das komplexe Phänomen der Softwarealterung umfassender zu betrachten. Sie sprechen in ihrem Artikel über Softwarealterung auch Faktoren wie Kosten für Änderungen und sich verändernden Nutzeranforderungen an [BF09]. Die Symptome der Softwarealterung können vielfältig sein: fehlende Unterstützung für Dateiformate, Kommunikationsstandards und neue Hardware oder graphische Benutzeroberflächen, die nicht die Anforderungen der Nutzer erfüllen [BF09]. Dieses Verständnis von Softwarealterung verdeutlicht die Abhängigkeit zwischen Software und Umwelt. Software ist nach ihrer Entwicklung in eine bestehende Anwendungslandschaft integriert, wird in Betrieb genommen und von unterschiedlichen Personen verwendet. Insgesamt fehlen in den identifizierten Quellen Indikatoren wie steigende Entwicklungskosten oder der (negative) Einfluss einer inkonsistenten Dokumentation auf die Softwareentwicklung. Diese sind jedoch für das IT-Management relevant, da es für die Steuerung und Aufrechterhaltung der Informationsinfrastruktur unter ökonomischen Gesichtspunkten verantwortlich ist [Kr10] [HL05]. Es fehlen Ansätze, die die Ursachen und Auswirkungen von Softwarealterung frühzeitig identifizieren können und sie mess- und bewertbar machen. Diese Ansätze müssen zudem in heterogenen Anwendungslandschaften für eigenentwickelte und zugekaufte Systeme anwendbar sein.

3 Methode und Vorgehen

Die Ergebnisse der Literaturrecherche zeigen, dass Softwarealterung aus der Sicht des IT-Managements von der Forschung bisher nicht explizit behandelt wurde. Um diesem Defizit zu begegnen, sollen bisherige Erkenntnisse aus der Forschung und Praxis miteinander in Beziehung gesetzt werden. Langfristiges Ziel ist es, einen neuen systematischen Bewertungsansatz für Softwarealterung zu entwickeln, der das IT-Management in seinen Aufgaben unterstützt. Um diesem Ziel näher zu kommen wurden in einer empirischen Studie Indikatoren von Softwarealterung und potentielle Anforderungen für den Bewertungsprototypen erhoben. Als Branche wurde die Finanzindustrie ausgewählt, da diese dafür bekannt ist, sehr große Anwendungslandschaften zu betreiben, deren Systeme teilweise bereits sehr lange im Einsatz sind. Das Forschungsparadigma Design Science nach Hevner et al. [He04] bietet einen geeigneten Rahmen, um dieser Herausforderung systematisch zu begegnen. Es untergliedert den Forschungsprozess in die vier Phasen Analyse, Entwurf eines Artefaktes, dessen Evaluation sowie die Diffusion der Forschungsergebnisse. In dem vorliegenden Artikel liegt der Fokus auf der Analyse-Phase. Am Ende des gesamten iterativ gestalteten Prozesses soll ein Ergebnis stehen, welches einen Beitrag zur Forschung und Praxis liefert [He04] [Ba08]. Dem Design Science-Paradigma folgend bietet die Knowledge Base geeignete Methoden für die Datenerhebung und -auswertung an [HC10]. Nachfolgend werden die für die Datenerhebung und -auswertung ausgewählten Methoden sowie das Vorgehen bei ihrer Anwendung beschrieben.

Vorgehensweise bei der Datenerhebung: Um die Sicht des IT-Managements von Banken und Versicherungen auf das Phänomen der Softwarealterung zu erheben, wurde das leitfadengestützte Experteninterview als Methode für eine direkte und kontextbezogene Datenerhebung ausgewählt [Tr05] [MN05]. Insgesamt wurden in 27 Interviews 28 Experten[1] befragt, wovon 16 dem Schwerpunkt Banken und 12 dem Schwerpunkt Versicherungen zuzuordnen sind. Die befragten Experten sind folgenden Rollen zuzuordnen: Führungsposition (11 Personen, Vorstand, Mitglied der Geschäftsführung, Chief Information Officer, IT-Leiter), IT-Bereichsleitung (11 Personen, Teilverantwortung für einen Bereich, z. B. Unternehmensarchitektur, interne Anwendungsentwicklung), IT-Abteilungsleitung (3 Personen, Verantwortung für ein Softwareprodukt oder Entwicklungsteam), IT-Beratung (3 Personen, berät verschiedene Unternehmen und hat einen Überblick über deren IT-Management). Für die Experteninterviews wurde ein Interviewleitfaden mit offenen Fragen erstellt und iterativ weiterentwickelt [GL10] [LT09]. Für die Konstruktion des Interviewleitfadens wurden die Empfehlungen aus der Literatur berücksichtigt [SF12] [MN05] [LT09]. Die Interviewpartner erhielten den Interviewleitfaden vor den Interviewterminen zur Vorbereitung. Die Interviews wurden vor Ort in den jeweiligen Unternehmen geführt und für die Auswertung digital aufgezeichnet [LT09].

Vorgehensweise bei der Datenauswertung: Für die Auswertung der erhobenen Daten wurde die qualitative Inhaltsanalyse ausgewählt, um typische Gemeinsamkeiten sowie die Unterschiede und Besonderheiten zu analysieren [MN05]. Konkret orientierte sich die Auswertung der Experteninterviews an dem Ablaufmodel von Gläser und Laudel [GL10]. Die aufgenommenen Interviews wurden vollständig transkribiert und anonymisiert [LT09]. Die Transkription führte zu 285 Seiten Interviewmaterial, das vollständig mithilfe der Software ATLAS.ti kodiert wurde. Erste Kodes wurden deduktiv aus den Interviewleitfäden und der recherchierten Literatur entwickelt [LS07]. Neue Kodes wurden während der Kodierung induktiv erstellt und im Kodierschema ergänzt. Im Gegensatz zu der Empfehlung von Gläser und Laudel wurde ein Probedurchlauf zur Kodierung vollzogen und das Kodierschema iterativ überarbeitet [GL10]. Dabei wurden Fehler korrigiert. Bedeutungsgleiche Kodes ein weiteres Mal geprüft und teilweise zusammengefasst oder untergliedert [GL10]. Im letzten Schritt wurden die gewonnen Informationen in Hinblick auf die Forschungsfrage verdichtet [GL10].

4 Softwarealterung in der Finanzindustrie: Indikatoren und Maßnahmen

In diesem Abschnitt werden die Ergebnisse der empirischen Untersuchung beschrieben. Zunächst wird die Sicht der Praktiker auf das Phänomen der Softwarealterung zusammengefasst. Anschließend werden die in der Praxis relevanten Indikatoren für Softwarealterung beschrieben. In Anlehnung an die Literatur zum IT-Management [HL05] [Kr10] werden diese in eine technische, eine organisatorische, eine finanzielle und eine strategische Dimension eingeordnet. Abschließend werden Maßnahmen beschrieben, die die Experten für den Umgang mit Softwarealterung vorschlagen.

[1] Alle Experten waren männlich, daher wird im Folgenden nur die männliche Form verwendet.

4.1 Der Begriff Softwarealterung aus Sicht des IT-Managements

In den Interviews wurde den Befragten keine Definition zum Begriff Softwarealterung genannt. Viele der befragten Experten verstehen das Softwarealter als eine Metapher, die für nicht oder schwer wartbare und langfristig problematische Software steht. Für sie geschieht *„Alterung […] dur*ch technologische, organisatorische und fachliche *Änderungen“*. Zum einen kann Software altern, wenn diese zu oft geändert wurde und ohne geeignete Gegenmaßnahmen mit der Zeit unwartbar wird. Zum anderen führen Änderungen außerhalb der Software zum Altern. Zu diesen Änderungen gehören die eingesetzten Technologien in der Anwendungslandschaft, die fachlichen Geschäftsprozesse oder die Organisation. Unter den Experten bestand Einigkeit darin, dass eine Software nicht schlecht sein muss, nur weil diese ein hohes Alter (in Jahren) aufweist. Eine alte Software ist oftmals stabiler als eine neue Software, da viele Fehler im Verlauf der Nutzung und Weiterentwicklung korrigiert wurden. Die Experten wiesen zudem auf einen engen Zusammenhang von Softwarealter und Softwarequalität hin.

4.2 Technische Dimension

Häufige Änderungen des Quellcodes und der Architektur, z. B. durch neue fachliche Anforderungen, können zu einer **hohen Komplexität und einer schlechteren Wartbarkeit** führen. Dies tritt insbesondere dann ein, wenn die Architektur nicht flexibel für Änderungen entworfen wurde. Es ist aus Sicht der Experten eine Herausforderung, die Anforderungen nachhaltig und optimal in die vorhandene Architektur zu integrieren. Dabei weist eine neue Software noch eine klare funktionale Trennung der Module auf. Zitat: *„Architektur und Schnittstellentechniken spielen* dann auch eine Rolle. Das ist sicherlich bei Produkten, die sich weiterentwickeln, ein Problem. Dazu zählt, wie aufgeräumt und strukturiert der Code ist und ob es Redundan*zen gibt.“* Zeit- und Kostendruck führen in der Praxis jedoch schnell dazu, dass diese Trennung aufgelöst wird. Auf diese Weise verliert die Architektur nach einigen Jahren ihre Konsistenz und wird dadurch schlechter wartbar.

Die **Aktualität von Architekturkonzepten, Standards und Best Practices** zeigt, dass zuvor gewählte Ansätze als überholt gelten. Ob die Software den aktuellen Standards entspricht, erfahren die Experten durch Befragung der Entwickler. Ebenso wie sich Best Practices weiterentwickeln, entwickeln sich Programmiersprachen im Laufe der Zeit weiter. Zitat: *„Die Software ist ja in einer Programmiersprache entwickelt. In der Regel* entwickeln sich Programmiersprachen weiter. Von daher kann die verwendete Programmiersprache nicht mehr zeitgemäß sein.*“* Früher wurden vorwiegend Programmiersprachen wie z. B. COBOL, PL/I oder Smalltalk eingesetzt. Diese gelten heute als nicht mehr zeitgemäß. Insbesondere altert die Software durch die Programmiersprache, wenn die Integration innovativer softwaretechnischer Konzepte oder neuer Plattformen nicht möglich ist.

Neben der **eingesetzten Programmiersprache** haben auch die **verwendeten Technologien** einen Einfluss auf das Softwarealter. Werden z. B. Bibliotheken, Frameworks oder Hardware nicht regelmäßig geprüft und aktualisiert, können Fehler auf Grund von Inkompatibilitäten auftreten oder auch Sicherheitslücken entstehen, weil

nicht die aktuellsten Updates installiert wurden. **Auslaufende Wartungsverträge** von Basis-Plattformen und deren Austausch können die Ursache dafür sein, dass Anpassungen an der Software notwendig sind. Ohne diese Anpassungen käme es zu Fehlern und die Software wäre auf Dauer nicht mehr einsatzfähig. Die resultierenden Fehler und der eintretende Alterungsprozess können durch **fehlgeschlagene Tests** angezeigt werden. Werden diese ersten Hinweise ignoriert, ist auch eine zunehmende Fehlerhäufigkeit im operativen Betrieb der Software zu beobachten.

4.3 Organisatorische Dimension

Aus der organisatorischen Perspektive betrachtet, hat eine häufig **wechselnde Zusammensetzung des Entwicklungsteams**, etwa durch Fluktuation oder dem bevorstehenden Ruhestand einiger Teammitglieder, einen Einfluss auf die Softwarealterung. Bei der Weiterentwicklung der Software kann es dann der Fall sein, dass die **Dokumentation nicht vollständig oder konsistent** ist und Mitarbeiter wichtige Aspekte im Laufe der Zeit vergessen können. Zitat: *„Da altert die Software, weil man* Personalwechsel hat. Das ist schon ein wichtiger Faktor. Denn man kann so viel dokumentieren wie man will, denn ein wesentlicher Teil steckt immer in den Köpfen der Mitarbeiter *[...]*. Auch durch das Wechseln von Mitarbeitern gehen wichtige *Informationen verloren.“* Daher sind langjährige Mitarbeiter für den Erhalt und die Weiterentwicklung der Software wichtig. Sie kennen sich besonders gut mit der Software aus und können entsprechend der Architektur die geforderten Änderungen umsetzen. Auf der anderen Seite können diese Teammitglieder ihre Macht ausnutzen, wenn sie sich über den Wert ihres Wissens im Projekt bewusst sind und es nicht an ihre Kollegen weitergeben.

Ein weiterer wesentlicher Faktor ist die **Motivation der Mitarbeiter**. Es kann vorkommen, dass das IT-Management seine Mitarbeiter unter Druck setzt, neue Anforderungen zu integrieren oder Fehler in der alten Software schneller zu beheben. Doch die **Bearbeitungsdauer** nimmt auf Grund der sinkenden Softwarequalität zu. In der Folge können sich die Mitarbeiter weigern, die neuen Anforderungen in die Software zu integrieren, weil dies zu aufwändig ist. Es kann aber auch Mitarbeiter geben, die aus Gründen des Arbeitsplatzerhalts neue Anforderungen in die Software integrieren wollen. Projekte können sich somit zu einer politischen Angelegenheit entwickeln, in denen nicht immer die optimale Lösung für die Software gefunden wird.

Softwarealterung wird von vielen Experten mit dem **Alter der Teammitglieder** und den von ihnen **gelernten Programmiersprachen und Technologien** verbunden. Zitat: *„Außerdem sterben die Sprachen aus und wenn du keinen Entwickler mehr has*t, der diese Sprache beherrscht. Das ist so ziemlich das Schlimmste, was du dir vorstellen kannst. Du hast dann eine stillgelegte Anwendung mit ziemlich vielen Kellern, Balkonen, Anbauten und hast dann keine *Leute, die sich damit auskennen.“* Mit den älteren Mitarbeitern, die das Unternehmen verlassen, *„sterben“* die verwendeten alten Programmiersprachen aus. An den Hochschulen lernen die Studierenden vielfach nur moderne Technologien kennen und müssen nach ihrem Abschluss Schulungen im Unternehmen besuchen, um alte Programmiersprachen zu lernen. Dieses scheitert

oftmals an der fehlenden Motivation, etwas nicht Innovatives zu erlernen. Daher ist es eine Herausforderung für die Unternehmen, den Bedarf an Fachkenntnissen hinsichtlich der Programmiersprachen in Einklang mit der Personalplanung zu bringen. Die meisten Experten schätzen diese Entwicklung als problematisch ein. Sie sind der Ansicht, dass es nicht genügend Personen gibt, die alte Programmiersprachen beherrschen bzw. neu erlernen. Durch die hohe Nachfrage können diese Mitarbeiter im Vergleich etwas teurer sein.

4.4 Finanzielle Dimension

In den Interviews wurden das zur Verfügung stehende **IT-Budget und dessen Verteilung** sehr häufig thematisiert. Es wird sehr genau abgewogen, wie viel des Budgets in die Umsetzung regulatorischer Anforderungen sowie in Innovation und neue Funktionalitäten investiert werden kann. Oft bleibt nur ein geringer Anteil für den Erhalt und die Erneuerung von Software übrig. Deshalb sollten die gesamten Kosten bei der Entwicklung einer neuen Software, also die Total Cost of Ownership, betrachtet werden. Ist die Software flexibel für Änderungen entwickelt worden, so können langfristig die Betriebskosten gering bleiben. Häufig wird in der Finanzbranche ein bestimmter Anteil des **IT-Budgets für Refactorings** geplant. Jedoch muss der Mehrwert von den Fachbereichen anerkannt und dafür ausreichend Budget zur Verfügung gestellt werden. Kleinere Refactorings lassen sich oftmals im Rahmen von Wartungsverträgen durchführen. Dennoch sind die Verhandlungen von Wartungsverträgen eine Gradwanderung für das IT-Management, in denen sie Kosten und Nutzen zueinander ins Verhältnis setzen müssen.

Einige Experten aus der Bankenbranche berichteten, dass teilweise bis zu 70% des IT-Budgets für die Umsetzung von regulatorischen Anforderungen investiert wird. In alter Software lassen sich diese neuen **regulatorischen Anforderungen nicht so schnell umsetzen**. Das verbleibende Budget wird nur noch nach Kritikalität und Kosten-Nutzen-Verhältnis verteilt, womit nur die dringendsten Aufgaben abgedeckt werden können. Daher bleibt selten Budget für verjüngende Maßnahmen übrig und der Alterungsprozess kann oft nicht aufgehalten werden. Das Hauptproblem, das fast alle Befragten während der Interviews nannten, ist, dass die **Fachbereiche kein Geld für Erneuerungsprojekte oder Wartungsmaßnahmen** zur Verfügung stellen, da diese keinen Mehrwert in den Maßnahmen erkennen. Dieses ist in der Banken- und Versicherungsbranche gleichermaßen deutlich erkennbar. Wird kein Budget zur Verfügung gestellt, so wird die IT „kaputt gespart“. Um diese schwierige Situation zu vermeiden, werden zum Teil indirekt Aufwände für Wartungsmaßnahmen in neue Funktionalitäten einkalkuliert.

4.5 Strategische Dimension

Softwarealterung hat einen **Einfluss auf den Wettbewerb und die Zufriedenheit der Kunden**. Die IT-Abteilung muss mit einer angemessenen **Umsetzungsgeschwindigkeit** neue Anforderungen und Produkte kostengünstig in ihre bisherige Software integrieren können, um wettbewerbsfähig zu bleiben. Zitat: *„Den Banken ist es wi*chtig, auf Änderungen zu reagieren und diese möglichst effizient umzusetzen. Effizienz wird

letztendlich in Kosten gemessen. Wir sehen einen Wettbewerb um die Kundschaft und das gilt auch für uns. Es ist äußerst wichtig, die Änderungen kostenarm und effizient *umzusetzen."* Dies ist eine besondere Herausforderung zu Zeiten großer Dynamik und großen Drucks durch die Gesetzgebung. Anforderungen müssen schnell und kostengünstig zugleich umgesetzt werden, da sonst die Software schnell veraltet.

Ebenso ist es im Wettbewerb wichtig, die vorhandenen Kunden zu erhalten und neue zu gewinnen. Aus diesem Grund messen einige Unternehmen die **Kundenzufriedenheit** über Befragungen, die z. B. die Leistungsfähigkeit und die Benutzbarkeit der Software bewerten. Eine alte Software kann die Bedürfnisse der Kunden nicht mehr vollständig erfüllen und dadurch sinkt die Zufriedenheit. Aus der Sicht der Produktentwicklung eines IT-Herstellers kann es sogar gewünscht sein, dass Software veraltet. Dies ist aber nur der Fall, wenn neue Innovationen bereits entwickelt wurden und das Interesse der Kunden zum Kaufen geweckt werden soll. Ein weiterer Indikator besteht darin, ob für eine Software noch **neue Kunden gewonnen werden können**.

4.6 Maßnahmen gegen Softwarealterung

Die Maßnahmen, die in der Praxis eingesetzt werden, um der Softwarealterung entgegenzuwirken, können einer technologischen und einer organisatorischen Ebene zugeordnet werden. Auf der technologischen Ebene wurde vorgeschlagen, bei jeder Weiterentwicklung **Refactoring-Maßnahmen** einzuplanen, damit die Software auch für die Zukunft gut wartbar bleibt und gut getestet werden kann. Deshalb ist es von Bedeutung, Budget und Zeit für Refactorings einzuplanen. Die Softwareentwicklung kann im Rahmen von **Code Reviews** auf Verbesserungen hinweisen, wenn der Code durch zu viele Änderungen seine Struktur verloren hat. Diese Verbesserungsvorschläge können später für die Realisierung priorisiert und geplant werden. Durch Refactoring-Maßnahmen werden die Effekte von Alterung minimiert und die Weiterentwicklungskosten für die Zukunft bleiben vergleichsweise gering.

Eine weitere Strategie, die insbesondere Unternehmen aus der Versicherungsbranche anwenden, ist die **Konsolidierung von Anwendungslandschaften und das Ablösen alter Software**. Durch vorhergegangene Fusionen sollen alte und zum Teil redundante Applikationen systematisch abgeschaltet und die Anwendungslandschaft beispielsweise mittels einer Service-orientierten Architektur vereinheitlicht werden. Dadurch sollen eine bessere Übersichtlichkeit erzeugt und die Kosten für Lizenzen und Infrastruktur reduziert werden. Nach der Planung einer Soll-Anwendungslandschaft können gezielt Systeme abgeschaltet werden. Die Experten plädieren für eine **vollständige Prüfung der Fach- und Anwendungsarchitektur auf Alterungseffekte**.

Eine andere mehrfach genannte Maßnahme zur Reduzierung der Komplexität und zur Ablösung alter Software, die zum Teil umstritten ist, besteht in der **Einführung von Standardsoftware**. Als Grund für den Einsatz von Standardsoftware gaben die IT-Manager sehr oft an, dass damit die Verantwortung weitestgehend an den Softwarehersteller übertragen werden kann. Der Softwarehersteller muss seine Standardsoftware in Release-Zyklen kontinuierlich weiterentwickeln, sodass die Software fachlich und technologisch für den Kunden aktuell bleibt. Speziell in der

Bankenbranche werden die Kernbankensysteme, die eine lange Lebenserwartung besitzen, derzeit in einigen Banken durch Standardsoftware ersetzt. Die Experten versprechen sich davon langfristig eine Komplexitätsreduzierung und Investitionssicherheit. Aus der organisatorischen Perspektive ist das **Fachwissen der Mitarbeiter** für das IT-Management von besonderer Bedeutung. Die eingesetzten Technologien in den Unternehmen haben eine erwartete Lebenszeit von teilweise mehreren Jahrzehnten. Daher ist es wichtig, dass dieses Fachwissen erhalten bleibt und sogar ausgeweitet wird, indem die Mitarbeiter regelmäßig geschult werden.

In den meisten Unternehmen werden fachliche und technische Anforderungen sowie Wartungsmaßnahmen im Rahmen der **jährlichen Planungsprozesse** oder Strategieüberprüfungen besprochen. Diese münden in Projekt- und Budget-Planungen für das folgende Jahr sowie in einer strategischen Planung mit langfristigen Zielen. Diese Prozesse sollen sicherstellen, dass die Wettbewerbsfähigkeit und die Innovationskraft des Unternehmens erhalten bleiben und es nicht zu kritischen Situationen kommt, in denen mit hohen Investitionen ein Nachholbedarf ausgeglichen werden muss. Zwar wird Softwarealterung in diesen Prozessen noch nicht explizit berücksichtigt, dies wäre aber nach Meinung vieler Experten sinnvoll. Zitat: „Man muss sicher regelmäßig seine Softwarelandschaft prüfen, welche Systeme auf keinen Fall altern dürfen und bei welchen Systemen akzeptieren wir das.“ Lediglich eines der befragten Unternehmen hat bisher eine systematische Überprüfung des Zustandes der Software der gesamten Anwendungslandschaft durchgeführt. Eine systematische Wiederholung dieser Überprüfung und die wiederkehrende Nutzung der dabei gewonnenen Informationen in den Planungsprozessen wurden in diesem Fall jedoch ebenfalls nicht praktiziert.

5 Zusammenfassung und Ausblick

In den letzten 20 Jahren wurde Softwarealterung primär aus der Perspektive der Softwareentwicklung erforscht, die sich auf Indikatoren wie steigende Fehlerrate, Performanceverlust und Ausfälle beschränkt. Praxistaugliche Ansätze für die Bewertung des Softwarealters aus der Sicht des IT-Managements, welche selbst für große Anwendungslandschaften anwendbar sind, stehen bisher nicht zur Verfügung. Die Literatur beschreibt Softwarealterung als einen Prozess, in dem die Software fehleranfälliger wird und daraufhin Abstürze und Ausfälle auftreten. Dagegen sehen die befragten Personen aus dem IT-Management die Ursache von Softwarealterung in der Veränderung der Software selbst, die unter anderem durch die Integration neuer Anforderungen und der damit einhergehenden steigenden Komplexität verursacht wird. Zugleich hat auch die Veränderung der Umwelt einen Einfluss auf die Softwarealterung, da Software nicht isoliert zu betrachten ist, sondern in einem komplexen System aus Abhängigkeiten zu sich ändernden Anforderungen, anderen Anwendungen, Mitarbeitern, Kunden und Lieferanten eingebettet ist.

Trotz der großen Bedeutung wird das Alter von Software in den befragten Unternehmen bisher nicht explizit regelmäßig bewertet. Eine professionelle Strategie zur Planung und Umsetzung von Gegenmaßnahmen zur Softwarealterung ist bisher nicht etabliert. Die

Strategien, wie mit Softwarealterung umgegangen wird, unterscheiden sich dabei zwischen Literatur und Empirie deutlich. Die Ansätze der Softwareverjüngung aus der Literatur sind sehr kurzfristig, wie z. B. ein Neustart der Software, und beziehen sich überwiegend nur auf einzelne Applikationen. Dagegen beschreibt das befragte IT-Management Strategien, die die Effekte von Softwarealterung langfristig beheben sollen, wie z. B. die Auswahl von Software für regelmäßige Refactorings, die Konsolidierung der Anwendungslandschaft, Weiterbildungsangebote für Mitarbeiter oder auch die eine Veränderung der Prozesse in der Budget- und Jahresplanung. Zukünftig müssen neue Ansätze entwickelt werden, welche die verschiedenen Dimensionen der Softwarealterung berücksichtigen und eine bessere Entscheidungsgrundlage für das IT-Management bieten.

Literaturverzeichnis

[An11] Andrade, E. C. et al.: Modeling and Analyzing Server System with Rejuvenation through SysML and Stochastic Reward Nets. In: Proc. ARES 2011, 2011; S. 161-168.

[Ba08] Baskerville, R.: What design science is not. European Journal of Information Systems (17:5), 2008; S. 441-443.

[BF09] Bombardieri, M.; Fontana, F.: Software aging assessment through a specialization of the SQuaRE quality model. In: Proc. WOSQ '09, ICSE Workshop, 2009; S. 33-38.

[Bo11] Bovenzi, A. et al.: Workload Characterization for Software Aging Analysis. In: Proc. ISSRE 2011, Hiroshima, 2011; S. 240-249.

[Br13] Brink, A.: Anfertigung wissenschaftlicher Arbeiten: Ein prozessorientierter Leitfaden zur Erstellung von Bachelor-, Master- und Diplomarbeiten. Springer Gabler, Wiesbaden, 2013.

[Ca01] Castelli, V. et al.: Proactive Management of Software Aging. IBM Journal of Research and Development (45:2), 2001; S. 311-332.

[Co11] Cotroneo, D. et al.: Software Aging und Rejuvenation: Where We Are and Where We Are Going. In: Proc. 3rd Int. Workshop on Software Aging and Rejuvenation, 2011; S. 1-6.

[CT02] Calzarossa, M. C.; Tucci, S. (Hrsg.): Performance Evaluation of Complex Systems: Techniques and Tools. Springer, Berlin, 2002.

[EDB08] El-Shishiny, H.; Deraz, S.; Bahy, O.: Mining Software Aging Patterns by Artificial Neural Networks. In (Marinai, S.; Schwenker, F. Hrsg.): Artificial Neural Networks in Pattern Recognition. Springer, Berlin, 2008; S. 252-262.

[Ga98] Garg, S. et al.: A methodology for detection und estimation of software aging, In: Proc. 9th Int. Symp. on Software Reliability, 1998; S. 283-292.

[GBB02] Gross, K. C.; Bhardwaj, V.; Bickford, R.: Proactive detection of software aging mechanisms in performance critical computers. In: Proc. 27th Annual NASA Software Engineering Workshop, 2002; S. 17-23.

[GL10] Gläser, J.; Laudel, G.: Experteninterviews und qualitative Inhaltsanalyse: Als Instrumente rekonstruierender Untersuchungen. VS Verlag für Sozialwissenschaften, Wiesbaden, 2010.

[Ji08] Jia, Y.-F. et al.: On the Relationship between Software Aging and Related Parameters. In: Proc. QSIC '08, Oxford, 2008; S. 241-246.

[HC10] Hevner, A.; Chatterjee, S. (Hrsg.): Design Research in Information Systems: Theory and Practice. Springer, Berlin, 2010.

[He04] Hevner, A. R. et al.: Design Science in Information Systems Research. MIS Quarterly (28:1), 2004; S. 75-105.

[HL05] Heinrich, L. J.; Lehner, F.: Informationsmanagement: Planung, Überwachung und Steuerung der Informationsinfrastruktur. Oldenbourg, München, 2005.
[Hu95] Huang, Y. et al.: Software rejuvenation: analysis, module und applications. In: Proc. 25th Int. Symp. on Fault-Tolerant Computing, 1995; S. 381-390.
[Jo07] Jones, C.: Geriatric Issues of Aging Software. In: Crosstalk, The Journal of Defense Software Engineering (20:12), 2007.
[Kr10] Krcmar, H.: Informationsmanagement. Springer, Berlin, 2010.
[KW06] Knackstedt, R.; Winkelmann, A.: Online-Literaturdatenbanken im Bereich der Wirtschaftsinformatik. Wirtschaftsinformatik (48:1), 2006; S. 47-59.
[LS07] Lewins, A.; Silver, C.: Using software in qualitative research: A step-by-step guide. Sage, Los Angeles, 2007.
[LT09] Liebold, R.; Trinczek, R.: Experteninterview. In (Kühl, S. Hrsg.): Handbuch Methoden der Organisationsforschung. VS Verlag für Sozialwissenschaften, Wiesbaden, 2009; S. 32-56.
[LXJ10] Lei, X.; Xue, K.; Jia, Y.: A comprehensive metric to evaluating software aging. In: Proc. ICCASM 2010, Taiyuan, 2010; S. V11-518-V11-520.
[MN05] Meuser, M.; Nagel, U.: ExpertInneninterviews - vielfach erprobt, wenig bedacht. In (Bogner, A. Hrsg.): Das Experteninterview : Theorie, Methode, Anwendung. VS, Verlag für Sozialwissenschaften, Wiesbaden, 2005; S. 71-93.
[Pa94] Parnas, D.: Software Aging. In: Proc. Int. Conf. on Software Engineering, 1994; S. 279-287.
[RD07] Rinsaka, K.; Dohi, T.: A Faster Estimation Algorithm for Periodic Preventive Rejuvenation Schedule Maximizing System Availability. In (Malek, M. et al. Hrsg.): Service Availability, Springer, Berlin, 2007; S. 94-109.
[SD00] Swanson, E. B.; Dans, E.: System Life Expectancy and the Maintenance Effort: Exploring Their Equilibration. MIS Quarterly (24:2), 2000; S. 277-297.
[SF12] Stigler, H.; Felbinger, G.: Der Interviewleitfaden im qualitativen Interview. In (Stigler, H.; Reicher, H., Hrsg.): Praxisbuch empirische Sozialforschung in den Erziehungs- und Bildungswissenschaften. Studien-Verlag, Innsbruck, 2012; S. 129-134.
[SLM10] Salfner, F.; Lenk, M.; Malek, M.: A survey of online failure prediction methods. ACM Comput. Surv. (42:3), 2010; S. 1-42.
[Tr05] Trinczek, R.: Wie befrage ich Manager? In (Bogner, A. Hrsg.): Das Experteninterview: Theorie, Methode, Anwendung. VS, Verlag für Sozialwissenschaften, Wiesbaden, 2005; S. 225-238.
[TV02] Trivedi, K. S.; Vaidyanathan, K.: Software Reliability and Rejuvenation: Modeling and Analysis. In (Calzarossa, M. C.; Salvatore, T. Hrsg.): Performance Evaluation of Complex Systems: Techniques and Tools. Springer, Berlin, 2002; S. 318-345.
[VG03] Vaidyanathan, K.; Gross, K.: MSET Performance Optimization for Detection of Software Aging. In: Proc. 14th IEEE Int. Symp. on Software Reliability Engineering, 2003.
[VT05] Vaidyanathan, K.; Trivedi, K.: A Comprehensive Model for Software Rejuvenation. In: IEEE Transact. on Dependable and Secure Computing (2:2), 2005; S. 124-137.
[WHG09] Wu, Q. E.; Han, Z. Y.; Guo, T. S.: Application of an Uncertain Reasoning Approach to Software Aging Detection. In: Proc. NCM '09, Seoul, 2009, S. 1592-1597.
[ZS08] Zhao, L.; Song, Q.: Availability and Cost Analysis of a Fault-Tolerant Software System with Rejuvenation. In: Proc. ICACTE '08, 2008; S. 261-265.

Konzeption eines SOA-Repository mit Analysefähigkeiten

Thomas Bauer [a], Stephan Buchwald [b], Julian Tiedeken [c], Manfred Reichert [c]

[a] Hochschule Neu-Ulm
Fakultät
Informationsmanagement
Wileystr. 1
89231 Neu-Ulm
thomas.bauer@
hs-neu-ulm.de

[b] T-Systems International
Delivery Unit Automotive
& Manufacturing Industries
Olgastr. 63
89073 Ulm
stephan.buchwald@
t-systems.com

[c] Universität Ulm
Institut für Datenbanken
und Informationssysteme
James-Franck-Ring
89081 Ulm
{manfred.reichert, julian.
tiedeken}@uni-ulm.de

Abstract: In einer SOA werden bereitgestellte Services von Servicenutzern, etwa fremden IT-Systemen konsumiert. Eine Serviceänderung bzw. -abschaltung kann solche Servicenutzer schwer beeinträchtigen. Deshalb müssen diese vor einem solchen Eingriff ermittelt werden, wozu das SOA-Repository genutzt werden kann. Allerdings ist bei umfangreichen SOA-Repositories schwer erkennbar, welche IT-Systeme (evtl. indirekt) in welchem Zeitraum betroffen sein werden. Da diese Problemstellung in der bisherigen SOA-Literatur nicht betrachtet wird, stellen wir ein Lösungskonzept vor. Es ermöglicht automatisierte Analysen, um in den Daten enthaltene Problemsituationen zu identifizieren (Ist-Analysen), und auch, um Auswirkungen zukünftig durchzuführender Serviceänderungen zu simulieren (Planspiele).

1 Einleitung

Service-orientierte Architekturen (SOA) führen zu einer Erhöhung der Flexibilität. So ermöglicht beispielsweise die lose Kopplung beim Service-Aufruf, schnell zur Verwendung eines anderen Services zu wechseln. Dadurch dass ein Proxy (Enterprise Service Bus, ESB) zwischen die aufrufende Applikation und den verwendeten Service geschaltet wird, muss für diesen Wechsel lediglich der Proxy verändert werden, jedoch nicht die konsumierende Applikation selbst. Wir nutzen im Projekt ENPROSO[1] zahlreiche solcher Mechanismen zur Erhöhung der Flexibilität [BBR11, Bu12].

Die hohe Flexibilität bei Service-Aufrufen kann jedoch zu einer Starre beim Betrieb der Services führen: So werden manche Veränderungen an bereitgestellten (d.h. betriebenen) Services erschwert. Dies gilt insb. für die Abschaltung von Services, da dies die Funktionsfähigkeit der Servicenutzer (Consumer) beeinträchtigen kann. Ein solcher Fall muss unbedingt vermieden werden, weil selbst ein Ausfall einzelner Funktionalitäten der Service-nutzenden Applikationen meist nicht akzeptabel ist. Andererseits verursacht ein lange andauernder Weiterbetrieb vieler veralteter Services hohe Kosten. Das bedeutet, dass ein Konzept entwickelt werden muss, um ungenutzte Services zuverlässig identifizieren und dann abschalten zu können (Service Sun-down). Solche Services können

[1] Enhanced Process Management by Service Orientation, durchgeführt bei und finanziert von der Daimler AG.

mit Informationen aus dem SOA-Repository identifiziert werden. Als Grundlage hierfür haben wir im Projekt ENPROSO ein sehr umfassendes Metamodell für ein SOA-Repository konzipiert [BTBR10, Bu12, Ti10]. Dieses stellt detaillierte Information zu Service-Versionen und -Installationen ebenso bereit wie zu Service-Nutzungsvereinbarungen (Contracts). Mit diesen und weiteren Informationen ist erkennbar, welche konkreten Service-Versionen und -Installationen tatsächlich genutzt werden, bis wann dies erfolgt und welche verzichtbar sind bzw. in Zukunft abgeschaltet werden können. Damit kann der Anpassungsaufwand für die konsumierende Applikation abgeschätzt werden.

Hierdurch ergibt sich ein umfangreiches SOA-Repository mit vielen verschiedenen Objekttypen und zudem jeweils vielen -Instanzen. Kritisch daran ist, dass zu erwartende Problemsituationen für menschliche Benutzer aufgrund der Datenmenge kaum mehr direkt erkennbar sind. Deshalb werden automatisierte Analysemöglichkeiten benötigt: Basierend auf den existierenden SOA-Repository-Daten müssen insbesondere zukünftig auftretende Problemsituationen entdeckt werden. Da es zu solchen automatischen Analyseverfahren bisher keine SOA-Literatur gibt, stellen wir in diesem Beitrag entsprechende Verfahren vor, um z.B. Services zu finden, deren Abschaltung vorgesehen ist, bevor ihre Nutzungsvereinbarung ausläuft.

Als Erweiterung zur Erkennung von solchen – aufgrund der existierenden Daten bereits „vorprogrammierten“ – Problemsituationen (d.h. Ist-Analysen) sollte ein SOA-Repository auch Planspiele (What-if-Analysen) ermöglichen: Um einen geeigneten Zeitpunkt für eine Service-Abschaltung wählen zu können, müssen die jeweiligen Folgen analysierbar sein. Dazu müssen mehrere Abschaltzeitpunkte betrachtet werden, wobei jeweils die betroffenen Service-Consumer und Lösungsalternativen identifiziert werden (Impact-Analysen). Da die Bewertung mehrerer potentieller Abschaltzeitpunkte basierend auf den SOA-Repository-Daten einen großen Aufwand verursachen würde, stellen wir in diesem Beitrag (ebenfalls originäre) Verfahren für automatisierte Impact-Analysen vor.

Als Grundlage für die Analysen stellen wir in Abschnitt 2 unser SOA-Repository-Metamodell auszugsweise vor. Abschnitt 3 beschreibt exemplarisch einige Analysen basierend auf aktuell existierenden Repository-Daten (Ist-Analysen). Planspiele und die zugehörigen Impact-Analysen werden in Abschnitt 4 betrachtet. Eine Diskussion des aktuellen Standes der Literatur sowie existierende Plattformen für SOA-Repositories findet sich in Abschnitt 5. Der Beitrag schließt mit einer Zusammenfassung.

2 Grundlagen: Metamodell des SOA-Repository

Für die Ermittlung der Anforderungen an ein SOA-Repository wurden in mehreren Geschäftsbereichen der Daimler AG Anforderungsanalysen durchgeführt [BTBR10, Bu12]. Hierbei ergab sich außer dem Bedarf an Analysen (siehe Abschnitte 3 und 4) auch die Notwendigkeit zur Speicherung diverser Objekt- und Beziehungstypen in einem SOA-Repository. Das entsprechende Metamodell zur Datenspeicherung ist in Abbildung 1 als Entity-Relationship-Modell dargestellt. Zur Beschreibung von Kardinalitäten wird die (min,max)-Notation verwendet [Ab74].

Das Metamodell umfasst sowohl fachliche (linke Hälfte) wie auch technische (rechte Hälfte) Objekttypen und ist auch bzgl. der konkreten Objekttypen sehr umfangreich. So werden u.A. fachliche und technische *DataObjectVersions*[2] berücksichtigt, weil auch diese für den Anwender z.B. beim Suchen bzw. Browsen nach verwendbaren Services hilfreich sein können. Ziel dieses Metamodells ist, sehr viele der häufig in einer SOA benötigten Informationen abzubilden, jedoch können abhängig vom jeweiligen Unternehmen und seiner SOA-Philosophie zusätzliche Objekt- und Beziehungstypen erforderlich werden bzw. dann irrelevante entfallen. Im Folgenden erläutern wir aus Platzgründen nur die für das Verständnis dieses Beitrags erforderlichen Objekt- und Beziehungstypen. Für eine Erläuterung der anderen und auch für eine Diskussion der jeweils relevanten Attribute wird auf [Bu12, Ti10] verwiesen.

Zentrales Objekt in einer SOA ist der *Service* und findet sich dementsprechend auch im Metamodell. Allerdings ist ein Service nur eine „Bündelung" von unterschiedlichen Service-Versionen, da sich ein „konkreter Service" im Laufe der Zeit weiterentwickelt. Die entstehenden Service-Versionen können unterschiedliche Ein-/Ausgabeparameter oder sogar Operationen haben, weshalb sie für den Consumer inkompatibel sind. Wir unterscheiden zudem in fachliche (*BusinessServiceVersion*, verbunden mit Beziehungstyp *isBusinessServiceVersion*) und technische Service-Versionen (*TechnicalServiceVersion*, verbunden mit *isTechnicalServiceVersion*). Fachlich wird ein Service so beschrieben, dass Anwender aus Fachabteilungen und Manager den Zweck und groben Inhalt des Service verstehen und bewerten können. Erst nachdem eine Realisierung der fachlichen Service-Version beschlossen wurde, entsteht eine technische Service-Version inkl. einer detaillierten Spezifikation u.a. der Schnittstelle (WSDL). Eine fachliche Service-Version kann auf unterschiedliche Arten umgesetzt werden – auch mehrfach – wobei sich die zugehörigen technischen Service-Versionen z.B. in den Datentypen der Parameter oder der Anzahl der Operationen unterscheiden können. Da technische Service-Versionen inhaltlich dieselbe Aufgabe erledigen, sich aber in technischen Details unterscheiden, werden mehrere von ihnen mittels *hasTechnServiceVersImplementation* derselben fachlichen Service-Version zugeordnet.

Nach Implementierung und Rollout (Deployment) einer technischen Service-Version entsteht eine Service-Installation (*ServiceInstall*), die ihr zugeordnet wird. Da die Service-Installation von einem konkreten IT-System bereitgestellt wird, ist sie zudem mit der Beziehung *providesServiceInstall* genau diesem *System* zugeordnet. Die Beziehung *hasProvider* beschreibt etwas „ungenauer", dass das IT-System diesen Service bereitstellt, wobei ein Service i.Allg. von unterschiedlichen IT-Systemen angeboten werden und mehrere Installationen besitzen kann, die evtl. zu verschiedenen Service-Versionen gehören.

Soll ein Service tatsächlich verwendet werden, so muss zwischen dem Anbieter und dem Consumer ein Nutzungsvertrag (*Contract*) geschlossen werden. Die Zuordnung des Contract zu dem Consumer erfolgt mittels der Beziehung *hasContract*. Die Information zu diesem Contract entsteht typischerweise sukzessive, so dass zuerst nur festgelegt wird, welche fachliche Service-Version genutzt werden soll (Zuordnung mittels *has-*

[2] Versionen von Datentypen, die in Service-Operationen als Ein- oder Ausgabeparameter verwendet werden.

BusinServiceVersConsumer). Zu diesem Zeitpunkt muss noch gar keine technische Service-Version hierzu existieren. Wird dann auch eine solche ausgewählt, so wird der Contract ergänzt (evtl. auch um technische Service-Level-Agreements) und die Zuordnung zum Contract erfolgt mittels *hasTechnServiceVersConsumer*.

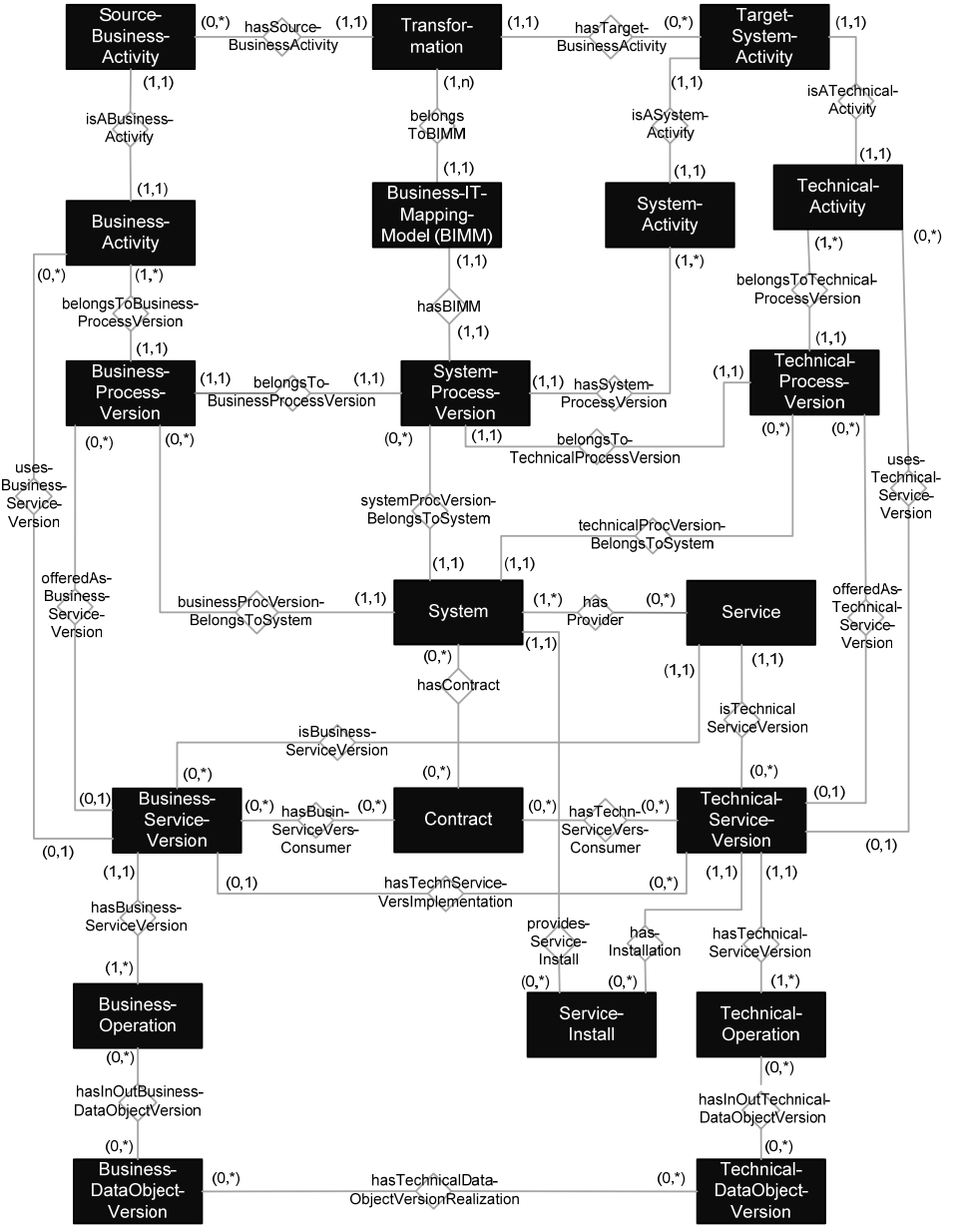


Abbildung 1: Metamodell des SOA-Repository als Entity-Relationship-Diagramm

Die soeben vorgestellten Entitäts- und Beziehungstypen werden wir im Folgenden verwenden, wofür noch die Notation erläutert werden soll: $e \in E$ ist erfüllt gdw. im SOA-Repository die Entität e in der Entitätsmenge E vorhanden ist. Für den Beziehungstyp b zwischen Entitätstyp E und E' mit $e \in E$ und $e' \in E'$ liefert die Funktion $b(e, e')$ wahr gdw. eine entsprechende Beziehung zwischen e und e' im SOA-Repository existiert.

3 Analyse bereits existierender SOA-Repository-Daten (Ist-Analysen)

Ist-Analysen basieren auf bereits im SOA-Repository vorhandenen Daten. Da Contracts jedoch auch in der Zukunft liegende Termine für Beginn und Ende einer Service-Nutzung beinhalten, können dadurch nicht nur bereits bestehende, sondern auch noch zu erwartende Problemsituationen identifiziert werden. Deshalb ermöglichen bereits Ist-Analysen die rechtzeitige Reaktion auf zu erwartende Problemsituationen. In diesem Abschnitt werden exemplarisch einige Ist-Analysen vorgestellt (weitere in [Bu12, Ti10]).

3.1 Technische Service-Version ohne Installation bzw. fachliche Service-Version

Ein sehr einfaches Analyseszenario ist die Identifikation von technischen Service-Versionen, für die keine Service-Installation existiert. Dieser Fall ist interessant, da eine solche technische Service-Version aktuell nicht nutzbar ist. Deshalb muss im Rahmen der SOA-Governance entschieden werden, ob eine entsprechende Service-Implementierung und -Installation durchgeführt werden soll oder die Service-Version nicht benötigt wird, also gelöscht werden kann. Wie alle Analysen basiert die Ermittlung solcher technischer Service-Versionen *TSV* auf den Daten aus dem SOA-Repository:
$TSV = \{tsv \in TechnicalServiceVersion \mid \nexists\, si \in ServiceInstall$ mit $hasInstallation(tsv, si)\}$

3.2 Nicht nutzbare Service-Versionen

Mit ähnlichen Regeln ist basierend auf den SOA-Repository-Daten ebenfalls einfach analysierbar, welche fachlichen und technischen Services aktuell nicht nutzbar sind. So kann zu jedem Service ausgelesen werden, welche BusinessServiceVersion existieren (mittels Beziehung isBusinessServiceVersion in Abbildung 1). In dem in Abbildung 2 dargestellten Beispiel ergeben sich für ServiceX die fachlichen Service-Versionen BSV1, BSV2 und BSV3. Falls einem dieser keine technische Service-Version zugeordnet ist, so ist dieser nicht nutzbar (BSV3 bzw. das unten dargestellte Szenario 3c). Für BSV1 existieren die technischen Service-Versionen TSV1a und TSV1b, für BSV2 existieren TSV2a und TSV2b. Ebenso ist erkennbar, dass z.B. TSV1b keine Service-Installation zugeordnet ist, so dass diese technische Service-Version nicht direkt nutzbar ist (siehe Szenarios 3a und 3b).

Wir betrachten nun, wann eine nicht nutzbare fachliche bzw. technische Service-Version zu Problemen führt und welche Maßnahmen zu deren Lösung ggf. ergriffen werden können. Die Analyse kann hierbei automatisch erfolgen; das Festlegen von Maßnahmen

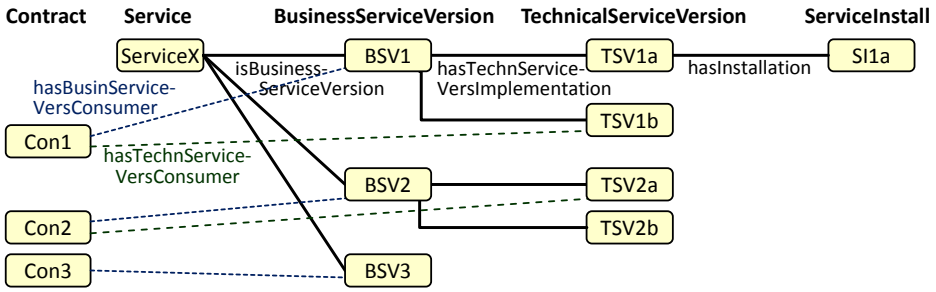


Abbildung 2: Beispieldaten aus dem SOA-Repository zu Contracts, Services, etc.

bleibt selbstverständlich eine intellektuelle Leistung, die z.B. von einem Governance-Gremium übernommen wird. Wir unterscheiden folgende Szenarien:

Szenario 1 (Kein Contract für eine nicht nutzbare fachliche Service-Version): In Abbildung 2 ist BSV2 nicht nutzbar, weil ihm keine Service-Installation zugeordnet ist (über TSV2a oder TSV2b). Angenommen, für BSV2 gibt es keinen relevanten Contract (z.B. weil das in Con2 vereinbarte Nutzungsende in der Vergangenheit liegt), dann besteht kein Problem für die Stabilität der Service-Consumer. Allerdings ist BSV2 nicht nutzbar, so dass die aktuelle Situation verändert werden sollte: Einerseits kann beschlossen werden, dass für TSV2a oder TSV2b eine Service-Installation entwickelt wird. Andererseits kann BSV2 auch für unnötig befunden und deshalb sein „Sun-down" beschlossen werden (inkl. TSV2a und TSV2b). Dies ist z.B. der Fall, wenn sich BSV2 als ungeeignet erwiesen hat, weshalb die zu TSV2a/b gehörenden Service-Installationen bereits abgeschaltet und aus dem SOA-Repository entfernt wurden.

Szenario 2 (Kein Contract für nicht nutzbare technische Service-Version): Gibt es für eine nicht nutzbare technische Service-Version (z.B. TSV1b in Abbildung 2) keinen gültigen Contract, dann besteht wieder kein unmittelbares Problem. Das Governance-Gremium kann entscheiden, ob für TSV1b eine Service-Installation erstellt werden soll (bzw. bereits in Entwicklung ist) oder ob diese technische Service-Version verzichtbar ist, weil sie gegenüber der alternativ verwendbaren TSV1a kaum oder keine Vorteile hat.

Szenario 3 (Contract für nicht nutzbare fachliche bzw. technische Service-Version): Interessanter ist der Fall, wenn die in Abbildung 2 dargestellten Contracts noch nicht abgelaufene Nutzungszeiten beinhalten. Dann existieren relevante Contracts für nicht nutzbare Service-Version. Dieses Szenario unterscheiden weiter nach dem Kriterium, ob alternativ nutzbare Service-Versionen existieren:

Szenario 3a (Technische Service-Version mit nutzbarer alternativer technischer Service-Version): In Abbildung 2 existiert der Contract Con1 für TSV1b, so dass es für TSV1b auch tatsächlich einen Consumer gibt bzw. geben wird. Dies ist problematisch, da TSV1b über keine Service-Installation verfügt. Allerdings gibt es eine weitere technische Service-Version TSV1a, die derselben fachlichen Service-Version zugeordnet ist und zudem nutzbar ist (wg. SI1a). Solche technischen Service-Versionen *TSV* können mittels der Repository-Daten durch die folgenden Kriterien automatisch erkannt werden:

$TSV = \{tsv \in TechnicalServiceVersion \mid$
$\exists\ con \in Contract$ mit $hasTechnServiceVersConsumer(con, tsv) \wedge$
$\nexists\ si \in ServiceInstall$ mit $hasInstallation(tsv, si) \wedge$
$\exists\ bsv \in BusinessServiceVersion$ mit $hasTechnServiceVersImplementation(bsv, tsv) \wedge$
$\exists\ tsv' \in TechnicalServiceVersion$ mit $hasTechnServiceVersImplementation(bsv, tsv') \wedge$
$\exists\ si' \in ServiceInstall$ mit $hasInstallation(tsv', si')$

Als Lösung kann in diesem Szenario von der nicht nutzbaren technischen Service-Version TSV1b zur nutzbaren TSV1a gewechselt werden. Dies ist sinnvoll, wenn TSV1b gegenüber TSV1a kaum Vorteile hat, erfordert aber eine Anpassung von Con1 und der zugehörigen Consumer-Applikation (falls diese bereits realisiert wurde). Alternativ kann auch die Entwicklung einer Service-Installation für TSV1b beschlossen oder auf den Abschluss einer ohnehin stattfindenden Entwicklung gewartet werden. Unterscheiden sich TSV1a und TSV1b nur geringfügig (z.B. andere, aber ähnliche Datenstrukturen für Ein-/Ausgabeparameter), so kann eine Service-Installation für TSV1b auch dadurch realisiert werden, dass SI1a mittels eines Proxy (ESB) gerufen wird, der die Unterschiede zwischen Consumer und Service durch geeignete Transformationen (der Datenstrukturen) ausgleicht. Eine solche Lösung nutzt die durch eine SOA ermöglichte Flexibilität [BBR11].

Szenario 3b (Technische Service-Version ohne nutzbare alternative technische Service-Version) In Abbildung 2 referenziert Con2 die nicht nutzbare technische Service-Version TSV2a. Die einzige ebenfalls zu BSV2 gehörende technische Service-Version TSV2b ist als Alternative nicht nutzbar, da sie über keine Service-Installation verfügt. Solche technischen Service-Versionen *TSV* sind wie folgt erkennbar:
$TSV = \{tsv \in TechnicalServiceVersion \mid$
$\exists\ con \in Contract$ mit $hasTechnServiceVersConsumer(con, tsv) \wedge$
$\nexists\ si \in ServiceInstall$ mit $hasInstallation(tsv, si) \wedge$
$\exists\ bsv \in BusinessServiceVersion$ mit $hasTechnServiceVersImplementation(bsv, tsv) \wedge$
$\nexists\ tsv' \in TechnicalServiceVersion$ mit ($hasTechnServiceVersImplementation(bsv, tsv')$
$\wedge \exists\ si' \in ServiceInstall$ mit $hasInstallation(tsv', si')$)

Selbstverständlich kann auch in diesem Szenario eine geeignete Service-Installation für TSV2a realisiert werden. Soll dies vermieden werden, so sind die Lösungsmöglichkeiten eingeschränkt: Da keine zur selben fachlichen Service-Version gehörende und zudem nutzbare alternative technische Service-Version existiert, bleibt nur der Wechsel zu einer anderen fachlichen Service-Version. Da diese nutzbar sein soll, muss dies in dem in Abbildung 2 dargestellten Beispiel BSV1 sein. Zu BSV1 kann gewechselt werden, indem Con2 und ggf. die Consumer-Applikation entsprechend angepasst werden. Die Veränderungen für letztere sind typischerweise deutlich umfangreicher als bei Szenario 3a, weil sich fachliche Service-Versionen i.d.R. viel stärker unterscheiden als technische.

Szenario 3c (Fachliche Service-Version ohne technische Service-Versionen): Für den Contract Con3 existiert lediglich eine fachliche Service-Version, aber keine korrespondierende technische Service-Version. Dies kann auftreten, wenn die zu realisierende technische Service-Version noch nicht endgültig beschlossen oder spezifiziert wurde. Prinzipiell muss dann auf eine entsprechende technische Service-Version gewartet und

dann der Contract vervollständigt werden. Alternativ kann wie in Szenario 3b zu einer anderen (bereits nutzbaren) fachlichen Service-Version gewechselt werden. Auf die formalen Kriterien zur Erkennung dieses Szenarios verzichten wir aus Platzgründen.

4 Planspiele mit zukünftig möglichen Daten (Impact-Analysen)

Die bisher vorgestellten Analysen betrachten ausschließlich die gegenwärtig im SOA-Repository dargestellte Situation. Die im Folgenden exemplarisch (weitere in [Ti10]) beschriebenen Impact-Analysen betrachten darüber hinaus Auswirkungen möglicher Änderungen in der Zukunft. Mittels solcher Analysen können Szenarien verglichen werden, was dann die Grundlage für Entscheidungen bildet. So kann z.B. analysiert werden, welche Service-Consumer von einer Service-Abschaltung zu einem bestimmten Zeitpunkt betroffen wären, und damit, welcher Abschaltzeitpunkt am günstigsten ist.

4.1 Indirekte Abhängigkeiten (Ripple-Effect)

Die hohe Wichtigkeit von Impact-Analysen ergibt sich, weil von z.B. der Abschaltung einer Service-Installation nicht nur deren unmittelbare Consumer betroffen sind: Diese Consumer sind IT-Systeme (System in Abbildung 1), die selbst Services bereitstellen können. Muss ein solches IT-System verändert werden, weil etwa auf einen alternativen (zu konsumierenden) Service ausgewichen werden muss, so können sich auch Veränderungen an bereitgestellten Services ergeben. Gibt es keinen alternativen Service, dann können Services evtl. überhaupt nicht mehr bereitgestellt werden. In beiden Fällen sind deren Consumer davon betroffen, so dass sich die Änderung über mehrere Stufen zu indirekt abhängigen IT-Systemen „fortpflanzen“ kann (Ripple-Effect).

Solche indirekten Abhängigkeiten lassen sich mit den in Abschnitt 2 beschrieben Repository-Daten automatisch erkennen. In dem in Abbildung 3 dargestellten Beispiel soll die Service-Installation SI1 in naher Zukunft abgeschaltet werden. Sie sei die einzige Realisierung der technischen Service-Version TSV1, wodurch aufgrund der Contracts Con1a, Con1b und Con1c die IT-Systeme Sys1a, Sys1b und Sys1c nicht mehr funktionsfähig sind und deshalb rechtzeitig angepasst werden müssen. Verändern sich dadurch z.B. die von Sys1b angebotenen Services Serv2 und Serv3 bzw. deren technische Service-Version TSV2 und TSV3, so pflanzt sich die Änderung fort: Die hiervon abhängigen Contracts Con2, Con3a und Con3b müssen ebenso angepasst werden, wie die Service-konsumierenden IT-Systeme, die natürlich selbst wieder Services anbieten können, etc.

Die von einer abzuschaltenden oder zu verändernden Service-Installation *si* direkt betroffenen IT-Systeme lassen sich mittels folgendem Kriterium automatisch ermitteln:
AffectedSys(si) = {*sys* ∈ *System* | ∃ *tsv* ∈ *TechnicalServiceVersion* mit
hasInstallation (*tsv*, *si*) für die gilt: ∃ *con* ∈ *Contract* mit
hasTechnServiceVersConsumer(*con*, *tsv*) und *hasContract* (*sys*, *con*) }

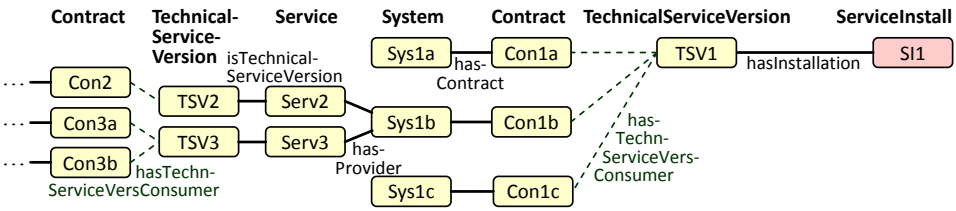


Abbildung 3: Beispieldaten aus dem SOA-Repository mit Ripple-Effect

Indirekt betroffene Systeme lassen sich dann ermitteln, indem man mit der Beziehung *providesServiceInstall* alle von einem System *sys* ∈ *AffectedSys(si)* bereitgestellten Service-Installationen *AffectedServiceInstall* ermittelt und für jedes *si'* ∈ *AffectedService-Install* eine entsprechende Menge *AffectedSys(si')* der nächsten Stufe berechnet, etc.

Wenn man lediglich die Beziehungen zwischen IT-Systemen (System) und Ausprägungen von Services (Service-Install, etc.) betrachtet, dann können ungünstigerweise auch Abhängigkeiten berechnet werden, die in Wirklichkeit überhaupt nicht existieren. Dass die Programmierung eines IT-Systems (irgendwo) einen bestimmten Service aufruft, bedeutet noch nicht, dass dieser für die Bereitstellung jedes angebotenen Services notwendig ist. So sind in dem erläuterten Beispiel das System Sys1b und damit die bereitgestellten Service Serv2 und Serv3 von der abzuschaltenden Service-Installation SI1 abhängig. Fall 1: Angenommen, die Implementierung des konkreten Services TSV2/Serv2 basiert auf TSV1 und SI1. Falls dann die Anpassung der Implementierung von Sys1b zu einer Veränderung der angebotenen Service-Schnittstelle führt, so hat dies Auswirkungen auf den Contract Con2 und die konsumierende Applikation. Fall 2: Angenommen, die Implementierung von TSV3/Serv3 verwendet keinerlei fremde Services oder zumindest nicht den abzuschaltenden SI1. Dann kann die technische Service-Version TSV3 unverändert bleiben, so dass es keinerlei Auswirkungen auf die Contracts Con3a und Con3b gibt. Da die konsumierenden IT-Systeme also nicht betroffen sind, pflanzt sich die Veränderung auf diesem Weg nicht weiter fort. Die in heutigen SOA-Repositories gespeicherten Informationen sind jedoch nicht detailliert genug, um diese beiden Fälle unterscheiden zu können. Deshalb würde eine Analyse stets irrtümlich zu viele Problemsituationen erkennen.

Das diesem Beitrag zugrunde gelegte SOA-Repository (vgl. Abbildung 1) ermöglicht es, die beiden Fälle zu unterscheiden: Durch die Beziehung *usesTechnicalServiceVersion* ist es möglich, exakt diejenigen Prozessschritte (*TechnicalActivity*) eines Geschäftsprozesses zu bestimmen, die wegen eines entsprechenden Serviceaufrufs von einer abzuschaltenden technischen Service-Version abhängig sind. Mittels *belongsToTechnical-ProcessVersion* lässt sich die zugehörige *TechnicalProcessVersion* (der ausführbare Geschäftsprozess / Workflow) ermitteln. Wird dieser wiederum als Service angeboten, so liefert die Beziehung *offeredAsTechnicalServiceVersion* die zugehörige *Technical-ServiceVersion*. Hat schließlich ein IT-System hierzu einen Contract, so ist es tatsächlich von der Änderung betroffen, es liegt also Fall 1 vor. Wir erhalten also das korrekte Analyseergebnis. Dadurch sind sehr viel bessere Planungen und Entscheidungen möglich, als wenn nur Informationen über die insgesamt von einem IT-System konsumierten und bereitgestellten Services im SOA-Repository existieren.

Die für das SOA-Repository benötigten Informationen entstehen bei Verwendung des ENPROSO-Konzepts ohnehin durch die zugrundeliegende durchgängige Modellierung vom fachlichen Geschäftsprozess, über technische Workflows, bis hin zu Services der SOA-Applikationen [BBR12, Bu12]. Da diese Informationen automatisch ins SOA-Repository übertragen werden können, entsteht kaum Mehraufwand für die Datenerfassung. Wird nicht von prozess-orientierten SOA-Applikationen ausgegangen, ist der Ansatz generell auch z.B. auf UML-basierte Modellierung von Services, Service-Implementierungen und deren Consumer übertragbar.

4.2 Analysen möglicher Szenarien für zukünftige Zeitpunkte (Planspiele)

Wie stark die Auswirkungen einer in der Zukunft durchzuführenden Änderung sind, hängt vom genauen Zeitpunkt ihrer Umsetzung ab. Dementsprechend soll der Zeitverlauf in die Analysen einbezogen werden. Indem Analysen für mehrere Zeitpunkte durchgeführt werden, können unterschiedliche Szenarien durchgespielt und analysiert werden. Solche Planspiele möchten wir nun exemplarisch vorstellen.

Angenommen, es soll die aktuell einzige Service-Installation SI0 der technischen Service-Version TSV0 aus Kostengründen in ca. 6 Monaten abgeschaltet werden. Es ist nun zu analysieren, ob dieser Termin aus Sicht des Gesamtunternehmens günstig ist. Als ersten Schritt bietet es sich an, eine graphische Übersicht der Service-Nutzungsverträge (Contracts) im Verlauf der Zeit zu erstellen (vgl. Abbildung 4). Eine solche Darstellung kann aus den Daten des SOA-Repositories automatisch generiert werden. Im Diagramm ist sehr einfach erkennbar, dass aktuell 4 Contracts für TSV0 existieren und dementsprechend 4 IT-Systeme von SI0 abhängig sind. Nach ca. 6, 10, 17 und 24 Monaten laufen Contracts aus. Diese Termine bieten sich als Abschaltzeitpunkte für TSV0 an, da dann jeweils weniger IT-Systeme betroffen sind.

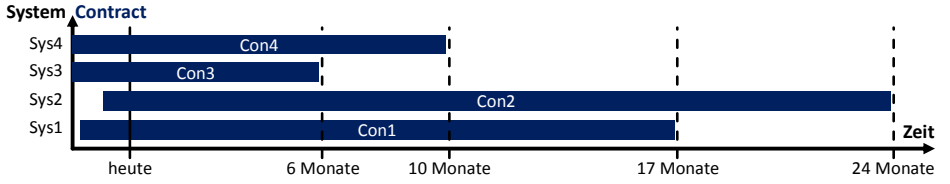


Abbildung 4: Graphische Darstellung der Contracts für TSV0 im Zeitverlauf

Die eben betrachtete Analyse berücksichtigt jedoch noch keine indirekten Abhängigkeiten (Ripple-Effect), so dass die Menge der von der Abschaltung betroffenen IT-Systeme noch unvollständig ist. Der in Abbildung 5a dargestellt Impact-Graph stellt auch solche Abhängigkeiten dar, indem er jede Indirektionsstufe als separate Schale visualisiert. Hierbei werden die zu einem bestimmten Zeitpunkt gültigen Contracts berücksichtigt, bei Abbildung 5a die in 6 Monaten noch gültigen Contracts. Auch dieses Diagramm lässt sich automatisch aus den Daten des SOA-Repositories generieren, wobei die in Abbildung 5b dargestellten Abhängigkeiten angenommen werden: Außer den wegen Con1, Con2, Con4 abhängigen Systemen Sys1, Sys2, Sys4 gibt es noch eine indirekte Abhängigkeit zu Sys5. Das IT-System Sys4 bietet nämlich die Service-Installation SI4 an. Diese implementiert als einzige die technische Service-Version

TSV4, die über Con5 an Sys5 gebunden ist. Eine Abschaltung von SI0 in 6 Monaten würde also die 4 genannten IT-Systeme beeinträchtigen. Mit dieser Information kann auf die Systemeigner zugegangen werden, so dass diese frühzeitig die Folgen der Service-Abschaltung abschätzen und ggf. ihre Systeme geeignet anpassen können.

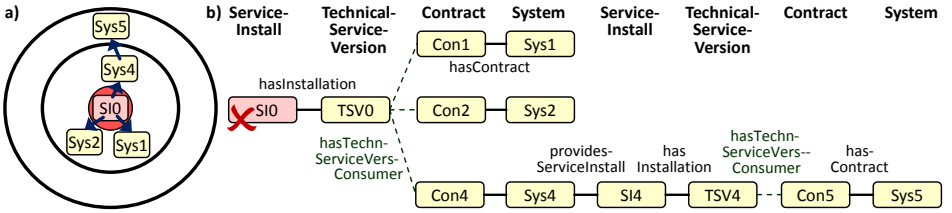


Abbildung 5: a) Impact-Graph und b) Abhängigkeiten von SI0 in 6 Monaten

Erweist sich der Anpassungsaufwand für diese 4 Systeme im Vergleich zu den Betriebskosten von SI0 als zu hoch, so sollte ein längerer Betrieb von SI0 in Erwägung gezogen werden. Als nächster Abschaltzeitpunkt bietet sich heute in 10 Monaten an, weil dann Con4 ausläuft und Sys4 nicht mehr betroffen ist (vgl. Abbildung 4). Dann kann auch das IT-System Sys5 nicht mehr (indirekt) betroffen sein, weil dessen Abhängigkeit ausschließlich über Sys4 bestand. Für die Darstellung des resultierenden Abhängigkeitsgraphen verweisen wir aus Platzgründen ebenso auf [BBTR14], wie für die Analyse der weiteren Abschaltzeitpunkte.

Durch solche Planspiele wird es möglich, abzuwägen, ob eine Service-Abschaltung eher früh umgesetzt werden soll (geringere Betriebskosten), oder ob es aus betriebswirtschaftlicher Sicht Sinn macht, länger zu warten (geringere Migrationskosten). Diese Entscheidung ist von dem zuständigen SOA-Governance-Gremium zu treffen, das durch automatisch aus dem SOA-Repository abgeleitete Informationen unterstützt wird.

5 Diskussion

In der wissenschaftlichen SOA-Literatur werden die hier vorgestellten Themen nicht unmittelbar adressiert (vgl. [BBTR14]). Jedoch gibt es zur Realisierung von SOA-Repositories Standards und Produkte: Das Metamodell des standardisierten Verzeichnisdienstes UDDI [OA02] ist sehr eingeschränkt und sieht auch keine Erweiterungen vor. Außer einigen vordefinierten Abfragen sind keine Analysen vorgesehen. Das Produkt IBM WSRR [Sa07] ermöglicht die Speicherung von technischen und stark eingeschränkt auch fachlichen Objekten. Allerdings ist WSRR bzgl. der Objekttypen und Beziehungen erweiterbar, so dass es als Basis zur Umsetzung des vorgestellten Metamodells verwendbar wäre. Die eher rudimentären, vorhanden Analysen sind um benutzerdefinierte Analysen erweiterbar. Ähnliches gilt für CentraSite Active SOA [Ro06], sowohl was die Objekttypen, die Erweiterbarkeit als auch was die Analysen betrifft. Sowohl Oracle Enterprise Repository [Or08] wie auch HP SOA Systinet [He10] basieren auf UDDI. Beide ermöglichen eine gegenüber dem UDDI-Standard erweiterte Speicherung von Objekttypen, so dass auch fachliche Objekte verwaltet werden können. Die Mächtigkeit der Visualisierung von Abhängigkeitsdiagrammen und der Analysen bleibt weit hinter

den in diesem Beitrag vorgestellten zurück. Das ARIS-Repository [ID08] ermöglicht es die Speicherung von sehr vielen (insb. fachlichen) Objekt- und Beziehungstypen. ARIS unterstützt (eingeschränkt mächtige) Analysen, die auch durch selbst realisierte Analysearten erweiterbar sind (vgl. [Bu12]).

6 Zusammenfassung und Ausblick

Wir haben ein umfassendes Metamodell für ein SOA-Repository vorgestellt, das als Basis für Analyse dient. So kann mit Ist-Analysen geprüft werden, ob aufgrund der gespeicherten Daten mit Problemen (z.B. wegfallende Service-Installation) zu rechnen ist. Mit Impact-Analysen kann simuliert werden, welche Auswirkungen zukünftige Änderungen haben werden, wobei auch die Fortpflanzung von Problemen (Ripple-Effect) berücksichtigt wird. All dies geht deutlich über die bisher in der SOA-Literatur erwähnte bzw. in Standards und Produkten vorgesehene Funktionalität hinaus. Allerdings ermöglicht die Erweiterbarkeit vieler Produkte die Umsetzung der vorgestellten Konzepte. So wurden insbesondere die Analyseverfahren mittels eines technischen Prototyps getestet [Ti10]. Eine vollständige Umsetzung des Metamodells und aller Analysen in einem SOA-Repository-Produkt sowie eine Erprobung des Ansatzes mittels einer Fallstudie wären wünschenswert, konnten bisher aber noch nicht realisiert werden.

Literaturverzeichnis

[Ab74] Abrial, J.R.: Data Semantics. Proc. IFIP Working Conf. Data Base Management, 1974.

[BBR11] Buchwald, S.; Bauer, T.; Reichert, M.: Flexible Prozessapplikationen in Service-orientierten Architekturen - Ein Überblick. EMISA Forum, 31(3), 2011.

[BBR12] Buchwald, S.; Bauer, T.; Reichert, M.: Bridging the Gap Between Business Process Models and Service Composition Specifications. In (Lee et.al., Hrsg): Service Life Cycle Tools and Technologies: Methods, Trends, and Advances. IGI Global, Hershey, 2012.

[BBTR14] Bauer, T.; Buchwald, S.; Tiedeken, J.; Reichert, M.: Erhöhung der System-Stabilität und -Flexibilität durch ein SOA-Repository mit Analysefähigkeiten. HNU Working Paper Nr. 32, Hochschule Neu-Ulm, 2014.

[BTBR10] Buchwald, S.; Tiedeken, J.; Bauer, T.; Reichert, M.: Anforderungen an ein Metamodell für SOA-Repositories. Proc. 2nd Central-European Workshop on Services and their Composition, 2010.

[Bu12] Buchwald, S.: Erhöhung der Durchgängigkeit und Flexibilität prozessorientierter Applikationen mittels Service-Orientierung. Dissertation, Universität Ulm, 2012.

[He10] Hewlett-Packard Development Company: HP SOA Systinet Software Data Sheet, 2010.

[ID08] IDS Scheer: ARIS SOA Architect - Geschäftsprozesse als Grundlage für Service-orientierte Architekturen. IDS Scheer, 2008.

[OA02] OASIS: Universal Description, Discovery, and Integration (UDDI). Version 3.0, 2002.

[Or08] Oracle: Oracle Enterprise Repository and Oracle Service Registry for the SOA-Lifecycle. Oracle Data Sheet, 2008.

[Ro06] Rogers, S.: CentraSite: An Integrated SOA Registry and Repository. Software AG, 2006.

[Sa07] Sachdeva, N.: Customize the WebSphere Service Registry and Repository Governance Profile. IBM developerWorks, 2007.

[Ti10] Tiedeken, J.: Konzeption und Realisierung eines logisch zentralen SOA-Repositories. Diplomarbeit, Universität Ulm, 2010.

Forensic Accounting - Potenziale von Big Data und Predictive Analytics im Rahmen der IT Governance

Rainer Schmidt[1], Michael Möhring[2], Robert Rieg[3], Stefan Maier[3], Gianfranco Walsh[2], Ralf-Christian Härting[3]

[1]Fakultät für Informatik und Mathematik
Hochschule München
Lothstraße 64, 80335 München
Rainer.Schmidt@hm.edu

[2]Wirtschaftswissenschaftliche Fakultät - Lehrstuhl für Allgemeine Betriebswirtschaftslehre und Marketing
Friedrich-Schiller-Universität Jena, Carl-Zeiß-Straße 3, 07443 Jena,
Michael.Moehring@uni-jena.de, Walsh@uni-jena.de

[3]Fakultät für Wirtschaftswissenschaften
Hochschule Aalen
Beethovenstraße 1, 73430 Aalen
Robert.Rieg@htw-aalen.de, Stefan.Maier@htw-aalen.de, Ralf.Haerting@htw-aalen.de

Abstract: Die Informationstechnik stellt ein wertvolles Instrumentarium für die Aufdeckung doloser Handlungen und die Prüfung von Regeleinhaltungen im Rechnungswesen bereit. Mit der exponentiellen Zunahme relevanter Daten für das Rechnungswesen, sowohl in strukturierter als auch wenig bis unstrukturierter Form, kommen bisherige Werkzeuge und Verfahren jedoch an ihre Grenzen in Bezug auf die Rechenzeit und Verarbeitungsmöglichkeit. Vor diesem Hintergrund bietet es sich an, zwei bisher nicht verbundene Felder miteinander zu verbinden, die Forensik im Rechnungswesen mit Big Data und Business Analytics. Der vorliegende Beitrag stellt die wesentlichen IT-Entwicklungen vor und diskutiert ihre Potentiale und Einsatzmöglichkeiten für das Rechnungswesen und die IT-Governance. Des Weiteren wird ein Bezugsrahmen für den Einsatz von Big Data und Predictive Analytics vorgestellt, der vor allem großen Unternehmen mit Massendaten, Prüfungsgesellschaften und Prüfungsbehörden Hinweise auf eine prototypische Vorgehensweise der datengetriebenen forensischen Analyse bietet.

1 Ausgangssituation Forensic Accounting

Die Wirtschaftskriminalität in deutschen Unternehmen und der daraus entstehende monetäre Schaden stellen hohe Risiken für Organisationen dar [Pric11]. In Deutschland wurden im Jahr 2012 knapp 82.000 Fälle doloser Handlungen mit einer kumulierten Schadenssumme von ca. 3,7 Mrd. Euro verzeichnet. Dolose Handlungen sind illegale Handlungen, welche durch vorsätzliche Täuschung, Verschleierung oder Vertrauensmissbrauch gekennzeichnet sind [Frei08]. Darunter fallen beispielsweise

Zahlungen, welche Mitarbeiter nicht konform vornehmen und Dritten gegenüber ausbezahlen. Durch sie kann der Unternehmung bzw. Organisation ein erheblicher Schaden entstehen und die Erreichung der Unternehmensziele gefährdet werden [Frei08]. Die Erkennung und das (präventive) Management doloser Handlungen stellt eine wichtige Aufgabe der internen Revision dar [Frei08]. Im Rahmen der IT Governance muss u. a. sichergestellt werden, dass alle IT-gestützten Unternehmensprozesse gesetzlichen und unternehmensspezifischen Regulatorien (Compliance) entsprechen und Geschäftsziele unterstützen [MeZK03]. Neben innerorganisatorischen Regularien nimmt die Anzahl an gesetzlichen Bestimmungen, wie der Sarbanes-Oxley Act, Basel III, KonTrAG und Solvency II, zu und betrifft damit neben den Geschäftsabläufen auch die IT-Governance [JoGo06]. IT-Forensik stellt damit ein wesentlicher Bestandteil zur Aufdeckung von Compliance-Verstößen im Rahmen der Governance dar [Kirm12]. Unternehmen müssen gemäß ihrer Sorgfaltspflicht [Bund65] und weiteren gesetzlichen Bestimmungen (wie bspw. der Verhinderung von Geldwäsche [Bund08]) betrügerischen Handlungen entgegen wirken. Bei der Abschlussprüfung müssen weiterhin Prüfungshandlungen und Ergebnisse mit Bezug auf betrügerische Handlungen dokumentiert werden [Idw06] [Prüf02].

Dolose Handlungen können auf unterschiedliche Weise erfolgen, wobei oft IT-Systeme genutzt werden [KLSH04] [Herz06] [BeKu12]. So können bspw. Zahlungen über ein Enterprise-Ressource-Planning (ERP) System veranlasst werden oder die Planung und Durchführung von dolosen Handlungen kann in elektronischen Texten wie bspw. E-Mails erfolgen. Da die angebotenen Funktionen und die Menge der verfügbaren Daten in IT-Systemen kontinuierlich steigen [GaRe11], ist auch bei der Suche nach dolosen Handlungen der IT-Einsatz unumgänglich und damit die Durchführung automatisierter IT-gestützter Analyseverfahren unerlässlich.

Bisherige Datenanalyseverfahren zur Missbrauchserkennung können die erforderliche Prüfungstiefe immer weniger gewährleisten. So werden isolierte Verfahren der explorativen Statistik sowie zur Jahresabschlussprüfung durch die Menge der zu verarbeitenden Daten überfordert. Sie können zudem die Prüfungen nicht mehr in der verfügbaren Zeit durchführen und haben grundsätzliche Schwierigkeiten bei der Analyse unstrukturierter Daten (wie bspw. E-Mails zwischen Kunden und Lieferanten). Weiterhin ist bei der Analyse unstrukturierter Daten auf die Einhaltung von gesetzlichen Bestimmungen zu achten.

Ziel des vorliegenden Beitrags ist es deshalb, einen neuartigen datengetriebenen Ansatz des forensischen Rechnungswesens im Rahmen des IT-Prüfungswesens vorzustellen. Er stützt sich dabei auf Fortschritte der Datenverarbeitung und -analyse, insbesondere bei den Faktoren Volumen, der Geschwindigkeit und der Varianztoleranz, die unter dem Schlagwort „Big Data“ subsummiert werden. Der Einsatz von neuen technologischen Entwicklungen im Rahmen von Big Data kann die Analyse von großen Datensätzen, unterschiedlichster Struktur in kurzer Zeit ermöglichen. Bisherige Forschung im Kontext des forensischen Rechnungswesens konzentrierte sich auf die Analyse strukturierter Daten ohne Anwendung von Big-Data-Analysen [PLSG10], [Herz06]. Über den Einsatz von Big Data in deutschen Unternehmen liegen nur wenige Forschungsergebnisse und dokumentierte praktische Erfahrungen vor [PLSG10], [Herz06].

Zunächst sollen dazu, Begriffe im Umfeld von Big Data definiert und in ihrem Verhältnis geklärt werden. Danach werden Methoden der datenbasierten Forensik vorgestellt. Es folgt die Einführung eines Bezugsrahmens zur IT basierten Unterstützung forensischer Methoden im Rechnungswesen mit Big Data.

2 Big Data zur betrieblichen Entscheidungsunterstützung

Die verstärkte Nutzung von Daten zur Entscheidungsfindung ist in jüngerer Zeit verstärkt in den Fokus praxeologischer und wissenschaftlicher Diskussionen gerückt [BuCM13]. Auslöser ist die schnelle Verbreitung von Technologien und Konzepten unter dem Schlagwort Big Data [MSMK13]. Häufig werden Beispiele aus dem Marketing herangezogen, um die Vorzüge von Big Data zu belegen [Hays04], [SMMP14]. Insbesondere wird auf die Steigerung der Produktivität, bspw. bei der Neukundenakquisition, durch den Einsatz datengetriebener Entscheidungsverfahren verwiesen [BrHK11].

Im Zusammenhang mit Big Data tauchen eine Reihe weiterer Begriffe auf, die teilweise überdeckend verwendet werden und die es deshalb abzugrenzen gilt. Unter Business Analytics [ChCS12] werden in erster Linie vorhersagende (predictive), aber auch beschreibende (descriptive) statistische Verfahren verstanden, die die Grundlage für Optimierungsmaßnahmen und die Unterstützung für Entscheidungsprozesse liefern. Das Objekt von Business Analytics kann variieren. Wichtige Bereiche sind die Prozess- und die Service-Analytik. Bei der Prozessanalytik werden Größen wie Kosten und Zeiten während der Durchführung von Geschäftsprozessen erfasst. Eine Sonderform stellt das Process Mining [KeMU06] dar, das der Rekonstruktion von Prozessmodellen dient. Die Service-Analytik wird insbesondere im Rahmen der Überwachung von Dienstgütevereinbarungen verwendet.

Business Analytics ist häufig eng verbunden mit der Visualisierung von Analyseergebnissen. Business Analytics stützt sich häufig auf eine Kombination von Kompetenzen aus Statistik, Datenbanken und Domänenwissen, die mit dem Begriff Data Science [KeMU06] zusammengefasst wird. Von diesen Begriffen ist außerdem noch Business Intelligence [KeMU06] zu unterscheiden. Hierbei handelt es sich um einen ganzheitlichen Management-Ansatz, der bestrebt ist, alle für die betriebliche Entscheidungsfindung erforderlichen Faktoren zu koordinieren und zu optimieren.

Big Data

Im Allgemeinen wird unter Big Data eine deutliche Steigerung der Verarbeitungsfähigkeit von Daten bezüglich der Dimensionen Volumen (Volume), Geschwindigkeit (Velocity) und Vielfalt (Variety) verstanden [MCBB11]. Big Data löst daher bestehende Konzepte wie Business Intelligence [KeMU06] oder Business Analytics [DaHM10] nicht ab. Big Data ist keine spezifische Technologie und kein Algorithmus, sondern steht vielmehr für das Zusammenlaufen einer Reihe von technologischen Fortschritten. Die dabei auftretenden Steigerungen der Verarbeitungsfähigkeiten sind jedoch so umfangreich, dass

von einer neuen Qualität der Entscheidungsunterstützung gesprochen werden kann [DaBB12]. Typisch für Big Data ist der Verzicht auf die Leistungssteigerung einzelner Rechensysteme (Scale up) und stattdessen die Nutzung von möglichst vielen, parallel arbeitenden, einfachen Systemen (Scale Out) [AgDE11]. Eine häufig angewandte Strategie ist, die Bewegung von Daten zu vermeiden und stattdessen mehrere Berechnungsschritte lokal durchzuführen [AgDE11] .

Mit Big Data geht eine deutliche Steigerung der Fähigkeit einher, Daten mit großer Vielfalt zu verarbeiten [Bitk12]. So war die betriebliche Datenverarbeitung lange darauf beschränkt, klar strukturierte Informationen, die in Datenbanken abgelegt waren, zu verarbeiten. Durch den Einsatz von Big Data Technologien ist nun auch die Verarbeitung von Daten mit variabler Struktur möglich, d.h. auch Daten zu erfassen, deren Struktur nicht fix ist, sondern immer wieder neu in Erfahrung gebracht werden muss. Typische Einsatzmöglichkeiten von Big Data sind beispielsweise die Erzeugung von Produktempfehlungen die außerhalb der bisherigen Suchvorgänge des Kunden liegen, indem die Ähnlichkeit des Kunden mit anderen Kunden genutzt wird, als auch die Erhöhung der Prognosegenauigkeit [KeMU06]. Auch kann Big Data dazu eingesetzt werden, die Retourenzahl im E-Commerce zu reduzieren [MSKW13]. Inwieweit Big Data Anwendungen für Zwecke des forensischen Rechnungswesens eingesetzt werden können ist in der betriebswirtschaftlichen Literatur bislang wenig erforscht.

3 Methoden der datenbasierten Forensik

Die datenbasierte Forensik [Oliv09] hat sich bisher stark auf Daten aus relativ leicht zu erschließenden Quellen konzentriert. So werden bevorzugt Daten aus betrieblichen Anwendungssystemen wie ERP und CRM herangezogen. Das Schwergewicht bilden dabei transaktionale Daten, die geschäftliche Transaktionen widerspiegeln. Diese sind meist in Datenbanksystemen gespeichert und verfügen über eine klar definierte Struktur und Semantik. Die Einbeziehung weiterer Informationsquellen wird durch zwei Barrieren behindert, eine strukturelle und eine semantische (siehe Abbildung 1). Die strukturelle Barriere bezeichnet den Umstand, dass große Mengen von Daten bisher nicht genutzt wurden, da sie nur in semistrukturierter oder unstrukturierter Form vorlagen, d.h. ihre Verwendung erfordert aufwändige Import- und Aufbereitungsvorgänge [Abit97].

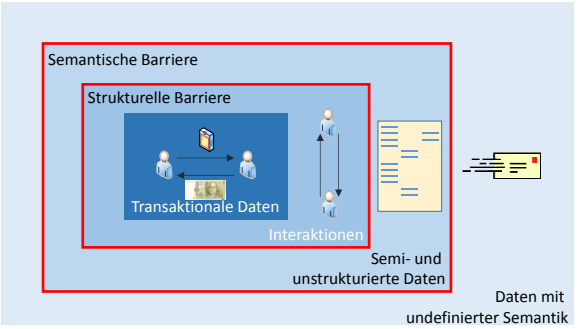


Abbildung 1: Barrieren für die Forensik

Die semantische Barriere verhindert die Nutzung von potentiell nützlichen Daten, da diese keine klar definierte Semantik besitzen. Dieser Umstand wird bspw. bei E-Mails deutlich. Diese sind im Gegensatz zu den durch die Schemadefinition klar beschriebenen Inhalten einer Datenbank kaum automatisch verarbeitbar.

Erweiterung der für die Forensik bereitstehenden Daten durch Big Data

Big Data erweitert deutlich die Menge der für die Forensik bereitstehenden Daten durch Überwindung der strukturellen Barriere und eine deutliche Verschiebung der semantischen Barriere, wie in Abbildung 2 dargestellt. So werden neben den strukturierten Daten beispielsweise solchen aus ERP-Systemen, auch weitere Datenklassen erfasst. Diese umfassen Daten außerhalb von geschäftlichen Transaktionen wie Interaktionen im Vorfeld oder im Nachlauf sowie benutzerzeugte Inhalte. Zu Interaktionen gehören beispielsweise Produkte deren Kauf zwar erwogen, aber dann doch noch verworfen wurde, indem sie aus dem Warenkorb entfernt wurden. Daten die Interaktionen wiedergeben, sind oft nur semi-strukturiert.

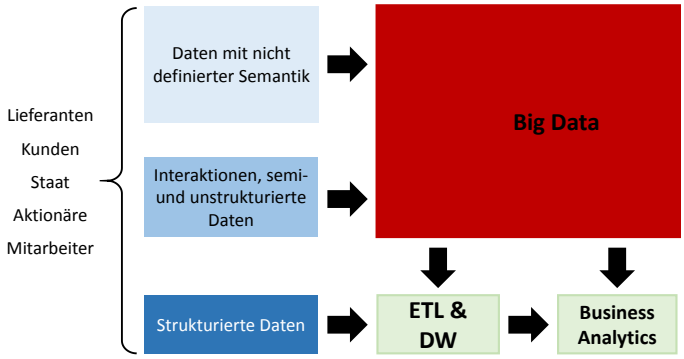


Abbildung 2: Big Data zur Vergrößerung des forensischen Potentials - in Anlehnung an [SMZW14]

Eine deutliche Verschiebung der semantischen Barriere findet bei der Einbeziehung benutzererzeugter Inhalte statt. Sind die bisher diskutierten Daten unter der Kontrolle des Unternehmens, können die Entstehung und Gestaltung benutzergenerierter Inhalte nicht mehr vom Unternehmen gesteuert werden. Erst recht ist es nicht möglich, die Bedeutung der verwendeten Begriffe zu standardisieren und somit einer direkten automatischen Bearbeitung zugänglich zu machen. Es liegt also eine sogenannte semantische Heterogenität [KeMU06] vor. Hierunter ist die unterschiedliche Interpretation von Begriffen in unterschiedlichen Kontexten zu verstehen. Bei den Daten unter der Kontrolle des Unternehmens ist eine eindeutige Interpretation möglich, bei den benutzergenerierten Inhalten ist dies nicht mehr der Fall. So kann es Unterschiede in der unternehmensseitigen Interpretation sowohl zwischen den unternehmenseigenen Daten und den benutzergenerierten Daten, als auch innerhalb der benutzergenerierten Inhalte geben. Letzteres bedeutet, dass der gleiche Begriff innerhalb der benutzergenerierten Inhalte unterschiedlich interpretiert werden muss.

4 Bezugsrahmen zur IT basierten Unterstützung forensischer Methoden im Rechnungswesen mit Big Data

Um dolose Handlungen wie etwa das Veruntreuen von Budgetteilen für organisationsfremde Zwecke zu erkennen, wird nachfolgender Bezugsrahmen zur IT gestützten Forensik im Rechnungswesen mit Big Data konzipiert (siehe Abbildung 3). Dolose Handlungen können im Unternehmen und in Beziehungen mit externen Partnern (wie bspw. Kunden, Lieferanten, Kapitalgebern und Behörden) auftreten [PLSG10] [Herz06]. Im Fokus dieses Bezugsrahmens werden derartige Handlungen in Zahlungsströmen bzw. Buchungen und Äußerungen der Beteiligten in Texten, wie z. B. (elektronischen) Handelsbriefen, untersucht. Auslöser dieser kriminellen Handlungen können externe Geschäftspartner, Fachmitarbeiter oder Manager sein [PLSG10] [Herz06].

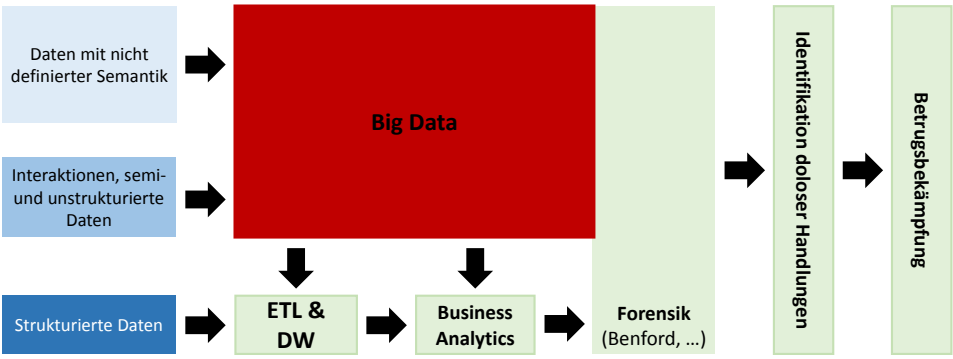


Abbildung 3: Bezugsrahmen zur IT basierten Unterstützung forensischer Methoden in Anlehnung an [SMZW14]

Die Verwendung von Big Data erweitert deutlich die für die Forensik und damit die Identifikation doloser Handlunge zur Verfügung stehende Datenmenge wie in Abbildung 3 dargestellt. Wurden früher nur strukturierte Daten verwandet, die über die Zwischenstationen Datawarehouse mit Extract, Transform & Load der Business Analytics zur Verfügung bereitgestellt wurden, kann heute auf ein deutlich größeres Spektrum von Daten zurückgegriffen werden. Auch können Big-Data zugehörige Technologien zur direkten Unterstützung der Forensik verwendet werden. Der Einsatz von Big Data ermöglicht es große Mengen von semi- und unstrukturierten, sowie Daten mit nicht definierter Semantik für die die Forensik zu erschließen. Diese stammen in großen Teilen von externen Partnern, wie Lieferanten, Kunden usw. Somit kann eine durchgängige Erkennung von Betrugsfällen im Rechnungswesen organisationsübergreifend sichergestellt werden. Anschließend werden verschiedene Analysemethoden zur Erkennung doloser Handlungen angewandt. Dies kann für unstrukturierte Daten, wie elektronische Handelsbriefe, mittels Text-Mining bzw. Text-Klassifikationsverfahren erfolgen. Für strukturierte Daten, wie bspw. bei Buchungssätzen, werden Schwellenwert bzw. Musterverfahren (bspw. Benford) angewandt. Zur beschleunigten Verarbeitung dieser großen Datenmengen variabler Struktur, werden Technologien aus dem Big Data

Umfeld (wie bspw. In-Memory-Computing oder Hadoop) angewandt. Diese Verfahren werden zur Steigerung der Robustheit, je nach Anwendungsfall, kombiniert. Auf Basis der Analyseergebnisse können mögliche Betrugsfälle identifiziert und in einem weiteren Schritt diesen entgegengewirkt werden.

Im Folgenden werden einige Verfahren beispielhaft kurz erläutert. Die Möglichkeiten der Unterstützung durch Big Data je Verfahren im Rahmen des Bezugsrahmens sind am Ende des Abschnitts beschrieben und in Tabelle 3 dargestellt.

Analyse mittels Text-Mining

Oft werden dolose Handlungen im (elektronischen) Schriftverkehr von den jeweiligen Beteiligten kodifiziert [BBBK11]. Vor allem der elektronische Geschäftsverkehr via E-Mail zwischen Lieferanten und Kunden kann derartige Muster aufweisen. Jedoch stellt die Auswertung von unstrukturierten Daten (wie E-Mails) eine große Herausforderung für Unternehmen und Organisationen dar. Neue technologische Entwicklungen im Big Data Umfeld ermöglichen eine schnellere Verarbeitung von Texten, die semantische Heterogenität aufweisen, bspw. unterschiedliche Begriffsverwendungen. Zur Auswertung und Erkennung von Betrugsfällen kann eine Analyse mittels Text-Mining durchgeführt werden. Als Erweiterung des Data Mining dient Text-Mining der Extraktion von Informationen bzw. Wissen aus unstrukturierten Daten, wie Texten [Tan99]. Um dieses Wissen über mögliche dolose Handlungen aus Texten zu extrahieren, wird nachfolgender Text-Miningprozess nach Hippner und Rentzmann [HiRe06] vorgeschlagen (siehe Tabelle 1).

Prozessschritt	**Bezug zur Forensik im Rechnungswesen**
1. Aufgabendefinition	Erkennung von dolosen Handlungen in E-Mails
2. Dokumentselektion	E-Mails und Texte von Austauschlaufwerken mit Potenzial für Regelverstöße werden untersucht.
3. Dokumentaufbereitung	Aufbereitung durch Zerlegung der Texte in einzelne Wörter/Wortgruppen unter Berücksichtigung der Stammformreduktion (Stemming) und Löschung von Bindewörtern
4. (Text-) Mining Methoden	Filterung von betrugsrelevanten Wort- bzw. Wortgruppen
5. Interpretation / Evaluation	Interpretation ob ein Betrugsfall vorliegt
6. Anwendung	Ansprache der Beteiligten und Implementation von Gegenmaßnahmen

Tabelle 1: Text-Miningprozess [HiRe06]

Nach Abarbeitung der grundlegenden Prozessschritte kann im Schritt vier des Text-Mining-Prozesses nach Wörtern und Wortgruppen gesucht werden, welche Hinweise auf dolose Handlung darstellen. Dies können bspw. Wörter wie "stehlen", "verschieben", "hinterziehen", "Bestechung" sein. Diese Wörter und Wortgruppen sollten branchenspezifisch durch Analyse erkannter Betrugsfälle extrahiert werden.

Neben elektronischen Handelsbriefen (wie bspw. Lieferantenemails) können auch Lageberichte von Kunden und Lieferanten [HlKH11] anhand von spezifischen Wörtern wie z.B. Personalpronomina wie „wir“, „unserer Einschätzung nach“ etc. ausgewertet werden [HlKH11]. Dadurch können falsche Angaben von Lieferanten und Kunden oder Tochter- sowie verbundenen Unternehmen aufgedeckt werden. Zur Textanalyse werden häufig auch praktikable (Text-)Cluster zusammengestellt und gebildet [BeKu12].

Weiterhin sind bei der Analyse derartiger unstrukturierter Daten (wie E-Mails, Texte auf Firmenlaufwerken) gesetzliche Bestimmungen zu beachten [Schu10]. Je nach Anwendungsfall und vorhergehender Einverständniserklärung der Mitarbeiter ist eine derartige Analyse in der Bundesrepublik Deutschland weitestgehend möglich [Schu10].

Neben der Analyse unstrukturierter Daten können strukturierte Daten über verschiedene Methodiken analysiert werden, welche im nachfolgenden vorgestellt werden.

Analyse mittels Verteilungsfunktionen

Das Benford Gesetz [Benf38] [RaTh07] stellt eine Verteilungsfunktion des Auftretens von (führenden) Ziffern in der Natur dar. Danach ist das Vorkommen bestimmter Häufigkeiten von (führenden) Ziffern in der Natur nach folgender logarithmischer Verteilungsfunktion definiert:

$$p(Ziffer) = log_{10}(1 + \frac{1}{Ziffer})$$

Formel 1: Benford-Verteilung [Benf38]

Auf Basis der Untersuchung von Buchungstransaktionen bspw. im ERP-System können nun verdächtige Transaktionen in einem definierten Zeitraum ermittelt werden, wenn die führenden Ziffern der jeweiligen Buchungsbeträge weit von der Benford-Verteilung abweichen. Wenn beispielsweise oft Transaktionen unter der zustimmungspflichtigen Grenze von 1000 Euro (bspw. 970 Euro) durchgeführt werden, würde dies aufgrund der Abweichung zum normalen Verteilungsvorkommen (bei Ziffer 9: 4,57%) auffallen und könnte auf eine dolose Handlung hinweisen. Ein derartiges Beispiel ist in nachfolgender Grafik (Abbildung 4) dargestellt. Daten für eine solche Analyse können sowohl direkt aus dem IT-System extrahiert oder über die Verwendung von Datenaustauschformaten wie der XBRL [GeHe04] [DFOP09] etwa mit Daten von Tochterunternehmern gewonnen werden.

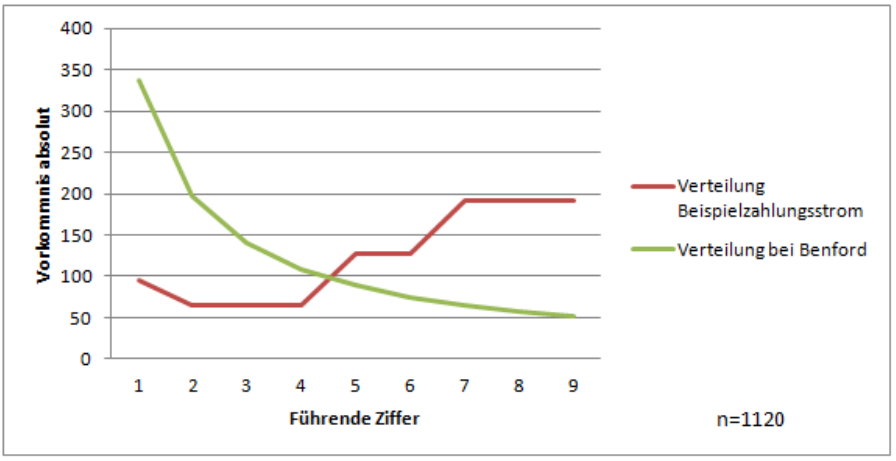


Abbildung 4: Beispielhafte Datenströme in Relation zur Benford-Verteilung

Zur Überprüfung der Passgenauigkeit der realen Zahlungsströme zu der Benford-Verteilung können Anpassungstests implementiert werden [Voge12]. Die Verarbeitung großer Mengen an Buchungstransaktionen aus verschiedenen IT-System und deren Überprüfung auf die Benford-Verteilung kann durch Big Data Technologien beschleunigt und somit die Betrugsbekämpfung verbessert werden.

Als ein weiteres Testmaß kann zur Überprüfung der Chi-Quadrat-Test für eine Abweichung der Vorkommastelle bei Zahlungen verwendet werden [BHRW07] [BeKu12].

Weitere Analyseverfahren

Weitere daten-getriebene Verfahren zur Mustererkennung bzw. Ermittlung von Schwellenwerten, ab wann eine dolose Handlung vorliegt, sind beispielsweise [KLSH04], [PLSG10]:

- Künstliche Neuronale Netze
- Expertensysteme
- Data-Mining
- Model-based Reasoning

Diese Verfahren können mit bestehenden Methoden kombiniert werden, um die Robustheit zu steigern. Als Ergebnis werden diese in Regelsets definiert um eine einheitliche Vorgehensweise zu ermöglichen [KLSH04], [PLSG10].

Die vorgestellten Verfahren können zum einen durch eine schnelle Verarbeitung größerer Datenvolumina in kürzerer Zeit (wie bspw. bei der Anwendung bei Benford) oder durch die bessere Analyse von unstrukturierten Daten (wie bspw. beim Text-Mining) durch Big

Data Technologien unterstützt werden. Somit können dolose Handlungen besser und schneller als bisher erkannt werden und ihnen präventiv entgegen gewirkt werden. Dieses Nutzenpotenzial (+: hohes Nutzenpotenzial; -: geringes Nutzenpotenzial) ist in nachfolgender Tabelle zusammengefasst.

	Big Data Unterstützung		
Analysemethode	Volumen	Geschwindigkeit	(Hohe) Strukturvarianz
Benford-Analyse	+	+	-
Chi-Quadrat-Test	+	+	-
Text-Mining	+	+	+
Künstliche Neuronale-Netze	+	+	+
Entscheidungsbaum	+	+	+

Tabelle 2: Big Data Unterstützung

Bei der Analyse unstrukturierter Daten wie bspw. E-Mails und Handelsbriefe können Verfahren des Text-Mining unter Berücksichtigung von In-Memory-Computing und neuen Verarbeitungsalgorithmen wie Hadoop zur Aufdeckung doloser Handlungen verwendet werden. Somit kann ein großes Volumina an Daten, unstrukturierter Form in Kürze verarbeitet werden, um betrügerische Handlungen nach gegebenen Muster aufzudecken. Weiterhin können beispielsweise strukturierte Daten (wie etwa Kontoauszahlungen) im ERP-System ebenfalls durch Verwendung von obigen Big Data Technologien auf dolose Handlungen hin schneller als bisher überprüft und präventiv entgegen gewirkt werden. Diese Anwendungsszenarien können gemäß der Kriterien und Methodiken in Tabelle 2 vielfältig ergänzt werden.

5 Fazit und Diskussion

Durch Anwendung von IT-basierten Analyseverfahren können dolose Handlungen in Unternehmungen besser aufgedeckt werden als bisher. Damit kann die IT-Governance besser Geschäftsziele und Compliance-Bestimmungen einhalten und (IT-)Risiken minimieren. Weiterhin ermöglicht die Anwendung von Big Data Technologien große Volumen an Daten in kürzerer Zeit kosteneffektiv zu analysieren und auch unstrukturierte Daten (wie bspw. E-Mails zu Kunden und Lieferanten) hinsichtlich Compliance-Verletzungen zu analysieren. Bei der Anwendung dieser Verfahren ist jedoch zu beachten, dass einerseits nicht alle Betrugsfälle fehlerfrei erkannt werden und andererseits alle erkannten Fälle direkt einer dolosen Handlung entsprechen. Zum einen können neuartige Muster auftauchen, die die Algorithmen bisher nicht erkennen. Zum anderen können diese Verfahren nie alle Transaktionen fehlerfrei klassifizieren. Daher kann es vorkommen, dass normale Transaktionen als Betrug gekennzeichnet werden. Entsprechend sollte je nach Gewichtung des potenziellen Betrugsfalls eine manuelle Überprüfung stattfinden. Weiterhin sollte jede Unternehmung vorab prüfen, ob eine automatisierte Überprüfung und die damit verbundene IT-Integration wirtschaftlich sind. So können Transaktionen

aus Produktionssystemen (wie bspw. Betriebsdatenerfassungssysteme) ggf. keinen hohen Beitrag zu Betrugserkennung von Transaktionen liefern, wie dies etwa Accounting-Systeme können. Weiterhin sind bei der Analyse von unstrukturierten Daten wie Texte aus E-Mails oder Firmenlaufwerken gesetzliche Bestimmungen zu beachten und gegebenenfalls vorab Einverständniserklärungen einzuholen. Zukünftig sollten die Verfahren in praktischen Szenarien implementiert und auf branchenabhängige Determinanten untersucht und angepasst werden. So haben beispielsweise Transaktionen in Finanzunternehmen andere Schwellenwerte zur Betrugserkennung als dies bspw. in mittelständischen Maschinenbauunternehmen der Anwendungsfall ist.

Literaturverzeichnis

[Abit97] Abiteboul, S.: Querying semi-structured data. Springer (1997).

[AgDE11] Agrawal, D. et al.: Big data and cloud computing: current state and future opportunities. Proceedings of the 14th International Conference on Extending Database Technology. pp. 530–533 ACM (2011).

[BBBK11] Bitzer, B. et al.: Tax Fraud & Forensic Accounting: Umgang mit Wirtschaftskriminalität. Springer DE (2011).

[BeKu12] Berwanger, J., Kullmann, S.: Interne Revision. Springer DE (2012).

[Benf38] Benford, F.: The law of anomalous numbers. Proc. Am. Philos. Soc. 551–572 (1938).

[BHRW07] Boenner, D.-K.A. et al.: „STAAN: Standard Audit Analysis "–Teil II. ZIR Z. Interne Revis. Erich Schmidt Verl. Berl. Ausgaben. 6, 2006, 03 (2007).

[Bitk12] BITKOM: Big Data im Praxiseinsatz – Szenarien, Beispiele, Effekte, http://www.bitkom.org/files/documents/Big_Data_BITKOM-Leitfaden_Sept.2012.pdf, (2012).

[BrHK11] Brynjolfsson, E. et al.: Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? (2011).

[BuCM13] Bughin, J. et al.: Ten IT-enabled business trends for the decade ahead | McKinsey & Company, http://www.mckinsey.com/insights/high_tech_telecoms_internet/ten_it-enabled_business_trends_for_the_decade_ahead.

[Bund08] Bundesrepublik Deutschland: Gesetz über das Aufspüren von Gewinnen aus schweren Straftaten. (2008).

[Bund65] Bundesrepublik Deutschland: Aktiengesetz. (1965).

[ChCS12] Chen, H. et al.: Business intelligence and analytics: From big data to big impact. MIS Q. 36, 4, 1165–1188 (2012).

[DaBB12] Davenport, T.H. et al.: How "Big Data" is Different. MIT Sloan Manag. Rev. 54, 1, 21–24 (2012).

[DaHM10] Davenport, T.H. et al.: Analytics at Work: Smarter Decisions. Better Results Harv. Bus. Sch. Press Boston MA. (2010).

[DFOP09] Debreceny, R. et al.: XBRL for interactive data. Springer (2009).

[Frei08] Freidank, C.-C.: Corporate Governance und interne Revision: Handbuch für die Neuausrichtung des Internal auditings. Erich Schmidt, Berlin (2008).

[GaRe11] Gantz, J., Reinsel, D.: Extracting value from chaos. IDC IView. 1–12 (2011).

[GeHe04] Gehra, B., Hess, T.: XBRL—Ein Kommunikationsstandard für das Financial Reporting. Control. Manag. 48, 6, 403–405 (2004).

[Hays04] Hays, C.L.: What Wal-Mart Knows About Customers' Habits, http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html, (2004).

[Herz06] Herzog, H.: Risk & Fraud Manage-ment als erweiterte Managementaufgabe. ZRFG–Risk Fraud Gov. 1, 2006, 11–15 (2006).

[HiRe06] Hippner, H., Rentzmann, R.: Text mining. Inform.-Spektrum. 29, 4, 287–290 (2006).

[HlKH11] Hlavica, C. et al.: Fraud: Aufdeckung und Prävention durch Forensic Accounting. Tax Fraud & Forensic Accounting. pp. 176–246 Springer (2011).

[Idw06] IDW, P.: 261: Feststellung und Beurteilung von Fehlerrisiken und Reaktionen des Abschlussprüfers auf die beurteilten Fehlerrisiken. Prüfungsstandard Inst. Wirtsch. Stand. 6, (2006).

[JoGo06] Johannsen, W., Goeken, M.: IT-Governance–neue Aufgaben des IT-Managements. HMD-Prax. Wirtsch. 43, 250 (2006).

[KeMU06] Kemper, H.-G. et al.: Business Intelligence: Grundlagen Und Praktische Anwendungen: Eine Einführung in Die IT-Basierte Managementunterstützung. Springer DE (2006).

[Kirm12] Kirmes, R.: Private IT-Forensik und private Ermittlungen, zwei Seiten einer Medaille? eine Analyse der Begriffe, Rollen und legalen Betätigungsfelder für private IT-Forensik, zugleich Grundlegung für ein Berufsrecht der privaten IT-Forensik. Eul (2012).

[KLSH04] Kou, Y. et al.: Survey of fraud detection techniques. Networking, sensing and control, 2004 IEEE international conference on. pp. 749–754 (2004).

[MCBB11] Manyika, J. et al.: Big data: The next frontier for innovation, competition, and productivity. McKinsey Glob. Inst. 1–137 (2011).

[MeZK03] Meyer, M. et al.: IT-Governance. Wirtschaftsinformatik. 45, 4, 445–448 (2003).

[MSKW13] Möhring, M. et al.: Big Data – neue Möglichkeiten im E-Commerce. Wirtsch. Manag. 2013, 2, 48–56 (2013).

[MSMK13] Möhring, M. et al.: Big Data – Implikationen für die Betriebswirtschaft. Wist - Wirtsch. Stud. 2013, 8, 456–458 (2013).

[Oliv09] Olivier, M.S.: On metadata context in Database Forensics. Digit. Investig. 5, 3–4, 115–123 (2009).

[PLSG10] Phua, C. et al.: A comprehensive survey of data mining-based fraud detection research. ArXiv Prepr. ArXiv10096119. (2010).

[Pric11] Price Waterhouse Coopers: Studie zur Wirtschaftskriminalität 2011: Kommissar Zufall deckt am meisten auf, http://www.pwc.de/de/risiko-management/studie-zur-wirtschaftskriminalitaet-2011-kommissar-zufall-deckt-am-meisten-auf.jhtml.

[Prüf02] Prüfungsstandard, I.D.W.: Abschlußprüfung bei Einsatz der Informationstechnologie (IDW PS 330). Wirtsch. 55, 21 (2002).

[RaTh07] Rafeld, H., Then Bergh, F.R.: Digitale Ziffernanalyse in deutschen Rechnungslegungsdaten. ZIR. 2, 26–33 (2007).

[Schu10] Schuster, F.P.: IT-gestützte interne Ermittlungen in Unternehmen – Strafbarkeitsrisiken nach den §§ 202a, 206 StGB. Z. Für Int. Strafdogmatik. (2010).

[SMMP14] Schmidt, R. et al.: Big Data as Strategic Enabler - Insights from Central European Enterprises. In: Abramowicz, W., Kokkinaki, A. (Hrsg.). 17th International Conference on Business Information Systems (BIS) (in press). Springer LNBIP. pp. 50–60 Springer Berlin Heidelberg, Larnaca, Cyprus (2014).

[SMZW14] Schmidt, R. et al.: Towards a Framework for Enterprise Architecture Analytics. EDOC Conference Workshop, 2014. EDOC'14. , Ulm (2014).

[Tan99] Tan, A.-H.: Text mining: The state of the art and the challenges. Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. pp. 65–70 (1999).

[Voge12] Vogel, J.: Die Güte von Anpassungstests auf Benfordverteilung Arbeitsbericht Nr. 2012-01, Juli 2012. Technische Universität Ilmenau Fakultät für Wirtschaftswissenschaften Institut für Wirtschaftsinformatik, Illmenau (2012).

Elektromobilität, Agenten und Smart Grids

(EASG)

Workshop „Elektromobilität, Agenten und Smart Grids“ (EASG)

Martin Matzner, Daniel Beverungen, WWU Münster, ERCIS

Hansjörg Fromm, Carola Stryja, Karlsruher Institut für Technologie

Jan-Marco Leimeister, Philipp Menschner, Universität Kassel, Universität St. Gallen

Helmut Krcmar, Michael Schermann, TU München

Sebastian Lehnhoff, OFFIS – Institut für Informatik sowie Sprecher der GI-FG Energieinformationssysteme (EINS)

Paul J. Kühn, Universität Stuttgart

Ralf Lehnert, TU Dresden

Hermann de Meer, Universität Passau

Kurt Rothermel, Universität Stuttgart

Michael Sonnenschein, Universität Oldenburg

Die Beiträge dieser beiden Workshops und thematisch verwandter Einreichungen der MATES’14 wurden im Rahmen eines gemeinsamen Thementags „Elektromobilität, Agenten und Smart Grids“ auf der 44. Jahrestagung der Gesellschaft für Informatik präsentiert.

IT-gestützte Dienstleistungen für die Elektromobilität

Unter dem Schlagwort Elektromobilität werden technologische und organisatorische Innovationen zusammengefasst, die einen Beitrag für nachhaltige und klimaangepasste Verkehrsinfrastrukturen leisten. Die Bundesregierung verfolgt das Ziel, Deutschland als Leitanbieter und Leitmarkt für die Elektromobilität zu positionieren. So sollen bis 2020 eine Million Elektrofahrzeuge in Deutschland betrieben werden. Voraussetzungen zur Erreichung dieses Ziel sind einerseits die Entwicklung innovativer Fahrzeug-, Energiespeicher- und Infrastrukturkonzepte. Andererseits entstehen im Rahmen der Elektromobilität auch neue Möglichkeiten zur Entwicklung IT-gestützter Dienstleistungskonzepte, mit deren Hilfe diese physischen Komponenten zu bedarfsgerechten Mobilitätslösungen erweitert werden. Beispielhaft sind Dienstleistungen für alternative Ladeinfrastrukturen, integrierte multi-modale Verkehrskonzepte und End-of-Life-Lösungen für Traktionsbatterien aus Elektrofahrzeugen zu nennen. Im Rahmen des Workshops werden innovative IT-

gestützte Dienstleistungen für die Elektromobilität sowie Methoden zur Entwicklung dieser Dienstleistungen diskutiert.

Smart Grids

Die Energiewende mit ihren Zielen für die Umstellung des Energiesystems auf nachhaltige Energieträger bis zum Jahr 2050 führt zu einschneidenden Veränderungen im bisherigen elektrischen Energieversorgungssystem bzw. zu einem Paradigmenwechsel bei Planung und Betrieb der Netze. Wenn bisher die Erzeugung im Wesentlichen am Bedarf ausgerichtet war (Lastfolgebetrieb), so wird man in Zukunft vermehrt versuchen, den Bedarf an das volatile Dargebot der erneuerbaren Energien anzupassen. Da die hierfür zu flexibilisierenden Verbrauchsanlagen – z.B. elektrische Fahrzeuge, Wärmepumpen, Mikro-BHKWs – zu einem signifikanten Anteil in den Verteilnetzen angeschlossen sind, ist mit maßgeblichen Investitionen in Primär- sowie Sekundärtechnik und intelligente Leitsysteme in diesem Bereich zu rechnen. Es wird allgemein davon ausgegangen, dass durch den Ausbau historisch gewachsener Netzinfrastruktur in ein IKT-geführtes intelligentes Netz, das die Flexibilität von Lasten und Erzeugern ausnutzt, ein signifikanter volkswirtschaftlicher Nutzen generiert werden kann. Bei diesem Transformationsprozess spielen jedoch nicht nur technologische Aspekte, sondern auch die Einbindung der Konsumenten, die oftmals gleichzeitig Produzenten sind (sog. Prosumer), eine wichtige Rolle. Daher liegt das Augenmerk nicht nur auf der Entwicklung, Bewertung und Anwendung neuer Technologien, sondern auch auf der Wechselwirkung dieser mit den Nutzern dieser Komponenten. Relevante Konzepte und Komponenten in der Energieversorgung erfordern auch eine Berücksichtigung und Anpassung der Regulierung, um das Potenzial des Smart Grid vollständig zu heben. Die Herausforderungen sind dabei nicht auf die Domäne Strom begrenzt, vielmehr steigt der Bedarf, die Herausforderungen von Strom-, Gas- und Wärmeversorgung sowie Mobilität integriert anzugehen.

Die Energieinformatik steht bei beiden Themen nicht nur in der Verantwortung, die „Systemintelligenz“ zu liefern – Algorithmen zur adaptiven Steuerung und kontinuierlichen dynamischen Optimierung des komplexen und sehr umfangreichen (europäischen) Energieversorgungssystems – sondern stellt auch die Methoden bereit, „Gesamtsystemkompetenz“ zu schaffen und zu orchestrieren. Denn Informatik als Strukturwissenschaft erlaubt Komplexitätsbeherrschung durch Dekomposition und Abstraktion, Identifikation von und Fokussierung auf verallgemeinerbare Prinzipien, Suchen von Entkopplungspunkten für effektive Governance, Vermeidung von Bottlenecks etc.

Ein Total Value-of-Ownership Modell zur Bewertung elektrifizierter Fahrzeugflotten

Thomas Görzen, Christian Meier, Dennis Kundisch
Lehrstuhl für Wirtschaftsinformatik, Information Management & E-Finance
Universität Paderborn
Warburger Str. 100
33098 Paderborn
Thomas.Goerzen@wiwi.uni-paderborn.de
Christian.Meier@wiwi.uni-paderborn.de
Dennis.Kundisch@wiwi.uni-paderborn.de

Abstract: Elektromobilität nimmt als zukünftige Form der Mobilität eine zentrale Rolle ein. Dabei existieren allerdings einige Hürden die eine Verbreitung dieser Mobilitätsform gegenüberstehen. Gewerbliche Fahrzeugflotten bieten in diesem Zusammenhang die Möglichkeit eine zunehmende Verbreitung zu unterstützen, da ihre Rahmenbedingungen teilweise schon jetzt den Einsatz von Elektrofahrzeugen begünstigen. Um eine ökonomische Bewertung des Elektrifizierungspotenzials zu ermöglichen, wird auf Basis bereits vorhandener ökonomischer Bewertungsmodelle aus der Literatur ein integriertes Modell zur umfassenden Berücksichtigung von Kosten- und Nutzenparametern resultierend aus dem Einsatz von Elektrofahrzeugen in gewerblichen Flotten entwickelt. Weiterführend stellt dieses Modell die Basis für die Entwicklung eines webbasierten Entscheidungsunterstützungssystems dar, welches Betreiber gewerblicher Flotten bei der Bewertung des Elektrifizierungspotenzials unterstützen soll.

1 Motivation

„Die Zukunft gehört dem, der als erster die Kraft der Sonne in den Tank packt, mit Wasserstoff überholt oder CO2*-frei vorankommt.“*

(Horst Köhler)

Elektromobilität als neue Form der Fortbewegung ist bereits im Zentrum der gesellschaftlichen Diskussion über Alternativen zur zukünftigen Mobilität angekommen. Die Gründe für die Suche nach Alternativen zu konventionell betriebenen Verbrennerfahrzeugen liegen vor allem in den stetig steigenden Preisen für fossile Kraftstoffe einhergehend mit einem absehbaren Ende deren Vorrats. „Die Elektrifizierung der Antriebe ist eine ganz wesentliche Stellschraube für eine zukunftsfähige Mobilität. Sie bietet die Chance, die Abhängigkeit vom Öl zu reduzieren, die Emissionen zu minimieren und die Fahrzeuge besser in ein multimodales Verkehrssystem zu integrieren.“ [Ne09, S.6]. Neben elektrisch unterstützten Zweirädern

befinden sich mittlerweile verschiedene batteriebetriebene Elektrofahrzeuge (BEV) auf dem Markt, die bereits in ersten Serien produziert werden. BEV stellen im Gegensatz zu Hybridfahrzeugen eine Form der Mobilität dar, die für den Antrieb gänzlich auf den Einsatz von fossilen Kraftstoffen verzichtet und somit eine langfristige Lösung darstellt.[1] Hinzu kommt vor allem auch in Deutschland ein erhöhtes Bewusstsein des weltweiten klimatischen Wandels, hervorgerufen durch den weltweit steigenden Ausstoß von Kohlenstoffdioxid und Feinstaub [DOK12], [Un98]. Laut Ergebnissen der Studie „Trend beim Autokauf 2013“ schreiben 75% der Befragten dem Elektroauto am ehesten die Fähigkeit zu, in den nächsten zehn Jahren am meisten für den Umweltschutz zu leisten. Der flächendeckenden Verbreitung der BEV stehen jedoch einige zu überwindende Hürden gegenüber. So antworteten 85% der Teilnehmer der Studie, dass ihr nächstes Fahrzeug voraussichtlich wieder mit einem Diesel- oder Ottomotor angetrieben werden wird [A13]. Lediglich 1% der Befragten zog ein elektrisch betriebenes Fahrzeug für den nächsten Kauf in Erwägung. Als Gründe dafür sind neben einer bisher nicht ausreichend vorhandenen Ladeinfrastruktur für Elektrofahrzeuge häufig die im Vergleich zu konventionell betriebenen Fahrzeugen mangelnde Reichweite sowie der deutliche höhere Anschaffungspreis zu identifizieren [EY11], [Te12].

In Anbetracht die ser Schwierigkeiten stellen gewerblich betriebene Flotten in diesem Zusammenhang kurz- und mittelfristig die *„vorrangig relevante* Nutzergrup*pe dar“* [NPE10, S.33f.]. Gewerbliche Flotten haben einen Anteil von über 13 % am gesamten mobilisierten Individualverkehr in Deutschland [MiD08] und machen ca. 30% des Neuwagenmarktes aus [Pl13]. Durch Fahrprofile mit einem hohen Anteil an planbaren, mit den Reichweiten aktueller BEV zu bewältigenden Routen können diese Hürden potenziell verringert werden [Pl13]. Jedoch bestehen aus Sicht gewerblicher Flottenbetreiber bei einer Einführung von BEV in den Betriebsalltag immer noch ökonomische Unsicherheiten. Um diese Unsicherheiten zu reduzieren und das Potenzial elektrisch betriebener Fahrzeuge im gewerblichen Flotteneinsatz ökonomisch bewerten zu können, werden umfassende ökonomische Modelle zur Erfassung der verschiedenen Kostenkategorien benötigt. Zu diesem Zweck wurden bereits zahlreiche Total-Cost-of Ownership (TCO), Life-Cycle-Costing (LCC) sowie Total Value-of-Ownership (TVO) Modelle in der Literatur vorgeschlagen [DRD05], [El95], [FP02], [HW04]. Die vorgeschlagenen Modelle sind in der Lage, über die reinen Anschaffungskosten hinaus eine Vielzahl von Kostenkategorien über den gesamten Lebenszyklus (meist eines einzelnen Fahrzeugs) abzubilden. Jedoch reicht die reine kostenseitige Betrachtung zur umfassenden ökonomischen Bewertung nicht aus, da vor allem die Vorteile die potenziell aus der Nutzung der neuen Technologien (z.B. Netzunterstützungsleistungen) den momentan noch substanziell höheren Anschaffungskosten im Vergleich zu konventionell betriebenen Fahrzeugen noch nicht umfassend gegenübergestellt werden können. Zwar werden diese Nutzenpotenziale in der Literatur bereits isoliert betrachtet, allerdings ist den Autoren aktuell kein umfassendes ökonomisches Bewertungsmodell über alle Phasen des Lebenszyklus bekannt, welches umfänglich sowohl Kosten als auch aus dem Einsatz der neuen Technologien resultierende Nutzenpotenziale in gewerblichen Fahrzeugflotten berücksichtigen kann. Der vorliegende Beitrag verfolgt

[1] Aus diesem Grund werden im Rahmen dieser Arbeit rein elektrisch betriebene Fahrzeuge fokussiert, das resultierende Bewertungsmodell jedoch ist durch seine modulare Struktur ebenfalls zur Bewertung von Hybridfahrzeugen geeignet.

damit das Ziel, diese Forschungslücke zu schließen, indem er die folgende Forschungsfrage adressiert: Wie sollte ein integriertes Modell zur ökonomischen Bewertung von elektrifizierten gewerblichen Fahrzeugflotten unter umfassendem Einbezug relevanter Kosten- und Nutzenkategorien gestaltet werden?

Im Rahmen eines gestaltungsorientierten Forschungsparadigmas [He04] wird zur Beantwortung dieser Frage ein Artefakt in Form eines Modells entwickelt. Dieses Modell wird zukünftig als Basis für die Entwicklung eines webbasierten Entscheidungsunterstützungssystems für Flottenbetreiber dienen und im Rahmen des öffentlich geförderten interdisziplinären Forschungsprojektes Pfleg!E-mobil[2], gemeinsam mit einem Flottenbetreiber parametrisiert, instanziiert und evaluiert.

2 Anforderungen an ein integriertes Kosten-Nutzenmodell der Elektromobilität und Stand der Forschung

Die Kostenbetrachtung von Elektrofahrzeugen stellt einen zentralen Aspekt bei der Beurteilung der Eignung von Elektromobilität für den Einsatz in gewerblichen Fahrzeugflotten dar. Im Gegensatz zu privaten Nutzern, die bei der Kaufentscheidung eines Fahrzeugs auch Wert auf Aspekte wie Marke, Design oder spezifische Fahrleistungen legen, spielen bei gewerblichen Haltern insbesondere der Anschaffungspreis und darüber hinaus die Gesamtkosten des Fahrzeugs eine zentrale Rolle bei der Kaufentscheidung [HA11]. Gerade im Bereich des Anschaffungspreises liegen elektrisch betriebene Fahrzeuge aktuell noch deutlich über konventionell betriebenen Fahrzeugen. Für einen BEV-Kleinwagen beispielsweise, können die Anschaffungskosten bis zu 40% über denen eines vergleichbaren konventionell betriebenen Dieselfahrzeugs liegen [Pl13]. Im Gegensatz dazu sind die Betriebskosten bei BEV potentiell geringer. Aus diesem Grund ist eine Gesamtkostenbetrachtung für eine fundierte Bewertung notwendig. Die Betrachtung der Gesamtkosten von Gütern stellt allgemein keinen neuen Ansatz dar. Bereits 1928 wurde die Bedeutung der Kostenbetrachtung über den eigentlichen Kaufpreis hinaus hervorgehoben [Ha28]. Im Laufe der Zeit wurden verschiedene Ansätze der Gesamtkostenbetrachtung, z.B. in Form von Total-Cost-of-Ownership Modellen entwickelt, welche seither auch ein stetiger Diskussionsgegenstand der Wirtschaftsinformatik sind [Ge09a].

In der Vergangenheit wurde eine Vielzahl verschiedener allgemeiner und kontextspezifischer Bewertungsmodelle in der Literatur vorgeschlagen ([DRD05], [El93], [FP02], [WH00], [NPE11], [Bi11]). Die vorliegenden Ansätze unterscheiden sich in ihrem grundsätzlichen Aufbau und in der Bestimmung der relevanten Kostentreiber sowie deren Kategorisierung teils deutlich, so dass diese verschiedenen Modelle zu differierenden Ergebnissen führen können [Kr07]. Um eine Basis für eine fundierte und vergleichbare ökonomische Bewertung elektrifizierter Fahrzeugflotten treffen zu können, ist es daher notwendig, die bestehenden Modelle auf die Anforderungen hin zu

[2] Im Rahmen der wissenschaftlichen Begleitforschung im Projekt Pfleg!E-mobil wird sowohl eine technische, soziologische als auch ökonomische Untersuchung durchgeführt. Das Projekt wird im rahmen des Ziel2.NRW Programms durch das Land Nordrhein-Westfalen sowie durch die Europäische Union gefördert.

überprüfen, die aus den besonderen Charakteristika elektrifizierter Fahrzeuge im Flotteneinsatz entstehen. Im Folgenden werden diese Anforderungen aus der einschlägigen Literatur sowie auf Basis von Interviews mit den Projektpartnern und Betreibern gewerblicher Fahrzeugflotten[3] argumentativ-deduktiv abgeleitet. Anschließend werden bestehende Modelle aus der Literatur anhand dieser Anforderungen bewertet.

2.1 Ableitung der Anforderungen an ein integriertes Bewertungsmodell der Elektromobilität in gewerblichen Flotten

Durch die Nutzung von (Elektro-)Fahrzeugen entstehen unterschiedliche Arten von Kosten. Neben den einmalig anfallenden Anschaffungskosten (im Falle eines Kaufs) sind typischerweise Steuern und Versicherungsgebühren zu entrichten. Diese Kosten sind üblicherweise nutzungsunabhängig, das heißt sie variieren nicht mit der Intensität des Gebrauchs. Dem gegenüber stehen zum Beispiel Betriebs- und Wartungskosten, welche stark von der Intensität der Nutzung abhängig sind. Da sich BEV innerhalb dieser Kategorien teils stark von konventionell betriebenen Fahrzeugen unterscheiden [Pl13], ist es nicht zuletzt aus Transparenzgründen notwendig, dass ein Bewertungsmodell in der Lage ist, zwischen diesen fixen und variablen Kosten zu unterscheiden.

Anforderung A1: Ein Bewertungsmodell der Elektromobilität in gewerblichen Flotten muss eine Differenzierung in fixe und variable Kosten erlauben.

Bereits vor der Kaufentscheidung für einen bestimmten Fahrzeugtypen fallen Kosten (z.B. für die Selektion eines geeigneten Fahrzeugs) an. Weiterhin fallen nach der Nutzungszeit ggf. Entsorgungskosten an. Da diese Kosten aufgrund des unterschiedlichen Angebots und den unterschiedlichen Maßen an Vertrautheit mit den jeweiligen Technologien bei BEV und konventionell betriebenen Fahrzeugen potenziell variieren können, sollte ein umfassendes Bewertungsmodell in der Lage sein diese Kosten berücksichtigen zu können. Daraus leitet sich die Anforderung ab, dass ein umfassendes Bewertungsmodell alle Phasen, von der Beschaffung bis zur Entsorgung, abbilden kann. Da ein Bewertungsmodell für die Nutzung im Kontext von Fahrzeugflotten erstellt werden soll, werden dabei an dieser Stelle explizit Kosten für die Produktion (im Verständnis von LCC Modellen) ausgeklammert[4].

Anforderung A2: Ein Bewertungsmodell der Elektromobilität in gewerblichen Flotten muss in der Lage sein, alle Phasen des Nutzungszyklus abzubilden.

Bei der Anschaffung von konventionell betriebenen Fahrzeugen betrifft die Anschaffungsentscheidung meist nur das Fahrzeug selbst, da Ladeinfrastruktur typischerweise in Form einer öffentlichen Tankstelleninfrastruktur verfügbar ist. Anders

[3] Es wurden im Rahmen des öffentlich geförderten Projektes drei semistrukturierte Leitfadeninterviews mit dem Geschäftsführer eines ambulanten Pflegedienstes, einem Fuhrparkmanager eines weiteren ambulanten Pflegedienstes, sowie einem Fuhrparkmanager eines regionalen Energieversorgers geführt.

[4] Wir nehmen vereinfachend an, dass derartige Produktions- und Entwicklungskosten aus Sicht des in dieser Untersuchung stehenden Flottenbetreibers bereits im Anschaffungspreis Berücksichtigung finden.

verhält es sich aktuell für BEV. Hier müssen die Ladeinfrastruktur sowie eventuelle Abrechnungssysteme ebenfalls bei der Bewertung Berücksichtigung finden. Aus diesem Grund reicht eine singuläre Betrachtung des reinen Fahrzeugs nicht aus.

Anforderung A3: Ein Bewertungsmodell der Elektromobilität in gewerblichen Flotten muss in der Lage sein, alle Infrastrukturkomponente, die für den Betrieb einer Fahrzeugflotte notwendig sind, abbilden zu können.

Fahrzeuge dienen in gewerblichen Flotten dem Zweck, eine Produkt- oder Dienstleistungserbringung zu ermöglichen (z.B. Lieferdienstfahrten). Fallen ein oder mehrere Fahrzeuge (oder Teile der benötigten Infrastruktur) aus, so entstehen neben den direkten Kosten für Instandsetzung oder Ersatzbeschaffung auch indirekte Kosten durch den Ausfall des Fahrzeugs. Nach dem TCO-Modell der Gartner-Group werden die indirekten Kosten als effizienzhemmende Vorgänge im Rahmen der Nutzung des jeweiligen Produkts verstanden [WH00]. Beispielsweise kann aus dem Ausfall eines benötigten Fahrzeugs bei einem innerstädtischen Pflegedienst eine nicht erbrachte Dienstleistung und somit eine Schmälerung der Erlöse (vllt. sogar verbunden mit Strafzahlungen) resultieren.

Anforderung A4: Ein Bewertungsmodell der Elektromobilität in gewerblichen Flotten muss sowohl direkte als auch indirekte[5] Kosten berücksichtigen können.

Neben der Ermöglichung der Dienstleistungserbringung in gewerblichen Flotten können Fahrzeuge auch sekundären Zwecken dienen. So ist es z.B. vorstellbar, dass das Fahrzeug als Werbeträger sogar Einnahmen generieren kann. Im Falle von BEV kommt zudem die Möglichkeit der Bereitstellung von Systemdienstleistungen für Energieversorgungsunternehmen in Betracht (Netzunterstützungsleistungen)[QZB09]. Die Bereitstellung einer Flotte von mobilen Energiespeichern kann unter Umständen den ökonomischen Effekt des höheren Anschaffungspreis im Vergleich zu konventionellen Fahrzeugen abschwächen. Diese und mögliche weitere Potenziale zur Wertgenerierung sollten in eine Bewertung ebenfalls mit einbezogen werden, um ein umfassendes Bild der ökonomischen Potenziale darstellen zu können.

Anforderung A5: Ein Bewertungsmodell der Elektromobilität in gewerblichen Flotten muss die Möglichkeit bieten, sowohl umfassende Kosten- als auch Erlöskategorien zu berücksichtigen.

2.2 Vergleich der Anforderungen mit bestehenden Modellen

Um einen Vergleich verschiedener Ansätze aus der Literatur anzustellen, wurde in kumulativer Forschungstradition eine Bewertung 20 verschiedener Modelle von [Ge09b] herangezogen. Aufbauend auf dem umfassenden Vergleich von branchespezifischen und branchenunabhängigen TCO und LCC Modellen wurden für den weiteren Verlauf dieser Arbeit die geeignetsten Modelle aus [Ge09b] ausgewählt und im Rahmen einer Literaturrecherche um branchenspezifische Modelle speziell für die Elektromobilität

[5] Im Sinne von indirekten Kosten als effizienzhemmende Vorgänge

ergänzt. Dabei wurden diejenigen Modelle aus [Ge09b] von der weiteren Untersuchung ausgeschlossen, die aufgrund ihres Branchenfokus nicht hinreichend generalisierbar sind. Von den verbleibenden, allgemein formulierten Modellen in der Literatur wurden für die weitere Betrachtung im Rahmen dieser Arbeit die Modelle von [DRD05] und [El93] ausgewählt, da sie in Bezug auf die Vollständigkeit sowie die weiteren von [Ge09b] angelegten Kriterien den höchsten Zielerreichungsgrad aufweisen. Zudem wurden die branchenspezifischen Modelle von [FP02] und [WH00] in die weitere Betrachtung mit aufgenommen, da sie eine umfangreiche Aufschlüsselung empirisch erhobener Kostenkategorien aufweisen [FP02] bzw. trotz Branchenfokus den höchsten Zielerreichungsgrad [WH00] der Kriterien von [Ge09b] erreicht haben und hinreichend generalisierbar sind. Diese vier Modelle wurden um im Rahmen einer Literaturrecherche ermittelte spezifische Ansätze zur Bewertung von Elektrofahrzeugen erweitert. Es werden dabei auf Basis einer Vorab-Begutachtung durch die Autoren ebenfalls die beiden umfangreichsten und vielversprechendsten Modelle ([NPE11], [Bi11]) in dieser Arbeit berücksichtigt. Tabelle 1 stellt die Erfüllungsgrade der definierten Anforderungen tabellarisch dar. Die Bewertung der jeweiligen Ansätze im Hinblick auf die Erfüllungsgrade wurde durch die Autoren des Beitrags vorgenommen. Dabei bewerteten die Autoren die aus der Literatur extrahierten Modelle jeweils unabhängig voneinander anhand der identifizierten Anforderungen und teilten sie bezüglich der Erfüllungsgrade dieser ein. Anschließend wurden unterschiedliche Einschätzungen diskutiert und die Ergebnisse zusammengefasst.

Tabelle 1: Erfüllungsgrade der definierten Anforderungen

	Anf. 1		Anf. 2	Anf. 3		Anf. 4	Anf. 5
	Fix	Variabel	Verschiedene Nutzungsphasen	Direkt	Indirekt	Gesamte Infrastruktur	Erlöse
[DRD05]	X	o	X	o	o		
[El93]	X	o	o	o	o		o
[FP02]	X	o	o	o	o	X	o
[WH00]	X	o		X	X	o	
[NPE11]	o	o					o
[Bi11]	X	X	o			o	X

Legende: X: Explizit im Modell vorgesehen
o : Nicht explizit im Modell vorgesehen, aber grundsätzlich im Modell umsetzbar

Die Analyse der Ansätze im Hinblick auf die definierten Anforderungen zeigt, dass die vorhandenen Modelle in Summe bereits viele der Anforderungen explizit oder implizit erfüllen können. Die verschiedenen Modelle legen dabei unterschiedliche Schwerpunkte und variieren im Hinblick auf die Granularität der definierten Kostenkategorien und konkreten Kostentreiber teils stark. Am ehesten ist das Modell von [Fe02] in der Lage, die definierten Anforderungen umfänglich zu erfüllen. Jedoch finden die Anforderung nach Berücksichtigung variabler Kostenarten, verschiedener Nutzungsphasen im Lebenszyklus, direkte und indirekte Kosten sowie die Berücksichtigung von Erlösen nur implizit Einzug im Modell [FP02]. Das bedeutet, dass zum Beispiel Opportunitätskosten (als eine Ausprägung indirekter Kosten) im Model berücksichtigt werden. Jedoch wird keine explizite Unterscheidung zwischen direkten und indirekten Kosten vorgenommen.

Eine strukturierte, explizite Darstellung der verschiedenen Anforderungen ist aus unserer Sicht jedoch nicht nur hilfreich, sondern auch notwendig, um den Nutzer eines solchen Bewertungsmodells bestmöglich bei der Identifikation und der Bewertung der konkreten Kosten in den einzelnen Kostenarten zu unterstützen. Im Folgenden wird daher auf Basis der Stärken der bestehenden Modelle ein integriertes Modell zur ökonomischen Bewertung für den Einsatz von Elektrofahrzeugen in gewerblichen Flotten entwickelt. Das Modell übernimmt Komponenten und Konzepte der bereits vorhandenen Modelle und überführt diese in ein Gesamtmodell mit dem Anspruch, alle definierten Anforderungen explizit um umfassend adressieren zu können.

3. Entwicklung eines integrierten TVO-Modells

Verschiedene Nutzungsphasen und Betrachtung der gesamten der Infrastruktur

Die grundlegende Struktur des Modells orientiert sich an dem Modell nach [DRD05]. Im Unterschied zu den rein kostenorientierten Modellen des LCC und TCO umfasst das integrierte Modell ebenso mögliche Erlöskategorien die bei der Elektrifizierung einer Fahrzeugflotte auftreten können. Für das Modell wird eine Matrix der einzelnen Kostenkategorien erstellt, die in den Spalten die verschiedenen Nutzungsphasen der Betrachtungsobjekte umfasst (vgl. Abbildung 1). Die Phasen in den Spalten umfassen die Beschaffung, den Empfang, den Besitz, die Nutzung und die Entsorgung der Betrachtungsgegenstände. Somit kann die Anforderung 2 erfüllt werden, alle relevanten Nutzungsphasen der Betrachtungsobjekte zu erfassen. In den Zeilen werden die fünf relevanten Betrachtungsobjekte aufgeführt, die notwendig sind, um eine Elektromobile Infrastruktur zu betreiben. Diese stellen neben dem Fahrzeug auch die Energie (Energieversorgung, Tarif etc.), die Ladeinfrastruktur, die Batterie und die für die Abrechnung der Energie benötigte Infrastruktur (z.B. handybasiertes Ladesystem, RFID Lösung etc.) dar.

Fixe und variable Kosten

Die einzelnen Kosten- und Erlöskategorien des Modells werden, Anf. 1 folgend, in fixe und variable Kategorien unterteilt (vgl. [BK11]). Dabei wird neben der Unterscheidung nach fixen (Anschaffungskosten pro Fahrzeug, Ladeinfrastrukturkomponente etc.) und variablen Kosten (Betriebskosten für Fahrzeug, nutzungsabhängige Wartungskosten, etc.) noch eine weitere Kategorie, die der flottenfixen Kosten, eingeführt. Fixe Kosten oder Erlöse fallen pro Betrachtungsgegenstand in feststellbarer, fixer Höhe an. Variable Kosten und Erlöse sind abhängig von der Nutzungsintensität des bzw. der Betrachtungsobjekte und variieren mit deren Nutzung (z.B. Treibstoff- oder Energiekosten). Im Unterschied zu privaten Nutzern, die einzelne Betrachtungsobjekte wie z.B. die Ladeinfrastruktur nur für ein Fahrzeug nutzen, können Flottenbetreiber bestimmte Betrachtungsobjekte für mehrere Fahrzeuge nutzen wodurch diese Kategorie (weitestgehend) unabhängig von der Nutzungsintensität ist. Daher ist die Differenzierung dieser Kategorie im Modell sinnvoll um Flottenbetreibern bei der ökonomischen Betrachtung diese Kosten nachvollziehbar aufzeigen zu können.

Indirekte und direkte Kosten

Da gerade bei der Betrachtung einer gewerblichen Flotte auch Kosten relevant sind, die durch ungeplante Ausfälle der Fahrzeuge und der Infrastrukturkomponenten entstehen können, wurde das Modell in Anlehnung an das bestehende TCO-Modell der Gartner Group [WH00] um direkte und indirekte Kosten erweitert (vgl. Anf. 4). Dabei werden unter direkten Kosten solche verstanden, die zur Bereitstellung der Leistung direkt den einzelnen Betrachtungsobjekten zugeordnet werden können [WH00]. „Die direkten Kosten können mit einem relativ hohen Genauigkeitsgrad bestimmt werden und größtenteils auch durch Belege oder Rechnungen in ihrer Höhe schriftlich erfasst *werden"* [WH00, S.11]. Beispiele für direkte Kosten sind Anschaffungskosten oder Leasinggebühren für einzelne Betrachtungsobjekte. Indirekte Kosten können als Wertverzehr definiert werden, *„der aus effizienzhemmenden Vorgängen im Rahmen der Nutzung einer bestimmten Technologie entstehen"* [WH00, S.11]. Diese indirekten Kosten umfassen typischerweise Opportunitätskosten durch Produktivitätseinbußen und können dem Ausgangsmodell entsprechend in zwei Teile gegliedert werden. Zum einen sind dies Kosten, die durch Zeiten entstehen, in denen die einzelnen Betrachtungsobjekte, wie z.B. die Elektrofahrzeuge oder die Ladeinfrastruktur, nicht genutzt werden können. Dazu zählen vor allem Ausfall- oder Wartungszeiten aber auch Verzögerungszeiten bei der Bereitstellung der Betrachtungsobjekte. Der andere Teilbereich der indirekten Kosten umfasst Kosten, die durch nicht produktiv verbrachte Arbeitszeit der Nutzer verursacht werden. Dazu gehören Zeiten, die für das Erlernen des Umgangs mit den Betrachtungsobjekten, z.B. den Fahrzeugen oder der Ladesäule, benötigt werden. In diesen Zeiten, die für das Lernen am Arbeitsplatz, Schulungen oder Peer-to-Peer Support (z.B. fachfremde Hilfestellung von Mitnutzern) aufgewendet werden, können die Mitarbeiter ihrer eigentlichen Tätigkeit nicht nachgehen wodurch (Opportunitäts-)Kosten entstehen.

Integration potenzieller Erlösfaktoren

Um die Anforderung nach einer Betrachtung der möglichen Erlösfaktoren zu erfüllen, und um die geäußerte Kritik an Modellen mit reinem Kostenfokus [HW04] aufzugreifen, wird das Modell um potenziell relevante Erlösfaktoren erweitert. Im Kontext der Elektromobilität stellen diese „revenue-enhancing-effects" [HW04], also jene Effekte, die zu möglichen Erlösen durch die Nutzung eines bestimmten Gutes führen können, vor allem Netzunterstützungsleistungen der Elektrofahrzeuge (Vehicle-to-Grid) dar [Bi11]. Insbesondere bei gewerblichen Fahrzeugflotten besitzen diese Erlösfaktoren aufgrund der Anzahl an Fahrzeugen sowie den planbaren Fahr-und Standzeiten ein relevantes Potenzial [QZB09]. Somit wird das rein kostenorientierte TCO-Modell in ein kosten- und nutzenorientiertes Total Value-of-Ownership Modell überführt.

Parametrisierung der Kategorien

Um eine umfangreiche und komplette Parametrisierung des erstellten Modells zu gewährleisten wurde eine Dokumentation der einzelnen Kosten-und Erlöskategorien erstellt. Dabei wurde für jede Kategorie ein Steckbrief erstellt, der die wichtigsten Informationen jeder einzelnen Kategorie enthält. Zunächst wird der einzelne Kosten-

oder Erlöstreiber sowohl der Phase als auch dem Betrachtungsobjekt zugeordnet. Beispielhaft ist in Tabelle 2 die Kostenkategorie „Anschaffungskosten“ des Betrachtungsobjekts „Fahrzeug“ dargestellt, welche innerhalb des TVO-Modells der Phase „Beschaffung“ zuzuordnen ist. Die Definition der Kostenkategorie stellt sicher, dass mögliche Unklarheiten, die anhand einer reinen Betrachtung des Steckbrieftitels entstehen könnten, ausgeräumt werden. Die „Art“ gibt an, ob es sich bei der betrachteten Kosten- bzw. Erlöskategorie um einen variablen, fixen oder flottenfixen Betrag handelt. Die „Berechnungsvorschrift“ beschreibt die Grundlage der monetären Quantifizierung der Kosten- oder Erlöse. Dabei wird definiert, wie die Höhe dieser festzulegen bzw. zu ermitteln ist. Häufig stellt der Wert die Anschaffungskosten des Betrachtungsobjekts dar, in Einzelfällen (z.B. für die Ermittlung der der Energie- oder Wartungskosten pro Fahrzeug) wird jedoch eine spezifische Berechnungsvorschrift benötigt, um eine möglichst genaue Ermittlung der Kosten oder Erlöse zu bestimmen.

Die Zeile Parametrisierung enthält eine Angabe zur monetären Höhe des Betrachtungsobjekts, welche mittels der Berechnungsvorschrift bestimmt werden kann. Die materialistische Grenze beschreibt eine bestimmte (und im Projektverlauf noch empirisch zu bestimmende) monetäre Größe die ein Betrachtungsgegenstand überschreiten muss, um als relevanter Faktor für die Wirtschaftlichkeitsbetrachtung angesehen zu werden. Wird diese Grenze nicht überschritten, so fließt diese Kosten-bzw. Erlöskategorie nicht in die ökonomische Betrachtung ein, wird aber zur möglichst vollständigen Betrachtung aufgeführt. Als Beispiel können die Zahlungskosten (bspw. für eine Überweisung anfallende Bankgebühren) der einzelnen Betrachtungsobjekte, die in der Phase „Beschaffung“ anfallen, angeführt werden. Im letzten Teil des Steckbriefs wird erläutert ob die Kosten- oder Erlöskategorie elektrofahrzeug-spezifisch ist, d.h. nur im Kontext der Elektromobilität auftritt oder auch für vergleichbare, konventionell angetriebene Fahrzeuge entsteht. Dieser Aspekt ist von Relevanz, da dadurch ein Vergleich zwischen den Kosten- und Erlösfaktoren von Elektro- und konventionell angetriebenen Fahrzeugen vereinfacht wird bzw. Unterschiede verdeutlicht werden können.

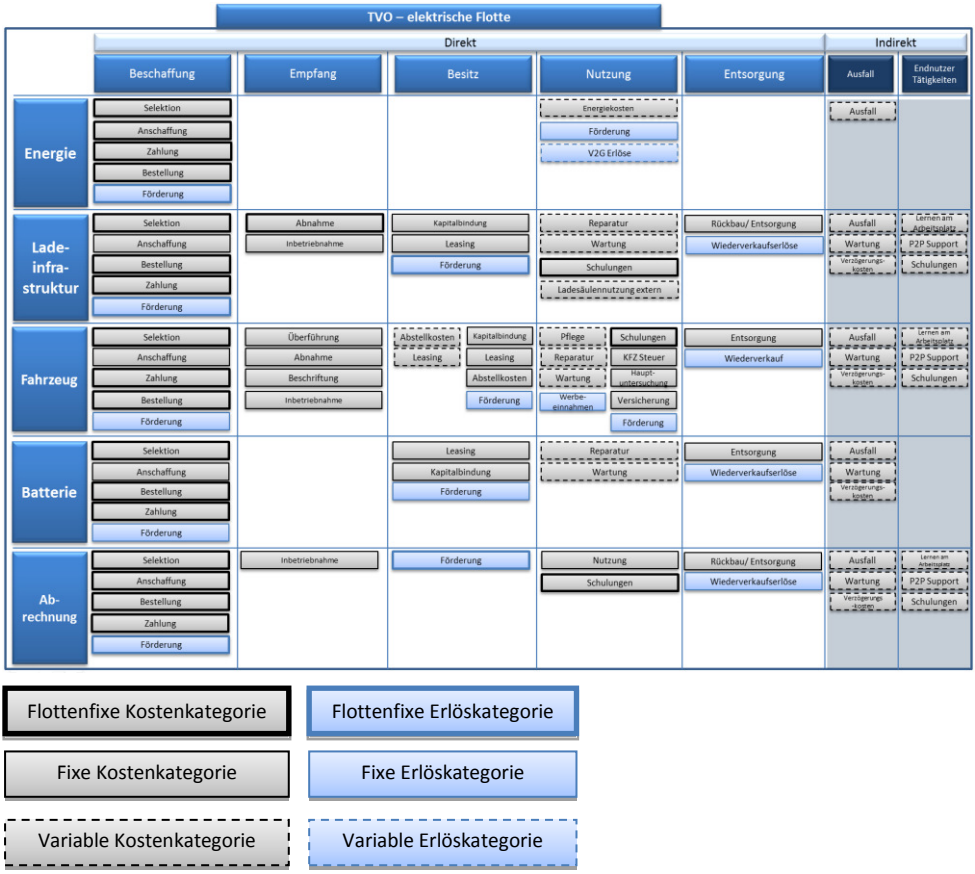
Tabelle 2: Steckbrief der Kosten-und Erlöskategorien

	Fahrzeug/Beschaffung/Anschaffungskosten
Definition	Die Anschaffungskosten für ein Fahrzeug bestehen aus dem Kaufpreis. Die Kosten werden hier ohne Kauf der Batterie betrachtet.
Art	Fix
Berechnungsvorschrift	Anschaffungskosten = Kaufpreis
Parametrisierung	Durchschnittlicher Nettolistenpreis DIESEL Kleinwagen: 12.592 € Durchschnittlicher Anschaffungspreis BEV Kleinwagen: 20.020€ Quelle: [Pl13]
Materialistische Grenze überschritten	Ja
Elektroauto-spezifische Kostenkategorie	Nein

Das konzeptionelle Gesamtmodell ist in Abbildung 1 dargestellt. Eine vollständige und umfangreiche Dokumentation der konkreten Kostentreiber innerhalb der dargestellten Kostenkategorien (in Form der Darstellung von Tabelle 2) wird im Rahmen der

Zwischen- und Abschlussberichte des öffentlich geförderten Forschungsprojektes, im Rahmen dessen dieses Modell erarbeitet wurde, veröffentlicht und ist somit nicht Teil des vorliegenden Beitrags. Die Spalten der Abbildung stellen die verschiedenen Nutzungsphasen dar, während die Zeilen das jeweilige Betrachtungsobjekt referenziert.

Abbildung 1: Integriertes TVO-Modell



4. Zusammenfassung und Ausblick

Elektromobilität ist eines der dominierenden Themen bei der Betrachtung zukünftiger Mobilitätskonzepte und bereits im Zentrum der gesellschaftlichen Diskussion angekommen. Dabei konnte diese Form der Mobilität allerdings bisher noch nicht die (auch durch die Bundesregierung im [Ne09] formulierte) angestrebte Verbreitung erreichen. Gewerbliche Fahrzeugflotten besitzen das Potenzial, eine wichtige Rolle bei der Verbreitung dieser Form der Mobilität einzunehmen, da sie aufgrund ihrer Eigenschaften (wie z.B. planbaren Routen mit Streckenlängen, die auch mit bestehenden Batterien bereits bewältig werden können) und ihrer hohen Marktrelevanz als vielversprechende Erstnutzer in Frage kommen. Um eine differenzierte und vollständige

ökonomische Bewertung der Elektrifizierung einer gewerblichen Fahrzeugflotte zu ermöglichen, wurde im Rahmen der vorliegenden Arbeit ein integriertes Kosten-Erlös-Modell für diesen Kontext entwickelt und konzeptionell dargestellt. Zu diesem Zweck wurden im ersten Schritt spezielle Anforderungen zur Elektrifizierung gewerblicher Fahrzeugflotten definiert. Im Folgenden wurden bestehende Bewertungsansätze aus der Literatur identifiziert und deren Eignung für den Einsatz zur Bewertung des Elektrifizierungspotenzials evaluiert. Auf Basis der wertvollen Ansätze der Literatur wurde ein integriertes Kosten-Nutzen-Modell entwickelt, welches in der Lage ist die definierten Anforderungen umfänglich zu erfüllen. Das resultierende Modell der Elektromobilität zielt darauf ab, Betreiber gewerblicher Flotten strukturiert bei der Identifikation und Bewertung der Elektrifizierungspotenziale zu unterstützen. Das Modell umfasst fünf Nutzungsphasen und ist in der Lage neben der Bewertung des Fahrzeugs ebenso die benötigten Infrastrukturkomponenten (Ladeinfrastruktur, Abrechnungsinfrastruktur etc.) zu berücksichtigen. Des Weiteren werden neben direkten Kosten auch indirekte Kosten resultierend aus effizienzhemmenden Vorgängen (Fahrzeugausfall etc.) einbezogen. Zudem können aus den spezifischen Charakteristika aus der Elektromobilität resultierende Erlöspotenziale (Netzunterstützungsleistung etc.) mit in die Bewertung einbezogen werden. Das in dieser Arbeit vorgestellte konzeptionelle Modell wird im Rahmen eines öffentlich geförderten Forschungsprojektes gemeinsam mit einem gewerblichen Flottenbetreiber parametrisiert und im Zuge dessen zur Bewertung des Elektrifizierungspotenzials einer konkreten gewerblichen Flotte eingesetzt. Auf Basis des umfangreichen Modells wird weiterhin zukünftig ein webbasiertes Entscheidungsunterstützungssystem entwickelt, welches Flottenbetreiber in Form eines „Quick-Checks“ sowie eines vollständigen Systems bei der Bestimmung des Elektrifizierungspotenzials ihrer konkreten Flotte unterstützen soll.

Literaturverzeichnis

[BK11] Brickert, S.; Kuckshinrichs, W: Electromobility as a Technical Concept in an Ecological Mobility Sector? An Analysis of Costs. In Proceedings of the 9th International Conference of the European Society for Ecological Economics (ESEE 2011), Istanbul, Turkey, 2011; Verfügbar online: http://www.esee2011.org/registration/fullpapers/esee2011_54ff9e_2_1307627504_8295_2080.pdf

[DRD05] Degreave, Z.; Roodhooft, F.; van Doveren, B.: The use of total cost of ownership for strategic procurement: a company-wide management information system. In Journal of Operational Research Society 56, 2005, S. 51-59.

[DOK12] Dijk, M.; Orsato, R.; Kemp, R.: The emergence of an electric mobility trajectory. In Energy Policy 52, 2012, S. 135-145.

[EY11] Ernst Young: European Automotive Survey, 2011, S. 16.

[El95] Ellram, L.M.: Total cost of ownership: An analysis approach for purchasing. In International Journal of Physical Distribution & Logistics Management 25 (8), 1995, S. 4-23.

[FP02] Ferrin, B.G.; Plank, R.E.: Total Cost of Ownership Models: An Exploratory Study. In Journal of Supply Chain Management 38 (2), 2002, S.18-29.

[Ge09a] Geissdörfer, K.: Total Cost of Ownership (TCO) und Life Cycle Costing (LCC). Einsatz und Modelle: Ein Vergleich zwischen Deutschland und USA. 1. Aufl. Münster, Westf: LIT.

[Ge09b] Geissdörfer, K.; Gleich, R.; Wald, A: Standardisierungspotentiale lebenszyklusbasierter Modelle des strategischen Kostenmanagements. In Zeitschrift für Betriebswirtschaft 79, 2009, S. 693–715.

[Ha28] Harriman, N.F.: Principles of Scientific Purchasing. New York: McGraw-Hill, 1928.

[He04] Hevner, A.R.; March, S.T.; Park, J.; Ram, S.: Design science in information systems research. MIS quarterly 28 (1), 2004, S. 75-105.

[HW04] Hurkens, W.; Wynstra, F.: The Concept 'Total Value of Ownership': A Case Study Approach, Erasmus Research Institute of Management, Rotterdam, 2004: S. 51- 62.

[Kr07] Krämer, S.: Total Cost of Ownership. Konzept, Anwendung und Bedeutung im Beschaffungsmanagement deutscher Industrieunternehmen. Saarbrücken: VDM Verl. Dr. Müller, 2007.

[MiD08] Mobilität in Deutschland 2008, Tabellenband, S. 113 ff. http://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CDMQFjAA&url=http%3A%2F%2Fwww.mobilitaet-in-deutschland.de%2Fpdf%2FMiD2008_Tabellenband.pdf&ei=Y6pnU4-0GIWiO5e1gagB&usg=AFQjCNEq3VYlo_eJrhEgX1vJsHLfzaVQ-A&sig2=dSGJwcWpaXJk0iYQ6trHPg&bvm=bv.65788261,d.ZWU. Letzter Abruf: 05.05.14.

[NPE10] Zwischenbericht der nationalen Plattform Elektromobilität, 30. November 2010. Letzter http://www.bmub.bund.de/service/publikationen/downloads/details/artikel/zwischenberichte-der-nationalen-plattform-elektromobilitaet/?tx_ttnews[backPid]=708 Letzter Abruf: 05.05.14.

[NPE11] Nationale Plattform Elektromobilität: Zweiter Bericht der Nationalen Plattform Elektromobilität. 2011 Anhang. Hg. v. Gemeinsame Geschäftsstelle Elektromobilität der Bundesregierung. Berlin. Online verfügbar unter http://www.bmub.bund.de/fileadmin/bmu-import/files/pdfs/allgemein/application/pdf/bericht_emob_2.pdf, Letzter Abruf 08.05.14

[Ne09] Nationaler Entwicklungsplan der Elektromobilität der Bundesregierung, 2009. Online verfügbar unter: www.bmbf.de/pubRD/nationaler_entwicklungsplan_elektromobilitaet.pdf, Letzter Abruf 08.05.14

[Pl13] Plötz, P.; Gnann, T.; Kühn, A.; Wietschel, M: Markthochlaufszenarien für Elektrofahrzeuge. Fraunhofer ISI, 2013.

[Un98] United Nations: Kyoto Protocol to the United Nations Framework Convention on Climate Change, 1998; Online verfügbar unter: http://unfccc.int/resource/docs/convkp/kpeng.pdf. Letzter Abruf 06.04.14

[Ha11] Hacker, F.: Öko-Institut e.V.: Betrachtung der Umweltentlastungspotenziale durch den verstärkten Einsatz von kleinen, batterieelektrischen Fahrzeugen im Rahmen des Projekts „E-Mobility Berlin", Status-Seminar Elektromobilität Berlin-Brandenburg, 2011, Öko-Institut: Berlin

[QZB09] Quinn, C.; Zimmerle, D.; Bradley, T.-H.: The effect of communication architecture on the availability, relaibility, and economics of plug-in hybrid electric vehicle-to-grid ancillary services. In Journal of Power Sources 195 (5), 2009, S.1500-1509.

[A13] Aral Studie: Trend beim Autokauf 2013. www.aral.de/content/dam/aral/pdf/Brosch%C3%BCren/aral_studie_trends_beim_autokauf_2013.pdf, Letzter Abruf: 05.05.14.

[Te12] Teichmann, GA; Trützschler, J; Hahn, C; Schäfer, PK; Hermann, A; Höhne, K: Elektromobilität Normen bringen die Zukunft in Fahrt. 2012; http://www.pwc.de/de_DE/de/offentliche-unternehmen/assets/studie_normung.pdf. Letzter Abruf 15.04.14

[WH00] Wild, M.; Herges, S.: Total Cost of Ownership (TCO) - Ein Überblick. Mainz: Arbeitspapiere WI, Universität Mainz, 2000.

Unterstützung von KMU bei der Erbringung komplexer Mobilitätsservices in Koopkurrenznetzwerken

Dr. Andreas Helferich

highQ Computerlösungen GmbH
Schwimmbadstraße 26, 79100 Freiburg
a.helferich@highQ.de

Norman Pelzl

Universität Stuttgart, Betriebswirtschaftliches Institut
Lehrstuhl für ABWL und Wirtschaftsinformatik II
Keplerstr. 17, 70174 Stuttgart
pelzl@wius.bwi.uni-stuttgart.de

Sixten Schockert

Universität Stuttgart, Betriebswirtschaftliches Institut
Lehrstuhl für ABWL und Wirtschaftsinformatik II
Keplerstr. 17, 70174 Stuttgart
und
Fachhochschule für Ökonomie und Management (FOM)
Rotebühlstraße 121, 70178 Stuttgart
schockert@wius.bwi.uni-stuttgart.de

Abstract: Kein einzelnes Mobilitätskonzept kann den Mobilitätsbedarf der Kunden alleine abdecken. Dies gilt gleichermaßen für den klassischen Öffentlichen Personennahverkehr wie auch für alternative Mobilitätskonzepte zur Elektromobilität oder das Carsharing. Daher profitieren Anbieter in diesen Bereichen davon, sich zu sog. Koopkurrenznetzwerken zusammenzuschließen und ihre Dienstleistungen im Verbund anzubieten. Die Teilnahme an einem solchen Koopkurrenznetzwerk stellt sowohl hohe Anforderungen an die Hintergrundsysteme als auch an die Front-Office-Systeme zur Anbindung der Endnutzermedien wie Smartphones oder Smartcards. Große Mobilitätsdienstleister sind dabei in der Lage, eigene organisatorische Einheiten und Rechenzentren zu nutzen. Kleine und mittelständische Mobilitätsdienstleister hingegen weisen i. d. R. keine leistungsfähige IT-Infrastruktur auf und sind nur eingeschränkt dazu fähig und bereit, entsprechendes Know-how aufzubauen. Vor diesem Hintergrund ist das Ziel des Forschungsprojekts Aprikose kleinen und mittleren Unternehmen im Bereich von Mobilitäts- sowie komplementärer Dienstleistungen eine Möglichkeit zu geben, sich einfach, kostengünstig und sicher als Anbieter an einem Koopkurrenznetzwerk zu beteiligen. Dieser Beitrag berichtet über erste Ergebnisse des vom BmBF geförderten Projekts.

1 Kontext und Motivation

Aufgrund der hohen Anschaffungskosten und einiger wahrgenommener Nachteile wie z. B. die geringe Reichweite der Fahrzeuge oder die geringe Dichte an Ladesäulen/Lademöglichkeiten ist die Wahrscheinlichkeit, dass sich Käufer für ein Elektrofahrzeug entscheiden, aktuell gering. Zudem zeigen verschiedene Studien der letzten Jahre zumindest für Großstädte einen Einstellungswandel im Bereich der Mobilität auf: nachdem lange Jahre die Anzahl der PKWs pro 1000 Einwohner stieg, stagniert dieser heute eher, während insbesondere junge Erwachsene seltener einen PKW besitzen [FS13]. Um somit viele potenzielle Nutzer mit Elektromobilität in Verbindung zu bringen, ist es notwendig, die Einstiegshürde möglichst niedrig zu halten. Insbesondere scheint die Kombination mit Angeboten der kollaborativen Mobilität wie (freefloating oder stationsgebundenes) Carsharing oder Miet-/Leihsystemen ratsam, bei denen die Bürger einfach und risikoarm ein Elektrofahrzeug testen können [BB13]. Keines dieser Mobilitätskonzepte kann allerdings alleine alle Mobilitätsbedürfnisse der potentiellen Kunden decken. So sind Leihfahrräder und –pedelecs für Langstrecken nicht geeignet, stationsgebundenes Carsharing für spontane Fahrten oftmals zu unflexibel und freefloating Carsharing für längere Reisen bislang preislich nicht konkurrenzfähig. Um Mobilität „von Tür zu Tür“ zu erreichen, muss der Kunde daher verschiedene Angebote nahtlos und flexibel zu einer integrierten Reisekette aus verschiedenen Verkehrsmitteln kombinieren können.

Dazu ist allerdings eine Kooperation mehrerer Anbieter notwendig, die in vielen Fällen gleichzeitig in Konkurrenz zueinander stehen. Dieser Zustand wird als Koopkurrenz (engl. Coopetition) bezeichnet und beschreibt die zunehmend vorkommende Marktsituation, dass Unternehmen einerseits Wettbewerber, andererseits Partner sind [BN11]. Besonders einfach und damit attraktiv für den Kunden ist die Nutzung derartiger kombinierter Angebote, wenn sowohl Fahrplanauskunft („Wie komme ich am besten von A nach B?“), als auch Preisauskunft („Was kostet mich das?“) als auch Zugang („Wie öffne ich das Fahrzeug? Habe ich die richtige Fahrberechtigung?“) und Abrechnung („Wie und wo muss ich bezahlen?“) integriert erfolgen. Um dies zu ermöglichen, müssen die beteiligten Anbieter Daten miteinander austauschen. Da sie im Wettbewerbsverhältnis bleiben und der Datenschutz gewahrt bleiben muss, soll ein Datenaustausch aber nur die jeweils notwendigen Daten umfassen. Diese Daten sicher, zuverlässig und jederzeit aktuell bereitzustellen, stellt die Anbieter IT-seitig vor große Herausforderungen – insbesondere kleinere Anbieter wie örtliche Busunternehmer oder ein als e.V. organisierter Carsharing-Anbieter.

An dieser Stelle setzt das Forschungsprojekt „Aprikose“ an.[1] Ziel von Aprikose ist es, kleinen und mittleren Unternehmen im Bereich von Mobilitäts- sowie komplementärer Dienstleistungen eine Möglichkeit zu geben, sich einfach, kostengünstig und sicher als Anbieter an einem Koopkurrenznetzwerk zu beteiligen. Dieser Beitrag berichtet über erste Ergebnisse des vom BmBF geförderten Projekts. Kapitel 2 führt dazu aus theoretischer und praktischer Sicht in die zentrale Rahmenbedingung des Koopkurrenznetwerks ein.

[1] Das Projekt „Aprikose“ wird im Rahmen des Programms „kmu innovativ“ vom BMBF unter dem Förderkennzeichen 01IS12015 gefördert. Weitere Informationen unter www.aprikose.wi.uni-stuttgart.de.

Kapitel 3 beschreibt das Vorgehen zur Ermittlung der Anforderungen an die IT-Infrastruktur in einem Koopkurrenznetzwerk. Kapitel 4 erläutert darauf aufbauend die Eigenschaften und Fähigkeiten der realisierten IT-Lösung für KMU zur Teilnahme an einem solchen Koopkurrenznetzwerk. Nachdem in Kapitel 5 erste Erfahrungen aus dem operativen Betrieb dargelegt werden schließt Kapitel 6 mit einem kurzen Fazit und Ausblick auf den weiteren Forschungsbedarf sowie zukünftige Einsatzzwecke.

2 Koopkurrenznetzwerke

2.1 Definition

Der Begriff der Koopkurrenz ist die deutsche Übersetzung des englischen Coopetition, eines Kunstworts, bestehend aus Cooperation und Competition. Zurückgeführt wird der Begriff auf Raymond Noorda, Gründer des Netzwerksoftwareherstellers Novell [BN11]. Wie der Name schon impliziert, bezeichnet Koopkurrenz den Zustand, dass zwei oder mehrere Unternehmen ein konkurrierendes und kooperierendes Wettbewerbsverhalten zeigen. Nach Beck lassen sich in Unternehmensnetzwerken drei Spielarten der Koopkurrenz unterscheiden [Be98]:

- Kombination kooperativer und wettbewerblicher Prinzipien in einer geschäftsbezogenen Netzbeziehung z. B. kooperieren Fluglinien untereinander, um Kunden ein breiteres Sortiment an Flugzielen zu bieten, konkurrieren aber auch miteinander auf manchen Flugrouten (Intra-Netzwettbewerb als „Koopkurrenz i. e. S.“).
- Schnelles Abwechseln von Kooperation und Konkurrenz auf Basis präkompetitiver Zusammenarbeit in ein und demselben Geschäftsfeld z. B. gemeinsame Forschungs- und Entwicklungsarbeiten oder Einkaufsaktivitäten (oft überlappend).
- Unternehmensnetzwerk als Ausdruck einer kollektiven Wettbewerbsstrategie, mit der Mitglieder individuelle Wettbewerbsvorteile gegenüber Nicht-Netzmitgliedern anstreben („cooperate to compete“); hiermit ist oft ein Wettbewerb zwischen Netzwerkmitgliedern verbunden („Koopkurrenz i. w. S.“), so z. B. innerhalb der Betriebssystem-Plattformen von mobilen Endgeräten.

Dabei können Koopkurrenz i. e. S. und i. w. S. durchaus gleichzeitig existieren: zwei an anderer Stelle konkurrierende Partner kooperieren innerhalb eines Unternehmensnetzwerkes miteinander und versuchen, sich damit Vorteile gegenüber weiteren Wettbewerbern, die nicht an diesem Unternehmensnetzwerk teilnehmen, zu verschaffen [RB00].

Das Koopkurrenzverhältnis innerhalb des Unternehmensnetzwerks stellt neue, enorme Herausforderungen an die IT-Infrastrukturen der beteiligten Unternehmungen. Einerseits müssen alle Anbieter über ein geeignetes IT-Netzwerk verbunden werden. Andererseits muss aufgrund des weiterhin bestehenden Konkurrenzverhältnisses sowie zum Zwecke des Schutzes von kritischen Unternehmensdaten sowie der Privatsphäre der Kunden sichergestellt werden, dass jeder Partner (ggf. anonymisiert) lediglich Zugriff auf diejenigen Daten erhält, die er zur Erfüllung seiner Aufgaben zwingend benötigt.

2.2 Beispiel für ein Koopkurrenznetzwerk: das (((eTicket Deutschland

Um die Konkurrenzfähigkeit und Attraktivität des ÖPNV zu steigern, betreiben der Verband deutscher Verkehrsunternehmen (VDV), das Bundesministerium für Verkehr, Bau und Stadtentwicklung (BMVBS) und eine ganze Reihe von Verkehrsunternehmen und -verbünden den Aufbau eines bundesweit einsetzbaren elektronischen Tickets. Dieses sog. (((eTicket Deutschland (www.eticket-deutschland.de) soll die Dienstleistungsqualität und Kundenfreundlichkeit des ÖPNV Länder- und Tarifgrenzen-übergreifend erhöhen. So können durch das (((eTicket den Verkehrsteilnehmern u. a. verbesserte Informationen und neuartige Dienstleistungen angeboten und darüber hinaus die Nutzung auch fremder Verkehrssysteme erleichtert werden. Der in vom BMVBS geförderten Projekten entwickelte Standard für das (((eTicket, die sog. VDV-Kernapplikation (kurz: VDV-KA) [VDV10], unterstützt sowohl Smartcards als auch entsprechend ausgerüstete Handys als Nutzermedien, auf denen elektronische Fahrberechtigungen aufgebracht werden können.

Aufbauend auf der VDV-KA können dem Verkehrsteilnehmer künftig komplexere (Mobilitäts-)Dienstleistungen angeboten werden, die sich aus mehreren, von verschiedenen Anbietern erbrachten, Einzeldienstleistungen zusammensetzen [Mo14]. Neben dieser Kooperation zur gemeinsamen Erbringung neuartiger Dienstleistungen verbleiben die Anbieter aber gleichzeitig Konkurrenten um die Fahrgäste (Koopkurrenz i. e. S.). Damit der Verkehrsteilnehmer diese Dienstleistungen einfach und komfortabel nutzen kann, müssen nicht nur entsprechende Nutzermedien für Zugang, Identifikation, Zahlung sowie Berechtigung zur Verfügung stehen. Auch die Hintergrundsysteme (HGS) der unterschiedlichen Dienstleister müssen dies unterstützen. Um die Vernetzung zu ermöglichen, stellt die VDV-Kernapplikations GmbH & Co. KG als zentrale Instanz verschiedene zentrale Services im sog. Interoperabilitäts-Netzwerk (ION) zur Verfügung. Um dem Datenschutz (und auch der Datensparsamkeit) sowie den Geschäftsinteressen der teilnehmenden Verkehrsverbünde und –unternehmen gerecht zu werden, wurde ein Rollenmodell entwickelt, auf dessen Basis jedes Unternehmen nur diejenigen Daten zu einem Kunden und seinen Berechtigungen („Fahrscheine“) und Fahrten erhält, die es zur Erfüllung seiner Aufgaben benötigt [VDV10]. Zu den Rollen gehören z. B. die des Kundenvertragspartners, der die Abrechnung und Kommunikation mit den Fahrgästen verantwortet oder die des Produktverantwortlichen, der die Tarifprodukte definiert und abrechnet.

Große Mobilitätsdienstleister haben die Möglichkeit, entsprechendes Know-how, eigene organisatorische Einheiten und eigene Rechenzentren zu nutzen, um sich an diese zentralen Services anzuschließen. Doch besonders die mehreren tausend kleinen und mittelständischen Mobilitätsdienstleister stehen durch das Koopkurrenzverhältnis vor Anforderungen an ihre IT-Infrastrukturen, welchen sie organisatorisch wie finanziell nicht gewachsen sind. Sie können in der Regel keine leistungsfähige IT-Infrastruktur aufweisen und sind nur eingeschränkt fähig, entsprechendes Know-how aufzubauen. Da von den ca. 6.000 Unternehmen im ÖPNV in Deutschland ca. 5.500 den KMU zuzuordnen sind [St10] und kleinere Anbieter mit neuartigen Mobilitätskonzepten zu erwarten sind, wird ein flächendeckendes Koopkurrenznetzwerk ohne ihre Einbindung nicht möglich sein. Ähnliche, aber regional begrenzte und stärker auf die intermodale Mobilität ausgerichtete Koopkurrenznetzwerke entstehen aktuell in Deutschland in verschiedenen Regionen, z. B. in Stuttgart und Hamburg (vgl. das Projekt „Stuttgart Services“ [Hi13a] und [SS13]).

3 Ermittlung von Anforderungen an die IT-Infrastruktur in einem Koopkurrenznetzwerk

Aufgrund des hohen Innovationsgrads des Vorhabens wurde ein dreistufiges Vorgehen im Sinne einer Produktlinienentwicklung gewählt. Stufe 1 hat die Entwicklung einer auf den ÖPNV beschränkten prototypischen Lösung zum Ziel. Der Vorteil ist hier, dass mit der VDV-KA [VDV10] ein branchenweiter auf der ISO 24014-1 basierenden Standard existiert, der die Interoperabilität innerhalb des ÖPNV ermöglicht, gleichzeitig aber eine Offenheit für die Erweiterung im Bereich der sog. Multiapplikation vorsieht. In Stufe 2 kommen neue Mobilitätskonzepte hinzu, um die Einsatzdomäne auf intermodale Wegeketten auszudehnen und Koopkurrenznetzwerke zwischen ÖPNV und Anbietern neuer Mobilitätskonzepte zu unterstützen. In Stufe 3 wird der Anwendungsbereich um mobilitätsnahe, komplementäre Dienstleistungen (insb. Tourismusdienstleistungen) ausgeweitet.

Zur Anforderungsermittlung kam die Produktplanungsmethode Quality Function Deployment (QFD) zum Einsatz (siehe z. B. [FC09] oder [He00]), da sie die Kundenanforderungen in den Mittelpunkt aller Bemühungen stellt: ein Produkt soll ausschließlich die vom Kunden gewünschten und nicht alle technisch möglichen Merkmalen aufweisen („fitness for use“). In vielen Unternehmen, aber auch in der Innovationsmanagement-Literatur wird QFD als prototypische Methode gesehen, um eine hohe Kundenorientierung als übergeordneten Erfolgsfaktor im Innovationsmanagement zu verwirklichen [RP09]. Im Projekt wird das sog. Continuous QFD (CQFD) eingesetzt, das speziell auf die inkrementelle Entwicklung in neuen, dynamischen und unsicheren Anwendungsdomänen ausgerichtet ist. [He00, Sc13]. Abbildung 1 gibt einen Überblick über die Schritte des Continuous QFD.

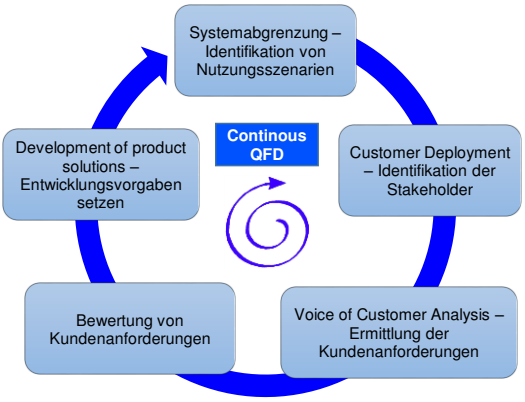


Abbildung 1: Inkrementelle Entwicklung mit Continuous QFD

Auf der Basis moderierter Workshops mit Vertretern potenzieller Partnerunternehmen wurden die Anforderungen erhoben, kategorisiert, priorisiert und mit Zielwerten versehen. Hierzu wurden für die erste Stufe auf Basis des VDV-KA-Rollenmodells Vertreter dreier Verkehrsverbünde sowie einer Landesnahverkehrsgesellschaft eingeladen (insgesamt 13 Kundenvertreter) und für die Stufen zwei und drei zusätzlich ein Carsharingunternehmen, ein Parkraumbetreiber sowie ein Vertreter der mit dem örtlichen Nahverkehr kooperieren-

den Stadtwerke, die ebenfalls Parkhäuser sowie Ladesäulen für Elektrofahrzeuge betreiben. In den Workshops ergab sich vor allem die soziale Herausforderung, die Beteiligten zur „wirklichen“ Kooperation zu ermuntern statt weiter ihrem bisherigen Konkurrenzverhalten zu folgen [Sc13]. Die Kunden priorisierten die Anforderungen, denn nur mit dieser Information – gepaart mit den Ergebnissen einer technischen Risikoklärung – war es möglich, die Funktionalitäten der ersten Prototypen festzulegen. Die Priorisierung gestaltete sich in diesem innovativen Umfeld schwierig, da das zukünftige Produkt für die Kunden neu und damit nur schwer vorstellbar ist. Um zu validen Bewertungen zu kommen, wurde neben der „normalen“ Wichtigkeitsabfrage auch die potentielle Wirkung einer unzureichenden Berücksichtigung einer Kundenanforderung erhoben (in Anlehnung an das KANO-Modell zur Klassifikation von Anforderungen [Ka84]). Zudem erfolgte sowohl eine Bewertung der Einzelanforderungen als auch eine auf Gruppenebene [Sc13]. Beide Maßnahmen dienten dazu, Inkonsistenzen in der Bewertung, potentielle Unsicherheiten und Missverständnisse möglichst frühzeitig unter allen Beteiligten auszuräumen.

Die 71 Anforderungen konnten auf diese Weise in fünf Ranggruppen geordnet werden. Z. B. bilden aus Kundensicht die Erfüllung der Nachvollziehbarkeit des Informationsflusses im Netzwerk, eine hohe Verfügbarkeit in der Kommunikation, die Interoperabilität mit existierenden Drittsystemen sowie die kassentechnische Sicherheit der Abrechnungsdaten unabdingbare Bestandteile der Ausbaustufe 1. Die Ergebnisse aus Kundensicht wurden dann der Perspektive der Entwickler gegenübergestellt. Es wurden für die insgesamt 18 Anforderungen der beiden höchsten Ranggruppe insgesamt 59 mögliche Funktionen bzw. Services und Leistungs- bzw. Qualitätsmerkmale identifiziert. Die möglichen Lösungen mit dem höchsten Beitrag zur Erfüllung der wichtigsten Anforderungen bildeten dann die Vorgabe für die Entwicklung. Die Analyse des Beitrages, den diese möglichen IT-Lösungsmerkmale zur Erfüllung der Kundenanforderungen haben, erfolgte mittels CQFD und ist in aggregierter Form in folgender Matrixdarstellung (einem sog. House of Quality entsprechend, [FC09] bzw. [He00]) ersichtlich.

Legende: Die Erfüllung des Kriteriums [oben] trägt 9 : zwangsläufig und sehr stark 3 : merkbar aber evtl. eingeschränkt 1 : möglicherweise und nur bedingt 0 : nur mittelbar oder ggf. gar nicht zur Erfüllung der Anforderung [links] bei.		Lösungsgruppen									
Kundenanforderungen	**Wichtigkeit**	**Anzeige**	**Ausfallsicherung**	**Datenhaltung**	**Datensicherheit und Datenschutz**	**Installation und Konfiguration**	**Kontrolle und Prüfung**	**Modularität**	**Schnittstellen und Kommunikation**	**Überwachung**	**Updates**
Administration	**13,60%**	9	3	3	1	9	3	3		9	9
Event Handling	**12,90%**			1			9		9	9	
Fachliche Betriebsführung	**12,40%**	9	3		3	1	3		3	3	1
Flexibilität	**11,10%**			3		3		9	9		3
Sicherheit	**9,30%**	1	3	3	9	1	1			3	3
Compliance	**8,00%**		3	3	9		1			1	
Verfügbarkeit	**7,60%**		9				3	1		9	3
Standards verwalten	**6,50%**					3		1	9		9
Individualisierung	**5,70%**	3		3		3		3			
Usability	**5,40%**	9				9	3			3	3
Abrechnungsunterstützung	**4,40%**	1					9		3	3	
Support	**2,20%**	1				3	1			1	3
Preismodell	**0,80%**							3			
Absolute Bedeutung		**3,2**	**2,0**	**1,6**	**2,1**	**2,7**	**2,9**	**1,7**	**3,2**	**4,1**	**3,0**
Relative Bedeutung		**11,9%**	**7,5%**	**5,9%**	**7,8%**	**10,2%**	**11,0%**	**6,6%**	**12,3%**	**15,5%**	**11,3%**
Rang		**3**	**8**	**10**	**7**	**6**	**5**	**9**	**2**	**1**	**4**

Abbildung 2: House of Quality (gemäß QFD-Methodik) der Gegenüberstellung von Anforderungsgruppen und Lösungsgruppen

Die bedeutendste Lösungsgruppe (in den Spalten in Abb. 2) zur Überwachung zielt dabei auf die Sicherstellung einer inhaltlich und technisch zuverlässigen Arbeitsweise der IT-Lösung ab. Die zweitwichtigste Lösungsgruppe fokussiert auf die Schnittstellen zur Kommunikation der verschiedenen Hintergrundsysteme der Mobilitätsanbieter untereinander, vor allem innerhalb der Welt der VDV-KA. Die vier folgenden Lösungsgruppen zur Anzeige, zu den Updates, zu Kontrolle und Prüfung sowie zur Installation und Konfiguration berücksichtigen dann vor allem den aufwendungsarmen und sicheren operativen Betrieb.

4 Einfache Anbindung an die IT-Infrastruktur in einem Koopkurrenznetzwerk: eine Appliance

Die Anforderungsworkshops zeigten deutlich, dass die Praxis der KMU eine einfach zu installierende und zu wartende IT-Lösung für die Teilnahme an einem Koopkurrenznetzwerk wie der (((eTicket-Welt benötigt. Ideal wäre ein zuverlässig funktionierendes, ausfall- und datensicheres „Rundum-Sorglos-Paket“, das keine zusätzlichen IT-Ressourcen für die KMU bindet. Gerade diese Zuverlässigkeits- und Sicherheitsaspekte sprechen für eine Lösung, bei der alle benötigten Komponenten wie Überwachungs- und Sicherheitsmechanismen integriert und an einer Stelle konfiguriert und kontrolliert werden können. Eine „out-of-the-box“-Lösung im Stile einer Appliance, also eine einzelne Einheit oft bestehend aus Hard- und Software, die aber (nahezu) keine Handlungen von außen während des Betriebs benötigt, erfüllt diese Voraussetzung bestens [Hi13b]. Ähnlich einem WLAN Router für den Internetzugang im privaten Haushalt soll eine Appliance als vorinstalliertes, vorkonfiguriertes, direkt einsetzbares Endgerät für Mobilitätsdienstleister die Nutzung des VDV-KA-Standards d. h. den Beitritt in die (((eTicket Deutschland Welt leisten.

4.1 Basisfunktionen einer Appliance

Neben der Kernfunktionalität der Appliance, die in erster Linie im Datenaustausch und der Datenkonvertierung sowie der Verbindung von unterschiedlichen Systemlandschaften innerhalb des Koopkurrenznetzwerks liegt, bietet die Appliance verschiedene Basisfunktionen, die für alle drei Ausbaustufen von Bedeutung sind. Hierunter sind als erstes umfassende Funktionen für die Systemüberwachung (Monitoring) zu nennen, mit denen insbesondere der Zustand der Hardware (CPU, Festplatte, etc.), aber auch Daten und Software automatisiert auf Fehler überprüft werden können. Zusätzlich zu einem Logging der Ergebnisse dieser Überprüfungen, kann mittels einer verschlüsselten Verbindung der Status weitergemeldet werden. Somit kann ein Systemadministrator alle Appliances derjenigen Kunden, die sich für die Dienstleistung System-Monitoring entschieden haben, überwachen. Die Möglichkeit eines Monitoring von Transaktionsdaten besteht mittels einer verteilten Open Source Lösung (ICINGA). Dadurch können Kommunikationsfehler mit anderen Systemen und auch mögliche Betrugsversuche aufgedeckt werden. Es wird ein Backup der Daten – wahlweise in die Cloud oder im LAN – angeboten, mit dem die Daten gesichert und nach einer festzulegenden Zeit datenschutzkonform archiviert werden. Funktionen für ein automatisiertes Remote-Update (inkl. eines Rollbacks für den Fall, dass Fehler auftreten) und eine Auslagerung von Konfigurationsdateien zur Vereinfachung der Administration sind vorhanden. Des Weiteren werden verschiedene VDV-KA-Versionen

unterstützt und eine Konverterfunktion angeboten, um Daten bzw. Transaktionen von einer Version in eine andere zu konvertieren und so die Abwärtskompatibilität zu erhöhen.

4.2 Gewährleistung der Systemsicherheit und des Datenschutzes

Innerhalb eines Koopkurrenznetzwerks ist ein hohes Maß an Sicherheit unumgänglich. Ein zentraler Baustein, um Sicherheit, Verfügbarkeit, und Skalierbarkeit sicherzustellen, ist die Nutzung von Cloud Computing. Insbesondere für kleine und mittlere Unternehmen (oftmals ohne dedizierter IT-Administrator mit unzureichend gesicherter Serverraum) stellt der Bezug von Software, Rechen- und Speicherkapazität „aus der Cloud" einen deutlichen Fortschritt im Hinblick auf Verfügbarkeit, Sicherheit und Skalierbarkeit dar.

Ein weiteres wesentliches Sicherheitsmerkmal stellen hier hardwaregebundene Zertifikate dar (sog. SAMs), da durch sie der Datenaustausch mit den ION-Vermittlungsstellen (Cloud) gewährleistet werden kann. Die Zertifikate können beispielsweise auf sogenannten Chipschlüsseln (Security-Token) basieren und haben den Vorteil, dass mit ihrer Hilfe eine erheblich höhere Sicherheitsstufe der Kommunikation im Gesamtnetz erreicht werden kann. Heutige Sicherheitsmedien sind bis Common Criteria (Common Criteria for Information Technology Security Evaluation; kurz CC) EAL 5+ (Evaluation Assurance Level) verfügbar und besitzen somit eine sehr hohe Sicherheitseinstufung. Folgende Funktionalitäten können mittels dieser Sicherheitsmedien sichergestellt werden:

- Das Medium wird mit der Appliance verbunden, um Zugangsdaten, -parameter, sowie die benötigten Zertifikate und Schlüssel gesichert in die Appliance zu übertragen (d. h. nahezu konfigurationsfreie Inbetriebnahme). Anschließend kann das Medium wieder an einem sicheren Ort aufbewahrt werden (z. B. in einem Tresor).
- Als Diebstahls- und Manipulationsschutz werden beim Ausschalten der Appliance alle Daten entfernt und müssen neu geladen werden, um überhaupt am Netzwerk teilnehmen zu können.
- Bei Verlust kann das Sicherheitsmedium zentral gesperrt werden.
- Die Datenübertragung zwischen der Vermittlungsstelle (ION-VS) und einer Appliance kann durch eine Verschlüsselungsmöglichkeit eines solchen Sicherheitsmediums zusätzlich geschützt werden.
- Ebenso kann der sichere Datenaustausch innerhalb des Unternehmens durch eine weitere Appliance gewährleistet werden, z. B. von zwei Filialen, in denen je eine Appliance Typ B vorhanden ist, an die Zentrale (mit Appliance Typ A) gezeigt.
- Implementierung des Sperrlistenmanagements, das die Prozesse zur Sperrung von Nutzermedien, Applikationen und Berechtigungen regelt. Aufgrund des engen Zeitfensters bei der Abholung der Sperrlisten durch die Verkehrsunternehmen erlaubt die Appliance eine deutliche Entzerrung des Datentransfers, da sie als Zwischenspeicher auch die Übertragung von Differenzsperrlisten und ein deutlich verbessertes Scheduling ermöglicht [VDV13a], [VDV13b].

In Verbindung mit dem ausgereiften Rollenmodell der VDV-KA kann zudem sichergestellt werden, dass schützenswerte Daten der einzelnen Unternehmen (z. B. Kunden- und Nutzungsdaten) nur dort wo sie für die Leistungserbringung oder Abrechnung notwendig sind, mit den anderen Unternehmen im Koopkurrenznetzwerk geteilt werden.

5 Erfahrungen aus dem operativem Betrieb

Die Appliance in Ausbaustufe 1 ist seit Juli 2013 beim dem Kreisverkehr Schwäbisch Hall (KVSH) sowie seit November 2013 bei dem Heilbronner-Hohenloher-Haller Nahverkehr (H3NV) im Piloteinsatz. Erfreulich ist dabei, dass Supportanfragen aus IT-Sicht bisher nahezu ausbleiben. Auch der testweise Austausch einer Aprikose ohne (!) IT-Spezialisten vor Ort im Sinne eines „plug & play" hat mit geringem telefonischen Support (0,5h) problemlos funktioniert. In Stufe 2 ist eine Aprikose ebenfalls in Heilbronn in Betrieb, dort wird die Abrechnung von Fahrten mit einem Taxi unterstützt, das in nachfragearmen Zeiten als sog. Rufbus/Ruftaxi eine Buslinie bedient. Der Fahrgast checkt – ganz so wie er es in Heilbronn inzwischen auch in Bus und Bahn macht – beim Ein- und Ausstieg in das Taxi ein und aus, nur eben an einem NFC-fähigen Smartphone, das der Taxifahrer bei sich hat und nicht an einer Bake, die am Bahnsteig oder im Bus fest verbaut ist. Die Erfahrungen aus der Erstinstallation der Aprikose sind ebenfalls positiv. Unter der Voraussetzung, dass die Hardware-Komponenten verfügbar sind, ist es möglich eine Aprikose selbst in der umfangreichsten Produktvariante innerhalb von fünf Arbeitstagen zum Einsatz zu bringen. Kundenseitige Voraussetzung ist dabei die Verfügbarkeit einer Datenbank bzw. entsprechender umfassender HGS zur Speicherung der Daten, die über die Aprikose geleitet werden. Für die Zukunft ist technisch wie lizenzrechtlich zu prüfen, ob auch z. B. eine MySQL-Datenbank innerhalb der Appliance mit VPN-Tunnel zur externen Kommunikation denkbar ist.

Variante	Module (M1-M8) / Szenarien (S1-S8)	M1: KVP-Anbindung	M2: PV-Anbindung	M3: Handy-Server	M4: SAM-Server	M5: Lizenz-Server	M6: Technische Datenkommunikation	M7: Proprietäre Schnittstellen	M8: Monitoring
Basis	**S1: Klassisches ÖPNV (Stufe 1)**	X	(X)				Webservice SOAP (Filetransfer) (Rest-WS)	(CSV) (HuSST)	technisch (fachlich)
	S2: CiCo – ÖPNV (Stufe 1)	X	X				Webservice SOAP Filetransfer	(CSV) (HuSST)	technisch (fachlich)
Mobil	**S3: Handy-Ticket (Stufe 1)**	X		X	X		Webservice SOAP Filetransfer Rest-WS	CSV JSON	technisch
	S4: Anschlussmobilität mit VDV-KA (St. 2)	X		X	X		Webservice SOAP Rest-WS	JSON	technisch
Lizenz	**S5: Handy-Kontrolle (Stufe 1)**	X		X	X	X	Webservice SOAP Rest-WS	CSV	technisch
	S6: CiCo-Taxi (Stufe 2)	X		X	X	X	Webservice SOAP Rest-WS	CSV HuSST JSON	technisch
Potential	**S7: Anschlussmobilität ohne VDV-KA (Stufe 2)**	Individuallösungen in Abhängigkeit der vorhandenen HGS					(Webservice SOAP) (Filetransfer) (Rest-WS)	JSON	technisch
	S8: Kooperation Einzelhandel (Stufe 3)	Individuallösungen in Abhängigkeit der vorhandenen HGS					(Webservice SOAP) (Filetransfer) (Rest-WS)	HuSST	technisch

Abbildung 3: Produktabgrenzung der Aprikose durch Gegenüberstellung zentraler Szenarien und Module (Angaben in Klammern entsprechen optionalen Erweiterungen)

Die Rückmeldungen aus dem technischen Betrieb vor Ort führten auch zu einer klareren internen Modulstruktur, die gleichzeitig den Weg zur eindeutigeren fachlichen Produktabgrenzung ebnet (siehe Abb. 3). Die acht Szenarien lassen sich in vier prinzipielle Produktvarianten einer Aprikose-Produktlinie zusammenführen. Den Kern bildet das Basissystem, welches die Szenarien S1 und S2, die klassischen ÖPNV-Szenarien mit Anschluss an VDV-KA-konforme HGS abbilden. Die obligatorische Anbindung an ein System eines Kundenvertragspartners (i. d. R. ein Verkehrsunternehmen, M1) kann dabei ergänzt werden um eine Anbindung an ein System eines Produktverantwortlichen (PV, i. d. R. ein Verkehrsverbund, M2). Dies passiert insbesondere dann, wenn eine Form der Fahrkettenbildung für das Aufzeichnen eines haltestellengenauen Ein- und Ausstiegs (sog. Check-in-Check-out – CiCo) benötigt wird. In Verbindung mit einem Identifikationsmedium (z. B. eine Kundenkarte) und entsprechenden Identifizierungsstellen in den Verkehrsmitteln können Fahrgäste jederzeit Ein- und Aussteigen und die (zusätzlich anzuschaffende) PV-Software berechnet den korrekten Preis für die Fahrt. Die Abrechnung kann dann beispielsweise über eine Kundenkarte mit Prepaid-Funktionalität erfolgen.

Die Szenarien S3 und S4 erweitern das Basissystem um die Möglichkeit des Handytickets bzw. genauer um die Möglichkeit, Tickets auf mobilen Endgeräten z. B. in Form eines 2-D-Barcodes anzuzeigen. Sinnvoll ist dabei die Anbindung an Auskunftssysteme der entsprechenden Verkehrsverbünde, um neben Preis- auch Routinginformationen zu liefern. Technisch kommen hier zwingend Handy- und Sicherheits-Server (sog. SAM-Server) zur sicheren Datenkommunikation hinzu (Module M3 und M4). Wenn neben der Nutzung von mobilen Endgeräten auf Endkundenseite diese auch zur Fahrgast-Kontrolle auf Dienstleisterseite zum Einsatz kommen (Szenarien S5 und S6), dann entspricht dies dem Lizenzpaket. Hier ist auf technischer Seite zusätzlich ein Lizenzserver (M5) zur Verifikation der mobilen Endgeräte nötig. Die Szenarien S7 und S8 nehmen eine Sonderstellung ein, da sie eine individuelle Anpassung an proprietäre HGS benötigen. Dafür haben beide ein besonderes Potenzial für die Anbahnung von neuen Koopkurrenznetzwerken, da sie die Kosten für deren IT-Unterstützung senken und somit ein Business Case zur Einrichtung des Netzwerks ggf. erst positiv beurteilt wird. So hat ein Carsharing- oder Fernbusanbieter, der seinen Kunden eine Anschlussmobilität ähnlich dem City-Ticket der Deutschen Bahn [DB14] erlauben möchte, nicht nur Verhandlungen über die kommerziellen Konditionen mit jedem Verkehrsverbund zu führen, sondern auch jeweils IT-Schnittstellen zu klären. Letzteres kann durch S7 mit entsprechenden Schnittstellen deutlich vereinfacht werden.

6 Fazit und Ausblick

Sowohl aufgrund von Änderungen im Kundenverhalten als auch aufgrund neuartiger Angebote im Bereich der Mobilitätskonzepte wie (freefloating oder stationsgebundenes) Carsharing, Fahrradverleihsysteme oder Systeme der kollaborativen Mobilität werden sich Kunden zur Deckung ihrer Mobilitätsbedürfnisse einer zunehmenden Anzahl an Anbietern bedienen. Durch Kooperationen können diese Unternehmen – oftmals auch zusammen mit dem ÖPNV – Einstiegshürden gerade für die Nutzung von Elektromobilität verringern und Nutzer enger an sich binden. Dazu ist eine Kooperation mehrerer Anbieter notwendig, die in vielen Fällen gleichzeitig in Konkurrenz zueinander stehen. Die entstehenden Koopkurrenznetzwerke stellen vor allem IT-seitig für KMU eine Herausforderung

dar. Hier setzt das Projekt Aprikose an, indem eine Appliance für den einfachen, sicheren und vergleichsweise kostengünstigen Anschluss an eine solche IT-Infrastruktur entwickelt wird. Die Appliance wird kundenorientiert mittels der Produktplanungsmethode Quality Function Deployment (QFD) in seiner inkrementellen Variante Continuous QFD entwickelt. Die für CQFD notwendige offene und vertrauensvolle Atmosphäre in den Anforderungsworkshops mit kooperierenden, gleichzeitig aber auch konkurrierenden Stakeholdern zu schaffen, erforderte dabei intensive Kommunikation mit allen Beteiligten. Insbesondere wurde den Kunden explizit Zeit und Freiraum eingeräumt alle aufkommenden Anforderungen zu reflektieren und vor ihrem individuellen Hintergrund zu bewerten.

Aktuell weist die Appliance unterschiedliche Funktionen für die drei Ausbaustufen sowie Basisfunktionen, die in jeder Ausbaustufe relevant sind, auf. Erste Erfahrungen aus dem prototypischen Einsatz sind positiv. Aus technischer Sicht zeigt sich der Betrieb wie erhofft wartungsarm und zuverlässig. Dies entspricht genau einem wesentlichen Alleinstellungsmerkmal der Aprikose, dass sie eben als Appliance (nahezu) selbständig (fast) ohne IT-Kenntnisse vor Ort betrieben werden kann. Die Evaluierungen mit den Verkehrsunternehmen vor Ort und damit die Bewertung auf Basis der fachlichen Anforderungen ist derzeit im Gange. Erste Rückmeldungen zeigen dabei, dass der Ansatz des Projektes, den Fokus auf KMU zu legen und diesen durch die Nutzung einer Appliance den Anschluss an ein Koopkurrenznetzwerk zu vereinfachen, erfolgversprechend ist.

Im Bereich des intermodalen Verkehrs sind in Zukunft folgende Funktionen vorgesehen: neben einer Weiterentwicklung der Bereiche Datenkonvertierung und Zwischenspeicherung (im Falle des Ausfalls einer Gegenstelle) ist vor allem die Aufnahme weiterer Schnittstellen (insbesondere Schnittstellen zu den wichtigsten Anbietern, z. B. autonetzer, car2go, nextbike oder Flinkster) zu nennen. Ebenfalls sollen Funktionen im Bereich der Unterstützung eines übergeordneten Kartenmanagements entwickelt werden, durch welche die Ausgabe von neuen oder Ersatzkarten, Änderungen von Berechtigungen, Sperrungen (z. B. wegen Verlust, Diebstahl oder Zahlungsrückständen) oder Auswertungen hinsichtlich der Inanspruchnahme verschiedener Dienstleistungen zukünftig möglich werden. Zusätzlich zu diesen Funktionen, die auch im Rahmen der Unterstützung mobilitätsnaher, komplementärer Dienstleistungen zum Tragen kommen, werden für die Stufe 3 weitere Schnittstellen von/zu Buchungs- und Bonussystemen sowie ein rudimentäres, cloudbasiertes Buchungs- und Kundenverwaltungssystem angestrebt. Damit lassen sich sowohl Angebote im Bereich Tourismus (z. B. als sog. Touristenkarten bzw. Citycards) als auch regionale Kundenbindungsprogramme im Einzelhandel unterstützen [St13], [Tu11].

Insbesondere die Verknüpfung mit Angeboten für Touristen, beispielsweise in Form einer Karte, die ein spezifisch auf die Bedürfnisse von Wochenendtouristen angepasst ist, und beispielsweise einen SPA-Besuch, ein 4-Gänge-Menü und ein Zeitkontingent für ein elektrisch betriebenes Carsharing-Fahrzeug oder zwei Pedelecs kombiniert, erscheint vielversprechend, um in einem nichtalltäglichen Kontext den Nutzern zu demonstrieren, wie vernetzte Mobilität aussehen kann. Dadurch können Verhaltensänderungen induziert und Ängste hinsichtlich der Benutzung elektrisch betriebener Fahrzeuge abgebaut werden. Ähnliche Wirkungen könnten erzielt werden, wenn Zeitkontingente für (elektrisch betriebene) Carsharing-Fahrzeuge, Leihfahrräder oder –pedelecs als Prämien in einem regionalen Kundenbindungsprogramm ausgegeben werden.

Literatur

[Ar09] Armbrust, M.; Fox, A.; Griffith, R.; Joseph, A.D.; Katz, R.H.; Konwinski, A.; Lee, G.; Patterson, D.A.; Rabkin, A.; Stoica, I.; Zaharia, M.: Above the Clouds: a Berkeley View of Cloud Computing. Technical Report. Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, 2009.

[BB13] Beckmann; J.; Brügger, A.: Kollaborative Mobilität, Internationales Verkehrswesen (65) 3, 2013.

[Be98] Beck, T.C.: Coopetition bei der Netzwerkorganisation, In: Zeitschrift Führung + Organisation, 67, 1998; S.271-276.

[BN11] Brandenburger, A.; Nalebuff, B.: Co-opetition, Random House Digital, 2011.

[CN07] Clements, P.; Northrop, L.M: Software Product Lines: Practices and Patterns, Addison Wesley, 2007.

[DB14] Deutsche Bahn AG: City-Ticket für BahnCard-Inhaber - Bequem von Tür zu Tür, http://www.bahn.de/p/view/bahncard/vorteile/cityticket.shtml

[FC09] Ficalora, Cohen, L.: Quality Function Deployment and Six Sigma, A QFD Handbook, Upper Saddle River, NJ, 2009.

[FS13] Follmer, R.; Scholz, J.: Mobilität der Zukunft bedürfnisorientiert statt technikfixiert, Internationales Verkehrswesen (65) 3, 2013.

[He00] Herzwurm, G.; Schockert, S.; Mellis, W.: Joint Requirements Engineering. Vieweg, Braunschweig, 2000.

[Hi13a] Hinger, P.; Sonnentag, K.; Meier-Berberich, J.: Stuttgart Services -Die Bürgerkarte für alle Fälle. In: VDV (Hrsg.), Der Nahverkehr - Öffentlicher Personenverkehr in Stadt und Region, Augabe7-8, 31. Jg., Alba Fachverlag, 2013.

[Hi13b] Hitachi ID Systems, Inc.: Appliances vs. Traditional Servers: Pros and Cons. http://hitachi-id.com/docs/appliance-server-pros-cons.pdf, Zugriff am 17.08.2013.

[Ka84] Kano, N.: Attractive Quality and Must-be Quality; Journal of the Japanese Society for Quality Control, No. 4, S. 39-48, 1984.

[Mü09] Münzel, G. et. Al: Cloud Computing - Evolution in der Technik, Revolution im Business, BITKOM-Leitfaden, Berlin, S. 13-17, 2009.

[Mo14] Mobilität 21 - Das Portal für innovative Verkehrslösungen, http://www.mobilitaet21.de/

[RP09] Reichwald, R; Piller, F.T: Interaktive Wertschöpfung. 2. Aufl, Gabler, Wiesbaden, 2009.

[RB00] Reiß, M.; Beck, T.: Netzwerkorganisation im Zeichen der Koopkurrenz. In: Foschiani, St. u.a. (Hrsg.): Strategisches Management im Zeichen von Umbruch und Wandel, Festschrift für Prof. Dr. Erich Zahn zum 60. Geburtstag, Stuttgart, S. 315-340, 2000.

[Sc13] Schockert, S.; Herzwurm, G.; Helferich, A.: Application of QFD within a co-opetition network of public transport organizations. ISQFD 2013, Santa Fe, New Mexico, USA.

[St10] Statistisches Bundesamt Deutschland: Statistisches Jahrbuch 2010, Kap. 16 – Verkehr.

[St13] Steinecke, A.: Management und Marketing im Kulturtourismus - Basiswissen – Praxisbeispiele – Checklisten, Springer, 2013

[SS13] Stuttgart Services (2013): http://www.livinglab-bwe.de/Project.aspx?id=92AB70A80F ADB245 A51A3DFD8F5CC02A#.Uj9Sp4YRDrH

[Tu11] Tuikka, T.; Siira, E.; Saukko, M.: City service discovery and access with Near Field Communication, In: Information Science and Service Science (NISS), 5th International Conference on New Trends 2011, S. 24-26.

[Va07] Van der Linden, F; Schmid, K; Rommes, E: Product Lines in Action, Springer, 2007.

[VDV10] VDV-Kernapplikation - KA_Technische Spezifikation, http://www.eticket-deutschland.de/spec-hd-bomv1107.pdfx, 2010.

[VDV13a] VDV-KA KG & Co. GmbH: VDV-Kernapplikation Lastenheft Sperrlistenservice-System für das (((eTicket Deutschland (KOSES), 2013.

[VDV13b] VDV-KA KG & Co. GmbH: VDV-Kernapplikation - Lastenheft Vermittlungsstellen im Interoperabilitätsnetzwerk der VDV-Kernapplikation (ION-VS), 2013.

Innovatives eCarSharing-Konzept am Beispiel des Hildesheimer Tandem-Modells

Carola Gerwig, Dennis Behrens, Ralf Knackstedt, Helmut Lessing

Betriebswirtschaft und Wirtschaftsinformatik
Universität Hildesheim
Marienburger Platz 22
31141 Hildesheim
gerwig@uni-hildesheim.de
dennis.behrens@uni-hildesheim.de

Abstract: In einem innovativen und nachhaltigen Nutzungskonzept werden Elektroautos an der Universität Hildesheim abwechselnd im innerbetrieblichen Carsharing und privat genutzt. Dieses sogenannte Tandem-Modell vereint unterschiedliche Carsharing-Ansätze und kann auf andere Nutzungskontexte übertragen werden. Dargestellt und analysiert wird das Konzept in einer Business Model Canvas mit Hilfe von Concept Maps. Darauf aufbauend wird ein verallgemeinertes Dienstleistungssystem entworfen, um das Geschäftsmodell auch an anderen Standorten anbieten zu können. Dabei werden die Anforderungen an die einzelnen Komponenten des Modells, beispielsweise an die IT-Komponente, herausgearbeitet, welches die Restriktionen im Betrieb von Elektroautos handhaben kann. Das so modellierte Dienstleistungssystem begegnet vielen der aktuellen Hemmnisse in dem Betrieb von Elektroautos und kann an anderen Standorten eingesetzt werden.

1 Motivation

Carsharing und Elektromobilität gehören zu den Schlüsselelementen zukünftiger Mobilitätskonzepte. Zur weiteren Etablierung von Elektromobilität ist die Forschung gefordert, neue Nutzungskonzepte auszuarbeiten (vgl. [Umw11]). Deutschland hat sich als Ziel gesetzt, Leitanbieter und Leitmarkt für Elektromobilität zu werden und deshalb vier große regionale Schaufenster initiiert, die interdisziplinär und öffentlichkeitswirksam nach systemübergreifenden Lösungen suchen. Die dabei angestrebten Zielsetzungen stehen in Einklang mit Zielen der Energiewende: So können Elektromobile nicht nur mit regenerativem Strom betrieben werden, sondern bieten auch ein hohes Potenzial für Lastverschiebungen im Smart Grid (vgl. [RVRJ11]). Weitere Vorteile liegen in dem emissionsfreien Betrieb und den niedrigen Lärmbelastung vor Ort. Allerdings sind Elektroautos mit einer Stückzahl von 12.000 PKW in der Bundesrepublik noch kaum verbreitet (vgl. [KB14]), wobei die geringe Reichweite der Elektroautos, die lückenhafte Ladeinfrastruktur, die hohen Anschaffungskosten sowie Berührungsängste als Haupthemmnisse gelten (vgl. [TTH+12]). Die Methoden der Wirtschaftsinformatik stellen dabei verschie-

dene Ansätze bereit, um bestehende Systeme zu analysieren, zu erweitern und neue zu entwickeln. So fordert [KBBM14] in einer systemübergreifenden Auswertung von aktuellen Forschungsansätzen im Handlungsfeld der Elektromobilität einen Fokus der Wirtschaftsinformatik auf „rigorose gestaltungsorientierte Forschungsansätze“ zu legen.

Im Folgenden wird ein Ansatz vorgestellt, welcher eine innovative Nutzung im Rahmen eines Dienstleistungskonzeptes für Elektromobilität beinhaltet. Das sogenannte Tandem-Modell, welches bereits an der Universität Hildesheim umgesetzt und erprobt wird, stellt in seiner Kombination von Merkmalen aus unterschiedlichen Carsharingkonzepten eine neue Variante dar, welche die speziellen Anforderungen von Elektrofahrzeugen berücksichtigt. Es handelt sich dabei um ein Carsharing zwischen der Universität und Mitarbeitern verbunden mit einem innerbetrieblichen Carsharing. Innerhalb der Arbeitszeiten werden die Elektroautos für Dienstreisen genutzt; Außerhalb der Arbeitszeiten gehen die Autos in den privaten Gebrauch einzelner Mitarbeiter über, die hierfür einen monatlichen Pauschalbetrag zahlen. Durch die Gestaltung des Tandem-Modells werden gängige Akzeptanzhemmnisse entkräftet: Der Mitarbeiter muss keine Kauf- oder Leasingpreis tragen, der Leasinggeber trägt im Rahmen des Leasingvertrages das Risiko der Batterielebensdauer und eine geeignete Ladeinfrastruktur wird am Arbeitsplatz über einen Solarcarport und weitere Ladestationen bereitgestellt. Der Arbeitgeber wiederum zieht in diesem Modell den Vorteil aus günstigen Leasingkonditionen und der bequemen Organisation von Dienstfahrten. Zudem generiert er ein modernes Image und ein attraktives Angebot an seine Mitarbeiter.

Um das neue Nutzungskonzept einzuordnen, werden in Kapitel 2 zunächst verwandte Ansätze und Modellierungsverfahren beschrieben. Darauf aufbauend wird das Tandem-Modell, wie es an der Universität Hildesheim derzeit erprobt wird, dargestellt. Für die Modellierung des Nutzungskonzeptes wird die Business Model Canvas (BMC) nach [OP11] genutzt. Zusätzlich werden Concept Maps (Cmaps) verwendet, welche die Abläufe verdeutlichen. In einem weiteren Schritt wird das Dienstleistungssystem verallgemeinert und auf Anforderungen an das Geschäftsmodell untersucht. In einem Fazit werden die analysierten Stärken und Schwächen zusammengefasst sowie die Entwicklungspotenziale des Modells aufgezeigt.

2 Related Work

2.1 Aktuelle Carsharing-Modelle

Die Idee des Carsharing findet bereits seit mehr als 20 Jahre in verschiedenen Konzepten Anwendung und setzt auf eine organisierte, gemeinschaftliche Nutzung einer oder mehrerer Fahrzeuge. In Deutschland haben sich in nahezu allen großen Städten Carsharing-Anbieter etabliert. Vor allem *stationsgebundene* Carsharing-Modelle sind weit verbreitet. Das Auto wird an einer zentralen Station abgeholt und nach der Fahrt dort wieder abgegeben. Mehr Flexibilität versprechen *free-floating* Anbieter, bei welchen das Auto an jedem freien Parkplatz im vorgegebenen Nutzungsgebiet abgestellt und dadurch zurückgegeben werden kann. Eine weitere junge Entwicklung ist das Carsharing unter Privatleuten – das sogenannte *peer-to-peer* (P2P) Carsharing. Die Anbieter stellen Zusatzversicherungen und

eine Plattform, auf der sich die Nutzer vernetzen können, zur Verfügung. Insgesamt waren in Deutschland nach [Ce13] bereits 2012 über 11.000 Fahrzeuge im Carsharing Betrieb, die von ca. 450.000 Menschen in Anspruch genommen wurden. Die Anzahl der Nutzer nimmt jährlich zu und es gilt als ein Mobilitätstrend, auf das eigene Auto zu verzichten und stattdessen andere umweltfreundlichere und zugleich kostengünstige Modelle zu nutzen (vgl. [fM11]). Speziell auf diese neue Nutzergruppe können neue Geschäftsmodelle ausgerichtet werden.

Die hohe Auslastung der Autos im Carsharing ist ein Vorteil, welcher gerade für Elektroautos mit hohen Anschaffungskosten und geringen Betriebs- und Reparaturkosten attraktiv ist. Zudem unterstreicht Carsharing den ökologischen Anspruch, den die Einführung von Elektroautos verfolgt. Daher liegt es nahe, dass sich bereits Geschäftsmodelle entwickelt haben oder in der Erprobung befinden, die Elektroautos nutzen: Car2Go[1] ist ein Carsharing-Anbieter des Automobilherstellers Daimler. Das Unternehmen betreibt seit 2011 eine rein elektrische Flotte von 300 Smarts in Amsterdam. Es handelt sich um ein free-floating Modell, bei dem das Laden der Autos von einem Service Team übernommen wird. Dieses Modell wird nun auch in Kooperation mit dem Energieversorgungsunternehmen EnBW, welches 500 Ladepunkte und regenerativ erzeugten Strom zur Verfügung stellen möchte, in Stuttgart umgesetzt[2]. Auch die Carsharing-Anbieter anderer großer Autokonzerne ziehen nach: Quicar, das Carsharing von Volkswagen möchte Elektrofahrzeuge in seinen operativen Carsharing-Betrieb in Hannover einbinden. Dabei handelt sich um ein klassisches, stationsgebundenes Carsharing-Konzept. Das Projekt „Elektromobilität im Stauferland (EMiS)“[3] versucht, ein Modell für Städte mittlerer Größe zu erproben, und startet mit zwei stationsgebundenen Elektroautos in der Stadt Schwäbisch Gmünd.

In weiteren Projekten des Schaufensters Elektromobilität werden Geschäftsmodelle für bestimmte Gewerbegruppen entwickelt: So wird in Stuttgart ein Geschäftsmodell für den Einsatz von Elektrofahrzeugen im Taxiverkehr ausgearbeitet (GuEST)[4], in den Projekten „EMiS“und „AES“[5] werden Hybridfahrzeuge in der Abfallwirtschaft erprobt. Der Vorteil dabei ist, dass aufgrund der leisen Elektromotoren Leerungen außerhalb der üblichen Zeiten möglich sind. Dies ist auch die grundlegende Idee des Projektes „NANU!“[6], in dem Belieferungs- und Transportprozesse durch die Nutzung von Elektromotoren in die Nachtstunden gelegt werden können. Im Projekt „eCarsharing für Gewerbekunden“[7] werden Unternehmen im Ausbau ihrer gewerblichen Flotte von Elektrofahrzeuge beraten.

[1] car2go Nederland B. V.; URL: https://www.car2go.com/en/amsterdam/, Zugriff am 15.04.2014

[2] e-mobil BW GmbH. Ladeinfrastruktur Stuttgart und Region (LIS); URL:http://www.livinglab-bwe.de/projekt/lis/l, Zugriff am 15.04.2014

[3] EMiS – Emobilität im Stauferland,URL: http://www.emis-projekt.de/, Zugriff am 15.04.2014

[4] e-mobil BW GmbH. GuEST – Gemeinschaftsprojekt Nutzungsuntersuchungen von Elektrotaxis in Stuttgart; URL: http://www.livinglab-bwe.de/projekt/guest/, Zugriff am 16.04.2014

[5] Berliner Agentur für Elektromobilität eMO. ElektroAES Einsatz von drei E-Entsorgungsfahrzeugen in der Abfallwirtschaft; URL: http://www.emo-berlin.de/de/schaufenster/projekte/gueterverkehr/elektro-abfallentsorgung/, Zugriff am 16.04.2014

[6] Berliner Agentur für Elektromobilität eMO. NANU! Mehrschichtbetrieb und Nachtbelieferung mit elektrischen Nutzfahrzeugen; URL: http://www.emo-berlin.de/de/schaufenster/projekte/gueterverkehr/nanu/, Zugriff am 16.04.2014

[7] Hannover Braunschweig Göttingen Wolfsburg GmbH Metropolregion. eCarsharing für Gewerbekunden; URL: http://www.metropolregion.de/pages/themen/schaufenster_emobilitaet/projektbeschreibung/fahrzeuge_flotten_verkehr/subpages/ecarsharing_fuer_gewerbekunden_/index.html, Zugriff am 16.04.2014

Die Dienstleistung umfasst eine Bedarfsanalyse und die Erstellung eines individuellen Carsharing-Nutzungskonzeptes mit Elektrofahrzeugen. Das Projekt „RheinMobil“[8] setzt auf eine hohe Auslastung durch Dienst- und Pendlerfahrten bei Siemens und Michelin. Durch Schnellladetechnik wird eine hohe Fahrzeugverfügbarkeit errreicht. All diese Projekte erproben ein stationsgebundenes Flottenmanagement für eine bestimmte Nutzergruppe. Das Ziel ist jeweils, ein wirtschaftliches Einsatzprofil einer E-Flotte nachzuweisen.

Das Tandem-Modell ist ebenfalls ein stationsgebundenes Carsharing für eine bestimmte Nutzergruppe: Unternehmen mit einer hohen Nachfrage an Dienstreisen über eine kurze Distanz. Es ist unabhängig vom Standort, da sich die Größe der Flotte nach dem Bedarf des Unternehmens richtet und sie dadurch tagsüber ausgelastet ist. Ein Alleinstellungsmerkmal zu allen vorgestellten Projekten ist die Idee, dass der private Nutzer (der Tandem-Partner) außerhalb der Arbeitszeiten das Elektroauto zur freien Verfügung hat. Dies kommt dem ursprünglichen Gedanke des „Besitzens“ nahe. Damit kann das Modell Unternehmensmitarbeiter ansprechen, die bisher nicht zur Gruppe der Carsharing-Nutzer gehören, da sie in ihrer Freizeit frei über ein Auto verfügen wollen. Weiterhin steht die Nachhaltigkeit des Betriebs im Zentrum der Projektes: Als Wertversprechen des Tandem-Modell soll ein nachhaltiger und wirtschaftlicher Betrieb nachgewiesen werden. Hierfür müssen Möglichkeiten geschaffen werden, regenerativ erzeugten Strom zu laden.

2.2 Modellierungstools

Um das Nutzungskonzept orts- und kontextunabhängig zu beschreiben, wird zuerst das bestehende Modell analysiert, modelliert und anschließend verallgemeinert. Es existieren verschiedene Ansätze, um Geschäftsmodelle darzustellen und zu analysieren. Eine Möglichkeit stellt der Value Code bzw. Value Imaging [BLS01] dar. Hier wird das Geschäftsmodell als Wert bzw. auf Werte basierend verstanden. Die Werte werden im Unternehmen erhoben und anschließend visuell dargestellt (Value Imaging). Das Value Model [GA02] stellt ebenfalls die erhobenen Werte in den Vordergrund. Es fokussiert dabei auf „E-Business-Situationen“ [DL03]. Beide Verfahren haben als Zielgruppe den Vorstand bzw. die Entscheidungsebene innerhalb der Unternehmen. Neben diesen Möglichkeiten existieren weitere Ansätze wie die E-Business Model Schematics [WV13] und die Notation der Geschäftsmodelltypologien [Wir01]. Beide dienen jedoch überwiegend der Schulung bzw. Simulation und sind für die Bereiche Beschaffung und Absatz konzipiert [DL03]. Keiner dieser Ansätze genügt den Anforderungen, um das Hildesheimer Tandem-Modell zu analysieren und anschließend zu verallgemeinern.

Für die Analyse und Weiterentwicklung von neuartigen und innovativen Geschäftsmodellen bedarf es eines einfachen Konzeptes, welches auf einen Blick die wesentlichen Komponenten dar- und insbesondere die logischen Zusammenhänge herausstellt. Daher wird die Business Model Canvas von Osterwalder [OP11] als Grundlage zur Modellierung des Tandem-Modells gewählt. Darüber hinaus werden Conepts Maps verwendet, um die logischen Abläufe in der Canvas abzubilden. Sie sorgen für eine Visualisierung der Zusammenhänge im Modell und können gut in die BMC integriert werden (vgl. [Fri12]).

[8] BEM Bundesverband eMobiiliät e. V.; RheinMobil URL:http://www.bem-ev.de/projekt-rheinmobil/, Zugriff am 22.06.2014

Die BMC hat den Vorteil, dass die Darstellung übersichtlich ist. Sie besteht aus neun Sichten, welche miteinander in Verbindung stehen und jeweils einen Aspekt des Geschäftsmodells abdecken. Dies stellt eine Vorstrukturierung dar und ermöglicht damit eine Vergleichbarkeit, um das Geschäftsmodell einordnen zu können. Es berücksichtigt dabei, wie die anderen Modelle auch, Kosten und Erlösaspekte, ohne sich jedoch in der Entwicklung bereits auf einzelne Werte festzulegen. Die weiteren sieben Sichten der Canvas lassen sich grob in das zentrale Werteversprechen, den Kundenbereich auf der rechten Seite und die Zulieferungsaspekte auf der linken Seite unterteilen. Der Kundenbereich besteht dabei aus der Kundenbeziehung, den Vertriebskanälen und dem Kundensegment. Auf der Zuliefererseite sind die Schlüsselaktivitäten, -ressourcen und -partner zu finden. Die BMC stellt einen verbreiteten Ansatz dar und bietet somit eine gute Diskussionsgrundlage, da das grundlegende Konzept bekannt ist und die Diskussion auf das Geschäftsmodell und damit den Inhalt der BMC konzentriert werden kann.

3 Hildesheimer Tandem-Modell

An der Universität Hildesheim wird im Rahmen des bundesgeförderten Projektes „eAutarke Zukunft “[9] ein innovatives eCarsharing-Modell erprobt, welches eine nachhaltige Versorgung und eine hohe Auslastung der Elektromobile anstrebt. In dem sogenannten Tandem-Modell teilt sich ein Mitarbeiter mit der Universität je ein Elektroauto. Das Auto wird während der Arbeitszeit für Dienstfahrten von Universitätsmitarbeitern genutzt, außerhalb dieser Zeiten steht es dem jeweiligen Mitarbeiter (Tandem-Partner) für den privaten Gebrauch zur Verfügung (auch an den Wochenenden und im Urlaub). Hierfür zahlt er dem Arbeitgeber einen monatlichen Festpreis, der alle anfallenden Kosten (Leasing, Versicherung, Strom und Wartung) deckt. In einem festen Zeitfenster während der Arbeitszeit des Mitarbeiters (z. B. von 9-16 Uhr) steht das Auto anderen Universitätsmitarbeitern zur Verfügung, die damit Dienstreisen wahrnehmen oder zwischen den verschiedenen Universitätsstandorten hin- und herreisen können. Ein Webportal bietet eine Übersicht über die Belegzeiten und ermöglicht eine bequeme Reservierung. Ein Solarcarport am Hauptcampus der Universität versorgt die Elektroautos mit Strom. Zudem besteht mit Hilfe eines speziellen Adapters die Möglichkeit, das Elektroauto auch an der heimischen Steckdose zu laden.

Im Folgenden sollen nun die einzelnen Punkte bzw. Sichten des Geschäftsmodells nach Osterwalder, wie es in Abbildung 1 gezeigt wird, erläutert werden.

3.1 Wertversprechen

Das zentrale Wertversprechen des Geschäftsmodells ist der nachhaltige und wirtschaftliche Betrieb einer Firmenflotte mit Elektroautos sowie die Bereitstellung eines attraktiven Mobilitätsangebots für Mitarbeiter der Universität. Dabei muss garantiert werden, dass die

[9] Hannover Braunschweig Göttingen Wolfsburg GmbH Metropolregion. eAutarke Zukunft http://www.metropolregion.de/pages/themen/schaufenster_emobilitaet/projektbeschreibung/energie/subpages/eautarke_zukunft__loesungen_im_arealnetz_/index.html, Zugriff am 22.04.2014

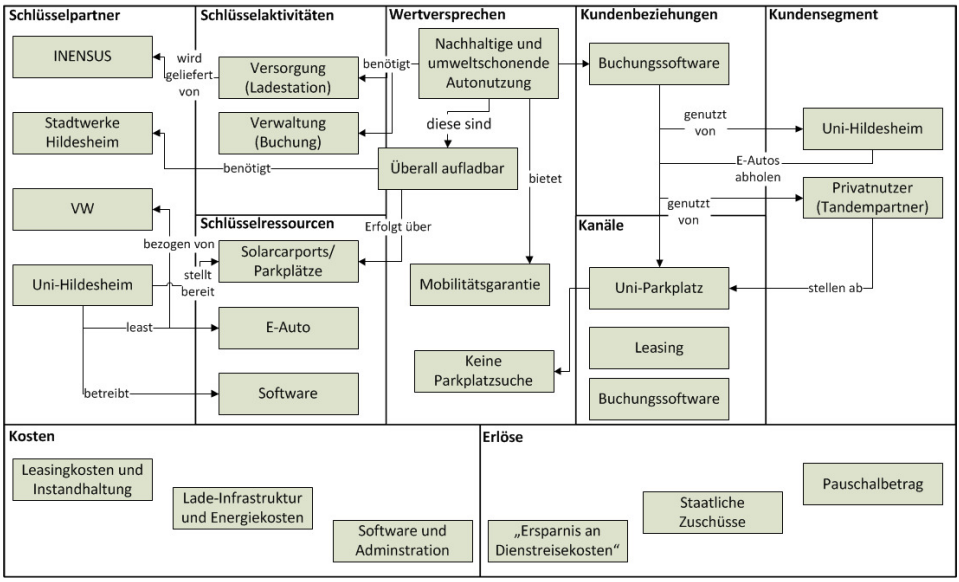


Abbildung 1: Das Hildesheimer Tandem-Modell in der Business Model Canvas

Autos für alle geplanten Fahrten ausreichend geladen sind. Für Dienstreisen stellt das eine Buchungssoftware sicher, welche die Fahrten koordiniert und nur bei entsprechenden Kapazitäten zulässt. In der Zeit, in der Autos ungenutzt sind, werden sie über ein Solarcarport oder das öffentliche Versorgungsnetz mit nachhaltigem Strom geladen. Um auch dem Tandem-Partner außerhalb der Dienstzeiten eine Mobilitätsgarantie geben zu können, wird ihm ein Adapter zur Verfügung gestellt, der auf jede beliebige konventionelle Steckdose aufgesetzt werden kann und diese in eine Ladestation „verwandelt“. Dies macht das Mobilitätsangebot insbesondere für Personen interessant, welche die Möglichkeit haben, eine heimische Steckdose in der Garage zum Laden zu nutzen. Im besten Fall nutzt auch der Tandem-Partner regenerativ erzeugten Strom zum Laden an seinem Wohnort.

Trotz der Lademöglichkeiten ist die begrenzte Reichweite ein limitierender Faktor, so dass das Modell hauptsächlich für Unternehmen und Mitarbeiter attraktiv ist, welche viele kurze Strecken zurücklegen müssen und die für Strecken über 100km auf andere Mobilitätsangebote zurückgreifen können.

3.2 Schlüsselaktivitäten, -ressourcen und -partner

Die wichtigsten Aktivitäten im Geschäftsmodell sind die Stromversorgung der Elektroautos und die Verwaltung derselben über eine Buchungssoftware. Die Ladestationen in Hildesheim werden von der Firma INENSUS geliefert. Die Buchungssoftware ist eine Eigenentwicklung der Universität und wird von dieser betrieben. Sowohl die Ladestationen als auch die Buchungssoftware stellen Schlüsselressourcen dar, ohne die das Geschäftsmodell

nicht funktionieren würde. Sogenannte SmartPlug-Adapter ermöglichen eine sichere und abrechenbare Aufladung an jeder Steckdose und bilden zusammen mit dem Solarcarport die Grundlage für eine durchgehende Mobilität. Einen weiteren Anteil an der Mobilitätsgarantie trägt die Buchungssoftware: Das Verwaltungssystem prognostiziert die voraussichtlichen Ladezustände auf der Grundlage von Reservierungsdaten und Erfahrungswerten. Eine eingehende Reservierung wird zunächst auf ihre Machbarkeit geprüft, bevor sie im positiven Fall einem geeigneten Elektroauto zugewiesen wird.

Eine weitere zentrale Schlüsselressource stellen die Elektroautos selbst dar. Durch die Möglichkeit eines $C0_2$-freien Betriebs sind sie maßgeblich für das Wertversprechen einer ökologischen Nachhaltigkeit verantwortlich. Damit sind allerdings Restriktionen für die Stromzufuhr verbunden. Abhängig davon, ob der Strom aus regenerativen Quellen selbst erzeugt, als Ökostrom aus dem Netz oder als herkömmlicher Strommix bezogen wird, ist das Aufladen der Autos und damit deren Nutzung mehr oder weniger umweltfreundlich. Im konreten Fall werden die Autos von der Volkswagen AG geleast und von einem nahen Autohaus im Wartungs- oder Schadensfall betreut.

3.3 Kundenbeziehungen und Vertriebskanäle

Zur erfolgreichen Durchfhrung des Projekts müssen zuerst Tandem-Partner gewonnen und das Interesse der reisenden Mitarbeiter geweckt werden. An der Universität wurde dies mit einer Onlineumfrage bzgl. des Bedarfs verbunden (s. Kapitel 3.5). Die weitere Kundenbeziehung erfolgt hauptsächlich über die Buchungssoftware und den angeschlossenen Support. Die Buchungssoftware wird hauptsächlich für die Buchung von Dienstreisen verwendet. Dabei sichert sie nicht nur die Bereitstellung eines ausreichend geladenen Elektroautos für zugelassene Fahrten, sondern gewährleistet auch den vertraglich zugesicherten Ladezustand für die Tandem-Partner nach Dienstschluss. Im Schadensfall wenden sich die Mitarbeiter ebenfalls an den Support, welcher das weitere Schadensmanagement übernimmt.

Die Kundenkanäle umfassen weiterhin die Bereitstellung von Ladestationen vor Ort, welche auf extra hierfür bereitgestellte Parkplätzen auf dem Gelände der Universität angebracht sind. Dadurch wird zum einen eine ausreichende Batterieladung gewährleistet und zum anderen den Tandem-Partnern und auch den Mitarbeitern der Universität die Parkplatzsuche erspart. Angeboten wird das gesamte Geschäftsmodell, bzw. die Dienstleistung über das Buchungsportal, so dass dieses ebenfalls bei den Kanälen aufritt.

3.4 Kostenstruktur und Erlöse

Die monatlichen Kosten setzen sich aus den Leasing- und Instandhaltungskosten für die Elektromobile, die Bereitstellung und dem Betrieb der Ladestationen sowie die Erstellung und die Administration der Buchungssoftware zusammen. Davon übernimmt der Tandem-Partner einen monatlichen Pauschalbetrag in Höhe von ca. 50% der kalkulierten Gesamtkosten. Hinzu kommen staatliche Zuschüsse im Rahmen des Verbundvorhabens.

3.5 Forschung und Evaluation des Hildesheimer Tandem-Modells

Mit den Tandem-Modell greift die Universität die aktuellsten Entwicklungen im Bereich der Mobilität auf: Die Verbreitung von umweltfreundlichen Elektroautos, die Ausweitung von Carsharing sowie die veränderten Mobilitätsbedürfnisse. Die Schwerpunkte liegen dabei im Bereich der nachhaltigen Stromversorgung, der Entwicklung eines Webportals zur Buchung der Fahrzeuge, der betriebswirtschaftlichen Evaluation des Tandem-Modells und der Nutzerakzeptanz sowie deren Entwicklung über den Verlauf der Systemeinführung.

Zunächst wurde in einer Onlineumfrage der Bedarf an Dienstfahrten in der Region ermittelt und zusätzlich die Einstellung gegenüber Elektroautos abgefragt. Dabei haben 30% aller Mitarbeiter an der Universität Hildesheim teilgenommen. Bislang hatten mehr als 90% der Befragten noch keine Erfahrungen mit Elektrofahrzeugen gemacht. Dennoch gaben knapp 50% der Befragten an, dass sie ein Elektroauto für eine gute Alternative zum traditionellen Pkw halten. Die Einschätzung und Einstellung gegenüber Carsharing ist ähnlich hoch: 46% halten Carsharing für eine gute Alternative zum eigenen Pkw, auch wenn insgesamt erst 17% der befragten Mitarbeiter Carsharing genutzt hatten. Die üblichen Strecken werden in der Region zurückgelegt. So haben 60% angegeben, seltener als einmal im Monat überhaupt eine Strecke von mehr als 100km zu fahren. Ferner würden 25% der Befragten das Tandem-Modell auch als Privatperson für 250 Euro im Monat nutzen.

Im weiteren Verlauf werden mit den Tandem-Partner qualitative, leitfadengestützte Face-to-Face-Interviews durchgeführt. Die Nutzer des Carsharing-Dienstes sollen anhand eines quantitativen Online-Fragebogens befragt werden. Anhand der Befragungen sollen Aspekte zur Kaufentscheidung, dem Informationsgrad und zur Bewertung von Elektromobilität ermittelt werden und zu einer allgemeinen Treiber-/Hemmnisanalyse führen.

Seit Dezember 2013 sind zwei Elektroautos im Betrieb des Tandemmodells, welche im März 2014 um zwei weitere ergänzt worden sind. Simultan zu den Elektroautos ist der erste Prototyp des Buchungssystems in den Betrieb geführt worden. Die weitere Entwicklung erfolgt sukzessive und die erste empirische, betriebswirtschaftliche Analyse des Nutzungskonzeptes wird gegen Ende des Jahres vorgenommen werden.

4 Innovatives Dienstleistungssystem auf Basis des Tandem-Modells

Das vorgestellte Dienstleistungssystem wird im Folgenden verallgemeinert, damit es nicht an den Standort Hildesheim gebunden ist und auch über den universitären Kontext hinaus genutzt werden kann. Hierfür werden die Schlüsselkomponenten des modellierten Konzeptes in der BMC analysiert, um so orts- oder unternehmensspezifische Entscheidungen zu identifizieren und durch verallgemeinerte Anforderungen an die einzelnen Komponenten zu ersetzen, durch welche das zentrale Werteversprechen einer nachhaltigen und ökonomischen Mobilität auch in der Verallgemeinerung bestehen bleiben kann. Daraus ergeben sich die Potentiale, aber auch Grenzen des Nutzungskonzeptes als dienstleistungsbasiertes Geschäftsmodell.

4.1 Schlüsselpartner

Im vorliegenden Geschäftsmodell werden e-up!s von VW verwendet. Die örtliche Nähe des Autokonzerns und dessen Entwicklungsabteilung ist für die Universität bezüglich des Datenmanagements vorteilhaft. Die Ladeinfrastruktur wird von der Firma INENSUS geliefert. Diese ist in Goslar angesiedelt und deshalb räumlich nah an der Region Hildesheim. Sie kann in Störungsfällen schnell reagieren, um diese zu beheben. An anderen Standorten kann es jedoch sinnvoll sein, die Elektrofahrzeuge von einem anderen Anbieter zu leasen bzw. die Ladestationen von einem anderen Hersteller zu beziehen.

4.2 Schlüsselressourcen

Damit das Geschäftsmodell nachhaltig betrieben werden kann, müssen die Autos mit nachhaltig erzeugtem Strom geladen werden. An der Universität ist dafür ein Solarcarport gewählt worden. Grundsätzlich sind Solarcarports mit Energiespeichern eine gute Möglichkeit, die nachhaltige Versorgung zumindest in Teilen zu gewährleisten. Eine weitere Möglichkeit ist die Anbindung an andere regenerative Erzeuger mit Speicher oder an ein Smart Grid, welches das Laden des Fahrzeugs innerhalb des Verbundes optimiert.

Ein wesentlicher Faktor für die Flexibilität des Konzeptes sind die umgesetzten Leistungsstufen an den Ladesäulen. Die Universität verwendet zunächst Wallboxen, an denen mit dem einheitlichen Typ 2 Standardstecker oder dem SmartPlug geladen werden kann (einphasig; 230V/16A, 3,7 A). Dies bedeutet, dass entsprechende Ladezeiten von der Buchungssoftware eingeplant werden müssen. Für ein Unternehmen mit einer hoch frequentierten und stromintensiven Nutzung der Elektroautos können Schnellladesäulen die bessere Wahl sein, da die Standzeiten der Autos entsprechend geringer sind.

Ein weiterer Gewinn an Flexibilität kann durch eine Fuhrparkerweiterung erreicht werden. Gerade bei einer hohen Anzahl an Kurzstrecken bietet es sich an, Pedelecs, E-Bikes und E-Roller in das Angebot zu integrieren. Die Buchungssoftware müsste an die spezifischen Parameter (Reichweiten, Ladezeiten) angepasst werden, um diese Erweiterung des Dienstleistungsangebots anbieten zu können. Der Organisation der Fahrzeuge könnte sich ebenfalls über einen monatlichen Pauschalbeitrag der Tandem-Partner amortisieren, welche die Mobile außerhalb der Arbeitszeit nutzen können. Darüber hinaus kann es vorteilhaft sein, für lange Strecke Autos mit Verbrenner- oder Hypridmotor in das Konzept mitaufzunehmen, wobei der Nachhaltigkeitsaspekt dabei zurückgestuft wird.

4.3 Schlüsselaktivitäten

Im Vorfeld ist es wichtig, den Bedarf an Dienstreisen im Unternehmen innerhalb eines bestimmten Radius um den Standort festzustellen, so dass die Größe der Elektroflotte entsprechend ausgelegt werden kann. Der Kreis der Mitarbeiter im Unternehmen, welche auf das Elektroauto Zugriff haben, sollte in einem entsprechenden Verhältnis zur Nutzungshäufigkeit und Anzahl der Elektroautos stehen.

Die IT-Infrastruktur, im Speziellen die Buchungssoftware, stellt einen zentralen Bestand-

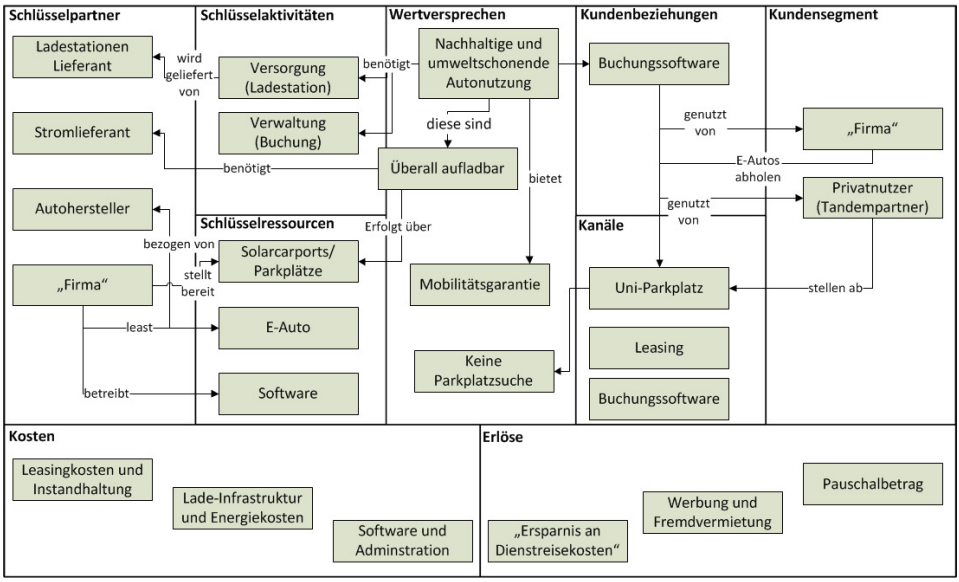


Abbildung 2: Das allgemeine Dienstleistungssystem in der BMC

teil in der Abwicklung und somit in der Umsetzung des Geschäftsmodells dar. Diese wird momentan an der Universität Hildesheim als Artefakt entwickelt und ist auf die lokalen Anforderungen zugeschnitten. Je nach Flotte, Art und Ort der Ladestationen kann das Verwaltungssystem entsprechend angepasst werden. Ganz allgemein ist den Nutzern eine möglichst reibungslose Buchung und Nutzung des Elektroautos zu ermöglichen. Zunächst muss es über eine Webseite oder eine App möglich sein, sich in das Buchungssystem einzuloggen und Reservierungen der Elektroautos bzw. der weiteren Fahrzeuge vorzunehmen. Dabei muss die Software prüfen, ob Fahrzeuge verfügbar und ob der voraussichtliche Ladezustand ausreichend ist. Ist eine Buchung vorgenommen, so wird aufgrund der geplanten Kilometerzahl der anschließende Ladezustand kalkuliert. Innerhalb der Parkzeiten werden die Elektromobile geladen, wofür je nach Ladesäule und Fahrzeug entsprechende Ladekurven zu berücksichtigen sind. Diese können anhand der historischen Daten für jedes Fahrzeug über Regressionsverfahren von der Verwaltungssoftware gelernt werden.

Darüber hinaus sind mehrere Erweiterungen des Geschäftsmodells denkbar, an welche das Verwaltungssystem angepasst werden kann. Eine Erweiterung des bisherigen Konzeptes wäre es, einen Tausch der Autos innerhalb der Tandem-Partner zu erlauben. Dies würde der Verwaltungssoftware mehr Spielraum in der Planung gestatten. Ein anderes Szenario ist die Vervielfachung der Ladeinfrastruktur an verschiedenen nahen Unternehmensstandorten: Sind an unterschiedlichen Standorten Stellplätze vorgesehen, so ergeben sich dadurch neue Lademöglichkeiten, aber auch Restriktionen in der Verfügbarkeit, welche von der Buchungssoftware berücksichtigt werden müssen. Hierbei muss abgeklärt werden, ob die Erweiterung der angebotenen Dienstleistungen im Verhältnis zu den Mehrkosten bzw.

dem Mehraufwand stehen.

Eine weitere IT-Komponente, welche bisher an der Universität Hildesheim noch nicht umgesetzt ist, ist die Idee des intelligenten Lastmanagements. Aktuell werden die Elektroautos, sobald sie an die Ladestation angeschlossen sind, aufgeladen und der Ladevorgang schaltet sich erst bei Erreichen der vollständigen Batterieladung aus. Alternativ könnte über das Buchungssystem der Zeitraum vorgegeben werden, in dem das Elektroauto eine bestimmte Ladeleistung aufnehmen soll. Hierdurch würde sich eine erhöhte Flexibilität ergeben, welche das Stromnetz entlasten und ggf. auch Kosten senken könnte (vgl. [BG14]). Dieses zu integrieren wird insbesondere bei flexiblen Stromtarifen attraktiv.

4.4 Kundensegment, -beziehungen und Vertriebskanäle

In der konkreten Umsetzung des Nutzungskonzeptes ist die Universität Hildesheim der Leasingnehmer und damit der Betreiber der Elektroautos. Der Bedarf an häufigen, kurzen Dienstfahrten ergibt sich durch die im Stadtgebiet verteilten Standorte und Dienstreisen zu lokalen Projektpartnern. Zudem wohnen die meisten der mehr als 700 Mitarbeiter in einem Umkreis von 50km und kommen damit als mögliche Tandem-Partner in Frage. Die hohe Auslastung der Elektroautos durch kurze Fahrten ist ein essentieller Faktor für den wirtschaftlichen Betrieb.

Die Bereitstellung der Ladeinfrastruktur ist eine wesentliche Komponente der Kundenkanäle. Zur Umsetzung muss gewährleistet werden, dass sowohl genug Parkmöglichkeiten als auch die Ladeinfrastruktur für die Elektroautos vorhanden sind. Auch bei den Tandem-Partnern muss privat eine Lademöglichkeit, zumindest in Form einer erreichbaren Schokusteckdose vorhanden sein. Dabei wird das zentrale Wertversprechen der Nachhaltigkeit eingehalten, wenn der Arbeitgeber bereit ist, am Standort des Unternehmens für eine nachhaltige Energieversorgung zu sorgen. Angeboten wird die Dienstleistung für die Mitarbeiter wie im Hildesheimer Modell über die Buchungssoftware, welche ggf. in das bestehende Firmennetz eingebunden werden kann.

Das verallgemeinerte Modell richtet sich an Unternehmen mittlerer Größe, welche einen hohen Bedarf an kurzen Dienstreisen, einen geeigneten Standort zum Aufbau der Ladestationen und Interesse an einem umweltfreundlichen Mobilitätskonzept haben.

4.5 Kostenstruktur und Erlöse

Zuletzt muss die Finanzierung angepasst werden. Zum einen entfallen die staatliche Zuschüsse, dafür könnten Werbeeinnahmen generiert werden, in dem die Autos als mobile Werbefläche genutzt werden. Auch sind zusätzliche Einnahmen denkbar, wie die kostenpflichtige Benutzung der Ladeinfrastruktur durch Dritte oder die Vermietung der Parkplätze. Der teuerste Bestandteil der Elektroautos ist die Batterie. Man geht allerdings davon aus, dass durch technische Weiterentwicklungen sowie Skaleneffekte der Batteriepreis in den nächsten Jahren deutlich sinken wird (vgl. [BBH$^+$09]).

5 Fazit und Ausblick

Das vorgestellte Nutzungskonzept für Elektromobile, welches sich aktuell in der Erprobungsphase an der Universität Hildesheim befindet, stellt ein neuartiges und innovatives Dienstleistungssystem dar. In dem kombinierten Carsharing-Modell werden geleaste Elektroautos abwechselnd von der Universität für Dienstfahrten und von Mitarbeitern im privaten Kontext genutzt. Mithilfe der Business Model Canvas in Verbindung mit Concept Maps wurde das Geschäftsmodell modelliert, analysiert und verallgemeinert.

Das verallgemeinerte Modell ist nicht mehr an die örtlichen Gegebenheiten in Hildesheim gebunden, sondern stellt vielmehr die Anforderungen an die jeweiligen Komponenten heraus, die erfüllt werden müssen, um das Werteversprechen eines nachhaltigen und wirtschaftlichen Mobilitätskonzeptes beizubehalten. Einige Möglichkeiten, die Anforderungen an anderen Standorten oder bei anderen Bedarfssituationen umzusetzen, wurden skizziert – für den Einzelfall müsste aus dem verallgemeinerten Modell wieder eine spezifiziertes Modell gemäß des Anforderungprofils abgeleitet werden. Das allgemeine Modell bietet hierfür eine übersichtliche Grundlage, so dass die verschiedensten Aspekte des Dienstleistungssystems diskutiert und variiert werden können.

Wie sich das Tandem-Modell in der Realität bewährt, wird sich in dem Feldversuch an der Universität Hildesheim zeigen. Anhand der Erfahrungen wird die Modellierung in der Business Model Canvas nachgeführt werden und ggf. eine der vorgeschlagenen Erweiterungen aus Kapitel vier umgesetzt werden. So zeichnet sich an der Universität Hildesheim sowohl im privaten als auch betrieblichen Bereich ein großes Interesse an Pedelecs ab, so dass deren Einführung im Tandem-Modell geplant ist.

Literatur

[BBH+09] Markus Blesl, David Bruchof, Niklas Hartmann, Enver Doruk Özdemir, Ulrich Fahl, Ludger Eltrop und Alfred Voß. Entwicklungsstand und Perspektiven der Elektromobilität: Endbericht. 2009.

[BG14] Dennis Behrens und Carola Gerwig. Selbstregulierende Verbraucher im Smart Grid: Design einer Infrastruktur mit Hilfe eines Multi-Agenten-Systems. *Multikonferenz Wirtschaftsinformatik 2014-Tagungsband der MKWI 2014*, Seiten 935–948, 2014.

[BLS01] Richard ES Boulton, Barry D Libert und Steve M Samek. *Value code-Werte schaffen in der neuen Wirtschaft: Erfolgsstrategien-Geschäftsmodelle-Praxisbeispiele*. Econ, 2001.

[Ce13] Bundesverband CarSharing eV. Jahresbericht 2012. 2013.

[DL03] Thomas Deelmann und Peter Loos. *Visuelle Methoden zur Darstellung von Geschäftsmodellen-Methodenvergleich, Anforderungsdefinition und exemplarischer Visualisierungsvorschlag*. Johannes Gutenberg Universität, 2003.

[fM11] Institut für Mobilitätsforschung. Mobilität junger Meschen im Wandel - multimodealer und weiblicher, 2011.

[Fri12] Thomas Frisendal. *Design Thinking for Business Analysis*. Springer, 2012.

[GA02] Jaap Gordijn und Hans Akkermans. Value Based Requirements Engineering: Exploring Innovative e-Commerce Ideas. *REQUIREMENTS ENGINEERING JOURNAL*, 8:114–134, 2002.

[KB14] Kraftfahrt-Bundesamt. Bestandsbarometer am 1.Januar 2014. website, 2014. available online at http://www.kba.de/cln_031/nn_125398/DE/Statistik/Fahrzeuge/Bestand/2014__b__bestandsbarometer__teil2__absolut.html , visited on April 14th 2014.

[KBBM14] Benjamin Klör, Sebastian Bräuer, Daniel Beverungen und Martin Matzner. IT-basierte Dienstleistungen für die Elektromobilität–Konzeptioneller Rahmen und Literaturanalyse. *Multikonferenz Wirtschaftsinformatik 2014-Tagungsband der MKWI 2014*, Seiten 2047–2068, 2014.

[OP11] Alexander Osterwalder und Yves Pigneur. *Business Model Generation: Ein Handbuch für Visionäre, Spielveränderer und Herausforderer*. Campus Verlag, 2011.

[RVRJ11] Sarvapali D Ramchurn, Perukrishnen Vytelingum, Alex Rogers und Nick Jennings. Agent-based control for decentralised demand side management in the smart grid. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, Seiten 5–12. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[TTH+12] GA Teichmann, J Trützschler, C Hahn, PK Schäfer, A Hermann und K Höhne. Elektromobilität Normen bringen die Zukunft in Fahrt. Website, 2012. Available online at http://www.pwc.de/de_DE/de/offentliche-unternehmen/assets/studie_normung.pdf; visited on April 8th 2014.

[Umw11] Wissenschaftlicher Beirat Globale Umweltveränderungen. *Welt im Wandel: Gesellschaftsvertrag für eine Große Transformation*. Wiss. Beirat d. Bundesregierung Globale Umweltveränderungen, 2011.

[Wir01] Bernd W Wirtz. *Electronic business*. Gabler Verlag, 2001.

[WV13] Peter Weill und Michael Vitale. *Place to space: Migrating to eBusiness Models*. Harvard Business Press, 2013.

A Business Process Model for the Reverse Logistics of Used Electric Vehicle Batteries

Benjamin Klör, Sebastian Bräuer, Daniel Beverungen

European Research Center for Information Systems
University of Münster
Leonardo-Campus 3
D-48149 Münster
benjamin.kloer@ercis.uni-muenster.de
sebastian.braeuer@ercis.uni-muenster.de
daniel.beverungen@ercis.uni-muenster.de

Abstract: Electric vehicles (EVs) are considered as a ground-breaking innovation for making transportation services more sustainable. Electric vehicle batteries (EVBs) are the major technological components of electric vehicles and account for up to one third of the EVs' initial price. Due to continued cell degradation, EVBs need to be removed from the EVs after some time, but might still be reused in different scenarios such as stationary applications. EVBs that cannot be reused must be recycled to retain valuable materials such as cobalt, lithium, and nickel. The purpose of this paper is to develop a business process model for the reverse logistics of EVBs, transporting them from car workshops to their next usage sites or to adequate recycling facilities. We develop the business process model in a design research endeavor, based on reviewing selected laws and regulations that need to be complied with when transporting high-voltage batteries.

1 Motivation

Electric mobility is supposed to be a key technology to cope with scarce resources and exceeding emissions caused by private and public mobility [Na12]. Powering cars, buses, trains, or motorcycles with electric energy is envisioned to reduce our dependency on fossil energy sources such as oil and might help to eliminate emissions caused by conventional combustion engines, including carbon dioxide, fine dust, and noise [Fe13, Na12].

Electric vehicle batteries (EVBs) are essential components of each electric-driven means of transportation. EVBs are high-voltage power storage devices for supplying the motor of electric vehicles (EVs) with energy. However, for their genuine application in EVs, EVBs become useless over time due to the battery cells' degradation and the growth of its internal resistance [Bu09]. Only replacing an EVB can sustain an EV's maximum performance.

Consequently, used batteries need to be returned back to their vendors or waste management services governed by public law (e.g. §§5 (1), 13 (1) BattG). The treatment of used EVBs is a challenging reverse logistics effort that touches on many ecological, economic,

legal, safety-related, and technical issues. Anecdotal evidence suggests that many of the existing reverse logistics processes do not conform to these challenges in the EVB domain.

The purpose of this paper is to identify legal requirements that govern the reverse logistics of used EVBs in Germany, based on which a business process is designed. In a design science approach, a business process model is developed and evaluated in an expert interview based on a subset of the identified legal regulations. With this contribution, we set out to provide an early business process model including the specific requirements and impacts for the reverse logistics of EVBs.

The remainder of the paper is structured as follows. In Section 2, we review functions and implications of EVBs and reverse logistics. In Section 3, the design science approach is sketched. In Section 4, selected legal aspects that govern the reverse logistics of EVBs are discussed. In Section 5, the business process model is designed using the Business Process Model and Notation (BPMN). The business process model has been evaluated by checking its compliance with several legal regulations, as depicted in Section 6. Finally, Section 7 concludes the paper by providing a brief outlook on the future development of the reverse logistics of used EVBs.

2 Related Work

EVBs are complex technical components located in EVs, transforming chemical energy into electrical energy to propel the vehicle. An EVB is a battery pack, consisting of several battery modules, each of which consists of several battery cells [ACG+09]. Heating and cooling applications are available to keep the EVB's temperature in an optimal range [YWWS13]. A battery management system (BMS) is an embedded system to monitor and control vital functions of the EVB, including load balancing, temperature control, event logging, etc [LHL+13].

Due to an EVB's cell degradation and an increasing internal resistance, the amount of energy that can be stored in an EVB decreases over time [SUN13]. In order to sustain an EV's range and other driving characteristics, an EVB must be removed from the car as soon as its state of health[1] (SOH) drops below 80% of its initial value [Bu09, CLP+03]. Despite the end of its first life, this amount of energy is sufficient for reusing the EVB in other scenarios that are less demanding [SML+12]. Such a second life is especially feasible when considering the expensiveness of EVBs in terms of their materials and manufacturing process. A SOH of 50% signifies the EVBs' end of life, such that they have to be decommissioned and recycled to retain valuable materials like lithium, nickel, and cobalt [Bu11].

To make EVBs available for reuse or recycling, they have to be collected and transported in a reverse logistics effort. Reverse logistics is an essential part of companies' supply chains and is defined by the Reverse Logistics Executive Council as "[t]he process of planning, implementing and controlling backward flows of raw materials, in process inventory,

[1] EVB's state of health is defined as the ratio of the remaining capacity to the initial capacity [CLH06, YWWS13].

packaging and finished goods, from a manufacturing, distribution or use point, to a point of recovery or point of proper disposal"[Re14, RTL98]. Reverse logistics commonly comprises four main processes, including collection, a combined inspection/selection/sorting process, recovery, and redistribution [dBD03]. Others add that reverse logistics involves a decision process for determining a disposition location [TR02]. Disposition activities such as refabrication, refurbishment, reuse, and recycling are in the scope of reverse logistics [dBD03, RTL98] and apply for the disposition of EVBs, too [Ni10].

However, the generic reverse logistics processes need to be adapted to the specific characteristics of EVBs. For instance, included materials such as lithium [LHL+13] lead to the declaration of EVBs as dangerous goods. In particular, leaking batteries can cause severe damage to the human body and to the natural environment. Therefore, reverse logistics processes need to be designed with close reference to legal regulations that govern the handling and transportation of high-voltage batteries.

A multi-lingual[2] literature search in Google Scholar (number of hits: 52, relevant hits: 0), Springerlink (9, 0), AISEL (0, 0), Scopus (0, 0), and Web of Science (0, 0) resulted in no existing business process models for the reverse logistics of EVBs. Additionally, we have identified research projects dealing with logistics and technical recycling of lithium batteries in general, e.g., LithoRec[3] and LithoRec II[4]. However, the lack of related published scientific literature requests for designing a consolidated business process model in the domain of used EVBs' reverse logistics reflecting relevant legal aspects.

3 Research Method

We set out to design a business process model for the reverse logistics of used EVBs that is compliant with selected legal requirements for transporting high-voltage batteries. A business process model is an IT artifact, which is further defined as language constructs, models, methods, and software implementations [MS95]. Design science research [HMPR04] refers to the design of IT artifacts as a science, building on Simon's [Si77] argument on sciences of the artificial.

In line with the design science research processes [HMPR04, PTRC07] and IT artifacts as prescriptive IS theories [GJ07, Gr06], we conceptualize our research process as follows: First, we conducted a literature review to identify the basic activities that need to be performed in reverse logistics in general. Second, based on these insights and our knowledge about EVBs, we drafted a high-level process model for the reverse logistics of used EVBs. This model was then evaluated and adapted during semi-structured expert interviews with two industry representatives from a recycling consulting company. Third, we identified legal regulations that the business process needs to comply with. The selected legal regulations mainly depend on the experts' recommendations. Fourth, we iteratively redesigned

[2] (("electric vehicle battery" OR "traction battery" OR "evb" OR "traktionsbatterie" OR "antriebsbatterie") AND ("reverse logistics" OR "reverse flow" OR "rückwärtslogistik")), conducted on March 11th 2014.

[3] http://www.lithorec.de

[4] http://www.lithorec2.de

and detailed the business process model with consideration of the identified legal regulations and in collaboration with the experts, until saturation of this detailed business process model was reached. Fifth, for emphasizing the validity of the business process model, a matching between business tasks and legal regulations was conducted. Further criteria (e.g., process performance, limitations of handovers) have not been considered at this early point of research.

4 Legal Regulations Influencing the Reverse Logistics of EVBs

Abundant European and country-specific recommendations, regulations, and laws govern reverse logistics processes of EVBs. German and the superordinate European regulations, rules, and recommendations refer to the materials composed in EVBs such as the REACH regulation, the 'Chemikaliengesetz' (ChemG) or the 'Giftinformationsverordnung bzgl. Chemikalien' (ChemGiftInfoV), to laboratory protection (e.g. the 'Gefahrenstoffverordnung' (GefStoffV)), whereas other rules apply for transporting dangerous goods (e.g. 'Accord européen relatif au transport international des marchandises Dangereuses par Route' (ADR) and 'Gefahrgutverordnung Straße, Eisenbahn und Binnenschifffahrt' (GGVSEB)), and still other regulations apply to the recycling of batteries ('Batteriegesetz' (BattG)), cars (e.g. 'Altfahrzeugverordnung' (AltfahrzeugV)), or goods in general (e.g. 'Abfallverbringungsgesetz' (AbfVerbrG), Kreislaufwirtschaftsgesetz (KrWG)). Additional rules are released by authorities such as the Federal Institute for Materials Research and Testing (BAM) or company networks such as the German Electrical and Electronic Manufacturers' Association (ZVEI) and the Association of German Engineers (VDI).

Since we develop a first business process model for the reverse logistics of EVBs, we focus on those sources that allow for providing a comprehensive overview about the redemption, storage, and transportation of EVBs. On the other hand, further aspects with regards to labor protection and details about applied chemical materials and their individual handling are left out here for the sake of simplification.

As an overriding authority, the legislation of the European Union (EU) shapes the national regulations of the EU member states. For reverse logistics of EVBs the EU directive 66/2006/EC is important since it stipulates the duty of EVBs' redemption of OEMs and defines recycling proportions. EU directive 2008/68/EC concerns the inland transportation of dangerous goods. Another EU related regulation and probably the most important one for transportation issues is the ADR. The ADR is a superior regulation regarding transportation policies of dangerous goods and concerns manifold domain-related aspects such as packing, dispatching, carriage, and receipt of these goods.

From a German perspective, the KrWG contains general regulations about avoidance, reuse, recycling, exploitation, and disposal of waste in Germany (§6 (1) KrWG). Originally, the KrWG was supplemented by a legal decree concerning the disposal of batteries. In 2009 the KrWG was refined and this legal decree was transformed into the BattG. Additionally, the EU directive 66/2006/EC was also implemented into national law by the BattG. With respect to EVBs these regulations mainly focus on the OEMs' duty of

redemption and exploitation (§9 (1) sentence 1 BattG) that also applies for waste management service providers governed by public law (§13 (1) BattG). If exploitation cannot be achieved, an EVB's disposal is mandatory (§14 BattG). In the context of the BattG, EVBs are classified as industrial batteries (§2 (5) BattG).

Four additional national regulations are crucial for the handling, storage, and transportation of EVBs. First, the AltfahrzeugV is a regulation that mainly contains procedural instructions for EVBs, e.g., regarding the immediate removal of EVBs from EVs on receipt (Appendix 3.2.2.1 AltfahrzeugV) and certain instructions for interim storage of EVBs (Appendix 3.1.2.4 AltfahrzeugV). Second, the GGVSEB contains regulations for the transportation of dangerous goods with regard to different stakeholders such as the carrier (§19 (2) GGVSEB), the vehicle driver (§28 GGVSEB) and the consignee (§20 GGVSEB). GGVSEB transposes EU directive 2008/68/EC to national law and, thus, mainly refers to ADR regulations. Third, in accordance with §8 GGVSEB and 1.5.1.1 ADR the Federal Institute for Materials Research and Testing (BAM) is declared as the German national competent authority for assessing the adequate (pre-)processing of dangerous goods.[5] As a consequence for (reverse) logistics of EVBs, the 'Allgemeinverfügung für beschädigte Lithiumbatterien' (AVLIB), released by BAM, regulates the proper handling of damaged EVBs in general. For these batteries the ordinance covers concrete procedural instructions such as the safe packing of damaged EVBs (articles 4 and 5), adequate authority's notification of scheduled transportation (articles 6 and 7) or labeling of the shipped items as "damaged or defective lithium batteries" (article 8). In contrast, the ADR specifies handling and transportation instructions for undamaged EVBs that can be classified according to the UN numbers 3090, 3091, 3480 and 3481. Fourth and finally, the AbfVerbrG deals with regulations concerning national and transnational transportations of waste. For this work, especially the general obligation for labeling transportation vehicles is a crucial regulation written down to §10 AbfVerbrG.

5 A Business Process Model for the Reverse Logistics of EVBs

With respect to more general reverse logistics models [dBD03, TR02], we derive a high-level business process for the reverse logistics of EVBs (Figure 1) to illustrate the complexity of this endeavor.[6]

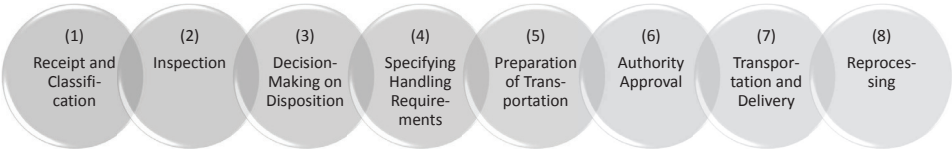


Figure 1: High-level process for the reverse logistics of EVBs

[5] http://www.unece.org/trans/danger/publi/adr/country-info_e.html#Germany

[6] Four crucial processes have been adapted from the theoretical reverse logistics models: 'receipt and classification' (1), 'inspection' (2), 'decision-making on disposition' (3), and 'reprocessing' (8).

Four abstract stakeholders (consignor, approving authority, carrier, and consignee as described by ADR) need to be considered for modeling the reverse logistics process. A consignor is an actor that takes back used goods. A carrier is responsible for preparing the transportation and for shipping the goods to a consignee who further processes the goods. The transport of dangerous goods needs to be permitted by an approving authority in advance.

The high-level process is triggered by the consignor (for instance, a car workshop) following the handover of an EV for replacing an EVB (1). The dismantled EVB has then to be inspected to detect any damages (2). This step is crucial to determine an appropriate disposition destination (3). Subsequently, handling requirements have to be specified (4) that also influence the preparation of the EVB's transportation (5). Depending on an EVB's condition, the approving authority is contacted to permit the transport (6). The transport itself is conducted by a qualified carrier (7). Finally, the consignee accepts the shipped item, such that all necessary reprocessing steps can be applied to the EVB (8).

Since the routing of used EVBs through this process needs to comply with the identified legal regulations and the process needs to be implemented into a software system, a more detailed process model is required. For refining the high-level process, we acquired domain-specific knowledge by industry representatives from a consulting company, who frequently deal with the design and implementation of recycling networks in their day-to-day work. Based on their feedback, the high-level process was detailed into a business process model designed with BPMN (Figure 2).

In the following, the detailed business process is described with reference to the involved stakeholders and the activities. In comparison to the high-level process, the only change regarding the stakeholders is the level of abstraction with respect to the roles of both the consignor and the authority. The consignor needs to be split up into the roles 'regular employee' and 'electrically qualified person'. This is necessary since only specifically trained personnel may operate high-voltage facilities. The authority 'federal institute for materials research and testing (BAM)' is divided into the roles 'risk prevention officer' and 'general authority' as there are operational and administrative tasks conducted by the BAM.

The detailed process resembles the abstract process and is started by the consignor. The first high-level process step 'receipt and classification' is represented by the tasks a regular employee carries out. A regular employee (for instance, a mechanic) of the consignor accepts an electric vehicle provided by a customer. The next step is to identify the containing EVB's type and specification, either by checking the EV's OEM manual or via database lookup in an OEM's knowledge base, or both. This identification is necessary since the further treatment of the EVB depends on its type (UN number). Subsequently, the EV is connected to a computer via an interface (e.g. CAN bus) to read out the EVB's data stored in the BMS. Among others, this data is required for making a sound decision for the optimal disposition of the EVB later on. The job of the regular employee is finished by contacting an electrically qualified person who is able to deal with the EVB on a technical level.

The next three high-level process steps 'inspection', 'decision-making on disposition', and 'specifying handling requirements' are carried out by an electrically qualified person[7] of the consignor or a risk prevention officer of the authority. These three processes are highly complex due to the involvement of several stakeholders and process decisions that have to be made.

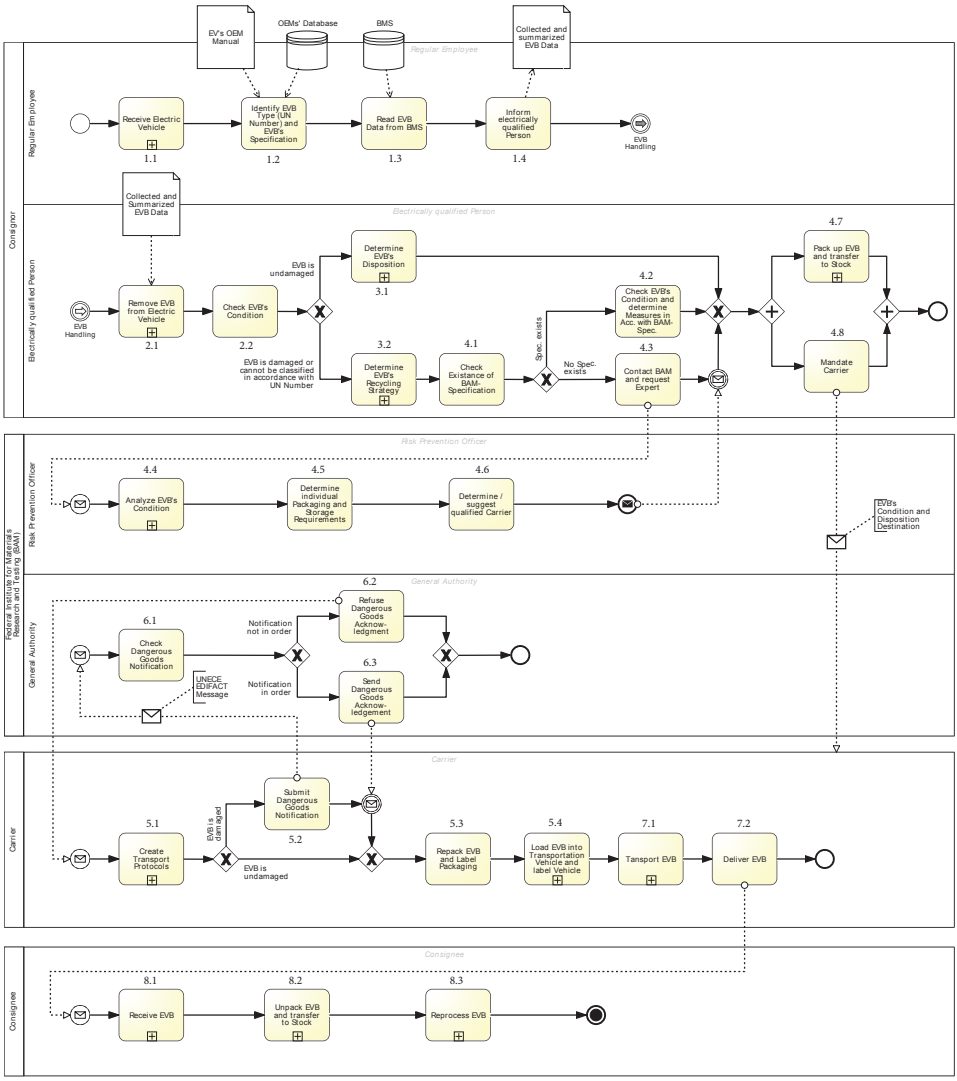


Figure 2: Detailed business process model for the reverse logistics of EVBs

[7]In Germany, a person with a proper qualification level is called 'Elektrofachkraft für Hochvolt-Systeme in Kraftfahrzeugen' as declared by BGI/GUV-I 8686.

The tasks concerning an EVB's removal and condition check are subordinated to the high-level process 'inspection'. An EVB's inspection implies operating with high-voltage systems. Conducting such highly specialized operational tasks requires a sophisticated education of an electrically qualified person. Based on the collected and summarized EVB data the removal of the battery is conducted since each OEM applies an individual design to the EVBs. Before removing the EVB, the qualified worker has to consider security rules, e.g., pulling the service-plug for unlocking the EVB or prevention of turning the EVB back on. Then, after unloosing screw connections the removal of the entire EVB is carried out. Afterwards, the EVB is checked for its general condition that is roughly divided into two states: (1) damaged and (2) undamaged. Checking an EVB's condition is carried out by visual control. The aim is to detect damages such as liquid or gaseous outflows caused by, e.g., leakages. This evidence is highly important for the rest of the reverse logistics process since a damaged battery triggers a special logistics processing of the EVB in comparison to an undamaged battery.[8] Treated insufficiently, a damaged EVB may cause severe and undesired effects such as fire, electric shocks, intoxication, chemical burns, or other serious harms to humans and the environment. Hence, it is necessary to apply extended reverse logistics services like sealing, boxing, etc. to damaged EVBs.

Considering the high-level process 'decision-making on disposition', determining an EVB's disposition is a binary decision. Based on the EVB's condition figured out before, the battery either has to be recycled or can be forwarded to a different disposition such as reuse, refabrication, and refurbishment. Furthermore, this decision is important for the high-level process 'specifying handling requirements' that is next.

If the EVB is found to be undamaged, there are no specific handling requirements other than described by common law dealing with the transportation of dangerous goods. Otherwise, if the battery is found to be damaged, recycling must take place. For that to happen, an adequate recycling strategy is selected. Then, a lookup for a BAM specification (special approval for the transportation of damaged EVBs) of the battery type is conducted. If such a specification is available, the containing procedural instructions apply, e.g., for storing, packing, and labeling the battery. If there is no BAM specification available, a risk prevention officer of the BAM has to be contacted for analyzing the damaged EVB's condition in detail, determining certain procedural instructions which should apply and suggesting a qualified carrier. Finally, operational processes of packing up the EVB and transferring it to the interim storage take place. Additionally, the suggested carrier is mandated who is responsible for the 'preparation of transportation'.

The 'preparation of transportation' is conducted by the suggested carrier. After creating the transport protocols, the EVB's condition again is the basic criterion for the carrier to proceed. If the battery is damaged, the high-level process 'authority approval' comes into play and the general authority has to be informed about the upcoming transportation by sending a (standardized) dangerous goods notification.[9] This federal instance then performs a contentual check of this notification and either accepts or declines it. If declined, the carrier potentially has to overwork the suggested packaging and labeling and needs to edit the document accordingly. Otherwise, the carrier is allowed to proceed with the

[8] This special processing also applies for EVBs that cannot be classified in accordance with the UN numbers.
[9] http://www.unece.org/trade/untdid/d01a/trmd/iftdgn_c.htm

preparation of transportation as declared in the transport protocols. In general, if the EVB is in an undamaged condition and has been classified according to an UN number, there is no authority approval needed for the transportation. If circumstances so require, the carrier repacks the battery for transportation and applies labeling to the completed packaging. The preparation is finished by loading and labeling the transportation vehicle.

Finally, the high-level processes 'transportation and delivery' and 'reprocessing' are conducted. A qualified driver of the carrier conducts the EVB's transportation from start to target destination under consideration of all operational requirements stipulated by traffic law. After EVB's delivery, the consignee concludes the entire reverse logistics process by unpacking, storing, and reprocessing the EVB.

6 Evaluation of Business Process Model

In the previous sections we presented relevant legal regulations and a detailed business process model for transportation of EVBs. In the following we conduct an evaluation based on a particularized mapping of aforementioned legal regulations to the detailed business process model. For that to happen, each process task is provided with a unique identifier (see Figure 2). The mapping is realized by a table that shows all process task identifiers on the left and matching legal regulations and their impact on the right (Table 1).

ID	Legal regulation	Impact
1.1	§5 (1) BattG; §6 (1) KrWG	Duty of OEM for redemption of EVBs
1.2	ADR 2013 2.2.9.1.7	Requirement for classifying EVBs according to an UN number allowing a handling according to ADR
	UNECE Manual of Tests and Criteria - Rev.5 subsection 38.3	Every cell and battery has to comply with a type, for which the testing requirements as specified in this definition have to be fulfilled. Otherwise, a transport is only possible on special permit.
	ADR 2013 3.3 SP 188	If cells and batteries meet requirements of SP 188, they are not affected by other rules of ADR.
1.3	e.g. D/BAM/GGVSEB (ADR)/2.2-256/12-rev.1, Annex	Testing of cells requires reading out data from battery management system for getting information about defective cells.
1.4	See 2.1	
2.1	BGI/GUV-I 8686 1. Anwendungsbereich	Necessity for proper qualification to deal with high-voltage facilities
	Appendix 3.2.2.1 AltfahrzeugV	EVBs have to be removed immediately on receipt.

continued ...

ID	Legal regulation	Impact
2.2	ADR 2013 1.5.1 → Multilateral Agreement M259 (2.2) ADR 2013 3.3 SP 661 AVLIB 3.1	Requirement for differentiating between damaged and undamaged EVBs In case of a damaged or defect battery or cell, the previous (erroneous) application has to be taken into account.
3.1	-	
3.2	ADR 2013 3.3 SP 661; Exceptions to SP 661 are possible under certain requirements ADR 2013 3.3 SP 636	Transportation of damaged EVBs requires approval of competent authority, including possible restrictions of packaging methods and transportation routes (tunnel restrictions).
4.1	ADR 2013 3.3 SP 661	Transportation of damaged EVBs requires approval of competent authority.
4.2	e.g. D/BAM/GGVSEB (ADR)/2.2-256/12-rev.1, Annex	BAM exception specifications contain detailed requirements concerning to be performed tests, packaging, driver, vehicles, labeling, etc.
4.3	See 4.1	
4.4	See 4.7	EVB's detailed condition determines packaging requirements (reacted vs. not reacted cells and batteries).
4.5	ADR 2013 3.3 SP 661	Necessity for competent authority to determine packaging, storage and transportation requirements. During granting, any recommendations of UNECE have to be taken into account.
4.6	-	
4.7	All: Appendix 3.1.2.4 AltfahrzeugV Undamaged: ADR 2013 1.4.1 (1.4.1.1), 1.4.3.2, 4.1.1, 4.1.3, 4.1.4.1 Damaged (not reacted): Multilateral Agreement M259, AVLIB 4 and 5 Damaged (reacted): ADR 2013 4.1.4.3 and M259, AVLIB 4 and 5	Depending on condition of EVB, specific packing and storage requirements apply (e.g. acid-proof packaging, storage location).
4.8	-	
5.1	ADR 2013 5.4.1 Damaged: AVLIB 7.1	Transport of EVBs has to be accompanied with structurally predefined transport protocols. If AVLIB is applied for transporting a damaged EVB, a copy of AVLIB has to be added to the transport documents.

continued ...

ID	Legal regulation	Impact
5.2	ADR 2013 3.3 SP 661 AVLIB 6.1.1	An acknowledgment is required for the transport of damaged EVBs.
5.3	All: ADR 2013 5.2.1, 5.3.1, §10 Damaged: AVLIB 8.1	Depending on battery state, the transport of EVBs requires special labeling of packaging and vehicles.
5.4	ADR 2013 1.4.3.1, 1.4.2.1 §10 AbfVerbrG	Duties for loader and consignor Requirements for labeling of the vehicle
6.1	AVLIB 6.1.3	Requirement for a proper dangerous goods notification and call for acknowledgment
6.2	AVLIB 6.1.2 and 6.1.3	Non-compliance of transport intent/message
6.3	AVLIB 6.1.2 and 6.1.3	Compliance of transport intent/message
7.1	§19 (2) GGVSEB; §28 GGVSEB Damaged: ADR 2013 3.3 SP 661	Existence of special duties of carrier and vehicle driver concerning road safety In case of damaged EVBs, a copy of the approval or a reference to the competent authority has to be included.
7.2	Damaged: AVLIB 6.1.5	If any incident occurred during transportation, reporting to BAM is required.
8.1	ADR 1.4.2.3, 1.4.3.7; §20 GGVSEB	Duties of consignee and unloader
8.2	Appendix 3.1.2.4 AltfahrzeugV	Requirements for storing EVBs
8.3	§6 (1) KrWG; §14 BattG	Duty for exploitation of waste

Table 1: Mapping of legal regulations to process tasks

7 Limitations and Outlook

The theoretical and managerial contribution offered by this paper is a business process model for the reverse logistics of EVBs that complies with some of the most important legal regulations in Germany. The work, therefore, extends previous thinking on this issue that has started to develop initial insights, but refrained from designing a detailed business process model with reference to selected legal regulations. Thus, our results are envisioned to inform subsequent research related to EVBs and electric mobility in general.

In the spirit of the design of IT artifacts as a search process [HMPR04], the business process model is envisioned to be a first step towards realizing compliant business processes and implementing them into software tools. Since the reverse logistics of EVBs is subject to abundant laws and regulations, the process must be further developed to comply with more regulations than could be reviewed here, capitalizing on the expert knowledge on lawyers and professionals for handling dangerous goods. Before putting the process into action, we suggest to perform a process of certification in addition, supervised by an independent certification agency.

Acknowledgements

This article was developed in the research project 'EOL-IS' (End-of-Life Solutions for Electric Vehicle Batteries – Development of Product Service Systems and Information Systems for Decision Support; funding label: 01FE13023) in the context of the funding program "Service Innovations for Electric Mobility" funded by the German Federal Ministry of Education and Research (BMBF). We thank the project management agency Projektträger im Deutschen Zentrum für Luft- und Raumfahrt (DLR) for their support. Furthermore, we acknowledge the support and advice of Mr. André Gierke and Mr. Timo Ratzke, both of whom work for Hellmann Process Management GmbH & Co. KG.

References

[ACG+09] Amirault, J., Chien, J., Garg, S., Gibbons, D., Ross, B., Tang, M., Xing, J., Sidhu, I., Kaminsky, P., and Tenderich, B. The Electric Vehicle Battery Landscape: Opportunities and Challenges. 2009.

[Bu09] Burke, A.: Performance, Charging, and Second-use Considerations for Lithium Batteries for Plug-in Electric Vehicles. Technical report. UC Davis: Institute of Transportation Studies (UCD). 2009.

[Bu11] Buchmann, I. Battery Testing at Garages Simplified, `http://www.buchmann.ca/article36-page1.asp` (Accessed on: April 20 2014). 2011.

[CLH06] Coleman, M., Lee, C., and Hurley, W.: State of Health Determination: Two Pulse Load Test for a VRLA Battery. In: *37th IEEE Power Electronics Specialists Conference*. Number 1. pp. 1–6. Jeju. 2006.

[CLP+03] Cready, E., Lippert, J., Pihl, J., Weinstock, I., and Symons, P.: Technical and Economic Feasibility of Applying Used EV Batteries in Stationary Applications. Technical report. Sandia National Laboratories. Albuquerque. 2003.

[dBD03] de Brito, M. P. and Dekker, R.: A Framework for Reverse Logistics. *ERIM REPORT SERIES RESEARCH IN MANAGEMENT*. ERS-2003-0:1–29. 2003.

[Fe13] Federal Ministry of Education and Research (BMBF): Electric Mobility – Rethinking the Car. Technical report. Bonn. 2013.

[GJ07] Gregor, S. and Jones, D.: The Anatomy of a Design Theory. *Journal of the Association for Information Systems*. 8(5):312–335. 2007.

[Gr06] Gregor, S.: The nature of theory in information systems. *Mis Quarterly*. 30(3):611–642. 2006.

[HMPR04] Hevner, A. R., March, S. T., Park, J., and Ram, S.: Design science in information systems research. *MIS quarterly*. 28(1):75–105. 2004.

[LHL+13] Lu, L., Han, X., Li, J., Hua, J., and Ouyang, M.: A review on the key issues for lithium-ion battery management in electric vehicles. *Journal of Power Sources*. 226:272–288. 2013.

[MS95] March, S. T. and Smith, G. F.: Design and natural science research on information technology. *Decision Support Systems*. 15(4):251–266. 1995.

[Na12] National Platform for Electric Mobility: Progress Report of the German National Platform for Electric Mobility (Third Report). Technical report. Berlin. 2012.

[Ni10] Nissan Motor Co. Ltd.: NISSAN Sustainability Report 2010. Technical report. 2010.

[PTRC07] Peffers, K., Tuunanen, T., Rothenberger, M. a., and Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*. 24(3):45–77. December 2007.

[Re14] Reverse Logistics Executive Council. Glossary: What is Reverse Logistics?, `http://www.rlec.org/glossary.html` (Accessed on: April 04 2014). 2014.

[RTL98] Rogers, D. S. and Tibben-Lembke, R. S.: *Going Backwards: Reverse Logistics Trends and Practices*. Reverse Logistics Executive Council. Reno, NV. 1998.

[Si77] Simon, H.: *The new science of management decision*. Prentice Hall. Englewood Cliffs. 1977.

[SML+12] Sachenbacher, M., Mayer, T., Leucker, M., Brand, M., and Jossen, A.: Towards 2nd-Life Application of Lithium-Ion Batteries for Stationary Energy Storage in Photovoltaic Systems. In: *Proceedings of the "International Conference on Solar energy for MENA region (INCOSOL)"*. pp. 22–23. Amman, Jordan. 2012.

[SUN13] Sasaki, T., Ukyo, Y., and Novák, P.: Memory effect in a lithium-ion battery. *Nature materials*. 12(6):569–75. June 2013.

[TR02] Tibben Lembke, R. S. and Rogers, D. S.: Differences between forward and reverse logistics in a retail environment. *Supply Chain Management: An International Journal*. 7(5):271–282. 2002.

[YWWS13] Young, K., Wang, C., Wang, L. Y., and Strunz, K.: Electric Vehicle Battery Technologies. In: Garcia-Valle, R. and Peças Lopes, J. a. A. (Eds.), *Electric Vehicle Integration into Modern Power Networks*. pp. 15–56. Springer New York. New York, NY. 2013.

Eigenverbrauchs- und Kostenoptimierung zur Untersuchung des Einflusses dynamischer Stromtarife auf die Netznutzung

Armin Wolf, Steffen Unger, Marc Schultz

IT4Energy
Fraunhofer FOKUS
Kaiserin-Augusta-Allee 31
10589 Berlin
{armin.wolf | steffen.unger | marc.schultz}@fokus.fraunhofer.de

Abstract: Im Folgenden wird eine Softwarekomponente für ein „Demand-Side-Management“ vorgestellt, die unter Nutzung von Batteriespeichern und verschiebbarer Lasten den Eigenverbrauch von selbst erzeugtem Strom – vorzugsweise aus regenerativen Quellen – optimiert, wobei der Bezug aus dem lokalen Stromnetz möglichst kostengünstig durch die Nutzung dynamischer Strompreise erfolgt. Diese Optimierungskomponente ist Bestandteil eines verteilten „Living Lab“, das unter anderem dazu dient zu untersuchen, welchen Einfluss dynamische Tarife auf das Nutzerverhalten in verschiedenen Kontexten und Anwendungsszenarien haben.

1 Einleitung und Motivation

Die informations- und kommunikationstechnische Integration dezentraler Energieanlagen in einem auf erneuerbaren Energien basierenden Stromnetz wird als „Smart Grid“ bezeichnet. Eine starke Nutzung erneuerbarer, volatiler Energiequellen führt zu einem Paradigmenwechsel in der Energieversorgung. Während die bisherige Versorgung nachfrageorientiert war, sind neue Mechanismen gefragt, die den Energieverbrauch an der schwankenden Erzeugung ausrichten. Das heißt, den idealerweise reduzierten Energieverbrauch der klimafreundlichen Erzeugung so anzupassen, dass möglichst viel der verfügbaren Wind- und Sonnen- oder Biomassenenergie genutzt werden kann. Das Potenzial zur Anpassung des Energieverbrauchs an die Erzeugung, durch sogenannte „verschiebbare Lasten“ ist laut einer VDE-Studie [ETG12] durchschnittlich 8,5 GW (Stand 2011), was einem jährlichen Energiebedarf von 74,5 TWh entspricht und damit ca. 12 % des Gesamtstrombedarfs ausmacht.[1] Um dieses Potenzial zu nutzen, müssen die Stromkunden allerdings in irgendeiner Form davon Kenntnis erhalten, zu welchen Zeiten es günstig ist, den Strom zu verbrauchen – nämlich dann, wenn viel erneuerbare Energie zur Verfügung steht – und zu welchen Zeiten eher nicht. Dabei ist „günstig“ durchaus auch im wörtlichen Sinn von „preisgünstig“

[1] laut BDEW betrug der Bruttostromverbrauch 602,6 TWh in 2011

zu verstehen, denn niedrige Strompreise zu bestimmten Zeiten stellen für Stromkunden attraktive Anreize dar, um ihr Stromnutzungsverhalten entsprechend anzupassen.

Setzt man die flächendeckende Einführung von last- und zeitvariablen Tarifen und einer entsprechenden Mess-Infrastruktur zur Abrechnung (sog. „Smart Meter“) voraus, bedarf es jedoch weiterer informationstechnischer Unterstützung bei der kostengünstigen Verschiebung von Lasten: Es ist nicht zu erwarten, dass ein Stromkunde eigenständig ein Lastmanagement betreibt und seine elektrischen Verbraucher (Geräte, Maschinen, aber auch Elektrofahrzeuge etc.) zu bestimmten Niedrigtarifzeiten manuell koordiniert und schaltet. Im weiteren Verlauf wird daher eine Demand-Side-Managementsoftware vorgestellt, die mittels mathematischer Optimierung und unter Berücksichtigung variabler Stromtarife und individueller Vorgaben die Nutzung auch von eigenerzeugtem Strom (Verbrauch, Speicherung, Einspeisung) kostengünstig plant. Diese Software wird in einem größeren Projektkontext eingesetzt, um den Einfluss variabler Stromtarife auf das Verbrauchsverhalten zu simulieren und zu untersuchen, ob sich z.B. „Überschwing-Effekte“ einstellen, d.h. ein zu hoher Strombedarf durch Lastverschiebung auch zu Zeiten hoher Stromverfügbarkeit. Dies soll helfen, Maßnahmen zu entwickeln, um solchen Effekten entgegenzuwirken, z.B. durch Anheben des Strompreises an bestimmten Leistungsgrenzen, z.B. durch sogenannte *Inclining Block Rates* (IBR) (vgl. [MRLG10]).

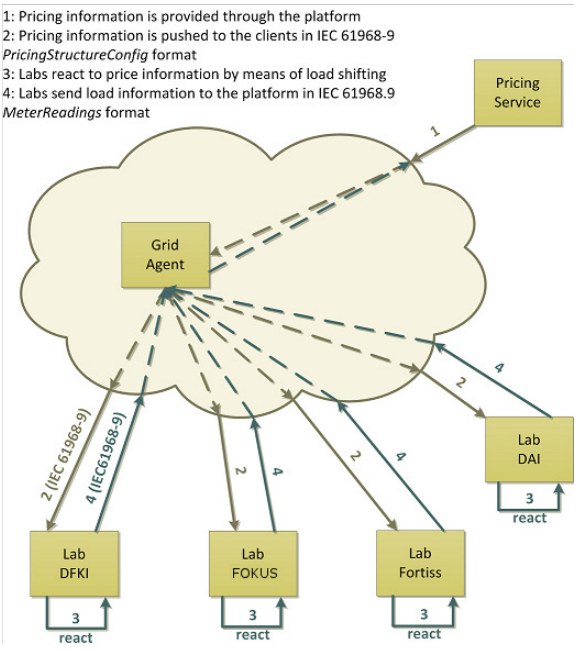


Abbildung 1: Architektur des „Open SES Experience Labs”

Bei dem Projektkontext handelt es sich um eine „Experience Lab“-Aktivität des European Institute of Technology (EIT) in der „Action Line Smart Energy Systems (SES)“,

nämlich um Aktivität 11831 „Open SES Experience Labs“. Dort wird eine dezentrale Infrastruktur entwickelt, um verteilte Energieexperimente durchführen zu können. Die Infrastruktur erlaubt es, verschiedene Laboratorien und Simulationsplattformen der Partner zu verbinden, um mit definierten Szenarien die Wechselwirkung zwischen im Markt aus Angebot und Bedarf ermittelten Preisen und lokalen (in den angeschlossenen Laboratorien der Partner) vorliegenden Produktions- und Verbrauchsprofilen abzubilden und zu erforschen. Eine Übersicht über die Struktur des Systems findet sich in Abbildung 1. Die lokalen Komponenten der Partner realisieren dabei verschiedene Anwendungsfälle:

- Reale „Living Labs“ mit Produktions- und Verbrauchskomponenten (Photovoltaik, IT-Geräte, Beleuchtung usw.) und Sensoren zur Messung von aktuellem Energieverbrauch und -produktion,
- Ladestationen für Elektroautos,
- Simulationskomponenten zur Vorhersage der Eigenproduktion aus regenerativen Energiequellen, zur Speicherung von Energie in Batterien und zur Planung des Verbrauchs im Büro und im Haushalt.

Die im Folgenden beschriebene Komponente gehört zur letzten Kategorie.

Das verteilte System aus realen und simulativen Komponenten erlaubt die Definition vielfältiger Szenarien zum Studium des Verhaltens und der Wechselwirkungen dieses komplexen Systems von Energiemarkt, Energieproduktion und Energieverbrauch. Im einfachsten Fall (s. Abb. 1) besteht das System aus verschiedenen Simulationskomponenten zur Ermittlung eines Energiepreises für jeweils die nächsten 24 Stunden, die Marktmechanismen nachbilden und dabei Energieproduktion und Energiebedarf berücksichtigen; weiterführende Information zur Ermittlung von Energiepreisen und Marktmechanismen ist in [CHL+13, LCP+14] zu finden.[2]

Diese Komponenten ermitteln ein Preisprofil, dass den angeschlossenen Partnerlaboratorien zur Verfügung gestellt wird. Diese verteilten lokalen Komponenten verfügen je nach Ausprägung über eine Vorhersage der eigenen Produktion erneuerbarer Energie, über Möglichkeiten zur Energiespeicherung und ein Profil fixer und verschiebbarer Lasten. Basierend auf den Preisinformationen und den lokalen Erfordernissen bestimmen sie ein lokales Lastprofil ebenfalls für die nächsten 24 Stunden und stellen dieses den Preissimulatoren zur Verfügung. Preisinformationen und daraus resultierende Lastprofile werden unter Nutzung des IEC 61968-9-Standards ausgetauscht. Die gegebenenfalls veränderten Lasten können zu einer bedarfsabhängigen Anpassung des Preisprofils führen, das wiederum den Klienten zur Verfügung gestellt wird. Sowohl auf Seite der Preisbestimmung als auch in den lokalen Laboratorien können nun verschiedene Kriterien zur Optimierung des Energieverbrauchs, zur Kostenreduktion, zur Erhöhung des Eigenverbrauchs usw. angesetzt werden und dadurch eine Vielzahl von Wechselwirkungen im Gesamtsystem untersucht werden. Die hier vorgestellte Demand-Side-Managementsoftware kann durch Parametrierung der beteiligten Produktions-, Speicher- und Verbrauchskomponenten eine Vielzahl verschiedener Klienten abbilden und damit ein System mit vielen Produzenten und Verbrauchern abbilden.

[2] siehe auch `http://virtualsmartgrid.project.cwi.nl/wiki/Market_Garden`

2 Verwandte Arbeiten

Während bei den eigenen Vorarbeiten [WLS13] nicht-unterbrechbare aber zeitlich verschiebbare Lastprofile mittels constraintbasierter Planung so terminiert werden, dass vordefinierte zeitliche Abhängigkeiten erfüllt und deren Kosten durch Nutzung eines zeitvariablen Strompreises minimiert werden, wird im Folgenden ein wesentlich komplexeres Stromnutzungsszenario betrachtet. Ergänzend wird nämlich neben einem fixen Verbrauchs- und Erzeugungsprofil (z.B. durch eine Photovoltaik-Anlage) auch Energiespeicherung mittels Batterien und die Möglichkeit der Netzeinspeisung berücksichtigt. Die dabei eingesetzte mathematische Modellierung ist ähnlich der in [STvdBH13] verwendeten; dort werden allerdings noch elektrische Heizungen und Kühlungen berücksichtigt und eine Optimierung unter unsicherem Verbrauchsverhalten und Strompreisen durchgeführt. Die resultierenden Optimierungsprobleme werden mit Hilfe gemischt-ganzzahliger linearer Programmierung (*Mixed Integer Linear Programming* – MILP) gelöst, die auch zur Modellierung und Lösung der verwandten, jedoch spezielleren *Unit Commitment Problems* eingesetzt wird (vgl. [SDSK13])[3].

Ein ähnliches Vorgehen ist auch bei [MRLG10] zu finden. Der Artikel beschreibt ein Verbrauchsmodell, bei dem der Energiebedarf von Verbrauchern innerhalb eines Zeitintervalls zu decken ist, wobei bei Vorgabe von dynamischen lastabhängigen Preisstufen nach Kosten und nach Verzögerung optimiert wird. Das Modell sieht allerdings weder gerätenoch anlagenspezifische Verbrauchsprofile und auch keine temporalen Abhängigkeiten zwischen den entsprechenden Aktivitäten vor. Interessant ist jedoch die Berücksichtigung von lastabhängigen Preisstufen, um insbesondere die Attraktivität von Niedrigpreisperioden bei steigender Last zu senken. Zwar wird die Nutzung von Batteriespeichern (hier von Elektrofahrzeugen) als Energiepuffer zur Einspeisung bei Spitzenlastzeiten andiskutiert, jedoch nicht im Detail ausgearbeitet. Experimentelle Untersuchungen zeigen auf, dass durch die Optimierung z.T. signifikant Kosten eingespart werden können und insbesondere lastabhängige Preisabstufungen helfen, Lastsynchronisationen zu vermeiden (da zuviel Last „bestraft“ wird).

In [LKK12] wird ein auf Tiefensuche basierender Scheduling-Ansatz beschrieben, um sowohl nicht-unterbrechbare als auch unterbrechbare Lastprofile so zu planen, dass die Lastspitze über einen diskreten Zeitraum minimiert wird, allerdings werden zeitvariable Preise dabei nicht berücksichtigt. Zur Optimierung wird eine Art *Branch-and-Bound*-Ansatz verfolgt, so dass Suchzweige nicht weiter verfolgt werden, wenn die aktuell beste Lastspitze überschritten wird. Experimentelle Ergebnisse zeigen, dass das genutzte Suchverfahren ohne zusätzliche Suchraumeinschränkung für größere Probleme aufgrund langer Laufzeiten nicht praktikabel einsetzbar ist.

Die Notwendigkeit zur Erstellung von geräte- und betriebsspezifischen Lastprofilen, die Grundlage für ein Demand-Side-Management (DSM) sind, wird in [PKRT14] motiviert. Des weiteren wird auf die DSM-Potenziale verschiedener Geräteklassen eingegangen, hinsichtlich Unterbrechbarkeit und Verschiebbarkeit. Größtes Potenzial haben im Haushalt Wäschetrockner, da sie periodisch einen hohen Verbrauch haben, ihre Nutzung lange ver-

[3] Hier werden neben Erzeugern auch Verbraucher und Speicher betrachtet.

schiebbar ist und die Heizung bis zu 30 Minuten unterbrechbar ist. Danach kommen Klimaanlagen, da Leistung und Betriebsdauer bei einem hohem Strombedarf variabel sind, sowie Geschirrspüler, da diese bei einem relativ hohen Stromverbrauch ein großes Verschiebepotenzial haben.

3 Die Optimierungskomponente

Die Optimierungskomponente unterstützt die Modellierung verschiedener, teils steuerbarer Stromerzeuger, -verbraucher und -speicher in einem „Micro-Grid", das an ein Verteilnetz angeschlossen ist und von dort Strom beziehen aber auch in dieses einspeisen kann. Damit lassen sich sowohl Wohn- und Geschäftsgebäude aber auch Gewerbebetriebe modellieren. Ziel der Optimierung ist es, unter der Vorgabe eines dynamischen Strompreises (*Real Time Price* - RTP) und unter Nutzung von Freiheitsgraden, die Stromnutzung kostenminimal zu steuern. Dabei wird davon ausgegangen, dass die Stromerzeugung durch erneuerbare Energien (z.B. durch eine Photovoltaik-Anlage oder ein Windrad) erfolgt und wie ein Teil des Stromverbrauchs zwar vorhersagbar aber nicht beeinflusst wird. Die Möglichkeit der Optimierung ist jedoch durch zeitliche Verschiebbarkeit von Lasten und die Nutzung des eigenerzeugten Stroms und des Netzes als Stromquelle und -senke sowie das Laden und Entladen von Batteriespeichern gegeben. Die Optimierung unterstützt das Energiemanagement im „Micro-Grid" dadurch, dass sie bestimmt, welche Anteile von Stromüberschüssen wie zu verbrauchen, zwischenzuspeichern oder ins Netz einzuspeisen sind und welche Anteile von Strombedarfen aus dem Netz oder aus der Batterie gedeckt werden, so dass die energetischen Randbedingungen erfüllt sind und dies möglichst kostengünstig ist.

3.1 Modellierung

Es wird angenommen, dass ein dynamischer Strompreis $r : H \rightarrow \mathbb{Q}$ für ein festes Zeitintervall (den *Planungshorizont*) $H = [u, w)$ mit $u, w \in \mathbb{Q}$ gegeben und stückweise konstant ist, d.h. dass es äquidistante diskrete Zeitschritte[4] $u = t_1 < t_2, \ldots < t_n = w$ mit $t_i \in \mathbb{Q}$ für $i = 1, \ldots, n$ gibt, so dass gelte

$$\exists \Delta t \in \mathbb{Q} \forall i \in \{1, \ldots, n-1\} \quad : \quad \Delta t = t_{i+1} - t_i \tag{1}$$

$$\forall i \in \{1, \ldots, n-1\} \forall t \in [t_i, t_{i+1}) \exists r_i \in \mathbb{Q} \quad : \quad r(t) = r_i \ . \tag{2}$$

Es wird weiterhin davon ausgegangen, dass die Steuerung zur Minimierung der Stromkosten zu den Zeitschritten $t_1, \ldots t_{n-1}$ für die jeweiligen Zeitintervalle $[t_i, t_{i+1})$ erfolgt und jede Stromquelle oder -senke (Erzeuger, Verbraucher, Speicher, Netz) in jedem Zeitintervall $[t_i, t_{i+1})$ des Planungshorizontes eine durchschnittliche Leistungsabgabe oder

[4] Die Äquidistanz ist o.B.d.A. durch die geeignete Granularität aufgrund der Diskretisierung immer herstellbar.

-aufnahme p_i hat, wobei das Vorzeichen Auskunft über Abgabe bzw. Aufnahme gibt.[5] Dabei sei $(p)_+ = |p|$, falls $p > 0$ ist und $(p)_+ = 0$ ansonsten. In Analogie sei $(p)_- = |p|$, falls $p < 0$ ist und $(p)_- = 0$ ansonsten, wobei $p \in \mathbb{Q}$ gelte (vgl. [STvdBH13]).

Wesentlich für die Optimierung ist, dass die Leistungen innerhalb des modellierten Micro-Grids in jedem Zeitintervall ausgeglichen sind, d.h. wenn S die Menge aller Stromquellen und -senken ist, dann soll gelten:

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad \sum_{s \in S} p_i^s = 0 \ . \tag{3}$$

Ferner ist sowohl die Summe der Stromzuflüsse als auch der -abflüsse zu jedem Zeitraum entsprechend der Leistungskapazität der lokalen Infrastruktur (Leitungen, Schalter, Sicherungen, etc.) beschränkt:

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad \sum_{s \in S} (p_i)_+^s \leq C \wedge \sum_{s \in S} (p_i)_-^s \leq C \ , \tag{4}$$

wobei C eine fest gegebene Leistungsgrenze sei.

3.1.1 Netzbezug und -einspeisung

Das Hauptziel der Optimierung ist es, die Kosten für den Bezug von Strom aus dem Netz zu minimieren, wobei eine Einspeisung ins Stromnetz dabei zusätzlich kostensenkend wirkt. Im Gegensatz zu einem variablen Bezugspreis wird gegenwärtig von einer festen Vergütung v innerhalb des betrachteten Planungshorizonts ausgegangen. Eine Erweiterung des Modells um variable Einspeisevergütungen wäre einfach umsetzbar.[6] Durch die Optimierung ist folglich ein Bezugs-/Einspeiseprofil $p_1, \ldots, p_{n-1} \in \mathbb{Q}$ zu bestimmen, so dass der Zielfunktionswert

$$\mathsf{cost} = \sum_{i=1}^{n-1} (r_i (p_i)_+ - v (p_i)_-) \quad \rightarrow \quad \min! \tag{5}$$

minimiert wird, wobei negative Leistungswerte Einspeisungen und positive Werte Bezüge darstellen.

3.1.2 Eigenerzeugung und fester Verbrauch

Eigenstromerzeugung mittels Photovoltaik-Anlagen oder Windrädern wird als vorhersagbar aber in dem betrachteten Kontext als nicht beeinflussbar angenommen. Gleiches gilt für einen Teil des Stromverbrauchs, der meist ad-hoc und abhängig vom aktuellen Bedarf erfolgt. Darunter fallen Aktivitäten wir Kochen, Duschen, Fernsehen aber auch Beleuchtung. Dies resultiert in Summe in einem bei der Optimierung fest vorgegebenen

[5] Die genaue Bedeutung wird im weiteren Verlauf konkretisiert.
[6] in Analogie zu den zeitvariablen Strompreisen.

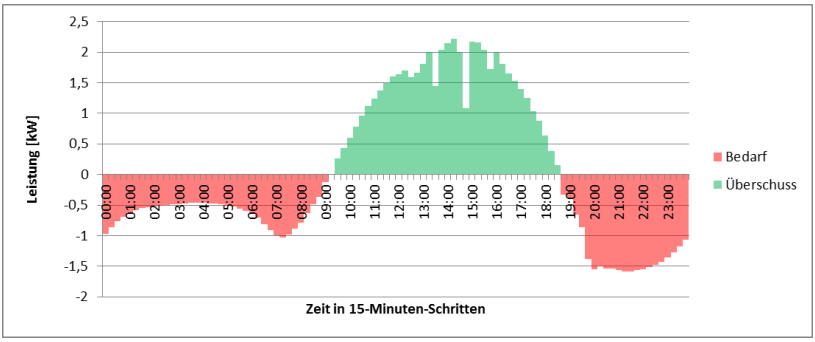


Abbildung 2: Bedarf-Überschuss-Profil eines Einfamilienhauses mit PV-Anlage im Sommer

Leistungsprofil $p_1, \ldots, p_{n-1} \in \mathbb{Q}$, wobei negative Leistungswerte einen entsprechenden Strombedarf und positive Werte einen Stromüberschuss widerspiegeln. Abbildung 2 zeigt ein typisches Leistungsprofil für einen sonnigen Sommertag mit einer zeitlichen Auflösung von 15 Minuten. Während am Morgen, am Abend und in der Nacht der nichtverschiebbare Strombedarf nicht durch die PV-Anlage gedeckt wird, besteht über den Tag ein Stromüberschuss mit einer Leistungsspitze am frühen Nachmittag.

Ist durch regulatorische Vorgaben gefordert, dass eine Netzeinspeisung nur durch Eigenproduktion erfolgen darf (und z.B. nicht aus einem Batteriespeicher), kann man dieser Forderung durch die Bedingungen

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad (p_i^{\mathsf{grid}})_- \leq (p_i^{\mathsf{own}})_+ \tag{6}$$

nachkommen, wobei $(p_i^{\mathsf{grid}})_-$ die Einspeisung im Zeitfenster i ist und $(p_i^{\mathsf{own}})_+$ die Eigenproduktion zur selben Zeit darstellt.

3.1.3 Batterie

Eine Batterie hat idealerweise einen Ladezustand $E \in [E_{\mathsf{dod}}, E_{\max}] \subseteq \mathbb{Q}$, wobei $E_{\mathsf{dod}} > 0$ eine Entlade-Untergrenze (*depth-of-discharge*) ist, die nicht unterschritten werden sollte und $E_{\max} > E_{\mathsf{dod}}$ ihre maximale Ladekapazität ist. Auch die maximale Lade- und Entladeleistung einer Batterie ist batteriespezifisch gegeben: $p_i \in [p_{\mathsf{charge}}, p_{\mathsf{discharge}}]$ für $i = 1, \ldots, n-1$, wobei eine negative Leistung ein Laden bedeutet und eine positive Leistung ein Entladen. Weiterhin hat jede Batterie einen Wirkungsgrad $\eta \in [0, 1)$, so dass sich der Ladezustand einer Batterie zum Zeitpunkt t_{i+1} aus dem Ladezustand zum Zeitpunkt t_i wie folgt ergibt:

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad E_{i+1} = E_i + \Delta t(\eta(p_i)_- - (p_i)_+) \ . \tag{7}$$

Der Einfachheit halber wird die Selbstentladung einer Batterie nicht berücksichtigt, da sie im Allgemeinen pro Tag nur wenige Prozent beträgt und bei einer rollierenden Tagespla-

nung Soll-Ist-Abweichungen durch Messwerte aus dem Batteriemonitoring ausgeglichen werden können.

3.1.4 Verschiebbare Lasten

Eine zeitlich verschiebbare Aktivität („Last“) mit der Dauer m sei durch ihr Lastprofil $l_1, \ldots, l_m \in \mathbb{Q}^+$ gegeben, das innerhalb eines Zeitfenster $[t_{\mathsf{est}}, t_{\mathsf{lct}})$ in diskreten Schritten verschiebbar ist, wobei t_{est} ihre früheste Startzeit und t_{lct} ihr spätestes Ende ist. Es wird dabei angenommen, dass die Last nicht-unterbrechbar ist und zu den Zeitpunkten t_s mit $s = \mathsf{est}, \ldots, \mathsf{lct} - m$ starten kann. Der Strombedarf einer verschiebbaren Last über dem Planungshorizont ist dann durch folgende Bedingung

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad p_i = \sum_{j=1}^{m} l_j \cdot x_{\max(0, i-j+1)} \tag{8}$$

definiert, wobei die Booleschen Schaltvariablen $x_0, \ldots, x_{n-1} \in \{0, 1\}$ den Start der verschiebbaren Last bestimmen, d.h. $x_i = 1$ genau dann gilt, wenn die Last zum Zeitpunkt t_i startet. Folglich gilt $x_i = 0$ für $i = 0, \ldots, \mathsf{est} - 1$ und $i = \mathsf{lct} - m + 1, \ldots, n - 1$ sowie

$$\sum_{i=1}^{n-1} x_i = 1 \quad \text{und} \quad s = \sum_{i=1}^{n-1} i \cdot x_i \ \text{, wobei } t_s \text{ die Startzeit der Last ist.} \tag{9}$$

Sollen sich mehrere Aktivitäten $a \in A$ zeitlich wechselseitig ausschließen, da diese z.B. eine gemeinsame Ressource (z.B. eine bestimmte Maschine oder Arbeitskraft) benötigen, wird dies durch die Ungleichung

$$\forall i \in \{1, \ldots, n-1\} \quad : \quad \sum_{a \in A} \sum_{j=1}^{m^a} : x^a_{\max(0, i-j+1)} \leq 1 \tag{10}$$

ausgedrückt. Diese besagt, dass zu jedem Zeitschritt höchstens eine der Aktivitäten aktiv ist (vgl. [Hoo12], S. 507, (7.199)).

Sollen zwei verschiedene Aktivitäten a und b zeitlich aufeinander folgen (z.B. Schweißen vor Lackieren oder Waschen vor Trocknen), so lässt sich dies mit Hilfe der Ungleichung $s^a + m^a \leq s^b$ beschreiben. Sollen zwischen dem Ende einer Aktivität a und dem Start einer anderen Aktivität b höchstens k Zeitschritte (z.B. zwischen Waschen und Trocken) vergehen, dann ist dies durch die Ungleichung $s^a + m^a + k \geq s^b$ darstellbar; sollen es mindestens k Zeitschritte sein (z.B. zwischen Lackieren und Montage), dann durch die Ungleichung $s^a + m^a + k \leq s^b$.

3.2 Implementierung

Zur Erstellung und Lösung individueller Optimierungsprobleme, wie in Abschnitt 3.1 beschrieben, wurde in der objektorientierten Programmiersprache JAVA[7] eine Software-Komponente mit einer entsprechenden API (*application programming interface*) realisiert. Neben Methoden zur Festlegung der Größe der Zeitschritte (in Minuten), der variablen Preise (pro kWh) und des Erzeugungs-/Verbrauchsprofils (in kW) gibt es Methoden zur Definition von Batterien und ihren Charakteristika sowie zur Spezifikation von verschiebbaren Lasten sowie den zeitlichen Abhängigkeiten zwischen ihnen. Die entsprechenden linearen Gleichungen und Ungleichungen werden in die Sprache CMPL (*Coin Mathematical Programming Language*)[8] übersetzt und mit Hilfe des MILP-Solvers Cbc[9] und der verfügbaren JAVA-API jCMPL gelöst.

Eine fortlaufende Planung über mehrere aneinandergrenzender Planungshorizonte wird dadurch erreicht, dass der Batterieladestand am Ende eines Horizontes den Initialwert zu Beginn des nächsten definiert. Der Teil eines Lastprofils einer geplanten Last, der von einem Horizont in den nächsten „hineinragt“, wird bei der Planung des nächsten Horizonts als Last mit fester Startzeit zu dessen Beginn betrachtet. Zu berücksichtigende zeitliche Abhängigkeiten mit anderen noch zu planenden verschiebbaren Lasten (ggf. aus dem vorherigen Planungshorizont) sind entsprechend anzupassen. Um ein solches „Hineinragen“ zu ermöglichen, werden die Horizonte bei der Planung um die größte Länge der zu planenden Lasten verlängert.

4 Experimentelle Ergebnisse

Die implementierte Optimierungskomponente wurde hinsichtlich ihres Laufzeitverhaltens mit Hilfe eines typischen Eigenheimszenarios evaluiert. Der Planungshorizont ist ein Sommertag (24 h) mit einer zeitlichen Auflösung von 15 Minuten. Für diesen Tag sind neben den variablen Preisen ein Bedarf-/Überschuss-Profil gegeben, welches für ein Einfamilienhaus mit PV-Anlage typisch ist (vgl. Abb. 2). Ferner verfügt das Einfamilienhaus in diesem Szenario über eine Batterie mit einer nutzbaren Kapazität von 4 KWh, die zu Beginn des Planungshorizontes über eine nutzbare Kapazität von 1,3 KWh verfügt. Des weiteren sind folgende verschiebbare Lasten gegeben, die ggf. auf bestimmte Zeitintervalle beschränkt sein können:

- Waschen: Schonwaschgang (sensitiv) und Vollwaschgang (intensiv),
- Trocknen: mit niedriger und hoher Temperatur (sensitiv/intensiv),
- Geschirrspülen: mit niedriger und hoher Temperatur (sensitiv/intensiv),
- Bügeln und Reinigen (Staubsaugen).

[7] http://www.java.com/de/
[8] https://projects.coin-or.org/Cmpl
[9] https://projects.coin-or.org/Cbc

Sowohl die beiden Wasch- und Trocknen- und Spülaktivitäten als auch Bügeln und Reinigen schließen sich wechselseitig aus (jeweils nur eine Maschine bzw. Person pro Aktivität verfügbar) und das Trocknen folgt auf das Waschen. Diese zeitlichen Randbedingungen sind frei definierbar; so kann im einem Mehrpersonenhaushalt der wechselseitige Ausschluss von Bügeln und Reinigen entfallen.

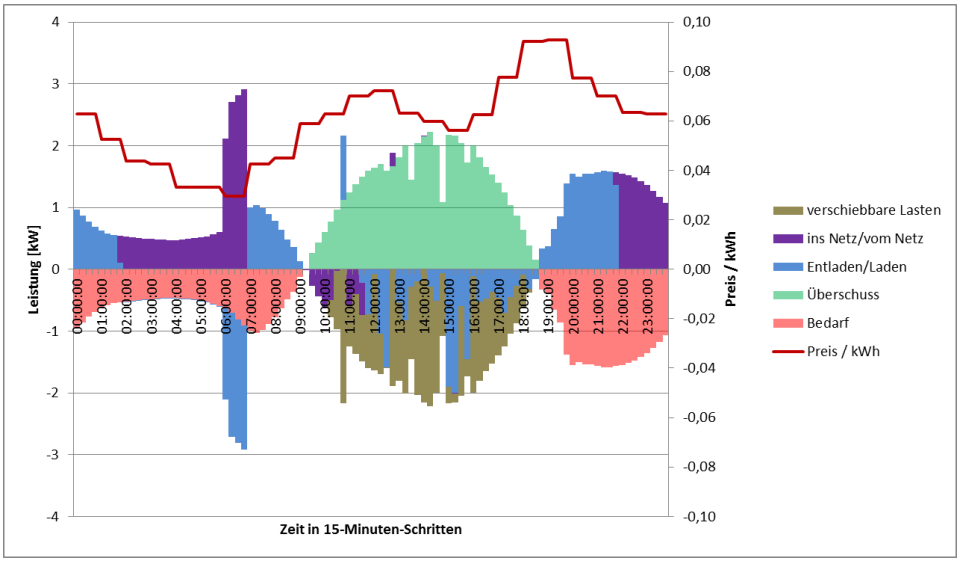


Abbildung 3: Kostenoptimierte Nutzung von Batteriespeicher mit Lastverschiebung

Abbildung 3 zeigt nun das Ergebnis der Kostenoptimierung in diesem Szenario, welches in ca. 4 Sekunden auf einem modernen Windows-PC (Intel i7, 8 GB, Windows 7) berechnet wird. Die Achsensymmetrie des Diagramms zeigt, dass die gefundene Lösung energetisch ausgeglichen ist - d.h. die Summe von Strombedarf und -verfügbarkeit gleich Null ist. Zudem ist deutlich zu erkennen, dass die Optimierung die niedrigen Strompreise in den Morgenstunden nutzt, um die zuvor in der Nacht entladene Batterie voll aufzuladen. Insbesondere wird dabei das Zeitintervall von 6:00 h bis 7:00 h intensiv genutzt, die Batterie mit preiswerter Energie aus dem Netz zu laden. Des weiteren sieht man, dass der über den Tag eigenerzeugte Strom nahezu komplett genutzt wird, um die verschiebbaren Lasten zu decken oder die Batterie nachzuladen. Dadurch entsteht kurzfristig eine Lastspitze zwischen 10:45 h und 11:00 h, die sich aufgrund der gewählten Zielfunktion (reine Kostenreduzierung) ergibt. Ansonsten wird durch die Kostenminimierung nur ein geringer Teil des eigenerzeugten Stroms über den Tagesverlauf ins Netz eingespeist. In den frühen Abendstunden und bei hohem Strompreis wird der in der Batterie gespeicherte Strom genutzt. Erst in den späten Abendstunden, wenn die Batterie entladen ist, wird Strom zu einem günstigeren Preis aus dem Netz bezogen.

5 Integration in das „Living Lab“

Jedem „Living Lab“ wird zur einheitlichen Kommunikation über ein abgestimmtes Protokoll eine Java-API durch eine separate Programmierbibliothek zur Verfügung gestellt, die einen XML-basierten standardisierten Datenaustausch von zeitvariablen Strompreisen und Lastprofilen (eigentlich sind es Smart-Meter-Daten) ermöglicht. Diese Entwicklung bietet für alle Szenarien eine vereinfachte und standardisierte Schnittstellen zu den anderen Labs an. Die zur gemeinsamen Kommunikation realisierten Schnittstellen setzen sich aus mehreren Java-XML-Webservices zusammen. Über diese ist es möglich, JAXB-Objekte (*Java XML Bindings*) zu übertragen, die sich mittels der Programmierbibliothek erzeugen, bearbeiten, empfangen und versenden lassen. Die Implementierung setzt dabei den bereits genannten IEC-Standard 61968-9 für *Meter Reading & Control* um. Die zugehörigen JAXB-Klassen und XML-Schemata entsprechen diesem Standard.

5.1 Durchführung verteilter Experimente

Die verteilten Labs und Simulationsmodelle der Partner können, wie in Abbildung 1 dargestellt, miteinander verbunden werden – die lokalen Komponenten erhalten Preisinformationen und publizieren ihre Verbrauchsprofile. Das verteilte System kann verschiedene Experimente realisieren: in Realzeit – wobei die Messungen in den Labs mit einbezogen werden – aber auch im Zeitraffer, wobei historische Messwerte aus den Labs ausgewertet werden können. Ein Preisdienst stellt gemäß dem IEC-Standard 61968-9 ein Preisprofil für die nächsten 24 Stunden bereit. Die lokalen Komponenten erhalten dieses Profil in dem vom Standard definierten Format. Nach einer möglichen Umplanung der eigenen Verbrauchsprofile und einer eventuellen Korrektur der Vorhersage für die Eigenproduktion aus erneuerbaren Energiequellen liefern sie ein adaptiertes Verbrauchsprofil für die nächsten 24 Stunden im IEC-standardisierten Format an das System zurück.

6 Zusammenfassung und Ausblick

In dieser Arbeit wurde eine Demand-Side-Management-Software sowie deren Integration in ein „Living Lab“ vorgestellt, um u.a. die Auswirkungen von variablen Strompreisen auf das Nutzerverhalten von „Prosumern“ zu analysieren. Die Software zur Eigenverbrauchsoptimierung und zur Stromkostensenkung basiert auf einem linearen Optimierungsmodell, das mit Hilfe eines MILP-Solvers gelöst wird. Es ist vorgesehen, in zukünftigen Erweiterungen der Software neben zeitvariablen auch lastabhängige Stromkosten, sogenannte *Inclining Block Rates* (IBR) zu berücksichtigen; in [MRLG10]wird gezeigt, wie dies linear zu modellieren ist. Zukünftig sollen auch Batteriekosten und berücksichtigt werden – sowohl anteilig unter Berücksichtigung von Lebensdauer (vgl. [STvdBH13]) und deren möglichen Verlängerung (vgl. [TP14]). Des weiteren sollen bei der erweiterten Optimierung auch übergeordnete Kriterien wie eine netzverträgliche Einspeisung berücksichtigt

werden. Ferner wird eine alternative constraint-basierte Realisierung der Optimierungskomponente angestrebt, die auf speziellen intervallbasierten Verfahren zur Suchraumeinschränkung beruht und so leistungsfähig ist, dass sie auch in einem eingebetteten Hardwaresystem performant lauffähig ist und in der Praxis eingesetzt werden kann.

Literatur

[CHL+13] Felix Claessen, Nicolas Höning, Bart Liefers, Han La Poutré und Peter A. N. Bosman. Market Garden: A Scalable Research Environment for Heterogeneous Electricity Markets. *ERCIM News*, 2013(92), 2013.

[ETG12] ETG-Task Force Demand Side Management. Demand Side Integration – Lastverschiebungspotenziale in Deutschland. Bericht, VDE, 2012.

[Hoo12] John Hooker. *Integrated Methods for Optimization*, Jgg. 170 of *International Series in Operations Research & Management Science*. Springer Verlag, 2. Auflage, 2012.

[LCP+14] Bart Liefers, Felix Claessen, Eric J. Pauwels, Peter A. N. Bosman und Han La Poutré. Market Garden: A Simulation Environment for Research and User Experience in Smart Grids. In Yves Demazeau, Franco Zambonelli, Juan M. Corchado und Javier Bajo, Hrsg., *Advances in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection - 12th International Conference, PAAMS 2014, Salamanca, Spain, June 4-6, 2014. Proceedings*, Jgg. 8473 of *Lecture Notes in Computer Science*, Seiten 351–354. Springer, 2014.

[LKK12] Junghoon Lee, Hye-Jin Kim und Gyung-Leen Park Mikyung Kang. Energy Consumption Scheduler for Demand Response Systems in the Smart Grid. *Journal of Information Science and Engineering*, 28(5):955–969, 2012.

[MRLG10] Amir-Hamed Mohsenian-Rad und Alberto Leon-Garcia. Optimal Residential Load Control with Price Prediction in Real-Rime Electricity Pricing Environments. *IEEE Transactions on Smart Grid*, 1(2):120–133, September 2010.

[PKRT14] Manisa Pipattanasomporn, Murat Kuzlu, Saifur Rahman und Yonael Teklu. Load Profiles of Selected Major Household Appliances and Their Demand Response Opportunities. *IEEE Transactions on Smart Grid*, 5(2):742–750, March 2014.

[SDSK13] B. Saravanan, Siddharth Das, Surbhi Sikri und D. P. Kothari. A solution to the unit commitment problema review. *Frontiers in Energy*, 7(2):223–236, June 2013.

[STvdBH13] Paul Scott, Sylvie Thiébaux, Menkes van den Briel und Pascal Van Hentenryck. Residential Demand Response under Uncertainty. In Christian Schulte, Hrsg., *Principles and Practice of Constraint Programming - 19th International Conference, CP 2013, Uppsala, Sweden, September 16-20, 2013. Proceedings*, Jgg. 8124 of *Lecture Notes in Computer Science*, Seiten 645–660. Springer Verlag, 2013.

[TP14] Malte Thomann und Florin Popescu. Estimating the effect of domestic load and renewable supply variability on battery capacity requirements for decentralized microgrids. In *The 4th International Conference on Sustainable Energy Information Technology (SEIT-2014)*, Jgg. 32 of *Procedia Computer Science*, Seiten 715–722. Elsevier, 2014.

[WLS13] Armin Wolf, Thomas Luckenbach und Mario Schuster. Smart Demand-Side Response at Home. *ERCIM News*, 2013(92), 2013.

Beyond Data: Mastering Management of Complex Research Data and Processing Environments

(BeyondData14)

Beyond Data: Mastering Management of Complex Research Data and Processing Environments (BeyondData2014)*

Management of research data is becoming a crucial requirement for research- and memory-institutions as well as commercial R&D. In order to foster scientific innovation and simultaneously reduce redundant spending on data generation, efficient access to research results as well as their fundamental data is indispensable. Re-using research data is not only economically worthwhile, but convenient access to research data may also spur new research ideas and interdisciplinary innovation. Furthermore, with the rise of networked functional services, e.g. Cloud offerings such as software-as-a-service and dedicated data services (Big Data), a process-oriented, holistic approach to research data management becomes more and more important.

The full-day workshop „Beyond Data: Mastering Management of Complex Research Data and Processing Environments“ covers current and upcoming challenges faced by memory institutions but also by research institutes and industry. The information age has changed the economic landscape and research environments thoroughly. Nearly every piece of valuable scientific or business information is created, transferred and stored digitally. Keeping digital research data available and usable has been proven to be more difficult than anticipated [VAA+14].

Moreover, information is not necessarily bound to single, easy identifiable digital objects but can be embedded in complex processes, which may include multiple interconnected machines and rely on external data service. For digital research data (respectively their technical data format) together with necessary processing environments even ten years are a huge timespan when trying to keep them accessible and usable as they are threatened by constantly changing technologies and the associated risk of hardware and software obsolescence. Even slight changes in configuration or soft/hardware environments may alter results significantly [GHJ+12].

References

[GHJ+12] H.B.M. Gronenschild, P. Habets, H.I.L. Jacobs, R. Menglers, N. Rozendaal, J. van Os, and M. Marcelis. The effects of FreeSurfer version, workstation type, and Macintosh

*Beyond Data: Mastering Management of Complex Research Data and Processing Environments (2014), http://beyonddata2014.wordpress.com [last access: 21.07.2014]

operating system version on anatomical volume and cortical thickness measurements. *PLoS One*, 7(6), 2012.

[VAA+14] Timothy H. Vines, Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. The availability of research data declines rapidly with article age. *Current Biology*, 24(1):94–97, 2014.

Organisation: Dr. Klaus Rechert (Universität Freiburg), Dr. Dirk von Suchodoletz, Dennis Wehrle (Universität Freiburg), Prof. Dr. Achim Streit (Karlsruher Institut für Technologie), Prof. Dr. Andreas Rauber (TU Wien), Dr. Pia Schmücker (Universität Ulm), Franck Borel (Universitätsbibliothek Freiburg)

Operating The HPI Future SOC Lab - An Infrastructure Laboratory For Researchers - A Practitioners' Report

Rami Akkad, Henning Schmitz, André Deienno Pansani

Innovation Center
SAP AG
Konrad-Zuse-Ring 10
14469 Potsdam
rami.akkad@sap.com
henning.schmitz@sap.com
andre.pansani@sap.com

Abstract: The Hasso-Plattner Institute (HPI) Future Service Oriented Computing (SOC) Lab is a research institution of the HPI in Potsdam. It aims at enabling and fostering the exchange between industry and academia. Researchers are provided with free-of-charge access to state-of-the art software and hardware. The HPI Future SOC Lab hosts events for researchers and industry representatives to create a forum in which research ideas and results are discussed. The administration and operation of the lab requires processes for data life cycle management, privacy and security concepts, data backups and restores, provisioning techniques, and data center monitoring. In this report we will give insights into how the FSOC Lab is operated and what processes are in place.

1 Introduction

The Hasso-Plattner Institute (HPI) Future Service Oriented Computing (SOC) Lab is a research institution at the HPI in Potsdam[1]. It was founded in 2010 as a collaboration of the HPI and four partners from industry; HP, EMC, Fujitsu, and SAP. The lab's overall goal is to enable and foster the exchange between industry and the research community. The HPI Future SOC Lab provides researchers with free-of-charge access to an infrastructure of state-of-the art software and hardware. This infrastructure is funded by the collaboration partners and operated at the HPI. Some highlights of the infrastructure are a 1000 core compute cluster, multi-core servers with up to 64 cores and up to 2TB RAM, GPU systems, a cloud computing infrastructure with 32 blades, and SAP software products. SAP sponsors three full-time employees and a student position at the lab, while the HPI funds one full-time employee, a student position, and the operational costs of the lab.

[1] HPI Future SOC Lab Homepage, http://hpi.de/en/research/future-soc-lab.html [last access 12.06.2014]

The Future SOC Lab also organizes and hosts events for researchers and representatives from industry in which research ideas and results can be discussed. These events include scientific workshops on questions regarding in-memory computing and cloud operations, regular HPI Future SOC Lab Days (see **2.2**), and a workshop on Big Data Benchmarking[2]. Main focus areas of the research activities at the HPI Future SOC Lab are defined by the partners of the lab and currently include in-memory computing, cloud computing, and GPU computing.

The activities at the HPI Future SOC Lab provide the partners from industry with insights into latest research trends, contacts to researchers, and a forum to position their own research questions in academia. Researchers get the infrastructure means to conduct their research while getting a chance to network with other researchers and potential partners from industry.

1.2 How The HPI Future SOC Lab Is Different To Other Infrastructure Labs

There a number of institutions worldwide that provide infrastructure services to researchers. Some of them are mentioned in the following with a focus on a highlighting one or more differences to the HPI Future SOC Lab.

The HPI Future SOC Lab's compute power can be considered small compared with what the Leibniz Supercomputing Centre offers with its SUPERMUC and other systems[3]. Hence it is better suited to support research in high performance computing than the HPI Future SOC Lab. The infrastructure of the HPI Future SOC Lab is located at the HPI in Potsdam and not distributed such as the infrastructure of the FutureGrid Project, which is a collaborative effort of many institutions[4]. The HPI Future SOC Lab is free-of-charge for researchers from any institution, which differentiates it from labs such as the GWDG in Göttingen, Germany[5], or the Konrad-Zuse-Zentrum in Berlin, Germany[6], which provide free services to certain institutions only. Providing datasets and workloads from industry is currently not in scope of the HPI Future SOC Lab either. By contrast, the Smart Data Innovation Lab with its focus on big data research intends to offer exactly such data sets[7].

2 The HPI Future SOC Lab Periods

The work of the HPI Future SOC Lab is organized in periods of which two are conducted per year. The following chapter provides details about these periods, which comprise a proposal submission period, the HPI Future SOC Lab day as the main event of a Future SOC Lab period, and the project phase, in which projects are conducted.

[2] Future SOC Lab Event Overview, http://hpi.de/en/research/future-soc-lab/events.html [last access 12.06.2014]
[3] Leibniz Supercomputing Centre Homepage, http://www.lrz.de/services/compute [last access 12.06.2014]
[4] Futuregrid Homepage, https://portal.futuregrid.org/about [last access 12.06.2014]
[5] GWDG Homepage, http://www.gwdg.de/index.php?id=wir_ueber_uns [last access 12.06.2014]
[6] ZIB Homepage, http://www.zib.de/de/institut/struktur-und-ziele-des-zib.html [last access 12.06.2014]
[7] Smart Data Innovation Lab Homepage, http://www.sdil.de/en [last access 12.06.2014]

2.1 Proposal Submission Period

Prior to the beginning of a period a call-for-proposals is published, which invites researchers to hand in project proposals to apply for infrastructure usage.

These proposals contain a description of the research purpose, some background about the researcher, the list of required infrastructure resources, the time for which access is needed, and whether this access is needed exclusively. The submission period lasts for roughly eight weeks. In exceptional cases project proposals can be submitted outside the official submission period. These are then handled in a separate process.

2.2 The HPI Future SOC Lab Days

At the end of the submission period a conference is hosted, the HPI Future SOC Lab day. This day marks the end of one and the beginning of a new Future SOC Lab period. Two Future SOC Lab Days are hosted annually, one in spring, usually at the beginning of April, and one in autumn, usually mid-October.

Representatives from the industrial partners, researchers, and guests meet to review and discuss the results of the projects on the day. Selected researchers present their results to the auditorium, while all researchers are asked to prepare posters and hand in reports summarizing their results. The reports are later published in HPI technical reports [Me10, Me11].

The HPI Future SOC Lab steering committee, which comprises representatives from the HPI and the partners from industry, approves or rejects the submitted project proposals during the Future SOC Lab Day as well.

2.3 The Future SOC Lab Period

The teams of accepted projects get access to the HPI Future SOC Lab for the following Future SOC Lab period, which is defined by the time between two subsequent Future SOC Lab Days and usually lasts six months. The principal investigator is required to accept the HPI terms of use, which includes the commitment to write a report and to create a poster by the end of the Future SOC Lab period. Afterwards the principal investigator provides the HPI Future SOC Lab team with the list of users.

For each project the required infrastructure is provisioned when needed by the project team. An average of 30 active projects and some 150 to 200 users access the infrastructure at any time. At the end of each Future SOC Lab period the users are locked. In some cases projects continue their work in subsequent Future SOC Lab periods. This requires the approval of a new project proposal but also operational consideration in terms of data and infrastructure handling, as a reuse shall be supported.

3 Technical Activities

The HPI Future SOC Lab comprises a highly heterogeneous infrastructure that is used for different purposes. To ensure a stable running environment and to simplify processes, several steps regarding resource sharing, security management, infrastructure provisioning, data- and user management, and monitoring have been implemented.

3.1 Resource Sharing

As projects require different access models to the resources the HPI Future SOC Lab offers, the lab implements two models:

1. Exclusive access to one or more requested servers / software services.

   Due to high demand exclusive access to a server is usually granted for a limited amount of time, e.g. one week or one day, on a first come first served basis. Projects need to inform the lab two weeks in advance if they require exclusive access. Access to a software service, e.g. SAP HANA, is always exclusive and granted throughout one Future SOC Lab period.

2. Shared access to one or more requested servers.

   Depending on the resource, shared access is realized through granting multiple projects access to one system without limitations (e.g. used in our GPU systems) or through limiting the host system's resources for a project with a control mechanism like Cgroups[8] (e.g. used in our 1000 core compute cluster).

The lab uses active directory groups to automate the assignment of project teams to a specific server. Each server has one active directory group and all the server's group members are allowed to access the server.

3.2 Security And Data Management

After the Steering Committee accepted a project, all project related user accounts are created with an expiration date that matches the end of the current Future SOC Lab period. To access the lab's infrastructure a user needs to connect to a virtual private network (VPN). The lab avoids public IP addresses to avert public disclosure of our servers. To minimize the possibility of a brute-force attack, a user account is automatically disabled after five failed login attempts and automatically re-enabled after one hour.

As projects may deal with sensitive data the HPI Future SOC Lab has security guidelines to prohibit unauthorized access to files. However, the lab does neither follow ISO

[8] kernel.org Cgroups specification, https://www.kernel.org/doc/Documentation/cgroups/memory.txt [last access 12.06.2014]

security policies nor does it guarantee data safety. It does not take responsibility in any case of data loss to third parties.

As described in 3.1, a project will get either exclusive or shared access to a resource. Depending on the operating system or offered service, the project is able to access the resource after successfully connecting to the lab's VPN through a protocol (e.g. SSH / RDP). If a project needs to execute root commands on a shared resource with no root access it needs to inform a lab admin who will carefully audition whether an execution of the command has any impacts on the security of the lab and who will perform the necessary operation.

To transfer huge amount of data the HPI Future SOC Lab runs an SFTP server that has a public IP address. Access to this server can be requested by projects at any time. All project related data will be deleted after the end of the Future SOC Lab period.

3.3 Infrastructure Provisioning

The lab uses different techniques to provision services, servers and virtual machines to meet the diverse requirements of projects running at the HPI Future SOC Lab.

Physical servers and virtual machines are deployed using Hewlett-Packard's private cloud solution Converged Cloud, in which pre-defined templates can be used to rapidly deploy operating systems and networks.

Most of the offered services comprise newest SAP software, such as SAP HANA. SAP HANA is currently installed on 1 TB RAM servers. Projects receive full control of their SAP HANA system. They have sufficient authorizations to perform customization and to develop, run, and debug their programs.

3.4 Backups

The HPI Future SOC Lab conducts backups on a voluntary basis and thereby does neither implement a standardized backup strategy nor does it guarantee the availability of fallback systems in case of data loss. All projects are responsible to create backups on their own. The lab does provide support in transferring data for backup reasons (e.g. when approaching the end of a Future SOC Lab period).

3.5 Monitoring

The HPI Future SOC Lab uses Icinga[9] to monitor its infrastructure to ensure that servers and services run in a stable mode and to react on hardware failures.

The lab uses built-in Nagios plugins such as check_ssh and check_root to perform monitoring operations. Additionally the lab adapted Icinga to monitor SAP HANA by

[9] Icinga Homepage, https://www.icinga.org [last access 12.06.2014]

writing a custom plugin that periodically checks that vital processes are running and that sufficient disk space is available on the relevant disks.

Icinga is also used as an internal reporting tool to generate statistics regarding load of servers and uptime of services. This helps the lab to understand problems and optimize daily operations.

4 Outlook And Conclusion

The HPI Future SOC Lab is a research institution at the HPI in Potsdam with the overall goal of enabling and fostering the exchange between industry and the research community. The HPI Future SOC Lab successfully provided over 200 [Bo11, LT11, SS14] projects from more than 40 institutes with free-of-charge access to an infrastructure of state-of-the-art software and hardware since it was founded in 2010. It hosted 12 events such as the HPI Future SOC Lab to create a hub in which industry and academia can connect.

The HPI Future SOC Lab will continue its work throughout the next years. To keep its infrastructure up to date and to attract and support cutting-edge research, the lab has to continuously update and improve its offerings and processes with the help of its partners.

This refers to both infrastructure and internal processes, particularly in regard to data and user management as well as automatic physical and virtual machine provisioning to enhance productivity and reliability.

References

[Me10] C. Meinel, A. Polze, A. Zeier, G. Oswald, D. Herzog, V. Smid, D. D'Errico, and Z. Hussain: Proceedings of the Fall 2010 Future SOC Lab Day, Technische Berichte Nr. 42 des Hasso-Plattner-Instituts für Softwaresystemtechnik an der Universität Potsdam, Potsdam, 2011.

[Me11] C. Meinel, A. Polze, G. Oswald, R. Strotman, U. Seibold, and D. D'Errico: HPI Future SOC Lab – Proceedings 2011, Technische Berichte Nr. 70 des Hasso-Plattner-Instituts für Softwaresystemtechnik an der Universität Potsdam, Potsdam, 2011.

[Bo11] M. Boehm, P. Lehmann, P. Volk, and W. Lehner: Query Processing on Prefix Trees, Dresden, 2011

[LT11] M. von Löwis, and P. Tröger: Safe and Scalable Build Automation through Virtualization, Tagungsband der INFORMATIK 2011, Berlin, 2011

[SS14] J. Spillner, and A. Schill: Towards Dispersed Cloud Computing, 2nd IEEE International Black Sea Conference on Communications and Networking, Chişinău, 2014

LAUDATIO-Repository.org – A Modular Approach of an Open Access Research Data Repository for a discipline

Dennis Zielke

Computer and Media Service
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
zielkede@cms.hu-berlin.de

Abstract: The presentation aims at presenting the LAUDATIO repository, a data repository for research data of historical corpus linguistics. It has a technical focus concerning the requirements that have been taken into consideration for the development of a repository infrastructure based on Fedora.

Research data collections are expensive in terms of time, money and knowledge that is necessary to compile them. The scientists' effort can be reduced by sharing these collections with a broader scientific community for re-use and re-evaluation. A common approach of creating sustainable and shareable data is the definition of existing and future open standards for the data creation and maintenance. The implementation and usage of digital repositories enables the management and distribution of that data. However, large collections of linguistic data, like corpora of deeply annotated text, pose unique challenges to repository design, which mainly arise due to the complexity of the applied data models. In the presentation a historical corpus linguistic data repository that is based on community-specific requirements and distinct user scenarios will be presented. Its main focus is on German historical texts including all dialects of time periods ranging from the 9th to the 19th century.

The LAUDATIO repository is developed in the cause of a research data infrastructure project funded by the German Research Foundation from 2011 until 2014 at Humboldt-Universität zu Berlin within a cooperation of Computer and Media Service, the Departments of Historical Linguistics and Corpus Linguistics at HU Berlin, as well as INRIA, France. [1]

The main objective for the technical implementation of the repository was to ensure the long-term access and preservation as well as the accessibility and the (re-)use of the heterogeneous research data. The requirements specification is essentially based on criteria such as the use of a complex data model for linguistic research data [Od13], their management, the structural collocation, the searchability, and long-term availability of the data. [Zi14]

Exemplary technical requirements regarding these aspects and the implemented features of the repository infrastructure include:

[1] LAUDATIO Repository: http://www.laudatio-repository.org [last access 26.06.2014]

- Compatibility with a complex data model
- Data indexing and search
- Data Versioning
- Registration and management of Persistent Identifiers (PID)
- Graphical User Interface (GUI) for data presentation
- Search and visualization with the help of ANNIS[2]

Fedora Commons[3] being an open-source repository framework that is well established in the repository community turned out to be the well-suited solution for the storage and management of linguistic research data considering the defined requirements. In the digital humanities domain Fedora is a widespread repository solution, and is for instance recommended and used by the CLARIN research infrastructure and its participating institutions. The LAUDATIO repository infrastructure is furthermore based on generalizable software modules such as the graphical user interface, the data exchange module between research data and Fedora REST-interface[4], and an indexed and faceted search based on Lucene-based[5] technology.[6] Naturally, the repository is oriented towards international guidelines concerning organizational, legal as well as technical aspects of research data repositories such as the Data Seal of Approval[7].

The presentation aims at a discussion on the re-usability of the LAUDATIO repository infrastructure modules for other initiatives providing research data infrastructures as well.

References

[Od13] Odebrecht, C.: Modellierung von linguistischen Forschungsdaten. Korpuslinguistik Kolloquium 13.11.2013, Berlin., https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/mitarbeiter-innen-en/carolin/modellierung-von-linguistischen-forschungsdaten/view

[Zi14] Zielke, D.: Open Access Research Data Repository for Corpus Linguistic Data – A Modular Approach Approach.Open Repositories 2014 Conference. 9.-13.6.2014 Helsinki, Finland, http://www.laudatio-repository.org/laudatio/wp-admin/tmp/2014/04/LAUDATIO-OR2014.pdf.

[2] ANNIS: http://www.sfb632.uni-potsdam.de/annis/ [last access 26.06.2014]
[3] Fedora Commons: http://www.fedora-commons.org/ [last access 26.06.2014]
[4] Representational State Transfer
[5] Lucene: An open source full-featured text search engine library, http://lucene.apache.org/core/ [last access 26.06.2014]
[6] Technical documentation of the LAUDATIO repository: http://www.laudatio-repository.org/repository/technical-documentation/; Sourcecode: https://github.com/DZielke/laudatio [last access 26.06.2014]
[7] Data Seal of Approval: http://datasealofapproval.org/en/ [last access 26.06.2014]

Status Report of bwFDM-Communities – A State Wide Research Data Management Initiative

Frank Tristram[*], Dennis Wehrle[**], Uğur Çayoğlu[*], Jessica Rex[***], and Dirk von Suchodoletz[**]

[*]Karlsruhe Institute of Technology, Steinbuch Centre for Computing (SCC)
[**]Albert-Ludwigs University Freiburg, Professorship in Communication Systems
[***]University of Constance, Communication-, Information,- Mediacenter

Abstract: Research data are valuable goods that are often only reproducible with significant effort or, in the case of unique observations, not at all. Scientists focus on data analysis and its results. By now, data exploration is accepted as a fourth scientific pillar (next to experiments, theory, and simulation). A main prerequisite for easy data exploration is successful data management. A holistic approach includes all phases of a data lifecycle: data generation, data analysis, data ingest, data preservation, data access, reusage and long term preservation. Tackling the challenge of increasing complexity in managing research data, the objective of bwFDM-Communities is to expose problems of research communities.[1] To achieve this goal, the project's key account managers enter into a dialogue with all relevant research groups at each university in Baden-Württemberg. Next to the identification of *best practices*, possible developments will be determined together with the scientists.

1 Project Motivation

Research data are generally understood as data that are generated during scientific work and they are building the basis for scientific results.[2] Such data can be very heterogeneous. They differ in origin, size and format, but share scientific significance. Research Data Management (RDM) includes all technical and organizational aspects of handling research data. This includes analysis, access, migration, integrity, metadata, visualization, and archiving, as well as cost models and legal aspects. In general, RDM attempts to increase scientific insight. This is usually difficult because research domains are very heterogeneous with respect to gaining knowledge and data demands. While some disciplines need to manage and analyze an enormous flow of data with each measurement they make, other disciplines only produce a few files in the course of their research. Therefore, most researchers do not focus on the management of their data. Nonetheless, the awareness that

[1]Especially the long tail of science, see e.g. `https://www.ci.uchicago.edu/blog/unwinding-long-tail-science` [last access 30.06.2014]

[2]DFG 2010 Call: "Informationsstrukturen für Forschungsdaten", `http://www.dfg.de/download/pdf/foerderung/programme/lis/ausschreibung_forschungsdaten_1001.pdf` [last access 30.06.2014]

a lot of accessible research data is not consistent, poorly connected, not traceable, and can hardly be integrated into one's own work is something most researchers share.

The current European political vision concerning research data can be summarized by a statement from the European High Level Expert Group on Scientific Data for 2030: "Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure..." [3] A strong German and European Open-Access-Initiative is following this political incentive that availability of high quality research data is an important key for science in Germany and the whole of Europe. The European Union, Germany and the German Research Foundation have announced their research strategy and development plans in several publications.[3,4] These developments suggest that external funding bodies will increase their requirements for data management and support more "Open Data" projects in the future. This will be realized by adjusting funding standards to new demands as well as through gradually increasing the general demands for all publicly funded research projects. In the near future it is likely that an obligatory publication of primary data is requested from all such projects.[5]

The project bwFDM-Communities is directly getting into contact with scientific communities in order to estimate their demand for services and infrastructure to handle research data at Baden-Württemberg's universities. The objective of the project is to gain knowledge about how relevant research groups at Baden-Württemberg' universities handle their research data and to identify their problems in order to suggest best practices. No such qualitative analysis of similar size and depth has been attempted in Germany before.[6]

2 Project Structure

At each university in Baden-Württemberg a key account manager is established. The key account manager's main task is creating a link between infrastructure suppliers and IT service institutions, libraries, and other university research institutions. Furthermore, the key account manager tries to gather information about the demand of the scientists in a continuous dialogue. Thereby the full spectrum of scientific data from diverse research fields is covered. Engaging in a continuous dialogue with suppliers of infrastructure and

[3]Riding the wave – How Europe can gain from the rising tide of scientific data, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf [last access 30.06.2014]

[4]Concept for IT infrastructure in Germany, http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf [last access 30.06.2014]; Information processing in higher education – Organization, Services and Systems, Recommendations of the Commission for IT Infrastructure 2011-2015, http://www.dfg.de/download/pdf/foerderung/programme/wgi/empfehlungen_kfr_2011_2015.pdf [last access 30.06.2014]

[5]Horizon 2020 – Multi-beneficiary General Model Grant Agreement, http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf [last access 30.06.2014]

[6]Some universities have carried out single surveys among their scientists. The WissGrid project (http://www.wissgrid.de/) and the Nestor (http://www.langzeitarchivierung.de/) community previously tried to bring experts and scientists together to discuss RDM problems within the scientific communities in Germany.

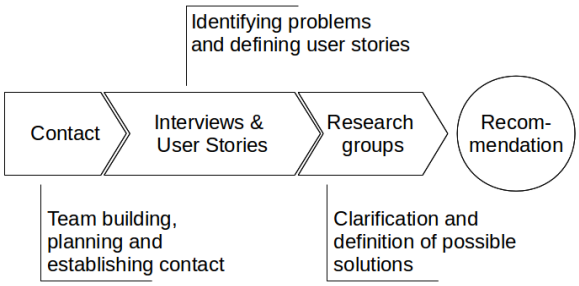


Figure 1: Phases of bwFDM-Communities project.

services the key account manager compiles all planned services. Ideally, the key account manager can hint at respectively helpful services and infrastructures for the scientist.

The project is divided into four phases (c.f. figure 1). The first phase, which has been reserved for team building, gaining knowledge, planning, establishing contacts and performing test interviews, ended in April 2014. The second phase, running from May 2014 till November 2014 is labelled as interview phase. Here, all nine key account managers are talking with scientists in order to identify scientific requirements and problems in connection with data management. Starting in December 2014, the clarification phase will follow, during which interested scientists and experts will come together in small groups and talk about possible solutions to specific problems that affect a larger number of scientists. In the last phase, which will cover the last three months of the project, a report will be drawn up and the results will be published.

Contact Phase In the first phase the key account managers' tasks were to identify the local data services at their universities, to create lists of all research groups at their respective universities and to identify a contact person for each research group. Furthermore, the next steps had to be planned and prepared taking special care of setting up an interview guideline properly. The guideline will be published at the end of the project.

Interview Phase Within six months, the current and future needs, problems and wishes of the scientists are identified. For this purpose a detailed conversation with each scientific group, concerning their present state and planned or expected developments, is necessary. As a result from the interviews user stories[7] will be written and – bringing together the various results – story maps[8] will be created.

Clarification Phase In the clarification phase topics of general importance as well as general demands will be defined and presented. This definition will be based on all user stories and the derived story maps as well as on the work done in the joint working groups.

[7] User stories are short paragraphs that explain a role, a wish (an instrument) and a purpose.

[8] A story map is a subsumption of related user stories that contains the main statements and also provides an embedding frame for each story.

Emerging questions and issues will be entered into a second questionnaire that mainly addresses technical details and aims at eliminating redundant stories efficiently.

Completion Phase The completion phase focuses on the evaluation, summary, and documentation of the project's results. On the basis of a substantiated demand analysis, detailed recommendations regarding concrete measures (e.g. building infrastructure, developing technology, transferring knowledge, ...), which can help to secure a sustainable research infrastructure for Baden-Württemberg's universities, will be provided. The recommended measures will feed into the second phase of the concept bwDATA as well as into the e-science strategy of Baden-Württemberg and complement the strategy by elaborating on concrete measures [HWC13].

3 Preliminary Findings

We have developed an interview guideline containing several elements we expected to come up frequently. Those can be marked like multiple choice answers by the interviewer. On the basis of these answers we will later build statistical plots and attempt to establish correlations. The structured interview covers the following topics of research data management: field of research, background and kind of data, data processing, storage, publication, data sharing and more general questions.[9] The qualitative information from the interviews is condensed into user stories. From a single interview typically two to three user stories, that include a specific need, requirement or problem, can be extracted by average. The user stories are stored in a database and reviewed by project members with regards to clarity and conciseness. The review is performed by joint working groups which are specialized on different topics and will ensure that the stories are well usable in the future.

As expected, our very first impression is that knowledge, structures, problems and solutions show a strong heterogeneity between the universities and disciplines. Some groups are not aware of the increasing importance and standards of data management set by the DFG or the EU. The experience from our interviews shows that even some of the professors do not know about DFG's guidelines and policies, a fact that has been proven by other studies before [SKS13]. Although data management plans are often formulated for project proposals, their implementation into the daily research routine is often missing. A reliable data storage system for a minimum of ten years is not provided in most cases. We also observed that some scientists do not care about what they write into their proposal. They write what they feel is expected of them but do neither take care of data management nor of the preservation of their data for at least ten years, as it is demanded by the DFG guidelines on good scientific practice. This may be partly caused by a lack of knowledge, motivation or incentive as well as a lack of reasons. This structural problem seems to be derived from unclear responsibilities. The scientist might see his institution in the position to care for

[9] E.g. whether the scientist feels well informed about research data management and which channel of information he or she would prefer.

a long term preservation of his or her data. The institution in turn, might not have the resources, or also might not feel responsible for providing long term access to all data. A solution might be university policies that clarify who is responsible for fulfilling the demands set by the funding agencies. Scientists are usually aware of all of these deficiencies themselves, for instance, the reputation of a small scientific group hardly increases when sharing their raw data, but they often worry about larger groups that might then dominate their field of research. Such dilemmas cannot be solved by official requirements alone. Apart from scientists lacking a thorough research data management, there are, of course, groups that show a high quality of data management.

However, aside from large collaborations, an exemplary data management behavior seems to either be directly important for the scientist's daily work or else depends on single persons who insist on good practices with research data. Proper tools, services and guidelines can at least create a very supporting environment for good practices and are also often requested by the scientists. Currently there is a perceived gap between demand and received support in research data management for small scientific communities.

Acknowledgments

The work presented in this publication is part of the project "bwFDM-Communities - Wissenschaftliches Datenmanagement an den Universitäten des Landes Baden-Württemberg" sponsored by the Ministry for Science, Research and Arts of the federal state of Baden-Württemberg, Germany.

The work would not have been possible without the support of many other people. The authors wish to express their gratitude to all additional key account managers: Dieta-Frauke Svoboda, Peter Bamberger, Jörg Hertzer, Johannes Knopp, Jonas Kratzke, Fabian Schwabe and Denis Shcherbakov. We furthermore want to thank Wiebke van Ekeris, Claudia Kramer, Pia Daniela Schmücker, Uli Hahn, Hans-Jürgen Goebelbecker and Karlheinz Pappenberger from the university libraries of Freiburg, KIT, Konstanz and Ulm for their time and contributions to our project. We are finally highly indebted to thank Marcus Hardt, Christopher Jung and Achim Streit from SCC at KIT for their supervision and for providing necessary information regarding the project. They have set up and guided the project in the lead time and the starting month.

References

[HWC13] Hannes Hartenstein, Thomas Walter, and Peter Castellaz. Aktuelle Umsetzungskonzepte der Universitäten des Landes Baden-Württemberg für Hochleistungsrechnen und datenintensive Dienste. *PIK-Praxis der Informationsverarbeitung und Kommunikation*, 36(2):99–108, 2013.

[SKS13] Elena Simukovic, Maxi Kindling, and Peter Schirmbacher. Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin. 2013.

Looking Beyond Data Through the Ages

Dennis Wehrle[*], Wiebke van Ekeris[**] and Uli Hahn[***]

[*]Albert-Ludwigs University Freiburg, Professorship in Communication Systems
[**]Albert-Ludwigs University Freiburg, University Library Freiburg
[***]Ulm University, Communication- and Information Center

Abstract: Accessing data has always been crucial in order to get access to (cultural) knowledge. But in the new ages different challenges arise to keep complex data accessible. Concerning these challenges a linguistic use case, which has been set up in 1993, serves as an example to show that it might be hard to re-enact an old system in order to reuse its research data and processes. Re-enacting the old system showed that other crucial points emerge in the context of research data management that need attention. In order to increase the probability of gaining access to research data in the future Data Management Plans can be identified as central tool. Although preservation of and creating access to data is a global task, the scientific management of data is still strongly shaped through national regulations and practices and hence delivers a quite fragmented picture.

1 Introduction

Access to data has always been crucial in order to get access to knowledge. Through cave painting one can learn a lot about the past. This was also practiced in the ancient world where data was stored in books and scrolls with the purpose to get access to cultural knowledge and to preserve it for the future. In science the accessibility of data is important, too, not only to ensure re-examination of publications. One of the most important points, is to facilitate re-usage of research data in order to prevent duplication of effort. The European Union, with their framework programme for research and innovation, or the DFG on a national basis in Germany are driving forces to encourage re-usage. But "experiments and numerical calculations can only be repeated if all important steps are reproducible. For this purpose, they must be recorded" [For13]. Thus the accessibility to research data needs to be ensured to foster new research. This also enables interdisciplinary research. One example for this interdisciplinarity is the climate database called *tambora*[1]. For a given time and place, one gets the weather condition when a correspondent entry in the map is given. This data is based on letters, ship diaries, newspaper articles, etc. Shared with other research areas, this data could be reused for other purposes, too. But (old) research data also have value to be reused for further research, as it is the case for our use case. The research project at hand was collecting local dialects of the south-western region of Germany, which were of interest to be reused for further research again. There are

[1]Tambora, `http://www.tambora.org` [last access 27.06.2014]

different approaches to granting accessibility to this data. The approach described in this paper uses the method of emulation in order to re-enact an outdated networked research environment. However this use case shows the problem which arises from the technical progress. Nowadays data is generated and stored mostly digitally. Compared to the past, the lifetime of systems and formats is changing fast. Additionally getting access to, and understanding of, data is more complex compared to books. Thus it is not sufficient to only ensure access to the data, because without having the tools to interpret the data, knowledge will be lost.

In order to see how knowledge, respectively data, is dependent on its access to it, first a historic vision of cases in the analog age is given in section two. With these examples in mind, a shift to the digital age shows in section three, which aspects of access to data changed from the analog to the digital age. Furthermore, the methods of research data management and Emulation-as-a-Service (EaaS) are introduced. The linguistic use case shows in section four, how a functional approach is used, to grant access to old research data. Section five finally takes an abstract stance on EaaS trying to typify this method within the field of research data management in general.

2 Data Management in the Past

When looking into the past the loss of data and cultural knowledge is omnipresent. In the following, two historical examples are given in order to emphasize the problem of the loss of data due to natural or man-made incidents: cave paintings from the Stone Age and the Library of Alexandria in the ancient world. One consequence of this loss is that sustainable accessibility of knowledge and data is not possible anymore.

Stone Age Cave paintings are the oldest preserved type of man-made data. On the basis of these paintings characteristics like the actual age, the fauna and flora at that time, as well as the use of tools can be studied. In Europe the most prominent places for cave paintings are southern France and northern Spain [LB00]. The cave paintings were formed in the Palaeolithic, around 40,000 BP to 12,000 BP.

However, it is unknown how many cave paintings are lost, destroyed or not yet discovered. As a consequence, the preserved cave paintings might not be representative. Another aspect is that the accessibility of cave paintings is, in general, limited. Because the original cave paintings cannot be stored in archives like scrolls and books, copies of the paintings (in form of photographs etc.) have to be produced and stored for the purpose of providing broad access to the data. In any case, information is lost by the process of copying, because certain information like the technique of scratching or the composition of color is not passed on.

Ancient World The Library of Alexandria was the most famous and largest library in the ancient world [Mac00]. More than 400,000 mixed and 90,000 single scrolls are said to have been stored in the main library, as well as about 40,000 books in the satellite-daughter

library. Also the first subject catalogue of the world with six sections and around 120,000 scrolls of classical poetry and prose was build up there. The scrolls were stored in linen or leather jackets and were kept in racks in the hall or in the cloisters and lasted for at least 300 years until a great fire burnt most of the library down. Unfortunately, most of the scrolls and books were lost. The exact date and the reason for the fire remains unknown. Three incidents or a combination of them are probable, though: 48 BC during the fighting between Julius Caesar and Cleopatra, 272 AD after an attack of Aurelian, the emperor, or 391 AD, while Theophilus (the Christian patriarch of Alexandria) tried to turn one of the temples in the city into a church [Mac00].

However, the story of the Library of Alexandria shows that due to disasters – natural or man-made – the loss of data and cultural knowledge can happen at any time.

3 Data Management Today

Nowadays, in the age of digital data, natural catastrophes like fires are not the main threats with respect to the loss of knowledge and data, compared to the Stone Age and the Ancient World. The short lifetime of data storage media and formats, their diversity and complexity have to be considered as modern threats with respect to archiving. Another aspect is the permanent access to and the interpretability of data, which has become a new challenge in the digital age.

According to Neuroth et al. (2010) the lifetime of digital storage media is – with around 100 years – considerably shorter than the lifetime of ancient scrolls or cave paintings, which are still preserved since the Paleolithic and the ancient times [NOS+10]. Not only the lifetime of storage media, but also the lifetime of data formats is limited, because data formats are in flux. In the analog age, different media were building around the data. For example, since the 20th century libraries and archives have increased the diversity of media to be stored: in addition to books and journals, microfilms, microfiches, disc records, audio and video tapes, and finally digital media were considered [PRSU11]. In the digital age, the digital medium becomes predominant and replaces the various analog media step-by-step. Additionally, the complexity of digital data has increased. For example, climate data and models cannot be described only by text. Therefore, the access to data has changed, because complete structures of these complex systems have to be represented. As a consequence, the access to knowledge has changed, too, because it is no longer possible or useful to copy all data into books. One functional approach to enable access to knowledge and data is emulation, which is described later on in this section and is the basis for our use case in section four. In order to provide permanent access to, and ensure interpretability of, digital data, the performance of research data management is of general interest, while in the case of emulation it is of explicit interest.

3.1 Research Data Management

The access to data has generally changed over times: In the ancient world, the traditional libraries were responsible for providing access to scrolls and books for their respective audience. Nowadays, their role has changed. The main task of libraries still is to provide access to books, articles, and other text-based publications. Additionally the fields of digital data, research data, and e-Science have become more and more important. These new fields come along with new scopes of duties like licencing, data management, long-term preservation of digital data, retrodigitalisation, and open access [BW11]. In order to cope with these new fields and to provide broad access to research data, a worldwide repository landscape has shaped. At the moment this landscape is still fragmented. There are many isolated solutions depending on the different institutions and the special fields of science, because the requirements with regard to technology, content, and archiving differ tremendously [BHM11].

Platforms like re3data[2] or Databib[3] resemble powerful search engines that bring an organizational structure into the scattered repository landscape. Their existence is proof enough that an individual overview over the repositories is not possible anymore. Apart from the infrastructural side, there is another aspect worth mentioning in connection with research data management. In their "Guideline on Research Data Management" (Leitfaden zum Forschungsdaten-Management), Jens Ludwig and Harry Enke describe the management of research data as measures which guarantee that digital research data are usable. What measures are necessary depends on the purpose of the use. Ludwig and Enke define four cases: (1) using the research data as a working copy for the scientific process, (2) re-using the research data for future research, (3) archiving documents for complying with the good scientific practice and (4) archiving the data for complying with legal or other non-scientific demands [LE13]. Research data management does not only imply the handling of data but also the process of data generation. Therefore both the scientist's side as well as the motivation to publish the generated data has to be considered. Most common may be the publishing of data due to regulations by external funding agencies, or because it is common practice in the scientist's respective research field. The priority initiative 'Digital Information' by the Alliance of German Science Organisations has postulated principles for the handling of research data. The principles are formulated very cautiously: While long-term preservation of, and the principle of open access to, data is supported by the alliance, at the same time this principle "shall be balanced against the scientific and legal interests of the researchers." [4] External funding organizations like the German Research Foundation (DFG), who are part of the Alliance of German Science Foundations, ask for compliance with the principles of good scientific practice, including the storage of and the access to all primary research data for at least ten years [For13]. Here, a divide between publishing and storing research data becomes obvious. While it is easy to formulate clear guidelines in order to maintain the quality of German research as such, it is not so easy to impose open-access-publishing for research data upon scientists. Although the Alliance

[2] Registry of Research Data Repositories, `http://re3data.org` [last access 27.06.2014]

[3] Databib, `http://databib.org` [last access 27.06.2014]

[4] Principles for the Handling of Research Data, `http://www.allianzinitiative.de/en/core_activities/research_data/principles/` [last access 27.06.2014]

argues that publicly funded research needs to be made accessible publicly, the legal situation is far more complicated, especially concerning cooperations with private partners. Yet, in both cases – be it to follow the DFG's recommendations for safeguarding good scientific practice, or to follow the alliance's ideal of publishing research data as open access – data management plans should stand at the beginning of every research project that is collecting data. "A data management plan is a document outlining how the research data collected or generated will be handled during a research project, and after it is completed, describing what data will be collected / generated and following what methodology and standards, whether and how this data will be shared and / or made open, and how it will be curated and preserved ..." [Com13a] Data management plans are already common practice in the US or the UK with whole institutions forming around the idea of preserving digital data.[5] With data management planning tools like DMPonline from the DCC[6], where scientists get help with formulating their data management plans according to their national and their funder's demands, the discrepancies between the UK and the US on the one side, and countries like Germany on the other side become apparent. In Germany such a DMP-tool that helps scientists in the funding stage of their project to create a data management plan only is offered at Bielefeld University. Of course, there is also a lot of activity when it comes to advocating the necessity of caring about research data, for example the NESTOR-initiative taking care of long term preservation and the above mentioned DFG formulating guidelines for scientists and various projects like the WissGrid initiative seek to improve the situation in Germany. In the UK the Digital Curation Center (DCC) puts significant effort into the sensitization of scientists but they also give lots of practical information that helps to improve the scientists' workflows, something that in Germany is not practiced to this extend. With the advent of Horizon 2020, data management plans might get more popular in Germany as well. Research projects that participated in the Open Research Data Pilot already contained data management plans. The optional article 29.3 of the Model Grant Agreement in Horizon 2020 now provides the legal requirements for such projects to integrate data management plans in the new framework program. Article 29.3 consists of two parts: Part a) is addressing the compulsory tagging of the research data with metadata, necessary for open access or long-term publishing in a repository. Part b) describes what is needed apart from the metadata in order to maintain the chance of re-using the concerned data sets: "Regarding the digital research data generated in the action ('data'), the beneficiaries must [...] provide information – via the repository – about tools and instruments at the disposal of the beneficiaries and necessary for validating the results (and – where possible – provide the tools and instruments themselves)."[Com13b] As shown in section five, a data management plan does not infringe the author's copyright. Yet, tagging the data, and giving a full subset of the used programs, hardware and software, will remain the responsibility of the scientist. By demanding data management plans from the scientists, a whole subset of positive effects could be achieved. In order to render long term preservation a valuable action, data has to be kept interpretable and tools to access the data should not be forgotten. How important this is can especially be seen when looking closely at preservation (or re-enacting) methods like Emulation-as-a-Service.

[5]RDM guidance webpages, `http://www.dcc.ac.uk/resources/policy-and-legal/rdm-guidance-webpages/rdm-guidance-webpages` [last access 27.06.2014]

[6]DMPonline, `https://dmponline.dcc.ac.uk/` [last access 27.06.2014]

3.2 Emulation-as-a-Service (EaaS)

Emulation is becoming a key strategy for digital preservation and securing access to digital artifacts. By digitally recreating computer systems, emulation ensures that digital objects can be rendered in their native environments and thus maintain their original "look and feel." In many cases it is only feasible to use the original environment in order to reuse research data and its processes, because a migration to a newer system might not be successful. Nevertheless, deploying operational emulation setups is a complex and laborious task requiring specific technical expertise and knowledge of how to maintain obsolete computer systems. Thus up to now emulation has been perceived as a domain reserved for technical experts only. In order to address both technical and organizational challenges, a scalable emulation-based service model has been developed: Emulation-as-a-Service (EaaS). EaaS is focusing on a functional approach. Accordingly the goal is to make digital artifacts sustainably accessible, which in turn means that the original system may be (slightly) changed without loosing the correctness and completeness of the digital artifact.

The fundamental building blocks of the EaaS architecture are abstract emulation components to standardize deployment and to hide individual emulator complexity. Each emulation component encapsulates a specific emulator type as an abstract component with a unified set of software interfaces for data I/O and technical as well as user-machine interaction. This way, different classes of emulators become interoperable. This means that various emulation technologies like qemu[7] and different virtualization technologies, for instance Oracles VirtualBox[8], could be used simultaneously. Currently, emulation components for all major past and present desktop systems, e.g. PPC, Sparc, M68k, Intel-based x86 and major operating systems, e.g. OS/2, MS Windows, Mac OS 7 and Linux are available for deployment. A detailed technical description of the EaaS framework and its workflows can be found in earlier work [RVvL12]. With EaaS it is not only possible to preserve a single system but to preserve a complete networked environment, too. Thus, preservation of a scientific desktop including its research data and processes which have dependencies on other machines, is feasible. After the system is successfully preserved, end users have access to the preserved system remotely through standard (web-)clients (e.g. web browsers).

Although emulation is certainly a key strategy in digital preservation still it does not seem to have found its place in the field of research data management. The "Handbook on Research Data Management" [BHM11], for example, does not list emulation per se; nonetheless emulation is hidden behind the scenes. When looking at the Curation-Lifecycle-Model[9] developed in 2010 by the DCC, it gets obvious that EaaS also fits perfectly into this lifecycle. After appraising and selecting what to preserve and ingest, which could be summarized as preparation step of EaaS, the emulation itself takes place covering the steps of preservation action and storage, ultimately providing for access, use, and reuse of the research data.

[7] QEMU open source processor emulator, `http://www.qemu.org` [last access 27.06.2014]]

[8] VirtualBox, `http://www.virtualbox.org` [last access 27.06.2014]

[9] DCC Curation Lifecycle Model, `http://www.dcc.ac.uk/resources/curation-lifecycle-model` [last access 27.06.2014]

4 Use Case: Linguistics

In the linguistics department at the University of Freiburg, a long running research project got recently shut down. Since the 1970s, the project had been collecting local dialects of the south-western region of Germany.[10] Later on it became one of the first projects in this field of the humanities to move from laborious paper based evaluation to a computerized environment. All collected data got added to a database. The data then was used to create customized dialect maps depending on various input parameters which got published in numerous publications and theses. Several specific workflows and even custom font and symbol shapes were created to produce Postscript output from a TeX file source. Many researchers and PhD students added to the project and refined the workflows over time. Unfortunately, no actual user of the system has a full understanding of how the system is set up and configured. As the contained research data has value to be reused for new research and cannot easily be extracted from the system, we took the system as a use case for preserving full scientific original environments.

In order to be able to reuse this research data, it is not only vital to preserve this data but to preserve the complete system, too, as it provides different procedures to process the data. With the EaaS approach, it is feasible to preserve not only a single system but a complete environment (e.g. a Client-Server infrastructure). In order to preserve a system an analysis and a preparation step, ideally carried out while the machine is still operational, is required. This includes an assessment of hardware, software (e.g. operating system, memory, network configuration, etc.) as well as its configuration to be able to migrate the machine to an emulated environment. In a second step, an imaging process of the physical hard disk is required [WRCv12]. Furthermore, to ensure proper re-enactment of the old system, it is not only vital to know explicit hardware requirements but also to determine explicit and implicit dependencies and their functional expectations. Such dependencies and expectations could be, for instance, a system with a hard coded address to connect to a database system within the local network, or, like in this use case, the system could expect network drives which will be initialized at startup. Thus, these network resources have to be available. Finally, the resulting re-enactment must be checked with regard to whether the expectations are met and the system is functionally working.

4.1 System Analysis

The system got set up in 1993 and consisted of one server machine running on OS/2 driving an IBM DB2 database and five client machines, offering access to the database over LAN. By the end of the project the server was fully functional and at least three clients were completely, two more partially, working (two to three hard disks were more or less unreadable, in another machine the floppy drive was failing or the system was not booting because of mainboard failure). The LAN infrastructure was running both TCP/IPv4 and NetBUI on a physical TokenRing infrastructure. All machines were x86

[10] The project was relevant enough to get added to the permanent exhibit of the Freiburg University Museum, `http://www.uniseum.de` [last access 27.06.2014]

hardware, featuring 486DX2 for the clients and a Pentium/Overdrive CPU (upgraded from 486) in the server. The machines were equipped with 8 to 16 MByte of RAM. All machines were connected to 15" CRT monitors powered by CirrusLogic VGA graphic adaptors with 1 MByte of RAM. The server hard disk was a SCSI disk of 1.1 GByte, the client hard disks of 240 MByte were using IDE adaptors. All clients were identically configured and the installation on the clients were originally identical. The system description illustrates how the technical configuration has changed over the last 20 years compared to today's computers. On basis of the derived information, a proper virtual machine was configured. As a next step to re-enact the system, images of the hard disks have to be created.

4.2 Challenges and Issues to Re-enact the System

There are several methods available to create an image of a hard disk: (1) in-system, non-intrusive method, which means to boot a networked mini linux operating system on the original system (preservation target) or (2) an intrusive method, which involves removing the hard disk for connecting and dumping it to a suitable imaging system. Such a dumping system has to provide appropriate connectors (e.g. SCSI or IDE), either internally (e.g. via mainboard) or externally (like USB-connectors). Because there was no suitable mini linux and no suitable network connection available, the intrusive method was applied. Therefore, the machines were disassembled to remove the drives. In order to create an image of the SCSI server hard disk and the IDE client hard disk, proper connectors had to be found. Both connectors are outdated today and are not used on new machines anymore. A working controller was found for SCSI, whereas different controllers were used for IDE. Interestingly the IDE system imaging procedure was more difficult than imaging the SCSI disk, because the logical layer changed over time. Therefore the produced disk image was not correct or the original disk was not even recognized. Finally, a controller could be found with which a correct image of the disk could be produced. Despite the imaging problem, there were other problems when starting the system for the first time. Because of the use of a virtualized environment the machines' expectations of the system were not fulfilled. This means that on a virtualized environment there are not exactly the same hardware components (e.g. graphic card, network card, etc.). Accordingly, we had to solve these issues too, by changing configuration parameters and installing new drivers manually.

4.3 Result

After solving the aforementioned problems, the two systems could finally be preserved and re-enacted. This includes the server machine which hosts the database for the linguistic data and a client machine which queries the database and processes the data (e.g. create a custom map). Despite the fact that the system could be preserved, knowledge is still lost (maybe just temporarily) as the location and the knowledge of how to use various tools is unknown. One possibility to process the data is to use a tool named `ListOASRV.cmd`,

which produces a custom map. But the application of this command is not intuitive as several steps have to be done, for example a key combination of ALT+Z, in order to produce the final result. Without the knowledge of how to use the tools, one might not be able to produce the desired result.

The use case clearly shows that it is possible to re-enact an old system for re-usage of research data. But it also illustrates that it might be hard to re-enact old systems properly as knowledge of the respective system is needed. Additionally, the IDE-Controller problem demonstrates that also out-dated hardware might be necessary to properly migrate the system.

5 Taking Care of the Future: The Necessity of Data Management Plans

This paper started out by stating that the loss of knowledge is – and has always been – omnipresent, be it in the form of paintings, writings or data. While in the analog age the loss of data could be linked to either man-made or natural causes, these boundaries are blurred in the digital age. Digital data, as well as its analog pendant, certainly suffer from natural decay. But, as the use case demonstrates, the decay of digital data may also have (unintended) man-made causes. Scientists can certainly not consider the various problems with which one is confronted by the preservation process of a system, which resembles a blindness of what is to come. This blindness calls for certain preventive steps when handling digital data that were started to be articulated with the practice of writing data management plans. The German Research Foundation introduced such data management plans in their 2010 proposal preparation instructions for project proposals, yet made those plans only mandatory for projects where "research data or information will be systematically produced."[11] The central question to such plans is, how scientific data can be kept accessible and, if the data is not accessible anymore, how accessibility to the scientifically relevant data can be achieved again. It becomes apparent that science cannot be reduced to text or data anymore, but that for achieving full accessibility (e.g. to a single system, but also to whole research environments like in the demonstrated use case) the tools that are used to make the data interpretable proof equally important.

As the use case shows, Emulation-as-a-Service can definitely serve as a solution to make old research data reusable again. At the same time, however, EaaS will certainly not be a solution for every outdated system. Although EaaS contains a certain "out-of-the-box"-component that results in a highly automated workflow, this only becomes possible if enough information about the system is available. In order for EaaS to work fluidly, data management plans that provide all relevant data in order to maintain the system are certainly helpful. Missing information on software, hardware, and data are the main reason, why re-enacting old systems is harder, than it will be with systems for which data management plans have been written. It may also be helpful to use (centrally) managed

[11] Leitfaden für die Antragstellung, http://www.dfg.de/formulare/54_01/54_01_de.pdf [last access 27.06.2014]

(virtual) research environments, where a standardized or partly standardized system will be used within a working group. If such environments are planed before or at the start of the project, it will be easier to migrate the system later or even just move the virtual environment into EaaS.

As EaaS needs a preparation step for the emulation to work properly, the first step to emulation will always be a decision making process: Is the re-enactment of the scientific data so relevant that the effort is paying off? The question which scientific data should be kept and if data is kept, for how long, is raised in other parts of the scientific domain as well. But while the preservation of scientific data seems to be fairly simple, the maintaining of its accessibility is much more complex. Repositories for scientific data are confronted with the same central questions as EaaS. Yet, while repositories can pertain a policy of "store first, ask later", EaaS works the other way round. It is essential to first consider the effort of re-enacting a system before taking action. This necessity proves that keeping data accessible does not only involve technical aspects but also entails a management aspect. By providing data management plans before or during scientific projects, EaaS can be implemented easier than stated in the use case above.

6 Conclusion

In this paper, the necessity of research data management was outlined by analyzing the ancient and modern threats with respect to the loss of cultural knowledge and data as well as the permanent access to data. Research data management is currently converted into a prerequisite for the transfer of benefits through the stipulation of data management plans by funding agencies like the DFG.

When looking into the future, an interconnection between (specialized) repositories and data centers combined with integrated EaaS becomes thinkable. Then a customized data curation, including permanent access and a sustainable long-term preservation, would be achieved. Technical and administrative interfaces are needed in order to setup such a repository network, e.g. OAI[12], service level agreements, and standardized workflows. Furthermore, local contact points for research data management and specialized IT infrastructure are required.

The project "Coordinated Structures for Verification and Efficient Reusability of Research Data Throughout the State of Baden-Württemberg" has the goal to realize the scenario mentioned above and to serve as a model for other institutions in the State of Baden-Württemberg. It is currently executed by the University of Freiburg's IT-Center and University Library, as well as by the Ulm Communication- and Information Center.

[12]Open Archives Initiative, `http://www.openarchives.org/` [last access 27.06.2014]

Acknowledgments

The work presented in this publication is part of the project "Coordinated Structures for Verification and Efficient Reusability of Research Data Throughout the State of Baden-Württemberg" ("Landesweit koordinierte Strukturen für Nachweis und effiziente Nachnutzung von Forschungsdaten") sponsored by the Ministry for Science, Research and Arts of the federal state of Baden-Württemberg, Germany.

References

[BHM11] Stephan Büttner, Hans-Christoph Hobohm, and Lars Müller. *Handbuch Forschungsdatenmanagement*. Bock+ Herchen, 2011.

[BW11] Sabine Brünger-Weilandt. *Gesamtkonzept für die Informationsinfrastruktur in Deutschland: Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder*. Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz e. V., 2011.

[Com13a] European Commission. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020, 2013.

[Com13b] European Commission. Multi-beneficiary General Model Grant Agreement, 2013.

[For13] Deutsche Forschungsgemeinschaft. *Proposals for Safeguarding Good Scientific Practice*. Wiley-VCH, 2013.

[LB00] Michel Lorblanchet and Gerhard Bosinski. *Höhlenmalerei*. Thorbecke, 2000.

[LE13] Jens Ludwig and Harry Enke. *Leitfaden zum Forschungsdaten-Management: Handreichungen aus dem WissGrid-Projekt*. Verlag Werner Hülsbusch, 2013.

[Mac00] Roy MacLeod. *The Library of Alexandria: Centre of Learning in the Ancient World*. IB Tauris, 2000.

[NOS+10] Heike Neuroth, Achim Oßwald, Regine Scheffel, Stefan Strathmann, and Karsten Huth. *Nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*. Verlag Werner Hülsbusch, 2010.

[PRSU11] Engelbert Plassmann, Hermann Rösch, Jürgen Seefeldt, and Konrad Umlauf. *Bibliotheken und Informationsgesellschaft in Deutschland: Eine Einführung*. Harrassowitz, 2011.

[RVvL12] Klaus Rechert, Isgandar Valizada, Dirk von Suchodoletz, and Johann Latocha. bwFLA – A Functional Approach to Digital Preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):259–267, 2012.

[WRCv12] Ian Welch, Niklas Rehfeld, Euan Cochrane, and Dirk von Suchodoletz. A Practical Approach to System Preservation Workflows. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):269–280, 2012.

Reviewing the Governance Structure of End-User Support in e-Science Infrastructures

Hashim Iqbal Chunpir and Thomas Ludwig

Department of Informatics, University of Hamburg
German Climate Computing Centre (DKRZ), Bundesstr. 45 a,
20146 Hamburg
chunpir@dkrz.de
ludwig @dkrz.de

Abstract: The research into e-Science infrastructures has changed the face of traditional scientific methods. The maturity of grid technology to get the data distributed around the globe has raised possibilities to perform ground-breaking research within a specific scientific community. However, the governance and management of user support services of e-Science infrastructures yet remain an under-studied topics. Currently, there is hardly a study undertaken to see what governance approach may be suitable to enrich the user support services that an e-Science infrastructure may provide. In this paper, the governance context of the user support services of a use case of a climate e-Science infrastructures: Earth System Grid Federation (ESGF) is taken as a case study. Amongst the plethora of challenges that the user support process in ESGF face, we focus on its governance and make recommendations to improve it. The recommendations are based on an empirical study conducted to find out the current state of user support governance.

1 Introduction

The utility and importance of user-support (also known as service-desk) in any organisation cannot be underestimated. However, the resources of an organisation, especially a research institute are mostly limited. This is the case in e-Science infrastructures, where the employees, hardware and software components of multiple national or global organisations form an e-Science infrastructure. The employees of the e-Science infrastructure have a limited time and budget to answer arriving user requests [SSW10, TSM12]. In Information Technology Service Management (ITSM) terminology the user requests are also known as incidents [Jä09].

In most of the cases, it is difficult to automatically process the incoming user-requests (incidents) because the application of (intelligent) technologies to automate the process either has limitations or is too expensive to implement. Therefore, in e-Science infrastructures, the employees of organisations that constitute a particular infrastructure handle incoming incidents in addition to their core activities, mostly on voluntarily basis [CBL14, CLB14a]. As per observations in Earth System Grid Federation (ESGF), some

employees participate in user support activities more than others [CLB14b]. ESGF is a well-known climate cyber-infrastructure [Vu13, Wi12]. Normally, the employees of cyber-infrastructures, for instance in ESGF (but not limited to that), are over-whelmed by their core tasks/activities and find little time and interest to process incidents [CLB14a].

The core activities of an e-Science infrastructure employee include: Research, programming, system administration, data curation, data management and others. These core activities are necessary to advance technology, cultivate standards and maintain the operations of the infrastructure. Processing an incident is not formally part of the job description and is considered as volunteer activity by the employees. Handling incidents depends on employee's free time, interest and motivation to get indulged in supporting users[1]. It seems appropriate to recognize the user support activities as part of the job description. However, in addition to this suggestion, some changes in the current governance set-up of ESGF have been proposed and elaborated in this paper, after interviewing some stakeholders and examining ESGF documents. The new governance structure suggests to include user support services in the current governance scheme of ESGF in the form of recommendations. Furthermore, it is suggested to recognize user-support-worker's effort that may be rewarded.

This paper is organized as follows: Section 2 describes the background of e-Science infrastructures, user support, the significance of user support in e-Science infrastructures and the existing models in optimizing employee/workers activities. Section 3 discusses the contemporary user support practices in ESGF as captured from diverse sources of information and the significance to introduce some changes in the current governance. Section 4 elaborates the current user support governance recommendations to be included in ESGF overall governance, followed by its discussion and future work in section 5. Finally, conclusion is deliberated in section 6.

2 Background

The related work of this paper is divided into the four main areas given in the sub-sections:

2.1 e-Science Infrastructures

An e-Science infrastructure is collaborative, relies on grid computing technologies to share and use data to perform research in different fields including humanities and science. e-Science infrastructures are also known as e-Research, cyber-Infrastructures that are deployed to access and share the data, high performance computing (HPC) facilities and human resources to facilitate intra-disciplinary and inter-disciplinary research to harvest knowledge. An e-Science infrastructure is formed through collaboration of many organisations across national and international boundaries. It is

[1] The word "users" can be replaced "customers". Customers pay for the services and may not necessarily be the users.

the synchronized integration of software, hardware, other technologies, as well as human expertise, required to support research in science, humanities and engineering. Networks that constitute e-Science infrastructures are complex: Users need an interface to access its resources usually data [Fr07]. The interface includes command line tools, web portals and Graphical User Interface (GUI) to access data assets; which are the main resources hosted by an infrastructure. However, during an interaction of a user with an e-Science infrastructure, a user may require help due to outages of some resources e.g. servers or any other anomaly. In addition, a user may require a particular scientific information. In order to meet these user support challenges, every e-Science infrastructure offers user support in the form of a help-desk [Ch13, CWC13, SSW10].

2.2 User support

User support has always been seen as an extra function to the core services of corporations. It was not up until beginning of 2000s that it was realized that customer support should be enhanced. This caused application of business process frameworks to improve service quality and customer satisfaction [Ke02]. Since then, diverse support models and structures are tested to match the corporation business model of servicing customer or end-user concerns. The first help desk (HD) in the 1980's had only a desk, pen and a telephone used by human support agent [Ke02, LL07].

The traditional HD afterwards had gone through different levels of evolution with the change in the commercial organizational set-up. In order to provide customer support different techniques were employed. These techniques include Automatic Call Distributions systems (ACD) [UHG03], Interactive Voice Response (IVR) systems [Cz98], help desk management system (HDMS) along with associated reporting tools [MM96]. Moreover, help desk expert systems, knowledge-management centric help desks [GGR05] are also used. In addition to it embedding case-based reasoning (CBR) engine in help desk [AP94], [Ro04], help desks based on corpus-based analysis (CBA) mechanisms [MZ05], [ZM06], use of remote control technologies to support end-users and web based e-support techniques with and without human support agents [DR04] have been employed to improve user support process. In the case study of ESGF, user support concept covers "help-desk" or "service-desk" of a distributed, multi-organisational research-oriented, non-commercial, collaborative environment. The current user support in ESGF is being performed by human support agents i.e. employees, that include top scientists [CLB14a].

2.3 User support in e-Science Infrastructures

In the last decade, the user-support in ESGF has been evolving mainly due to the changes in ESGF architecture. For instance; looking at the development history of ESGF, the technological plus organisational changes, introduction of new data projects served by the ESGF data archive system, the number of users and their needs have been on constant rise that impacted the user support structure and the behaviour of ESGF employees to treat user requests [BBB+05, CCM+12, CWC13, Wi12]. Consequently, up until now the employees of ESGF are performing the user support by handling incidents,

on a free will basis, on top of their core infrastructure development activities/ tasks. A recent survey questionnaire and mailing list analysis conducted revealed that up to 15% of the incidents were ignored by the employees. Therefore a need to enhance ESGF user support was felt and as a part of result changes in the organisation and particularly governance structure of ESGF were suggested.

2.4 Governance and User Support

Governance is a broad concept and is being used in different domains with different connotations. In this paper, governance of e-Science infrastructure is seen as the formal and informal institutions, mechanisms, relationships, and processes between the partners or organisations. They form the strategy and underlying processes defined by the e-Science infrastructure executive board to achieve a collective benefit or a common goal. This perception of e-Science infrastructure governance is based on well-established definitions of IT Governance [Gu04].

The key elements in the Governance of e-Science infrastructure are the alignment of the service goals with the IT instruments (i.e. technological components) that must lead to the achievement of value i.e. better service for the users and stakeholders. Better e-Science governance can be achieved by setting up an e-Science infrastructure governance framework. Such a framework must encompass practices of proposing structures and processes that introduce better user-support mechanisms.

3 Research Methods Applied

In this paper, single case study method is chosen as a research method. The information about current user support governance and practices in ESGF, was captured via survey-questionnaire, participatory observation of the first author, twelve interviews with the stakeholders; mainly employees and users of ESGF and C3Grid cyber-infrastructures having different backgrounds and roles, observing relevant documents such as reports, publications and archival analysis of user and staff communication within the user's mailing list of ESGF. The triangulation of sources of information was chosen to capture different perspective to validate and to contrast the findings [Yi94], [RBP+03], [BSC12]. The current governance structure of ESGF and the recommendations made to improve the current governance of ESGF is partly based on the research methods applied.

4 Significance of the ESGF as a Case-Study

An important practical use-case in the field of climate science e-Science infrastructure is ESGF (Earth System Grid Federation) project. ESGF is the first inter-agency and international effort in the domain of Climate Science used for Earth Science Modelling (ESM) [HT05], [WBC+13], [CCM+12]. At the moment, more than twenty-seven thousand researchers are accessing climate data from ESGF distributed data-archive

worldwide that makes it a key e-Research infrastructure that supports ESM. This is the main reason to take ESGF as a use-case for this research.

Moreover, ESGF facilitates to study climate change and impact of climate change on human society and Earth's ecosystem [CCM+12]. Since physical phenomenon that govern Earth's climate are so complex and diverse, it is the most important scientific challenges of our time to undergo sophisticated model simulations that generate huge amount of data, collect observational data from various sources and share that data at a global scale. This is made possible by ESGF to discover, analyse and access the Climate data sets which are stored at multiple geographic locations across the globe [CCM+12], [Ea13], [Is13], [Vu13].

5 Current Governance Structure of ESGF

ESGF consortium is an international body of collaborating institutions, with every institute having its own norms, culture, community, specialization, components (hardware/software), human resources, standards, standing (reputation), goals and expectations. Therefore, ESGF has its own governance scheme and policy based on the principles of "governance" developed over years. Every participating institute influences it. The user support in ESGF needs an explicit governance strategy. This governance structure is comparable to general principles of governance or the "constitution" found in arena of Political Science, adhering to the principles of: Critical ideas of fairness, transparency, measurability, responsibility, accountability and performance. ESGF governance structure is composed of the following constituents, see Fig. 1 on the next page.

ESGF has a review board known as "the ESGF Review Board (ERB)" which closely coordinates and monitors the three committees: Executive Committee (EC), Steering Committee (SC) and the Technical Committee (TC), see Fig. 1. The main functions of all governance units namely ERB, EC, SC and TC are depicted in the ovals near them in the same colour. The ESGF organization and the governance is focussed on the ESGF data archive software development primarily and does not include user support process strategy making and a general support services strategy explicitly. Since the ESGF infrastructure has become mature to leverage its services to users, therefore it is important to make policy that satisfies user concerns and makes ESGF service delivery efficient and ultimately lowers the costs.

6 Proposed Changes in the Governance Scheme of ESGF

The ESSC comprises of three task-forces namely: ESGF Developers Task Force (DTF) to cater the needs of developers of ESGF, User-support Services Task Force (USTF) to cater the needs of users of ESGF and finally ESGF Data Management and Publishing Task Force (DMPTF) to cater the requirements of ESGF data managers, as shown in Fig. 2. In order to address user support concerns in the current ESGF governance scheme,

user support committee or task force (USTF) is needed to be created, consisting primarily of the user support process owner (USPO), user support coordinators (USC), members of executive committee (EC) as well as technical committee (TC). In addition, USC for scientific data modelling projects can also take part in USTF.

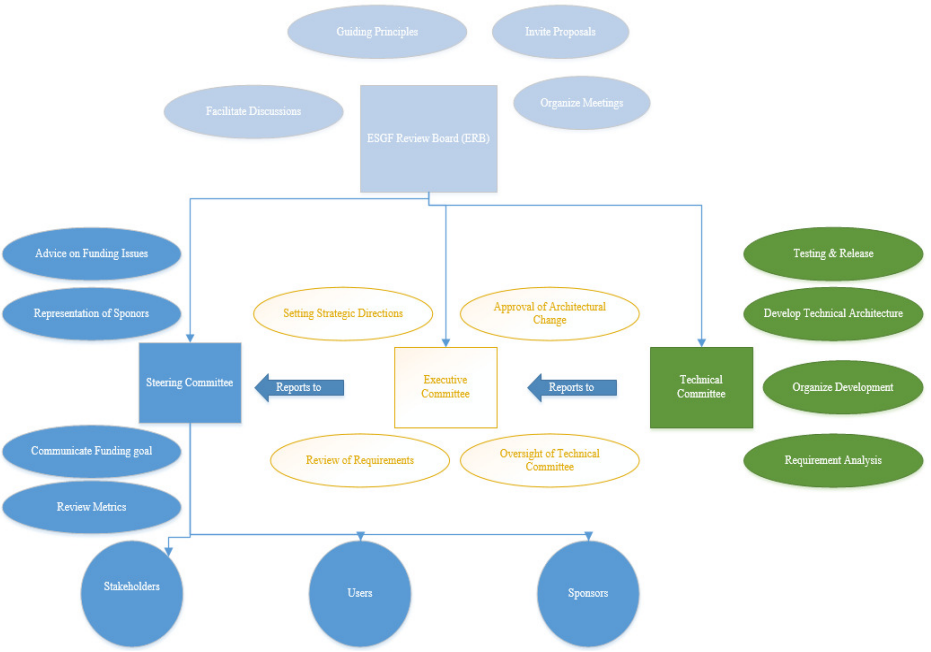


Fig. 1: The figure shows the current governance scheme of ESGF e-Science infrastructure. The squares depict the governing committees and the ovals are the main functions of each committee.

USTF is a special committee which is formed to serve the interests of users, attract new users for more usage and improve the usability (ease of use) of the ESGF. USTF interacts closely with the data publishing and management support (DPMS) and the ESGF software team. In the long run, USTF aims to have a comprehensive user support platform that covers the complete user support process, automatizing it whereever useful for the ESGF user community.

In this paper, only USTF is discussed in detail. USTF is composed of USPOs, USCs, EC members as well as TC members. All members of USTF, especially USPO, user support coordinators, first level support staff (FLS) get together at-least once a month and discuss the short-term and long-term activities for future and the progress achieved so far. USPO may liaise with other user support strategy developers in other e-Science infrastructures of other domains as well as industry.

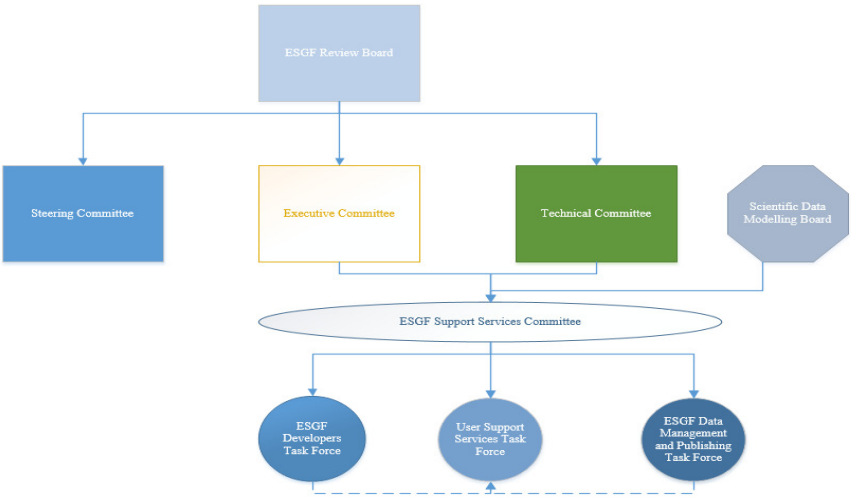


Fig. 2: The figure shows the new governance scheme where ESGF Support Services Committee (ESSC) has formed to address concerns of users of ESGF.

The USTF is a monitoring body that monitors the strategy of user support services and its implementation at the local administrative bodies or the institutes participating in ESGF consortium. Furthermore, USTF adapts the user support services to the changing needs of the e-Science infrastructure. The manifesto of USTF includes defining norms and standards in the form of SLA's[2] and OLA's[3] to make process reliable and monitoring these agreements at regular intervals. USTF must also look for adequate funding from the sponsors and set budget for user support services. The funding must be then invested in an efficient manner. In addition, USTF must investigate the need of providing automation in the user support process necessary.

Categories of user issues must be sorted out by unanimously agreeing on the categorization standards. All known and unknown problems must be documented. If possible a data-base can be created. These problems and their corresponding categories can then be shared with other e-Science infrastructures as well. It is useful to present a list of all problems that users encounter and the list can discussed in the meetings. Finally, a strategy can be made to reduce the user issues by solving the underlying root causes. The global user support directory (USD) must be maintained and updated at regular intervals. The USD will include all the FLS staffs, SLS staffs along with their components for which they are responsible and accountable.

It is vital to invite and encourage users to provide feedback on the ESGF operations and services. These suggestions may lead to inclusion of new features i.e. user driven development and enhancement of ESGF as well as its user support. In addition, short survey to capture user and support staffs satisfaction level must be conducted at regular

[2] Service Level Agreement (SLA) is an agreement between users who use the service and the service providers.
[3] Operation Level Agreement (OLA) is an agreement between different service providers within an organisation who provide operational services under the same organisation.

intervals. The users can be involved in training sessions, workshop, meetings, conferences and teleconferences at local as well as global levels. If researchers and open-source contributors are attracted to the investigation and development of user support process of ESGF, usability and scalability of ESGF can be improved drastically. Furthermore, data modelling centers can be contacted to provide more data as well as data support in the ESGF data archive system.

Possible incentives to the users and the support staffs may be introduced in order to improve the user support process. Responsibility and accountability must be allocated to staffs. To make the user support process interesting, gamification techniques can be applied. Similarly, self-help can be promoted by disseminating understandable task-oriented support information to the users via ESGF web-pages, wikis, in the form of documents and online tutorials. Moreover, user statistics must be collected and communicated to interested stakeholders of ESGF. Eye-tracking technologies can be applied to evaluate and enhance UI components of ESGF. must be proposed in the ESGF system. Commercial access may be granted to the commercial users with some fees. For instance insurance companies can use climate data. Finally, appropriate metrics to measure the quality of user support must be introduced.

7 Conclusion and Future Work

In this paper, the need to introduce improved user support process strategies in e-Science infrastructures are highlighted. These policies may vary from one e-Science infrastructure to another depending on their scientific domain and architecture. But these policies must be implemented in the governing scheme of e-Science infrastructures to enhance user support process; thus ultimately improving service desk function within e-Science infrastructures. More work is needed to be done to introduce a stable user support for the users of e-Science infrastructure. There is a need to understand importance of user support process and its inclusion in the overall governance in e-Research. This shall bring value to over-all productivity of the e-Science infrastructures and their stakeholders, particularly users.

References

[AP94] Agnar Aamodt and Enric Plaza. Case-Based Reasoning (CBR): Foundational issues, methodological variations, and system approaches. AI communications, 7(1):39–59, 1994.

[BBB+05] David Bernholdt, Shishir Bharathi, David Brown, Kasidit Chanchio, Meili Chen, Ann Chervenak, Luca Cinquini, Bob Drach, Ian Foster, Peter Fox et al. The earth system grid: Supporting the next generation of climate modeling research. Proceedings of the IEEE, 93(3): 485–495, 2005.

[BSC12] David A. Buchanan, G. Symon, and C. Cassel. Case studies in organisational research. Qualitative Organisational Research: Core Methods and Current Challenges, 2012.

[CCM+12] Luca Cinquini, Daniel Crichton, Chris Mattmann, John Harney, Galen Shipman, Feiyi Wang, Neil Miller et al. The earth system grid federation: An open infrastructure for access to distributed geospatial data. IEEE 8th International Conference on E-Science, 1–10, 2012.

[CBL14] Hashim Iqbal Chunpir, Amgad Badewi, and Thomas Ludwig. User support system in the complex environment. Human Computer Interaction-HCI International, 11:392-402, 2014.

[CLB14a] Hashim Iqbal Chunpir, Thomas Ludwig, and Amgad Badewi. A snap-shot of user support services in Earth System Grid Federation (ESGF): A use case of climate cyber- infrastructures. Proceedings of the Fifth International Conference on Applied Human Factors in Ergonomics (AHFE), 2014.

[CLB14b] Hashim Iqbal Chunpir, Thomas Ludwig, and Amgad Ali Badewi. Using Soft Systems Methodology (SSM) in understanding current user-support scenario in the climate science domain of cyber-infrastructures. Human Computer Interaction-HCI International, 10:495-506, 2014.

[Ch13] Hashim Chunpir. C3Grid-INAD: Support centre with semi-automatic notification and coordination for experts work. 2013.

[CWL13] Hashim Chunpir, Dean Williams, Luca Cinquini et al. 3rd Annual Earth System Grid Federation and Ultrascale Visualization Climate Data Analysis Tools Face-to-Face Meeting Report. December 2013.

[Cz98] Barbara Czegel. Help desk practitioner's handbook. John Wiley, 1998.

[DR04] Garett Dworman and Stephanie Rosenbaum. Helping users to use help: Improving interaction with help systems. 1717–1718, 2004 — ISBN 1581137036.

[Ea13] Earth System Grid Federation: ESGF. URL http://esgf.org/. - last accessed on 2013-06-10.

[Fr07] Peter A. Freeman. Is it possible to define cyberinfrastructure? First Monday, 12(6), 2007.

[GGR05] Luz Minerva González,Ronald E. Giachetti, and Guillermo Ramirez. Knowledge management-centric help desk: specification and performance evaluation. Decision Support Systems, 40(2): 389–405, 2005.

[Gu04] Erik Guldentops. Structures, processes and relational mechanisms for IT governance. Strategies for information technology governance. 1–36, 2004.

[HT05] Tony Hey and Anne E. Trefethen. Cyberinfrastructure for e-Science. Science (New York, N.Y.). 308(5723): 817–21, 2005.

[Is13] IS-ENES: IS-ENES portal. URL https://verc.enes.org/. - last accessed on 2013-08-22.

[Jä09] Marko Jäntti. Lessons learnt from the improvement of customer support processes: A case study on incident management. Proceedings of 10th International Conference, PROFES. 2009.

[Ke02] Howard Kendall: “Prehistoric Help Desk!!”. Support World. Help Desk Institute, 2002.

[LL07] N. K. Y. Leung and S. K. Lau. Information technology help desk survey: To identify the classification of simple and routine enquiries. Journal of Computer Information Systems 2007.

[MM96] Rita Marcella and Iain Middleton. The role of the help desk in the strategic management of information systems. OCLC Systems & Services, 12(4):4–19, 1996.

[MZ05] Yuval Marom and Ingrid Zukerman. Analysis and synthesis of help-desk responses, 890–897, 2005.

[RBP+03] Tonette S. Rocco, Linda A. Bliss, Aixa Pérez-Prado, and S. Gallagher. Taking the next step: Mixed methods research. Organizational Systems, 21(1): 19–29, 2003.

[Ro04] T. R. Roth-Berghofer. Learning from HOMER, a case-based help desk support system. Advances in Learning Software Organizations: 88–97, 2004.

[SSW10] Catherine Soehner, Catherine Steeves, and Jennifer Ward. E-Science and data support services. 2010.

[TSM12] Karl E. Taylor, Ronald J. Stouffer, and Gerald Meehl. An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society, 93(4): 485–498, 2012.

[UHG03] John A. Underwood, Deb Hegdahl, and Jeff Gimbel. A proper set of tools are needed to corral support. Proceedings of the 31st annual ACM SIGUCCS conference on User services - SIGUCCS: 23–26, 2003.

[Vu13] Linda Vu. Earth System Grid Federation (ESGF): A modern day “Silk Road” for climate data. URL https://es.net/news-and-publications/esnet-news/2012/ESGF/.— Energy Science Network - last accessed on 2013-08-23.

[WBC+13] Dean N. Williams, Gavin Bell, Luca Cinquini, Peter Fox, John Harney and, Robin Goldstone. Earth System Grid Federation (ESGF): Federated and integrated climate data from multiple sources, 6:61–77. 2013.

[Wi12] Dean N. Williams. Earth System Grid Federation (ESGF): Future and governance World Climate Research Programme (WCRP), Working Group on Coupled Modelling (WGCM)— Stakeholders and ESGF:1–17, 2012.

[Yi94] R.Yin. Case study research: Design and methods. Sage Publishing, 1994.

[ZM06] Ingrid Zukerman and Yuval Marom. A comparative study of information-gathering approaches for answering help-desk e-mail inquiries: 546–556, 2006.

Scientometrics: How to perform a Big Data Trend Analysis with ScienceMiner

Volker Frehe, Vilius Rugaitis, Frank Teuteberg

Osnabrück University
Accounting and Information Systems
Katharinenstr. 1
49074 Osnabrück
volker.frehe@uni-osnabrueck.de
vrugaiti@uni-osnabrueck.de
frank.teuteberg@uni-osnabrueck.de

Abstract: This paper describes the results of the implementation of an application that was designed under the design science principles. The purpose of this application is to identify trends in science. First, the status quo of similar applications as well as the knowledge base about data mining in the field of scientometrics is analyzed. Afterwards, the implementation as well as the evaluation of our application is described. Our web-based application allows to search for contributions (literature and internet, e.g., twitter, news), executes several data mining methods and visualizes the results in seven different ways. Each visualization has some filters and further control elements. It is the first application to provide the complete process from data acquisition to data visualization in an automated way.

1 Motivation

Independent of the research field, the literature review is an important and essential yet time-consuming method to gather the status quo in science. There are several indices, like the h-index [Hi05], by means of which authors can be rated and distinguished authors and literature can be identified. It is a broadly accepted method to separate relevant from irrelevant literature by means of the various variants of the h-index, e.g., the one for institutions [Ki07] or else completely new variants like the g-index [Eg06] that is also based on the h-index. But scientific knowledge is not only distributed in literature, it can also be found in the internet, e.g., in social networks like Twitter, Facebook, etc. As the knowledge base continues to increase, new methods need to be developed to capture it. There are already automated methods from the field of information retrieval (IR), that are used in scientific knowledge capturing, like co-classification [AG10] and co-word analysis [DCF01, Le08]. Moreover, it has been proven that automated citation analysis is able to reduce the workload of the scientists [Co06, Ma10].

Therefore, automation and the use of IR and data mining methods seem to be indispensable in the field of scientific research, especially for knowledge discovery.

This contribution describes the process of the development of an application for big data trend analysis in the area of science. As the process follows the design science principles [He04], for the implementation, we use a great knowledge base deriving from a literature review as well as information from the environment (e.g., from existing applications, round tables and surveys). There are already several similar applications, but our application is the first to provide the complete process from data acquisition to data visualization in an automated way.

The paper is structured as follows. At first, the literature search and similar applications are described. Afterwards, in section 3, the research methodology is explained. Section 4 provides an extensive literature review, and section 5 contains information about the implementation of the application. Before closing the paper with a conclusion, section 6 shows the results of the evaluation among scientists.

2 Status Quo and Related Work

A systematic literature review based on [WW02] was performed to gain information about the topic's trend analysis and scientometrics. We used several databases with specific search terms to get a broad body of literature. Following the guideline of [Br09] for transparency of the search process, the complete list of databases, search words and results can be downloaded[1]. Overall, our search resulted in 2,674 contributions. As the journal Scientometrics is solely about the "science of science" and thus is very important for our research, we also investigated 1,190 additional contributions from this journal. In order to consolidate the huge amount of 3,864 contributions, we first eliminated all non-academic contributions. Further, we sorted out the irrelevant papers by reading the titles, which resulted in a list of 594 contributions. In addition, we accepted papers exclusively in English and removed contributions without relevance to the topic by reading the abstracts. Ultimately, we used 289 contributions for the analysis.

Our literature search revealed several applications with some kind of scientometric analysis. The list of identified applications is shown in table 1, which also provides information about the type of the data that the application can analyze. We divided it into bibliometric data, altmetric data or other scientometric data. Most applications access bibliometric data, however, only few use altmetrics. The most common visualization methods are tables and diagrams, whereas methods like tag clouds, world maps or heat maps are under-represented. A detailed analysis of the applications can be downloaded[2].

Our application differs from these applications as, to the best of our knowledge, it is the first to comprise all functions, from data acquisition over data preparation and analysis to data visualization in form of a user-friendly web application.

[1] Literature Search Details & Results, http://uwi.uos.de/att/SM-LiteratureSearch.pdf [last access 27.06.2014]
[2] Investigated Tools, http://uwi.uos.de/att/SM-Tools.pdf [last access 27.06.2014]

	Type			Visualization								
Tool	**Bibliometrics**	**Scientometrics**	**Altmetrics**	**Table**	**Diagram**	**Landscape Map**	**Heat Map**	**Tag Cloud**	**World Map**	**Stream Graph**	**Strategy/ strategic map**	**Network Map**
Bibexcel	x											
CiteSpace	x				x				x			x
CoPalRed				x	x						x	x
IN-SPIRE				x	x	x				x		x
Leydesdorff's Software	x											
Network Workbench Tool		x			x							x
Science of Science (Sci²) Tool	x			x	x				x			x
VantagePoint	x			x	x							x
VOSViewer						x						x
Sitkis	x											
BiblioTools	x											
SAINT	x		x									
SciMAT	x			x							x	x
CATAR	x											
TEXTREND	x				x							x
ImpactStory			x	x								
Altmetric			x	x	x		x		x			
SciCombinator			x	x				x				
PlumX			x	x	x							

Table 1: Investigated Applications and Visualization Methods

3 Research Method

Our research is following the design science [He04] method, as we want to include scientific knowledge as well as information from the environment in the development, as shown in figure 1.

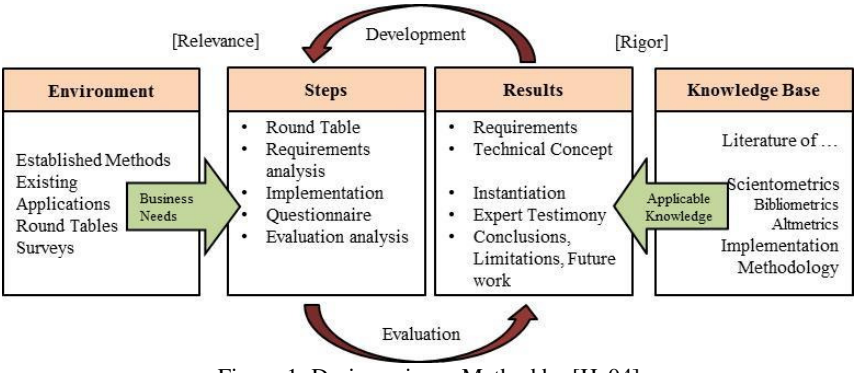


Figure 1: Design science Method by [He04]

The seven design science principles [He04] are met, as shown in table 2.

Guideline	Description
Design as an Artifact	The application, which is our developed artifact, follows the definition of [MS95].
Problem Relevance	The relevance of the topic is already mentioned in section 1 and further worked out in section 4.
Design Evaluation	At the moment, there is already a first evaluation (cf. section 6); further evaluations will follow.
Research Contributions	Our application is the first of this kind, providing all functions from data acquisition over data preparation and analysis to data visualization in form of a user-friendly web application.
Research Rigor	In our research, we follow the methodological requirements for literature research [WW02], design science [He04, MS95] and the evaluation by survey [My09]
Design as a Search Process	The iterative search process is shown in figure 1.
Communication of Research	The web application itself is free for everyone[3]. The results will be presented to the scientific community inter alia via this contribution.

Table 2: Design science Guidelines [He04]

4 Literature Review

Our literature search resulted in 289 contributions. We also conducted a cluster analysis via Rapidminer[4] and used the title, abstract and keywords of each contribution for clustering. Common stop words, as well as an extra list of 71 stop words, were eliminated from the list of words. We equally eliminated words with less than 4 and more than 25 characters. After the selection of the valid words, a word stemming has been performed. The cluster analysis was performed by a k-means algorithm and resulted in 9 clusters as displayed in table 3.

Most papers belong to clusters 6, 7 and 3, which represent some kind of bibliometric citation analysis (cluster 6), trend detection (cluster 7) and co-word/co-citation analysis. In the following, we want to get a deeper insight in each cluster.

Cluster 0 is about indices. Radicchi and Castellano (2013) investigate the relationship of the h-index and the amount of publications and citations of a scientist [FC13]. They detect a weak connection between amount of publications and h-index and a strong correlation between the h-index and the amount of citations. A study about the influence of age, field and uncitedness on author ranking is performed by [Am12]. It shows that the age and the field of the scientist have great influence on the ranking. **Cluster 1** is about social network analysis (SNA), like the contributions of [No12], who develops a method to detect trends in social networks (SN). Therefore, several self-organizing maps

[3] ScienceMiner, http://scienceminer.uwi.uos.de [last access 27.06.2014]
[4] Predictive Analytics, Data Mining, Self-service, Open source – RapidMiner, http://rapidminer.com [last access 27.06.2014]

(SOM) are created at various times who visualize the results of a cluster analysis. These SOMs can be used to identify trends in social networks.

Cluster	Description	# of Papers
0	h; h_index; index; indic; type; citat; number; public; individu; evalu; measur; scientist	23
1	network; social; social_network; mine; data; futur; analysi + analyz; domain; user; relationship; knowledg; techniqu; pattern	17
2	technolog; literatur; trend; bibliometr; emerg; analysi + analyz; model; bibliometr_analysi; appli + applic; system; network; develop; research; case; studi; citat	29
3	co; cluster; map; network; analysi + analyz; knowledg; structur; document; citat; tool; china; scienc; research; keyword; field	43
4	collabor; intern; scientif; author; institut; countri; level; network; china; nation; paper; co; field; scienc; analysi + analyz; public; bibliometr; output; impact; research; pattern; year	31
5	factor; journal; impact; citat; publish; cite; articl; paper; individu; author; web; discuss; effect	12
6	citat; public; countri; research; cite; scienc; bibliometr; output; evalu; indic; articl; product; analysi + analyz; number; scientif; scientist; area; disciplin; knowledg; journal; paper; assess; publish	68
7	detect; topic; user; trend; emerg; model; collect; data; social; inform; search; propos; approach; cluster; method; interest; dynam; system; appli + applic; retriev; analysi + analyz	44
8	author; rank; univers; citat; co; approach; research; measur; base; bibliometr; assess; compar; index; topic; propos	22

Table 3: Cluster description

Cluster 2 is also about trend analysis in networks, but from a more bibliometric point of view. Guille (2013) indicate that the mentioning frequencies (e.g., re-tweets) are a better indicator for the popularity of a topic than the global frequency of a topic [Gu13]. These indicators can be used to measure information diffusion in SN. Khan et al. (2011) created a concept (network of core, based on the mathematical graph theory), to discover hidden structures in scientific networks by the visualization of theoretical constructs, models and concepts of a specific scientific domain through a network [KMP11]. **Cluster 3** is mainly about co-word analysis und co-citation analysis. An analysis of co-citation performance of six retrieval methods has been conducted by [Et12]. A positive effect on performance could be found by using the co-citation context and the normalization technique of cited frequency. Yang et al. (2012) have combined several visualization techniques (cluster tree, strategy diagram and social network maps) of the co-word analysis to use the advantages of each technique and to weaken the disadvantages [YWL12]. A problem in the field of co-word analysis is the use of keywords as a weak point of literature search [NPS13, Wa12]. Solutions are to use the Knowledge Discovery Process (KDP) for cluster analysis of all available contribution data [NPS13] or to integrate expert knowledge into the co-word analysis in form of a new method, the semantic based co-word analysis [Wa12]. **Cluster 4** deals with the collaboration of scientists. Gazni et al. (2012) have investigated that collaborations between authors, institutions and countries have gradually increased in the past years

[GSD12], which indicates the importance of this topic. He et al. (2011) explore co-author networks via a context subgraph [HDN11]. Through this subgraph quantitative factors should be obtained by the integration of the author's background in the analysis. **Cluster 5** is concerned with journal impact factors. Vanclay (2012) critically study the Thomson Reuters Impact Factor (TRIF) and demonstrates the influence of wrong links, misspelling, missing cites and advocate a complete overhaul of the TRIF [Va12]. Thelwall (2012) additionally request for adding new indicators (altmetrics), like online readership indicators, social bookmarking indicators, link analysis, web citations and Twitter (tweets) in order to enhance the bibliometric indicators [Th12]. To avoid manipulation, a mixture of several indicators should be used. **Cluster 6** is about citation analysis. Franceschet (2009) conducted a correlation analysis to reduce quantitative, bibliometric indicators for scientist assessment [Fr09]. The analysis includes 13 indicators. The amount of papers (for productivity assessment), the amount of citations (for impact assessment), the average citation amount per paper (for relative impact assessment) and the m-quotient (for long-term impact assessment) are identified as the most important indicators. **Cluster 7** deals with trend analysis. Tseng et al. (2009) investigate several trend indices [Ts09]. It was figured out that the linear regression is best for timeline analysis, which supports the extensive usage of this method. Guo et al. (2011) use several indicators (increase of specific word usage, amount of new authors in research field and amount of interdisciplinary citations) in a mixed model [GWB11]. Their research indicates that new authors explore a new research field first, then, they reference interdisciplinary literature before they use some specific words more often. Through this information, new trends and hot topics can be identified. **Cluster 8** covers author rankings. Wang et al. (2012) identify that the co-citation analysis can also be used to recognize research patterns, find research communities and is in a position to identify hot topics in science [WQY12]. Ding (2011) criticizes that author rankings are field independent [Di11]. He proposes a new ranking which includes the authors' fields (topic-based PageRank for authors). The author-conference-topic model (ACT) is used to gain information about the authors' fields and it is integrated with the PageRank algorithm to enable a field dependent author ranking. The results of the literature review have been used in the conceptual phase of the implementation of our web application.

5 Implementation of the Prototype

The framework of the application and the interaction of the several modules are displayed in figure 2. As our framework is built on a modular basis, enhancements are possible in every step (e.g., adding new data sources or mining/visualization methods).

The developed artifact is a web application for automated trend detection via bibliometric and altmetric analysis. We follow the Knowledge Discovery in Databases (KDD) process of [FPS96], which consists of the steps selection, preprocessing, transformation, data mining and interpretation. Therefore, the web application is designed as user-friendly as possible. At first, the user states a keyword for a topic to search for. The application will execute the next steps in the background so that the user gets a result of the process in form of some visualizations.

The first step is the selection of data. Therefore, we integrated several data sources from the internet which are accessed through application programming interfaces (API). Because of usage and technical restrictions or bad quality of data, our prototype has access to Microsoft Academic Search[5] as source for bibliometric data and the service altmetric[6] for altmetric data. The service altmetric combines access to several sources like Facebook, Google+, Twitter, Reddit and several blogs and news sites. The data selection is performed via a batch process on the server side. This allows the user to state a query and leave the web application while the search query is executed in the background. This approach provides flexibility for the end-user, since most of the data sources suffer of technical and legal restrictions, which lead to a long execution time. This way the time-consuming queries can be initiated and then executed in the background without the need of permanent user presence. When logging in again, the user has access to all his executed queries. The batch process also enables multi-threading and parallel processing of various queries, which enhances the performance. The batch process also performs the second step of the KDD process (preprocessing). Irrelevant words (stop words) are eliminated, a word stemming is executed and synonyms are combined through the use of a dictionary. The mainly utilized entities like users, administrators, contributions and dictionaries in form of a Unified Modeling Language class diagram can be downloaded[7].

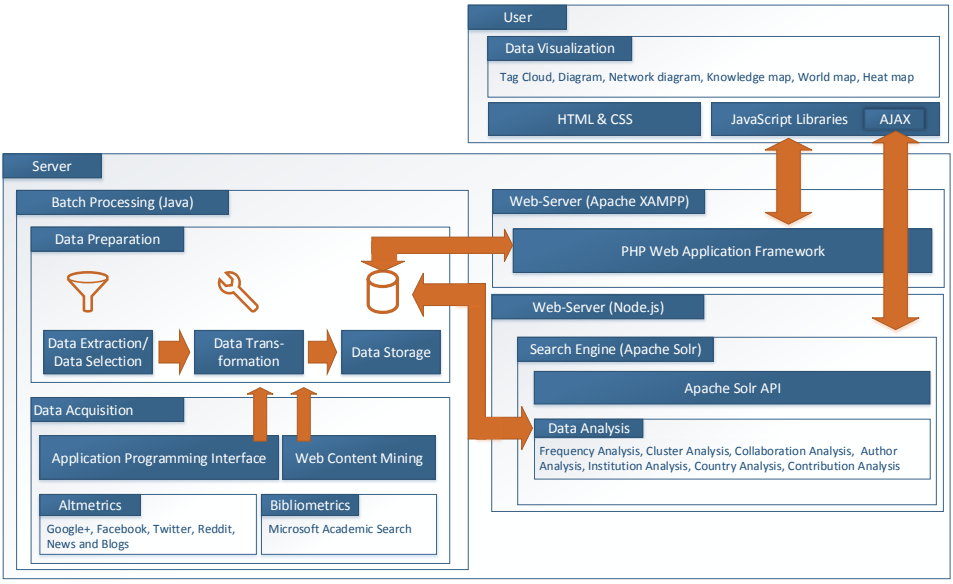


Figure 2: Architecture of application

After the data selection, the next step is the transformation which is also performed on the server. The contributions' data enriched with altmetric data are converted to the

[5] Microsoft Academic Search, http://academic.research.microsoft.com [last access 27.06.2014]
[6] Altmetric API documentation, http://api.altmetric.com [last access 27.06.2014]
[7] ScienceMiner UML diagram, http://uwi.uos.de/att/SM-UML.pdf [last access 27.06.2014]

needed format, if necessary merged and stored in a relational database. This is the last step of the batch process.

The next KDD process step (data mining) is done by Apache Solr[10] on the server. This product is suitable due to the provision of advanced text analysis methods, fast response times, import and export functionalities and enhancement possibilities. Because of performance reasons, the data is imported to Solr and is not analyzed in the relational database. This provides the flexibility which is needed for the interactive visualization of the results. However, Solr does not provide any security mechanisms for the data exchange. This is why we decided to use Node.js[11] as proxy server for Solr to handle the access. At the moment, only clustering and frequency analysis is used for data mining.
The last step of the KDD process (interpretation) involves the user again. The web application provides HTML and JavaScript functionalities that communicate with the web server via Asynchronous JavaScript and XML (AJAX). Several visualization possibilities are given, which aid the user interpreting the results.

The most important part for the user is the visualization of the results. There are several methods provided to display the mining results. An example of the user interface with a result of the query "Scientometrics" is shown in figure 3. At first, there is general information providing an overview of the data gathered by the query (e.g., how many publications and altmetric data have been found, the date of the first and last publication, etc.).

As our literature review reveals numerous visualization techniques, our application implements several of them. The tag cloud provides an overview about the most-relevant terms, keywords, authors, etc. The diagram allows to show a timeline of the publication dates and to also visually view authors, affiliations, etc. as well as the amount of their publications. The network map is a construct to visualize the connections between entities like authors, countries and affiliations. The topic map enables to cluster the contributions and show main topics and the associated keywords. The world map is a construct by which the origin (and amount) of the contributions is displayed on a world map. The heat map (cf. figure 3) shows the diffusion of several topics over time. Each visualization element has some controls. There are controls to specify the timeframe, choose the element to be analyzed (e.g. author vs. affiliation), specify the amount of elements to be shown, etc. Depending on the visualization element, the respective controls are depicted. A complete overview of all visualization elements can be downloaded[12].

Every method can be displayed or hidden and also the order of the methods can be changed. The left navigation panel can also be hidden in order to use the available space for the visualization elements.

[10] Apache Lucene - Apache Solr, https://lucene.apache.org/solr [last access 27.06.2014]
[11] node.js, http://nodejs.org [last access 27.06.2014]
[12] ScienceMiner Visualization, http://uwi.uos.de/att/SM-Visualization.pdf [last access 27.06.2014]

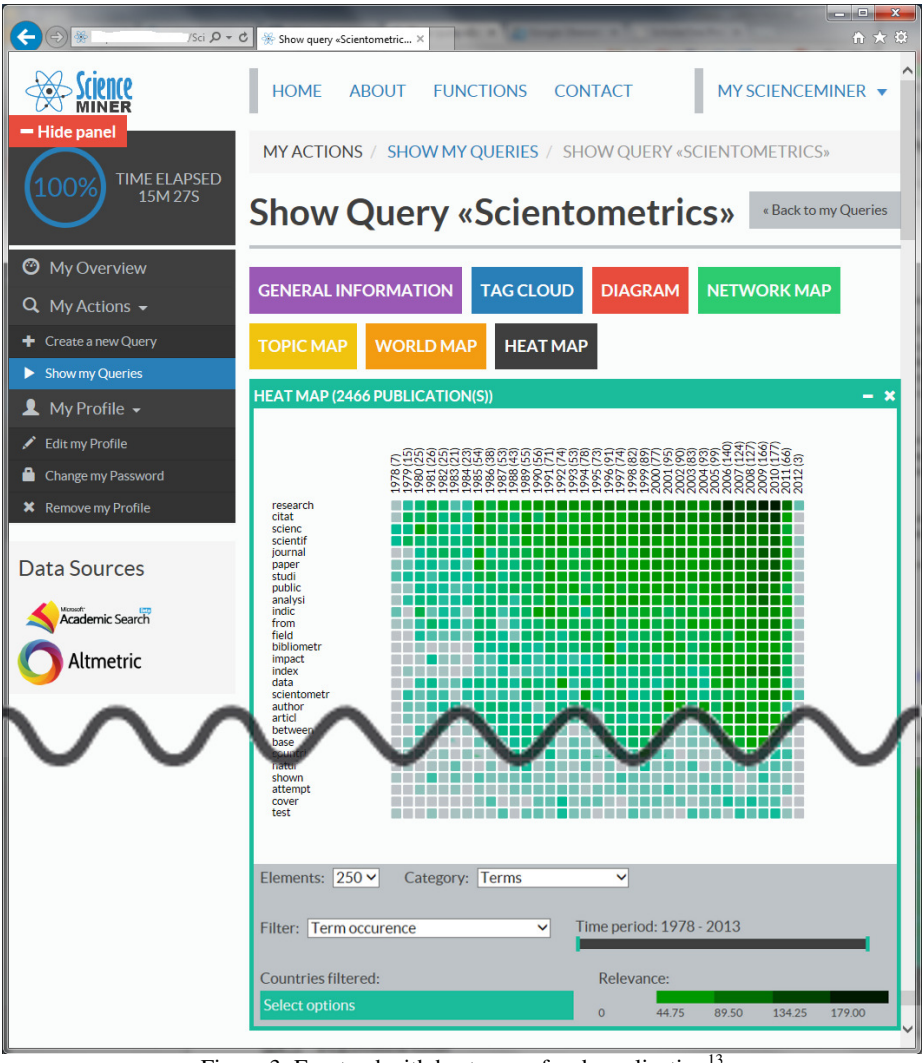


Figure 3: Frontend with heat map of web application[13]

6 Evaluation

After the experimentation phase, we invited 40 scientists and young researchers via e-mail to take part in an evaluation of the web application. We asked them to use the application and fill in an online survey. Apart from the integrated online help, no further support was given. Up to now, 14 of the invited scientists and researchers have completed the survey. The average age of the participants is 27.8 and all are male. Four of them are students, two graduates, seven research assistants and one professor. Of the

[13] The curved line indicates that we merged two screenshots into one.

respondents, 71.5% come from the IS field, 21.5% from the field of economics and 7% from other fields. The survey consists of 7-point Likert scale questions as well as free text fields for notes and recommendations. The Likert Scale reaches from "strongly agree" to "strongly disagree". The questions are grouped into clusters to evaluate the design, the content, the usability and the functionality of the application as well as to raise general questions about bibliometrics and altmetrics. Figure 4 shows one sample question for each category and the associated results. The survey shows good results for the design, the content, the usability and the functionality of the application. The bibliometrics and altmetrics seem to be accepted methods by scientists, but only in addition to other methods (see next paragraph). The complete survey consists of 46 questions; the results are comparable with the ones mentioned here. As the survey has not been concluded yet, the presented results only serve the purpose of giving first insights.

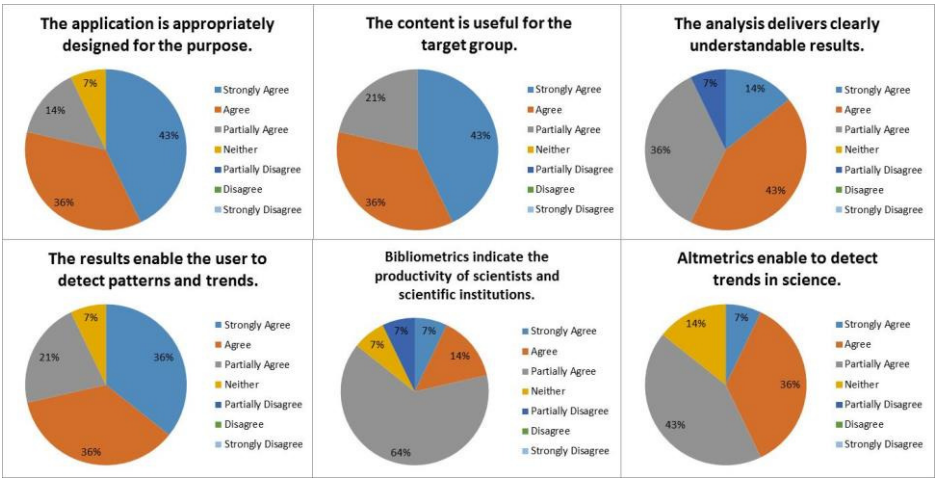


Figure 4: Survey results

However, already the annotations received so far provide some valuable elements of improvement for the application. Most people still perceive the qualitative review to be indispensable. According to them, the bibliometric and altmetric analyses can only be used in a subsidiary manner or just to identify relevant literature. Although our application is deemed useful, there are also some improvement suggestions, for instance, to integrate a spellchecker in the research as well as the inclusion of acronyms in the search. As the search is time-consuming, apparently there is a need for some kind of fast pre-search. Furthermore, some comments referred to the wish, that not only the abstracts should be investigated, but the entire contributions. Additionally, more search engines (like Google Scholar) should be integrated to obtain more results. One person asked for a list of all identified publications. However, this feature can due to legal restrictions not be integrated as it would be an imitation of the search engine's functionality. Two people asked for a comparison of two search results. Also, more visualization methods were wanted as well as the possibility to export the results. If procurable, all these recommendations will be implemented to further improve the application in the aim of design science (cf. section 3).

7 Conclusion and Future Work

Following the design science principles, the developed application proves how theoretical knowledge from scientometrics and data mining theories can be used in a practical way. The application can be used by scientists to get new insights into several fields of their research. The evaluation indicates that the application is practicable and useful. However, the automated data mining should only be used in addition to traditional literature research methods. Nevertheless, the developed application can be seen as an enhancement to the traditional methods and although it prods to new trends and discovers undetected contributions by the use of not only scientific contributions, but also information from the web (like Facebook, Twitter, etc). We are well aware of the fact, that our application has only been evaluated by 14 people so far, which represents a limitation. However, with this contribution we pursue the goal of stimulating a broad use of our prototype. Thereby, more scientists might work with it and we might obtain further meaningful recommendations from the science community.

Acknowledgments

The authors would like to thank the anonymous reviewers and Ms. Marita Imhorst, who provided valuable insights, help and substantive feedback during the research process.

References

[AG10] M.-R. Amini and C. Goutte: A co-classification approach to learning from multilingual corpora. Machine Learning, 79(1–2):105–121, 2010.

[Am12] L. Amez: Citation Measures at the Micro Level: Influence of Publication Age , Field , and Uncitedness. Journal of the American Society for Information Science and Technology, 63(7):1459–1465, 2012.

[Br09] J. vom Brocke, A. Simons, B. Niehaves, K. Riemer, R. Plattlauf and A. Cleven: Reconstructing the Giant: On the Importance of Rigour in Documenting. ECIS 2009 Proceedings, 2009.

[Co06] A. M. Cohen, W. R. Hersh, K. Peterson and P.-Y. Yen: Reducing workload in systematic review preparation using automated citation classification.Journal of the American Medical Informatics Association, 13(2):206–19, 2006.

[DCF01] Y. Ding, G. G. Chowdhury and S. Foo: Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing & Management, 37(6):817–842, 2001.

[Di11] Y. Ding: Topic-Based PageRank on Author Cocitation Networks. Journal of the American Society for Information Science and Technology, 62(3):449–466, 2011.

[Eg06] L. Egghe: Theory and practise of the g-index. Scientometrics, 69(1):131–152, 2006.

[Et12] M. Eto: Evaluations of context-based co-citation searching. Scientometrics, 94(2):651–673, 2012.

[FC13] F. Radicchi and C. Castellano: Analysis of bibliometric indicators for individual scholars in a large data set. Scientometrics, 97(3):627–637, 2013.

[FPS96] U. M. Fayyad, G. Piatetsky-Shapiro and P. Smyth: From data mining to knowledge discovery: An overview. (U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R.

Uthurusamy Eds.): Advances in knowledge discovery and data mining, pp. 1–34, AAAI Press, Menlo Park, 1996.
[Fr09] M. Franceschet: A Cluster Analysis of Scholar and Journal Bibliometric Indicators. Journal of the American Society for Information Science and Technology, 60(10):1950–1964, 2009.
[GSD12] A. Gazni, C. R. Sugimoto and F. Didegah: Mapping World Scientific Collaboration: Authors, Institutions, and Countries. Journal of the American Society for Information Science and Technology, 63(2):323–335, 2012.
[Gu13] A. Guille: Information Diffusion in Online Social Networks. SIGMOD Records, 42(2):17-28, 2013
[GWB11] H. Guo, S. Weingart and K. Börner: Mixed-indicators model for identifying emerging research areas. Scientometrics, 89(1):421–435, 2011.
[HDN11] B. He, Y. Ding and C. Ni: Mining Enriched Contextual Information of Scientific Collaboration?: A Meso Perspective. Journal of the American Society for Information Science and Technology. 62(5):831–845, 2011.
[He04] A. R. Hevner, S. T. March, J. Park and S. Ram: Design science in information systems research. MIS Quarterly, 28(1):75–105, 2004.
[Hi05] J. E. Hirsch: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46):16569–16572, 2005.
[Ki07] A. L. Kinney: National scientific facilities and their science impact on nonbiomedical research. Proceedings of the National Academy of Sciences of the United States of America, 104(46):17943–17947, 2007.
[KMP11] G. F. Khan, J. Moon and H. W. Park: Network of the core: mapping and visualizing the core of scientific domains. Scientometrics, 89(3):759–779, 2011.
[Le08] W. H. Lee: How to identify emerging research fields using scientometrics: An example in the field of Information Security. Scientometrics, 76(3): 503–525, 2008.
[Ma10] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza and P. O'Blenis: A new algorithm for reducing the workload of experts in performing systematic reviews. Journal of the American Medical Informatics, 17:446–453, 2010.
[MS95] S. T. March and G. F. Smith: Design and natural science research on information technology. Decision Support Systems, 15(4): 251–266, 1995.
[My09] M. D. Myers: Qualitative Research in Business & Management. London, 2009.
[No12] P. N. E. Nohuddin, F. Coenen, R. Christley, C. Setzkorn, Y. Patel and S. Williams: Finding "interesting" trends in social networks using frequent pattern mining and self organizing maps. Knowledge-Based Systems, 29:104–113, 2012.
[NPS13] P. Nieminen, I. Pölönen and T. Sipola: Research literature clustering using diffusion maps. Journal of Informetrics, 7(4):874–886, 2013.
[Th12] M. Thelwall: Journal impact evaluation: a webometric perspective. Scientometrics, 92(2): 429–441, 2012.
[Ts09] Y.-H. Tseng, Y.-I. Lin, Y.-Y. Lee, W.-C. Hung and C.-H. Lee: A comparison of methods for detecting hot topics. , 81(1):73–90, 2009.
[Va12] J. K. Vanclay: Impact factor: outdated artefact or stepping-stone to journal certification? Scientometrics, 92(2):211–238, 2012.
[Wa12] Z.-Y. Wang, G. Li, C.-Y. Li and A. Li: Research on the semantic-based co-word analysis. Scientometrics, 90(3): 855–875, 2012.
[WQY12] F. Wang, J. Qiu and H. Yu: Research on the cross-citation relationship of core authors in scientometrics. Scientometrics, 91(3):1011–1033, 2012.
[WW02] J. Webster and R. T. Watson: Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26(2):xiii–xxiii, 2002.
[YWL12] Y. Yang, M. Wu and L. Cui: Integration of three visualization methods based on co-word analysis. Scientometrics, 90(2):659–673, 2012.

Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)

Michael Diepenbroek, MARUM, Universität Bremen
Frank Oliver Glöckner, Jacobs University and MPI für Marine Mikrobiologie
Peter Grobe, Zoologisches Forschungsmuseum König, Bonn
Anton Güntsch, Botanischer Garten und Botanisches Museum Berlin-Dahlem
Robert Huber, MARUM, Universität Bremen
Birgitta König-Ries, Universität Jena and iDiv
Ivaylo Kostadinov, Jacobs Unversity
Jens Nieschulze, Universität Göttingen
Bernhard Seeger, Universität Marburg
Robert Tolksdorf, Freie Universität Berlin
Dagmar Triebel, Staatl. Naturwissenschaftl. Sammlungen Bayerns

Abstract: Biodiversity research brings together the many facets of biological environmental research. Its data management is characterized by integration and is particularly challenging due to the large volume and tremendous heterogeneity of the data. At the same time, it is particularly important: A lot of the data is not reproducible. Once it is gone, potential knowledge that could have been gained from it is irrevocably lost. In this paper, we describe challenges to biodiversity data management along the data life cycle and sketch the solution that is currently being developed within the GFBio project, a collaborative effort of nineteen German research institutions ranging from museums and archives to biodiversity researchers and computer scientists.

1 Introduction

Environmental and biological research is becoming central to major societal challenges related to the Earth's ecosystems and climate dynamics. To handle the scale and complexity of the scientific questions being addressed, there is a strong need to integrate knowledge. However, relevant data are currently scattered, difficult to share and often threatened to be lost. In this paper, we take a close look at these challenges and the state of the art. We then sketch a coherent infrastructure to improve scientific data integration and preservation that is currently being developed in the GFBio project. This project brings together national key players providing environmentally related biological data and services to develop the German Federation for the Curation of Biological Data (GFBio). The overall goal is to provide a sustainable, service oriented, national data infrastructure facilitating data sharing and stimulating data intensive science in the fields of biological and environmental

research. While GFBio is still in the early phases of implementation, it might well serve as blueprint for similar initiatives in other disciplines. The current setup and approach are the result of a long and thorough preparatory phase involving representatives of all important stakeholders in the field.

2 Importance of Biodiversity Data and its Management

With the rapid evolution of scientific methodologies and technologies during the last decades an enormous amount of data in bio-sciences has been and will continuously be generated. Further, complex scientific challenges require international, interdisciplinary approaches resulting in extremely heterogeneous data sources to be analyzed. This increase in experimental and observational data volumes and analytical complexity has leveraged a large shift in research practices from traditional hypothesis-driven approaches by domain experts towards data-driven approaches which involves computational analysis, networked sensors and data sources, data sharing, large numbers of researchers including crowd-sourcing, and interdisciplinary collaboration.

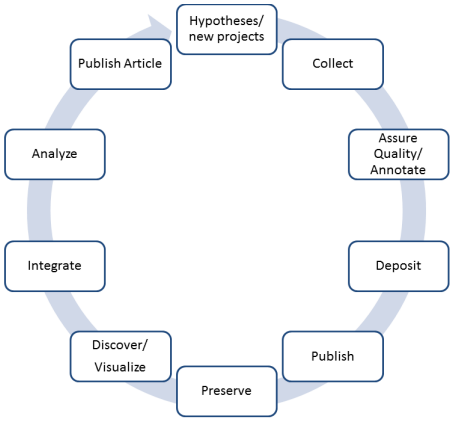


Figure 1: The Data Life Cycle

Ecosystem research, the study of organisms and their interactions with the environment, is one of the most fertile and receptive areas to such developments – this field is now key to meet global challenges such as climate change, particularly the loss of biodiversity and food resources and its impact on security. However, there is a significant lack of understanding of the complex interactions between organisms and the geo- and atmosphere. Environmental monitoring and observation systems show promising progress (e.g., GMES, EU-BON, GEO-BON, or ESONET/EMSO) but global data coverage is still patchy. Many species are still undiscovered or undescribed and even the extent of this knowledge gap is heavily disputed [MTA+11]. Our knowledge on the scales at which organisms and the environment interact is fragmentary. For any organism, a wide variety of information is of interest, describing for example their morphology, metabolism, genome, community structure and interactions, and ecological functions. An integrated analysis of this diverse knowledge requires a federated infrastructure that ensures general and long-term availability and usability of quality assured data and that supports the entire data life cycle as described by DataOne[1] (see Figure 1).

[1] DataOne: Tutorials on Data Management, `http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx` [last access: 26.06.2014]

3 State of the Art and Challenges in Biodiversity Data Management

In this section, we will follow the data life cycle and will discuss for each phase or group of phases what it is about, how it is addressed today and what the major open challenges are that will be addressed in GFBio. We will focus on the German research landscape.

Data Collection

What it is about: Biodiversity data is very diverse, ranging from remote sensing and next-generation-sequencing data to field observations and collection object-associated data linked with categorical or numerical data. The data are from ecophysiological studies, describe functional traits or are molecular data collected for phylogenetic treatments. They are gained automatically by data loggers (e.g., sensor data from e.g., climate stations) or gathered manually by individual researchers. The result is a great heterogeneity between the single data packages or data sets which might comprise data of varying granularity and quality. Often, the single data packages/data sources are linked with other data packages, but the documentation of the linkage might be scarce. Traditionally, biodiversity research data and ecological data are often produced in spreadsheet and specialised local databases, which are not appropriate to manage complex data sets, e.g., with relations to other data sets or multimedia data, e.g., audio files.

Currently a few database systems and workbenches are on the way to support biodiversity data gathering already in the field, later in the lab or finally in the collection magazine. Such tools have to be flexible and usable in the field without internet access but also be able to integrate into a virtual research environment using network resources and services.

How it is done today: There exist first tools to support human data collection in the field fulfilling a number of requirements listed by [JKN+09]. Examples include DiversityMobile[2] or the recently released OSD App[3], which allows both scientists and citizens to easily contribute to a global, orchestrated scientific event – the Ocean Sampling Day – by collecting standardized metadata. Also, tools not specifically developed in the context of biodiversity to support automatic data collection from different types of sensors etc. are in use.

Challenges addressed by GFBio: Data collection is a particularly crucial step when looking at the data life cycle, i.e., problems occurring in this step like errors, missing data, or missing meta data, introduce severe, expensive to overcome problems later on in the life cycle. Thus, it is very important to guide researchers from the very beginning of the study through the process of collecting well structured data with appropriate technical tools. In the best case, the tools should support the data collection, improve the data quality, maybe also enable the collection of a higher data volume without restricting the scientific goal. Challenges are thus both technical (develop better tools) and educational or sociological (training scientists in data management and creating a culture that acknowledges the importance of data management). GFBio will work on both issues by extending existing tools and by providing a virtual helpdesk with guidance and training.

[2]Diversity Mobile, `http://www.diversitymobile.net` [last access 26.06.2014]

[3]Ocean Sampling Day, `http://www.oceansamplingday.org` [last access 26.06.2014]

Quality Assurance

What it is about: Research data are often managed by single researchers without support by data scientists or IT experts. Thus, the data are often not well structured and the management concepts and solutions used are not appropriate. Often researchers spent little time on data management, because they are under pressure to publishing their results as quickly as possible. Unfortunately, long-term handling of research data has very low priority. Once an article has been published, the underlying data are no longer in focus. On the other hand, large portions of the raw data and the accompanying results are not being published for various reasons. As a result, errors and gaps in primary research data are often detected late, if ever. This seriously hinders reuse of data.

How it is done today: Today, biodiversity and ecological research is often organised in larger projects which have a dedicated data manager. These experts try to ensure high data quality by (mostly manually) checking data produced by the scientists in their projects. Tools supporting data quality management start to appear but are as of yet poorly integrated into the data collection tools.

Challenges addressed by GFBio: It is essential to offer flexible and user-friendly data management systems to single scientists and to larger research groups already at the beginning of their research activities. This will allow them to manage and quality-control their research data. Already at that time, the researchers have to be convinced that a maximised high quality data pool is relevant for their own research, now and in the future. The possibility for an early publication of high quality data with adequate citation mechanism (see below) might encourage the researchers to provide their data to the scientific community after the end of their research studies. Our goal with GFBio is to provide appropriate mechanisms so data is made public as early as possible. Also, GFBio aims to promote data collection and management tools that support and integrate data quality control.

Data Deposition

What it is about: Once data has been collected and the first level of management and quality assurance is done by the single researcher, it needs to be managed in a larger project context. Typically, this first management level is done for the duration of a project.

How it is done today: Traditionally, data management within small and medium scale projects is achieved by providing tailored common data management systems or simply by file sharing among partners. During the last years, many collaborative projects have developed their own data deposition or data management systems, including [LNB+12, NRB+13]. In addition to these project-specific tools, several flexible tools have been developed, which are highly configurable, reusable and can be adapted to the specific needs of research projects. Diversity Workbench[4] and BExIS 2[5] are examples for such multipurpose data management working environments. Both tools are currently used by a large number of organizations and consortia and provide an important basis for mid-term and long-term data management.

Challenges addressed by GFBio: Whether data is stored in a project-specific platform

[4] Diversity Workbench, `http://diversityworkbench.net` [last access 26.06.2014]

[5] BExIS++, `http://fusion.cs.uni-jena.de/bexis` [last access 26.06.2014]

or in one of the more adaptable ones, the long-term fate of data managed locally or with these tools is often unclear and dependent on the capacities of the organization hosting them. Many large scale projects are using tools for their internal data workflow without a clear long term data hosting strategy beyond their project lifetime. Therefore, a large amount of data needs to be preserved and transferred to a suitable long term data holding infrastructure and possibly to a data management infrastructure maintained by a scientific data center or a long-time institutional data repository. Strong support for the development of standard conform and advanced software for long-time data management and deposition as well as the establishing of content standards, management plans and archiving concepts is therefore urgently needed. Within GFBio we will extend BExIS and the Diversity Workbench to meet these requirements and to serve as blueprints for other tools.

Data Publication

What it is about: There is a cultural change in the research community promoting the idea of data publishing as a clear incentive for scientists to share their data. Only in the past few years have scientists began calling for data "citation" and referring to data "publication" rather than data "sharing" and "availability" [CMG+13].

How it is done today: Data publication can be similar to the conventional publication of articles in journals that includes online submission, quality checks, peer-review, editorial decisions, and an equivalent of page proofs. In fact, data storage in central databases is getting increasingly important or is yet mandatory for the acceptance of peer reviewed publications in specific fields of biodiversity research as e.g. molecular sciences or ecology [Whi11]. PANGAEA has built up an extensive collaboration with science publishers and implemented services to link articles and data. The impact on citation rates could be shown in a recent bibliometric study on science articles having supplementary data [Sea11]. Eventually an index for science data would be needed similar to Data Usage Index (DUI) [IC11]. Data publishing is also strongly supported by GBIF and the ICSU World Data System (WDS).

Challenges addressed by GFBio: One major challenge is to make data publication and access to published data as easy as possible for scientists. This requires a strong integration of data publication in data management tools which should support a "one click" publication. In addition to fostering such tools, GFBio's helpdesk will provide a wide range of information on options for data publication.

Data Preservation

What it is about: Much of biodiversity data is hard or impossible to reproduce. For instance, it is impossible to repeat an observation of the species richness in a certain area for a point of time in the past. It is thus extremely important that data is preserved beyond the end of projects. Since only a small percentage of data end up being part of a scientific article or a data publication, other means to ensure long-term availability of data are needed.

How it is done today: Globally, the backbone of long-term archives for environmentally related biological data is mostly restricted to specific disciplines, data types, and/or

regional coverage. Prominent are the International Nucleotide Sequence Database Collaboration (INSDC) with their archives. Other networked resources or platforms include the Global Biodiversity Information Facility (GBIF), Dryad and some others devoted to certain data domains [THR12]. Prominent global infrastructures for environmental data include the WMO and ICSU World Data Centers. Besides there are various national data centers, e. g. the Australian National Data Service (ANDS) and the US DataOne initiative.

In Germany, a number of data centers and archives are available to researchers: the University of Bremen jointly with the AWI operates PANGAEA, a data archive specialized on environmental data. Founded in 1992 PANGAEA has demonstrated its long term perspective by a certification of ICSU World Data System and was accredited by the WMO as Data Collection and Processing Center (DCPC). Currently, PANGAEA hosts more than 450.000 data sets comprising of more than 6 billion individual measurements collected during approx. 160 EU, international and national projects. Currently three main databases for molecular biodiversity analysis exist in Germany. They are mainly used for the treatment of sequences from microorganisms, which make up the majority of currently available molecular data. The SILVA rRNA database[6] provides quality-controlled sequence data sets for molecular biodiversity analysis. The Megx.net portal[7] integrates microbial genomic and environmental data based on georeferencing. Both systems are hosted and maintained at the Max Planck Institute for Marine Microbiology and Jacobs University Bremen. The Bacterial Diversity Metadatabase "BacDive"[8] is a highly curated, comprehensive data resource for all cultured Bacteria and Archaea in Biologial Resource Collections (BRC), currently under development by the German Collection of Microorganisms and Cell Cultures (DSMZ) in Braunschweig. The German Natural History Collections, organised in the Consortium of the Deutsche Naturwissenschaftliche Forschungssammlungen (DNFS) guarantee sustainability in hosting more than 140 million of life science objects together with adjacent information, mainly occurrence data of taxa in space and time, as well as all kinds of additional ecological data. The DNFS institutions also manage enormous amounts of biodiversity- and organism-related information, e.g., concerning DNA samples and data, monitoring data, taxon- and object-related metadata and multimedia data which also include (meta-) data from past and ongoing third-party-funded research projects. Promising progress has been achieved regarding standardization and networking among these national history collections which led to harmonized access to these archives with more than 10 million data records within e.g. GBIF and BioCASE.

Challenges addressed by GFBio: For individual projects, resources and funding are limited. That is one reason why – despite a growing number of projects having produced large volumes of data during the last years – only part of this data is stored in data centers such as the ones mentioned above and generally available for reuse. A lot of data is lost. Funding agencies as well as research organizations put increasing pressure on the research community and demand long term storage and open access on research data. The majority of researchers are generally willing to archive and share their data. However, to get acceptance by the researchers, the data archives and scientific data centers have

[6] Silva: High Quality Ribosomal RNA Databases, `http://www.arb-silva.de` [last access 26.06.2014]

[7] Megx - Marine Ecological Genomics, `http://www.megx.net` [last access 26.06.2014]

[8] BacDive - The Bacterial Diversity Metadatabase, `http://www.bacdive.dsmz.de` [last access 26.06.2014]

to provide appropriate functionalities and prove their trustworthiness as well as a long term perspective of operations. It is important for them to comply with international standards, technical formats and guidelines for long term archives. In Germany, only few data archives/data centers operate in a sustainable and organized way, especially in the fields of environmental sciences and natural history collections and museums. Germany's situation is essentially the same as the global one, where existing infrastructures are insufficient to cope with heterogeneous data from multidisciplinary projects, leading to the existence of numerous isolated, technically proprietary and unsustainable databases. A coordinated approach is needed to improve the current situation and expand the infrastructure of the archives and data centers. Such an approach is established in GFBio.

Discovery and Visualisation

What it is about: For biological sciences a major drawback is the lack of interconnections between the classical biodiversity knowledge resources such as natural history, environmental and molecular data providers as well as literature. Data generated in integrative biodiversity and ecosystem research are extremely heterogeneous, multidisciplinary, and collected by many different scientific disciplines using a large range of methods, technologies, and data standards. On the other hand, biodiversity research critically depends on a wide range of scientific disciplines providing the fundament for the understanding of organisms and their interactions with their environment in time and space. The lack of coordination between standards still hinders true interoperability which makes it impossible to enable unified mechanisms for discovery and visualisation across all relevant resources.

How it is done today: Harmonization and mapping of domain specific metadata standards as well as the integration of community profiles has to occur before interdisciplinary archives can successfully be realized. Several approaches have been initiated in the past e.g. by TDWG groups to integrate geospatial and biodiversity data or GEO-BON to integrate data resources from different domains. Domain specific approaches include the definition of complex namespaces and schemas such as the Diversity Workbench (DWB) or EDIT Common Data Model (CDM) suitable to be applied as a (maximum) standard for mapping biodiversity and ecological data. On the European level the INSPIRE initiative fosters the ISO19139 schema implementation. Efforts have been made to e.g. map between ISO19139 and other schemas such as Dublin Core or the GBIF metadata profile and to map genome and geographical information. In recent years, a number of initiatives started out to implement semantic services providing stable terminologies and ontologies supporting specific user demands in biodiversity research such as the "TDWG-Ontology" and the GBIF multilingual vocabulary server[9].

Challenges addressed by GFBio: Despite the progress made over the last decade, there is a general lack of formally represented knowledge about the relation between terms referring to similar entities used in different scientific communities. As a consequence, computer-based biodiversity studies often fail to exploit the potential of existing and highly valuable data resources ranging from data stemming from individual scientific projects to large scale data intensive biodiversity and environmental assessments. GfBio addresses

[9]TDWG LSID Vocabularies, `http://rs.tdwg.org/ontology/voc/Collection`, GBIF Vocabularies, `http://vocabularies.gbif.org/` [last access 26.06.2014]

this gap by developing a terminology server and services offering unified interfaces to existing and presently in-accessible vocabularies used in different relevant scientific disciplines. The terminology server will also provide a platform for ontology development with a special focus on collaborative aspects and consensus finding as well as maintenance processes in continuously evolving scenarios. In a community-based approach, GfBio works closely with experts in the different sub-disciplines providing prototypic domain-vocabularies and use-cases for the future terminology services.

Data Integration and Analysis

What it is about: Data integration provides means to access various data sources in a uniform manner to extract information that is implicitly hidden within their union. For biodiversity research, it is expected that the value of this implicit knowledge is large.

How it is done today: Today, the individual disciplines of biodiversity research collect large data collections, let it be by measurements or by digitalization of existing material. While these collections are large and technically advanced, they remain isolated and means for integrated access and analysis lack. Thus, today, data integration in biodiversity remains a largely manual task requiring considerable effort. An example for such an integration effort is the TRY database [KDL+11]. This is the largest compilation of plant trait data integrating data from more than a hundred contributors. Despite the clear thematic focus, integrating this data requires tremendous manual effort. The situation is even worse when trying to integrate data across disciplines. This is a clear hindrance to answering key questions of biodiversity research.

Challenges addressed by GFBio: Making archives interoperable and enriching data with (ontology-based) metadata is of little use, if there is no common interface that makes it easy for users to access the data. This means that sophisticated search interfaces are needed that allow searching in data and metadata, support browsing and faceted search and hide syntactic differences between the different data. The continuously increasing volume and complexity of biodiversity data this becomes more and more challenging. On the one hand, efficient harvesting and indexing techniques that scale with growing data volume are needed, on the other hand, new interfaces for navigating across search results are needed, including visualizations that allow a quick overview over potentially huge and complex search results. Community involvement is a proven measure to improve the overall quality of data services. Crowdsourcing has been tested in a variety of web pages and e.g. categorization and prioritization of content (so-called folksonomy) has a good potential to improve the find-ability of research data. Further, personalization can help to recommend contents and enable users to return to and build on earlier search results. Highly heterogeneous data in differing formats needs innovative integrative approaches such as text mining and indexing including Semantic Web technologies and Linked Data concepts using RDF and SPARQL. Such innovative concepts need to be further elaborated and adapted to the specific needs of the scientific community. Many of these new approaches have previously been evaluated for biodiversity research, however the adaptation of these technologies in a highly interdisciplinary data environment still is a challenge.

Data-intensive science [TTH11] is building on the integration of data from various sources. The real challenge in interdisciplinary data management is finding a common approach to

exchange data not only metadata. Only little progress has been made to agree on interdisciplinary content standards and corresponding exchange protocols for data. Even within one scientific community multiple, incompatible formats are used for machine-machine interaction such as NetCDF or OGC OM format. In addition there are a large number of web services offering proprietary formats. Emerging technologies such as grid or cloud technologies have made promising progress with regards to data exchange, storage, scalability and performance. But no successful attempt has been made to apply these technologies to integrate the complex and interdisciplinary data scenario related to biodiversity. However such concepts are a prerequisite for scalable handling of large data volumes and necessary to provide innovative commonly usable tools and services to analyze such data. GFBio will therefore investigate new ways to embed these technologies within the given data landscape in order to provide such added value services for the integration and analysis of heterogeneous environmentally related biological data.

4 GFBio – A proposal for an integrated platform

The need to address the challenges outlined above was recognized several years ago by the senate commission on biodiversity research of DFG. As a result, it encouraged a preliminary project to further investigate the conditions, in particular the willingness to share data and the availability of an appropriate infrastructure in Germany. Results of this project were published in [ETB+12, BSE+12] and showed clear need for action. This resulted in an initiative to build a national infrastructure for data management in biological and ecological research which brought together stakeholders from all involved communities, namely archives, national history collections, biodiversity research and computer science. This group of 21 principal investigators obtained funding by DFG to set up a "German Federation for the Curation of Biological Data (GFBio)".

Figure 2 provides an high-level overview of the GFBio components. GFBio will build on international standards and proven data archiving infrastructures and workflows such as the PANGAEA long term archive for environmental data, the resources of networked natural history collection data repositories and genome data archives. Interoperability among these resources, the integration of domain specific data bases as well as external, international data providers and a cooperation with relevant publishers are a key focus of the project. GFBio will follow innovative approaches to stimulate data submission among the scientific community. In particular it will pursue the promising efforts of data publication and cross-linking of research data with the scientific literature. Project data management and submission will further be supported by integrating and further developing innovative tools and platforms built on existing efforts such as the Biodiversity Exploratories Information System (BExIS) and Diversity Workbench (DWB).

Via the portal, we aim to establish a central point of contact for researchers and the interested public with respect to environmentally related biological data. The portal will offer the possibility to search across the connected data sources and will be closely integrated with visualization and analytical tools which will enable efficient data aggregation and analysis, including links between formerly non-connected data sets from different sub-

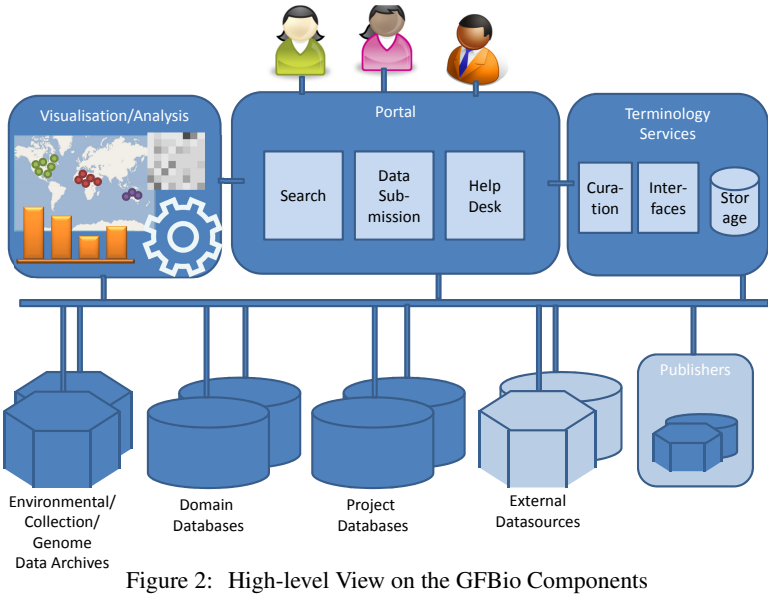


Figure 2: High-level View on the GFBio Components

disciplines. In addition, the portal will support data submission from project or domain databases to the underlying archives. Based on commonly defined workflows and procedures, the virtual helpdesk will address digital data curation issues as early as possible within the creation of research data and will provide researchers with guidance on how to do data management and what tools to use. The Terminology Services provide a semantic backbone supporting high quality data acquisition and data integration.

GFBio will thus provide an infrastructure that addresses not only the individual challenges identified above, but the data life cycle as a whole. We are convinced that only such an integrated solution with strong links to international initiatives has the potential to provide efficient, sustainable support for biodiversity (or any other) research. With the specific challenge to integrate data and metadata from several domains GFBio will also push relevant standards and innovative solutions. Special emphasis will be given to governance and organizational issues of the planned infrastructure. Taking into account the working conditions for every institution involved, a common organizational structure with common services and sustainable business models will be developed.

5 Summary and Conclusion

While research data management has gained a lot of interest over the last few years and initial building blocks for a solution exist, many challenges remain. In particular, up to now, integrated solutions that provide users with a ”one-stop” platform for all their needs

with respect to data management across all phases of the data life cycle are lacking. For Germany, GFBio aims at providing such a platform.

Acknowledgements: We thank all GFBio PIs (in addition to those co-authoring the paper these are: François Buscot, Johanna Eder, Stephan Frickenhaus, Christoph Häuser, Thomas Hickler, Jens Kattge, Heike Neuroth, Jörg Overmann, Jens Schumacher, Johann Wägele, Christian Wirth, Ramin Yahyapour) and everyone else involved in GFBio for their contributions. The work described here is funded by DFG.

References

[BSE+12] Kerstin Bach, Daniel Schäfer, Neela Enke, Bernhard Seeger, Birgit Gemeinholzer, and Jörg Bendix. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics*, 11:16–24, 2012.

[CMG+13] Mark J. Costello, William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, and Philip E. Bourne. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, 28(8):454–461, August 2013.

[ETB+12] Neela Enke, Anne Thessen, Kerstin Bach, Jörg Bendix, Bernhard Seeger, and Birgit Gemeinholzer. The user's view on biodiversity data sharing – Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11:25–33, 2012.

[IC11] Peter Ingwersen and Vishwas Chavan. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC bioinformatics*, 12(Suppl 15):S3, 2011.

[JKN+09] Stefan Jablonski, Alexandra Kehl, Dieter Neubacher, Peter Poschlod, Gerhard Rambold, Tobias Schneider, Dagmar Triebel, Bernhard Volz, and Markus Weiss. DiversityMobile – Mobile Data Retrieval Platform for Biodiversity Research Projects. In *GI Jahrestagung*, pages 610–624, 2009.

[KDL+11] Jens Kattge, Sandra Diaz, Sandra Lavorel, IC Prentice, P Leadley, G Bönisch, Eric Garnier, Mark Westoby, Peter B Reich, IJ Wright, et al. TRY–a global database of plant traits. *Global change biology*, 17(9):2905–2935, 2011.

[LNB+12] Thomas Lotz, Jens Nieschulze, Jörg Bendix, Maik Dobbermann, and Birgitta König-Ries. Diverse or uniform? – Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecological Informatics*, 8:10 – 19, 2012.

[MTA+11] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on Earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011.

[NRB+13] Karin Nadrowski, Sophia Ratcliffe, Gerhard Bönisch, Helge Bruelheide, Jens Kattge, Xiaojuan Liu, Lutz Maicher, Xiangcheng Mi, Michael Prilop, Daniel Seifarth, et al. Harmonizing, annotating and sharing data in biodiversity-ecosystem functioning research. *Methods in Ecology and Evolution*, 4(2):201–205, 2013.

[Sea11] JR Sears. Data sharing effect on article citation rate in paleoceanography. In *AGU Fall Meeting Abstracts*, volume 1, page 1628, 2011.

[THR12] Dagmar Triebel, Gregor Hagedorn, and Gerhard Rambold. An appraisal of megascience platforms for biodiversity information. *MycoKeys*, 5:45–63, 2012.

[TTH11] Kristin M Tolle, D Tansley, and Anthony JG Hey. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proceedings of the IEEE*, 99(8):1334–1337, 2011.

[Whi11] Michael C. Whitlock. Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution*, 26(2):61 – 65, 2011.

Modeling and Optimization of high-Complex systems

(MOC 2014)

Modeling and Optimization of high-Complex systems (MOC 2014)[1]

Abstract: Die aktuellen Entwicklungen und Herausforderungen in Gesellschaft und Wirtschaft stellen neue Anforderungen an die Informatik und Wirtschaftsinformatik. Nur durch die Weiterentwicklung und den gezielten Einsatz von IT-Technologien, -verfahren und -methoden sind die „Probleme unserer Zeit" nachhaltig zu lösen. Beispieldomänen hierfür sind die Energiewende, das flexible Steuern und Überwachen großer Verkehrsnetze und das Schaffen und Betreiben agiler effizienter Unternehmensverbünde. Wesentliche Aspekte hierfür sind „agilitätsbegünstigende" und validierbare Modellierung sowie die Optimierung der Prozesse und des Betriebes. Der Workshop adressiert Aspekte der Modellierung, der Optimierung und der Simulation hochkomplexer Systeme im o.g. Sinne, wobei insbesondere auch die zeitnahe Auswertung großer Datenmengen (Big Data) zur „optimalen Steuerung" der modellierten Prozesse thematisiert werden soll. Besonderes Interesse besteht dabei an Verfahren, die neben Effizienz auch Deklarativität und Interaktivität bei der Behandlung von Simulations- und Optimierungsproblemen sichern. Ansätze bieten hierfür zum Beispiel Constraint-Verfahren. Deklarativität - eine Beschreibung der gewünschten Lösung, anstelle einer prozeduralen Abarbeitungsvorschrift - scheint mit Effizienz, insbesondere vor dem Hintergrund der Behandlung von Optimierungsaufgaben, kaum verträglich zu sein. Dennoch ist Deklarativität wünschenswert, um eine Problembehandlung möglichst einfach (Nichtinformatiker) und transparent (Anpassbarkeit) darzustellen und Verständlichkeit, Wartbarkeit und Erweiterbarkeit/ Adaptierbarkeit der Software zu erleichtern. Die Eigenschaft der Interaktivität kann Unterstützung bieten, Expertenwissen während des Optimierungsprozesses zu nutzen, um in kürzerer Zeit eine Lösung zu finden. On-the-fly-Expertenwissen ist für Simulations- und Optimierungsverfahren schon deshalb ein unverzichtbares Potential, weil die zu behandelnden Probleme aufgrund wachsender Größe, steigender Suchraumgröße und Komplexität durch automatische Verfahren kaum oder nicht beherrschbar sind. Weiterhin erfordern viele Anwendungen, dass gute oder gar beste Lösungen innerhalb kürzester Zeit zu finden sind, um beispielsweise Betriebsabläufe bei unvorhergesehenen Störungen unmittelbar anzupassen und so die Wettbewerbsfähigkeit von Unternehmen und Unternehmensverbünden zu erhalten. Als erforderlich werden daher interaktive Simulations- und Optimierungssysteme angesehen, die auf Basis einer umfassenden Daten- und Erkenntnislage (siehe oben) Anwender bei ihren operativen und auch strategischen Entscheidungen, z.B. durch "was-wäre-wenn"-Szenarien oder durch Berechnen und Präsentieren von alternativen Lösungen, unterstützen.

Der Workshop soll Fachleuten und Anwendern die Möglichkeit zum Austausch und zur fruchtvollen Diskussion von Ideen, Ansätzen, Verfahren und Problemlösungen geben, wobei hierbei idealer Weise Grundlagen für zukünftige,

[1] http://www.constraint-programming.de/MOC-2014/

im o.g. Sinne zielführende, Forschungsarbeiten entstehen. Ein Block des Workshops soll für Tutorials genutzt werden, in denen für interessierte Anwender einige Grundlagen der Modellierung von industriellen und logistischen Optimierungsproblemen dargestellt werden.

Organisation: Ulrich Geske (Universität Potsdam), Ulrich John (Hochschule für Wirtschaft, Technik und Kultur, Berlin), Armin Wolf (Fraunhofer FOKUS, Berlin)

Programmkomitee: Christoph Beierle (FernUni Hagen), Ulrich Geske (Universität Potsdam), Hans-Joachim Goltz (Fraunhofer FOKUS, Berlin), Petra Hofstedt (Brandenburgische Technische Universität, Cottbus), Walter Hower (Hochschule Albstadt-Sigmaringen), Ulrich John (Hochschule für Wirtschaft, Technik und Kultur, Berlin), Andreas Kmoch (Toll Collect), Georg Ringwelski (Hochschule Zittau/Görlitz), Hans Schlenker (IBM Deutschland), Andreas Schutt (NICTA, Australien), Dietmar Seipel (Universität Würzburg), Armin Wolf (Fraunhofer FOKUS, Berlin)

Nine Interaction Guidelines for the Design of Human Computer Interaction in Detailed Scheduling Systems

Anna Prenzel, Georg Ringwelski

Hochschule Zittau/Görlitz
Fakultät Elektrotechnik und Informatik
Brückenstraße 1
02826 Görlitz
aprenzel@hszg.de
gringwelski@hszg.de

Abstract: With the help of modern software, the scheduling process in many fields can be automated nearly completely. However, these systems often provide insufficient interaction possibilities for the scheduler to incorporate his own expert knowledge. Finding and implementing suitable planning decisions is often a time-consuming and error-prone process that is not supported by the computer. This deficiency can impair the practicability of the schedules and hence the benefit of the planning system considerably. In this paper, we deduce 9 interaction guidelines that form a concept for the design of human-computer-interaction in scheduling systems. Unlike approaches that extend general HCI guidelines we derive the rules from recent research on human decision making. From that perspective we define computer support for both the formation and the implementation of planning decisions by the scheduler. The results of a user test show that the concept allows an effective incorporation of the human scheduler. Furthermore the concept can be applied to any detailed scheduling problem.

1 Introduction

1.1 Application area

The interaction guidelines described in this paper can be applied in any detailed scheduling system that supports the assignment of *jobs* to *resources*. Examples are the assignment of shifts to staff members, production items to machines or transport orders to vehicles. Detailed scheduling systems serve to make operative planning decisions (rather than tactical or strategic planning decisions, see [SFG+12]) prescribing the temporal arrangement of jobs within the schedules of the individual resources. Both assigning and sequencing of the jobs is usually subject to constraints that limit the scope of action. Typical constraints are *start time constraints* (including time windows, non-overlapping constraints and sequencing constraints) and *resource constraints* (including capacity constraints and resource qualifications). The goal of scheduling is typically to maximize profit and to reduce costs. Scheduling systems are able to create schedules that are valid with respect to

the given constraints and optimal with respect to one or more goal functions. The detailed scheduling problems our guidelines address are derivatives of the *resource constrained project scheduling problem (RCPSP)*. A formal definition of the RCPSP can be found in [HB10]. We consider detailed scheduling systems with regard to their *user experience*. We define it as the extent to which the work task of the user is supported by the computer.

1.2 State of the art

In the last 20 years, many field studies have been carried out in order to evaluate the user experience of scheduling systems [CvW12, GFW11]. The results showed that scheduling software is often designed to make all decisions autonomously. It is not expected that the human scheduler participates in the process of schedule creation. Therefore it is often hard to modifiy the underlying scheduling model. However, the field studies revealed that in practice the constraints for jobs and resources and the requirements on the properties of the schedule change frequently. For each change the scheduling model has to be adapted accordingly, otherwise the resulting schedules are not applicable. Furthermore, only the human scheduler has the expert knowledge to decide at which point of time which change has to be carried out. He needs a system that supports him with finding and incorporating the right decisions. It becomes clear that there is a contradiction between the actual needs of the users and the assumption of the software industry, that full automation is sufficient. Due to this insight, the design of human computer interaction for scheduling systems has become a matter of research. Initially it was the goal to identify user actions, which allow the scheduler to incorporate his expert knowledge. A comprehensive list can be found in [dNE05]. The most important user actions deal with adapting the scheduling model (*interactive optimization*): weighting goal functions, modifying input data, changing the constraints, selecting a solution and manually modifying a solution.
It is the task of the software developers to implement the actions according to the guidelines of usability (e.g. DIN EN ISO 9241-110). They lead to interactive systems in which both human and computer take part in the solution process. However, the interaction still has some drawbacks concerning the user experience. The user actions often induce a high cognitive workload caused by the huge amount of data that is generated by the computer. For example, it is difficult for the user to compare numerous alternative solutions. In [CvW11] it is criticized: On the one side there are sophisticated algorithms to create schedules, on the other side the user is completely left alone with his actions.
For this reason the concept of function allocation between human and computer was extended from *interactive* to *mixed initiative*. As the term "mixed" indicates, both human and computer can intervene in the solution process on their own initiative. While the user is active, the computer does not remain passive, but it helps him to successfully complete the action. This approach was applied in numerous scheduling prototypes, for example in [BZE+07]. According to [BBIM96] this approach opens a new way to make the knowledge represented in the computer available for the user: The human solves a problem collaborating with the computer like with a human partner that contributes suggestions and opinions.

1.3 Contribution

In the literature there are concepts and examples of human computer interaction that can be applied to detailed scheduling systems. However, we miss a design concept that is tailored to the requirements of general detailed scheduling problems. The existing concepts have been developed either for specific problems (e.g. production scheduling) or for general optimization problems with constraints. In both cases it is difficult to reuse the interaction design in the development of a new scheduling system.
In this paper, we define an interaction concept for general detailed scheduling problems that is based on the principles of interactive optimization and mixed initiative scheduling. It helps to provide an optimal user experience in scheduling systems. Our work is structured as follows: First, we investigate the process of scheduling in practice and characterize the decision process of the scheduler (section 2). From this, we conclude requirements on the design of the human computer interaction (section 3.1). In the sections 3.2 to 3.10 we define 9 interaction guidelines that show how the requirements can be fulfilled. In the last section 4 we describe the results of a user test that confirm the effectiveness of the guidelines.

2 Human Decision and Design Processes in Scheduling

2.1 Temporary Constraints

In practice there are several typical situations in which the original scheduling model has to be updated by adding, removing or modifying the constraints. We distinguish between two different kinds of constraints: *Persistent constraints* are contained in the original scheduling model. They describe the services available to carry out the jobs (e.g. working times, capacities of resources) and the scheduling requirements of the jobs (e.g. time windows, required resources). In the remaining work we do not consider changes to persistent constraints. *Temporary constraints* represent the scheduler's decision, which resource should carry out which job at a specific point of time (cf. [Hig01]). For example, the assignment of values to decision variables (e.g. start times, resources of jobs) can be regarded as development of temporary constraints. However, we generalize the concept of temporary constraints by encompassing all kinds of constraints that narrow the domain of the variables. This allows a better modeling of human scheduling strategies. They are called "temporary", because they are developed during a scheduling process and discarded afterwards.

2.2 Rule-Based and Knowledge-Based Scheduling

Schedulers use templates (e.g. existing schedules from past), decision rules (e.g. if-then-statements), dispatching rules (e.g. sorting jobs by earliest start date) and other strategies

(e.g. divide and conquer) to manually solve scheduling problems. These strategies process the scheduling model in its original state, thus they can be implemented in a software. Scheduling with the help of manual strategies or arbitrary algorithms is in the following called *rule-based scheduling*.
In practice, rule-based scheduling often cannot be applied readily, because the model does not map the reality at the moment of schedule creation adequately. For example, the scheduler has expert knowledge about changed conditions, under which the resources operate (e.g. "a machine in need of repair should be used less often") or about requirements on the properties of the schedule that are not included in the model (e.g. "the schedule should be comprehensible for the staff"). The conditions and requirements can be considered by complementing the model with temporary constraints. After that, the scheduling problem can again be solved rule-based. We refer to this process as *knowledge-based scheduling*.
Often expert knowledge cannot be directly translated into temporary constraints, as it is not clear from the start, with which decisions a condition or requirement is represented best. In this case the scheduler has to set up and evaluate different constraints. The process of developing, evaluating and choosing temporary constraints can be modeled as decision process with 4 main stages (cf. [TSD11]):

Intelligence: This stage is the entry point of the decision process. Here the scheduler collects the knowledge that has to be incorporated into the schedule.
Design: In this stage alternative temporary constraints are compared and evaluated.
Choice: The final decision for a set of constraints is made in the choice stage.
Implementation: Now the selected constraints can be added to the scheduling model. A solution can be found using a rule-based procedure.

The design stage can be left out if conditions and requirements can be translated into constraints immediately.

2.3 Constraint Processing during Knowledge-Based Scheduling

In the design stage of a decision process, the scheduler evaluates alternative sets of temporary constraints expressing his preferences. Viewing scheduling as a design problem [HGM12], each set is a "'design artifact"' created with the following operations of constraint management [Dar91, Vis06]: *constraint posting* and *constraint propagation* [HW07].
Design problems typically are not solved by immediately posting constraints that assign a value to a variable. Instead, the domains of the variables are narrowed step by step. After each step the consequences for the remaining variables are examined by constraint propagation. The current state of the domains shows the scope for continuing the design process. In the following steps the design can be refined by posting more constraints. This procedure allows the designer to gain insights about the relationships between the variables. The effects of propagation make him question his preferences and, if necessary, change or discard them.
Building on the insights of constraint-based design, we can develop our understanding of knowledge-based scheduling further. In the design stage, temporary constraints can be posted to delay the assignment of a concrete value to a variable. For example, the

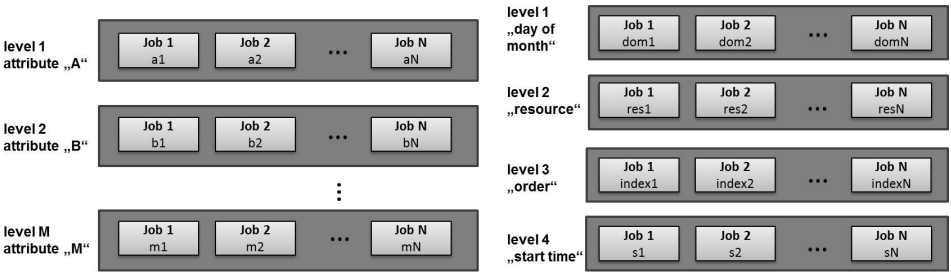


Figure 1: Template for an assignment hierarchy (left), example for fleet scheduling (right).

scheduler can define a certain time of day (e.g. morning or afternoon) before he assigns a concrete start time for a job. In doing so, an assignment for the variable "time of day" corresponds to an *abstract assignment* of a range of values to the variable "start time". It could be preceded by an even more abstract assignment, for example by an assignment for the variable "day of month". Generally, a variable x can be used as an abstract assignment for variable y, if both variables represent attributes of the same job i and if the propagation of an assignment for x removes certain values in the domain of y. Thus, the order in which the variables of a job are assigned determines which assignment is more abstract than another. Each specific order imposes the following *assignment rules*:

1. An assignment can only be made, if all previous abstract assignments have been completed.
2. If an abstract assignment is changed, all subsequent assignments have to be renewed.

The scheduler can choose the assignment order depending on his decision situation. For example, if he is interested in the consequences of an assignment for variable x on the remaining variables, this assignment will be made first. Furthermore, assignment orders can often be derived from rule-based scheduling strategies. For example, a typical strategy to solve vehicle routing problems is to first assign each job a driver, and secondly assign each job a start time. In knowledge-based scheduling, the scheduler can maintain this strategy if it allows him to better comprehend the problem.
We suggest a concept for visualizing typical assignment orders of a scheduling problem, the *assignment hierarchy*. In this paper it will be used to illustrate the computer support for knowledge-based scheduling. Figure 1 left shows the template for an assignment hierarchy: It consists of M levels, where each level represents an attribute. Each level i contains N variables (e.g. $a_1, .., a_n$ for attribute "A") as a placeholder for the attribute values of all N jobs. The order of the levels corresponds to the chosen assignment order and the state of the jobs complies with the assignment rules: At level i,$i < M$ a job j is in an abstract state, as its assignments at the previous levels have already been completed, but the assignments of the subsequent levels are pending. For a scheduling problem there possibly exist multiple assignment hierarchies. Figure 1 right shows an example assignment hierarchy for a fleet scheduling problem. For each day of month each driver is assigned to a tour which defines the order and start times of the jobs to be carried out.

2.4 Factors Influencing the Effort of Knowledge-Based Scheduling

After an assignment, propagation can cause one or more variables to have empty domains or domains that do not any more contain the desired values [HW07]. In order to prevent this the scheduler would have to look ahead on the consequences on the remaining variables before assigning a value to a variable. This kind of *anticipatory propagation* is generally not possible. The reason is, first, the limited capacity of the human working memory that does not allow to process the high number of variables in practical scheduling problems [Mac08]. Second, the scheduler usually does not know the constraints of the current scheduling problem. Consequently, he is not able to perform the propagation process manually.
The difficulty of propagation leads to a blind assignment, which can end up in an empty or unsuitable domain. In the moment the scheduler notices this the responsible assignment possibly lies far back. The scheduler has to perform a backtracking process to find out and change it, which can also be very time-consuming and difficult.
Even if the scheduler finds a valid set of temporary constraints not violating any constraints, it is not guaranteed that the modified scheduling model can be solved in the implementation stage. If the scheduler does not include all variables of the model in the design stage he can assess neither the solvability nor the optimality of the considered alternative solutions. This effect makes it impossible to work with a limited set of variables, for which preferences exist, in the design stage. Furthermore, it prevents the scheduler to evaluate and compare "schedule sketches" that are composed of a few abstract assignments. Therefore we refer to this effect as *loss of abstraction*.

3 Guidelines for the Development of Interactive Scheduling Systems

3.1 Requirements on the Computer Support

We define an interaction model that provides a computer support adapted to the knowledge-based scheduling process. The requirements can be deduced from the analysis: The computer should support the design and the evaluation of alternative solutions in the design stage (sec. 2.2, 2.3). According to the mixed initiative principle, the effort caused by constraint propagation and loss of abstraction should be reduced (sec. 2.4).

3.2 Guidelines G1 to G9

3.2.1 Guideline G1 - Determining the assignment hierarchy

Prior to the development of a scheduling system, typical assignment hierarchies should be identified.
The establishment of an assignment hierarchy helps to find out, which scheduling deci-

sions and which dependencies between them are relevant for the scheduler. Furthermore, it helps to identify abstract assignments that schedulers possibly use in their heads. This is the prerequisite to create a mixed initiative computer support for the user's actions.

3.2.2 Guideline G2 - Manual modification of a schedule

G2.1: *The scheduling system should allow a manual assignment of values to all variables of the scheduling model.*
G2.2: *The scheduling system should ensure that manual assignments comply with the assignment rules of a suitable assignment order.*
If several assignment hierarchies are to be supported, the scheduler should be able to choose the current assignment hierarchy.

3.2.3 Guideline G3 - Verifying the compliance with constraints

The scheduling system should verify, whether manual assignments comply with the constraints of the scheduling model. If a constraint is violated, the scheduler should be informed.
This guideline helps the scheduler to become acquainted with the scheduling model and to refine his preferences during the design of a scheduling alternative.

3.2.4 Guideline G4 - Verifying the compliance with requirements

The scheduling system should verify, whether the schedule complies with all requirements of the organisation. After each assignment, the quality scores should be updated.
The scheduling system should supervise the assignments of the scheduler to prevent him from making decisions that significantly impair the quality scores and other requirements.

3.2.5 Guideline G5 - Explicit fixation

G5.1: *The scheduling system should allow the scheduler to define temporary constraints for all variables (fixation). They should be taken into account by automated scheduling.*
G5.2: *The fixation has to be chained upwards automatically: If an assignment is fixed at assignment level i, the assignments for the variables at the levels 1 to $i-1$ of the same job have to be fixed as well.*
With this guideline the computer can take over the rule-based implementation stage after the scheduler has defined temporary constraints. Furthermore, it helps to avoid the loss of abstraction in the design stage. In order to evaluate a scheduling alternative and its quality, the scheduler can fix his decisions at the desired assignment levels and let the computer complete the remaining assignments.
If an existing schedule is to be changed, the scheduler can fix parts of the schedule that must not be modified any more.

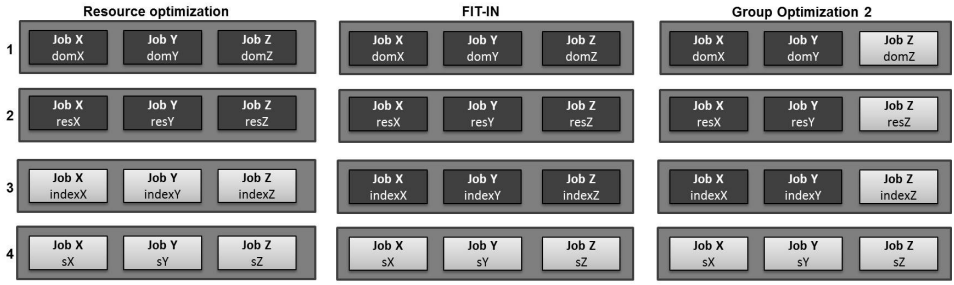


Figure 2: Templates for scheduling functions (dark grey color indicates fixed jobs).

3.2.6 Guideline G6 - Scheduling functions with implicit fixation

A scheduling system should provide scheduling functions that carry out multiple fixations and automatic scheduling in one step. They can be deduced from the assignment hierarchy of the scheduling problem.
Implicit fixation reduces the effort of manual fixation. For example, 7 different scheduling functions can be used in fleet scheduling (fixed levels/jobs are given in brackets, see figure 2): *resource optimization* (level 1 and 2), *resource optimization 2* for optimizing a tour (levels 1 and 2 for a selected resource, levels 1 to 4 for the rest of the schedule), *FIT-IN* for repairing a schedule with overlapping jobs (levels 1 to 3), *group optimization* for rescheduling selected jobs (levels 1 to 4 for non-selected jobs), *group optimization 2* where non-selected jobs can be shifted in time (level 1 to 3 for non-selected jobs) and *full optimization* where all jobs are scheduled (no fixation).

3.2.7 Guideline G7 - Support for value selection

The scheduling system should support the scheduler in selecting a value from the domain of a variable. The support can be implemented in three different ways:
G7.1: *The scheduling system suggests values that lie in the domain of the variable.*
G7.2: *After each assignment, the scheduling system removes all values from the domains of the unassigned variables that cause a conflict with the assigned variables. It suggests values that lie in the reduced domain of the variable.*
G7.3: *After each assignment, the scheduling system removes all values from the domains of the unassigned variables that are not part of a solution extending the current partial schedule. It suggests values that lie in the reduced domain of the variable.*
Guideline G7.1 prevents the scheduler from using a trial-and-error strategy for value selection, if he doesn't know the domain of a variable. Guideline G7.2 is useful if an assignment has to be changed in an existing schedule: The computer only suggests values that do not cause the violation of constraints. With guideline G7.3 the computer takes over the propagation of assignments. Values that will cause empty domains in future assignment steps are removed from the domain of the considered variable. If the scheduler selects one of the suggested values, he can be sure, that no backtracking will be necessary later. This way the

effort of knowledge-based scheduling is reduced. Furthermore, G7.3 allows the scheduler to delay the automatic completion of the schedule until the implementation stage, in which the temporary constraints are fixed (G5).
This guideline can be implemented with algorithms that achieve k-consistency with $k \geq 2, k \leq N$ for a constraint optimization problem with N variables [HW07]. The higher k the more values can be removed from the domains but the higher is the required computing time. For this reason, in practice it can be required to make a compromise, where as many values as possible are removed in an acceptable time.

3.2.8 Guideline G8 - Weak fixation

The scheduling system should allow the scheduler to define temporary constraints from which the computer can deviate during automated scheduling if the problem is otherwise not solvable (weak fixation). Thereby, the deviation from the original assignments should be as small as possible.
The weak fixation can be used to find a set of temporary constraints, for which the scheduling problem is solvable. If the domain of a variable does not contain the desired value any more because it was removed during the propagation of previous assignments, the scheduler has two options: He could fix the desired assignment of this variable normally and the assignments of the previously considered variables weakly. This way the values of the latter variables are adapted during automatic scheduling. Alternatively he could fix the values of all variables weakly and let the computer find a compromise with a small deviation. In both cases the scheduler is relieved from the effort of manual backtracking.

3.2.9 Guideline G9 - Extended support for value selection

If the scheduler wants to assign an value from the domain of a variable that is inconsistent with the current state of the schedule, the scheduling system should determine the assignment level, on which the schedule has to be changed in order to enable this assignment.
Guideline G9 helps to assess the extent to which an existing schedule has to be changed in order to enable an assignment currently violating one or more constraints. For example, consider a fleet scheduling problem where the scheduler wants to assign a job to a different resource. Three different situations can be distinguished, if there is no gap in the tour of this resource, where the job can be inserted:

- *Change on level 4:* In the tour of this resource there is a gap which can be widened by adjusting the start times of the jobs. However, the existing order of the jobs can remain unchanged.
- *Change on level 3:* The order of the jobs in the tour of this resource has to be changed before the new job can be inserted.
- *Change on level 2:* One or more jobs have to be removed from the tour before the new job can be inserted.

The required changes might influence the preference of the scheduler. If they are too big, the scheduler possibly does without the assignment.

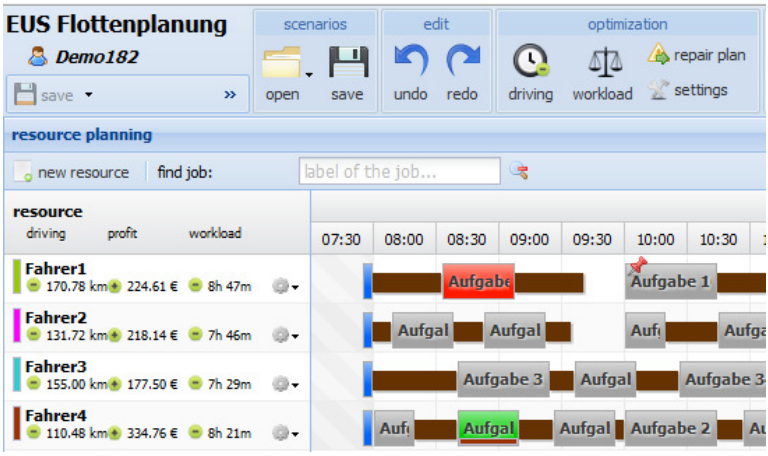


Figure 3: User interface for fleet scheduling.

4 Evaluation of the Guidelines

4.1 Test hypothesis

Conventional scheduling systems often provide one of these interaction models: *manual scheduling* (M1), *full automatic scheduling with subsequent manual modification of the schedule* (M2), *provision of alternative schedules by the computer*. The latter model is not considered in the test, as it doesn't allow knowledge-based scheduling. Models M1 and M2 allow knowledge-based scheduling, but they do not include computer support to reduce the effort (see section 2.4). We assume that depending on the problem structure, it can be difficult to incorporate temporary constraints, such that the resulting schedule is both valid (no constraints are violated) and optimal. Our hypothesis is that an interaction model based on our guidelines provides a better user experience to achieve these properties.

4.2 Test setup and results

We have carried out a user test in order to prove our test hypothesis. In this test, 45 participants (students from our university) solved 6 typical scheduling tasks in the application area of fleet scheduling. They had to change the resource assignments, the order and the start times of certain jobs in complete schedules or they had to create new schedules incorporating similar preferences. Constraints to consider were start time constraints (time windows for jobs) and resource constraints (limited choice of resources for some jobs). Furthermore, the participants should try to achieve a good quality score.
The participants were provided with a scheduling system, whose user interface is depicted in figure 3. The system was configured in 5 different test models M1 to M5. Test models

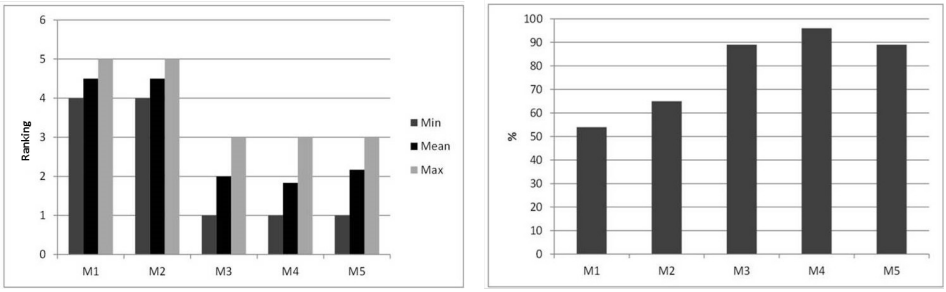


Figure 4: Ranking of the models with regard to the quality (left) and average task success (right).

M1 and M2 correspond to the interaction models M1 and M2 in section 4.1. Test models M3 to M5 represent our new interaction model in 3 different ways: In test model M3 the system provides fixation features at the levels 2 and 4[1]. The test models M4 and M5 additionally provided the scheduling functions *resource optimization 2, group optimization 1/2*, and *FIT-IN*. In order to encourage the use of the scheduling functions (see figure 2) *full optimization* was missing in M4. In all models a minimal support for manual scheduling was provided according to the guidelines G1 to G5 and G7.2[2]
Each scheduling task was solved using each model which allows us to compare the effectiveness of the models with regard to the metrics quality and task success. Figure 4 shows the overall ranking of the models with regard to the achieved quality scores (1 is the best, 6 the worst rank). The ranks were averaged over the six scheduling tasks. Models 3, 4 and 5 produce the shortest travel time. The worse quality achieved with M1 and M2 can be attributed to the high effort of manual scheduling. The participants often had to backtrack their assignments. Furthermore, in their effort to fulfill all constraints they were not able to simultaneously strive for a good quality.
A task was completed successfully, if all jobs were scheduled according to the given preferences and no constraints were violated. Figure 4 shows the distribution of the average success rate over all 5 models. The success rate is not more than 60%, if manual scheduling is required (M1, M2). The high success rate of the models M3 to M5 confirms our assumption that manual scheduling generally is inconvenient for humans.

4.3 Conclusions

The results show that the interaction models M1 and M2 are not sufficient for knowledge-based scheduling. They confirm our hypothesis that a high quality and task success rate can only be achieved if fixation facilities or scheduling functions at suitable assignment levels are provided.
As the participants were not scheduling experts, the preferences had to be given in the tasks. They did not carry out a search for constraints in the design stage, thus the guidelines

[1] Fixation at levels 1 and 3 was not necessary to perform the tasks.
[2] In a previous test nearly all participants failed if no support for value selection or only G7.1 was provided.

G7.3, G8 and G9 could not be evaluated. This has to be done in a field study.

References

[BBIM96] Mark H. Burstein, Bolt Beranek, Newman Inc, and Drew V. Mcdermott. Issues in the development of human-computer mixed-initiative planning. In *Cognitive Technology*, pages 285–303. Elsevier, 1996.

[BZE+07] Keith A. Butler, Jiajie Zhang, Chris Esposito, Ali Bahrami, Ron Hebron, and David E. Kieras. Work-centered design: a case study of a mixed-initiative scheduler. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 747–756, 2007.

[CvW11] Julien Cegarra and Wout van Wezel. Allocating Functions to Human and Algorithm in Scheduling. In Jan C. Fransoo, Toni Waefler, and John R. Wilson, editors, *Behavioral Operations in Planning and Scheduling*, pages 339–370. Springer Berlin Heidelberg, 2011.

[CvW12] Julien Cegarra and Wout van Wezel. Revisiting Decision Support Systems for Cognitive Readiness: A Contribution to Unstructured and Complex Scheduling Situations. *Journal of Cognitive Engineering and Decision Making*, 6(3):299–324, 2012.

[Dar91] Francoise Darses. The constraint satisfaction approach to design: A psychological investigation. *Acta Psychologica*, 78(1-3):307–325, 1991.

[dNE05] Hugo A.D. do Nascimento and Peter Eades. User hints: a framework for interactive optimization. *Future Generation Computer Systems*, 21(7):1177–1191, 2005.

[GFW11] Roland Gasser, Katrin Fischer, and Toni Wäfler. Decision Making in Planning and Scheduling: A Field Study of Planning Behaviour in Manufacturing. In Jan C. Fransoo, Toni Wäfler, and John R. Wilson, editors, *Behavioral Operations in Planning and Scheduling*, pages 11–30. Springer Berlin Heidelberg, 2011.

[HB10] Sönke Hartmann and Dirk Briskorn. A survey of variants and extensions of the resource-constrained project scheduling problem. *European Journal of Operational Research*, 207(1):1–14, 2010.

[HGM12] Jean-Michel Hoc, Clément Guerin, and Nasser Mebarki. The nature of expertise in scheduling: The case of timetabling. *Human Factors and Ergonomics in Manufacturing and Service Industries*, 2012.

[Hig01] Peter G. Higgins. Architecture and interface aspects of scheduling decision support. In *Human performance in planning and scheduling*, pages 245–280. Taylor & Francis, 2001.

[HW07] P. Hofstedt and A. Wolf. *Einführung in die Constraint-Programmierung: Grundlagen, Methoden, Sprachen, Anwendungen*. Springer, 2007.

[Mac08] M. MacDonald. *Your Brain: The Missing Manual*. Missing manual. O'Reilly Media, 2008.

[SFG+12] Hartmut Staedler, Bernhard Fleischmann, Martin Grunow, Herbert Meyr, and Christopher Sürie. *Advanced Planning in Supply Chains - Illustrating the Concepts Using an SAP APO Case Study*. Springer Berlin Heidelberg, 2012.

[TSD11] Efraim Turban, Ramesh Sharda, and Dursun Delen. *Decision Support and Business Intelligence Systems*. Pearson, 2011.

[Vis06] Willemien Visser. Designing as construction of representations: a dynamic viewpoint in cognitive design research. *Human-Computer Interaction*, 21(1):103–152, 2006.

Studentische Ausbildung und berufliche Weiterbildung in Angewandter Softwarequalitätssicherung

(SAG WAS)

SAG WAS – Studentische Ausbildung und berufliche Weiterbildung in Angewandter Softwarequalitätssicherung

Durch die Diskussion über agile Praktiken wie Test Driven Development (TDD) ist das Thema Softwarequalitätssicherung stärker in den Fokus der Softwareentwicklung gelangt. Doch wie bereiten wir Personen in der beruflichen Praxis vor, damit sie solche Praktiken gewinnbringend für die Qualität der Software einsetzen können? Mit Büchern? In Kursen? Learning by doing? Wie wird in der Hochschullehre darauf reagiert? Wie werden dort die Konzepte einer systematischen Qualitätssicherung vermittelt? In Vorlesungen? Vertieft durch studentische Projekte? In dem diesjährigen SAG WAS-Workshop suchen wir wieder den Austausch zwischen den beiden "Welten" der Vermittlung.

In der ersten Sitzung beginnen wir mit dem Thema Softwarequalitätssicherung in der beruflichen Weiterbildung. Norbert Kastner stellt mit seinem Vortrag „Berufliche Weiterbildung – Serum gegen Fehlerseuche“ verschiedene Weiterbildungskonzepte vor. Während Jörn Münzel die Frage stellt „Karriere als Tester – ist das möglich?“. Die zweite Sitzung setzt den Fokus auf studentische Ausbildung. Ulrich Hafner wird seine Erfahrung aus einer sehr praxisorientierten Veranstaltung im Masterstudium „Erfolgreiche Wissensvermittlung durch Learning by Doing in einem Open Source Projekt“ zur Diskussion stellen. Jens Grabowski und Steffen Herbold referieren, wie „Zertifizierungen als Teil der Universitären Ausbildung in der Software Qualitätssicherung“ integriert werden können. In der dritten Sitzung kommen wir zu den Anforderungen im täglichen beruflichen Alltag zurück. Die Frage wie Kollegen und Kolleginnen in den Projekten auf das Thema Softwarequalitätsicherung vorbereitet werden, steht im Vordergrund. Gerhard Müller und Alexandra Schladebeck beschreiben ihre Erfahrungen wie alle „Lernen von Kollegen“ und „Gemeinsam zur Qualität“ kommen.

In der abschließenden Postersitzung werden von Axel Böttcher, Veronika Thurner und Daniela Zehetmeyer, Barbara Dörsam, Karin Vosseberg sowie Andreas Spillner verschiedene Lehrveranstaltungskonzepte zum Thema Softwarequalitätssicherung vorgestellt. Die Konzepte liegen im vorliegenden Tagungsband als erweiterte Abstracts vor.

Programm- und Organisationskomitee

Prof. Dr. Axel Böttcher, Hochschule München,
Dr. Klaudia Dussa-Zieger, Method Park Erlangen,
Prof. Dr. Ina Schieferdecker, Freie Universität Berlin,
Prof. Dr. Andreas Spillner, Hochschule Bremen,
Prof. Dr. Veronika Thurner, Hochschule München,
Stefan Vogel, eXept Software AG und ASQF FG Softwaretest,
Prof. Dr. Karin Vosseberg, Hochschule Bremerhaven (Chair)

Der frühe Tester fängt den Bug

Axel Böttcher, Veronika Thurner, Daniela Zehetmeier*

Fakultät für Informatik und Mathematik – Hochschule München
Lothstraße 64 – 80335 München
a.boettcher@cs.hm.edu, thurner@hm.edu, daniela.zehetmeier@hm.edu

Abstract: Das Thema *Testen* spielt derzeit in der Grundausbildung der Softwareentwicklung oft nur eine untergeordnete Rolle. Entsprechend entwickeln die Studierenden wenig Erfahrung im Erstellen von Unit-Tests und kaum Bewusstsein für deren Bedeutung. Durch den Ansatz "Objects first, Tests second" rückt die Thematik des automatisierten Testens in den Vordergrund der Ausbildung in Softwareentwicklung.

1 Motivation und Grundproblem

In vielen Informatik-nahen Studiengängen nimmt die Softwareentwicklung einen hohen Anteil der Grundausbildung ein. Dabei werden meist allgemeine Kernkonzepte wie Variablen und Kontrollstrukturen behandelt, sowie Objektorientierung und weiterführende Themen wie Vererbung, Collections und Ausnahmen. Das Thema *Testen* kommt dabei – wenn überhaupt – erst gegen Ende der Grundausbildung in Softwareentwicklung zur Sprache. Auch die Lehrbuchlandschaft spiegelt dieses Phänomen wider. Nach unseren Recherchen gibt es auf dem Markt der Lehrbücher zur (objektorientierten) Softwareentwicklung kaum Literatur, die das Thema *Testen* von Beginn an in den Vordergrund stellt.[1]

In der Folge haben viele Studierenden das Testen in der Grundausbildung entweder gar nicht kennen gelernt, oder nur als eines von vielen weiterführenden Themen meist am Ende des zweiten Semesters. Entsprechend gering ist die gewonnene praktische Erfahrung. Ferner messen die Studierenden dem Testen oft nur eine untergeordnete Bedeutung bei – denn schließlich ging die Grundausbildung ja über weite Teile auch *ohne* Testen.

2 Zielsetzung und Vorgehensweise

Um diesem Problem entgegenzusteuern entwickeln wir einen Lehransatz, der Unit-Tests als zentralen Inhalt bereits in der Grundausbildung zur Softwareentwicklung verankert. Unser Ziel ist dabei, den Studierenden frühzeitig praktische Erfahrung im Testen zu vermitteln, den Nutzen des Testens für die Lösungsqualität erlebbar zu machen und so das Unit-Testen zu einer Selbstverständlichkeit werden zu lassen.

Dazu haben wir im Wintersemester 2013/14 die Veranstaltung "Softwareentwicklung 1" nach dem Ansatz "Objects first, Tests second" durchgeführt. Das heißt, wir haben zunächst die Grundidee und die Kernkonzepte der Objektorientierung eingeführt, sowohl anschaulich als auch implementatorisch (in Java). Unmittelbar danach führten wir die Grundlagen

*Gefördert durch das BMBF Förderkennzeichen 01PL11025 (Projekt "Für die Zukunft gerüstet"), im Programm "Qualitätspakt Lehre"

[1] Ausnahme: Jeff Langr. *Agile Java – Crafting Code with Test-Driven Development*. Pearson Education, 2005.

von Unit-Tests ein. Diese wurden dann über die Veranstaltung hinweg sukzessive weiter ausgebaut und erweitert, jeweils passend zu den gerade behandelten Konstrukten. So wurde z.B. in der Lerneinheit zu *Arrays* die Methode `assertArrayEquals()` eingeführt.

Im Rahmen des Praktikums mussten die Studierenden von Beginn an ihre Lösungen so gestalten, dass sie testbar sind, sowie entsprechende Unit-Tests mit entwickeln. Ergänzend gaben die Lehrenden bei einigen Aufgaben die zu erfüllenden Tests als Teil der Anforderungsdefinition mit vor, hinführend auf die Praxis der testgetriebenen Entwicklung.

3 Erfahrungsbericht und Lessons Learned

Eine wesentliche Herausforderung des Ansatzes bestand darin, alle in der Lehrveranstaltung verwendeten Beispiele und Aufgaben so zu gestalten, dass sie konsequent testbar sind. Insbesondere müssen dabei die erstellten Methoden als Ergebnis einen Datenwert zurückliefern und nicht lediglich irgendwelche Informationen auf der Konsole ausgeben. Entsprechend wurde das Konstrukt `System.out.println()` nahezu bedeutungslos.

Diese Forderung schränkt die mögliche Begleitliteratur drastisch ein, da kaum ein anfängertaugliches Java-Lehrbuch ohne exzessive Nutzung der Konsolenausgabe auskommt. Diese wird überwiegend dazu verwendet, um die im Programm aktiven Objekte und deren Veränderung über die Zeit für den Lernenden sichtbar darzustellen.

Verzichtet man auf die testunfreundliche (und praxisferne) Verwendung der Konsolenausgabe, fällt dieser Einblick in den Programmzustand weg, und damit ein Teil der Anschaulichkeit. Als Ersatz wurden GUIs zu den Beispielen entwickelt, die die Objekte darstellen.

Insgesamt haben die Studierenden am Ende des ersten Semesters das Erstellen von Unit-Tests als zentralen Bestandteil jeglicher Softwareentwicklung akzeptiert und verinnerlicht. Dabei beherrschen Sie die behandelten Grundtechniken im Wesentlichen sicher.

Ein weiterer, indirekter Nutzen der frühzeitigen Einführung von Unit-Tests zeigte sich bei der automatisierten Bewertung der studentischen Lösungen im Praktikum. Diese basiert auf Unit-Tests, die die Lehrenden zu den einzelnen Übungsaufgaben entwickelt haben. Da die Studierenden mit der dahinter liegenden Technik vertraut sind war der automatisierte Bewertungsmechanismus für sie sofort nachvollziehbar und wurde gut akzeptiert.

4 Weiterführende kritische Fragen für den nächsten Durchlauf

Obwohl der "Objects first, Tests second"-Ansatz bereits beim ersten Versuch recht erfolgreich war, bieten einige Punkte beim nächsten Durchlauf Raum für Verbesserungen im didaktischen Ansatz. Die folgenden Fragestellungen spiegeln dies wider:

- Zu welchem Zeitpunkt und wie intensiv behandelt man die Frage, wieviele (und welche) Tests geschrieben werden müssen?
- Wie schafft die Lehrperson im Praktikum eine ausgewogene Balance zwischen dem Aufwand für die gezielte Verwendung neuer Konstrukte, die Konzeption der Logik und dem zugehörigen Testaufwand?
- Auf welche Weise werden allgemein Unit-Tests an sich qualitätsgesichert, und wie im Speziellen die von den Studierenden im Rahmen ihrer Aufgaben erstellten Tests?

Diese werden wir im nächsten Durchlauf der Veranstaltung gezielt angehen.

Softwaretests- und Software-Qualitätsthemen in der Hochschullehre

Prof. Dr. Barbara Dörsam

Digital Publishing
Hochschule der Medien
Nobelstr. 10
70569 Stuttgart
doersam@hdm-stuttgart.de

Abstract: Im folgenden Beitrag wird ein Ansatz zur praxisorientierten Lehre von Softwaretest- und Software-Qualitätsthemen an der Hochschule der Medien vorgestellt. Basierend auf der Analyse, warum viele Studenten oft keinen Zugang zu Testthemen finden, wird eine Vorgehensweise vorgestellt, die den Studenten einen offenen Umgang mit den Themen und einen Erkenntnisgewinn aus praktischen Beispielen ermöglicht.

1 Softwaretests im Hochschulunterricht

Während in industriellen Projekten die Themen Softwarequalität und Softwaretests immer mehr an Bedeutung gewinnen, sind an Hochschulen Veranstaltungen zu diesen Themen noch selten und dabei oft sehr theoretisch aufgebaut. Die Folge ist, dass Hochschulabsolventen relativ wenige Kenntnisse über Tests in die Firmen mitbringen. Da das Thema Testentwicklung trotz neuester Trends in vielen Unternehmen noch nicht denselben Stellenwert wie die produktive Softwareentwicklung hat, verfügen auch erfahrene Softwareentwickler oft nur über unzureichende Kenntnisse von Test- und Qualitätssicherungsmethoden in der Softwareentwicklung.

In diesem Diskussionsbeitrag werden neue Lehrmethoden für Test- und Qualitätsthemen vorgestellt, die an der HdM Stuttgart bereits erfolgreich eingesetzt werden und sich auch in die industrielle Praxis umsetzen lassen. Dabei stehen die folgenden Herausforderungen beim Lehren von Test- und Softwarequalitätsthemen im Vordergrund:

1. Die vorgefassten Meinungen der Studenten zum Thema Softwaretests: Studentische Projekte haben oft einen Prototypcharakter, da sie nie die Produktreife erreichen. Dadurch sammeln Studenten zwar viel Erfahrung in der Entwicklung von Software, aber so gut wie kein Feedback über die - oft nicht zufrieden stellende - Qualität ihrer Ergebnisse. Mit einer wachsenden Anzahl der durchgeführten Projekte sehen die Studenten immer weniger die Notwendigkeit von Tests für die eigene Software.

2. In der Theorie hören sich Testthemen oft einfach und nachvollziehbar an. Praktische Beispiele werden von den Studenten daher oft als konstruiert erlebt. Dies wird noch dadurch verstärkt, dass Studenten, die lediglich Erfahrung mit ihrer eigenen – ungetesteten – Software haben, selten die Erfahrung gemacht haben, dass auch einfache Software zahlreiche Fehler enthalten kann.
3. Daher sehen Studenten oft nur den Aufwand, den das zusätzliche, scheinbar unproduktive, Testen mit sich bringt, aber nicht den Nutzen für die eigene Arbeit.

2 Softwaretests – Lehre an der HdM

Als eine Möglichkeit, die oben aufgeführten Herausforderungen zu meistern, wird die explorative Lehrmethode vorgestellt, die im Masterstudiengang Computer Science and Media an der HdM Stuttgart eingesetzt wird. Den Kern der Methode stellt ein dreistufiges Verfahren dar:

1. Die Studenten fangen jedes Thema mit einer einfachen Aufgabe an. Bei den Aufgaben handelt es sich entweder um fertige Software, die sie testen müssen, oder um Beispiele, die sie erst entwickeln müssen. Dabei sind die Beispiele so beschrieben, dass die Wahrscheinlichkeit, bei der Entwicklung Fehler in die Software einzubauen, sehr hoch ist. Dadurch merken die Studenten bei der eigenen Arbeit,
    - dass es unter bestimmten Bedingungen nicht einfach ist, Tests so zu entwickeln, dass diese tatsächlich die relevanten Fehler finden, oder auch,
    - dass auch kleine, scheinbar einfache Entwicklungsaufgaben viele Fehlerquellen enthalten können.
2. Wurde dadurch einmal der vorgefassten Meinung, Tests wären einfach oder gar unnötig, widersprochen, kann eine konstruktive Diskussion der Ergebnisse und Vorstellung der theoretischen Ansätze für die behandelten Probleme mit den Studenten stattfinden.
3. Anschließend wiederholen die Studenten den ersten Schritt, allerdings mit komplexeren Beispielen und unter Anwendung der gerade erarbeiteten theoretischen Ansätze. In diesem Schritt vertiefen sie die gerade erlernte Theorie – allerdings mit der verfestigten Überzeugung, dass dieses Wissen für sie auch in der Praxis eine große Bedeutung hat.

3 Softwaretests – Potential für die Diskussion

Im Workshop können drei Aspekte dieses Ansatzes aufgegriffen werden:

- Konkrete Vorstellung und die Diskussion des Ansatzes an ausgewählten Beispielen, z.B. aus dem „Test Driven Development“ und am Beispiel von Testfällen für objektorientierte Ansätze.

- Diskussion der – positiven wie negativen – Erfahrungen aus der Lehrpraxis.

- Aufzeigen eines Weges, über den die Erfahrungen aus der Lehre in die Industriepraxis und Weiterbildung der eigenen Mitarbeiter übertragen werden kann.

Studienprojekte als Experimentierraum für QS

Karin Vosseberg

Hochschule Bremerhaven
Studiengänge Informatik und Wirtschaftsinformatik
An der Karlstadt 8
27568 Bremerhaven
karin.vosseberg@hs-bremerhaven.de

Abstract: Der „Test-First-Ansatz“ und damit eng verbunden eine Testautomatisierung wird mit agilen Vorgehensmodellen in der Softwareentwicklung vorangetrieben. Testautomatisierung ist jedoch mehr als nur Testfälle implementieren. Was müssen Studierende lernen, um Testautomatisierung gewinnbringend zur Verbesserung der Softwarequalität einzusetzen.

1 Motivation

Mit der Diskussion über „Test Driven Development“ in Verbindung mit agilen Vorgehensmodellen rücken die Aufgabenbereiche der Rollen „Tester“ und „Entwickler“ immer stärker zusammen. Testen wird nicht länger als separate Aufgabe am Ende des Projekts gesehen sondern als gemeinsame Projektaufgabe zur Verbesserung der Qualität des Softwareprodukts. Mit Aufgaben der Testautomatisierung brauchen Tester verstärkt auch Softwareentwicklungskompetenzen sowie Entwickler Wissen über systematisches Testen [VS13].

2 Studentische Projekte

In einjährigen Bachelorprojekten habe ich an der HS Bremerhaven eine Umgebung geschaffen, in der Studierende der Studiengänge Informatik und Wirtschaftsinformatik mit einem „Test Driven Development“-Ansatz in der Softwareentwicklung experimentieren konnten. Ziel der Projekte war im Rahmen der Entwicklung einer kleinen Applikation die Testautomatisierung der Komponenten-, Integrations- und Systemtestfälle bereits von Beginn des Projekts an zu berücksichtigen. Im Fokus der Projekte stand das Experimentieren mit Werkzeugen für die Einrichtung einer adäquaten Entwicklungs- und Testumgebung. Um eine realistische Projektumgebung zu simulieren, waren externe Auftraggeber involviert, z.B. Bredex GmbH, Braunschweig oder die Forschungstransferstelle der HS Bremerhaven.

Neben Eclipse als Entwicklungsumgebung haben die Studierenden Jubula, Junit und DBUnit zur Implementierung von Testfällen auf den verschiedenen Teststufen eingesetzt. Die Studierenden haben schnell erkannt, dass Testautomatisierung mehr ist als Testfälle zu implementieren. Testautomatisierung ist eng verbunden mit Konzepten der Versionskontrolle, des Daily-Builds und des Continous Integration. Daraufhin haben sie zusätzlich mit Werkzeugen wie Jenkins, Subversion, Maven, EclEmma eine Werkzeugkette für ihre nächtlichen Testläufe aufgebaut.

Flankiert wurden die Aktivitäten mit einer Vorlesung zum Thema „Systematisches Testen“ auf Basis des internationalen Lehrplans zum ISTQB® Certified Tester Foundation Level [CT11]. Die Studierenden hatten die Möglichkeit hier zusätzlich das Zertifikat zu erwerben.

Fazit

Projekte bieten genügend Spielraum, um die praktischen Erfahrungen in der Softwarequalitätssicherung sammeln zu können. Die Studierenden konnten ihre Projektergebnisse, z.B. auf der iqnite 2010 [VF10] oder dem Eclipse Testing Day 2011 [SWV11] vorstellen. Ein externer Auftraggeber ist sehr hilfreich, um den Realitätsbezug herzustellen, aber führt auch dazu, dass durch Termindruck die Entwicklung - wie in realen Projekten - wieder stärker in den Vordergrund rückt. Nach wie vor schwierig gestaltet sich die Frage, wie der theoretische Input zum systematischen Testen in die Bewältigung der Aufgabe zu integrieren ist, ohne das Projekt zu überfrachten. Aktuell lege ich in der Veranstaltung „Grundlagen Qualitätsmanagement“ den Fokus auf die Vermittlung der methodischen Grundlagen und versuche im Übungsbetrieb gemeinsam mit Kollegen der HEC GmbH, Bremen die Testautomatisierungskette zu thematisieren.

Literaturverzeichnis

[CT11] Certified Tester Foundation Level Syllabus, Version 2011 1.0.1, www.german-testing-board.info/service/information/lehrplaene.html#c72 25.4.2014.

[SWV11] Markus Sagurna, Jonas Wolniczak, Karin Vosseberg: Testautomatisierung – mehr als nur schreiben von Testfällen, Eclipse Testing Day 2011, Neuss.

VF10] Karin Vosseberg, Marco Fischer: Evaluation of Test Tools – a Student Education Project. iqnite 2010, Düsseldorf.

[VS13] Karin Vosseberg, Andreas Spillner: Traditionell cum Agil. Am Beispiel Test-Driven Development. ASQF SQ-Magazin, Juni 2013.

Certified Tester: Ein Lehrplan für berufliche Weiterbildung und studentische Ausbildung

Andreas Spillner

German Testing Board & Hochschule Bremen
Flughafenallee 10
28199 Bremen
andreas.spillner@hs-bremen.de

Abstract: Der Test von Software hat in den letzten Jahren immer mehr an Bedeutung gewonnen und somit auch die Aus- und Weiterbildung im diesem Bereich. Seit nahezu 15 Jahren existiert ein weltweit akzeptiertes Ausbildungsschema: der »ISTQB® Certified Tester«. In Deutschland besteht die Möglichkeit, den Lehrplan »Foundation Level« auch an den Hochschulen zu unterrichten. Das Ausbildungsschema und Lehrerfahrungen werden vorgestellt.

1 ISTQB® Certified Tester

Professionelles Prüfen und Testen von Software erfordert eine fundierte Ausbildung. Mit dem ISTQB® Certified Tester Programm existiert ein weltweit anerkanntes, standardisiertes Weiterbildungsschema für Softwaretester. „Das ISTQB (International Software Testing Qualifications Board, [ISTQB]) ist ein internationaler Dachverband, der sich dem Ziel der Professionalisierung des Testens widmet, d. h. der Etablierung des Testens als eigenständiger Disziplin und dessen Fortschreibung entlang neuer Erkenntnisse und Methoden. National vertreten wird der ISTQB durch Verbände, in Deutschland ist dies exklusiv das German Testing Board (GTB, [GTB]).“ [Si13] Das Schema besteht aus drei Ausbildungsstufen: Foundation Level, Advanced Level und Expert Level. Prüfungen zu den einzelnen Lehrinhalten werden von unabhängigen Zertifizierungsstellen angeboten und durchgeführt.

Das Weiterbildungsschema Certified Tester ist international sehr erfolgreich und ist in 70 Ländern vertreten. Knapp 30 Trainingsprovider bieten in Deutschland Weiterbildungskurse im Rahmen des Certified Tester Programms an. Derzeit gibt es in Deutschland 33.000 zertifizierte Tester. Weltweit sind es 320.000 Personen, die eine Prüfung abgelegt und bestanden haben.

2 Lehrveranstaltungen

An mehr als 30 Hochschulen in Deutschland gibt es Lehrveranstaltungen auf Basis des Certified Tester Lehrplans »Foundation Level«, der die Grundlagen zum Testen von Software umfasst. Folgende Themenblöcke werden behandelt: Grundlagen des Softwaretestens, Testen im Softwarelebenszyklus, Statischer Test, Testfallentwurfsverfahren, Testmanagement und Testwerkzeuge. Unter anderem werden Kurse an der Technische Universität Berlin, der Universität Göttingen, der Fachhochschule Köln, der HAW Hamburg und der Hochschule Bremen regelmäßig angeboten. Das GTB stellt den Lehrenden einen kostenlosen Foliensatz zur Verfügung, der alle Inhalte des Lehrplans umfasst. Die Studenten können im Anschluss an die Vorlesung die Prüfung zum Certified Tester »Foundation Level« ablegen. Bisher haben 1.700 Studierende diese Möglichkeit wahrgenommen und das Zertifikat erhalten.

3 Fazit

Sowohl bei den Weiterbildungsangeboten für Berufstätige als auch bei den Kursen an den Hochschulen wird der Lehrplan »Foundation Level« unterrichtet ([CT11], [SL12]). Damit existiert eine direkte Brücke zwischen den beiden Ausbildungswelten. Die Studierenden erhalten eine Ausbildung, die in der Praxis anerkannt ist und einen guten Start für die berufliche Laufbahn bietet.

Das GTB kooperiert mit dem Testing Board von Malaysia, um das erfolgreiche Hochschulprogramm auch in Malaysia zu etablieren. Kurse gab es bereits an knapp 10 Universitäten. Die verpflichtende Verankerung des ISTQB-Lehrplan in das Curriculum der Universitäten mit Informatikausbildung wird angestrebt. In Deutschland sind die ISTQB-Kurse in der Regel Wahl- und keine Pflichtfächer. Aber vielleicht ändert sich in naher Zukunft auch etwas.

Referenzen

[CT11] Certified Tester Foundation Level Syllabus, Version 2011 1.0.1, www.german-testing-board.info/service/information/lehrplaene.html -> Lehrplan Foundation Level, 12.5.2014

[GTB] German Testing Board, www.german-testing-board.info, 12.5.2014

[ISTQB] International Software Testing Qualifications Board, www.istqb.org, 30.4.2014

[Si13] Frank Simon: ISTQB-Certified-Tester: Eine erfolgreiche Geschichte in vier Akten. Objektspektrum Online Themenspezial Zertifizierungen 2013, www.sigs-datacom.de/fachzeitschriften/objektspektrum/archiv/artikelansicht.html?showYear=2013 -> Ausgabe Zertifizierungen/2013, 12.5.2014

[SL12] Andreas Spillner, Tilo Linz: Basiswissen Softwaretest. Aus- und Weiterbildung zum Certified Tester - Foundation Level nach ISTQB-Standard, dpunkt.verlag, 5., überarbeitete und aktualisierte Auflage, 2012

System Software Support for Big Data

(BigSys)

System Software Support for Big Data (BigSys) *

Abstract: Big Data is one of the most important challenges for computer science research in the upcoming decade. Most research efforts so far concentrate on algorithms for analyzing and processing Big Data. The system software and infrastructure implications of Big Data processing are so far underrepresented. We intend to improve on this situation with our BigSys workshop. In this workshop, Big Data-related topics from all areas of system software, such as operating systems, middleware, and databases, as well as related topics from the area of system architecture are to be considered. Of special interest are the impact of high-performance architectures for Big Data, system-level methods to improve data processing performance, challenges in system software design, and architecting systems for Big Data in general. We would especially like to encourage contributions on topics covering multiple layers of the hardware/software stack.

The aim of this workshop is to bring together practitioners and researchers from all of these areas to discuss today's limitations, promising new concepts and possible new synergies in the domain of system software support for Big Data.

Organizers: Michael Engel (School of Computing, Creative Technologies & Engineering, Leeds Beckett University), Ulf Hollberg (IBM Deutschland Management & Business Support GmbH), Wilhelm Messing (FUJITSU PDG ES&S SWE SOL), Olaf Spinczyk (Embedded System Software Group, Technische Universität Dortmund)

Program Committee: Michael Engel (Leeds Beckett University), Ulf Hollberg (IBM), Wilhelm Messing (Fujitsu), Andreas Polze (HPI Potsdam), Kai-Uwe Sattler (TU Ilmenau), Olaf Spinczyk (TU Dortmund), Jens Teubner (TU Dortmund)

*http://ess.cs.uni-dortmund.de/workshops/bigsys/

Unleashing the Hidden Power of Integrated-GPUs for Database Co-Processing

Edward Ching, Norbert Egi, Masood Mortazavi, Vincent Cheung, Guangyu Shi

FutureWei Technologies

Abstract

Modern high-performance server systems with their wide variety of compute resources (i.e. multi-core CPUs, GPUs, FPGAs, etc.) bring vast computing power to the fingertips of researchers and developers. To investigate modern CPU+GPU co-processing architectures and to discover their relative marginal return on resources ("bang for the buck"), we compare the different architectures with main focus on database query processing, provide performance optimization methods and analyze their performance using practical benchmarking approaches. Discrete GPU devices (d-GPUs) are connected to host memory through PCIe which is relatively slower than the internal bus which connects integrated GPU subsystems (i-GPUs) to the CPU and its caches. Our experimental analysis indicates that while massive processing capabilities have grown in d-GPUs, data locality, memory access patterns and I/O bottlenecks continue to govern and impede the speed of database processing functions. While the use of GPU systems (whether discrete or integrated) lead to significant speed-up in primitive data processing operations, the marginal performance returns on deploying teraflops d-GPU compute resources do not appear to be large enough to justify their cost or power consumption. At this point, integrated GPUs provide better cost-performance and energy-performance results in parallel database query processing.

1 Introduction

Discrete Graphics Processing Units (d-GPUs) have long been used for accelerating the execution of a wide-range of applications, most generally known in the space of scientific high-performance computing (HPC) [WGHP11, GS11]. In recent years, CPU+GPU co-processing have widely been adopted to improve processing performance of data analytic applications handling very large data sets using main memory database management systems [WZHC10, HLH13, HLY+09, PMK11, KLMV12, HSP+13, Bøg13]. While these GPUs posses an enormous amount of compute power, they typically connect to the CPU and the host memory through a relatively slow interconnect, such as PCI-Express (PCIe). Compared to HPC, database query processing requires less computation but on significantly more data where the performance of the interconnect between the CPU, host memory and the GPU has a more significant effect on the application's performance. However, recent CPUs with integrated GPUs (i-GPU) enabled for general-purpose computing provide an alternative architecture. While the computational power of the GPU cores is lower than that of d-GPUs, they are connected to the CPU and its caches through an internal bus

which is significantly faster than PCIe.

Earlier studies in parallel query processing generally focused on two kinds of systems: multi-core CPUs or many-core d-GPUs accessible through a PCI-e connection. Most studies on d-GPU co-processing have indicated PCI-e bandwidth issues as the main bottleneck in off-loading parallel computation to these co-processors [HLY+09, KLMV12, PMK11]. Micro-benchmarks and our own independent investigations in various settings have also confirmed this issue. To resolve this issue, Ocelot uses a memory management facility to transparently "pin" data arrays on d-GPU memory and to dynamically control and optimize the eviction of these data arrays. More recent work has paid greater attention to the use of i-GPU-based, fine-grained co-processing [HLH13, TK13] and have concluded that i-GPUs are promising alternatives to d-GPUs for data-intensive applications and functionalities, such as the JOIN operation. The work we report here is primarily experimental and focuses on the latter approach, i.e. on the use of modern i-GPUs for parallel data processing. We have selected benchmark queries and a set of primitive operations (essentially, micro-benchmarks) to compare the performance of the current, state-of-the-art systems. Our findings indicate a balancing trade-off (in gigaflops, I/O bandwidth and power consumption) that favors modern i-GPUs in query processing.

We focus on the discovery of system profiles and trade-offs that can help guide system designers to select among various architectural options for parallel processing of data analytic queries. We used modern multi-core, i-GPU and d-GPU systems and evaluated each system through the execution of analytic micro-benchmarks and a more complex TPC-H query. We target each system with implementations that are best suited to their architectural requirements. We found that high-gigaflops d-GPU computing systems may be an overkill when it comes to query processing and that performance gains for d-GPU compared to modern i-GPU systems may not justify the energy and capital costs for the purposes of data analytic processing. Next, we will discuss the relevant features of system architectures we investigated in Section 2, present the performance results in Section 3 and conclude in Section 4.

2 Architecture

In this section we analyze the key architectural differences of i-GPUs compared to d-GPUs and highlight their "hidden" potential for database query acceleration to be exploited by applications. Our analysis is based on two recent and widely-used GPUs, that is, Nvidia's GTX780 d-GPU and Intel's HD4600 i-GPU integrated into the Haswell i7-4770K CPU. Figure 1 illustrates the co-processing architecture using d-GPUs, while Figure 2 shows the internals of the Intel CPU with integrated GPU. The most performance critical differences between the two co-processing architectures are (1) the performance of interconnect(s) between the GPU, host memory and the CPU's caches, (2) the cache structure and (3) the computational capacity of the GPUs.

The performance of the interconnect(s) between the GPU and the location of the data prior to (i.e. data input) and right after (i.e. data output) being processed by the GPU becomes highly critical for data-intensive workloads that are collectively processed by the CPU and GPU, in which case the data typically resides in the CPU's cache or in the host memory and has to be moved from there to the GPU and finally back after the GPU processing

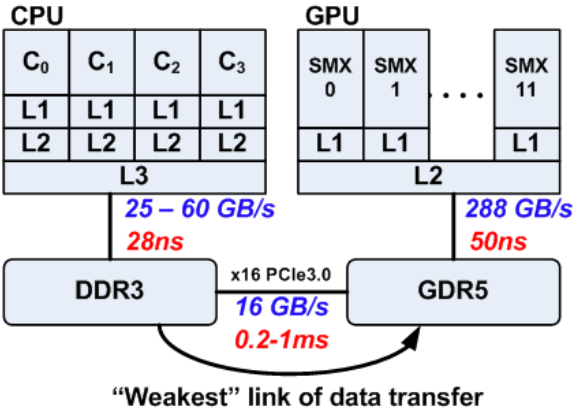


Figure 1: Discrete GPU architecture (NVidia GTX780)

has been completed. CPU+GPU co-processing of database queries demonstrate a data movement pattern very similar to that of above. After a query has been launched, a query execution plan is created and optimized for performance by the database management system (DBMS). The query execution plan determines the order the functionalities (i.e. query primitives) should be executed and the steps used to access data. Depending on the characteristics of functions in the execution plan, they provide better performance either on the CPU or GPU. Based on these characteristics and performance statistics of previous query executions, the optimizer decides which query primitives should be executed on the CPU and which ones on the GPU. The optimizer also aims for minimal amount of data transfer between the CPU and the GPU. However, even in the best case, data that is needed for processing at the GPU traverses over the interconnect between the CPU and GPU twice (once in each direction).

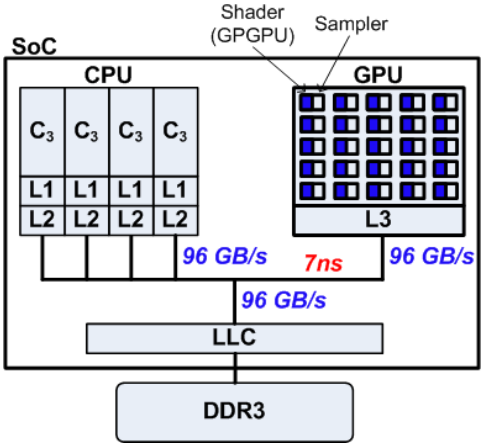


Figure 2: Integrated GPU architecture (Intel i7-4770K)

While d-GPUs connect to the CPU and host memory via a relatively slow (compared to memory speed) PCIe interconnect, i-GPUs connect to the CPU via a fast internal bus. In addition, i-GPUs have direct access to the Last-Level Cache (LLC) within the CPU die, providing even faster access to the data if it is present in the cache, a typical scenario with CPU+GPU co-processing. In our Intel i7-4770K chip the CPU and GPU shares the 8MB of LLC.

In addition to the speed of the interconnect between the CPU and GPU and whether the GPU has direct access to the CPU's cache, the GPU cores' computational capacity is a major contributing factor to applications' performance. In Table 1 we compare the specification of the above mentioned Nvidia GTX780 (d-GPU) and Intel HD4600 GPUs (i-GPU). The important parameters to note are, that the d-GPU outperforms the i-GPU by over a factor of 9x in terms of floating point operations. However, the i-GPU has nearly twice as many data lanes [1] it can run instructions on (i.e. maximum physical occupancy) and its clock speed is 25% higher than that of the d-GPU. However, the d-GPU has access to 3GB of local very fast GDR5 memory while the i-GPU's Last Layer Cache is only 8MB.[2]

	GTX780	**HD4600**
Cores	12	20
Threads/Core	6	7
Data lane/Thread	32	8 / 16 / 32
Max. physical occupancy	2304	4480[a]
Clock (GHz)	1.0	1.25
Power consumption (W)[b]	250	<30[c]
GFLOPS[d]	3977	432

Table 1: Hardware parameter comparison of Nvidia GTX780 d-GPU and Intel HD4600 i-GPU

[a] Using 32 data lanes per thread.
[b] As Thermal Design Power (TDP)
[c] Combined CPU and GPU equals to 84 Watts. [Des]
[d] Database processing typically requires little floating point operations, thus this metric does not directly reflect the real compute power ratio of the two architectures.

2.1 Data Transfer Mechanisms

In order to understand how the speed of interconnect affects the applications' performance, it is critical to understand the data transfer mechanisms supported by the different hardware architectures. In this section, we provide a brief overview of the key aspects of these mechanisms and quantify their effect on the performance of applications in Section 3.

Discrete GPUs support two types of data transfer mechanisms between the CPU and the GPU. On one hand, the more traditional model is using hardware supported Direct Memory Access (DMA) operations to transfer all the data to the GPU's memory (i.e. GDR5 in Figure 1) prior to processing it and DMA the data back after completion. While DMA op-

[1] Assuming SIMD32, the maximum number of data lanes we can have per thread.
[2] Next-generation integrated CPUs are expected to have 128MB of EDRAM (Embedded DRAM) on-chip as an additional layer of cache. [Tak]

erations are executed by hardware and do not need either the CPU's or GPU's intervention, it still happens over the relatively slow PCIe connection. As such, it can add to the total execution time of the application if the GPU cannot be utilized while the DMA operations take place.

On the other hand, a block of the main memory belonging to the user application on the CPU's side can be mapped into the GPU's virtual memory space via the GPU's Memory Management Unit (MMU). Programming the MMU takes only several cycles and thus is a much faster operation than DMA-ing the data between the main memory and the GPU's memory. Data mapped into the GPU's virtual memory space can be referenced directly and when it happens a Programmed I/O (PIO) operation is executed from the GPU's side, which moves the data from the main memory to the GPU's registers for processing. The advantage of this method is that only the part of the data that is processed is moved to the GPU, which might be significantly less than the total amount of application data. However, during a PIO transfer the GPU is stalled until the data transfer completes and because the transfer happens over the relatively slow PCIe connection this can slow the overall processing down except for suitable workloads where the GPU can overlap computation with kernel originated PCI-Express transactions.

In comparison, integrated GPUs, like that of the Intel i7-4770K, share the cache and main memory with the CPU and access memory through the GPU's own virtual memory space that is mapped by the GPU Address Translation Table (GTT) to the physical memory. This mechanism is similar to that of the memory mapped case of d-GPUs with the main exceptions that the data is fetched through a fast internal bus and it can also be retrieved from the shared LLC of the CPU, providing even faster access to the data. In this architecture, data buffers can be allocated by either the GPU or CPU and direct access can then be given to the other through memory address mapping.

3 Performance Analysis

In this section, we provide the results of our performance analysis study that quantifies the architectural differences between i-GPU and d-GPU based co-processing by using artificial and realistic workloads.

3.1 Raw data access

The objective of our first performance measurement was to analyze the speed of raw data access using the different data transfer mechanisms outlined in Section 2.1 (i.e. d-GPU DMA, d-GPU memory mapped, and i-GPU shared access). To this end, we created an artificial workload in OpenCL (i-GPU) and CUDA (d-GPU) where a fixed size array of integers was read through randomly. We selected random read over sequential to avoid prefetching that can hide some of the physical data access delays. Figure 3 shows the execution time of this workload based on the three different data transfer mechanisms.[3] Regardless of the size of the data structure, the i-GPU significantly outperforms both types of d-GPU based mechanisms by a factor of 2.2 to 44 thanks to the high bandwidth, low-latency internal bus and the lower latency of the DDR3 memory. The performance dif-

[3] Please note the logarithmic scaling.

ference compared to the DMA based mechanism varied between 2.2x and 2.6x, but the difference is close to constant regardless of the size of the data set [4], while the difference compared to the memory mapped mechanism is between 20x and 44x and is increasing with the data set size. As for the two d-GPU mechanisms, the DMA based execution provides a 14-fold (on average) performance difference over the memory-mapped execution, clearly demonstrating the drawbacks of memory mapping with d-GPUs in the case when multiple independent accesses are made to the main memory and the GPU cores are stalled while data is moved over the PCIe interconnect.

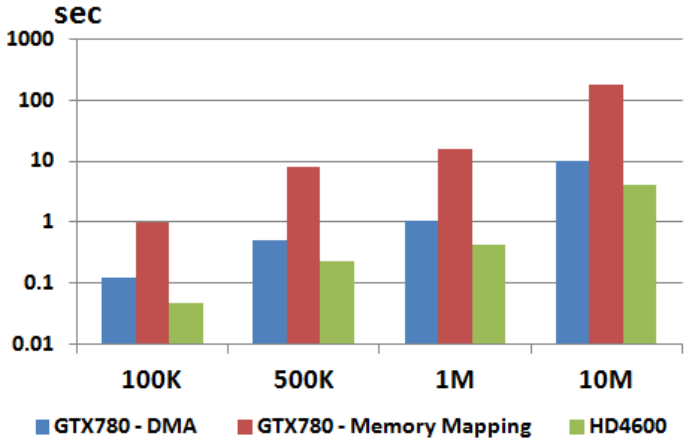


Figure 3: Performance (execution time) of different data transfer mechanisms

3.2 Micro-benchmarks

Previous studies have used database micro-benchmarks to evaluate the performance of database query co-processing [HLY+09]. We have implemented a selected set of these functionalities in OpenCL and CUDA and ran them on both the i-GPU (OpenCL only) and d-GPU platforms to understand their relative performance. Our selected micro-benchmarks represent workloads with:

- Simple optimal memory access patterns: Map, Reduce, Scan (a.k.a. prefix sum) [5];
- Randomized memory access pattern: Gather, Scatter;
- Combination of several sorting operations (i.e. randomized memory access, comparison, branching, limited integer arithmetic): Split, Bitonic and Radix Sort.

Figure 4 shows the execution time of the different micro-benchmarks with 32 Million entries on the i-GPU and d-GPU platforms. On the d-GPU platform we ran the benchmarks in OpenCL for a fairer comparison as well as in CUDA, which is highly optimized for Nvidia GPUs and thus provides the highest achievable performance. We used the DMA

[4] In the case of the DMA based mechanism the reported times include both the DMA transfer to the GPU and the random read of the array. The DMA transfer in all cases takes less than 0.1% of the reported time.

[5] The latter involves a large amount of data movement with some level of randomness though.

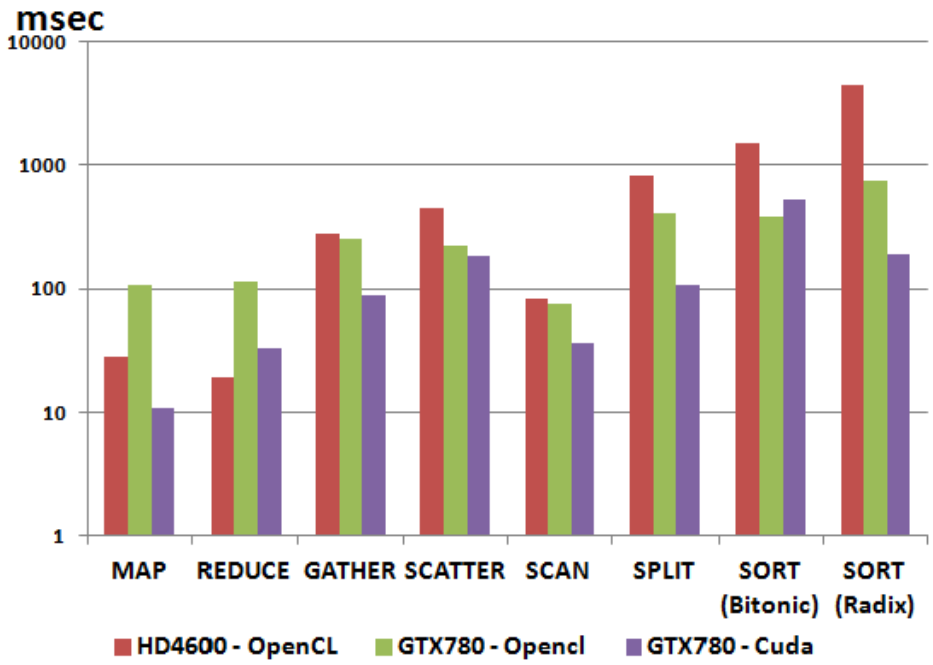


Figure 4: Micro-benchmark execution times

based data transfer mechanism with all the d-GPU measurements. Figure 5 shows the same results, but in terms of data throughput normalized to the power consumption of the GPUs. As far as raw performance is considered (i.e. Figure 4) the i-GPU platform provides better or near equal performance to the d-GPU platform with OpenCL for workloads that demonstrate randomized memory read and simple optimal memory access patterns to localized and coalesced data, which allows maximum leverage of shared memory and caches (i.e. map, reduce, gather and scan). The additional workloads require more computation and multiple iterations through the same data and thus favor the d-GPU platform more for its higher computational capability. For example, in the case of Radix Sort where the performance difference is the most significant in the favor of the d-GPU, the caching advantage of the i-GPU disappears due to the fact that many passes need to be made through a large set of data that does not fit into the cache. That is, the one-time transfer into the large memory unit of d-GPU is amortized over the multiple passes, while the i-GPU loses some of its advantage through frequent cache churns. However, if we take the power consumption of the GPUs into account, as shown in Figure 5, the i-GPU platform outperforms the d-GPU platform even with CUDA in all cases except with Radix Sort. Note again that CUDA uses libraries highly optimized for the Nvidia GPUs, while the OpenCL implementation does not have all of these highly optimized functions available, and hence the lower performance. Given that recent initiatives of new server design [Sea, HP , Dia] demonstrate the growing importance of power consumption in data center infrastructure, our results indicate that i-GPU based co-processing architectures have a high potential to become the

dominant database acceleration platform in the near future.

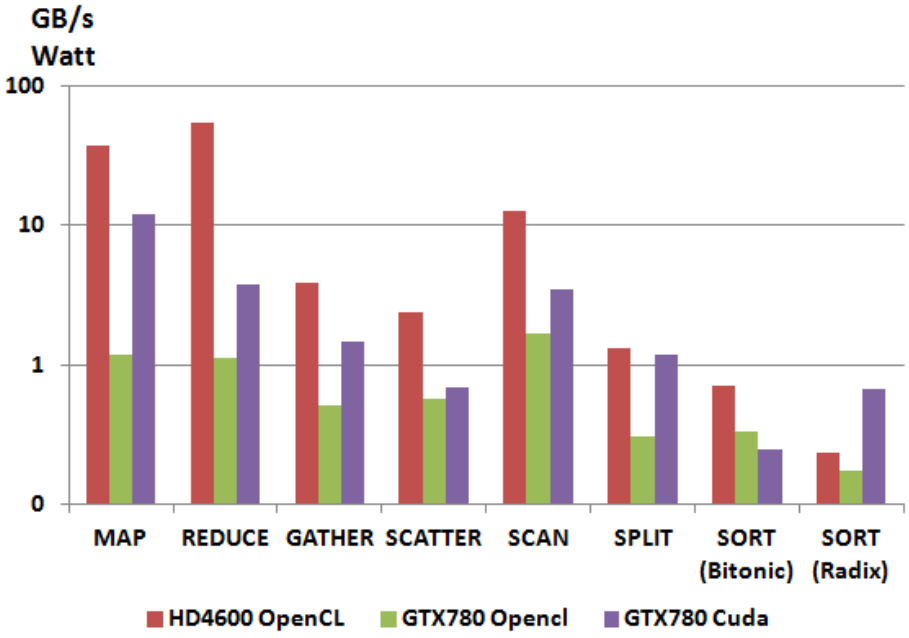


Figure 5: Micro-benchmark data processing throughput, normalized to power consumption

3.3 Database query

Our last set of performance analysis results were gathered by implementing and running Query-1 (Q1) and Query-9 (Q9) of the widely-used TPC-H decision support benchmark. We have implemented both TPC-H Q1 and Q9 as User Defined Functions (UDF). As far as TPC-H Q1 is concerned, we mofified a UDF for MonetDB suggested in [BZN05] by distributing the data set among workitems/threads on the targeted GPU to ensure full resource occupancy. As part of this query's execution plan, each workitem/thread performs sub-aggregation of the results followed by the CPU fully aggregating all the results. The data set was generated by setting the Scaling Factor (SF) to 1, which corresponds to slightly over 6 Million data records. We implemented the UDF in both OpenCL and CUDA. In terms of execution time and data throughput normalized to power consumption, Figures 6 and 7 show the similar trends as we observed earlier with micro-benchmarks. As shown by these results, the HD4600 i-GPU processes the data faster than the GTX780 d-GPU regardless of whether OpenCL or CUDA has been used as the programming environment for the d-GPU. Even more remarkably, if we consider the power consumption as the normalizing factor, the i-GPU outperforms the d-GPU platform by over a magnitude.

In addition to the results discussed above, we also ran TPC-H Query-1 in a hybrid manner in OpenCL, half of the data being processed by the CPU cores and the other half by the

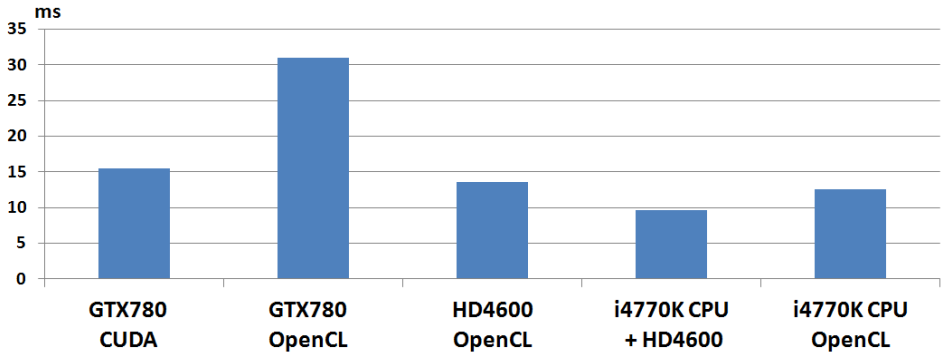


Figure 6: TPC-H Q1 (UDF) execution times

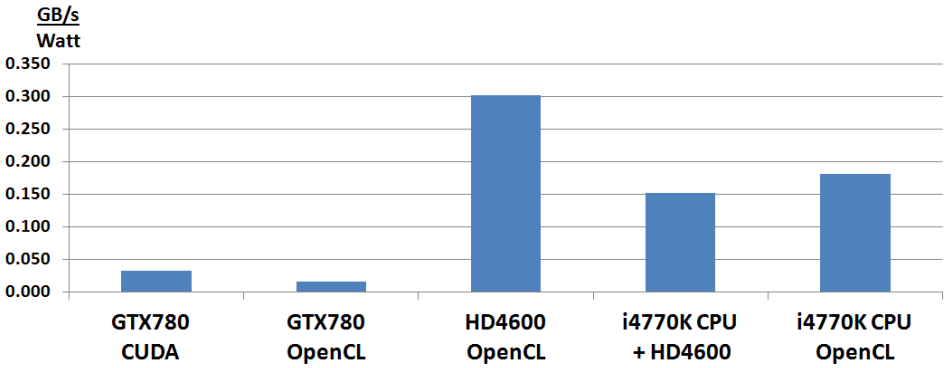
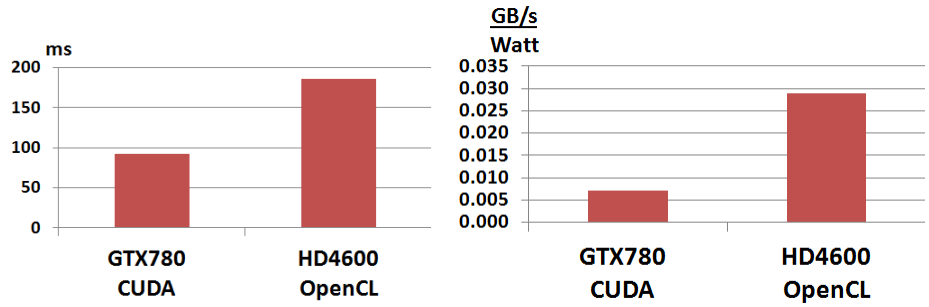


Figure 7: TPC-H Q1 (UDF) data processing throughput, normalized to power consumption

GPU cores, merging the results after completion of the query. As shown by the 4th bar in Figures 6 and 7, even more database processing power can be extracted from the i-GPU based co-processing platform using OpenCL and without increasing the complexity of the program at all. For comparison, we also included (as bar 5) the performance of the CPU-only execution of the TPC-H Q1 UDF which uses the program suggested in [BZN05] for targeting the CPU *only*.

As far as TPC-H Q9 is concerned, we implemented it as a UDF, similarly to Q1, using column based input. The key operations of this query include string matching, join, aggregation, group by and order by. As for string matching, we implemented a naïve linear string search algorithm and the Horspool algorithm [Hor80]. Our micro-benchmarks of this function indicated that the naïve algorithm outperforms the Horspool algorithm and its performance advantage increases with increasing record size. As such, we used the naïve approach in our final implementation. We performed aggregation and grouping using the

sort-merge join algorithm, and used the radix-sort algorithm for ordering. The hash join operation in our experiments took up almost half of the reported Q9 execution time. For hash join, we used an un-partitioned, bucket-chained hash join. It is possible that using a cache optimized partitioned hash join would improve the performance on our i-GPU platform although it has been reported that for off-load mode processing on i-GPU platforms, partitioned hash joins and un-partitioned hash joins show similar results [HLH13].



(a) Execution time

(b) Data processing throughput, normalized to power consumption

Figure 8: Performance comparison of TPC-H Q9 (UDF)

Figure 8 shows the results of our performance evaluation of the TPC-H Q9 UDF implementation, with Scaling Factor set to 1, using all the resources of both the Nvidia GTX780 d-GPU and the Intel HD4600 i-GPU. On the d-GPU platform we used CUDA to execute the query while on our i-GPU platform we used OpenCL. Looking at the performance from the execution time's perspective, the i-GPU based environment requires twice as much time to execute the query with the same number of records. However, in terms of data processing throughtput normalized to the power consumption, the i-GPU outperforms the d-GPU by four times, further strengthening our observation that i-GPUs provide a more power efficient acceleration environment for database processing compared to that of d-GPUs.

4 Conclusions

We have examined query and primitive operation processing on i-GPUs and d-GPUs of comparable performance attached to the system on PCIe 3.0. We used a series of varied micro-benchmarks and a more realistic data-analytic query from the TPC-H standard. We have found that while their compute resources are weaker, i-GPUs excel significantly in the speed of data access from the CPU. Furthermore, i-GPUs behave as "free" resources in some sense and consume far less power. Database management systems require CPUs to perform query parsing, plan generation, plan optimization and plan execution. They can also generate parallel execution plans targeted to multi-core CPU systems. The question of interest is whether dedicated parallel-processing subsystems can be used for off-loading such parallel query executions. Our work confirms earlier findings that such devices can be useful, but draws a further distinction between i-GPUs and d-GPUs, demonstrating that

i-GPUs have better marginal performance return on capital and operating costs. We have arrived at this conclusion through architectural analysis, micro-benchmarks using typical database query primitive operations and a typical complete query when each are optimally implemented for d-GPUs or i-GPUs. It may be worth exploring more complex queries (e.g., skyline queries) on larger data sets to examine whether this advantage of i-GPUs extends to a broader set of query processing scenarios.

References

[Bøg13] Kenneth Sejdenfaden Bøgh. Efficient Skycube Computation on the GPU. Master's thesis, Aarhus University, 2013.

[BZN05] Peter A. Boncz, Marcin Zukowski, and Niels Nes. MonetDB/X100: Hyper-Pipelining Query Execution. In *CIDR*, pages 225–237, 2005.

[Des] Desktop 4th Generation Intel Core Processor Family: Datasheet, Vol. 1. http://www.intel.com/content/www/us/en/processors/core/4th-gen-core-family-desktop-vol-1-datasheet.html.

[Dia] Diane Bryant. Architecting for Hyperscale Datacenter Efficiency. http://www.intel.com/newsroom/kits/atom/c2000/pdfs/ Architecting_ for_ Hyperscale_ DC_ Efficiency.pdf.

[GS11] Dominik Göddeke and Robert Strzodka. Cyclic Reduction Tridiagonal Solvers on GPUs Applied to Mixed Precision Multigrid. *IEEE Transactions on Parallel and Distributed Systems (TPDS), Special Issue: High Performance Computing with Accelerators*, 22(1):22–32, Jan 2011.

[HLH13] Jiong He, Mian Lu, and Bingsheng He. Revisiting Co-processing for Hash Joins on the Coupled CPU-GPU Architecture. *Proc. VLDB Endow.*, 6(10):889–900, August 2013.

[HLY+09] Bingsheng He, Mian Lu, Ke Yang, Rui Fang, Naga K. Govindaraju, Qiong Luo, and Pedro V. Sander. Relational Query Coprocessing on Graphics Processors. *ACM Trans. Database Syst.*, 34(4):21:1–21:39, December 2009.

[Hor80] Nigel R Horspool. Practical Fast Searching in Strings. In *Software - Practice & Experience*, volume 10, pages 501–506, 1980.

[HP] HP Moonshot System. http://www.hp.com/go/moonshot.

[HSP+13] Max Heimel, Michael Saecker, Holger Pirk, Stefan Manegold, and Volker Markl. Hardware-oblivious Parallelism for In-memory Column-stores. *Proc. VLDB Endow.*, 6(9):709–720, July 2013.

[KLMV12] Tim Kaldewey, Guy M. Lohman, Ren Mller, and Peter Benjamin Volk. GPU join processing revisited. In Shimin Chen and Stavros Harizopoulos, editors, *DaMoN*, pages 55–62. ACM, 2012.

[PMK11] Holger Pirk, Stefan Manegold, and Martin L. Kersten. Accelerating Foreign-Key Joins using Asymmetric Memory Channels. In Rajesh Bordawekar and Christian A. Lang, editors, *ADMS@VLDB*, pages 27–35, 2011.

[Sea] SeaMicro SM15000 Fabric Compute Systems and Freedom Fabric Storage. http://www.seamicro.com/products.

[Tak] Taking Advantage of Intel Graphics with OpenCL. http://download-software.intel.com/ sites/ default/ files/ managed/46/ba/ Webinar001_-_ Taking_ Advantage_of_ Intel_Graphics_ with_OpenCL.pdf.

[TK13] Dirk Habich Wolfgang Lehner Tomas Karnagel, Benjamin Schlegel. The HELLS-Join A Heterogeneous Stream join for ExtremeLy Large Windows. In *9th Ninth International Workshop on Data Management on New Hardware (DaMoN 2013)*, June 2013.

[WGHP11] Rick Weber, Akila Gothandaraman, Robert Hinde, and Gregory Peterson. Comparing Hardware Accelerators in Scientific Applications: A Case Study. *IEEE Trans. Parallel Distrib. Syst.*, 22(1):58–68, 2011.

[WZHC10] Ren Wu, Bin Zhang, Meichun Hsu, and Qiming Chen. GPU-Accelerated Predicate Evaluation on Column Store. In Lei Chen, Changjie Tang, Jun Yang, and Yunjun Gao, editors, *Web-Age Information Management*, volume 6184 of *Lecture Notes in Computer Science*, pages 570–581. Springer Berlin Heidelberg, 2010.

Pattern-guided Big Data Processing on Hybrid Parallel Architectures

Fahad Khalid, Frank Feinbube, Andreas Polze

Operating Systems and Middleware Group
Hasso Plattner Institute for Software Systems Engineering
Prof.-Dr.-Helmert-Str. 2-3
14482 Potsdam
fahad.khalid@hpi.uni-potsdam.de
frank.feinbube@hpi.uni-potsdam.de
andreas.polze@hpi.uni-potsdam.de

Abstract: The advent of hybrid CPU-GPU architectures has significantly increased the number of raw FLOP/s. However, it is not obvious how these can be put to use when processing Big Data. In this paper, we present an approach for designing Big Data simulations for hybrid architectures, which is based on a hierarchal application of design patterns in parallel programming. We provide a detailed account of the step by step approach that results in efficient utilization of processing and memory resources, while simultaneously improving developer productivity. Finally, we present our vision of automated tools that will further simplify the development of efficient parallel implementations for Big Data processing on hybrid architectures.

1 Introduction

Recent advances in experimental and and computational natural sciences have led to an unprecedented increase in data processing requirements. In domains such as biology, new insights into processes such as regulation of gene transcription, metabolism, signal transduction, and protein-protein interaction, have led to new models that require more complex computations on larger datasets for realistic simulations. Similarly, in geophysics, the increasing amounts of seismic data being captured due to a large number of sensors placed in seismically active regions; and a multitude of innovations in other fields have resulted in the generation of very large datasets that are used by simulations in order to produce accurate predictions. Such simulations not only require powerful processing resources, but demand novel and effective ways of managing very large amounts of data.

Algorithms used in such simulations can be broadly categorized based on whether they exhibit regular or irregular data access patterns. This paper focuses solely on algorithms with regular data access patterns that result in decomposable input domains. The current generation of hardware architectures provides performance high enough to satisfy compute requirements for such simulations. However, processing very large data sets does not just

require compute performance. It requires that the simulations be able to efficiently manage large amounts of data subject to the limited main memory and cache sizes. This situation is further complicated in hybrid architectures where the system consists of both CPUs and accelerators such as GPUs, whose massively parallel architecture is particularly suited to data parallel applications. In such cases, each processor type has an associated memory hierarchy different from the other. Even though such hybrid architectures contribute a high number of FLOP/s, managing high data volumes becomes quite challenging.

In this paper we propose an approach to effectively manage the processing of very large data sets on shared-memory hybrid architectures with CPUs and GPUs. We argue that a hierarchical and methodical application of *patterns for parallel programming* [KMMS10] can be used to maximize system performance, and provide a means to efficiently manage very large data sets. Moreover, we present an approach – complimented by our vision – that helps improve developer productivity; thereby reducing the overall cost of development for hybrid architectures.

The paper is organized as follows: Section 2 describes the pattern-based design approach for efficient resource utilization in hybrid architectures. Each subsection elaborates on a different level of hierarchy and how it can be applied. This is followed by a discussion on the applicability and limitations of the approach in Section 3. Section 4 describes the novel *Architecture-based Algorithm Decomposition Approach* proposed in this paper. This is followed by a discussion on how tools can assist in increasing developer productivity by automatically discerning parallel design patterns and exposing potential parallelism in serial code.

Note: Throughout the rest of the document, we use the term *Host* to refer to a combination of one or more CPUs available in a hybrid system. We use the term *Device* to refer to the GPU available in the same system.

2 Patterns, Productivity and Efficient Hybrid-Resource Utilization

In this section we describe a design approach based on hierarchical application of *patterns for parallel programming* to data parallel applications with regular data access patterns. We hypothesize that not only can our approach be used to improve simulation performance on a hybrid architecture, it also provides the foundation for developing frameworks and automation tools that can improve developer productivity. We call this method the *Efficient Hybrid-Resource Utilization (EHRU)* approach. This approach, as described in this paper, is based on previous work [KNTP13, KFP14].

2.1 Simulation as Pipeline

Let us consider a typical simulation scenario consisting of the following three steps:

1. Input data is read from file

2. Input data is processed using one or more computational kernels
3. Output is written to file

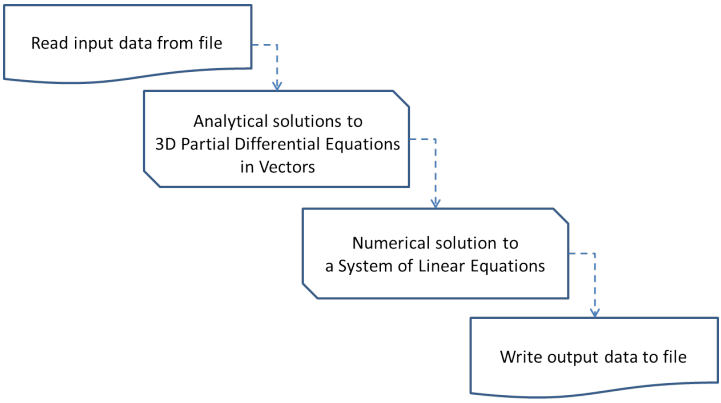


Figure 1: A hypothetical simulation depicted as pipeline.

A pictorial representation of the above mentioned steps is provided in Figure 1. In addition to the two file I/O stages, the simulation presented in Figure 1 comprises of two different computational kernels; each constituting a pipeline stage. Let us assume that the input dataset is very large, and only a fraction of the dataset can fit into main memory. In this case, the dataset can be partitioned into several small chunks, where each chunk can easily fit into main memory. Then, the first stage of the pipeline reads one chunk at a time and passes it on to the next stage. Each stage processes output from the previous stage and passes its output to the next stage, until the final result is written to file. This mechanism makes it possible to process a dataset too large to fit into the main memory.

It is worthwhile mentioning that a single chunk that completely fills the memory essentially serializes the pipeline. Therefore, determining the appropriate size of the input data chunk requires analyzing various properties of the different stages of the pipeline. We refer the reader to [KFP14] for a detailed account of optimizing pipeline performance.

The concept of parallel pipelining [DFF+02] is well established in parallel computing. Our intention here is not introduce pipelining as a contribution, but rather to lay the foundation for the use of the *pipeline* pattern [KMMS10, MRR12] for efficient big data processing on hybrid architectures.

2.2 Data Partitioning and Processing on a Hybrid Architecture

Representing the entire simulation as a pipeline is considered as the top-level pattern in our proposed hierarchy of patterns. The file I/O stages remain at this level of hierarchy.

The processing stages, however, can be further decomposed into two parts, i.e., 1) *Host-only processing*, and 2) *Hybrid Pipeline*. In order to distribute the input data chunk over the *Host-only processing* implementation and the *Hybrid Pipeline* implementation, the *geometric decomposition* [KMMS10] pattern and/or the *partition* [MRR12] pattern can be used. Once the data is partitioned, *Host-only processing* and *Hybrid pipeline* work as two parallel forks. The number of *Host* threads assigned to each fork depends on the nature of the algorithm and the resulting performance considerations. Once both forks have finished processing the assigned chunks of data, a join followed by a reduction can be used to consolidate the separately computed results. Therefore, at this level, in addition to geometric decomposition and partition, *fork/join* [MRR12] and *reduction* [MRR12] patterns are utilized. This constitutes level 2 of the pattern hierarchy.

(*L2*) EHRU { Patterns: *Geometric decomposition, Pipeline, Fork/Join, Reduction*
  ***(L3)* Hybrid Pipeline** { Patterns: *Pipeline*
    ***(L3.1)* Device** { Patterns: *Partition, <Stencil, Map>*
      ...
    }
    ***(L3.2)* Post-processing** { Patterns: *<Scan, Reduction>*
      ...
    }
  }
  ***(L3)* Host-only** { Patterns: *Geometric decomposition, <<Stencil>>, <<Scan>>*
    ...
  }
}

Figure 2: Pattern hierarchy.

2.3 Host-only Processing

The total number of available *Host* threads are divided into two groups; one for *Host-only* processing and the other for *Hybrid pipeline*. *Host-only* processing is based on an implementation that combines all the processing stages of the outer pipeline into a single *Host*-only execution sequence. This implementation can then use the *loop parallelism* [KMMS10] pattern and/or the *fork/join* pattern to process its share of data in parallel. High-level parallel programming models such as OpenMP [Boa11], Intel TBB [Rei07], Cilk [Rob13], OpenCL [Gro11] etc., can be used by the developer to easily implement these patterns. *Host-only processing* and *Hybrid pipeline* constitute level 3 of the pattern hierarchy.

2.4 Hybrid Pipeline

We use the term *Hybrid Pipeline* for an implementation of the parallel pipeline pattern where different stages of the pipeline are executed on different processor types, e.g., some on the *Host* and others on the *Device*. For an outer pipeline similar to the one depicted in Figure 1, where the processing algorithm comprises two stages, the two stages can be characterized as the *producer-consumer* dynamic. Such a produce-consumer dynamic can be implemented using a 3-stage hybrid pipeline, as shown in Figure 3.

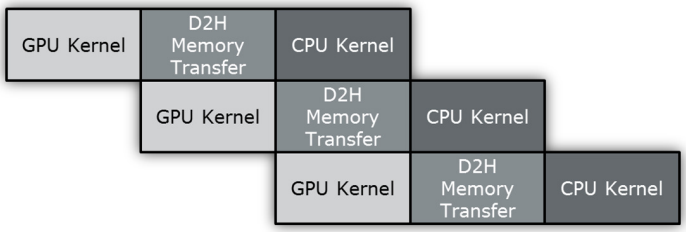


Figure 3: A 3-stage hybrid parallel pipeline.

Let us assume that the first processing stage is the most compute intensive stage of the two, and therefore constitutes the part of the simulation where most of the processing time is spent. Then, in order to improve the performance of this stage, the massively parallel nature of the *Device* can be exploited by implementing this stage as a *Device* kernel.

Results from the *Device* kernel can then be fed into the next stage which comprises of a less compute intensive post-processing algorithm. Since the computational requirements for this stage are not as high, sufficient performance can be gained by using a *Host*-only parallel implementation. Here, we assume that the time required to process this stage as a *Host*-only parallel implementation is less than the time required to process the *Device* kernel stage.

Since the two above mentioned stages execute on different processor types, results from the first stage must be copied from the *Device* memory to the *Host* memory. Such transfer of memory can be time consuming, and therefore must be overlapped with computation. Consequently, it is important to introduce a third stage between the two stages which implements the asynchronous transfer of results from *Device* memory to the *Host* memory.

All the above mentioned stages, when combined, form a 3-stage hybrid pipeline. The advantages of such a hybrid pipeline are two-fold. Assuming that the post-processing stage consumes less time than the *Device* stage, the time taken by the post-processing stage is outweighed by the *Device* stage for most of the pipeline iterations. For a large enough dataset, the post-processing stage works at almost no cost in terms of time. Moreover, since the post-processing stage is implemented on the *Host* using high level programming models, the development effort is much lower. Therefore, the hybrid pipeline provides efficient processing and high productivity. A generic framework for hybrid pipelining was

Table 1: Lists of suitable and unsuitable dwarfs.

Suitable Dwarfs	Unsuitable Dwarfs
Dense Linear Algebra	Sparse Linear Algebra
Structured Grids	Unstructured Grids
Monte Carlo	Graph Traversal

presented in [KFP14].

3 Feasibility and Limitations of the EHRU Approach

As mentioned in Section 1, our approach primarily focuses on data parallel applications with regular access patterns. In this section, we show that there are several broad categories of algorithms that fit these criteria. Moreover, we also highlight categories of algorithms for which our approach is not suitable. Both these categories are based on Berkeley Dwarfs [ABC+06]. Table 1 provides a comparative listing of dwarfs for which *EHRU* approach is suitable and unsuitable.

3.1 Applications for which the EHRU Approach is Suitable

The most important parameter that contributes to the applicability of the *EHRU* approach is decomposability of data. If the input data can be easily decomposed into several chunks that can be processed using synchronization-free parallelism, the *EHRU* approach can be used, regardless of the access patterns within the chunks.

Linear Algebra operations on dense vectors and matrices constitute one category of algorithms for which the *EHRU* approach is suitable. E.g., two very large dense matrices can be multiplied block-wise [Str03], without requiring communication between different processes that operate on different blocks. In certain simulations, Finite Element [ZM71] or Boundary Element [BB81] methods are used, where the same computations are applied to most elements in a finite grid. A structured grid can often be easily divided into many different parts that can be processed in parallel without the need for synchronization. Final results in such applications are computed by applying a post-processing reduction operation.

3.2 Applications for which the EHRU Approach is not Suitable

Many Big Data applications operate on input datasets which cannot be easily partitioned into chunks that can be processed independently. One category of such problems are those that operate on graph structures. Graph partitioning in itself is a hard problem [BMS+13].

Therefore, for simulations involving graph structures, it is not often possible to partition the input datasets in a way that chunks can be processed in a pipeline. This is because an element in one chunk might require access to elements spread over several other chunks.

Another possible negative case is where the simulation generates huge amounts of data at runtime, and data generated at each iteration is used as input to the next iteration. Therefore, the memory requirements increase at runtime. One example from the domain of computational geometry is the problem of *extreme ray enumeration in a polyhedral cone* [MRTT53]. In this problem, the memory consumption grows exponentially over each iteration. Moreover, the output generated during each iteration cannot be easily partitioned, i.e., every process requires access to the entire data. Therefore, pipelining the input data is not a feasible solution.

4 Architecture-based Algorithm Decomposition

The hybrid pipeline approach outlined in Section 2.4 raises an important question, *"how does one decide which stage of the hybrid pipeline to implement on which architecture?"*. In this section, we argue that it should be possible to structure the pipeline flow in such a way that different stages are executed by either the *Device* or the *Host* depending on which architecture is more suitable for that part of the kernel. In the following subsections we highlight the characteristics of a computational kernel that make it suitable for execution on a certain architecture. We then propose, that based on these characteristics, an algorithm can be decomposed into multiple kernels, where each kernel is suitable for execution on one specific processor type in a hybrid architecture. The overall approach is depicted in Figure 4.

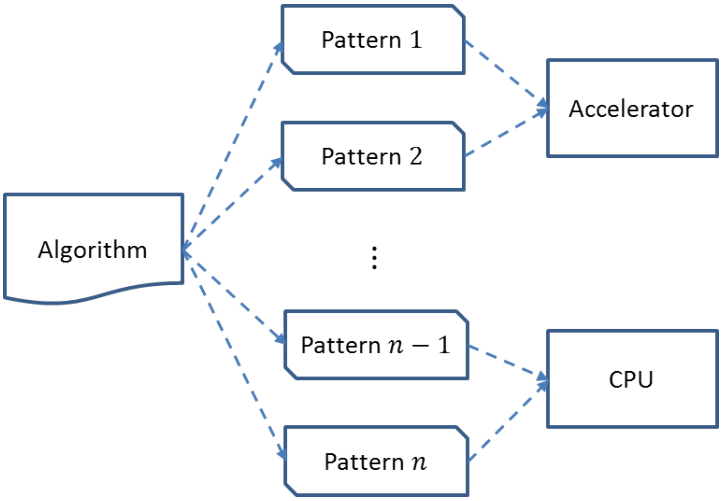


Figure 4: Architecture-based algorithm decomposition.

4.1 Characteristics of Computational Kernels

Host architectures (i.e., CPUs such as Intel Core i7, Intel Xeon series etc.) are equipped with a feature rich Instruction Set Architecture (ISA), Vector processing units, and advanced features such as branch prediction. This makes it possible for the *Host* to efficiently perform all kinds of computations. The major disadvantage as compared to *Device* architectures is the lack of massive parallelism.

Device architectures have the following major limitations:

- A high *Degree of Parallelism (DoP)* in a kernel is a fundamental requirement. If the degree of parallelism is not high enough, compute cycles are not fully utilized.
- Caches are much smaller on the *Device*. Therefore, kernels with low *arithmetic intensity* [PH13] are not particularly suitable.
- *Device* architectures do not support branch prediction, and are bounded by a minimum number of threads that must execute the same instruction in a single clock cycle. Therefore, excessive *control divergence* in the kernel can lead to wasted cycles.

It is argued in the following sub-sections that *Device* kernels for which the processing is dominated by one or more of the above mentioned limitations, can gain performance improvement if the appropriate parts of the kernel are executed on the suitable architecture.

4.2 Algorithm Decomposition and Design Patterns

An algorithm can be decomposed based on architectural features if the following conditions are satisfied:

- At least two design patterns can be identified, where one pattern is suitable for the *Device* architecture, and the other one is suitable for the *Host* architecture. Each pattern can then be implemented as a stage of computation.
- The *pipeline* pattern is applicable across the two stages of computation.

The design patterns for parallel programming can be used to identify the data and control flow in a computational kernel. This makes it possible to see whether a certain kernel would be suitable for a given architecture. The following subsections describe the patterns that are suitable for *Device* architectures, as well as those that are not suitable for *Device* architectures.

4.2.1 Patterns Suitable for *Device* Architectures

Some of the important patterns suitable for execution on the *Device* are:

- **Map**: The Map [MRR12] pattern can be seen as a fully unrolled loop, where each iteration of the loop is mapped to a different processing unit of a massively parallel processor. There are no dependencies between the loop iterations, which implies that each processing unit reads input values and writes output values that are completely independent of all other processing units. The degree of parallelism is proportional to the number of loop iterations, which means that the hardware utilization is maximal for a large number of iterations.

- **Stencil**: Similar to Map, there are no output related dependencies amongst the processing units. This pattern is also similar to a loop without dependencies. In the Stencil [MRR12] pattern, however, a processing unit needs access to input values from the neighboring processing units as well. In fact, Map can be seen as a special case of Stencil with the restriction that each processing unit only reads its own corresponding input values.

Both these patterns map well onto the massively parallel *Device* architecture. This is because each processing unit can perform computations completely independent of all other units. There is no need for elaborate locking and synchronization primitives, which could slow down the computation.

4.2.2 Patterns Suitable for *Host* Architectures

As mentioned in Section 4.1, the *Host* architecture is suitable for all computational kernels. The only limitation is that massive parallelism is not available on these architectures. The following patterns are recommended for execution on the *Host* because they do not exploit the massively parallel architecture of the *Device* efficiently.

- **Reduce**: The Reduce [MRR12] pattern is employed when a large number of values have to be reduced to a single value. E.g., a vector of 10 values is reduced to a scalar by summing all the elements of the vector. Furthermore, the Reduce pattern can be applied to any data type on which an associative binary operator is defined. The most efficient method for parallel reduction follows computation in a tree structure. For a sum reduction, it starts with all the values at the leaves, and proceeds upwards to the root by applying the operator at each level, and halving the number of operands. Eventually, only the final scalar value is left at the root, which is the desired result.

- **Scan**: An example of the Scan [MRR12] pattern is the prefix-sum operation. In this case, the output vector is the same size as the input vector. Therefore, the tree structure consists of two sweeps: 1) *Up-sweep* from the leaves to the root to compute partial reductions, and 2) *Down-sweep* from the root back to the leaves to complete the scan.

Both the Reduce and Scan patterns share the same problem when it comes to efficient resource utilization. The degree of parallelism varies at each step of computation. In Reduce, e.g., the number of processing units that can be used at the first step is equal to the number of elements in the vector. For a large vector, this amounts to maximum

resource utilization. However, at each successive step, the number of processing units required is halved; eventually leading to the point where most of the compute resources are idle.

5 Tool-guided Parallelization for Hybrid Architectures

As described in Section 3, the overall approach presented in this paper is suited primarily to problems with easily decomposable data and regular access patterns. It follows from these assumptions, that for such applications, most loops are composed of affine iteration spaces [LL97], as well as affine mapping functions [LL97] from iteration index to array element. This makes it possible to apply the existing theory of dependence analysis and loop transformation to these problems.

5.1 Discerning Parallel Patterns from Serial Source Code

The *architecture-based algorithm decomposition* approach (described in Section 4) relies on the assumption that the developer can easily discern parallel design patterns from serial source code. This, however, requires a significant amount of knowledge and experience on part of the developer. Therefore, we argue that an automated tool that can discern patterns from serial source code would further improve developer productivity.

It is our conjecture that the fundamental mathematical property that distinguishes different patterns is the *type of dependence among the iterations* of the corresponding *for* loop. The type of dependence can be computed using the already established theory of dependence analysis [KA01]. The result can then be mapped onto the corresponding pattern. A tool capable of discerning parallel patterns from serial source code will not only improve developer productivity, it will also connect theoretical/compiler level concepts to higher level abstractions of patterns that are traditionally studied at the level of software engineering.

5.2 Suggestions for Parallelization for Hybrid Architectures

Given our earlier assumption regarding affine spaces and functions, we propose that instead of the developer having to discover all different levels of parallelism, a tool imbued with the knowledge of patterns as well as discerning parallelism, can assist the developer by highlighting possibly parallelizable code segments, as well as suggesting appropriate optimizing transformations. Intel Parallel Advisor [BCS12] is a commercially available tool with similar features. However, neither does Intel Parallel Advisor discern patterns from source code, nor is it applicable to hybrid architectures with GPUs. We are currently working towards the development of such a tool for hybrid architectures.

6 Related Work and Our Contribution

As determined in a thorough literature review [KFP14] on the subject, apart from overlapping of computation and transfer of data from GPU memory to CPU memory, pipelining has also been used in hybrid architectures for improving resource utilization and load balancing. Most pipeline applications for hybrid architectures are focused on solving a particular problem. E.g., much work has been done in utilizing pipelining for hybrid *MapReduce* implementations [CQD+13, SO11]. In [CHA12], an efficient *MapReduce* scheduling approach is presented for a coupled CPU-GPU chip. This approach makes it possible to dynamically divide Map tasks onto CPU and GPU, thereby utilizing both types of processing resources. The *Moim* [XKB] framework for *MapReduce* supports simultaneous execution of the reduce phase on both CPU and GPU. Other than *MapReduce*, a 3-stage hybrid CPU-GPU pipeline has also been used for solving eigenvector and eigenvalue problems [GGV12].

As mentioned in Section 5.2, Intel Parallel Advisor [BCS12] is a commercially available tool that guides a developer through the process of parallelizing existing serial source code. There have also been other efforts with similar goals. Prospector [KKL10] is a tool designed to assist programmers in parallelizing legacy serial applications developed by other programmers. For GPUs, GROPHECY [MMK+11] is a performance projection framework that uses skeletonized pieces of CPU source code to determine if the performance gain from porting the application to GPU is worth the development effort.

To the best of our knowledge, the *Architecture-based Algorithm Decomposition* is a novel idea introduced in this paper. Moreover, despite various commendable efforts in developing automated tools for assisting in code parallelization for either CPUs or GPUs, one has yet to be developed for hybrid architectures. We believe that our ideas presented in Section 5 will help us develop an effective tool for parallelization assistance on hybrid architectures. Our hierarchical application of parallel patterns for efficient processing of very large datasets is also original, even though the general theme of optimizing resource utilization for hybrid architectures has already been explored from other perspectives.

7 Summary and Future Work

We have presented our *Efficient Hybrid-Resource Utilization* approach for processing very large datasets on hybrid CPU-GPU based architectures. It comprises data decomposition and distribution amongst different processor types within a shared-memory system. System efficiency and developer productivity are both improved by employing hierarchical application of patterns for parallel programming. While this approach has partly been implemented and published elsewhere, in this paper, we have extended this idea as a foundation for *Architecture-based Algorithm Decomposition*. We have argued that the suitability of execution of a particular computational kernel on a particular processor type can be determined by inspecting the pattern(s) of which the kernel consists. Certain characteristics of a pattern make it suitable for execution on one or the other processor type in a hybrid

architecture. In certain cases, in addition to improving system performance, this information can save the development effort required for implementing *Device* kernels, thereby improving developer productivity.

We have further argued, that it is plausible to assume, that for data parallel applications with regular access patterns, it is possible to use dependence analysis for discerning parallel patterns from serial source code. Automation of this process can further assist in *Architecture-based Algorithm Decomposition*. Moreover, it is plausible to assume that for such applications, an automated tool can be developed that highlights potential parallelism in a serial application, and guides the developer through the process of parallelization for hybrid architectures.

We are currently in the process of developing theoretical foundations for *discerning parallel patterns from serial source code*. Finishing this project is our short-term objective. The tool resulting from this project will then be applied to *Architecture-based Algorithm Decomposition*. Eventually, we hope to have an automated tool for assisting the developer in developing parallel applications for hybrid parallel architectures.

References

[ABC+06] Krste Asanovic, Ras Bodik, Bryan Christopher Catanzaro, Joseph James Gebis, Parry Husbands, Kurt Keutzer, David A. Patterson, William Lester Plishker, John Shalf, Samuel Webb Williams, and Katherine A. Yelick. The Landscape of Parallel Computing Research: A View from Berkeley. Technical Report UCB/EECS-2006-183, EECS Department, University of California, Berkeley, Dec 2006.

[BB81] Prasanta Kumar Banerjee and Roy Butterfield. *Boundary element methods in engineering science*, volume 17. McGraw-Hill London, 1981.

[BCS12] Stephen Blair-Chappell and Andrew Stokes. *Parallel Programming with Intel Parallel Studio XE*. John Wiley & Sons, 2012.

[BMS+13] Aydin Buluç, Henning Meyerhenke, Ilya Safro, Peter Sanders, and Christian Schulz. Recent Advances in Graph Partitioning. *CoRR*, abs/1311.3144, 2013.

[Boa11] OpenMP Architecture Review Board. OpenMP Application Program Interface. Standard specification, July 2011.

[CHA12] Linchuan Chen, Xin Huo, and Gagan Agrawal. Accelerating mapreduce on a coupled cpu-gpu architecture. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 25. IEEE Computer Society Press, 2012.

[CQD+13] Yi Chen, Zhi Qiao, Spencer Davis, Hai Jiang, and Kuan-Ching Li. Pipelined Multi-GPU MapReduce for Big-Data Processing. In *Computer and Information Science*, pages 231–246. Springer, 2013.

[DFF+02] Jack Dongarra, Ian Foster, Geoffrey C. Fox, William Gropp, Ken Kennedy, Linda Torczon, and Andy White, editors. *The Sourcebook of Parallel Computing*. Morgan Kaufmann, 2002.

[GGV12] Michael T Garba and HORACIO GONZÁLEZ-VÉLEZ. Asymptotic peak utilisation in heterogeneous parallel CPU/GPU pipelines: a decentralised queue monitoring strategy. *Parallel Processing Letters*, 22(02), 2012.

[Gro11] Khronos OpenCL Working Group. The OpenCL Specification. Standard specification, December 2011.

[KA01] Ken Kennedy and John R Allen. Optimizing compilers for modern architectures: a dependence-based approach. 2001.

[KFP14] Fahad Khalid, Frank Feinbube, and Andreas Polze. Hybrid CPU-GPU Pipeline Framework. In *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, 2014. To appear.

[KKL10] Minjang Kim, Hyesoon Kim, and Chi-Keung Luk. Prospector: A dynamic data-dependence profiler to help parallel programming. In *HotParâ10: Proceedings of the USENIX workshop on Hot Topics in parallelism*, 2010.

[KMMS10] Kurt Keutzer, Berna L. Massingill, Timothy G. Mattson, and Beverly A. Sanders. A design pattern language for engineering (parallel) software: merging the PLPP and OPL projects, 2010.

[KNTP13] Fahad Khalid, Zoran Nikoloski, Peter Tröger, and Andreas Polze. Heterogeneous Combinatorial Candidate Generation. In Felix Wolf, Bernd Mohr, and Dieter Mey, editors, *Euro-Par 2013 Parallel Processing*, volume 8097 of *Lecture Notes in Computer Science*, pages 751–762. Springer Berlin Heidelberg, 2013.

[LL97] Amy W. Lim and Monica S. Lam. Maximizing Parallelism and Minimizing Synchronization with Affine Transforms. In *Proceedings of the 24th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, POPL '97, pages 201–214, New York, NY, USA, 1997. ACM.

[MMK+11] Jiayuan Meng, Vitali A Morozov, Kalyan Kumaran, Venkatram Vishwanath, and Thomas D Uram. GROPHECY: GPU performance projection from CPU code skeletons. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, page 14. ACM, 2011.

[MRR12] Michael McCool, James Reinders, and Arch Robison. *Structured parallel programming: patterns for efficient computation*. Elsevier, 2012.

[MRTT53] T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall. The double description method. *Contributions to the Theory of Games*, 2:51–73, 1953.

[PH13] David A Patterson and John L Hennessy. *Computer organization and design: the hardware/software interface*. Newnes, 2013.

[Rei07] James Reinders. *Intel threading building blocks: outfitting C++ for multi-core processor parallelism*. O'Reilly Media, Inc., 2007.

[Rob13] Arch D Robison. Composable Parallel Patterns with Intel Cilk Plus. *Computing in Science & Engineering*, 15(2):0066–71, 2013.

[SO11] Jeff A Stuart and John D Owens. Multi-GPU MapReduce on GPU clusters. In *Parallel & Distributed Processing Symposium (IPDPS), 2011 IEEE International*, pages 1068–1079. IEEE, 2011.

[Str03] Gilbert Strang. *Introduction to linear algebra*. SIAM, 2003.

[XKB] Mengjun Xie, Kyoung-Don Kang, and Can Basaran. Moim: A Multi-GPU MapReduce Framework.

[ZM71] Olgierd Cecil Zienkiewicz and PB Morice. *The finite element method in engineering science*, volume 1977. McGraw-hill London, 1971.

BiDAl: Big Data Analyzer for Cluster Traces

Alkida Balliu, Dennis Olivetti, Ozalp Babaoglu, Moreno Marzolla, Alina Sîrbu

Department of Computer Science and Engineering, University of Bologna
Mura Anteo Zamboni 7, 40126 Bologna, Italy
{alkida.balliu, dennis.olivetti}@studio.unibo.it
{ozalp.babaoglu, moreno.marzolla, alina.sirbu}@unibo.it

Abstract: Modern data centers that provide Internet-scale services are stadium-size structures housing tens of thousands of heterogeneous devices (server clusters, networking equipment, power and cooling infrastructures) that must operate continuously and reliably. As part of their operation, these devices produce large amounts of data in the form of event and error logs that are essential not only for identifying problems but also for improving data center efficiency and management. These activities employ data analytics and often exploit hidden statistical patterns and correlations among different factors present in the data. Uncovering these patterns and correlations is challenging due to the sheer volume of data to be analyzed. This paper presents BiDAl, a prototype "log-data analysis framework" that incorporates various Big Data technologies to simplify the analysis of data traces from large clusters. BiDAl is written in Java with a modular and extensible architecture so that different storage backends (currently, HDFS and SQLite are supported), as well as different analysis languages (current implementation supports SQL, R and Hadoop MapReduce) can be easily selected as appropriate. We present the design of BiDAl and describe our experience using it to analyze several public traces of Google data clusters for building a simulation model capable of reproducing observed behavior.

1 Introduction

Modern Internet-based services such as cloud computing, social networks, online storage, media-sharing, etc., produce enormous amounts of data, not only in terms of user-generated content, but also in the form of usage activity and error logs produced by the devices implementing them. Data centers providing these services contain tens of thousands of computers and other components (e.g., networking equipment, power distribution, air conditioning) that may interact in subtle and unintended ways, making management of the global infrastructure far from straightforward. At the same time, services provided by these huge infrastructures have become vital not only to industry but to society in general, making failures extremely costly both for data center operators and their customers. In this light, monitoring and administering data centers become critical tasks. Some aspects of management, like job scheduling, can be highly automated while others, such as recovery from failures, remain highly dependent on human intervention. The "holy grail" of system management is to render data centers autonomous, self-managing and self-healing; ideally, the system should be capable of analyzing its state and use this information to

identify performance or reliability problems and correct them or alert system managers directing them to the root causes of the problem. Even better, the system should be capable of anticipating situations that may lead to performance problems or failures, allowing for proactive countermeasures to steer the system back towards desired operational states. Needless to say, these are very challenging goals [SLM10].

Given the size of modern data centers, the amount of log data they produce is growing steadily, making log management itself technically challenging. For instance, a 2010 Facebook study reports 60 Terabytes of log data being produced by its data centers each day [TSA+10]. For live monitoring of its systems and analyzing their log data, Facebook has developed a dedicated software called Scuba [AAB13] that uses a large in-memory database running on hundreds of servers with 144 GB of RAM each. This infrastructure needs to be upgraded every few weeks to keep up with the increasing computational power and storage requirements that Scuba generates. Log analysis falls within the class of Big Data applications: the data sets are so large that conventional storage and analysis techniques are not appropriate to process them. There is a real need to develop novel tools and techniques for analyzing logs, possibly incorporating data analytics to uncover hidden patterns and correlations that can help system administrators avoid critical states, or to identify the root cause of failures or performance problems.

Numerous studies have analyzed trace data from a variety of sources for different purposes, but typically without relying on an integrated software framework developed specifically for log analysis [CAK12, LC12, RTG+12a]. This is partially due to the sensitive nature of commercial log trace data prohibiting their publication, which in turn leads to fragmentation of analysis frameworks and difficulty in porting them to traces from other sources. One isolated example of an analysis framework is the Failure Trace Archive Toolkit [JKIE13], limited however to failure traces. Lack of a more general framework for log data analysis results in time being wasted by researchers in developing software for parsing, interpreting and analysing the data, repeatedly for each new trace [JKIE13].

In this paper we describe the Big Data Analyzer (BiDAl), a prototype software tool implementing a general framework, designed for statistical analysis of very large trace data sets. BiDAl integrates several built-in storage types and processing frameworks and can be easily extended to support others. The BiDAl prototype is publicly available through a GNU General Public License (GPL) [BOB+14]. We illustrate the actual use of BiDAl for analyzing publicly-available Google cluster trace data [Wil11] in order to extract parameters for a cluster simulator which we have implemented.

The contributions of this work are several fold. We first present the novel architecture of BiDAl resulting in extensibility and ease of use. BiDAl incorporates several advanced Big Data technologies to facilitate efficient processing of large datasets for data analytics. We then describe our experience with BiDAl when used to extract workload parameters from Google compute cluster traces. Finally, we describe a simulation model of the Google cluster that, when instantiated with the parameters obtained through BiDAl, is able to reproduce a set of behaviors very similar to those observed in the traces.

The rest of the paper is organized as follows. We provide a high level overview of the framework followed by a detailed description of its components in Section 2. The frame-

work is applied to characterize machines and workloads in a public Google cluster trace, and used in the development of a cluster simulator in Section 3. We discuss related work in Section 4 and conclude with new directions for future work in Section 5.

2 The Big Data Analyzer (BiDAl) prototype

2.1 General overview

BiDAl can import raw data in CSV format (Comma Separated Values, the typical format of trace data), and store it in different backends according to the user's preference. In the current prototype two backends are supported: SQLite and Hadoop File System (HDFS), the latter being particularly well suited for handling large amount of data using the Hadoop framework. Other backends can easily be supported, since BiDAl is based on a modular architecture that will be described in the next section. BiDAl uses a subset of the SQL language to handle the data (e.g., to create new tables or to apply simple filters to existing data). SQL queries are automatically translated into the query language supported by the underlying storage system (RSQLite or RHadoop).

BiDAl also has the ability to perform statistical data analysis using both R [R D08] and Hadoop MapReduce [SKRC10, DG10] commands. R commands are typically applied to the SQLite storage, while MapReduce to the Hadoop storage. However, the system allows mixed execution of both types of commands regardless of the storage used, being able to switch between backends (by exporting data) transparently to the user. For instance, after a MapReduce operation, it is possible to analyze the outcome using R; in this case, the software automatically exports the result obtained from the MapReduce step, and imports it to the SQLite storage where the analysis can continue using R commands. This is particularly useful for handling large datasets, since the volume of data can be reduced by applying a first processing step with Hadoop/MapReduce, and then using R to complete the analysis on the resulting (smaller) dataset.

2.2 Design

BiDAl is a modular application designed for extensibility and ease of use. It is written in Java, to facilitate portability across different Operating Systems, and uses a Graphical User Interface (GUI) based on the standard Model View Controller (MVC) architectural pattern. The View provides a Swing GUI, the Model manages different types of storage backends, and the Controller handles the interaction between the two. Figure 1 outlines the architecture using the UML class diagram.

The Controller class connects the GUI with the other components of the software. The Controller implements the Singleton pattern, with the one instance accessible from any part of the code. The interface to the different storage backends is given by the GenericStorage class, that has to be further specialized by any concrete backend developed. In

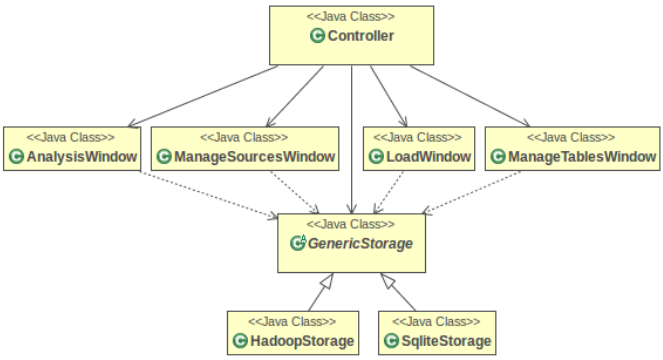


Figure 1: UML diagram of BiDAl classes.

our case, the two existing concrete storage backends are represented by the SqliteStorage class to support SQLite, and the HadoopStorage class, to support HDFS. Neither the Controller nor the GUI elements communicate directly with the concrete storage backends, but only with the abstract class GenericStorage. This simplifies the implementation of new backends without the need to change the Controller or GUI implementations.

The user can inspect and modify the data storage using a subset of SQL; the SqliteStorage and HadoopStorage classes use the open source SQL parser Akiban to convert the queries inserted by users into SQL trees that are further mapped to the native language (RSQLite or RHadoop) using the Visitor pattern. The HadoopStorage uses also a Bashexecuter that allows to load files on the HDFS using bash shell commands. A new storage class can be implemented by providing a suitable specialization of the GenericStorage class, including the mapping of the SQL tree to specific commands understood by the backend. Although the SQL parser supports the full SQL language, the developer must define a mapping of the SQL tree into the language supported by the underlying storage; this often limits the number of SQL statements that can be supported due to the difficulty of realizing such a mapping.

2.3 Functionality

The typical BiDAl workflow consists of three steps: instantiation of a storage backend (or opening an existing one), data selection and aggregation and data analysis. For storage creation, BiDAl is designed to import CSV files into an SQLite database or to a HDFS file system, depending on the type selected. Except for the CSV format, no other restrictions on the data type exist, so the platform can be easily used for data from various sources, as long as they can be viewed as CSV tables. Even though the storages currently implemented are based on the the concept of tables (stored in a relational database by SQLite and CSV files by Hadoop), in the future, other storage types can be supported by BiDAl. Indeed, Hadoop supports HBase, a non-relational database that works with <key, value> pairs.

Since Hadoop is already supported by BiDAl, a new storage that works on this type of non-relational databases can be easily added.

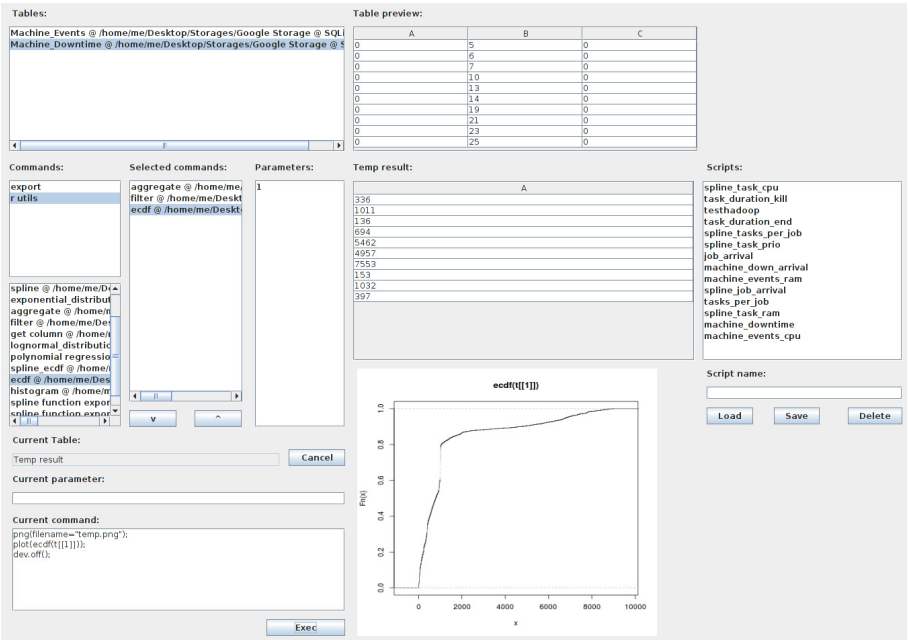


Figure 2: Screenshot of the BiDAl analysis console displaying R commands.

Selection and aggregation can be performed using queries expressed using a subset of SQL. At the moment, the supported statements are SELECT, FROM, WHERE and GROUP BY. For the SQLite storage, queries are executed through the RSQLite library of the R package (R is used quite extensively inside BiDAl, and executing SQLite queries through R simplified the internal structure of BiDAl as we could reuse some internal software components). For the Hadoop backend, GROUP BY queries are mapped to MapReduce operations. The Map function implements the GROUP BY part of the query, while the Reduce function deals with the WHERE and SELECT clauses. We used RHadoop as a wrapper so that we can access the Hadoop framework through R commands. This allows the implementation of Map and Reduce functions in R rather than Java code.

Data analysis can be performed by selecting different commands in the specific language of the storage and applying them to the selected dataset. There is a common set of operations provided by every storage. However it is possible to concatenate operations on different storage backends since BiDAl can automatically export data from one backend and import it on another. Therefore it is possible to use a MapReduce function on an SQLite table, or execute a R command on a HDFS store. This requires that the same data is duplicated into different storage types so, depending on the size of the dataset, additional storage space will be consumed. However, this operation does not generate consistency issues, since log data does not change once it is recorded.

R command	Description
get_column	Selects a column.
apply_1Col	Applies the desired function to each element of a column.
aggregate	Takes as input a column to group by; among all rows selects the ones that satisfies the specified condition; the result obtained is specified from the function given to the third parameter.
difference_between_ rows	Calculates the differences between consecutive rows.
filter	Filters the data after the specified condition.
exponential_distribution	Plots the fit of the exponential distribution to the data.
lognormal_distribution	Plots the fit of the lognormal distribution to the data.
polynomial_regression	Plots the fit of the n-grade polynomial regression to the data in the specified column.
ecdf	Plots the cumulative distribution function of the data in the specified column.
spline	Divides the data in the specified column in n intervals and for each range plots spline functions. Also allows to show a part of the plot or all of it.

Table 1: List of some R commands implemented by BiDAl.

Using R within BiDAl BiDAl provides a list of pre-defined operations, implemented in R, that can be selected by the user from a graphical interface (see Figure 2 for a screenshot and Table 1 for a full list of available commands). When an operation is selected, an input box appears asking the user to provide the parameters needed by that specific operation. Additionally, a text box (bottom left of Figure 2) allows the user to modify on the fly the R commands to be executed.

All operations are defined in an external text file, according to the following BNF grammar:

```
<file> ::= <command name> <newline> <number of parameters> <newline>
           <list of parameters> <newline> <command code>

<list of parameters> ::=  <parameter description> <newline>
     <list of parameters> | <empty>

<command code> ::= <text> | <command code> <parameter>
  <command code> | <empty>

<parameter> ::= '$PAR' <number of the parameter> '$'
```

New operations can therefore be added quite easily by simply adding them to the file.

Using Hadoop/MapReduce with BiDAl BiDAl provides also a list of Hadoop/MapReduce commands that allow to distribute computation across several machines. Usually, the Mapper and Reducer functions are implemented in Java, generating files that need to be compiled and then executed. However, BiDAl abstracts from this approach by using the RHadoop library which handles MapReduce job submission and permits to interact with

R command	Parameter type	Parameter value
get_column	column number	2
filter	condition	t[[1]]<11000.
log_histogram	column number, log step, log axis	1, 0.06, xy

Table 2: Commands used to generate Figure 3b.

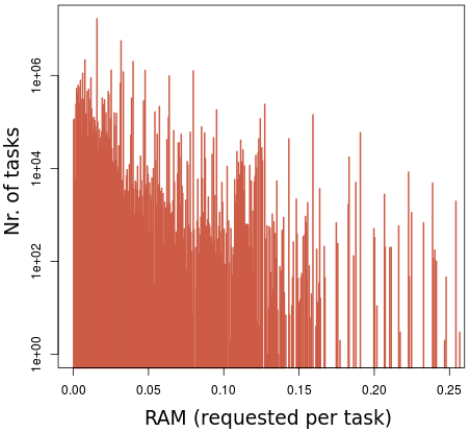
Hadoop's file system HDFS using R functions. Once the dataset of interest has been chosen, the user can execute the Map and Reduce functions already implemented in RHadoop or create new ones. Again, the MapReduce functions are saved in an external text files, using the same format described above, so the creation of new commands does not require any modification of BiDAl. At the moment, one Map function is implemented in BiDAl, which groups the data by the values of a column. The Reduce function counts the elements of each group. Other functions can be added by the user, similar to R commands.

3 Case study

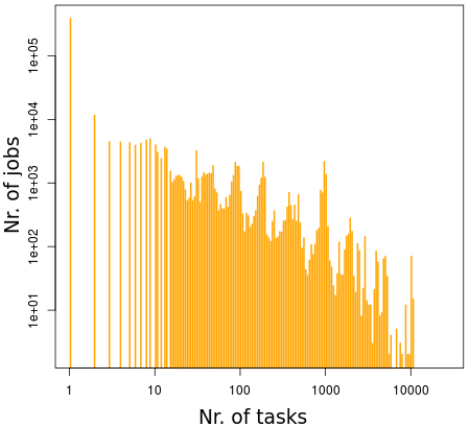
The development of BiDAl was initially motivated by the need to process large data traces of compute clusters, such as those publicly released by Google [Wil11]. The ultimate goal was to extract workload parameters from the traces in order to instantiate a simulation model of the compute cluster capable of reproducing the most important features observed in the real data. The simulation model, then, can be used to perform "what-if analyses" by exploring different scenarios where the workload parameters are different, or several types of faults are injected into the system. In this section we first describe the use of BiDAl for analyzing the Google traces, and then present the structure of the simulation model instantiated with the parameters obtained from the analysis phase.

3.1 Workload Characterization of the Google Cluster

To build a good simulation model of the Google cluster, we needed to extract some information from the traces. The data consist of a large amount of CSV files containing records about job and task events, resources used by tasks, task constraints, etc. There are more than 2000 files describing the workload and machine attributes for over 12000 cluster nodes, reaching a total *compressed size* of about 40 GB. In total, over 1.3 billion records are available. We used BiDAl to extract the arrival time distribution of each job, the distribution of the number of tasks per job, and the distributions of execution times of different types of tasks (e.g., jobs that successfully completed execution, jobs that are killed by the users, and so on). These distributions are used by the Job Arrival entity of the simulation model to generate jobs into the system. Additionally, we analyzed the distribution of machines downtime and of the time instants when servers are added / removed from the pool.



(a) RAM requested by tasks. Values are normalized by the maximum RAM available on a single node in the Google cluster.



(b) Number of tasks per job

Figure 3: Examples of distributions obtained with BiDAl.

Some of the results obtained with BiDAl are shown in the following (we are showing the actual plots that are produced by our software). Figure 3a shows the the amount of RAM requested by tasks, while Figure 3b shows the distribution of number of tasks per job.

To generate the graph in Figure 3b, we first extracted the relevant information from the trace files. Job and task IDs were required, therefore we generated a new table, called *job_task_id*, from the *task_events.csv* files released by Google [Wil11]. The query generation is automated by BiDAl which allows for simple selection of columns using the GUI. Since the DISTINCT clause is not yet implemented in BiDAl, we added it manually in the generated query. The final query used was:

```
SELECT DISTINCT V3 AS V1,V4 AS V2 FROM task_events
```

where V3 is the *job_id* column while V4 represents the *task_id*.On the resulting *job_task_id table*, we execute another query to estimate how many tasks each job has, generating a new table called *tasks_per_job*:

```
SELECT V1 AS V1, COUNT(V2) AS V2 FROM job_task_id GROUP BY V1
```

Three R commands were used on the *tasks_per_job* table to generate the graph. The first extracts the second column (job id), the second filters out some uninteresting data and the third plots the result. The BiDAl commands used are shown in Table 2.

The analysis was performed on a computer with 16 GB of RAM, a 2.7 GHz i7 quad core processor and a hard drive with simultaneous read/write speed of 60 MB/s. For the example above, importing the data was the most time consuming step, requiring 11 minutes to load 17 GB of data into the SQLite storage (which may be influenced by the disk speed). However, this step is required only once. The first SQL query took about 4 minutes to complete, while the second query and the R commands were almost instantaneous.

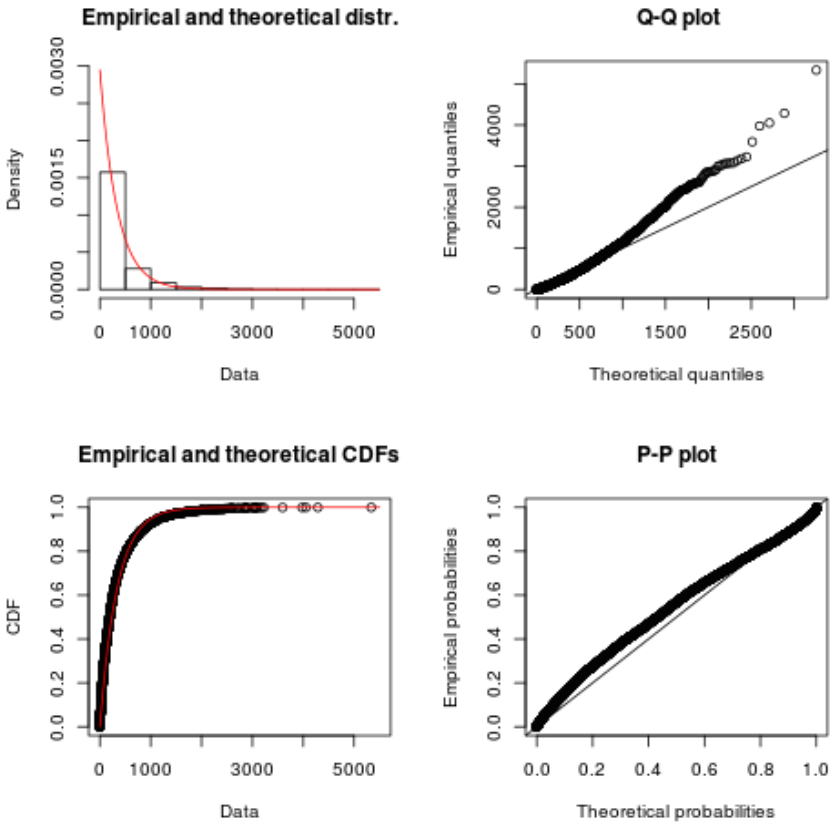
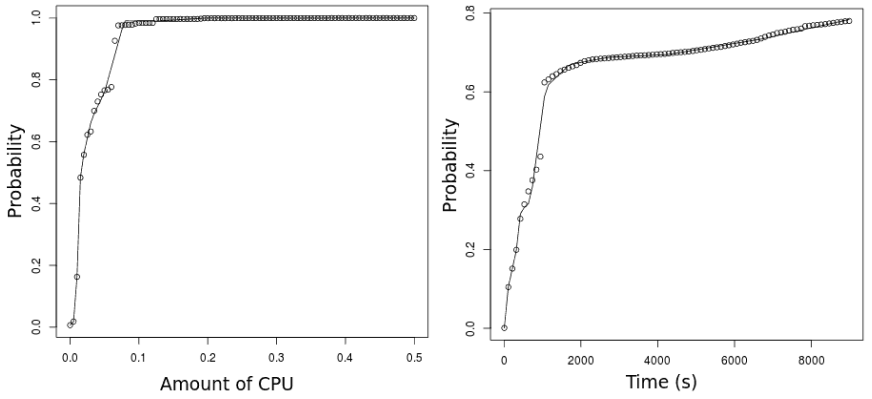


Figure 4: Machine update events, fitted with an exponential distribution. The left panels show the density and cumulative distribution functions, with the lines representing the exponential fitting and the bars/circles showing real data. The right panels show goodness of fit in Q-Q and P-P plots (straight lines show perfect fit).

In Figure 4 we tried to fit the time between consecutive machine update events (i.e., events that indicate that a machine has changed its list of resources) with an exponential distribution; the four standard plots for the goodness of fit show that the observed data is in good agreement with the fitted distribution.

Cumulative distribution functions (CDFs) have also been obtained from the data and fitted with sequences of splines, in those cases where the density functions were too noisy to be fitted with a known distribution. For instance, Figure 5a shows the distribution of CPU required by tasks while Figure 5b shows machine downtime, both generated with BiDAl. Several other distributions were generated, similar to CPU requirements, to enable simulation of the Google cluster: RAM required by tasks; Tasks priority; Duration of tasks that end normally; Duration of killed tasks; Tasks per job; Job inter-arrival time; Machine failure inter-arrival time; Machine CPU and RAM.



(a) CPU task requirements

(b) Machine downtime

Figure 5: Examples of CDFs fitted by sequences of splines, obtained with BiDAl. The circles represent the data, while the lines show the fitted splines.

3.2 Cluster Simulator

We built a discrete-event simulation model of the Google compute cluster corresponding to that from which the traces were obtained, using C++ and Omnet++. According to the information available, the Google cluster is basically a large batch system where computational tasks of different types are submitted and executed on a large server pool. Each job may describe constraints for its execution (e.g., a minimum amount of available RAM on the execution host); a scheduler is responsible for extracting jobs from the waiting queue, and dispatching them to a suitable execution host. As can be expected on a large infrastructure, jobs may fail and be resubmitted; moreover, execution hosts may fail and be temporarily removed from the pool, or new hosts can be added. The Google trace contains a list of timestamped events such as job arrival, job completion, activation of a new host and so on; additional (anonymized) information on job requirements is also provided.

The simulation model, shown in Figure 6, consists of several active and passive interacting entities. The passive entities (i.e., those that do exchange any message with other entities) are Jobs and Tasks. A task represents a process in execution, or ready to be executed; each task has an unique ID and the amount of resources required; a Job is a set of (likely dependent) tasks. Jobs can terminate either because all their tasks complete execution, or because they are aborted by the submitting user.

The active entities in the simulation are those that send and receive messages: Machine, Machine Arrival, Job Arrival, Scheduler and Network. The Machine entity represents an execution node in the compute cluster. Machine Arrival and Job Arrival generate events related to new execution nodes being added to the cluster, and new jobs being submitted, respectively. These arrival processes (as they are called in queueing theory) can be driven by the real trace logs, or by synthetic data generated from user-defined probability

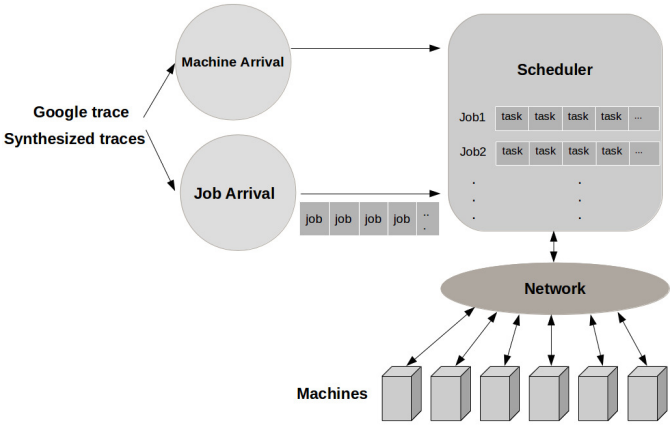


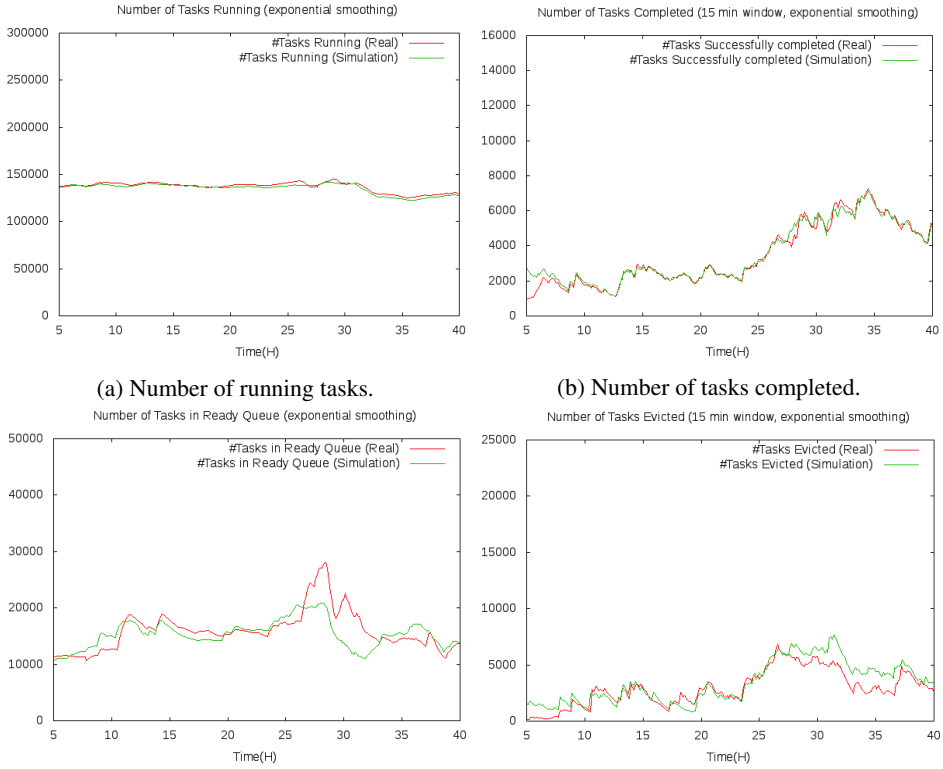
Figure 6: Architecture of simulation model.

distributions that can be identified using BiDAl. The Scheduler implements a simple job scheduling mechanism. Every time a job is created by the JobArrival entity, the scheduler inserts its tasks in the waiting queue. For each task, the scheduler examines which execution nodes (if any) match the task constraints; the task is eventually sent to a suitable execution node. Note that the scheduling policies implemented by the Google cluster allow a task with higher priority to evict an already running task with lower priority; this eviction priority mechanism is also implemented in our simulator. Finally, the Network entity is responsible for simulating the message exchanges between the other active entities.

3.3 Trace-Driven Simulation of the Google Cluster

We used the parameters extracted from the traces to instantiate and run the simulation model. From the the traces, it appeared that the average memory usage of the Google machines is more or less constant at 50%. According to Google, the remaining memory on each server is reserved to internal processes. Therefore, in the simulation we also set the maximum available memory on each server at half the actual amount of installed RAM.

The purpose of the simulation run was to validate the model by comparing the real traces with simulator results. Four metrics were considered: number of running tasks (Figure 7a), completed tasks (Figure 7b), waiting tasks (ready queue size, Figure 7c) and evicted tasks (Figure 7d). All plots show the time series extracted from the trace data (green lines) and those produced by our simulator (red lines), with the additional application of exponential smoothing to reduce transient fluctuations. The figures show a very good agreement between the simulation results and the actual data from the traces.



(a) Number of running tasks.

(b) Number of tasks completed.

(c) Number of tasks waiting.

(d) Number of tasks evicted.

Figure 7: Simulation and real data for four different metrics.

4 Related work

With the public availability of the two Google cluster traces [Wil11], numerous analyses of different aspects of the data have been reported. These provide general statistics about the workload and node state for such clusters [LC12, RTG+12a, RTG+12b] and identify high levels of heterogeneity and dynamicity of the system, especially in comparison to grid workloads [DKC12]. However, no unified tool for studying the different traces were introduced. BiDAl is one of the first such tools facilitating Big Data analysis of trace data, which underlines similar properties of the public Google traces as the previous studies. Other traces have been analyzed in the past [KTGN10, CGGK11, CAK12], but again without a dedicated tool available for further study.

BiDAl can be very useful in generating synthetic trace data. In general synthesising traces involves two phases: characterising the process by analyzing historical data and generation of new data. The aforementioned Google traces and log data from other sources have been successfully used for workload characterisation. In terms of resource usage,

classes of jobs and their prevalence can be used to characterize workloads and generate new ones [MHCD10, WBMG11], or real usage patterns can be replaced by the average utilization [ZHB11]. Placement constraints have also been synthesized using clustering for characterisation [SCH+11]. Our tool enables workload and cloud structure characterisation through fitting of distributions that can be further used for trace synthesis. The analysis is not restricted to one particular aspect, but the flexibility of our tool allows the the user to decide what phenomenon to characterize and then simulate.

Recently, the Failure Trace Archive (FTA) has published a toolkit for analysis of failure trace data [JKIE13]. This toolkit is implemented in Matlab and enables analysis of traces from the FTA repository, which consists of about 20 public traces. It is, to our knowledge, the only other tool for large scale trace data analysis. However, the analysis is only possible if traces are stored in the FTA format in a relational database, and is only available for traces containing failure information. BiDAl on the other hand provides two different storage options, including HDFS, with transfer among them transparent to the user, and is available for any trace data, regardless of what process it describes. Additionally, usage of FTA on new data requires publication of the data in their repository, while BiDAl can be used also for sensitive data that cannot be made public.

Although public tools for analysis of general trace data are scarce, several large corporations have reported building in-house applications for analysis of logs. These are, in general, used for live monitoring of the system, and analyze in real time large amounts of data to provide visualisation that helps operators make administrative decisions. While Facebook use Scuba [AAB13], mentioned before, Microsoft have developed the Autopilot system [Isa07], which helps administer their clusters. This has a component (Cockpit) that analyzes logs and provides real time statistics to operators. An example from Google is CPI2 [ZTH+13] which monitors Cycles per Instruction for running tasks to determine job performance interference. This helps in deciding task migration or throttling to maintain high performance of production jobs. All these tools are, however, not open, apply only to data of the corresponding company and sometimes require very large computational resources (e.g. Scuba). Our aim in this paper is to provide an open research tool that can be used also by smaller research groups that have more limited resources.

5 Conclusions

In this paper we presented BiDAl, a framework that facilitates use of Big Data tools and techniques for analyzing large cluster traces. BiDAl is based on a modular architecture, and currently supports two storage backends based on SQlite and Hadoop; other backends can be easily added. BiDAl uses a subset of SQL as a common query language that is automatically translated to the appropriate commands supported by each backend. Additionally, data analysis using R and Hadoop MapReduce is possible.

We have described a usage example of BiDAl that involved the analysis of Google trace data to derive parameters to be used in a simulation model of the Google cluster. Distributions of relevant quantities were easily computed using our tool, showing how this

facilitates Big Data analysis even to users less familiar with R or Hadoop. Using the computed distributions, the simulator produces results that are in good agreement with the observed data. Another possible usage of the platform is for application of machine learning tools for predicting abnormal behavior from log data. At the moment, BiDAl can be used for pre-processing and initial data exploration; however, in the future we plan to add new commands to perform this type of analysis directly. Both usage examples could provide new steps towards achieving self-* properties for large scale computing infrastructures in the spirit of Autonomic Computing. In its current implementation, BiDAl is useful for batch analysis of historical log data, which is important for modeling and initial training of machine learning algorithms. However, live log data analysis is also of interest, so we are investigating the addition of an interface to streaming data sources to our platform. Future work also includes implementation of other storage systems, especially to include non-relational models. Improvement of the GUI and general user experience will also be pursued.

References

[AAB13] Lior Abraham, John Allen, and O Barykin. Scuba: diving into data at facebook. *Proceedings of the VLDB Endowment*, 6(11):1057–1067, 2013.

[BOB+14] Alkida Balliu, Dennis Olivetti, Ozalp Babaoglu, Moreno Marzolla, and Alina Sîrbu. BiDAl source code, 2014. URL `http://cs.unibo.it/~sirbu/bidal.zip`.

[CAK12] Yanpei Chen, Sara Alspaugh, and Randy H Katz. Design Insights for MapReduce from Diverse Production Workloads. *Technical Report, University of California Berkeley*, UCB/EECS-2, 2012.

[CGGK11] Yanpei Chen, Archana Ganapathi, Rean Griffith, and Randy Katz. The Case for Evaluating MapReduce Performance Using Workload Suites. *2011 IEEE 19th Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems*, pages 390–399, July 2011.

[DG10] Jeffrey Dean and Sanjay Ghemawat. MapReduce: A Flexible Data Processing Tool. *Communications of the ACM*, 53(1):72–77, January 2010.

[DKC12] Sheng Di, Derrick Kondo, and Walfredo Cirne. Characterization and Comparison of Google Cloud Load versus Grids. In *International Conference on Cluster Computing (IEEE CLUSTER)*, pages 230–238, 2012.

[Isa07] Michael Isard. Autopilot: automatic data center management. *ACM SIGOPS Operating Systems Review*, 41(2):60–67, 2007.

[JKIE13] Bahman Javadi, Derrick Kondo, Alexandru Iosup, and Dick Epema. The Failure Trace Archive: Enabling the comparison of failure measurements and models of distributed systems. *Journal of Parallel and Distributed Computing*, 73(8), 2013.

[KTGN10] Soila Kavulya, Jiaqi Tan, Rajeev Gandhi, and Priya Narasimhan. An Analysis of Traces from a Production MapReduce Cluster. In *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing*, number December, 2010.

[LC12] Zitao Liu and Sangyeun Cho. Characterizing Machines and Workloads on a Google Cluster. In *8th International Workshop on Scheduling and Resource Management for Parallel and Distributed Systems (SRMPDS)*, 2012.

[MHCD10] Asit K Mishra, Joseph L Hellerstein, Walfredo Cirne, and Chita R. Das. Towards Characterizing Cloud Backend Workloads : Insights from Google Compute Clusters. *Sigmetrics performance evaluation review*, 37(4):34–41, 2010.

[R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.

[RTG+12a] Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Heterogeneity and Dynamicity of Clouds at Scale : Google Trace Analysis. In *ACM Symposium on Cloud Computing (SoCC)*, 2012.

[RTG+12b] Charles Reiss, Alexey Tumanov, Gregory R Ganger, Randy H Katz, and Michael A Kozuch. Towards understanding heterogeneous clouds at scale : Google trace analysis. *Carnegie Mellon University Technical Reports*, ISTC-CC-TR(12-101), 2012.

[SCH+11] Bikash Sharma, Victor Chudnovsky, Joseph L Hellerstein, Rasekh Rifaat, and Chita R. Das. Modeling and Synthesizing Task Placement Constraints in Google Compute Clusters. In *2nd ACM Symposium on Cloud Computing (SoCC)*, pages 3:1–3:14, 2011.

[SKRC10] K. Shvachko, Hairong Kuang, S. Radia, and R. Chansler. The Hadoop Distributed File System. In *Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium on*, pages 1–10, May 2010.

[SLM10] Felix Salfner, Maren Lenk, and Miroslaw Malek. A survey of online failure prediction methods. *ACM Computing Surveys (CSUR)*, 42(3):1–68, 2010.

[TSA+10] Ashish Thusoo, Zheng Shao, Suresh Anthony, Dhruba Borthakur, Namit Jain, Joydeep Sen Sarma, Raghotham Murthy, and Hao Liu. Data warehousing and analytics infrastructure at facebook. *Proceedings of the 2010 international conference on Management of data - SIGMOD '10*, page 1013, 2010.

[WBMG11] Guanying Wang, Ali R Butt, Henry Monti, and Karan Gupta. Towards Synthesizing Realistic Workload Traces for Studying the Hadoop Ecosystem. In *19th IEEE Annual International Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS)*, pages 400–408, 2011.

[Wil11] John Wilkes. More Google cluster data. Google research blog, November 2011. Posted at `http://googleresearch.blogspot.com/2011/11/more-google-cluster-data.html`.

[ZHB11] Qi Zhang, Joseph L Hellerstein, and Raouf Boutaba. Characterizing Task Usage Shapes in Google's Compute Clusters. In *Proceedings of the 5th International Workshop on Large Scale Distributed Systems and Middleware*, 2011.

[ZTH+13] Xiao Zhang, Eric Tune, Robers Hagmann, Rohit Jnagal, Vrigo Gokhale, and John Wilkes. CPI 2 : CPU performance isolation for shared compute clusters. In *Proceedings of the 8th ACM European Conference on Computer Systems*, pages 379–391. ACM, 2013.

Asynchronous Logging and Fast Recovery for a Large-Scale Distributed In-Memory Storage

Kevin Beineke, Florian Klein,
Michael Schöttner
Institut für Informatik, Heinrich-Heine-Universität Düsseldorf,
Universitätsstr. 1, 40225 Düsseldorf, Germany
E-Mail: Kevin.Beineke@uni-duesseldorf.de

Abstract: Large-scale interactive applications and online graph analytic processing require very fast data access to many small data objects. DXRAM addresses these challenges by keeping all data always in memory of potentially many nodes aggregated in a data center. Data loss in case of node failures is prevented by an asynchronous logging on flash disks. In this paper we present the architecture of a novel logging service designed to support billions of small data objects, flash disk characteristics and fast node recovery. The latter is challenging if 32-64 GB of in-memory data of a failed node needs to be recovered in seconds from the logs.

1 Introduction

Large-scale interactive applications and online graph computations must often manage many billions of small data objects. Facebook for example supports more than one billion users by keeping around 150 TB of data in more than 1,000 memcached servers [ORS+11]. Around 70% of these objects are smaller than 64 byte [NFG+13] and as a result each memcached server stores a few hundred million of objects. The sheer amount of objects and the small data sizes can be found in many other online graph applications, e.g. [SWW+12]. Another important aspect are the access patterns of these applications where reads dominate over writes [BAC+13], [AXF+12]. And depending on the type of data there are also deletes and updates but both less frequent than reads.

DXRAM is a distributed in-memory system aiming at managing billions of small data objects [KS13]. The core implements a key-value data model for binary data with basic get and put operations. The system design is open to execute algorithms on storage nodes or to run DXRAM as an ultra-fast back-end storage. A super-peer overlay is used for scalable node lookup and a paging-like translation scheme for efficiently mapping global unique keys to local virtual addresses. On top of the core, we plan extended data services, e.g. richer data models and naming/indexing services.
In this paper we present the design of our novel asynchronous logging approach, which is storing replicas of objects on remote flash disks to prevent data loss potentially caused by software and node failures. Replication in memory is not an option as memory is very expensive and power outages may draw down many nodes resulting in data losses. The

logging is designed carefully, taking into account SSD characteristics, to maximize write performance. The latter requires basically sequential writes, which results in a sequential log. This however requires to provide a cleaning mechanism similar to the work published in log-structured file systems [RO92].
Another aspect is that we design the log in a way to allow very fast recovery of nodes states (32-64 GB of data), which is very challenging [ORS+11]. Fast recovery will be even more challenging as we aim at supporting billions of small data objects.

This paper is structured as follows. In section 2 we give a short overview of the DXRAM system including the global meta-data and in-memory data management. Section 3 presents the logging approach, followed by section 4 and 5 where we describe the recovery process and the cleaning of logs. Related work is discussed in section 6 followed by the conclusions and an outlook on future work.

2 DXRAM Architecture Overview

DXRAM is a distributed in-memory storage system designed for the management of binary data in data center environments. We aim at supporting billions of small data objects (16-64 byte) which is for example required for managing very large graphs, e.g. social networks. By keeping all data always in RAM we can provide low-latency data access for all data. Furthermore, by providing transparent persistence we relieve programers from synchronizing caches with secondary storage.

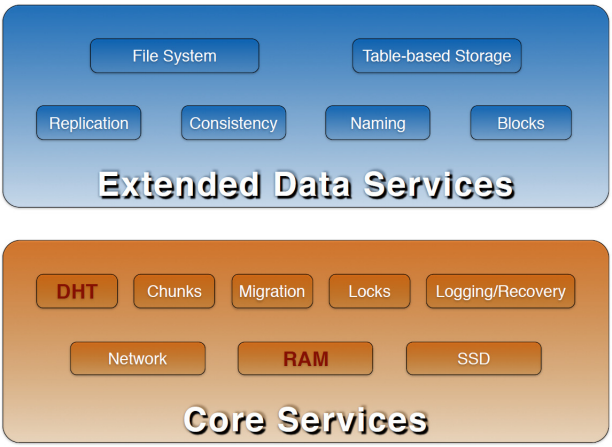


Figure 1: DXRAM architecture

Figure 1 shows the overall architecture of DXRAM. Extended data services include general services and extended data-models built upon core services. Not all extended services shown in the figure are implemented but we designed the core to be flexible allowing for example to support different data models as needed without requiring to re-implement

everything from scratch.

Core services provide functionality for the management, storage and transfer of chunks (key-value tuples). One of the main objectives in core services is to keep the functionality and the interface for high layers as compact as possible. The minimal interface for chunks includes following functions *create*, *delete*, *get*, *put* and *lock*.

The DXRAM core implements a key-value data model where a tuple is called chunk. A chunk has a 64 bit globally unique chunk ID (CID) followed by the value (raw bytes). The CID is composed of two parts. The first 16 bit is the node ID of the creating node (NID_C). The remaining 48 bit is a locally unique value, named LID, which is incremented during each local chunk creation. This results in a sequential CID generation scheme on every node. The key size is configurable but with the described numbers we can address 65,536 nodes each storing up to 280 trillion chunks. We decided to use this scheme in favor of user-defined keys to avoid hashing allowing us to keep IDs compact, to support locality-based access patterns and to avoid adjusting hash functions in case of nodes scaling up and down.
But the sequential ID generation is not a constraint for the applications. On the one hand most applications which use databases as persistent storage access data through auto-incremented row IDs similar to our LID. On the other hand DXRAM provides a name service to address a CID using a user-defined key. The super-peers store the mapping of keys and CIDs in a patricia-trie structure. The intention is that not each single object needs a user-defined key but only a subset, e.g. the user records in a social network.

Chunks have variable sizes defined during their creation always stored en bloc. The basic get and put operations on chunks read and update always full chunks. Because RAM is expensive we replicate chunks for fault tolerance on multiple backup nodes only in flash memory.

2.1 Global Meta-Data Management

A super-peer overlay is used for implementing a custom DHT allowing fast node lookup of chunks. Because of the high-speed network and the limited number of super-peers, e.g. 8-10% of all nodes we decided to keep them knowing each other in contrast to traditional internet-scale DHTs like Chord, CAN, etc. This in turn allows lookups with O(1) time complexity. In addition peers cache node-lookup results reducing load on super-peers. Meta-data is replicated on the neighboring super-peers to mask super-peer failures. However, we do not need a high replication factor, which would be very costly, as meta-data can be dynamically re-constructed.
In a controlled environment like a data center we expect only small node churn caused by failures. Of course a power outage may kill all our nodes and will require to restore all data from all logs, which will take considerable recovery time; but we expect this disaster to be very seldom. Otherwise we expect controlled up and down scaling of nodes as needed dynamically using the super-peer overlay. If we reach a defined threshold a new super-peer is promoted or respectively demoted. In both cases the hashing function for the DHT does

not need to be adjusted thus not requiring to re-hash entries (like for a traditional DHT).

The super-peer overlay can aggregate the meta-data of chunks because of the sequential CID creation and bundles multiple CIDs to a CID Range (start, end) together with the NID_O (o: actual owner). Note, the actual owner of an object may be different to the creator if an object has been migrated, e.g. because of load-balancing reasons.
For example, if we have 1,024 chunks with CIDs 1 to 1024 created on one node, then we have only one entry for this CID range on the associated super-peer. In addition to avoid gaps in the CID ranges through deletions of chunks, we reuse free CIDs for new chunks. As super-peers have to map billions of CIDs from all their associated peers this compact representation is very space efficient.

As mentioned above DXRAM also supports chunk migrations to relieve load on peers storing possibly several hot spots. For example in the context of social networks a hot spot can be a famous artist whose profile is very often visited (some artists have up to 40 million friends in Facebook [SWL13]). Although migration solves the load problem it results in a range split of the global ID range on the super-peer responsible for the migrated object. Still this is not a problem for the global meta-data management because we expect that such chunk migrations are seldom as we do not expect billions of hot spots. A fast lookup of CIDs in chunk ranges is ensured by implementing a B-tree-like data structured described in [KS13].

Beside meta-data management, super-peers are also responsible for recovery coordination of failed nodes which is discussed in Section 4 and for load monitoring which is beyond the scope of this paper.

2.2 In-memory Data Management

We aim at managing up to one billion of small objects on one node, which is challenging especially concerning the mapping of global IDs to local virtual addresses as well as the local memory management. Although both topics are well studied in literature the sheer amount of objects raise new challenges regarding space efficiency and access times.

Many systems use hash tables for the translation of global IDs to local addresses which tend to waste memory or get slow when the load factor increases (causing collisions). In contrast we implemented a paging-like translation scheme which is fast and space efficient.

Furthermore, most memory allocators are not optimized for many small allocations. On the one hand they align memory for cache performance reasons (4/8 byte granularity) and on the other hand they append a header (4 - 64 byte) depending on programming language and runtime system. Obviously, the overhead is too large, e.g. for one billion objects, each with a 16 byte header, we would need 16 GB RAM just for the headers. Because memory is expensive we have decided to design a memory management trading cache performance for minimizing internal fragmentation. Cache penalties are not critical for DXRAM (when used as back-end storage service) as the time for client operation requests is dominated by network transmissions.

The global ID mapping and the local memory management are described in a paper, currently submitted to another conference.

3 Organization of the Log

As mentioned before DXRAM keeps all data always in memory. In order to address node failures, power outages and not to burden the programmer with persistence management on secondary storage we introduce an asynchronous logging service described in this section. Each object is replicated on a configurable number of remote nodes called backup nodes. These backups are stored in a log optimized for SSDs (, which could also be applied to traditional disks). During fault-free execution only write operations of updates are executed on the log but no reads. In case of a recovery request, when a node crashed, the full log is read (described later in more detail). A fast recovery is only possible if the state of one node is spread over many backup nodes. This allows to aggregate SSD and network bandwidths during recovery [ORS+11]. Therefore, we introduce backup zones, similar to the segments in RAMcloud, although we do not use static sizes but allow dynamic size adaption. Each node providing RAM is also providing SSD backup-capacity.

Before we start with the presentation of the log service we discuss the characteristics of SSDs. In contrast to traditional disks, SSDs read and write clustered pages, each 4 KB and each on another NAND flash memory accessed via multiple channels. This means, writing one byte results in writing at least 4 KB. Obviously, as we aim at supporting many small data objects (16 - 64 byte), we need to introduce some buffering for bundling write accesses. To benefit from internal parallelism of SSDs it is even better to write multiple pages at once [MKC+12]. Furthermore, SSDs cannot re-write a page, but instead have to delete the block (64 to 128 pages) first. The write access will then be done on a page the SSD controller chooses. To manage the complexity of that fact the SSD controller uses a specific garbage collection and flash translation layer to hide the characteristics of flash memory. This results in a much higher bandwidth while writing sequentially than randomly [MKC+12]. Finally, two more aspects to be taken into consideration are to avoid delete operations whenever possible because of the large erase block granularity (depending on model, e.g. 256 KB - 2,048 KB) and not to mix read and write accesses as both can harm each other [CKZ09]. The latter is no problem as read and write operations are anyway separated (fault-free execution and recovery), see above.

3.1 Primary Log

DXRAM organizes logs into two levels, one primary log and several secondary logs, one for each node requesting backups. The objective of the primary log is to store incoming backup requests as soon as possible on SSD to avoid data loss caused by software or hardware failures. The primary log is organized as a ring buffer similar to SpriteFS' [RO92] and all incoming backup requests (from any node) are bundled together in the primary log.

In contrast to log-structured file-systems the log is never read during fault-free execution. Thus there is no need to store meta-data of the log in memory. However, the log is self describing, each entry has a header with a length field, NID and LID for identifying the object and optionally a CRC checksum, and is fully read-in during node recovery.

In order to avoid many small write accesses to the primary log, all incoming backup requests are bundled in a write buffer. The buffer itself is organized as a ring buffer and the access is performed using several producer and one consumer thread. The producer threads decouple network threads processing backup requests, however, the node requesting a backup may force a synchronous operation, if requested by the above application. This is optional in case of critical updates, under control by the applications.

It is important to note that the buffer will be filled by backup requests from potentially many nodes, which partially backup their state on a backup node. So, it is likely that there will soon be enough data in the buffer to flush one page out to the primary log or even multiples of a page, depending on the update frequency in the overall system. However, we set a configurable timeout ($<$1s) for the consumer thread to flush out data to SSD. The idea is to avoid an increased data-loss probability in case of very low update frequency.

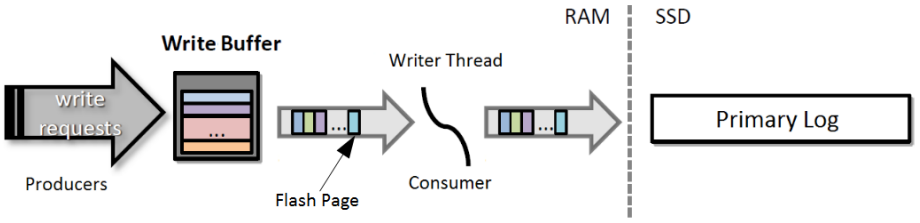


Figure 2: Interaction between write buffer and primary log

Obviously, there is a small time interval where an update is vulnerable to a power outage or failure of all nodes storing the update in memory. Of course in the worst case scenario the full data center could go down by a power outage. The latter is very unlikely, whereas the first may happen. As a consequence we provide applications a sync command to enforce write through for backups of critical updates.

Right now we would have to read in and analyze the full primary log during a recovery requests to detect log entries of the crashed node and skip all other entries. This is obviously inefficient, which in turn lead us to establishing the secondary log.

3.2 Secondary Logs

Each node distributes its state over numerous backup nodes, e.g. 64 GB RAM of one node is distributed over 100 or more backup nodes. Thus each node will service as storage node and as a backup node for potentially many other nodes. To speed up recovery we decided to sort backups into one separate secondary log for each node. Thus we can avoid reading in one large log storing backups from many nodes and analyzing which entries are needed

for a recovery.

After having written a backup request into the primary log, the data is also copied (in memory) into the secondary log buffer using the NID, see Figure 3. The secondary log buffer array grows dynamically to provide one write buffer for each node whose backups are to be stored. The reason for buffering writes to secondary logs is again to bundle them into at least one 4 KB page (similar to the primary log).

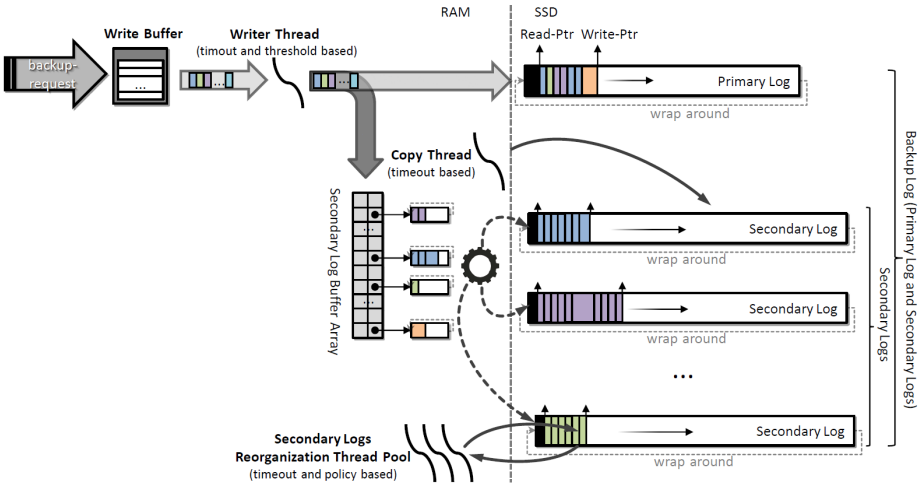


Figure 3: Log overview

It is important to note that the secondary log buffer does not introduce any risk of losing data because all backups are stored in the primary log on SSD. However, the write buffers are flushed periodically (less frequently than the primary write buffer) by the copy thread. The benefits are threefold: First, the primary log can remove all objects that are written to the corresponding secondary logs by adjusting the read pointer only. Second, the reorganization of secondary logs is more efficient with all objects in log (more in Section 5). And third, the recovery process is faster if no or only a few objects in the primary log have to be processed (more in Section 4).

In case of frequent updates from a node, backups may skip the primary log and secondary log buffers and are directly copied from the primary write buffer into the secondary log.

The secondary logs also allows us to reduce the object header per entry. First, there is obviously no need to store the NID_O in the headers because every object in a secondary log has the same NID_O. Furthermore for consecutive object IDs it is sufficient to only store the full LID of the first object and a flag for successor entries. We expect this case to occur quite often as we allocate objects with consecutive IDs on each node and expect updates to be rather seldom. Furthermore, during recovery we have to read in the full secondary log and can reconstruct consecutive IDs easily.

4 Recovery

In large-scale clusters node failures are a rule, not an exception. Therefore, DXRAM keeps the states of nodes in a log spread over backup nodes, see Section 3. Below we describe how the recovery works for a failed node.

Node failures are detected using timeouts in case of data/backup requests by any node. In addition we use the super-peer overlay (described in Section 2), to establish a hierarchical heart beat protocol between each peer and its associated super-peer. Any peer suspecting a peer failure informs the responsible super-peer, which may wait for the next heart beat. The recovery of a failed node is controlled by its associated super-peer, see Figure 4. This will allow distributing recovery control on different super-peers if several peers fail.

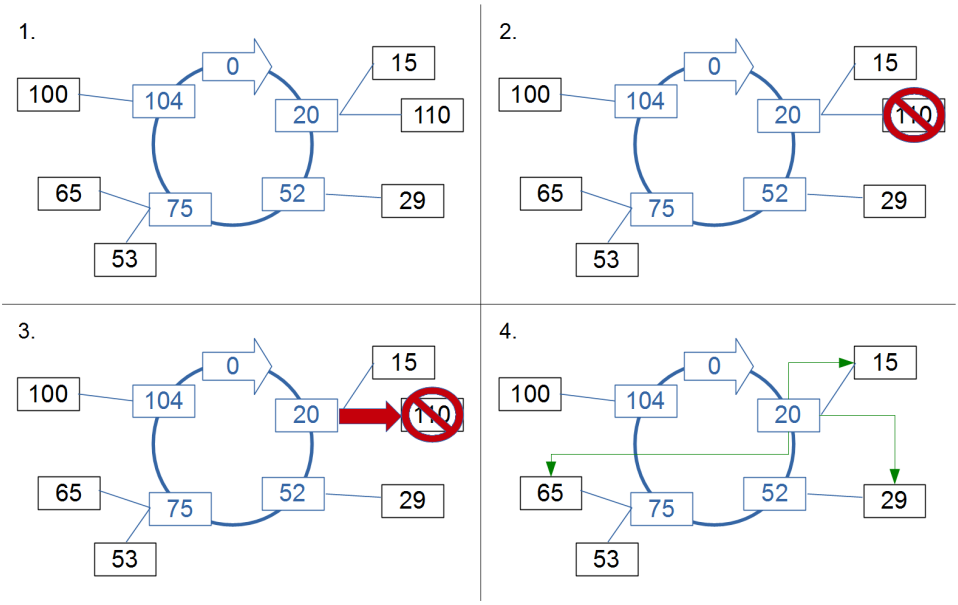


Figure 4: Recovery process: 1. Super-peer overlay with four super-peers (20, 52, 75, 104) and six peers (15, 110, 29, 53, 65, 100) 2. Failure of peer 110 3. Failure detection 4. Distribution of recovery messages

The first step of the super-peer during recovery is to determine backup nodes of the failed peer. As mentioned in Section 2, super-peers provide fast node lookup using CID ranges. These CID range entries do also include a list of backup nodes, typically 2-3 backup nodes. Thus the super-peer can easily determine, which backup nodes to contact to recover the full state of a crashed node. If objects have been migrated they belong to the node state of the new node and no longer to initial creator. As object migrations are also stored in the super-peers (resulting in a CID range split) this does not cause any problems during recovery.

If a super-peer crashes it is recovered by the super-peer overlay by a neighbor super-

peer, which is holding replicas of the meta-data. In order to not burden the system much by meta-data replication on super-peers, we only have 1-3 replica for each super-peer state. As we use dedicated nodes as super-peers, we expect enough memory for such a configuration. In rare cases several super-peers may crash simultaneously, resulting in meta-data loss. Still this disaster scenario will not lead to data loss and after the super-peer overlay structure has been repaired, the new super-peers will recover their meta-data through a multicast to gather CID ranges from their associated peers, which will also include backup nodes and migrated chunks. The chunk IDs of the latter cannot be re-used locally and thus we can use the entry in the paging-like map tables for storing the node where the chunk has been migrated to. This will however also require that if a chunk is migrated from the creator to another node and from there again to a new node, that the creator will be kept informed. This approach is also used after a full power outage, which would require to rebuilt all super-peers, which can be done in parallel like described above.

If a super-peer has gathered all backup nodes to be contacted it will always start with backup 1 and if this one is unavailable go to backup 2 and so on. The idea behind this sequential backup node selection is supported by the strict backup update order during fault-free execution, which is performed asynchronously for performance reasons. The latter ensures that the newest versions are always found in backup 1 and if unavailable we got for 2, where we still hope to find the most recent versions. Depending on the failure situation we might lose the newest version of an object because the object holder crashed and backup 1 crashed too and the other backups did not get updated because of message loss on the network. Obviously, this is a rather seldom event in a data center, it still could occur. In that situation we plan to recover some older version and inform the application. Again we want to point out that the core API will include a sync command, which allows applications to force synchronous backups for critical operations. But of course the overall system performance would be burdened if an application programmer will place after each put a sync.

After a backup node receives a recovery message for a given NID, it first flushes its write buffer and the corresponding secondary write buffer. This will ensure that all backup copies of the failed node are in its secondary log on SSD. It now depends on the configuration regarding the number of backup nodes, which will affect the size of the secondary log, e.g. if we have storage nodes with 32 GB RAM and distribute their state on 100 backup nodes, each backup node will potentially store 320 MB of data in its secondary log. To avoid pressure on the re-organization (described in the following section), we use at least double size for the secondary log, for the example above resulting in 640 MB for the secondary log per node. This is no problem as storage capacity on SSDs is increasing and getting cheaper, so we can expect at least 512 GB, which would allow each node to store logs for around 800 storage nodes. Depending on the free memory of the backup node, it might be possible to read in the full secondary log at once and analyze it in memory with multiple threads benefiting of multiple cores. Analyzing is necessary to determine the newest versions and also deleted chunks. Otherwise, if the backup node does not have enough memory to load the full secondary log, it needs to analyze it step by step, e.g. 16 MB per step.

Assuming all backup nodes have enough memory, they all could recover the secondary

log in their memories, could inform the super-peer, and could from then on serve chunk requests. This is the fastest possible recovery as this approach does run in parallel without any data transfers over the network. However, data locality is not preserved, as the state of the recovered node is now spread over numerous or many nodes. Therefore, it is planned to asynchronously move data from backup nodes to a fresh node to rebuilt the crashed node. In contrast to existing solutions, we plan to do this asynchronously, being able to serve requests, while we move larger chunk sets to the fresh node. So far we have implemented the local recovery, loading the full secondary log into the memory of the backup node and run a multi-threaded analysis.

5 Reorganization

The write buffers and the primary log are flushed periodically or when running full (beyond a given threshold), so there is no need for reorganizing them. However, the secondary logs are contiguously filled with update and delete log entries appended to the end of the log. For each chunk creation, update, and deletion we create one log entry.
Obviously, there is a strong need for reorganization in order to free space of outdated and deleted log entries, at latest when the log size reaches a predefined threshold. This is a known issue in log-structured file systems, which also benefit of the sequential high-speed write operations but have to spend considerable efforts for cleaning the log when running out of space.

Our current prototype uses a rather simple reorganization scheme. If the log size exceeds 75% of the overall capacity we trigger a log reorganization. The latter reads in a full secondary log and runs a multi-threaded cleaning using a thread pool with a work stealing mechanism. The cleaner threads analyze the full log for valid, deleted, and outdated entries. Because of the strict sequential write order we know that newer version are stored in the log after older ones.

While this basic in-memory solution performs well, we are currently studying incremental and more space-efficient approaches avoiding to load the full secondary log into memory. This in turn requires to reason more about the log layout as we cannot afford millions of small random gets and puts on SSD like in memory. Another aspect is related to the update patterns, depending on the applications. If an application updates a few objects many times with the described on-demand reorganization, the log is increasingly populated with many outdated log entries. This slows down the recovery process as all log entry headers need to be analyzed during recovery. We plan to address these problems by an incremental and periodical reorganization scheme.
In [RO92] the authors have shown that the reorganization overhead can be reduced by dividing the log into segments and monitoring write frequency of each segment. The authors then propose to prefer reorganizing segments that are updated less frequently as these segments tend to stay longer organized. This in turn avoids repeated moves of unchanged log entries and leads to a categorization of log entries depending on their update frequency (cold and hot segments). The selection of segments is based on the cost-benefit policy (equation 1) in which ”u” the utilization of the segment and ”age” the age of the youngest

log entry in that segment is. This means the desired segment contains a few, for a long unchanged entries, hence much space can be regained and the resulting new less-filled segment includes data that was changed less frequently in the past.

$$\frac{\text{benefit}}{\text{cost}} = \frac{\text{free space generated} * \text{age of data}}{\text{cost}} = \frac{(1-u) * \text{age}}{1+u} \tag{1}$$

While this strategy enables the reorganization of logs with low additional write costs there are also some shortcomings. First of all this approach requires maintaining some monitoring data in memory per log entry to determine the segment with best cost-benefit coefficient and to decide which log entries are alive. Furthermore, by reorganizing only a subset of log entries at a given time, the system can no longer determine the newest version of an log entry by its position in log (newer versions are nearer to the end of the log). Thus we need to add a version number to the log entry headers. Considering the sheer amount of chunks we expect, this might be an issue. On the other hand by giving up the strict ordering of log entries, techniques like threading and hole-plugging [MRC+97] or slack space recycling [OKC+10] are applicable.

Another interesting approach is to distribute log entries into different logs based on write statistics. This means chunks that are changed frequently are bundled in one log and those that are changed seldom in another one (more than two partitions conceivable). This creates hot and cold zones like with the cost-benefit policy but with shifting the complexity from reorganization to storing of backups. The advantage of this approach is the rather simple reorganization, because the segment selection is less important. However both variants are adaptable which will be an important aspect in finding a space and time efficient reorganization scheme for logging billions of small chunks.

6 Related Work

Logging has been used in different research areas but because of the limited space in this paper we discuss only the most relevant related work.

As DXRAM is inspired by RAMCloud [OAE+10] it is sharing many ideas but there are also important differences, including the logging approach. First of all, RAMCloud is providing a table-based data model and is not aiming at supporting billions of small data objects like DXRAM does. The in-memory layout is a 1:1 copy of the remote log which fits well for a table-based model but not as good for many small objects. The reorganization of the disk logs is controlled by the storage node reorganizing segments in memory and writing the compacted results to new segments on potentially new backup nodes. While the reorganization can run in memory this introduces a heavy network traffic. Therefore, we decided to allow backup nodes to do a reorganization for their secondary logs independent of other backup logs and the storage node. Furthermore RAMCloud does not have the two stage logging approach with the sorting of log entries per node as we have. Finally, it is worth to note that RAMCloud has successfully shown that it is possible to recover the state of node with 35 GB of in-memory data in around 1.6 seconds [ORS+11]. This is a lower

bound as the authors have used 1,000 backup nodes for recovering one node in parallel and used an Infiniband high-speed network.

The idea behind log-structured file systems like SpriteFS [RO92] is to minimize the movements of the actuator arm in traditional HDDs, because moving the actuator arm to the right position and waiting for the rotation until the desired block is accessible can take numerous milliseconds. As a consequence SpriteFS tried to write sequentially whenever possible, which in turn means that block updates are appended with higher version instead of replacing blocks in place. Read accesses are mostly served from the block cache avoiding heavy random reads on the HDD. With time the log becomes overfilled with old versions and lacks free space to store new objects or updates. SpriteFS addressed this by dividing the log into segments and execute a cleaner on segments periodically. During the cleaning, old versions are deleted and the segment is defragmented. After defragmentation the space of the old segment is freed and the remaining objects are appended to the log. In addition the cleaner does not choose segments with the same probability but uses a cost benefit policy to maximize the amount of space that is regained with every processed segment.

SFS is a more recent work transferring the advantages of a log-structured file system from HDDs to SSDs [MKC+12]. For better performance the segment size fits the block size of the SSDs. Another interesting aspect is that the hot and cold zones are generated proactively with statistics on block level to increase the efficiency of the defragmentation. It is future work for DXRAM to refine the cost benefit policy and to optimize it for billions of small data objects stored on SSDs.

Logging has also a long tradition in database systems, where typically a write-ahead log (WAL) is used. However, these logs are totally different from the one proposed in this paper. WAL logs do not store the data per se but the transaction's meta-data to apply transactions on backup databases or to repeat failed transactions. The data objects are stored on persistent memory anyway. If the log is overfilled old log entries are flushed to the database and will be overwritten. So, there is no need for a log cleaner.

Finally, there have been proposed many SSD-based key value stores, which however typically use SSDs as caches for HDDs in order to speed up data accesses. One example is Flashstore [DSL10], which uses a cuckoo-based hash-table with compact key signatures as index structure in RAM to improve access times for reading objects from SSD. While the related insights are interesting for meta-data management in DXRAM, they do not affect the log architecture.

7 Conclusion

DXRAM is a distributed in-memory system that is designed to manage billions of small data objects. By keeping all data always in memory it is providing low-latency data access to all data, while at the same time relieving programmers from keeping caches and secondary storage synchronized. The latter requires an efficient asynchronous logging of in-memory data on remote flash disks for providing fast recovery from node failures and

avoiding data loss.

We have presented a novel two stage logging approach sorting log entries per node allowing fast recovery of failed nodes. The latter is also supported by spreading the state of one node on potentially many backup node logs, which allows a parallel recovery from many nodes. As known from the systems literature, sequential logging is fast but requires a cleaner to remove outdated and deleted data and compact the log to free space. We believe that the cost-benefit approach like proposed for log-structured files systems is a good base for a more optimized custom solution for an in-memory based storage.

Future work includes the design of cleaner policies as well as evaluations with different application access patterns (including recovery times depending on different configurations). Currently, we are adopting the BGbenchmark [BG13], which allows to run actions of social network applications on top of different storage technologies (e.g. MongoDB and SQL-X with memcached). We plan to use this benchmark for comparing DXRAM performance with key-value and distributed in-memory caches (e.g. Gemfire and Hazelcast) as well as the recovery times for single and multiple node failures.

References

[AXF+12] Berk Atikoglu, Yuehai Xu, Eitan Frachtenberg, Song Jiang, and Mike Paleczny. Workload Analysis of a Large-scale Key-value Store. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '12, pages 53–64, New York, NY, USA, 2012. ACM.

[BAC+13] Nathan Bronson, Zach Amsden, George Cabrera, Prasad Chakka, Peter Dimov, Hui Ding, Jack Ferris, Anthony Giardullo, Sachin Kulkarni, Harry Li, Mark Marchukov, Dmitri Petrov, Lovro Puzar, Yee Jiun Song, and Venkat Venkataramani. TAO: Facebook's Distributed Data Store for the Social Graph. In *Proceedings of the 2013 USENIX Conference on Annual Technical Conference*, USENIX ATC'13, pages 49–60, Berkeley, CA, USA, 2013. USENIX Association.

[BG13] Sumita Barahmand and Shahram Ghandeharizadeh. BG: A Benchmark to Evaluate Interactive Social Networking Actions. In *CIDR*, 2013.

[CKZ09] Feng Chen, David A. Koufaty, and Xiaodong Zhang. Understanding Intrinsic Characteristics and System Implications of Flash Memory Based Solid State Drives. In *Proceedings of the Eleventh International Joint Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '09, pages 181–192, New York, NY, USA, 2009. ACM.

[DSL10] Biplob Debnath, Sudipta Sengupta, and Jin Li. FlashStore: high throughput persistent key-value store. *Proc. VLDB Endow.*, 3(1-2):1414–1425, September 2010.

[KS13] F. Klein and M. Schoettner. Dxram: A persistent in-memory storage for billions of small objects. In *Proceedings of the 14th International Conference on Parallel and Distributed Computing, Applications and Technologies*, PDCAT 13, 2013.

[MKC+12] Changwoo Min, Kangnyeon Kim, Hyunjin Cho, Sang-Won Lee, and Young Ik Eom. SFS: Random Write Considered Harmful in Solid State Drives. In *Proceedings of the*

10th USENIX Conference on File and Storage Technologies, FAST'12, pages 12–12, Berkeley, CA, USA, 2012. USENIX Association.

[MRC+97] Jeanna Neefe Matthews, Drew Roselli, Adam M. Costello, Randolph Y. Wang, and Thomas E. Anderson. Improving the Performance of Log-structured File Systems with Adaptive Methods. In *Proceedings of the Sixteenth ACM Symposium on Operating Systems Principles*, SOSP '97, pages 238–251, New York, NY, USA, 1997. ACM.

[NFG+13] Rajesh Nishtala, Hans Fugal, Steven Grimm, Marc Kwiatkowski, Herman Lee, Harry C. Li, Ryan McElroy, Mike Paleczny, Daniel Peek, Paul Saab, David Stafford, Tony Tung, and Venkateshwaran Venkataramani. Scaling Memcache at Facebook. In *Presented as part of the 10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*, pages 385–398, Lombard, IL, 2013. USENIX.

[OAE+10] John Ousterhout, Parag Agrawal, David Erickson, Christos Kozyrakis, Jacob Leverich, David Mazières, Subhasish Mitra, Aravind Narayanan, Guru Parulkar, Mendel Rosenblum, Stephen M. Rumble, Eric Stratmann, and Ryan Stutsman. The case for RAMClouds: scalable high-performance storage entirely in DRAM. *SIGOPS Oper. Syst. Rev.*, 43(4):92–105, January 2010.

[OKC+10] Yongseok Oh, Eunsam Kim, Jongmoo Choi, Donghee Lee, and Sam H. Noh. Optimizations of LFS with Slack Space Recycling and Lazy Indirect Block Update. In *Proceedings of the 3rd Annual Haifa Experimental Systems Conference*, SYSTOR '10, pages 2:1–2:9, New York, NY, USA, 2010. ACM.

[ORS+11] Diego Ongaro, Stephen M. Rumble, Ryan Stutsman, John Ousterhout, and Mendel Rosenblum. Fast Crash Recovery in RAMCloud. In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, SOSP '11, pages 29–41, New York, NY, USA, 2011. ACM.

[RO92] Mendel Rosenblum and John K. Ousterhout. The Design and Implementation of a Log-structured File System. *ACM Trans. Comput. Syst.*, 10(1):26–52, February 1992.

[SWL13] Bin Shao, Haixun Wang, and Yatao Li. Trinity: A Distributed Graph Engine on a Memory Cloud. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 505–516, New York, NY, USA, 2013. ACM.

[SWW+12] Zhao Sun, Hongzhi Wang, Haixun Wang, Bin Shao, and Jianzhong Li. Efficient Subgraph Matching on Billion Node Graphs. *Proc. VLDB Endow.*, 5(9):788–799, May 2012.

Hochschule 2020

Hochschule 2020:
Aktuelle Entwicklungen, Herausforderungen und Trends

Djamshid Tavangarian, Ulrike Lucke, Wolfgang Karl

Universität Rostock, Universität Potsdam, Karlsruher Institut für Technologie
http://www.cs.uni-potsdam.de/hochschule2020

Die wachsende Bedeutung der Mikro- und Makromobilität von Studierenden, Mitarbeitern und Geräten, allgegenwärtigem IT-Zugang, dynamischer Allokation von personalisierten Inhalten sowie flexibler Gestaltung und Organisation von Community- sowie individuellen Lern- und Arbeitsprozessen stellt moderne Bildungseinrichtungen bei ihrer stetigen Entwicklung vor technische, organisatorische und inhaltliche Herausforderungen. Die Workshop-Reihe „*Hochschule 2020*" stellt diese Themen zur Diskussion:

- Aus **technischer Sicht** fordern die breite Verfügbarkeit mobiler Netzwerktechnologien, smarte und kontextbewusste Nutzerendgeräte sowie verteilte, oft nutzergenerierte Inhalte eine adäquate ITInfrastruktur. Sie muss in der Lage sein, flexibel, transparent und dynamisch verschiedene Geräte- und Anwendungsklassen mit unterschiedlichen Dienstportfolios zu vernetzen, um einen verlässlichen Informationsaustausch zwischen Individuen und Communitys sicher zu stellen. In diesem Forschungs- und Entwicklungsgebiet gab und gibt es bereits vielfältige Ansätze, beispielsweise zur Einführung serviceorientierter Mechanismen und intelligenter Agenten. Ein Schwerpunkt des Workshops ist daher die Einführung adäquater technischer Infrastrukturen im Rahmen ganzheitlicher und systematischer Organisationskonzepte.

- Zudem sorgt die – durch Bologna geforderte – Mobilität der Studierenden und Mitarbeiter weltweit für diverse **organisatorische Herausforderungen**. Der hochschulübergreifende Austausch erworbener Studierendenleistungen, das Angebot von Schnittstellen für externe Bildungseinrichtungen, der effiziente Umgang mit persönlichen ePortfolios und die Entwicklung und Einführung flexibler Support-Strukturen sind nur einige Beispiele dafür, was ein modernes IT-basiertes Campus-Management neben der Bereitstellung und Verwaltung von mediengerechten Inhalten, insbesondere für mobile Szenarien leisten muss. Innovative Ansätze, aktuelle Entwicklungen und Ergebnisse in Forschung und Lehre sind der zweite Themenschwerpunkt des Workshops.

- Die Symbiose aus traditionellen und mediengerechten **Lehr-/Lernszenarien** und ihre Nutzung in unterschiedlichen Phasen von Bildungsprozessen (Präsenz, Remote, Blended etc.) stellen einen dritten Schwerpunkt des Workshops dar.

Sie werden heute nicht mehr nur lokal an Hochschulen entwickelt und vermittelt. Auch die – oftmals internationale – Übertragung und Anpassung der Inhalte unter Nutzung modernster Medien spielt im Zuge der Globalisierung eine wichtige Rolle. Sie ist die Grundlage für die Förderung und Stärkung der globalen Lehr-/Lern-Community und den akademischen Austausch. Allerdings verlangen gerade der internationale Charakter einer derartigen Gemeinschaft sowie die zunehmende Nutzung verschiedener Medien und Plattformen neue Lösungen an traditionellen Bildungseinrichtungen.

Die im Workshop präsentierten aktuellen Entwicklungen, Herausforderungen und Trends zu den Themen IT-Infrastrukturen, Organisationsstrukturen und Lehr-/ Lernszenarien für zukunftsorientierte Bildungseinrichtungen umfassen in diesem Jahr sechs Beiträge, die vom Programmkomitee begutachtet und ausgewählt wurden:

- Die technische Perspektive auf IT-Infrastrukturen an Hochschulen bilden die Beiträge „Hochschulrechenzentren als IT-Stadtwerke – Konzepte und kritische Diskussion innovativer IT-Dienstangebote von Hochschulen“ sowie „Eine IT-Referenzarchitektur für Hochschulen: Vom generischen Modell zur spezifischen Lösung“.

- Aus der organisatorischen Sicht auf innovative IT-Aspekte widmet sich der Workshop „CIO-Modellen an deutschen Hochschulen - Umsetzung der IT-Governance“ sowie „Hochschul-Apps im Überblick“.

- Lehren und Lernen sind durch zwei Beiträge aus dem Bereich Studienplanung repräsentiert: „Get your Study Plan“ und „Einsatz von MultiTouch-Tischen zur kooperativen Semesterorganisation“.

Die wissenschaftlichen Beiträge werden durch Diskussionen abgerundet und bereichert. Wir danken allen Einreichern für die sorgfältige Aufbereitung ihrer Arbeitsergebnisse sowie den Mitgliedern des Programmkomitees für die Mitwirkung bei der Begutachtung und Auswahl der Beiträge. Den Organisatoren der GI-Jahrestagung danken wir für die Unterstützung bei der Ausrichtung der Veranstaltung. Nicht zuletzt gilt Raphael Zender und Francis Zinke unser großer Dank für ihr unermüdliches Engagement und ihre große Hilfe rund um die Organisation des Workshops.

Programmkomitee: Arndt Bode (LRZ München), Michael Breitner (Uni Hannover), Marc Göcks (Multimedia Kontor Hamburg), Stefan Fischer (Uni Lübeck), Heinz-Gerd Hegering (LMU München), Markus von der Heyde (vdH-IT, Weimar), Christoph Igel (Uni des Saarlandes), Odej Kao (TU Berlin), Michael Koch (UniBW München), Bernd Krämer (FU Hagen), Hans Pongratz (TU München), Christa Radloff (Uni Rostock), Peter Schirmbacher (HU Berlin), Ulrik Schroeder (RWTH Aachen), Bernd E. Wolfinger (Uni Hamburg), Markus Zahn (Uni Augsburg)

Organisation: Raphael Zender (Uni Potsdam), Francis Zinke (Uni Potsdam)

CIO-Modelle an deutschen Hochschulen – Netzwerk der entscheidenden Faktoren

Dr. Markus von der Heyde

vdH-IT
Rosenweg 7
99425 Weimar
info@vdh-it.de

Abstract: Die Steuerung und Entscheidungen zum Einsatz von modernen Informations- und Kommunikationstechnologien (IuK) ist für Hochschulen zu einer zentralen Aufgabe geworden. Die gesamte Institution umfassende Projekte, Kooperationen zwischen Hochschulen und der stetig wachsende Finanzdruck führten seit mehr als zehn Jahren zu Empfehlungen der DFG, des DINI, des ZKI und anderen nationalen sowie föderalistischen Gremien und Institutionen. Die Einführung von für Hochschulen geeigneten CIO-Modellen ist keine Frage, sondern der klare Auftrag.

In einer vom ZKI beauftragten Studie wurden ca. die Hälfte aller bis Frühjahr 2014 etablierten Modelle untersucht. In einem strukturierten Interview wurden CIOs bzw. Mitglieder von CIO-Gremien telefonisch befragt und danach die Gespräche quantitativ ausgewertet. Die Analyse zeigt einige Korrelationen und Unterschiede zwischen den 28 Teilnehmenden der Studie. Dieser Beitrag fasst Kernfaktoren (Einfluss der Entscheidungsdomänen, für CIO-Aufgaben aufgebrachte Zeit und organisatorische Verankerung der CIO-Position) zusammen und stellt diese gegenseitig in Beziehung und damit in einen größeren Zusammenhang. Die Analyse konzentriert sich primär auf den Blickwinkel, wie die Faktoren der jeweils implementierten CIO-Modelle mit anderen Maßgrößen der Hochschulen im statistischen Zusammenhang stehen.

1 Motivation und Methode

In der Wirtschaft sind CIO-Modelle lange Jahre erfolgreich etabliert und wissenschaftlich untersucht worden. Die Forderung, ähnliche Modelle in deutschen Hochschulen zu etablieren, sind seit 2001 von verschiedenen Institutionen und Organisationen betont worden [MA01, Ga03, MA06, He08, BK10, GP11 und Hr13].

Weill und Ross führten 2004 [WR04] die Sichtweise der Entscheidungsdomänen zur Umsetzung einer IT-Governance ein und wiesen nach, dass die Steuerung von IT geeignete Entscheidungsstrukturen benötigt. Das aus diesen Erkenntnissen abgeleitete Framework ist 2008 als ISO Standard [Is08] erschienen.

In europäischen Universitäten wurden erste Analysen, wie weit höhere Bildungseinrichtungen überhaupt diesen Standard implementieren (können), in mehreren

Ländern durchgeführt. Zum Beispiel berichten Fernández und Llorens im Rahmen der EUNIS 2009 zu einer landesweit angelegten Initiative, die Steuerungsprozesse an das ISO38500 Framework anzupassen [FL09].

Von Schwabe wurde 2008 ein Vergleich des Reifegrads der Umsetzung von IT-Governance in Österreich, der Schweiz und Deutschland vorgenommen [Sc10] und festgehalten, dass die Umsetzung allenfalls einem wiederholbaren, aber keinem verbesserten Prozess entspricht. In zwei deutschen Hochschulen wurde der Stand 2010 von Börgmann und Bick untersucht und qualitativ beschrieben [BB11, Bi11].

Quantitative Aussagen zur Umsetzung der IT-Governance und Hinweise zu spezifischen Modellen an deutschen Hochschulen fehlten bisher.

Der ZKI e.V. hat daher eine Studie durchgeführt, in der zwei Leitfragen beantwortet werden sollten:

- Welche CIO-Modelle und Umsetzung einer IT-Governance haben sich an deutschen Hochschulen entwickelt?
- Welche quantitativen Faktoren tragen in welcher Weise zu einem Erfolg der Modelle bei?

Die Studie wurde im Zeitraum Februar und März 2014 mit 28 deutschen Hochschulen (6 Fachhochschulen und 22 Universitäten) durchgeführt. Insgesamt wurden zu diesem Zeitpunkt 56 Hochschulen identifiziert, an denen eine CIO-Struktur bestand bzw. im Aufbau war. Die Teilnehmer hatten sich freiwillig zur Teilnahme an einem Telefoninterview mit einem strukturierten Fragenkatalog bereit erklärt. Die Interviews hatten eine Länge zwischen 45 und 120 Minuten. In „CIOs und IT-Governance an deutschen Hochschulen“ [He14] werden weitere Details zu Methoden und Hintergründen dargestellt. Eine dort bereits vorgenommene statistische Analyse erläuterte einen Großteil der Unterschiede in den quantitativen Faktoren der Modelle. Weiter konzentriert sich [He14] auf den Vergleich zwischen den empfohlenen und den letztlich realisierten CIO-Modellen und schließt mit einer Sammlung von konkreten Empfehlungen zur Vorgehensweise bei dem Aufbau von CIO-Strukturen.

Die hier vorgestellten zusätzlichen statistischen Ergebnisse und die zusammenfassende Übersicht als Netzwerk sind nicht Gegenstand der Publikation im ZKI.

Um den Artikel sprachlich zu vereinfachen, wird oftmals vom CIO in männlicher Einzahl gesprochen. Dabei soll nicht angedeutet werden, dass nicht auch weibliche CIOs an der Studie teilgenommen haben und mit den Aussagen gemeint sind. Ebenfalls wird ein Interviewteilnehmer, wenn er oder sie einem CIO-Gremium angehört oder dieses repräsentiert, als CIO bezeichnet.

2 Analyse

Vorab und in den Interviews wurden kategoriale Variablen (= Faktoren) bzw. Zielvariablen (abhängige Variablen) erfasst. So entstammen z.B. die Zahlen zur Anzahl der Studierenden, dem Alter einer Institution oder die absolute Drittmittelhöhe von Universitäten öffentlich zugänglichen Quellen wie DFG und HRK.

In einer anonymen Auswertung dieser Variablen wurden, da für die meisten Größen keine für T-Tests hinreichende Annahme über die homogene Streuung und Gaußsche Normalverteiltheit anzunehmen ist, statistische Proben nach Mann-Whitneys U-Test (auch bekannt als Wilcoxon-Rangsummentest) gerechnet. Dabei wurde die Gesamtheit der 28 Teilnehmenden in jeweils zwei Gruppen anhand einer Größe (z.B. CIO berichtet an das gesamte Rektorat versus CIO berichtet an höchstens eine spezifische Person) geteilt. Für die übrigen Daten wurde dann geprüft, ob die Hypothese erfüllt ist, dass die Gruppen sich nicht statistisch unterscheiden. Mit anderen Worten, wenn diese Hypothese nicht erfüllt ist, wird geschlossen, dass die beiden Gruppen nicht einer Grundgesamtheit entstammen, sich also in einer Zielvariablen (z.B. in der Quote der Drittmittel vom Gesamthaushalt) im Rangordnungsverfahren unterscheiden. In diesem Fall kann also die Variable (Berichtsrichtung) als „Prädiktor“ und die Zielvariable (Drittmittelquote) als „Indikator“ interpretiert werden.

Es werden jeweils die Gruppenstärken, Mittelwerte der zwei Gruppen, der U-Wert des Tests sowie die Standardabweichung der Gesamtstichprobe berichtet. Die statistische Signifikanz des U-Tests wird mit (*) auf dem 5% Niveau der Irrtumswahrscheinlichkeit angezeigt. Mit (**) wird das 1% Niveau und mit (***) das 0.1% Niveau ausgezeichnet.

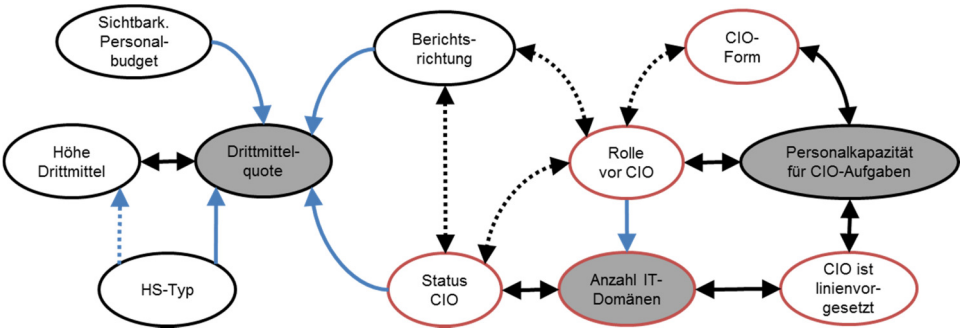


Abbildung 1: Netzwerk von Faktoren, die untereinander in Pfeilrichtung signifikante statistische Beziehungen besitzen. Schwarze Pfeile zeigen eine bidirektionale Signifikanz an. Blaue Pfeile zeigen die Signifikanz in nur eine Richtung an. Die rote Umrandung kennzeichnet diejenigen Faktoren, die Prädiktoren für das Alter der Institution sind. Gestrichelte Pfeile kennzeichnen statistische Resultate, die zwar bestehen, aber nicht näher erläutert werden.

Die folgenden Abschnitte erläutern einige zentrale Faktoren (graue Ellipsen aus Abbildung 1), die bei dieser Analyse auffällig stark wirksam bzw. sichtbar werden.

2.1 Anzahl der vom CIO entscheidungsbefugten IT-Domänen

Die Anzahl der IT-Domänen (nach dem IT-Governancemodell [WR04] bzw. der ISO 38500 [Is08] sind dies IT-Strategie, -Architektur, -Infrastruktur, -Applikationen, Priorisierung bzw. IT-Kosten), in denen der CIO abschließende Entscheidungen treffen kann, war statistisch signifikant als Indikator [StdA=2,04] für folgende Prädiktoren:

- Der CIO hat die direkte Linienverantwortung für alle zentralen IT-Einrichtungen [U=26**]. Gruppe 1: Linienverantwortung besteht [N1=7; M1=3,29]. Gruppe 2: Es besteht keine Verantwortung des CIO für die zentralen IT-Einrichtungen. In der Regel sind in diesem Fall RZ-Leitung und CIO gleichberechtigt bzw. CIO und RZ-Leitung werden in Personalunion ausgeübt [N2=21; M2=0,81].
- Organisatorischer Status / Rolle des Interviewpartners [U=46*]. Gruppe 1: CIO stammt aus der Forschung und Lehre bzw. ist Stabsstelle [N1=15; M1=2,40]. Gruppe 2: CIO hat die Leitung einer zentralen Einheit inne [N2=13; M2=0,31].
- Aufgabe / Beruf des CIO vor der Übernahme der aktuellen CIO-Rolle [U=40*]. Gruppe 1: CIO war zuvor in der IT-Versorgung von (meistens der aktuellen) Hochschulen tätig [N1=10; M1=0,00]. Gruppe 2: CIO war zuvor in der Wirtschaft tätig oder hat/hatte eine professorale Rolle in Forschung und Lehre [N2=18; M2=2,22].

Zusammenfassung: CIOs, die eine Linienverantwortung für alle zentralen Einheiten besitzen, eine professorale Rolle oder Stabsstelle bekleiden oder aus Forschung und Lehre bzw. der Wirtschaft stammen, bekommen überdurchschnittlich oft die Entscheidungskompetenzen für mindestens eine der IT-Domänen der aus dem IT-Governancemodell übertragen.

Gleichzeitig hat die Anzahl der IT-Domänen als Prädiktor mit Gruppe 1 [N1=18] „keine Domäne“ und Gruppe 2 [N2=10] „mindestens eine Gruppe wird vom CIO abschließend entschieden“ für folgende Indikatoren Signifikanz:

- CIO hat Linienverantwortung für alle zentralen IT-Einheiten [M1=6%; M2=60%; StdA=43%; U=41*].
- Der Interviewpartner hat professoralen Status [M1=44%; M2=100%; StdA=48%; U=40*].
- Der Interviewpartner ist Leiter einer zentralen Einheit [M1=67%; M2=10%; StdA=50%; U=39*].
- Alter der Institution in Jahren [M1=154; M2=63; StdA=146; U=46*].

Zusammenfassung: CIOs, die Entscheidungsbefugnisse besitzen, haben häufiger eine Linienverantwortung für alle Leiter von zentralen IT-Einheiten, besitzen den professoralen Status, sind nicht direkter Leiter einer zentralen Einheit und arbeiten an jüngeren Einrichtungen.

Interpretation: Da in der IT-Governance die Regelungen zu den Entscheidungsbefugnissen in den IT-Domänen von kritischer Bedeutung ist, spielen sie auch an

Hochschulen eine zentrale Rolle. Oftmals enthielten die schriftlichen Regelungen der untersuchten Hochschulen – so sie denn überhaupt existieren – keine klaren Aussagen in dieser Richtung. Offensichtlich ist eine der Hauptdeterminanten für den entscheidenden Schritt der Verlagerung von (Teil-)Verantwortung aus der Hochschulleitung hinaus an die Erfahrungen des CIO vor einer aktuellen Position gebunden. Zumindest zeigt die aktuelle Situation, dass die aus den eigenen Reihen der Hochschule stammenden IT-Mitarbeiter nicht in die Verantwortung genommen werden. Da die beiden anderen Faktoren („Status des CIO“ und „CIO ist linienvorgesetzt“) bidirektional mit der Anzahl der IT-Domänen verknüpft sind, ist aus dieser Analyse kein klares Wirkungsgefüge ableitbar.

2.2 Drittmittelquote als Anteil der Drittmittel vom Gesamthaushalt

Berechnet man den Anteil (nicht die absolute Höhe!) des Gesamthaushalts, den jede Institution durch die Einwerbung von Mitteln Dritter (z.B. Forschungsmittel von DFG, EU usw.), so ist dieser Indikator [StdA=9,4%] von folgenden Prädiktoren determiniert worden:

- Hochschultyp bzw. Promotionsrecht [U=9***]. Gruppe 1: Universitäten mit Promotionsrecht [N1=22; M1=20,5%]. Gruppe 2: Fachhochschulen ohne Promotionsrecht [N2=6; M2=7,8%].
- Organisatorischer Status / Rolle des Interviewpartners [U=51*]. Gruppe 1: CIO stammt aus der Forschung und Lehre bzw. ist Stabsstelle [N1=15; M1=21,7%]. Gruppe 2: CIO hat die Leitung einer zentralen Einheit inne [N2=13; M2=13,2%].
- Berichtsrichtung des CIO [U=44*]. Gruppe 1: Es wird (unspezifisch) an das gesamte Präsidium berichtet [N1=18; M1=15%]. Gruppe 2: Es wird höchstens eine konkrete Person im Präsidium berichtet [N2=10; M2=22,8%]. Dabei waren in der zweiten Gruppe sechs Fälle, bei denen direkt an den Präsidenten berichtet wurde, zwei Fälle berichteten an einen der Vizepräsidenten und zwei Interviewpartner gaben an, nicht zu berichten.
- Sichtbarkeit von IT-Personal für den CIO [U=46*]. Gruppe 1: CIO hat keine Übersicht bzw. kennt ausschließlich die Personalzahlen aus dem zentralen Rechenzentrum [N1=11; M1=12,9%]. Gruppe 2: der CIO hat (regelmäßige und aktuelle) Kenntnis der Personalzahlen aller zentralen Einheiten bzw. generell institutionsweit [N2=17; M2=20,9].
- Absolute Höhe der Drittmittel [U=50*]. Gruppe 1: weniger als 28 Mio € pro Jahr [N1=14, M1=13,5%]. Gruppe 2: mehr als 28 Mio € pro Jahr [N2=14; M2=22,0%].

Zusammenfassung: Hochschulen, die a) das Promotionsrecht besitzen (in unserem Fall alle Universitäten), b) bei denen der CIO eine Übersicht zum IT-Personal besitzt, die über das Rechenzentrum hinausgeht, c) bei denen der CIO nicht Leiter der zentralen IT-Einrichtung ist, d) der CIO nicht unspezifisch an das ganze Präsidium berichtet oder e) die eine absolutes Drittmittelaufkommen von mehr als 28 Mio. € besitzen, verfügen statistisch um eine um mehr als 8% höhere Drittmittelquote als andere Hochschulen.

Wiederum umgekehrt wurde die Drittmittelquote als Prädiktor getestet. Bei einer ausgeglichenen Gruppenstärke [N1=14; N2=14] ergeben sich keinerlei signifikante Indikatoren. Bei einer Teilung der Stichprobe in Gruppe 1 mit weniger als 17% [N1=16] und Gruppe 2 [N2=12] mehr als 17% Drittmittelquote zeigt sich die absolute Höhe der Drittmittel als signifikanter Indikator [M1=33 Mio €; M2=67 Mio €; StdA=53 Mio €, U=47*].

Interpretation: Einige der heute bestehenden Ranking-Modelle greifen die Höhe der eingeworbenen Drittmittel als Indikator auf. Die nach außen erzielte Sichtbarkeit ist damit direkt an den erzielten Forschungserfolg, zumindest aber an den Erfolg bei der Finanzierung von Forschungsleistungen, gekoppelt.

Die Mehrheit der erfolgsorientierten Mittelverteilungsmodelle in den Bundesländern hat die absolute Höhe der Drittmittel als Faktor im Verteilungsschlüssel verrechnet. Da somit ebenfalls die Höhe der verfügbaren Mittel in einem gewissen korrelierten Maß steigt oder sinkt, wurde in dieser Studie auch die abgeleitete Größe der „Drittmittelquote“ in die Analyse einbezogen.

Die drei mit den CIO-Modellen bzw. mit der IT-Governance verbundenen Faktoren („Status-CIO“, „Berichtsrichtung“ und „übergreifende Sichtbarkeit von IT-Person- bzw. IT-Sachmitteln“) sind jeweils unidirektional mit Wirkrichtung auf die Drittmittelquote sichtbar geworden. Somit erscheint die Schlussfolgerung legitim, dass nicht Hochschulen mit besonders hoher Drittmittelquote bestimmte CIO-Modelle implementieren, sondern andersherum, dass bestimmte Faktoren in den CIO-Modellen statistisch mit einer hohen Quote der erzielten Drittmittel verbunden sind. Klar dabei ist, dass hier kein direkter Wirkzusammenhang hergestellt werden kann. Es scheint vielmehr so zu sein, dass Hochschulen, die bestimmte Entscheidungen zu ihrer IT-Governance und dem etablierten CIO-Modell getroffen haben, ebenfalls förderliche Entscheidungen in Bezug auf Berufungspolitik, Forschungsausrichtung u.d.g. treffen.

2.3 Verfügbare Arbeitskraft für CIO-Aufgaben

Die insgesamt für CIO-Tätigkeiten verfügbare Arbeitskraft vom CIO und weiterem Personal, gemessen in Vollzeitäquivalenten (VZÄ), war statistisch signifikanter Indikator [StdA=0,67 VZÄ] für folgende Prädiktoren:

- Der CIO hat die direkte Linienverantwortung für alle zentralen IT-Einrichtungen [U=33,5*]. Gruppe 1: Linienverantwortung besteht [N1=7; M1=1,25]. Gruppe 2: Es besteht keine Weisungsbefugnis des CIO für die zentralen IT-Einrichtungen [N2=21; M2=0,67].
- Aufgabe / Beruf des CIO vor der Übernahme der aktuellen CIO-Rolle [U=44*]. Gruppe 1: CIO war zuvor in der IT-Versorgung von (meistens der aktuellen) Hochschulen tätig [N1=10; M1=0,41 VZÄ]. Gruppe 2: CIO war zuvor in der Wirtschaft tätig oder hat/hatte eine professorale Rolle in Forschung und Lehre [N2=18; M2=1,03 VZÄ].

- CIO-Form nach DFG Empfehlung [U=26**]. Gruppe 1: CIO als Gremium etabliert [N1=8; M1=0,26 VZÄ]. Gruppe 2: CIO als Personenmodell realisiert [N2=20; M2=1,03VZÄ].

Zusammenfassung: Bei Einrichtungen mit direkter Linienverantwortung für alle IT-Einheiten in der Hand des CIO oder bei Einrichtungen, bei denen der CIO nicht aus der eigenen IT stammt oder bei der Realisierung des CIO-Modells als Personen- und nicht Gremienmodell wird statistisch signifikant mehr Arbeitskraft für CIO und Governance-Aufgaben investiert.

Umgekehrt ist die für CIO-Tätigkeiten verfügbare Personalkapazität bei Unterteilung in Gruppe 1 [N1=15] mit weniger als 0,6 VZÄ bzw. Gruppe 2 [N2=13] mit mehr als 0,6 VZÄ Prädiktor für folgende Indikatoren:

- Linienverantwortung des CIO für alle zentrale IT-Abteilungen [M1=0%; M2=54%; StdA=43%, U=45*].
- CIO war zuvor in der IT-Einrichtung der eigenen Hochschule tätig [M1=40%; M2=92%; StdA=48%, U=46,5*].
- Es wurde ein Personenmodell für die CIO-Funktion gewählt [M1=47%; M2=100%; StdA=45%; U=45,5*].

Zusammenfassung: In Einrichtungen, die mehr als 0,6 Stellenanteilen (VZÄ) für CIO-Aufgaben aufwenden, wird häufiger eine Linienverantwortung für alle zentralen IT-Einrichtungen dem CIO übergeben, der CIO stammt seltener aus der eigenen IT-Einrichtung und es wird häufiger ein Personenmodell für die CIO-Funktion gewählt.

Interpretation: Alle drei hier analysierten Faktoren („CIO-Form“, „Rolle vor CIO“ und „CIO ist linienvorgesetzt“) haben bidirektionale Bezüge auf die insgesamt für CIO-Aufgaben zur Verfügung stehende Personalkapazität. Auch hier ist ein Wirkzusammenhang nicht direkt nachweisbar, liegt aber für alle mit Erfahrung in den hier untersuchten komplexen Hochschulgefügen auf der Hand. Eine zentrale Rolle wie den CIO mit zu wenig Zeit (Arbeitskraft) auszustatten muss zwangsläufig dazu führen, dass die Tätigkeiten, die nicht effektiver gestaltet werden können – nämlich die Kommunikation in kritischen Phasen/Aufgaben – zu kurz kommen. Sachaufgaben sind mit guter Organisation und Disziplin stets zu verbessern und (z.B. durch den Einsatz von IT) zu optimieren. Da die hier dargestellte Vernetzung der personellen Kapazität über die CIO-Rolle und die Linienverantwortung mit den IT-Domänen verknüpft ist, liegen die potentiell vorhandenen Wirkzusammenhänge hier konkret auf dem Tisch. Die CIO-Form (Gremien oder Person) selbst spielt als gestaltender Faktor eine untergeordnete Rolle, wogegen die mitgebrachte Erfahrung, wie das CIO-Amt zu gestalten ist, und die direkte Einflussnahme (Linienvorgesetztheit) als vermittelnde Faktoren in den Vordergrund treten.

3 Zusammenfassung

Die meisten Faktoren treten in dieser Analyse sowohl als Indikator als auch als Prädiktor in Erscheinung. Das Alter der jeweiligen Einrichtung ist ausschließlich Indikator einer Reihe von Faktoren. Die Drittmittelquote ist überwiegend Indikator. Die Anzahl der IT-Domänen, für die der CIO Entscheidungskompetenzen besitzt, ist Indikator der ehemaligen Beschäftigungsverhältnisses des CIO. Der Hochschultypus (Universität bzw. Fachhochschulen) ist in der Studie nicht vom Recht zur Promotionsvergabe zu unterscheiden. Beide Faktoren wirken prädiktiv, aber nicht als Indikator. Alle anderen hier beschrieben Faktoren wirken, wenn statistische Relevanz besteht, bidirektional.

Das hier beschriebene Netzwerk von Faktoren erlaubt keinen Nachweis einer Kausalkette in den beschriebenen Zusammenhängen. Es wird vielmehr eine Landschaft vielfältiger Beobachtungen gezeichnet.

Die Empfehlungen aus [He14] deuten nicht auf eine schlichte Umsetzung der „Erfolg versprechenden Prädiktoren“ hin. Mit dem hier vorgestellten Netzwerk aus Faktoren können dennoch jedoch (oder auch bereits realisierte) CIO-Modelle hinterfragt und vorab getestet werden. Der bewusste Umgang mit diesen Faktoren erhöht die Chance, dass ein CIO-Modell zu entscheidenden Vorteilen für eine Hochschule beiträgt. Ob für alle Hochschulen die Konzentration auf erfolgreiche Mitteleinwerbung ein profilierender Faktor ist, hängt selbstverständlich von der strategischen Ausrichtung der jeweiligen Einrichtung ab. Dennoch erlaubt eine Gestaltung einer CIO-Form mit den hier beschriebenen Faktoren zumindest eine erfolgreiche Umsetzung eines IT-Governance-Modells. Die Früchte dieser Umsetzung hängen nicht tief, sondern erfordern eine kontinuierliche Kommunikation und Pflege.

Es besteht für Hochschulen weiterhin die Herausforderung, ein für die jeweilige Institution und handelnden Personen passendes CIO- bzw. IT-Governance Modell zu implementieren.

Literaturverzeichnis

[Bi11] Markus Bick: Zwischen Fachwissen und strategischer Entscheidung - Was muss die Hochschulleitung über IT wissen? CIO/IT-Governance-Modelle in deutschen Hochschulen, in HIS: Forum Hochschule 4|2013, http://www.his-he.de/pdf/pub_fh/fh-201304.pdf, zuletzt abgerufen am 21.03.2014

[BB11] Katrin Börgmann, Markus Bick: IT-Governance in deutschen Hochschulen - eine qualitative Untersuchung. Hochschulmanagement (HM) 6(2):47–53, 2011, http://www.universitaetsverlagwebler.de/inhalte/hm-2-2011.pdf, zuletzt abgerufen am 21.03.2014

[BK10] Werner Bröcker, Frank König: Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme, Empfehlungen der Kommission für IT-Infrastruktur für 2011–2015, Deutsche Forschungsgemeinschaft 2010, http://www.dfg.de/download/pdf/foerderung/programme/wgi/empfehlungen_kfr_2011_2015.pdf, zuletzt abgerufen am 21.03.2014

[FL09] Antonio Martínez Fernández, Faraón Largo Llorens: "An IT-Governance framework for universities in Spain". In: Proceedings of the EUNIS 2009 Conference: June 23rd to 26th 2009, University of Santiago de Compostela, 2009. ISBN 978-84-9887-138-8, http://www.gti4u.es/pdf/an_it_governance_framework_for_universities_in_spain.pdf, Abruf zuletzt am 21.03.2014

[Ga03] Rudolph Gardill: Empfehlungen des ZKI e.V. zur Etablierung eines CIO-Funktionsträgers in Hochschulen, Vorlage für die 13. Mitgliederversammlung 2004 in Würzburg, 2003, https://www.zki.de/fileadmin/zki/Gremien/vorstand/veroeffentlichungen/CIO-Papier.pdf, Abruf nur für Berechtigte.

[GP11] Simone Görl, Johanna Puhl, Manfred Thaller: Empfehlungen für die weitere Entwicklung der Wissenschaftlichen Informationsversorgung des Landes NRW, ISBN: 978-3-8442-0694-4, Köln, 2011

[He08] Markus von der Heyde et al.: IuK-Strukturkonzepte für Hochschulen - Empfehlungen zu Zielsetzung, Aufbau und Inhalt, http://www.zki.de/fileadmin/zki/Publikationen/IuK-SK_Gesamt-V1_0_02.pdf, zuletzt abgerufen am 21.03.2014

[He12] Markus von der Heyde et al.: Treiben oder getrieben werden? - IT-Strategie in Hochschulen", Abschlussbericht des IT-Strategieforums, Jan 2012, http://www.zki.de/fileadmin/zki/Arbeitskreise/IT-SF/webdav/web-public/Publikation/EndredaktionPapier_120130-final.pdf, Abruf zuletzt am 21.03.2014

[He14] Markus von der Heyde: CIOs und IT-Governance an deutschen Hochschulen, In Kürze verfügbar unter: http://www.zki.de/publikationen/

[Hr13] Entschließung der 13. Mitgliederversammlung der HRK am 20. November 2012 in Göttingen: Hochschule im digitalen Zeitalter: Informationskompetenz neu begreifen - Prozesse anders steuern, erschienen in Beiträge zur Hochschulpolitik 1/2013, http://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-10-Publikationsdatenbank/Beitr-2013-01_Informationskompetenz.pdf, zuletzt abgerufen am 21.03.2014

[Is08] ISO/IEC 38500:2008 Corporate Governance of Information Technology", ISO/IEC. 2008, from: www.iso.org

[Kl11] Frank Klapper: Geschäftsprozessmanagement unter dem Fokus des IT-Managements in Prozessorientierte Hochschule - Allgemeine Aspekte zur prozessorientierten Hochschule, DINI 2011, http://dini.de/fileadmin/docs/Prozessorientierte_Hochschule_ 2011.pdf, zuletzt abgerufen am 21.03.2014

[MA01] Joachim Mönkediek, Peter Arends: Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme, Empfehlungen der Kommission für Rechenanlagen für 2001–2005, Deutsche Forschungsgemeinschaft 2001, http://www.dfg.de/download/pdf/foerderung/programme/wgi/wgi_kfr_empf.pdf, Abruf am 21.03.2014

[MA06] Joachim Mönkediek, Peter Arends: Informationsverarbeitung an Hochschulen – Organisation, Dienste und Systeme, Empfehlungen der Kommission für Rechenanlagen für 2006–2010, Deutsche Forschungsgemeinschaft 2006

[Sc10] Gerhard Schwabe: IT-Governance an Universitäten in Deutschland, Schweiz und Österreich, in ZKI-Herbsttagung 2010 – IT-Governance und IT-Infrastruktur-management in europäischer Ausprägung, September 2010, http://www-docs.tu-cottbus.de/ikmz/public/files/Veranstaltungen/Tagungswoche_09_2010/Vortraege/ZKI/02_Schwabe_IT-Governance_VM_Teil_2.pdf, Abruf am 21.03.2014

[WR04] Peter Weill, Jeanne W. Ross: IT Governance: How Top Performers Manage IT Decision Rights for Superior Results". Harvard Business School Press, Boston 2004, ISBN 978-1591392538.

Hochschulrechenzentren als IT-Stadtwerke – Konzeption und kritische Diskussion innovativer IT-Dienstangebote von Hochschulen

Steven Strehl, B.Sc.

ZE Hochschulrechenzentrum
Hochschule für Technik und
Wirtschaft Berlin
steven.strehl@htw-berlin.de

Prof. Dr. Andreas König

School of Management and Law
Zürcher Hochschule für
Angewandte Wissenschaften
andreas.koenig@zhaw.ch

Abstract: Der Aufsatz befasst sich mit der Frage, wie Hochschulrechenzentren neue und innovative Produkt-Portfolios entwickeln und sich damit neue Märkte und Kunden erschließen können. Angestrebt ist eine kritische Diskussion im Kollegium mit Blick auf die hier aufgezeigten Chancen und Potentiale. Es ist hingegen noch nicht Aufgabe des Artikels, Lösungsstrategien zu zeigen, da diese stets konkret, organisations- bzw. kompetenzbezogen sein müssten. Einleitend wird die Diskussion zusammengefasst, die sich mit den Ansprüchen an die Informatikdienste an Hochschulen sowohl in technischer wie auch in wirtschaftlicher und strategischer Sicht befasst. Aktuelle Angebotsstrukturen von Informatikdiensten deutscher Hochschulen werden im zweiten Abschnitt untersucht und verglichen. Eine Analyse exemplarischer Zukunfts- und Trendstudien im dritten Abschnitt zeigt auf, welche neuen Anforderungen auf die Hochschule im Allgemeinen und ihre IT im Besonderen zukommen. Auf diesen Vorarbeiten entwirft der vierte Abschnitt das Modell eines neuen Serviceangebotes der Hochschulinformatik. Die Autoren erörtern dazu, welche Leistungen mit der bereits bestehenden Infrastruktur für hochschulnahe und hochschulexterne Anspruchsgruppen angeboten werden können. Hochschulrechenzentren können zum Anlaufpunkt für Bürger mit Beratungs- und Informationsbedarf werden und dadurch zur geforderten positiven Wahrnehmung und Wirkung der Hochschule über den Campus hinaus beitragen. Die neuen IT-Dienste dienen als Inkubatoren für Spin-Offs, studentische Start-Ups und die Verbreitung von Open-Source-Software. Im fünften Abschnitt wird dieser Entwurf kritisch diskutiert mit Blick auf seine mögliche Bedingungen und Wirkungen. Dabei werden insbesondere Aspekte des Managements, des Personalwesens und der Abgrenzung der Hochschule nach außen angesprochen. Zum Schluss wird das Szenario mit Blick auf seine Chancen und Erträge kurz bilanziert.

1 Informatikdienste an der Schnittstelle von Hochschulen – eine andauernde Debatte

Dieser Abschnitt fasst kursorisch zusammen, welchen Anforderungen sich die Informatikdienste der Hochschulen gegenüber sehen und sahen. Die deutsche Diskussion zum Thema im engeren Sinne ist sehr begrenzt (vgl. [BBS07], [Le12], [RSL10], [Vo03]) und meist entweder auf die Lehre oder Fragen der Hochschulentwicklung allgemein (vgl. [EK12], [Mu05]) bezogen. Im englischen Sprachraum findet hingegen eine breite Debatte statt, wie noch zu zeigen ist. Insofern versucht unser Beitrag, eine Diskussion anzuregen, aber auch in den deutschen Raum zu übertragen.

Wir unterscheiden Funktionen der Informatikdienste in hochschulinterne und hochschulexterne. Bei den hochschulinternen Funktionen ist sicher die Lehre im Fokus der Diskussion. Köhler und Neumann [KN11] versammeln Beiträge zur Diskussion der Öffnung der Hochschullehre und -forschung durch digitale Medien. Frey [Fr13] geht speziell auf die sozialen Medien ein, die die Hochschullehre stark beeinflussen. Auch Grell und Rau [GR11] behandeln Probleme und Potentiale sozialer Medien für Hochschulen. Die Wechselwirkung neuer Medien und auch neuer Anforderungen der Studenten mit dem Hochschulunterricht thematisieren Reinmann et al. [RES13]. Brugger [Br13] fokussiert, wie sich die digitalen Arbeitsgewohnheiten der Studenten verändert haben und reflektiert dann, welche Auswirkungen dies auf die Informatikdienste des Hochschulsektors haben wird. Pollara et al. [PB11] fragen ebenfalls nach den Erwartungen und Arbeitsgewohnheiten der Studenten speziell mit Fokus auf den Bereich des mobilen Lernens und Arbeitens.

Auch Entwicklungen der Hochschulpolitik und -struktur sind maßgeblich für Anpassungen der Informatikdienste. So analysiert Kühl [Ku12] die Wirkung bürokratischer Faktoren, allen voran die wachsende Notwendigkeit, die in Leistungspunkten ausgedrückten Veranstaltungen, Prüfungen und Module stimmig miteinander kombinieren zu können. Arbeiten, die speziell die Anpassung der Informatik für neue und erweiterte IT-Dienste vorschlagen, legen beispielsweise Du et al. [Du12a] sowie [Du12b] vor. Sie wollen auf dem Campus bestehende Infrastrukturen vernetzen und so erweitern, dass Forschung und Lehre, aber auch private Nutzungen der Hochschulangehörigen profitieren. Rekik et al. [Re06] untersuchen, wie webgestützte Kollaborationsprozesse speziell für Laborarbeiten von Ingenieuren und gemeinsame Forschungsnutzung von Laboren durch eine .Net-Umgebung verbessert werden. Wang et al. [Wa12] entwickeln schließlich ein Evaluationssystem für die Qualität der Informatikdienste der Hochschule. Sowohl akademische wie nicht-akademische studentische Dienste sollen in Hinblick auf Verlässlichkeit, Verfügbarkeit, Sicherheit und Angemessenheit bewertet und optimiert werden können.

Prozessoptimierung ist das Thema der Lerntechnologieabteilung der Universität Basel [Le12]. Ihr Projekt „IT-Service-Integration in Studium und Lehre (ITSI)“ soll die „...vorhandenen Informations- und Kommunikationstechnologien sowie die damit verbundenen Dienstleistungen zur Unterstützung von Lehre und Studium besser aufeinander abstimmen und benutzerorientiert ausrichten“.

Weiter gewinnen betriebswirtschaftliche Aspekte in der Hochschulinformatik eine wachsende Bedeutung. Beispiele für die Ökonomisierung der Hochschulinformatik nach innen gibt Voros [Vo12]. Er beschreibt, wie die UC Berkeley ein Labor zur Herstellung von Computerhardware erfolgreich betrieben hat. Für den Erfolg war maßgeblich, dass Ressourcen hochschulweit in einem Pool geteilt und die Kosten transparent und fair ermittelt und verrechnet wurden. Wie bei einer kollaborativen Nutzung geteilter Infrastrukturen eine faire und transparente Kostenallokation für die Nutzer mit Hilfe von realistischen Schätzungen aufgrund von Ressourcenprofilen erreicht werden kann, hat Brandl aufgezeigt [BBS07]. Von der Markt- und Kundenorientierung her denken Rudakova et al. [RSL10] die Anforderungen an das IT-Service-Management. Informatikabteilungen können betriebswirtschaftlich nicht nur als Cost Center betrieben, sondern zu einem tatsächlichen Wettbewerbsfaktor der Organisation gemacht werden, wie bereits 2003 von Eltz am Beispiel des entsprechenden Change-Projektes im Wacker-Konzerns aufgezeigt hat [Vo13].

Vor allem amerikanische Hochschulen haben traditionell Erfahrungen mit der Öffnung für externe Publika, indem sie auch die Hochschulinformatik für Unternehmenszwecke geöffnet haben. Rindos [Ri12] stellt der Zusammenarbeit von IBM mit amerikanischen Hochschulen gute Noten für das gemeinsam entwickelte WebSphere-Produktportfolio aus. Produktinnovation, Testing, proof-of-concepts und Showcases, aber auch Talententwicklung und Rekrutierung im HR-Bereich sowie verbesserte Verkaufsmöglichkeiten und Sichtbarkeit im Marketing erkennt er als positive Wirkungen der Kooperation. Siebenhofer [Si09] urteilt ähnlich positiv über die Kooperation von Siemens IT Solutions and Services mit chinesischen Hochschulen. In zahlreichen Praktika seien nicht nur Fachkräfte ausgebildet, sondern auch Impulse für curriculare Entwicklung gegeben und Projekte gefördert worden.

Europäische Hochschulen praktizieren diese Öffnung im Bereich der Lehre und Beratung. Ein Beispiel der Fernuniversität Hagen [Le08] steht stellvertretend dafür. In diesem Fall wird soziologisches Wissen an Laien weitergegeben, die außerhalb des Wissenschaftssystems stehen. In die gleiche Richtung zeigen Initiativen wie die Bürgeruniversität und Angebote an Kinder oder Senioren. Robinson et al. [Ro14] führen die Öffnung der Hochschule zur Gesellschaft an der Universität von Durham vor. Forscher engagieren sich aktiv in Debatten um die Einführung von und den Umgang mit neuen Technologien und befähigen Bürger, in einem verantwortungsvollen und demokratischen Diskurs rationale Entscheidungen zu treffen.

Im angelsächsischen Sprachraum wird schließlich eine Debatte um die „Community Informatics“ (nachfolgend „CI“) geführt, die ebenfalls in unseren Kontext gehört. Die zahlreichen Umsetzungsformen belegen beispielsweise Gurstein [Gu00], Bradley [Br06], Stillman [St09] oder Stillman et al. [SJF09]. Stillman versteht CI als das Phänomen der Aneignung von Informationstechnologien durch lokale Gemeinden und problematisiert dabei die ambivalente Rolle des Staates, der einerseits zwar Infrastrukturen und Dienste anbietet, andererseits aber so auch die Autonomie sowohl der Gemeinden wie auch der Bürger kompromittieren könnte. Es ist überflüssig zu betonen, dass genau die Universitäten an dieser Stelle eine hilfreiche Brückenfunktion einnehmen und die Kontroverse entschärfen könnten.

Konkrete und bürger- wie gemeindenahe Projekte der CI beschreiben weiter Ezenkwu et al. [EOK13]. Sie veranschaulichen, wie mit Hilfe von geographischen Netzwerken Gemeinden in Nigeria z.B. Kriminalität besser bekämpfen können. Verwandte Beispiele für CI in der Regionalentwicklung liefern Marshall et al. [MTY04]. Die Autoren wollen Führungskräften, Forschern, Studenten und Gemeindearbeitern erfolgreiche Strategien für CI in der Regionalentwicklung an die Hand geben. Attwood et al. [At13] schildern den Nutzen sogenannter „Telecenters“ in Entwicklungs- oder Schwellenländern (hier RSA) und wie sie zur Verbesserung des Lebensstandards der Gemeinde messbar beitragen. Dass CI keinesfalls nur auf Entwicklungsländer beschränkt ist, zeigt Goodwin [Go07] am Beispiel des Birmingham City Council / UK. Dessen Projekt, den Bürgern wertvolle Internetdienste anzubieten und sich neu als Partner zu positionieren, bilanziert er als sehr erfolgreich, kritisiert aber zugleich den Legitimitätsverlust, der aus dem Versuch der Regionalregierung resultiert, Kontrolle und Management dieser Dienste zu usurpieren.

Die hier zusammengefassten Diskussionen zeigen, dass Hochschulen bereits heute ein breites Portfolio von Informatikdiensten aufweisen, dieses immer wieder auf neue Anforderungen hin rekonfigurieren und auch bereits an zahlreichen Stellen in die Gesellschaft hineintragen. Im nächsten Abschnitt wird nun zu untersuchen sein, welche konkreten Dienste die HRZ heute anbieten, um danach weiter zu reflektieren, welche Anwendungsmöglichkeiten auf diesen Grundlagen möglich erscheinen, um die HRZ als „Informatik-Stadtwerke“ im Sinne einer hochschulgetragenen CI zu positionieren.

2 Informatikdienste deutscher Hochschulen – das aktuelle Portfolio

Die Hochschulrechenzentren (im weiteren: „HRZ“) sind zunächst für den zuverlässigen Betrieb einer sich ständig weiterentwickelnden IT-Infrastruktur am Hochschulstandort zuständig. Im Hochschulalltag erbringen sie schwerpunktmäßig IT-Dienstleistungen für das Campus-Management (Student-Life-Cycle-Management und Hochschulverwaltung), sowie für Forschung und Lehre.

Neben einem Basisportfolio variiert das Angebotsspektrum je nach Fächerschwerpunkt und Größe der Hochschule. Zudem operieren HRZ nicht immer als geschlossene Organisationseinheit: je nach Größe und Organisationsstruktur übernehmen Fakultäten, Fachbereiche und bei der Verwaltung angesiedelte Abteilungen dediziert Teilaufgaben.

Rudakova et al. [RSL10] definieren die „Prozessunterstützung bei der Schaffung von Forschungsergebnissen“ als Ziel der HRZ an Universitäten. Wir sehen die Aufgaben von HRZ an Universitäten und Fachhochschulen in den drei Kernleistungsbereichen Campus-Management, Lehre und Forschung, sowie Infrastruktur. Das Rückgrat der Hochschul-IT bilden die Datenzentren. Deren Mitarbeiter stellen die Zuverlässigkeit der Systeme, die Hochverfügbarkeit der Dienste und die Datensicherheit in Verwaltung und Student-Life-Cycle sicher.

Die stetig steigenden Studierendenzahlen erfordern einen fortwährenden Ausbau der IT-Kapazitäten an Hochschulen, um eine qualitativ gleichbleibenden Zuverlässigkeit zu

gewährleisten. Gleichzeitig verändert sich die Beziehung zwischen Universitäten und Studierenden: Bildung wird als Dienstleistungsangebot wahrgenommen, und die Qualität dieses Angebotes entscheidet über die Attraktivität eines jeden Standortes.

Um den Herausforderungen einer wachsenden Hochschul-IT gerecht zu werden, bedarf es hochqualifizierter Mitarbeiter. Personale und technische Einrichtungen der Hochschulen sehen sich einer großen Herausforderung gegenüber: sie müssen einen attraktiven Arbeitsplatz bieten, was im Vergleich zur freien Wirtschaft schwierig ist. Die Gehälter im öffentlichen Dienst sind niedriger, die Aufstiegs- und Entwicklungschancen sind geringer. Hochschulen müssen also punkten, wo die Privatwirtschaft es nicht oder nur schwerlich kann. Möglichkeiten dazu sehen wir in der Identifikation, vor allem in der gesellschaftlichen Implikation, der Sichtbarkeit und Verantwortung, die Hochschulen und ihre Angehörigen in der Gesellschaft einnehmen.

Hochschulrechenzentren sind Einrichtungen der Hochverfügbarkeit. Die Heterogenität der Angebote sorgt für großen Bedarf an qualifizierten Mitarbeitern mit Spezialwissen. Diese haben einerseits eine hohe Arbeitsbelastung und aufgrund der beschränkten öffentlichen Mittel oft keine Vertreter. Tarifverträge geben zudem wenig Raum für Arbeit am Wochenende. Stabilität und Erreichbarkeit der Systeme müssen also zunehmend automatisiert gewährleistet werden, da an Wochenenden, Feiertagen und in der Nacht Personaleinsätze nicht möglich sind. HRZ werden als Organisationen des Öffentlichen Dienstes zudem durch Bürokratie, lange Wartezeiten und komplexe Entscheidungsfindungs- und Bewilligungsprozesse eingeschränkt. Unser Szenario zeigt hier einiges Innovationspotenzial auf.

3 Informatikdienste und die zukünftigen Anforderungen

Die European Internet Foundation hat im März 2014 eine neue Langzeitprognose technologischer, wirtschaftlicher und soziopolitischer Trends im digitalen Bereich vorgelegt. Die Befunde fasst die Studie mit der Trendaussage zusammen: „Von der Massenkollaboration zur Wissensgesellschaft“. Die erwartete umfangreiche Digitalisierung und ihre Folgen werden sich, so die Prognosen, in Ökonomie und Gesellschaft tiefgreifend auswirken. Die Bedeutung der Digital Divide wird größer, aber nicht Besitz und auch nicht Verfügbarkeit werden 2030 noch die Trennung verursachen, sondern die Fähigkeit, digitale Medien sozial und ökonomisch einzusetzen. Daraus ist auf einen umfassenden Schulungs- und Beratungsbedarf zu schließen. Die Studie gibt zahlreiche Anhaltspunkte, dass neue Formen von sozialen und ökonomischen Beziehungen neue Formate an Diensten benötigen werden. Das Netzwerkunternehmen Cisco [Ci11] fokussiert die Wirtschaft, genauer den Arbeitsplatz von morgen, und erhebt dazu die Erwartungen und Ansprüche der nächsten Generation von Mitarbeitern bei 3000 Personen zwischen 18 und 29 Jahren in 14 Ländern. Sie gibt ein recht gutes Bild von der Digitalisierung des Arbeitsplatzes, die die kommende Generation erwartet.

Die jüngste Studie des Frankfurter Zukunftsinstitutes über Lebensstile [Hu14] erklärt, wie die Fragmentierung der Lebenswelten und -stile angesichts einer Explosion der Möglichkeiten weiter zugenommen hat. Sie kondensieren 10 Lebensstile, deren

Heterogenität Bezeichnungen belegen wie „Creativiteens: teenager unterwegs vom Mitmach-Web ins Mitmach-Leben“, „Business-Freestyler: Jung, gebildet, ehrgeizig, erfolgreich - und Lebenskünstler“ oder „Silverpreneure: Vom Beruf zur Berufung in der Anti-Ruhestandsgesellschaft“. Die Titel verweisen auf eine breite Palette von Situationen und Visionen, mit denen künftige Kunden ein „IT-Stadtwerk“ ansprechen könnten.

Der auf die E-Learning-Branche bezogene Trendbericht des MMB-Institutes [Mi13] nennt bei den mittelfristigen Technologietrends das Mobile Learning, Social Media bzw. Social Learning, MOOCs sowie Cloud Computing und Software as a Service. Der aktuelle Bericht aus dem gleichen Haus [Mi14] fokussiert lediglich einen Technologietrend, nämlich das adaptive Lernen bzw. intelligente tutorielle Systeme. Hierin deckt sich die Prognose mit der der European Internet Foundation, die auch auf die wachsende Bedeutung der Softwareentwicklung für das Lernen hinweist.

Die großen Zukunftstrends betreffen auch die Hochschulen. Bereits heute ist absehbar, dass die Globalisierung auf deutsche Hochschulen deutliche Auswirkungen hat. Gruß [Gr13] fasst wichtige Aspekte dieser Diskussion zusammen. Einerseits werden die Universitäten als Organisationen bezüglich Funktionen wie Rekrutierung, Fachkräfteausbildung, Marketing und Employer Branding neue Wege beschreiten müssen, um wettbewerbsfähig zu bleiben, ist doch das Umfeld der Hochschulen heute hoch kompetitiv auf globalem Niveau. Die Demographie erzwingt neue Rekrutierungsmethoden und eine Profilschärfung. Das Image bzw. der Employer Brand ist eine wichtige Größe auch für Hochschulen geworden, die ebenso wie Unternehmen im Wettbewerb um die besten Köpfe für Forschung und Lehre stehen.

Der eher technologieorientierte Horizon-Report [NE14] ermittelt für Hochschulen eine große Wirkung von Softwareentwicklungen und einen weiter wachsenden Qualifizierungs- und Kompetenzbildungsbedarf sowie neue Lehrmodelle. Auch hier prognostizieren die Experten eine deutliche Öffnung der Hochschulen bei gleichzeitigem Bemühen um die Erhaltung der Relevanz von Hochschulbildung. Diese auf Bildungstechnologien bezogenen Zukunftsstudien ergeben im übrigen auch zu älteren Studien ein kongruentes Bild (vgl. z.B. [Mu05], [Pu05] [Mi06]).

Auf strategische und kontextuelle Entwicklungen beziehen sich Staley und Trinkle [ST11]. Die Hochschuldozenten würden globaler und die Studenten mobiler werden. Die Hochschule werde „unsichtbarer“, womit Bezug genommen wird auf die Entgrenzung der Hochschule. Weitere wichtige Einflussfaktoren seien die neuen demographischen Merkmale der Studentenschaft und der wachsende Druck, einen klaren Mehrwert im Studium zu zeigen. Neue Lernangebote müssten auch mittlere Qualifizierungsniveaus ansprechen und lebenslange Lernpartnerschaften angestrebt werden. Die Aussagen von Quitney et al. [QBR12] decken sich mit diesen Prognosen weitgehend.

Elkana und Klöpper [EK12] diskutieren die Auswirkungen gängiger neoliberaler Wirtschaftsauffassung auf das (künftige) Handeln der Hochschulen und empfehlen eine humanistisch inspirierte Reorientierung auf die Einheit von Lehre, Forschung und Gesellschaft. In den Abschnitten über Demokratie und Bildung sowie über die digitale

Universität argumentieren sie dabei durchaus auch in einem Sinne, der unsere Position legitimiert.

4 Informatikdienste der Hochschule auf dem Weg zum „IT-Stadtwerk“

Dieser Abschnitt skizziert grob eine Neukonzeption von HRZ als „IT-Stadtwerke“. Für die Entwicklung dieses Szenarios motiviert uns auch eine derzeit vehement geführte Diskussion, die die Durchdringung praktisch aller Bereiche von Wirtschaft und Gesellschaft durch große Provider wie Google, Facebook, Yahoo und andere thematisiert. Exemplarisch belegen dies die Kritiken von Dams et al. [Da14] (ebenso auch [We09] und [KR12]) bezüglich der kritischen Folgen für Privatsphäre und Datenschutz der Verbraucher und für die Wettbewerbsbedingungen der (deutschen) Wirtschaft.

Auf dem Weg vom HRZ zum IT-Stadtwerk mit „digitalem Versorgungsauftrag“ sehen wir vorrangig zwei Bereiche für das Leistungsportfolio, und zwar Innovation und Community Informatics.

Der erste Bereich, Innovation, umfasst Dienstleistungen für hochschulinterne und -nahe Anspruchsgruppen. Spin-Offs, studentische Start-Ups und Initiativen profitieren von der Hochschule als alternativem Inkubator. Im Lichte unserer Diskussion sind hier „As a Service“-Dienstleitungen denkbar, die bereits heute im Angebotsportfolio einiger HRZ enthalten sind. Diese können unbürokratischer und möglichst automatisiert zur Verfügung gestellt werden. Projektmittel mit einem niedrigen dreistelligen Betrag (wie es bereits auf Studiengangsebene möglich ist) sollen kurzfristig Handlungsspielraum geben. Räume für professionelle Meetings und Präsentation sowie professionelle Coachingangebote durch Lehrbeauftragte decken den kurzfristigen Bedarf für die Professionalisierung von Studienprojekten ab. Zu klären ist hierbei (auch auf juristischer Ebene), inwiefern die unterstützten Studierenden zu Gegenleistungen verpflichtet werden können. Bestehende Start-Up-Kompetenzzentren der Hochschulen können Innovationen weiter fördern. Synergien sehen wir darin, dass Studierenden ein breites Leistungsportfolio rund um die Gründung eines Unternehmens geboten wird.

Forschungsprojekte in den Human- und Geisteswissenschaften leiden oft unter mangelndem Zugang zu IT-Kompetenz. Die „IT-Stadtwerke“ können Forschungsprojekten unbürokratisch Informatikerleistungen bereitstellten. Projekte mit breiter empirischer Basis können bspw. mit „Data Mining“-Expertise arbeiten, ohne eine volle Stelle dafür in Ihrem Projekt einplanen zu müssen. Eine solche organisationale Struktur begünstigt das Einwerben von Drittmitteln für inter- und transdisziplinäre Projekte und fördert die Attraktivität eines Forschungsstandortes.

Der zweite Bereich, Community Informatics, richtet sich an hochschulexterne Anspruchsgruppen. Der Grundgedanke ist eine Art Bürgerbüro, das als Anlaufstelle für Fragen zur digitalen Gesellschaft dient und den niederschwelligen Einstieg in die Verwendung von Open-Source-Programmen ermöglicht. Die NSA-Affäre, der

„Heartbleed-Bug“ und massenhaft gestohlene Passwörter haben einen Vertrauensverlust in die digitale Wirtschaft, aber auch in öffentliche Organisationen und die Regierung bewirkt. Eine bürgernahe physische Anlaufstelle würde helfen, Internettechnologien begreifbar zu machen.

Auf die Bedürfnisse und Fähigkeiten technischer Laien ausgerichtete Informationsangebote stärken die gesellschaftspolitische Wahrnehmung der Hochschule und kommunizieren deren Verantwortung über den akademischen Betrieb hinaus. Gleichzeitig intensiviert die Hochschule so ihren Austausch mit der Gesellschaft und wird auch für Nichtakademiker als Lernort attraktiv. „Kryptoparties“ beispielsweise stellen einen niederschwelligen Zugang zum Verständnis von Datensicherheit und der praktischen Umsetzung von Datenschutzmaßnahmen mit Open-Source-Software wie beispielsweise mit der E-Mail-Verschlüsselung mit PGP dar.

Erste Schritte in die Richtung Community Informatics macht die TU Berlin. Heise.de [He13] berichtet, die TU Berlin betreibe eine der größten Installationen von ownCloud weltweit. Auch die ETH Zürich setzt die Software unter dem Namen „polybox“ ein. ownCloud ist eine Open-Source Alternative für Clouddienste, wie sie auch von Google, Apple, Microsoft und DropBox angeboten werden. Der Funktionsumfang reicht von Cloudspeicher (ähnlich Google Drive, DropBox, SkyDrive) bis hin zu Adressbüchern (Google Contacts) und Kalendern (Google Calendar), die man über eine Weboberfläche erreicht. Die Dateien, Kalender und Adressdaten liegen in diesem Fall bei der TU Berlin und nicht mehr bei einem Anbieter in den USA. Über einen Desktop-Client, CalDAV-, WebDAV- und CardDAV-Schnittstellen können verschiedene Geräten Dateien, Kontakte und Termine synchronisieren und teilen (genau wie mit der Produktfamilie von Google). Studierende werden so unabhängig von proprietären Diensten, deren Geschäftsbedingungen die Datensicherheit, den Datenschutz und die Urheberrechte kompromittieren.

Im Sinne der Community Informatics kann das „IT-Stadtwerk“ darüber hinaus als zentraler Anbieter für eine kommunale Open-Data-Plattform fungieren und über hochschulrelevante Themen hinaus Schnittstellen (APIs) anbieten. Damit trüge es zur Anwendung des §12EGovG (E-Government-Gesetz) bei, welches die Bereitstellung von maschinenlesbaren Datenbeständen fordert.

Bruckner et al. [Br14] zeigen ein weiteres potentielles Nutzungsfeld im Sinne einer Communiy Informatics, nämlich die zunehmende Informatisierung von Gebäuden, auch und gerade im privaten Bereich. Die Autoren beschreiben, wie Computernetzwerke aufgebaut, aber auch Gebäude- und Wohnfunktionen informatisiert werden. Dies können Funktionen sein, die von einem potentiellem Informatik-Stadtwerk geschult, unterstützt oder mitgetragen werden.

Das „IT-Stadtwerk“ richtet das HRZ auf Bürgerdienste und gesamtgesellschaftliches Engagement hin aus und eröffnet so den Bürgern einen niederschwelligen Zugang zu Informatik-Kompetenz und den Hochschulen neue Möglichkeiten der Finanzierung und Positionierung in der Öffentlichkeit.

Dabei muss ein „IT-Stadtwerk“ nicht an einem Ort allein bestehen. HRZ einer Region können sich zusammenschließen und dem damit geschaffenen IT-Stadtwerksverbund gegenseitig ihre Services anbieten. So können sie ihre Zuständigkeits- und Aufgabenspektren erweitern und das zu bewältigende Arbeitspensum aufteilen. Vorstellbar ist, dass ein Mitglied beispielsweise Moodle als Service und den Nutzersupport dazu anbietet, ein anderes betreibt ownCloud-Server. Die Nutzung dieser Services durch hochschulinterne und -externe Gruppen im Verbund kann über ein Single-Sign-On-Verfahren realisiert werden. Die Heimathochschulen können, wie auch schon bei DFN und eduroam, als Identity Provider fungieren.

5 Schluss: Kritische Perspektiven auf das Szenario

Das Konzept der „IT-Stadtwerke“ beinhaltet eine strategische Neuorientierung der HRZ auf Innovation und Community Informatics. In der Funktionsbezeichnung Community Informatics ist bereits angesprochen, dass solche Einrichtungen den großen (US-amerikanischen) Providern nicht die Stirn bieten können. Aber sie können sicher zur Sensibilisierung und Kompetenzbildung ihrer Umgesellschaft beitragen. Klar ist weiterhin, dass eine solche Ausrichtung die bestehende Heterogenität und Komplexität in Struktur und Angebot sowie häufig existierende Überlasten von HRZ weiter herausfordert. Daher ist denkbar, dass die IT-Stadtwerke als zusätzliche Abteilung bestehender HRZ geführt werden, statt dass diese insgesamt umgestaltet würden. Essentielle Vorbedingung für deren Wirken wird dann sicher die Abklärung rechtlicher Rahmenbedingungen sein.

Eine solche Umstrukturierung kann und soll hier nicht mehr konkret ausgearbeitet werden. Im Bedarfsfall müsste dies gemeinsam mit Partnern etwa eines Pilotprojektes in Angriff genommen werden. Die im Artikel angesprochenen Beobachtungen verweisen zuerst auf eine veränderte Kommunikation und Führung im Betrieb. Eine innovative IT-Strategie braucht Maßnahmen zur Neuregelung interner Kommunikation und Entscheidung, bei der Führungskräfteentwicklung und der Ausprägung einer Kundenorientierung (vgl. [Vo03]). Rechenzentren werden zunehmend auch gefordert sein, das Unvorhersehbare zu managen. Ihr Management muss daher nicht nur betriebswirtschaftlich und kundenorientiert, sondern auch veränderungsfähig sein (vgl. [De12]). Weick und Sutcliffe [WS10] geben Ratschläge für die Führung von Organisationen, die – wie auch die HRZ – eine hohe Verlässlichkeit sicherstellen müssen. Die zitierte Lebensstilstudie [Hu14] lässt weiter vermuten, dass rasch wechselnde und auch konfligierende Erwartungen der Mitarbeiter, Kunden und Partner hohe Ansprüche an das Management stellen werden.

Mit neuen Produkten, Prozessen, Kunden, Partnern und Mitarbeitern werden die Struktur, aber auch die Grenzziehung der Organisation neu zu definieren sein. Jung [Ju10] problematisiert, wie viel oder wenig Grenze eine Organisation verträgt und braucht. Er empfiehlt, gezielt und variantenreich die Außenbeziehungen zu gestalten und die Öffnung der eigenen Grenzen aktiv zu gestalten. Organisationale Grenzen werden auch neue Beschäftigungsformen bedeuten. Mayrhofer et al. [MM02] haben die neuen Selbstständigen in Zusammenhang mit der organisationalen Grenzziehung der

Organisation gebracht. Die Frage prekärer Beschäftigungsverhältnisse wird auch hier Thema werden.

Die große Komplexität des Prozesses lässt eine Umsetzung frühestens in mittlerer Sicht, also in einem Zeitraum von 3-5 Jahren erwarten. Gesamtgesellschaftliche Entwicklungen bspw. der Digitalisierung und Automatisierung lassen jedoch erwarten, dass der Handlungsdruck zunimmt und bereits vorher strategische Initiativen wie diese etwa in Prototypen umgesetzt und evaluiert werden.

Literaturverzeichnis

[At13] Attwood, H., Diga, K., Braathen, E., May, J., 2013. Telecentre functionality in South Africa: re-enabling the community ICT access environment. J. Community Inform. 9, 1–10. doi:http://ci-journal.net/index.php/ciej/article/view/970/1060

[Br06] Bradley, G., 2006. Social and community informatics: humans on the net. Routledge, Abingdon, Oxon ; New York.

[BBS07] Brandl, R., Bichler, M., Ströbel, M., 2007. Cost accounting for shared IT infrastructures: Estimating resource utilization in distributed IT architectures. WIRTSCHAFTSINFORMATIK 49, 83–94. doi:10.1007/s11576-007-0030-9

[Br14] Bruckner, D., Dillon, T., Hu, S., Palensky, P., Wei, T., 2014. Guest Editorial Special Section on Building Automation, Smart Homes, and Communities. IEEE Trans. Ind. Inform. 10, 676–679. doi:10.1109/TII.2013.2281939

[Br13] Brugger, R., 2013. A digital shift. Switch J. April 2013, 21.

[Ci11] Cisco, 2011. 2011 Cisco Connected World Technology Report.

[Da14] Dams, J., Ettel, A., Gassmann, M., Heuzeroth, T., Jost, S., Kunz, A., Seibel, K., Tauber, A., 2014. Digitalisierung - Google greift ins Herz der deutschen Wirtschaft ein. Hambg. Abendblatt.

[De12] Dell, C., 2012. Die improvisierende Organisation. Management nach dem Ende der Planbarkeit. Transcript Verlag, Bielefeld.

[Du12a] Du, Z., Fu, X., Zhao, C., Liu, T., 2012a. Multi-Domain Cloud Social Network Service Platform Supporting Online Collaborations on Campus, in: Proceedings of IEEE CCIS2012. Presented at the Proceedings of IEEE CCIS 2012, IEEE, pp. 287–291.

[Du12b] Du, Z., Fu, X., Zhao, C., Liu, T., 2012b. University campus social network system for knowledge sharing, in: Systems and Informatics (ICSAI), 2012 International Conference on. Presented at the Systems and Informatics (ICSAI), 2012 International Conference on, IEEE, pp. 2505–2508. doi:978-1-4673-0199-2/12

[EK12] Elkana, Y., Klöpper, H., 2012. Die Universität im 21. Jahrhundert: für eine neue Einheit von Lehre, Forschung und Gesellschaft. Edition Körber-Stiftung, Hamburg.

[EOK13] Ezenkwu, C.P., Ozuomba, S., Kalu, C., 2013. Strategies for improving community policing in Nigeria through Community Informatics Social Network, in: Emerging & Sustainable Technologies for Power & ICT in a Developing Society (NIGERCON), 2013 IEEE International Conference on. Presented at the Emerging & Sustainable Technologies for Power & ICT in a Developing Society (NIGERCON), 2013 IEEE International Conference on, IEEE, Owerri, pp. 163–168. doi:10.1109/NIGERCON.2013.6715650

[Fr13] Frey, J.-C., 2013. Social Media an Hochschulen, Internet-Technologie und Gesellschaft. Book on Demand.

[Go07] Goodwin, I., 2007. COMMUNITY INFORMATICS AND THE LOCAL STATE IN THE UK: Facilitating or assimilating an agenda for change? Inf. Commun. Soc. 10, 194–218. doi:10.1080/13691180701307446

[GR11] Grell, P., Rau, F., 2011. Partizipationslücken – Social Software in der Hochschullehre. Medienpädagogik. Zeitschrift für Theorie und Praxis der Medienbildung 2011, 1–23.

[Gr13] Gruss, P., 2013. Der Wohlfahrtsstaat braucht Grundlagenforschung... FAZ Frankf. Allg. Ztg. 20130426 7.

[Gu00] Gurstein, M. (Ed.), 2000. Community informatics: enabling communities with information and communications technologies. Idea Group Pub, Hershey, PA.

[He13] Riesen-Wolke an der TU Berlin. News-Meldung von heise.de vom 08.05.2013. http://www.heise.de/ix/meldung/Riesen-Wolke-an-der-TU-Berlin-1859537.html - besucht 12.05.2014.

[Hu14] Huber, T., Rauch, C., Schuldt, C., Seitz, J., Steinle, A., 2014. Lebensstile für Morgen. Das neue Modell für Gesellschaft, Marketing und Konsum. Zukunftsinstitut, Frankfurt.

[Ju10] Jung, D., 2010. Grenzenmanagement und Organisationsentwicklung. OrganisationsEntwicklung 4, 41–47.

[KN11] Köhler, T., Neumann, J. (Eds.), 2011. Wissensgemeinschaften digitale Medien - Öffnung und Offenheit in Forschung und Lehre. Waxmann, Münster; New York, NY; München; Berlin.

[Ku12] Kühl, S., 2012. Der Sudoku-Effekt. Hochschulen im Teufelskreis der Bürokratie. Eine Streitschrift. Transcript Verlag, Bielefeld.

[KR12] Kurz, C., Rieger, F., 2012. Die Datenfresser wie Internetfirmen und Staat sich unsere persönlichen Daten einverleiben und wie wir die Kontrolle darüber zurückerlangen. Fischer-Taschenbuch-Verl., Frankfurt, M.

[Le12] LearnTechNet, 2012. Projekt IT-Service Integration in Studium und Lehre (ITSI). Abschlussbericht LTN-. LTN-BBiT (LearnTechNet – Bereich Bildungstechnologien) Universität Basel – Vizerektorat Lehre, Basel.

[Le08] Lengfeld, H., 2008. Sociology goes public. About laymen, a citizen's university and a chair in sociology. Soziologie 37, 389–406.

[MTY04] Marshall, S., Taylor, W., Yu, X.H. (Eds.), 2004. Using community informatics to transform regions. Idea Group Pub, Hershey, PA.

[MM02] Mayrhofer, W., Meyer, M., 2002. "No more shall we part?" Neue Selbständige und neue Formen der Koppelung zwischen Organisationen und ihrem Personal. Z. Für Pers. 16, 599–615.

[Mi06] Michel, L.P., 2006. Digitales Lernen, Forschung - Praxis - Märkte. Norderstedt.

[Mi13] Michel, L.P., 2013. MMB-Trendmonitor I/2013. Weiterbildung und Digitales Lernen heute und in drei Jahren: Präsenzlernen wird virtuell. Ergebnisse der Trendstudie MMB Learning Delphi 2013.

[Mi14] Michel, L.P., 2014. MMB-Trendmonitor I/2014. Wenn der digitale Lernassistent Sie an die Hand nimmt Zukunftstrend Adaptives Lernen – ein Überblick.

[Mu05] Multimedia Kontor Hamburg, M., 2005. Hochschulen im Digitalen Wandel. Chancen, Trends, Praxis. Multimedia Kontor Hamburg, Hamburg.

[NE14] New Media Consortium, EDUCAUSE Learning Initiative, Johnson, L., Adams Becker, S., Estrada, V., Freeman, A., 2014. Horizon Report 2014. Higher Education Edition. Deutsche Ausgabe. New Media Consortium, Austin, Texas.

[PB11] Pollara, P., Broussard, K.K., 2011. Student Perceptions of Mobile Learning: A Review of Current Research, in: Education, S. for I.T.& T. (Ed.), Proceedings of Society for Information Technology & Teacher Education International Conference. AACE, Chesapeake, VA, pp. 1643–1650.

[Pu05] Pulichino, J., 2005. Current Trends in e-Learning Research Report. The eLearning Guild.

[QBR12] Quitney Anderson, J., Boyles, J.L., Rainie, L., 2012. The future impact of the Internet on higher education. Pew Research Center, Washington D.C.

[RES13] Reinmann, G., Ebner, M., Schön, S. (Eds.), 2013. Hochschuldidaktik im Zeichen von Heterogenität und Vielfalt. Doppelfestschrift für Peter Baumgartner und Rolf Schulmeister. Books on Demand GmbH, Norderstedt.

[Re06] Rekik, Y., Gillet, D., Nguyen-Ngoc, A.V., Guillaume-Gentil, T., 2006. Framework for Sustaining Collaboration in Laboratory-Oriented Communities of Practice, in: Information Technology Based Higher Education and Training, 2006. ITHET '06. 7th International Conference on. Presented at the Information Technology Based Higher Education and Training, 2006. ITHET '06. 7th International Conference on, IEEE, pp. 594–599. doi:10.1109/ITHET.2006.339673

[Ri12] Rindos, A., 2012. Innovation within IBM (and WebSphere) Development through University Collaborations, in: Proceedings of the ITI 2012 34th Int. Conf. on Information Technology Interfaces. Presented at the ITI 2012 34th Int. Conf. on Information Technology Interfaces, ITI 2012 34th Int. Conf. on Information Technology Interfaces, Cavtat / Croatia, pp. 11–12. doi:10.2498/iti.2012.0485

[Ro14] Robinson, P.A., Macnaghten, P., Banks, S., Bickersteth, J., Kennedy, A., Richardson, Y., Shaw, S., Sylvestre, I., 2014. Responsible scientists and a citizens' panel: new storylines for creative engagement between science and the public: Commentary. Geogr. J. 180, 83–88. doi:10.1111/geoj.12042

[RSL10] Rudakova, L., Söllner, M., Leimeister, J.M., 2010. Warum Hochschulrechenzentren andere IT-Service-Management Konzepte brauchen – eine konzeptionelle Analyse, in: 40. Jahrestagung der Gesellschaft für Informatik. Gesellschaft für Informatik, Leipzig.

[Si09] Siebenhofer, R., 2009. Systematic Cooperation between Industry and Universities. The Experience from Siemens IT Solutions and Services, in: Secure Software Integration and Reliability Improvement, 2009. SSIRI 2009. Third IEEE International Conference on. Presented at the Secure Software Integration and Reliability Improvement, 2009. SSIRI 2009. Third IEEE International Conference on, IEEE, pp. 324–324. doi:10.1109/SSIRI.2009.64

[ST11] Staley, D.J., Trinkle, D.A., 2011. The Changing Landscape of Higher Education (EDUCAUSE Review). Educ. Rev. 46, 16–33.

[St09] Stillman, L., 2009. What Is Community Informatics (and Why Does It Matter)? Inf. Commun. Soc. 12, 754–756. doi:10.1080/13691180902865986

[SJF09] Stillman, L., Johanson, G., French, R. (Eds.), 2009. Communities in action: papers in community informatics. Cambridge Scholars Pub, Newcastle upon Tyne, UK.

[Vo03] Von der Eltz, A., 2003. Running IT as a Business - Veränderung des IT-Bereichs vom Cost Center zum Wettbewerbsfaktor. OrganisationsEntwicklung 2003 / 4, 18 – 27.

[Vo12] Voros, K., 2012. Financial Analysis of a Successful Multi-User Academic Laboratory, in: Micro/Nano Symposium (UGIM), 2012 19th Biennial. Presented at the Micro/Nano Symposium (UGIM), 2012 19th Biennial, IEEE, pp. 1–1. doi:10.1109/UGIM.2012.6247051

[Wa12] Wang, Y., Zhang, L.-J., Cai, H., Sun, J., Li, N., 2012. Evaluating the Quality of Distance Education Services by Using Modern Information Technology, in: Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific. Presented at the Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific, IEEE, pp. 192–199. doi:10.1109/APSCC.2012.50

[We09] Weber-Wulff, D. (Ed.), 2009. Gewissensbisse: ethische Probleme der Informatik: Biometrie, Datenschutz, geistiges Eigentum, Kultur- und Medientheorie. Transcript, Bielefeld.

[WS10] Weick, K.E., Sutcliffe, K.M., 2010. Das Unerwartete Managen. Wie Unternehmen aus Extremsituationen lernen. Schäffer Poeschel, Stuttgart.

Eine IT-Referenzarchitektur für Hochschulen: Vom generischen Modell zur spezifischen Lösung

Andreas Hartmann, Wilhelm Rossak

Friedrich-Schiller-Universität Jena
Lehrstuhl für Softwaretechnik
Ernst-Abbe-Platz 2
07743 Jena
andreas.hartmann@uni-jena.de
wilhelm.rossak@uni-jena.de

Abstract: Der vorliegende Beitrag zielt auf die heterogene IT-Landschaft von Hochschulen und sucht nach einer zielführenden Möglichkeit, die Abstimmung zwischen geschäftlicher und technischer Seite zu optimieren. Hierbei soll eine weitgehend generische Referenzarchitektur vorgestellt werden, die auf der einen Seite die Hochschule aus geschäftlicher Sicht beschreibt und auf der anderen Seite eine IT-Sicht ableitet, die Strukturen optimal positioniert und folglich auch zukunftsfähig aufstellt. Dabei wird neben den zentralen Prozessen auch die Vereinfachung der IT-Landschaft im Allgemeinen adressiert.

1 Einleitung

Hochschulen sind geprägt von einer heterogenen sowie historisch gewachsenen IT-Landschaft und zahlreichen individuellen IT-Lösungen. „Kunden" sind primär Studierende und Wissenschaftler in Forschung und Lehre – selbst getrieben von einem hohen und auch notwendigen Maße an Individualität. Darüber hinaus muss die IT-Landschaft den Anforderungen einer modernen Verwaltung genügen. Einführungsprojekte von Campus Management Systemen (CMS) wie HISinOne verdeutlichen das Spannungsfeld zwischen individuellen Wünschen der Wissenschaftler auf der einen Seite und homogener, bezahlbarer Technologie auf der anderen Seite. Eine transparente und nachvollziehbare Referenzarchitektur, mit verschiedenen Aspekten und Schichten, kann dabei helfen dieses Spannungsfeld zu reduzieren. Insbesondere kann eine solche Architektur unterstützen, die IT vom Getriebenen[1] zum treibenden Innovationsfaktor zu entwickeln. Eine reale Umsetzung des hier präsentierten Ansatzes ist an der FSU Jena bereits erfolgt und wird an der Universität Erfurt aktuell durchgeführt.

Unsere Referenzarchitektur wird schrittweise von der Geschäftssicht hin zur IT-Sicht entwickelt und beschrieben (siehe Abschnitt 2 dieses Beitrags): Beginnend mit den strategischen Zielen der Hochschule wird eine allgemeine Struktur definiert, aus der dann

[1] Beitrag des IT-Strategieforums des ZKI e.V. [He12]

weiter ein übergreifendes Prozessmodell abgeleitet wird. Die Organisation folgt den Prozessen. Übergehend zur IT-Sicht werden zunächst Technologien allgemein beschrieben. Strategische Ziele und Struktur definieren die Rahmenbedingungen. Als Rahmen der IT-Landschaft wird eine Applikationslandkarte erstellt. Sie führt Technologien und Struktur schließlich im Referenzmodell zusammen. Neben dieser Übersicht werden einzelne Modelle zur detaillierten Beschreibung von IT-Systemen eingeführt. Abschließend formuliert ein Dienstkatalog das zu leistende Angebot und definiert dieses auch nach außen. Jeder der Schritte auf dem Weg zum Dienstkatalog entspricht dabei einem bestimmten Aspekt des Modells der Referenzarchitektur.

2 Lösungsansatz für die Referenzarchitektur

Die Entwicklung der geschäftlichen Sicht der Referenzarchitektur erfolgte unter Verwendung von Quasar Enterprise [En08], einem Framework für Enterprise Architektur Management. Für die Modellierung der Referenzen sowie die Beschreibung der technischen Sicht kamen ARIS Methode & Toolset [IDS10] zur Anwendung. Alle Referenzmodelle stehen in einer Referenzdatenbank zur Verfügung.

2.1 Schritt 1: Die Geschäftsarchitektur

Die Referenzarchitektur startet mit den strategischen Geschäftszielen der Hochschule, die sich im Vergleich zum Unternehmen unterscheiden. Auf einer Seite steht bei Hochschulen nicht die Gewinnmaximierung im Vordergrund, sondern z.B. die Exzellenz in der Forschung und Lehre. Ein klares Ziel der heute international konkurrierenden Hochschulen ist die Gewinnung und Bindung von herausragenden Forschern sowie Studierenden. Es entstehen prozessbezogene Ziele und eine Übertragung in die IT auf der anderen Seite, die sehr wohl Ähnlichkeiten verzeichnen. Abbildung 1 zeigt die abgeleitete Übersicht der strategischen Ziele. Deutlich zu erkennen ist, wie sich aus den Bereichen der Forschung und Lehre insbesondere Agilität und Innovation übertragen. Dieser Punkt ist bei der weiteren Gestaltung der Referenzarchitektur zu berücksichtigen, da sich diese Ziele im Vergleich zu Effizienz oder Korrektheit anders auswirken werden. Als Beispiel kann die Anwendung einer Projektmethode genannt werden. Im Bereich der Verwaltung wird ein klassisches Projektvorgehen in der Regel die bessere Wahl sein. Zum Ende des Projekts werden neu eingeführte IT-Systeme vom Test- in den Produktivbetrieb übergeben. Effizienz, Effektivität und Korrektheit sind treibende Faktoren. Anders sind bei einem Projekt in der Forschung Innovation und Agilität nicht nur gefragt, sondern vielmehr eine Voraussetzung. Während des Projekts eingeführte IT-Systeme erreichen häufig nicht einmal den Produktivstatus. Im Ergebnis kann festgehalten werden, dass die Hochschularchitektur maßgeblich durch übergreifende Ziele der Forschung, der Lehre und der Verwaltung geprägt wird. Die drei genannten Bereiche differenzieren sich jedoch später bei einer detaillierten Betrachtung hinsichtlich der Prozesse und der zur Unterstützung eingesetzten IT. Angelehnt an die Methode von [En08] lohnt es sich, einen gezielten Blick auf diese Geschäftsbereiche zu werfen und dort nach wiederkehrenden bzw. allgemeinen Strukturen zu suchen. Die Struktur liefert später die Basis für weitere Referenzmodelle.

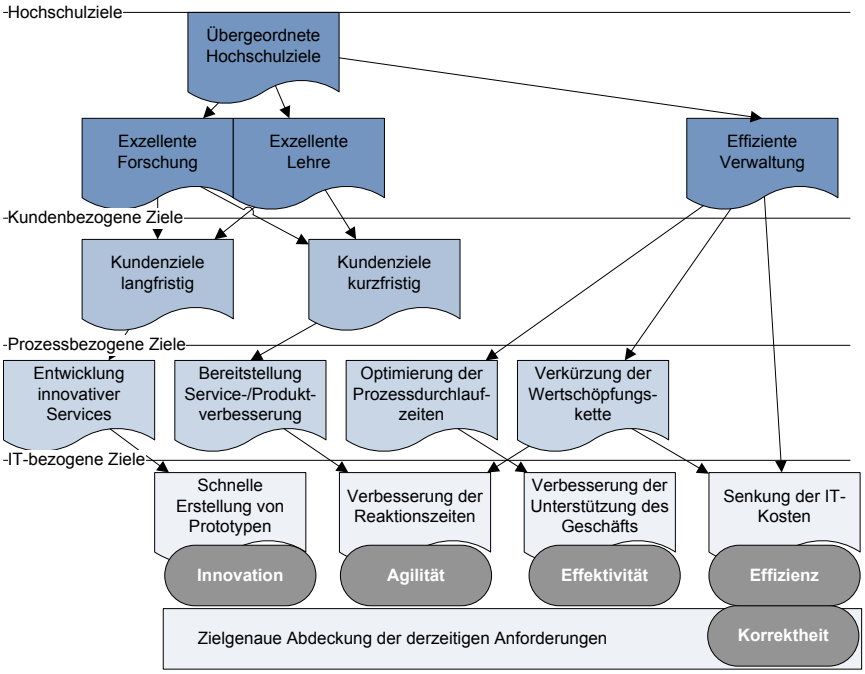


Abbildung 1 Strategische Ziele der Hochschule (nach [En08])

Die in Abbildung 1 dargestellten Ziele geben einen ersten Hinweis auf die Geschäftsbereiche der Hochschule. Das sind die Forschung, die Lehre und die Verwaltung. In der klassischen Organisation finden sich diese Bereiche tatsächlich in Form von Fakultäten oder Departments bzw. zentralen Universitätsverwaltungen wieder. Daneben gibt es noch weitere Geschäftsbereiche mit Anforderungen an IT-Unterstützung. Viele Hochschulen verfügen über eine Bibliothek (z.B. Universitätsbibliothek). Sowohl Forschungsprojekte sind damit verbunden als insbesondere auch die Lehre, die auf das dort vorhandene Wissen in Form von Büchern oder elektronischen Medien zugreift. Letztlich können noch die Universitätskliniken genannt werden, wobei hier auch rechtlich eigenständige Organisationsformen anzutreffen sind. Sie werden in der weiteren Betrachtung daher zunächst nicht weiter berücksichtigt.

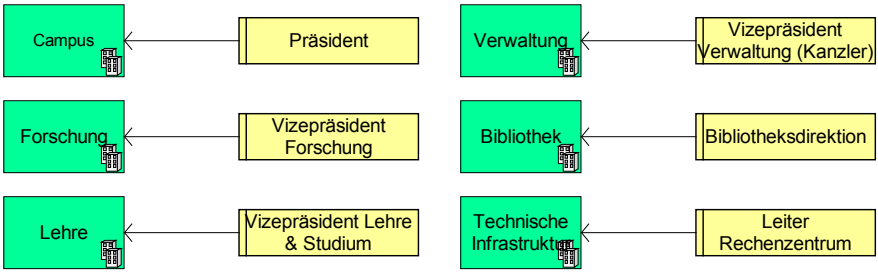


Abbildung 2 Geschäftsstruktur mit Bereichen und Stakeholdern

Abbildung 2 zeigt die abgeleiteten vier Geschäftsbereiche in einer Übersicht. Als Querschnitt kommen die Bereiche Campus und Infrastruktur hinzu. Der Bereich Campus steht für übergreifende Aspekte, die sich auf die gesamte Hochschule auswirken. Analog ist die Infrastruktur zu betrachten. Neben der primären Geschäftsstruktur stellt die Abbildung einen Bezug zur Organisation her bzw. eröffnet den Blick auf die Steuerungssicht. Entscheider sind in Form von typischen Rollen zugeordnet, die Kanten stehen für die Verantwortlichkeit. Rektorate (Präsidien) setzen sich meist aus eben jenen Stellen zusammen und bilden die Hochschulleitung. Wird das Thema IT-Governance auf dieser Basis betrachtet, so können gemäß [WR04] Archetypen abgeleitet werden. Unterschiede zwischen den Domänen Forschung & Lehre (business monarchy/feudal) sowie die der Verwaltung & Infrastruktur (feudal/anarchy) lassen erahnen, wie differenziert die IT-Steuerung in den einzelnen Geschäftsbereichen ausgeprägt ist.

2.2 Schritt 2: Das Prozessmodell des IT-Dienstleisters

[GKH10] stellt ein bewährtes Organisationsmodell für die IT in der Öffentlichen Verwaltung vor. Auf Basis von [GKH10] und mit Hilfe der Geschäftsstruktur wird ein Prozessmodell als Referenz für den IT-Dienstleister der Hochschule abgeleitet. Hierbei ist der Dienst als logisches Konstrukt zu verstehen, welches zwischen dem Anwender („was wird benötigt“) und der IT-Abteilung („was wird angeboten“) vermittelt. Der Dienst selbst wird als Leistung beschrieben und kann somit über ein Service Level Agreement (SLA) zwischen beteiligten Parteien vereinbart werden. Im Unterschied zum Dienst definiert der Prozess das „wie“ des IT-Dienstleisters und nicht die Leistung. So können mehrere Prozesse an der Erbringung eines Dienstes beteiligt sein. Darüber hinaus gibt es Prozesse für die Planung und Entwicklung von Diensten. Abbildung 3 zeigt das Referenzprozessmodell mit sechs Hauptprozessen.

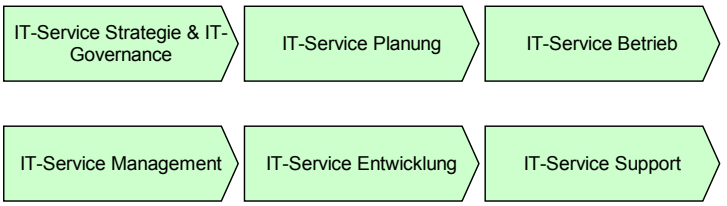


Abbildung 3 Abgeleitetes Prozessmodell mit sechs Hauptprozessen

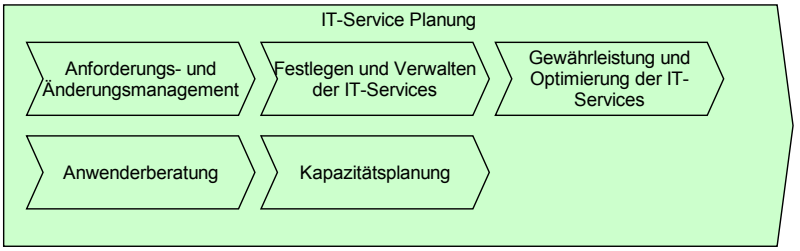


Abbildung 4 Abgeleitetes Prozessmodell, Auszug IT-Service Planung

Jeder Hauptprozess wird detailliert beschrieben und anschließend mit organisatorischen und technischen Zuordnungen ergänzt. In Abbildung 4 wird der Hauptprozess IT-Service Planung erweitert. Das vollständige Referenzmodell stellt nach [IDS10] alle sechs Hauptprozesse bis auf Ebene 3 der Funktionszuordnung dar. Einzelne Prozessbeschreibungen mittels EPK knüpfen an dieser Stelle an. Berücksichtigt werden im Referenzmodell auch Managementaufgaben, die im operativen Betrieb oft zu kurz kommen.

Im Idealfall nach [GKH10] würde jetzt die Organisation horizontal entlang der Hauptprozesse abgeleitet werden. In zahlreichen Hochschulen ist allerdings nicht genug Personal dafür vorhanden. So ist eine eigene Entwicklungsabteilung gerade für kleine Hochschulen und Fachhochschulen nicht finanzierbar. Vielmehr verteilen sich Aufgaben zusätzlich auf die vorhandenen Ressourcen. Das Prozess- und das Organisationsmodell berücksichtigen diesen Aspekt. Eine andere Anpassung betrifft den operativen Betrieb des IT-Dienstleisters. Hier unterscheidet das Referenzmodell den Betrieb von Anwendungssystemen und die IT-Infrastruktur. Die Trennung wurde unter Berücksichtigung der zuvor genannten Ressourcen/Organisation vorgenommen, d.h. wenn Aufgaben des Betriebs und der Entwicklung im selben Team zugeordnet sind. Die Betreuung der Anwendungssysteme ist hier sehr eng mit Geschäftsprozessen der Fachabteilung gekoppelt. In der Regel wird ein zusätzlicher Koordinator notwendig sein, der zwischen Fachverfahren und technischer Lösung vermittelt. Das Fachverfahren steht dabei im Vordergrund. Im Bereich der IT-Infrastruktur besitzt der IT-Dienstleister eine deutlich höhere Entscheidungskompetenz. Die Technologie und deren Einbettung in die IT-Landschaft sind treibend. Folglich unterscheiden sich die Aufgaben in einigen Punkten und die Organisation kann entsprechend angepasst werden, um hier effizienter zu arbeiten.

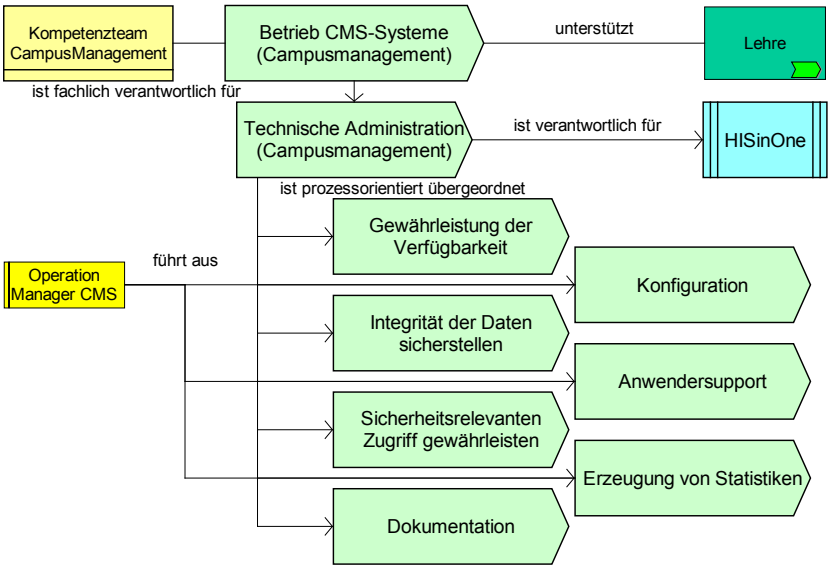


Abbildung 5 Abgeleitetes Prozessmodell für den IT-Service Betrieb, Campusmanagement

Abbildung 5 zeigt das Prozessmodell für die Bereitstellung eines CMS. Die Lesart des Modells ist dabei wie folgt. Das Kompetenzteam ist für den Betrieb und die Konfiguration eines Campus-Management-Systems fachlich verantwortlich, im Beispiel handelt es sich um das System HISinOne der HIS e.G. Der Prozess unterstützt den Geschäftsbereich Lehre, wodurch eine Verknüpfung zu den strategischen Hochschulzielen hergestellt wird. Folglich ist der fachliche Stakeholder in der Rolle des Vizepräsidenten für Lehre zu finden. Zum Betrieb des Systems gehört die technische Administration, die sich in standardisierte Teilaufgaben unterteilt (nach [GHK10]). Die Teilaufgaben werden vom Anwendungssystemadministrator (Rolle: Operation-Manager) wahrgenommen. Das Referenzmodell kann mit geringen Anpassungen auf andere Bereiche übertragen werden. Darüber hinaus beschreibt es typische Arbeitsvorgänge für Stellen der Rolle.

2.3 Schritt 3: Das Organisationsmodell des IT-Dienstleisters

Auf Basis des Prozessmodells wird jetzt das zugehörige Organisationsmodell abgeleitet. Von Vorteil ist, dass der Prozess Ausgangspunkt für die Organisation ist und nicht umgekehrt. Im ersten Schritt werden hierzu geeignete Unterstrukturen (Abteilungen) identifiziert. Rollen fassen ähnliche Aufgabenprofile zusammen und bilden die Grundlage für Kompetenzteams und den weiteren Stellenplan. Schließlich kann über eine Stellenbedarfsermittlung (u.a. Ausführhäufigkeit) die benötigte Quantität der Stellen ermittelt werden. Umgekehrt bringt die Prozesskostenrechnung Transparenz in den Personaleinsatz. Im Prozessmodell sind alle Aufgaben des IT-Dienstleisters zusammengefasst. Der proportionale Anstieg von Prozessinstanzen (=Personalbedarf) im Vergleich mit zu betreuenden IT-Lösungen unterstreicht das Optimierungspotenzial. So macht es einen deutlichen Unterschied, ob fünf oder 9 Datenbanktechnologien unterstützt werden sollen. Der IT-Dienstleister kann seinen Wunsch nach einer homogenen IT-Landschaft argumentieren – das strategische Ziel „Effiziente Verwaltung“.

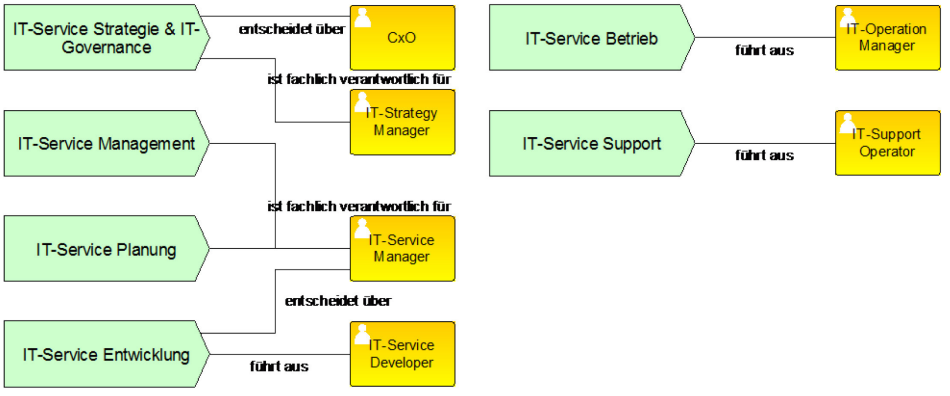


Abbildung 6 Abgeleitete Rollen aus dem Prozessmodell

In Abbildung 6 werden zunächst sechs verschiedene Rollen definiert, wie sie zur Wahrnehmung vergleichbarer Aufgaben benötigt werden. Die Rollen sind angelehnt an ITIL und werden durch dieses ergänzt. Es lassen sich entsprechende Stellentypen zum Aufbau

der Organisation ableiten. Abbildung 7 zeigt das Referenzmodell des Organigramms für den IT-Dienstleister mit Abteilungen und entsprechend des Prozessmodells (Auszug). Die dargestellten Organisationseinheiten werden nach Bedarfsermittlung und auf Basis der Rollen mit Stellen aufgebaut. Zudem unterstützen dezentrale Kompetenzteams den Bedarf an individueller Beratung und werden dynamisch in das Organigramm eingebunden. Sie stellen den wichtigen Kontakt zu den zentralen Teams her und können somit Standardlösungen vermitteln. Die Kompetenzteams bestehen z.B. aus einer Anzahl von Operation Managern sowie mindestens einem IT-Service Manager, der auch die Rolle des IT-Beraters übernehmen kann. Empfohlen wird, mit dem Referenzmodell noch einmal die Archetypen der IT-Governance zu bewegen, denn hier lässt sich jetzt *federal* anstelle von *anarchy* anstreben.

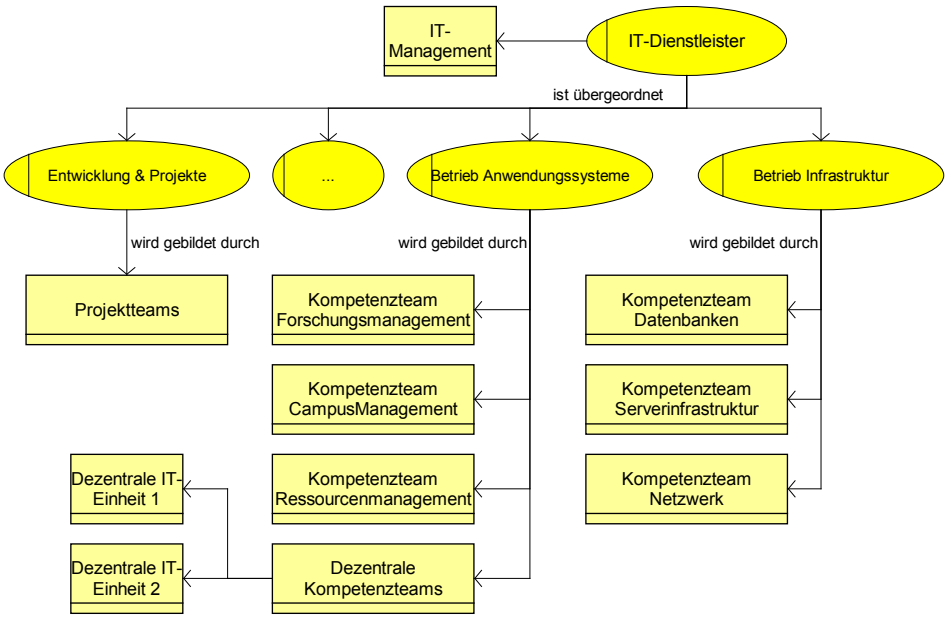


Abbildung 7 Auszug des abgeleiteten Organigramms, IT-Dienstleister

2.4 Schritt 4: Das Technologieverzeichnis

Ein essentieller Bestandteil der IT sind die Technologien. Über sie wird faktisch die technische Unterstützung der Hochschule hergestellt. Aber welche Technologien werden eingesetzt und was wird wirklich benötigt? Der Architekturbaukasten nach [IDS10] gibt eine erste Antwort. In der Rolle des IT-Architekten kann eine verantwortliche Person den aktuellen Stand aufnehmen, konsolidieren und prüfen, was in der Zukunft benötigt wird. Der Architekturbaukasten differenziert gemäß der Geschäftsstruktur in geschäftsprozess-spezifische Komponenten, Standardtechnologien und Infrastruktur. Ein weiterer Bereich Entwicklungstechnologien wird nicht bei jeder Hochschule anzutreffen sein. Zum Beispiel finden sich individuelle Technologien der Forschung im oberen Bereich.

Ein ERP-System würde folglich den unterstützenden Geschäftsprozessen zugeordnet werden. Email und DBMS gehören klar zu den Standardtechnologien. Abbildung 8 zeigt den Architekturbaukasten als Referenz mit einer umfangreichen Auswahl an Unterkategorien. Gängige Lösungen sind bereits verzeichnet und darüber hinaus können individuelle Bedarfe der Hochschule ergänzt werden (siehe auch Abbildung 10).

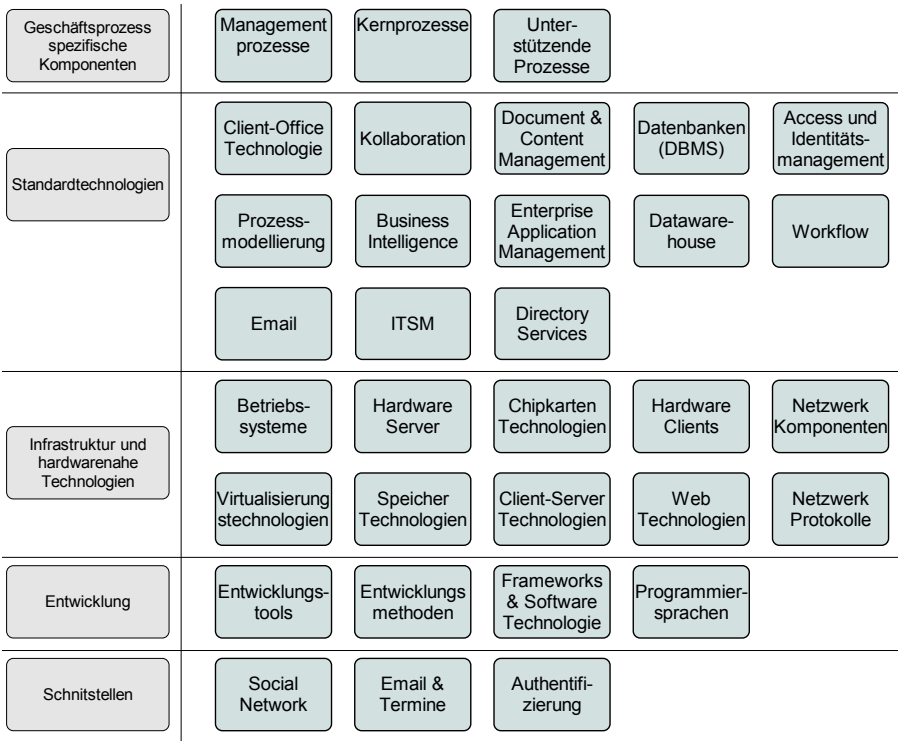


Abbildung 8 Architekturbaukasten mit Technologieverzeichnis

Neben den Technologien können im Baukasten auch weitere Standards verwaltet werden. Das können z.B. Schnittstellen sein, die von der Hochschule grundsätzlich und unabhängig von den dahinter stehenden (implementierenden) IT-Systemen stehen. [He13] zeigt zukunftsweisend, wie das Bild der IT-Landschaft durch Schnittstellenstandards definiert wird. Der IT-Architekt kann dieser Entwicklung Rechnung tragen und die Schnittstellen in sein Portfolio aufnehmen (z.B. ics, iCal, calDAV, SyncML).

2.5 Schritt 5: Die Applikationslandkarte

Im Architekturbaukasten wurden Technologien eingetragen und eine Kategorisierung vorgenommen. Hierfür wurde die Struktur aus Abbildung 2 einbezogen. Die Referenz der Applikationslandkarte kann nun erstellt werden, indem die Differenzierung konsequent und ausgerichtet am Geschäft weiter verfolgt wird (Business-IT Alignment). Alle Technologien werden zunächst mit einer allgemeinen und möglichst produktunabhängi-

gen Bezeichnung versehen. Sie werden später mit konkreten Lösungen untersetzt. Die Domänen der Applikationslandkarte werden analog zur Geschäftsstruktur definiert und dabei insbesondere individuelle Lösungen aus dem Kerngeschäftsfeld von Standardtechnologien unterschieden. Im Ergebnis wird die Applikationslandkarte (Abbildung 9) erstellt. Die Technologien aus dem Baukasten werden eingetragen und über weitere Technologiemodelle konkretisiert. Die Technologien werden in diesem Fall als Klassen modelliert und die konkreten Lösungen den Klassen zugeordnet.

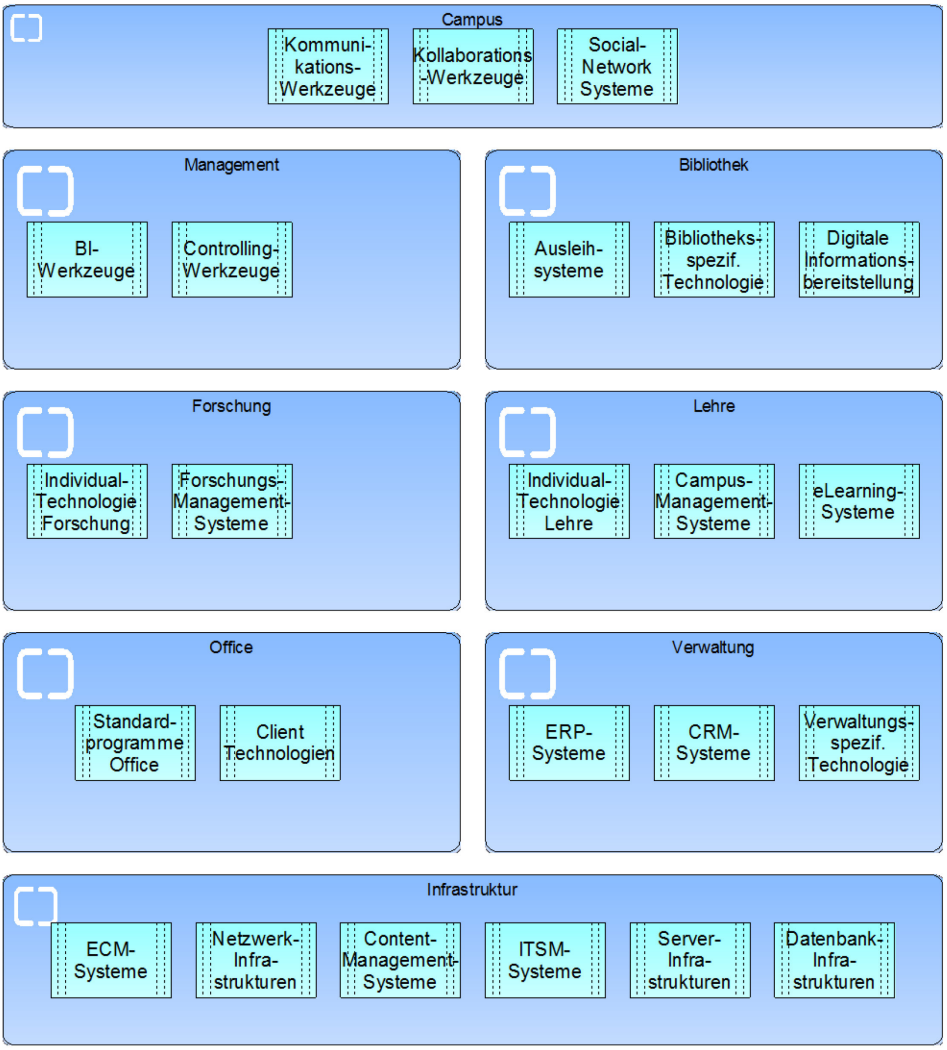


Abbildung 9 Auszug der abgeleiteten Applikationslandkarte

Abbildung 10 zeigt ein Technologiemodell, in welchem die Klasse ERP-Systeme mit den zur Verfügung stehenden Produkten untersetzt wird. Der IT-Architekt soll hier aller-

dings nicht das komplette Marktangebot eintragen, sondern nur freigegebene Technologien aus dem Architekturbaukasten. Bei einigen Hochschulen würde demnach nur die Lösung der HIS GmbH eingetragen.

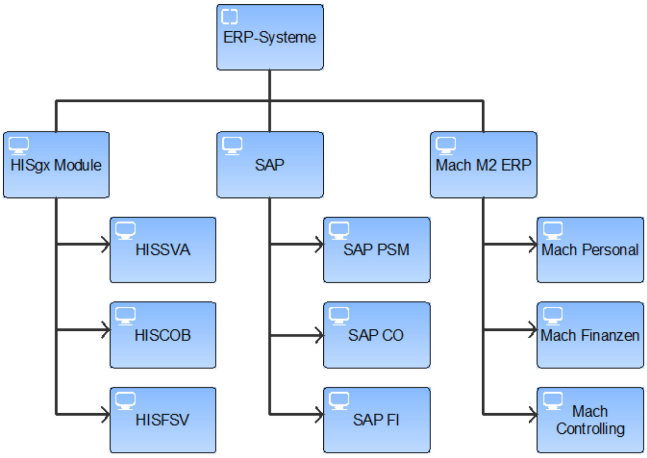


Abbildung 10 Technologiemodell für konkrete IT-Systeme (Auszug)

2.6 Schritt 6: Die IT-Systeme

Nachdem die Technologien und die Applikationslandkarte festgelegt wurden, können die einzelnen Lösungen detaillierter beschrieben werden. Hierzu gehören u.a. Referenzarchitekturen für die Installation von IT-Systemen. Entsprechende Standards sind definiert und verfügbar (siehe [BMI11] ff.). Zu unterscheiden ist die Technologieklasse, die konkrete IT-Lösung und die installierten Instanzen.

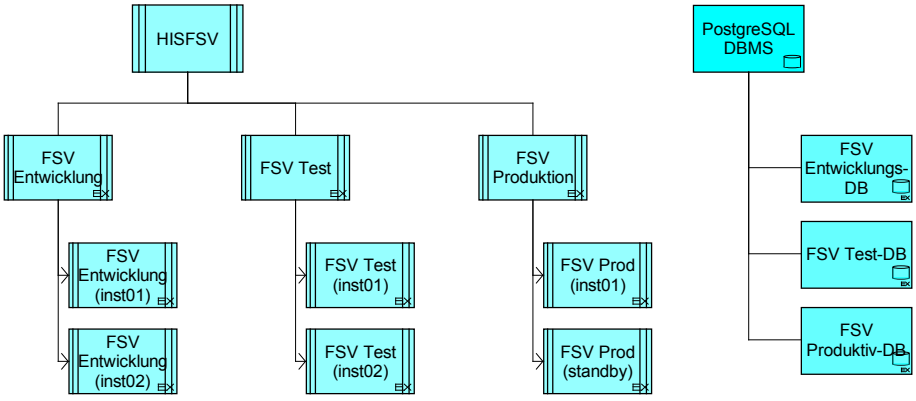


Abbildung 11 Beispiel eines Referenzmodells für IT-Systeme

In Abbildung 11 wird beispielhaft das Finanzverwaltungssystem HISFSV der HIS GmbH dargestellt. Es gehört zur Technologieklasse ERP-Systeme. Der IT-Architekt legt als Referenzarchitektur die Installation in mindestens drei Instanzen fest, wobei nach Entwicklungs-, Test- und Produktivsystem zu unterscheiden ist. Darüber hinaus wird entsprechend des Prozessmodells das Release- und Änderungsmanagement der Instanzen definiert. Neben dem IT-System können auch eingesetzte Datenbanken und Betriebssysteme auf diese Art näher spezifiziert werden. Im Beispiel gehört PostgreSQL zur Technologieklasse DBMS aus den Standardtechnologien.

2.7 Schritt 7: Der Dienstkatalog

Aus Prozessmodell und Applikationslandkarte kann nun ein Dienstkatalog für den IT-Dienstleister abgeleitet werden. Einige Prozesse wie z.B. im Bereich des IT-Supports dienen hierbei als direkter Ausgangspunkt für einen anzubietenden Dienst (IT-Umzüge, Technikverleih, usw.). Andere Dienste leiten sich aus den zu betreibenden IT-Lösungen ab, wobei der Betrieb im Vordergrund steht. Der IT-Dienstleister ist hier als Anbieter im Sinne von Hosting-Dienstleistungen zu verstehen. Es kann dabei frei entschieden werden, ob man einen allgemeinen Dienst definiert, der für alle Lösungen einer Technologieklasse oder gar einer Domäne angeboten wird. Der Vorteil wäre ein entsprechend geringer Dokumentationsaufwand und nur ein SLA für mehrere IT-Lösungen. Auf der anderen Seite könnte weiter differenziert werden und für jedes konkrete IT-System ein SLA entwickelt werden.

Eine Rückkopplung zur Geschäftssicht kann problemlos nach [En08] durchgeführt werden, indem IT-Dienste zur Unterstützung der Geschäftsdienste mit einer entsprechenden Kante im Modell verknüpft werden. Im Ergebnis zeigt sich mit dem Dienst als Bindeglied eine aussagekräftige Übersicht, für welche Geschäftsbereiche welche IT eingesetzt wird.

3 Realisierung, Fazit und Ausblick

Die Referenzarchitektur wurde im Bereich der FSU Jena vollständig umgesetzt. Die ehemalige Verwaltungs-IT konnte dabei neu organisiert und zukunftsfähig ausgerichtet werden. Das Personal wurde deutlich effizienter eingesetzt und anstelle von Stellenkürzungen ein größeres Aufgabenportfolio übernommen. Das bedeutete bei mehr Aufgaben eine Einsparung von 2,5 VZÄ (entspricht 12,5%). Heute arbeitet das Team vorausschauend anstelle reagierend, wodurch die Anzahl von Havarien maßgeblich reduziert wurde. Das zugehörige Modell wurde vollständig in ARIS modelliert und kann auf Anfrage weitgehend zugänglich gemacht werden. Es wurde jedoch nicht öffentlich publiziert.

Der vorliegende Beitrag stellt eine Referenzarchitektur für Hochschulen vor, mit besonderem Blick auf den IT-Dienstleister und die IT-Architektur. Auch wenn die Hochschule in ihren Kernbereichen der Forschung und Lehre von einem hohen Maß an Individualität geprägt ist, zeigt sich jedoch, dass durchaus Standards und generische Muster eingeführt werden können. Die Hochschule unterscheidet sich zudem in einigen Bereichen maßgeb-

lich von Unternehmen, was die Übertragung von Industriestandards erschwert. So gibt es z.B. keine Gewinnsteigerung als Geschäftsziel und sie hat deutlich andere Entscheidungsstrukturen. Die vorgestellte Methode berücksichtigt diese Unterschiede und macht sie darüber hinaus transparent. Dabei muss allerdings auf eine sorgfältige und konsistente Differenzierung der Belange geachtet werden. Das Referenzmodell orientiert sich dazu am Geschäftsmodell und trägt dessen Grundstrukturen in alle anderen Bereiche. Die Umsetzung an der FSU Jena zeigt, dass mit einer prozessorientierten Ausrichtung der Organisation und dem gezielten Einsatz der IT einige Optimierungen machbar sind. Hochschulen können dieses Potenzial nutzen, um in die Unterstützung der individuellen Bereiche von Forschung und Lehre zu investieren. Im Wettbewerb mit internationalen Konkurrenten sollte das Potenzial keinesfalls ungenutzt bleiben.

Die Referenzarchitektur bietet noch zahlreiche Möglichkeiten der Erweiterung, z.B. durch konkrete Teilarchitekturen und Modelle. Darüber hinaus können Empfehlungen und Richtlinien der DFG oder des ZKI e.V. implementiert und übersichtlich zur Verfügung gestellt werden. Anwender können Richtlinien somit direkt auf ihre lokalen Umgebungen adaptieren. Eine interessante Weiterentwicklung eröffnet sich über das ARIS Toolset. Da die Referenzarchitektur als Modell dort implementiert ist, kann z.B. eine Qualitätsprüfung technisch unterstützt werden.

Derzeit arbeiten wir an einer weiteren Verallgemeinerung unseres Ansatzes, dessen Abgleich mit existierenden Standards und spezifischen Lösungsmodellen an anderen Hochschulen, sowie an einer Abbildung auf weitere Anwendungsbereiche. Erarbeitet wurde z.B. eine dienstorientierte Adaption auf Landesebene, wobei gleichartige Dienste der Verwaltung und Infrastruktur konsolidiert wurden. Die Geschäftsbereiche der Forschung und Lehre wurden mit den entsprechenden individualisierten Diensten in einem gemeinsamen Modell erweitert. Das Ergebnis zeigt einen skalierbaren Ansatz zur Konsolidierung von neun Thüringer Hochschulrechenzentren auf ein bis maximal zwei Struktureinheiten – mit klarem Einsparpotenzial und Serviceverbesserung.

Literaturverzeichnis

[BMI11] Das Architekturmanagement der IT-Steuerung Bund, Bundesministerium des Inneren, Berlin, 2011

[En08] Engels, H. H.: Quasar Enterprise. dpunkt.verlag GmbH, Heidelberg, 2008.

[GKH10] Gündüz R.; Köpp P; Hahn M.: Entwicklung einer IT-Organisation. In (itSMF e.V., Hrsg.): Organisationsmodell für die IT in der Öffentlichen Verwaltung. Symposion Publishing GmbH, 2010; S. 45-61

[He12] von der Heyde, M.: Treiben oder getrieben werden. In (ZKI e.V., Hrsg.): IT-Strategie in Hochschulen, Selbstverlag des ZKI, Weimar, 2012

[He13] von der Heyde, M.: Integration persönlich genutzter Services in den Hochschulalltag – simply bring your own service (BYOS). In (Universität Hamburg, Hrsg.): Universitätskolleg-Schriften. Hamburg, 2013

[IDS10] IDS Scheer AG. Benutzerhandbuch IT-Architekt 7.1. IDS Scheer AG, Saarbrücken, 1992-2010

[WR04] Weill, P.; Ross, J.: IT-Governance: how top performers manage IT decision rights for superior results. Harvard Business School Press, Boston, 2004

Hochschul-Apps im Überblick

Dietmar Zoerner, Florian Gößler, Ulrike Lucke

Universität Potsdam
Institut für Informatik und Computational Science, A.-Bebel-Str. 89, 14482 Potsdam
vorname.nachname@uni-potsdam.de

Abstract: Der Beitrag gibt einen Überblick über vorhandene mobile Apps an deutschen Hochschulen. Dafür werden inhaltliche, technische und organisatorische Aspekte ausgewertet. Aus dem Vergleich bzw. der Kategorisierung ergibt sich eine Aufstellung von zu bedenkenden Fragen und möglichen Realisierungswegen, die eine Entscheidungshilfe für Hochschulen bei der Entwicklung eigener Apps bietet.

1 Motivation

„Smartphone statt Megafone“[1] titelte kürzlich die Webseite zum Wissenschaftsjahr der digitalen Gesellschaft. Moderne Technologien durchdringen unseren Alltag, und mobile Mehrwertdienste bieten eine Vielzahl neuer Möglichkeiten [Ca+13]. Die Digitalisierung der Hochschule brachte neue Herausforderungen mit sich (z.B. E-Learning [LS12]), die sowohl zu zukunftsweisenden Neuerungen im Studium [Ke13] als auch zu kritischer Reflexion [Be11] führten. Auch in Forschung und Verwaltung ist die Digitalisierung nicht aufzuhalten [BL13][Ka+13][Za+12]. Gleichzeitig wird im Zuge der globalisierten Bildung von den Hochschulen ein deutlich wachsendes Maß an Kundenorientierung und Flexibilisierung verlangt. Dies umfasst alle Akteure in Lehre, Forschung und Verwaltung, aber in besonderem Maße ist dies auf das Gewinnen und Halten von Studierenden gerichtet. Mobile Angebote sind daher ein Marketinginstrument von wachsender Bedeutung, wozu neben per Smartphone zugreifbaren Werkzeugen wie Lernplattformen[2] oder Campus Management Systemen[3] (die allein noch keine umfassende Sicht auf die Hochschule bieten) zunehmend auch Apps zur Integration mehrerer Dienste zählen.

Zwar sind Apps, die einen mobilen Zugang zu oft ohnehin im Web verfügbaren Informationen bieten sollen, schnell implementiert, doch an eine professionelle Lösung sind weitergehende *Anforderungen* zu stellen [Mr13]: ein konsistentes Layout über verschiedene Plattformen hinweg (Corporate Identity), ansprechende Icons, Bilder und Animationen, eine flüssige Performance und eine intuitive Benutzung sollten selbstverständlich sein. Darüber hinaus sollte ein hinreichender Umfang an Funktionen „aus einer Hand“ geboten werden, wohingegen auf simple WebViews verzichtet werden sollte, die lediglich Webseiten einbinden die oft nicht einmal für mobile Geräte gestaltet

[1] https://www.digital-ist.de/aktuelles/digital-aktuell-der-newsticker/smartphone-statt-megafone.html
[2] z.B. http://docs.moodle.org/25/de/Mobile_App
[3] z.B. http://www.datenlotsen.de/produkte-2013/campusnet-mobile/ueberblick

wurden. Auch eine attraktive und aussagekräftige Präsentation im App-Store ist ein nicht zu unterschätzender Faktor für den Erfolg einer App. Eine besondere Herausforderung stellt zudem die Anpassung an die Spezifika einer mobilen Plattform dar. Beispielsweise unterscheiden sich Apple/iOS und Google/Android in der Handhabung z.T. deutlich. Gute App-Entwickler können Richtlinien der Corporate Identity einer Hochschule zusammenführen mit den Design Guidelines konkreter Plattformen sowie deren Besonderheiten von Hardware, Benutzerführung und Look & Feel.

Für eine Hochschule, die eine eigene App gestalten möchte, sind also zahlreiche konzeptionelle Überlegungen anzustellen. Neben funktionalen und gestalterischen sind auch technische und organisatorische Entscheidungen zu treffen. Welche Features sollen in die App hinein? Für welche dieser Funktionen gibt es bereits Schnittstellen in der vorhandenen Campus-Software? Lohnt sich eine Eigenentwicklung, oder sollte das Projekt extern vergeben werden? Kann auf einen Baukasten zurück gegriffen werden? Wer bietet die App im Store an? Welche Kosten entstehen, sowohl für die Veröffentlichung als auch in den Folgejahren? Für welche Plattformen soll die App angeboten werden, oder doch nur als mobile Webseite? Für diese und weitere Fragen liefert der vorliegende Beitrag eine Entscheidungshilfe, indem der Stand der Forschung i.S.v. bestehenden Erkenntnissen guten Software-Designs [Wi+14] abgeglichen wird mit dem Stand der Technik i.S.v. bestehenden App-Lösungen.

Der Beitrag stellt zunächst in Abschnitt 2 bestehende Veröffentlichungen zu diesem Thema vor und leitet daraus den Bedarf einer umfassenderen Analyse ab. Anschließend werden in Abschnitt 3 die untersuchten Apps dargestellt, und in den Abschnitten 4 bis 6 werden sie aus inhaltlicher, technischer und organisatorischer Perspektive gegenüber gestellt. Der Beitrag schließt mit einer Zusammenfassung der Ergebnisse in Abschnitt 7.

2 Vorhandene Studien

Aus wissenschaftlicher Sicht gibt es Annäherungen an das Thema Hochschul-Apps v.a. aus den Perspektiven IT-Strategie [He+09] und Akzeptanz mobiler Dienste [Bü+11]. Dabei wurden bislang mobile Anwendungen allgemein betrachtet, und nicht spezielle Apps um die Web-Dienste einer Hochschule auf Smartphones zugängig zu machen. Das Thema schien bislang vornehmlich Computerzeitschriften und Community-Plattformen vorbehalten zu sein, in denen diverse Rankings[4][5] veröffentlicht wurden. Diese Listen sind jedoch von einer meist willkürlich erscheinenden Auswahl geprägt. Auch in populären Magazinen [We13] findet sich das Thema zunehmend wieder.

Auf Fachkonferenzen konnten wir bislang nur einen Beitrag [Bi13] finden. Dieser Überblick war jedoch auf Apple/iOS beschränkt, weil es zum damaligen Zeitpunkt dort die meisten Apps für Hochschulen gab. Dass dies nicht mehr so ist, kann der vorliegende Beitrag zeigen. Zudem erfolgte die Auswertung i.W. als Überblick der angebotenen Funktionen, was allein noch keine hinreichende Entscheidungshilfe für andere

[4] http://www.studis-online.de/Studieren/studi-apps.php
[5] http://www.chip.de/news/Apps-fuer-Studenten-Downloads-zum-Uni-Start_52200476.html

Hochschulen bei der Konzeption einer App liefert. Darüber hinaus gibt es in der Fachliteratur diverse Vorstellungen von Einzelsystemen, etwa um den Campus zu erkunden [Pe+12][Ze+14]. Der Fokus dieses Beitrags liegt jedoch auf mobilen Zugängen zu Regeldiensten wie Campus Management oder Lernplattform. Über die vorhandenen Ranking hinaus sollen derartige Apps daher nachfolgend aus inhaltlichen, technischen und organisatorischen Gesichtspunkten analysiert werden.

3 Datengrundlage

Für die Identifikation relevanter Apps wurden die Stores von Apple und Google nach Suchbegriffen wie „Universität", „Fachhochschule" etc. durchsucht. (Die Windows-Plattform wurde nicht berücksichtigt, da die Marktanteile bedeutend geringer sind [Co+14].) Dabei wurden v.a. für Android sehr viele ähnliche Begriffe aus Fremdsprachen gefunden, die irrelevante Ergebnisse lieferten und manuell aussortiert werden mussten. Ergänzend wurden Apps, die in einem Store gefunden wurden, gezielt auch im anderen gesucht. Einige Apps wurden zudem vom Store mittels Auto-Vervollständigung im Suchfeld vorgeschlagen und dadurch gefunden.

Aufgrund der inzwischen sehr großen Zahl an Apps (nicht selten gibt es mehrere konkurrierende Lösungen für eine Hochschule) fokussiert die nachfolgende Analyse auf die weiter entwickelten Apps, die mehr als nur Mensa-Pläne bieten. (Eine Stichprobe ergab, dass sich v.a. Windows-Apps im Wesentlichen auf diese Funktion beschränken.) Somit flossen iOS- und Android-Apps der folgenden 58 deutschen Hochschulen in die Untersuchung ein:

- RWTH Aachen
- Uni Augsburg
- Uni Bayreuth
- TU Berlin
- HU Berlin
- FU Berlin
- Uni Bielefeld
- Uni Bonn
- Uni Bremen
- TU Darmstadt
- Uni Dortmund
- FH Dortmund
- TU Dresden
- Uni Duisburg-Essen
- Uni Düsseldorf
- FH Erfurt
- Uni Erlangen-Nürnberg
- Uni Flensburg
- Uni Frankfurt/Main
- Uni Frankfurt/Oder
- *TU Freiberg*
- Uni Freiburg
- Uni Göttingen
- Uni Halle
- Uni Hamburg
- Uni Hannover
- *Uni Heidelberg*
- Uni Hildesheim
- FH Hof
- *Uni Hohenheim*
- FH Jena
- KIT Karlsruhe
- Uni Kiel
- Uni Koblenz-Landau
- Uni Köln
- Uni Konstanz
- Uni Leipzig
- Uni Lüneburg
- Uni Magdeburg
- FH Mainz
- *Uni Mannheim*
- *HS Mittweida*
- TU München
- Uni Münster
- *HS Niederrhein*
- FH Offenburg
- Uni Osnabrück
- *Uni Paderborn*
- Uni Passau
- Uni Potsdam
- Uni Regensburg
- Uni Saarbrücken
- Uni Stuttgart
- Uni Trier
- Uni Tübingen
- Uni Ulm
- Uni Weimar
- FH Westküste

Den in Abschnitt 1 genannten Anforderungen werden dabei insbesondere die kursiv hervorgehobenen Apps gerecht. Web-Apps bzw. responsive Webseiten sind in der Analyse nicht enthalten, da sie nicht über zentrale App-Stores vertrieben werden. Auch Apps, die über einen dedizierten Store einer Hochschule vertrieben werden, konnten so nicht gefunden werden. Aufgrund der hohen Dynamik der App-/Web-Landschaft wäre eine manuelle Suche zu aufwändig und dadurch schnell veraltet. Auch nicht-hochschulspezifische Apps[6] sowie die Nutzung anderer Apps durch Hochschulen (etwa für die Distribution von Vorlesungsaufzeichnungen[7] oder für Chats zum Studierendenmarketing[8]) wurden hier nicht betrachtet.

4 Inhaltliche Analyse

Die am häufigsten (aber inzwischen nicht mehr von allen Hochschul-Apps) angebotene Funktion ist nach wie vor der Mensa-Speiseplan, dicht gefolgt von lokalen News. Der über mobile Geräte erzielbare Mehrwert ortsbezogener Informationen (i.S.v. Raum- oder Gebäudeplänen) stellt ebenfalls eine wichtige Funktion dar. Interaktive, individuelle und unmittelbar studienrelevante Aspekte wie Vorlesungsverzeichnis, Stundenplan, Leistungsübersicht (d.h. Zugang zum Campus Management System) oder E-Learning (d.h. Zugang zur Lernplattform) finden sich jedoch immer noch selten. Manche Apps bieten bereits Schnittstellen zum lokalen ÖPNV oder zur Bibliothek.

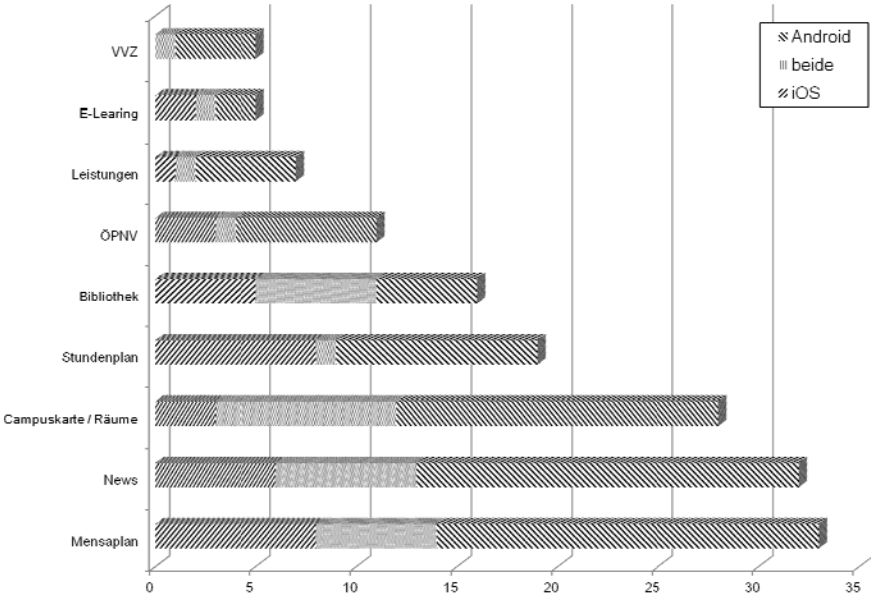


Abbildung 1: Absolute Häufigkeit der vorhandenen Funktionen pro Betriebssystem

[6] z.B. http://www.campus-app.de/
[7] z.B. https://itunes.apple.com/de/app/itunes-u/id490217893
[8] http://www.yaez.de/Studium/3332-Whats-App-Chat-Studieren-in-Ostdeutschland.html

Aus der in Abbildung 1 nach Betriebssystem differenzierten Verteilung der Funktionen ist ersichtlich, dass es nur noch marginale funktionale Unterschiede zwischen Android- und iOS-Apps der Hochschulen gibt. Zudem gibt es vereinzelt weitere Funktionen wie Campus Radio / TV, Stellenmarkt, Hochschulsport, Freizeitangebote, Personensuche oder FAQ.

Für fast alle Hochschulstandorte existieren mehrere Apps für Mensapläne. Es gibt sowohl Allrounder, die mehrere Hochschulen innerhalb einer App abbilden (z.B. mensa.app, Mensaplan, My.Mensa, Mensa-Jäger), als auch Apps eines zentralen Hersteller die jeweils nur eine konkrete Mensa bzw. Hochschule unterstützen. Desweiteren gibt es oft kleinere Apps von Einzelpersonen für einzelne Hochschulen. Diese reinen Mensa-Apps wurden in der vorliegenden Analyse nicht betrachtet.

Als weitere spezialisierte Apps sind ED Sync (iOS) zur Abfrage rund um die Bücherausleihe in der Bibliothek sowie eStudy (Android) zum Zugriff auf die Lernplattform Stud.IP zu nennen. Letztere zeichnet sich insbesondere durch ein fundiertes Konzept für die Wiederverwendbarkeit und Portabilität aus.

Unter den Baukästen mit umfassender Funktionalität (die jedoch nicht überall ausgeschöpft wird) sind Campus App und Campus News (beide Android) zu erwähnen.

5 Technische Analyse

Für die nachfolgend beschriebenen Untersuchungen wurden die o.g. Apps auf Testgeräten installiert und – soweit möglich – ihr Code und ihr Verhalten analysiert.

5.1 Softwaredesign

Bei der Konzeption einer App stellt sich die grundlegende Frage nach den zu unterstützenden Betriebssystemen, wie z.B. Android und iOS als die derzeit am weitesten verbreiteten Systeme. Zur Entwicklung nativer Apps für bestimmte Betriebssysteme stehen umfassende Hilfsmittel zur Verfügung. Weitgehend offene Ansätze hinsichtlich der unterstützten Betriebssysteme laufen dagegen entweder direkt in einem Browser oder basieren auf Browser-basierten Ansätzen, wie z.B. phonegap[9] oder chayns[10]. Solche Betriebssystem-unabhängigen Ansätze besitzen den Vorteil, der großen Vielfalt der verschiedenen Endgeräte der Studierenden eine breite Unterstützung zu bieten und sind zudem unabhängiger in Bezug auf neue Betriebssysteme oder Betriebssystem-Updates. Als Nachteil dieser Ansätze müssen Kompromisse bzgl. der möglichen Unterstützung von Betriebssystem-spezifischen Interaktionsformen, Design-prinzipien und Funktionen angesehen werden, die von Betriebssystem-spezifischen Lösungen deutlich besser berücksichtigt werden. Eine sorgfältige Abwägung für diese grundlegende Entscheidung muss deshalb die erwartete Vielfalt der unter den Nutzern

[9] http://phonegap.com
[10] http://www.tobit.de/chayns

(d.h. Studierenden, Mitarbeitern und Gästen) verbreiteten mobilen Endgeräte dem für die angestrebte Funktionalität erforderlichen Bedarf an spezifischer Unterstützung von Betriebssystem-Merkmalen gegenüber stellen.

Von den untersuchten 58 Apps sind 37 für Android und 39 für iOS verfügbar. Native Apps stellen also die überwiegende Mehrheit dar. Insbesondere viele kleinere, hier nicht berücksichtigte Apps mit nur einer Funktion sind vorwiegend unter Android verfügbar, was vermutlich auf die günstigeren und somit v.a. für Studierende attraktiveren Gebühren des Google Play Store zurückzuführen ist. Eine Analyse des Codes bzw. des Erscheinungsbilds ergab, dass offenbar nur eine Handvoll Apps mit Frameworks wie phonegap oder chayns Plattform-übergreifend entwickelt wurden.

Weiterhin haben die erläuterten Alternativen zum Softwaredesign erheblichen Einfluss hinsichtlich der Aufwände bei der Softwareentwicklung- und Pflege. Grundsätzlich ließe sich durch die Verwendung Betriebssystem-unabhängiger Ansätze der Entwicklungsaufwand reduzieren, denn es muss nicht für jedes Betriebssystem eine eigene Software entwickelt werden. Ob im konkreten Projekt eine Aufwandsreduzierung tatsächlich realisierbar ist, hängt neben den gewünschten Features aber vor allem von den bereits vorhandenen Erfahrungen der beteiligten Softwareentwickler ab.

Eine Alternative zu einer kompletten Eigenentwicklung können Baukasten-Systeme wie Campus App[11] bilden, aus deren vordefinierten Modulen die gewünschte Funktionalität zusammen gestellt und an Design und Standard-Software der jeweiligen Hochschule angepasst werden kann. Besonders hervorzuheben sind in diesem Zusammenhang die modularen Architekturkonzepte der Apps aus Hohenheim[12] und Potsdam [Ki+14], die verschiedene native und Web-basierte Komponenten zu einer durchgängigen Lösung vereinen.

Letztlich stellt auch die externe Vergabe eines Entwicklungs- und Wartungsauftrags einen möglichen Weg dar. Aus Kostengründen wird dies jedoch angesichts knapper öffentlicher Kassen und oft hoher App-Kompetenzen im IT-Team bzw. unter Studierenden der Hochschulen selten die beste Wahl sein. Auch vor dem Hintergrund der sparsamen Verwendung von Steuermitteln kann dies als kritisch eingestuft werden. Inwiefern die untersuchten Hochschul-Apps über einen Auftrag extern entwickelt wurden, ließ sich im Rahmen der durchgeführten Untersuchung nicht ermitteln; hierfür hätten die Hochschulen bzw. App-Anbieter aufwändig direkt befragt werden müssen.

5.2 Authentifikation und Autorisierung

Autorisierung wird benötigt, wenn Funktionen nicht frei zugänglich sein sollen, sondern nur für bestimmte Nutzergruppen. Beispielsweise können bestimmte E-Learning-Angebote nur für Studierende und Mitarbeiter einer Hochschule bestimmt sein. Dies erfordert ein System zur Verwaltung von Nutzern, Rechten und Rollen im Hintergrund, wie es an Hochschulen i.d.R. ohnehin im Einsatz ist. Authentifikation ist zudem für alle

[11] http://www.campus-app.de/
[12] https://www.uni-hohenheim.de/app

Funktionen notwendig, die persönliche Daten verwenden sollen. Ein Beispiel hierfür wäre die Anzeige von erbrachten Prüfungsleistungen, welche nur für den Studierenden selbst zugänglich sein dürfen.

Zur Realisierung von Funktionen, welche über eine allgemeine Informationsvermittlung (wie z.B. ein Mensaplan) hinausgehen, sowie bei Zugriff auf persönliche Daten werden daher Mechanismen für Authentifikation und Autorisierung erforderlich. Hierfür sollte auf vorhandene IT-Systeme (z.B. LDAP-Service, Single-Sign-On oder Hochschul-übergreifende Authentifikation wie z.B. mit Shibboleth[13] bzw. DFN-AAI[14]) aufgebaut werden. Die Hochschul-App muss mit diesem Dienst verbunden sein, damit Nutzer einheitliche Kennwörter verwenden können und ihre Identität mit der in den konventionellen IT-Systemen der Hochschule übereinstimmt.

Der bisherige Funktionsumfang eines Großteils der in diesem Rahmen untersuchten Apps bedingt nicht die Notwendigkeit einer Authentifizierung und Autorisierung, da es sich um allgemein zugängliche Informationen (Mensaplan, News, Räume, Stundenpläne) handelt. Lediglich Apps, die personenbezogene Daten verarbeiten oder personalisierte Daten anbieten, sind auf die Anwendung von Authentifizierungs- und Autorisierungsmechanismen angewiesen. Dies ist beispielsweise bei Diensten im Bereich E-Learning und bei der Information über Leistungen in der Universität erforderlich, die jedoch nur von ca. 10% der Apps offeriert werden. Eine auf Shibboleth basierende Lösung wird dabei teilweise schon eingesetzt [PD14].

5.3 Verstetigung

Die kontinuierliche Pflege einer App stellt besondere fachliche sowie organisatorische Anforderungen an eine Hochschule. Pflegebedarf entsteht insbesondere durch sich ändernde Infrastrukturen, sich wandelnde technische Plattformen hinsichtlich der mobilen Endgeräte sowie die fortwährende Entwicklung der Nutzung moderner Technik. Im folgende wird auf diese Herausforderungen und auf geeignete Lösungskonzepte näher eingegangen.

Infrastrukturen ändern sich z.B. durch die Einführung neuer Systeme oder im ungünstigen Fall sogar bei der Aktualisierung vorhandener Systeme, wenn dabei Schnittstellen betroffen sind. Durch die Kapselung der Schnittstellen aller übergreifend genutzten Systeme kann der Pflegeaufwand hinsichtlich der App wie auch anderer gekoppelter Systeme deutlich reduziert werden [KLZ14]. Für eine solche Kapselung eignen sich Konzepte und Lösungen aus dem Bereich der Service-orientierten Architekturen (SOA), wie z.B. ein Enterprise-Service-Bus (ESB), welcher insbesondere Protokolle, Formate und Abrufadressen verbergen, wandeln und abstrahieren kann. Die Verbreitung derartiger Lösungen ist jedoch noch gering, wie sich an dem bei fast allen untersuchten Apps derzeit noch oft fehlenden Zugriff auf Dienste bestehender Campus-Infrastrukturen ablesen lässt.

[13] http://shibboleth.net/
[14] https://www.aai.dfn.de/der-dienst

Auch die technischen Plattformen der mobilen Endgeräte verändern sich. Neben weiter entwickelter Hardware sind vor allem die relevanten Betriebssysteme der mobilen Endgeräte einem fortwährendem Wandel unterzogen. Auf der einen Seite können neben den beiden wichtigsten Betriebssystemen Android und iOS auch weitere Alternativen an Relevanz gewinnen, wie z.B. derzeit Windows Phone. Weiterhin kann durch die Verbreitung neuerer Versionen der Betriebssysteme auch ein Pflegebedarf hinsichtlich einer App entstehen. Selbst bei grundsätzlicher Kompatibilität neuerer Betriebssysteme hinsichtlich älterer Apps können dennoch aktualisierte Features, wie z.B. neue Interaktionsformen, zu einem Pflegebedarf führen, damit eine App auch zukünftig den gewohnten Bedienparadigmen der Zielplattform folgen kann. Durch den Einsatz von Browser-basierten Interaktionskomponenten wird die Wartbarkeit hinsichtlich der sich ändernden Betriebssysteme deutlich verbessert. Browser-basierte Komponenten können entweder mittels dynamischem HTML unmittelbar in eingebetteten Browsern laufen oder mittels Frameworks zur Betriebssystem-übergreifenden App-Entwicklung, wie z.B. phonegap, so mit Betriebssystem-unabhängigen Funktionen erweitert werden, dass komplexere Funktionen unterstützt werden. Im Vergleich zur Entwicklung nativer Apps für jedes unterstützte Betriebssystem kann damit der Entwicklungs- und Pflegeaufwand deutlich reduziert werden. Jedoch werden solche Betriebssystem-unabhängigen Apps nicht immer den Interaktionsparadigmen der jeweiligen Plattform vollständig genügen und können daher Kompromisse bzgl. des Nutzerkomforts mit sich bringen.

Die Nutzung moderner Technik entwickelt sich fortlaufend. So erhielten mobile Endgeräte in den letzten Jahren z.B. neuartige Sensoren, wurden durch stromsparende Technik und verbesserte Akkus zunehmend mobiler, und mobiles Internet nimmt in seiner Netzabdeckung immer weiter zu. All dies hat Rückwirkungen auf die Art, wie mobile Endgeräte von den Menschen genutzt und verstanden werden. Neben wandelnden Nutzungsverhalten durch die Allgegenwärtigkeit des Internets seien hier als Beispiele Touch-Gesten und Spracheingaben genannt, welche die üblichen Interaktionsformen maßgeblich beeinflusst haben. Wandlungen dieser Art haben großen Einfluss auf die Software, welche auf den mobilen Endgeräten zum Einsatz kommt. Die Wartbarkeit und Anpassbarkeit mobiler Apps hinsichtlich wandelnder Nutzung kann durch Software-Architekturen unterstützt werden, welche Interaktionsformen von Logik und Datenhaltung entkoppelt.

Ein weiterer Aspekt der sich wandelnden Nutzung betrifft den unterstützten Funktionsumfang. Derzeit wird z.B. der Bedarf an mobil nutzbaren Lernwerkzeugen als gering eingeschätzt. Zukünftig könnten Geräte mit größeren Displays, wie z.B. Tablets, eine stärkere Verbreitung finden, wodurch auch sich auch der Bedarf an mobil nutzbaren Lernwerkzeugen verändern könnte. Ein modularer Aufbau einer App wird die Anpassbarkeit bzgl. des Umfangsumfangs maßgeblich vereinfachen. Dieser Ansatz wird aber von der Mehrheit der untersuchten Apps derzeit noch nicht verfolgt.

Aufgrund des bisher eher geringen Funktionsumfangs der meisten Apps und der marginalen technischen Komplexität der angebotenen Funktionen ist der Wartungs- und Pflegeaufwand sowie die Umsetzung der Updates und Weiterentwicklung für die Apps noch nicht erkennbar. Die bisherigen Aktualisierungszyklen erlauben hierfür noch keine

verlässliche Aussage. Eine Einschätzung des Entwicklungs- und Pflegebedarfs wird sich nach einer Reifung des Markts und einer ersten Sättigung und Stabilisierung der Nutzerzahlen durchführen lassen.

5.4 Ressourcenbedarf

Mobile Endgeräte verfügen bereits über oft beeindruckende Rechenleistung, Akku-Kapazität und Speicherplatz. Eine App für die Nutzung im Hochschulumfeld muss dennoch unterschiedlichsten Hardwaregegebenheiten gerecht werden und sollte daher sparsam mit diesen Ressourcen umgehen. Da die untersuchten Apps überwiegend einen sehr kleinen und bzgl. des Ressourcenbedarfs unkritischen Funktionsumfang besitzen, ist dieser Faktor derzeit zu vernachlässigen.

Weiterhin steht vielen Nutzern mobiles Internet nicht oder nur mit eingeschränkter Bandbreite oder Performance zur Verfügung. Aus diesem Grund sollte eine Hochschul-App auch einen offline-Modus anbieten, bei dem dynamische Daten zuvor über WLAN geladen wurden. Diese Funktion wird teilweise von den untersuchten Apps angeboten (beispielsweise von der App myHHU[15]).

6 Organisatorische Analyse

Die untersuchten Apps sind in aller Regel für den Nutzer kostenfrei. Als einzige Ausnahme fand sich die App "BibAlarm" an der Uni Münster, welche jedoch von einem Drittanbieter bereitgestellt wird. Die Finanzierung einer Hochschul-App durch Nutzergebühren scheint derzeit also nicht vielversprechend.

Viele Hochschul-Apps sind derzeit offenbar durch private Initiative entstanden oder im Rahmen von studentischen Projekten auf den Weg gebracht worden. Einige Account-Inhaber der App-Stores, über die die untersuchten Hochschul-Apps vertrieben werden, sind Mitarbeiter aus zentralen Einrichtungen oder aus dem wissenschaftlichen Bereich. Dieser dezentrale Entstehungshintergrund erklärt auch die durchschnittlich geringe und einseitige Funktionsabdeckung. Entsprechend sind Design, Professionalität und plattformübergreifende Kompatibilität oft zu bemängeln bzw. uneinheitlich. Nur wenige Apps wurden durch offizielle Stellen der Hochschulen entwickelt oder zumindest im Store veröffentlicht, insbesondere tub2go[16], ovgu2go[17], myUDE[18], eStudentLBS[19], Uni Mannheim[20] und Uni Tübingen[21].

[15] http://www.uni-duesseldorf.de/home/shortcuts/App
[16] http://tub2go.tu-berlin.de/
[17] http://www.uniapp.ovgu.de/
[18] https://www.uni-due.de/myude/
[19] https://play.google.com/store/apps/details?id=de.uni.bremen.estudent.lbs
[20] https://play.google.com/store/apps/details?id=de.klm_mobilesolutions.unima
[21] https://play.google.com/store/apps/details?id=de.uni_tuebingen

Die Veröffentlichung einer eigenen App ist für eine Hochschule in vielerlei Hinsicht schwierig. Neben der Finanzierung von Entwicklung, Pflege und Betrieb ist die häufig quer zu anderen IT-Projekten (z.B. Campus Management, E-Learning, Pressearbeit, Studentenwerk) liegende Koordination zu berücksichtigen. Weiterhin müssen die Apps über die offiziellen Plattformen (in sogenannten App-Stores) angeboten werden. Hierfür müssen personenbezogene Accounts und auf Kreditkarten beschränkte Abrechnungen verwendet werden, die in der öffentlichen Verwaltung nicht immer einfach abzuwickeln sind. Je nach den lokalen Gegebenheiten kann die Transferstelle einer Hochschule hierbei behilflich sein, wie z.B. bei der App der Uni Mannheim. Negativ zu erwähnen sind v.a. der überkomplexe Registrierungsprozess und die jährlichen Gebühren des iOS-Shops, die (im Unterschied zu der einmaligen und geringeren Gebühr sowie der zügigen Abwicklung bei Android) den Ablauf erschweren.

Die Gewichtung der dauerhaften Pflege sollte gegenüber den anfänglichen Entwicklungskosten nicht unterschätzt werden. Es muss nicht nur eine Lösung zur Abdeckung des initialen Aufwands gefunden werden, sondern auch die Aufwände der dauerhaften Betreuung müssen getragen werden.

7 Zusammenfassung

Die Planung zur Entwicklung einer Hochschul-App muss zunächst inhaltlich die Rahmen der zu unterstützenden Funktionsbereiche abstecken; dabei können die von bestehenden Apps angebotenen Funktionen eine Orientierungshilfe bilden. – Technisch sind besonders die Aspekte der Plattform-spezifischen gegenüber der Plattform-unabhängigen Softwareentwicklung bzw. der damit verbundenen Entwicklungs-umgebungen zu berücksichtigen. Android und iOS als die beiden am meisten vertretenen Plattformen weisen kaum noch Unterschiede in der Konzeption der untersuchten Hochschul-Apps auf, sodass in die Entscheidung außer finanziellen Erwägungen auch die vorhandenen Kompetenzen der Entwickler einfließen können. Außerdem sind technische Lösungen für verstetigte Schnittstellen zu Standardsoftware sowie für Authentifikation und Autorisierung notwendig, um die App mit den vorhandenen IT-Systemen der Hochschule zu koppeln. – Aus organisatorischer Sicht müssen neben der Finanzierung und Projektplanung/-steuerung vor allem die Veröffentlichung, der Betrieb und die dauerhafte Pflege der App abgedeckt werden. Die dauerhafte Betreuung kann ggf. durch Personal mit speziellem KnowHow im Rechenzentrum übernommen werden, was eine Erweiterung oder Umschichtung des Personalbudgets notwendig machen kann.

Die zu diesen Fragen in dem vorliegenden Beitrag dargestellten Einzelfragen und Realisierungswege können als Entscheidungshilfe für die Gestaltung einer Hochschul-App genutzt werden. Die Kriterien basieren auf dem bekannten Stand der Forschung, der anhand vorhandener Hochschul-Apps zu einer Dokumentation des aktuellen Stands der Technik zusammen gefasst und konkretisiert wurde. Für eine weiterführende Analyse könnten zudem die Entwickler bzw. Anbieter der untersuchten Apps direkt befragt werden. Aus den im vorliegenden Beitrag präsentierten Untersuchungsergebnissen können dafür wesentliche Eckpfeiler zur Konstruktion eines geeigneten Fragebogens abgeleitet werden. Eine derartige Befragung könnte zudem auf die technischen

Verantwortlichen (RZ-Leiter/CIOs) aller deutschen Hochschulen ausgeweitet werden, um auch Apps zu erfassen die noch in der Entwicklung sind oder bereits auf anderen Wegen vertrieben werden. Zusammen mit der hier vorgelegten Analyse könnten derartige Informationen zu einem Leitfaden für App-Entwicklungsprojekte an Hochschulen weiterentwickelt werden.

8 Literaturverzeichnis

[Be11] Beaucamp, G.: Im Bann des Bildschirms. In: Forschung und Lehre, Nr. 18, Juni 2011, S. 448-449.

[BL13] Buck, D.; Lübbe, S.: Forschungsmanagement – eine Querschnittsaufgabe. In: Wissenschaftsmanagement: Zeitschrift für Innovation, 19(1), 2013, S. 13.

[Bi13] Bias, A.: Campus mobil mit Hochschul-Apps. Vortrag auf der 11. Tagung der DFN-Nutzergruppe Hochschulverwaltung, Mannheim, Mai 2013. online verfügbar unter www.hochschulverwaltung.de/tagung/2013/folien/DFN-820_Bias_Hochschul-Apps.pdf

[Bü+11] Bührig, J.; Guhr, N.; Breitner, M.H.: Technologieakzeptanz mobiler Applikationen für Campus-Management-Systeme. In: Proc. 41. Jahrestagung der Gesellschaft für Informatik, Bonn : Köllen, September 2011, S. 461.

[Ca+13] Callewaert, P.; Combes, C.; Choukhman, A.; van der Heijden, B.; Joosten, B.; De Jaeger, T.: CIONET Mobility Survey - Key takeaways to advance in mobile. Deloitte, 2013. online verfügbar unter http://www.cionet.com/Data/files/groups/Mobileweb.pdf

[Co+14] Cozza, R.; Durand, I.; Gupta, A.: Market Share: Ultramobiles by Region, OS and Form Factor. Gartner, 2014.

[He+09] von der Heyde, M.; Hotzel, H.; Olbrich, S.; Stenzel, H.; Weckmann, H.-D.; Wimmer, M.: Strukturkonzepte für die Informations- und Kommunikationsversorgung von Hochschulen. In: Praxis der Informationsverarbeitung und Kommunikation 32(3), 2009, S. 167-173.

[Ka+13] Kaiser, S.; Kuhnt, E.; Lemcke, S.; Lucke, U.: Web-basierte Beschaffung. In: Proc. 43. Jahrestagung der Gesellschaft für Informatik, Bonn : Köllen, September 2013, S. 308-319.

[Ke13] Keil, R.: Contextualising Change. International Innovation. In: ResearchMedia.EU, 2013, S. 42-44.

[KLZ14] Kiy, A.; Lucke, U.; Zoerner, D.: An Adaptive Personal Learning Environment Architecture. In: Proc. Architecture of Computing Systems (ARCS 2014), Berlin : Springer, März 2014, S. 60-71.

[Ki+14] Kiy, A.; Grünewald, F.; Zoerner, D.; Lucke, U.: Ein Hochschul-App Framework: Hybrid und modular. In: Proc. Die 12. e-Learning Fachtagung der Gesellschaft für Informatik (DeLFI 2014), Bonn : Köllen, im Druck.

[LS12] Lucke, U.; Schroeder, U.: Forschungsherausforderungen des E-Learning. In: i-com 11(1), Februar 2012, S. 1-2.

[Mr13] Mroz, R.: App-Marketing für iPhone und Android. MIT Press, 2013.

[PD14] Politze, M.; Decker, B.: Shibbolet & OAuth2 für die Autorisierung von Apps am Beispiel der RWTHApp. Vortrag auf der 60. DFN Betriebstagung, Berlin, März 2014. online verfügbar unter www.dfn.de/fileadmin/3Beratung/Betriebstagungen/bt60/1-aai_Marius_Politze-App.pdf

[Pe+12] Pérez-Sanagustín, M.; Ramírez González, G.; Hernández Leo, D.; Muñoz Organero, M.; Santos, P.; Blat, J.; Delgado Kloos, C.: Discovering the campus together: A mobile and computer-based learning experience. In: Network and Computer Applications, 35(1), 2012, S. 176-188.

[We13] Weddeling, B.: Campus to go: die wichtigsten Wissens-Apps. In: Focus, Nr. 12, März 2013, S. 98.

[Wi+14] Wilson, G.; Aruliah D.A.; Brown C.T.; Chue Hong N.P.; Davis M.; et al.: Best Practices for Scientific Computing. In: PLoS Biology, Vol. 12, Nr. 1, 2014.

[Za+12] Zakhariya, H.; Kosch, L.; Breitner M.H.: Elektronische Drittmittelakte in der Hochschulverwaltung – Erkenntnisse aus Fallstudien. In: Proc. 42. Jahrestagung der Gesellschaft für Informatik, Bonn : Köllen, September 2012, S. 613-626.

[Ze+14] Zender, R.; Metzler, R.; Lucke, U.: FreshUP - A Pervasive Educational Game for Freshmen. In: Pervasive and Mobile Computing, Elsevier, 2014 (im Druck).

Get Your Study Plan

Patrick Bittner, Christopher Ritter, Thomas Hildmann

tubIT – IT-Service-Center der TU-Berlin
Technische Universität Berlin
Einsteinufer 17
10587 Berlin
[vorname].[nachname]@tu-berlin.de

Abstract: Das moderne Studium stellt die Studierenden vor inhaltliche, organisatorische und technische Herausforderungen. Zur Unterstützung und Begleitung soll daher an der TU Berlin eine Plattform geschaffen werden, die eine geeignete, vorkonfigurierte Softwareinfrastruktur zur Nutzung von digitalen Lerninhalten bereitstellt. Ein Recommender-System soll dabei unterstützen, die richtigen Module, Vorlesungen und die hierfür geeigneten Materialien und Softwarewerkzeuge auszuwählen. Dafür kann sowohl auf Empfehlungen der Dozierenden als auch von Studierenden älterer Semester und auf gesammelte Metadaten des Studierenden zurückgegriffen und mit den Anforderungen aus der geltenden Studienprüfungsordnung abgeglichen werden. Im Laufe der Nutzung dieser Plattform ist es nicht nur möglich, immer bessere Vorschläge für den weiteren Verlauf zu erstellen, sondern auch einen Überblick über den Studienfortschritt und mögliche Optionen für den weiteren Verlauf zu veranschaulichen.

1 Einleitung

An den Universitäten ist IT-Unterstützung heute nicht mehr wegzudenken. Den Studierenden werden unterschiedliche Dienste angeboten, wie beispielsweise Online-Bewerbungen oder E-Mail-Accounts. Diese Dienste werden zum Teil stark genutzt. Allerdings bringen viele Studierenden bereits zum Studienanfang ihre eigene IT-Umgebung mit und verknüpfen die universitären Dienste zum Teil lediglich mit ihrer eigenen Struktur. Auch diese Entwicklung wird zunehmend von den Universitäten gefördert, um den Studierenden ein größtmögliches Maß an Komfort zu bieten. Im späteren Verlauf wird kurz der aktuelle Ansatz der TU Berlin hierzu dargestellt, da dieser die Grundlage für das hier beschriebene Verfahren bildet. Eine ausführlichere Beschreibung kann unter [RBK14] gefunden werden.

Trotzdem ist der Verwaltungsaufwand während eines Studiums groß. Unterschiedliche Studienordnungen schaffen Anforderungen, an die sich die Studierenden der unterschiedlichen Fachrichtungen zu halten haben. Module müssen gesichtet und einzelne Vorlesungen ausgewählt werden, um die nötigen Anforderungen letztlich erfüllen zu können. Dabei ist es nicht immer einfach, die genauen Anforderungen der

Prüfungsordnung oder auch nur eines bestimmten Moduls zu durchblicken. Bisher existieren wenige Systeme, die es den Studierenden leichter machen, herauszufinden, welche Module von ihnen noch zu belegen sind, und welche Voraussetzungen dafür mitzubringen sind. Ebenso wenig wird ihnen mitgeteilt, was ihnen überhaupt noch zum Abschluss seines Studiums fehlt.

Die Entwicklung eines solchen Systems würde die Selbstverwaltung von Studierenden erheblich vereinfachen. Dabei könnte nicht nur auf Studienordnungen und Modulhandbücher zurückgegriffen werden: Bezieht man die Meinungen anderer Studierender mit ein, welche einen ähnlichen Hintergrund aufweisen, ermöglicht man es den Studierenden, sehr viel gezielter Module auszuwählen, die ihnen fehlen und die für ihre Anforderungen und ihre Interessen passend sind.

In dieser Arbeit werden wir zuerst bisherige Ansätze vorstellen, welche versucht haben, die Studierenden bei der Selbstverwaltung zu unterstützen. Anschließend stellen wir ein System vor, mit dem die Studierenden bei der Modulwahl und der Übersicht über ihr eigenes Studium unterstützt und welches dabei Informationen von anderen Studierenden berücksichtigt.

Abschließend geben wir einen Ausblick über mögliche, weitere Ausbaumöglichkeiten dieses Systems.

2 Related Work

Bisherige Veröffentlichungen zum Thema Campus-Management beziehen sich vor allem auf die Einführung und den Aufbau von Campus-Management Systemen.

In [FH09] wird die Einführung des Campus Management Systems in der Universität Hamburg beschrieben. Dieses beinhaltet auch ein System, welches die Modulplanung an der Universität optimiert, um so Modulüberschneidungen, vor allem bei Kombinationsstudiengängen, zu minimieren. Dies ist eine deutliche Verbesserung für die Studierenden, da ihnen auf diese Weise ermöglicht wird, ihr Studium in der Regelstudienzeit zu beenden. Eine automatische Empfehlung an die Studierenden, welche Module für sie sinnvoll sind, ist jedoch nicht integriert.

[RHHL09] geht kurz darauf ein, dass in Ihrem System ein „Self Monitoring“ möglich ist, welches es den Studierenden ermöglicht, sich selbst besser einschätzen zu können. Dabei wird dem Studenten im Wesentlichen dargestellt, wie ihre erhaltene Note sich im Vergleich zu den Noten der anderen Studierenden präsentiert.

Ein „Self Assessment“ zur besseren Einschätzung der eigenen Stärken und Schwächen berücksichtigt den bisherigen Studienverlauf. Ein Ansatz, der auch von uns aufgegriffen wird. Allerdings ist der Verwendungszweck bei uns nicht die Darstellung der Stärken der Studierenden, sondern die Empfehlung, welche Module für sie sinnvoll erscheinen könnten.

In [Ra09] wird u.a. darauf eingegangen, dass eine personalisierte, konfigurierbare Benutzerumgebung für den Studierenden zur Verfügung gestellt wird. Diese Basis erscheint sehr komfortabel, geht allerdings nicht so weit wie der tubIT-Live Ansatz der TU Berlin, der ein komplettes, personalisiertes OS-Image liefert, und den Studierenden auf diese Weise auch eine Integration von Software in ihre universitäre Benutzerumgebung zu integrieren ermöglicht, ohne dabei ihr eigenes System neu konfigurieren zu müssen.

Der von uns im Folgenden dargestellte Ansatz wurde in dieser Form nach unserem Kenntnisstand bisher nicht näher betrachtet.

3 Get Your Study Plan

An vielen Universitäten wird das Wissen, welche Module zu belegen sind und welche sich sinnvoll miteinander kombinieren lassen, von den höheren Semestern an die neuen Studierenden weitergegeben. Oft geschieht das mündlich, wodurch jeder einzelne Studierende nur von einem Bruchteil der Erfahrungen älterer Studierender profitieren kann.
Sich ändernde Studienordnungen erschweren diese Kommunikation zusätzlich und oft ist nicht auf den ersten Blick ersichtlich, welche Module die Anforderungen des einzelnen Studiengangs erfüllen.

Get Your Study Plan setzt an dieser Stelle auf, um den Studierenden zu helfen, ihr Studium selbst zu organisieren. Aufbauend auf den vorliegenden Informationen zu Modulen und Prüfungsordnungen sowie Informationen über die bisherigen Leistungen der einzelnen Studierenden, kann ein System entwickelt werden, welches den bisherigen Stand des Studiums darstellen und offene Bereiche identifizieren kann.

3.1 Planmodule

Zur Übersicht über den eigenen Studienstand und noch zu belegende Module müssen die Informationen über die Studienordnung selbst sowie die bereits belegten Module vorliegen. Die Studienordnungen legen fest, wie viele Module in welcher Kombination belegt werden können, um der Studienordnung gerecht zu werden.
Dabei gibt es Module, die belegt werden müssen, jene, die belegt werden können, und Modulbereiche, in denen einige Module aus einer Auswahl gewählt werden können.

Die Wahlbereiche beinhalten hierbei eine Auswahl an Modulen und/oder weiteren Wahlbereichen, aus denen gewählt werden kann. Den Modulen wird eine Anzahl an Credits zugewiesen. Die Wahlbereiche können eine variable Anzahl an Credits aufweisen. So kann ein Wahlbereich dafür stehen, dass Module (und Wahlbereiche) aus dem Wahlbereich ausgewählt werden müssen, die in Summe eine Creditzahl beinhalten, welche in einem bestimmten Bereich liegt. Die Menge aller Anforderungen, die für einen Studiengang erfüllt werden müssen kann wie in Abbildung 1 vereinfacht als Regeln dargestellt werden.

Erreiche insgesamt 120 Punkte aus Bereich „alle Module“

1. Erreiche 30-30 Punkte aus Bereich „Masterthesis“
2. Erreiche 54-60 Punkte aus Bereich „Informatikmodule“
3. Erfülle 1-1 Regel aus den Regeln a-d
    a. Erreiche 30-60 Punkte aus Bereich „System Engineering“
    b. Erreiche 30-60 Punkte aus Bereich „Verlässliche Systeme“
    c. Erreiche 30-60 Punkte aus Bereich „Intelligente Systeme“
    d. Erreiche 30-60 Punkte aus Bereich „Kommunikationsbasierte Systeme“
4. Erreiche 30-36 Punkte aus dem Bereich „Anwendungsfächer und Sonstige“
5. Erfülle 1-1 Regel aus den Regeln a…
    a. Erreiche 18-24 Punkte aus Bereich „Elektrotechnik“
    b. Erreiche 18-24 Punkte aus Bereich „Mathematik“
    c. Erreiche 18-24 Punkte aus Bereich „[weiteres Anwendungsfach]“
    d. ..
6. Erreiche 1-beliebig Punkte aus Bereich „Seminare“
7. Erreiche 1-beliebig Punkte aus Bereich „Projekte“

Abbildung 1: Darstellung eines Studienverlaufs
am Beispiel des Studienfachs Informatik Master

Selbstverständlich ist hierbei darauf zu achten, dass keine Kreisschlüsse entstehen. Des Weiteren sollten Module nicht in mehreren, sich nicht ausschließenden Wahlbereichen vorkommen.

3.2 Eigene Einschränkungen

Die „Planmodule“ geben die natürliche Einschränkung des Studiums vor. Sie bestimmen, welche Modulkombinationen überhaupt möglich sind und in welchem Umfang sie vorliegen müssen.

Es wird zwar berücksichtigt, welche Studienrichtung betrachtet wird (abhängig von der Studienordnung), aber es wird nicht berücksichtigt, welche Module passend sind. Normalerweise gibt es keine Beschränkung, in welcher Reihenfolge Module belegt werden können. Allerdings ist nicht jede Reihenfolge sinnvoll und nicht jedes Modul ist für jedes Studienfach gleich gut geeignet.

Um dem Rechnung zu tragen, kann ein Recommender System verwendet werden, welches den Studenten Vorschläge liefert, welche Module für sie sinnvoll sein könnten.

Dabei können 3 unterschiedliche Herangehensweisen unterschieden werden (Abbildung 2):

1. Dozentengesteuerte Empfehlung
2. Metadaten gesteuerte Empfehlung und
3. Studentengetriebene Empfehlung

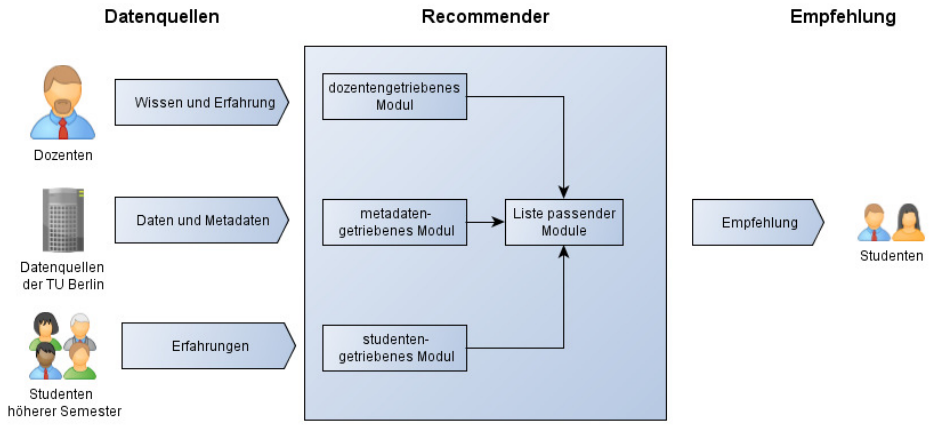


Abbildung 2: Darstellung der unterschiedlichen Herangehensweisen für das Recommender System

Gemeint ist hierbei die Herkunft der Empfehlung, also der Grund, warum gerade ein bestimmtes Modul empfohlen wird.

Bei der dozentengesteuerten Empfehlung ist der Ursprung direkt die Empfehlung eines verantwortlichen Dozenten.

Dabei wird auf Metakeys zurückgegriffen:

Es werden bestimmte Schlüssel (Metakeys) ausgewählt wie z.B. Semesterzahl, Studiengang und ähnliche. Auch die Prüfung, ob ein anderes Modul bereits (erfolgreich) belegt wurde, wäre denkbar. Dies sind Informationen über die Studierenden, die jederzeit vom System abrufbar sind. (Ein Abruf seitens der Dozenten ist aus datenschutzrechtlichen Gründen selbstverständlich nicht möglich). Auf dieser Basis kann der Dozent festlegen, dass ein bestimmtes Modul für Studierende eines bestimmten Studiengangs empfehlenswert ist. Über Kombinationen oder Bündel von solchen Einschränkungen lässt sich dies feiner eingrenzen.

So kann z.B. markiert werden, dass ein Modul für Studierende des Studiengangs BWL im 2. bis zum 4. Semester interessant sein könnte.

Metadatengesteuerte Empfehlung basiert auf den „Erfahrungen“ des Systems mit dem Studenten. Auf der Basis vom Verhalten, das die Studierenden in ihrem bisherigen

Studienleben an den Tag gelegt haben, wird ermittelt, welche Module für den Studenten passend wären. Hierbei wird z.B. ausgewertet, welche Module häufig mit welchen anderen kombiniert werden.

Hat ein Studierender bereits ein oder mehrere dieser Kombinationsfächer gewählt, werden ihm auch die anderen Module, die häufig damit assoziiert werden, empfohlen.

Analog dazu kann z.B. über Bewertungen in bestimmten Modulen oder Auslandssemester bestimmte Modulvorschläge ermittelt werden.

Diese Art der Empfehlung ist die individuellste, da sie direkt auf den Entscheidungen/Leistungen des Studierenden basiert. Allerdings benötigt diese Art der Empfehlung auch den größten Aufwand, da die individuellen Daten vorher gesammelt werden müssen, um so auch neue Entscheidungsmöglichkeiten ermitteln zu können.

Die studentengetriebene Empfehlung basiert auf den Erfahrungen und Informationen, die ältere Semester direkt ins System eingeben. Diese Form der Empfehlung ist der ursprünglichen Weitergabe von Informationen von Studierenden höherer Semester an neuere Studierende am nächsten. Allerdings sind die Informationen in diesem Fall allen neueren Studenten zugänglich, und nicht nur denen, denen die Informationen direkt weitergegeben werden.
Diese Art der Informationsweitergabe kann auch als eine Art Erweiterung der anderen beiden Arten verstanden werden: Studierende könnten wie Dozenten Empfehlungen abgeben, basierend auf den Metadaten der anderen Studierenden.
Eine andere Möglichkeit wäre es, das Studierende bestimmte Modulkombinationen vorschlagen, Reihenfolgen von Fächern empfehlen oder bestimmte Leistungen in anderen Fächern als hilfreich einstufen, so wie es auch der metadatengesteuerte Ansatz vorsieht.

4 Integrationsansätze innerhalb der TU Berlin

Das skizzierte Konzept beschreibt, wie die Modulwahl der Studierenden um Einschränkungen und Vorschläge angereichert werden kann. Allerdings ist es dafür unabdingbar, dass die Studierenden überhaupt in die Lage versetzt werden, diese Informationen auch zu erhalten und ggf. selbst auszuwerten.

In der TU Berlin wird das Konzept des Get Your Study Plan bereits in Ansätzen realisiert. Die aktuelle Lösung setzt dabei auf zwei, bereits an der TU-Berlin entwickelten Systemen auf. Zum Einen das dort entwickelte hochschulweite Identitäts-managementsystem (IdM) TUBIS und zum Anderen das ebenfalls noch am Anfang stehende tubIT-Live Projekt.

4.1 Hochschulweites IdM

Um jede Studierende mit Empfehlungen versorgen zu können ist es unabdingbar, dass jedes Mitglied der Universität eindeutig und mit allen Kontexten wie etwa Status

(Studierender, Mitarbeiter), Rolle (Dekan, FG-Leiter, Abteilungsleiter, …) oder Studiengang, bekannt ist. Dies ist auch notwendig, damit ein möglichst einfacher, einheitlicher Zugang zu allen für sie relevanten Diensten möglich wird; und dies gleichgültig, ob sich die Person im Hörsaal, am Arbeitsplatz, zu Hause oder auf einer Dienstreise befindet.

Die TU-Berlin hat zur Erfüllung der Aufgaben eines Identitätsmanagementsystems das universitätsweite, rollenbasierte Autorisierungssystem TUBIS entwickelt [HKR08, HR07]. Dieses integriert sich in die bestehende Infrastruktur der Campus-managementsysteme und überwacht die Identitäten dabei während des gesamten Lebenszyklus zwischen Eintritt uns Ausscheiden.

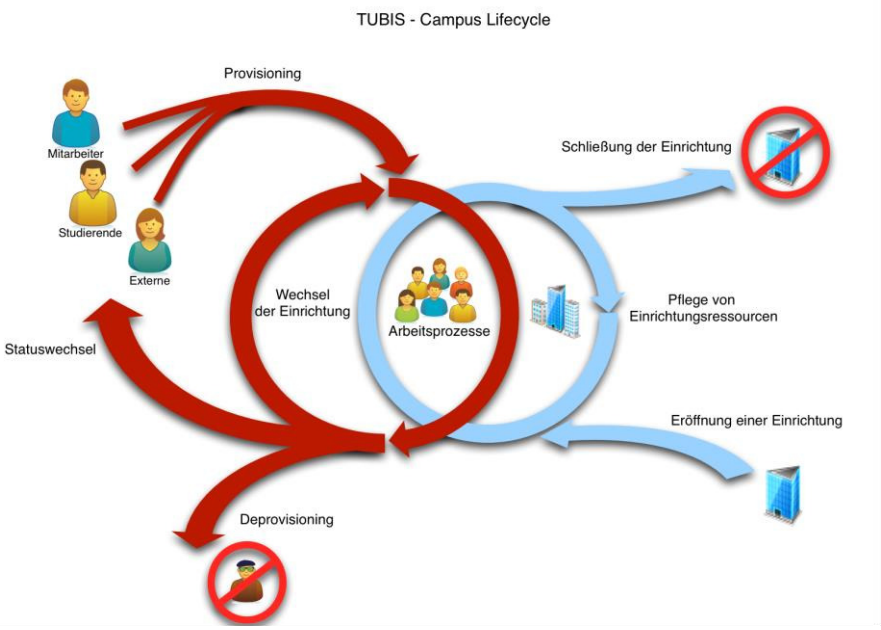


Abbildung 3. Lebenszyklus von Identitäten und Einrichtungen

Wie in Abbildung 3 als Übersicht skizziert wurde, werden dabei sowohl die Lebenszyklen der Personen, als auch die der Einrichtungen selbst berücksichtigt. Das IdM bietet den jeweiligen Einrichtungen die Möglichkeit der dezentralen Rechteverwaltung [RHK10]. Durch die direkte Anbindung an die Systeme der Studierendenverwaltung ist eine optimale Voraussetzung zur Auswertung von Metadaten für das Empfehlungssystem gegeben. Wechsel im Studiengang, des geplanten Abschluss oder auch Urlaubssemester werden sofort registriert und können für weitere Empfehlungen berücksichtigt werden. Auch die für spätere Versionen geplante Auswertung von erbrachten Studienleistungen ist technisch damit bereits möglich.

Fragen des Datenschutzes wurden bei der Konzeption des rollenbasierten IdM bereits ausführlich behandelt [GH00, HW05]. Eingebettet in das auf diesen Prinzipien

entwickelte Portal [HR07] wird die hier beschriebene Anwendung nun weiterentwickelt. Die Sichtbarkeit von Informationen und die Verteilung von Rechten müssen im Rahmen des Anwendungsdesigns genauer spezifiziert werden.

4.2 tubIT-Live

Während die Generationen vor den Digital Natives zum Zeitpunkt der Immatrikulation nur teilweise über eine eigene IT-Ausstattung verfügten und von der Hochschule entsprechend mit Hardware und IT-Diensten versorgt werden mussten, führte die stetig wachsende Zahl an Studierenden mit vorhandener IT-Ausstattung zum Paradigma des Bring Your Own Device (BYOD) und den damit verbundenen Schwierigkeiten der Einbindung in die vorhandenen IT-Infrastrukturen der Universität. Mit der Zunahme an mobilen Endgeräten sowie der kostengünstigen und flächendeckenden Verfügbarkeit unterschiedlichster IT-Dienste außerhalb der universitären Infrastruktur wurde das BYOD Paradigma zunehmend zum BYOE (Bring Your Own Environment) Paradigma. [RBK14] beschreibt dabei ein System, wie die Studierenden mit einer auf ihr Studium zugeschnittenen Umgebung ausgestattet werden können, ohne ihre bereits vorhandene, gewohnte Umgebung dabei aufgeben zu müssen.

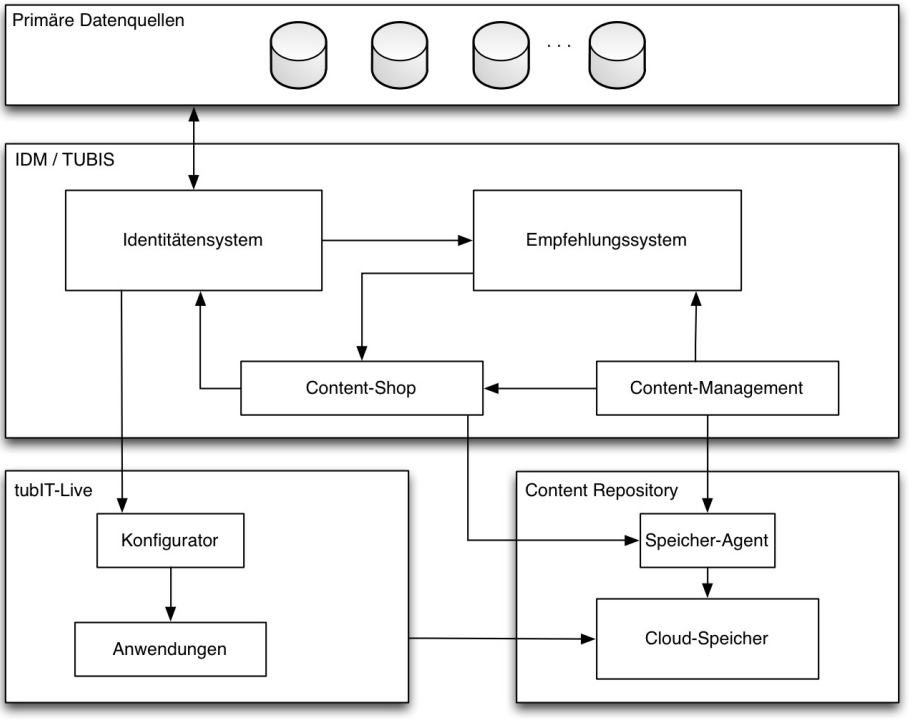


Abbildung 4: Framework zum Get Your Study Environment (GYSE)

An der TU-Berlin wird dieses System (Abbildung 4) im Rahmen des tubIT-Live Projektes realisiert. In der derzeitigen Entwicklungsphase handelt es sich dabei in erster Linie um ein personalisiertes OS-Image. Die Bereitstellung eines OS-Images ermöglicht es den Studierenden ein komplettes System in ihre eigene Umgebung zu integrieren, ohne selbst tiefgreifende Änderungen am eigenen System vornehmen zu müssen. Sie können das OS-Image auf ihren eigenen PCs oder Mobilen Endgeräten wie Laptops abspielen, und so auf ihre studienbezogene Umgebung zugreifen, ohne Kompatibilitätsprobleme befürchten zu müssen. Damit stellt das OS Image eine Erweiterung des BYOE (Bring Your Own Environment) Paradigmas dar. Das personalisierte Image kann dabei entweder auf einem USB-Stick oder direkt im persönlichen Cloud-Speicher Bereich abgelegt werden.

Durch den Betrieb eines Cloud-Speichersystems und deren nahtlose Integration in die personalisierten OS-Images wird dem Studierenden bereits die Möglichkeit geboten selbstständig Inhalte zu allen Modulen auszuwählen, um Zugriff auf diese zu erhalten.

Dabei wird bereits von einem Recommender System vorgeschlagen, welche Inhalte für den Studierenden relevant sein könnten. Das System ist relativ neu und berücksichtigt bisher ausschließlich einige dozentengetriebene Empfehlungen. Die Integration der metadatengesteuerten als auch der Studentengetriebenen Empfehlung ist jedoch geplant, die dozentengetriebene Empfehlung wird schrittweise weiter ausgebaut.

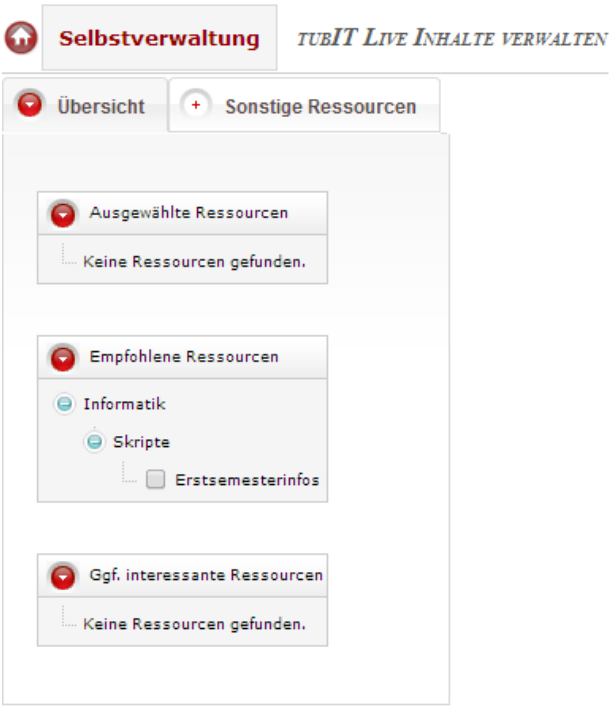


Abbildung 5: Screenshot aus dem Content-Shop für tubIT-LIVE

Die Auswahl der Inhalte (respektive Module) erfolgt hierbei über einen Content Shop, an welchem das Recommender System angeschlossen ist. Hier können die Module wie in einem Online-Shop ausgewählt werden. Die Auswahl eines Inhaltes im Content-Shop sorgt u.a. dafür, dass über Agenten der betreffende Content aus einem Clowd-Speicher mit dem Studenten geshared, und somit direkt im tubIT-LIVE Image integriert wird. Die Abwahl eines bereits gewählten Inhalts entfernt den Share entsprechend für den Studierenden.

Abbildung 5 zeigt den Auswahlbildschirm, in dem die gewählten Inhalte/Module gewählt werden können. Es ist auch zu erkennen, dass die Empfehlungen gänzlich ignoriert werden kann, um sich einen völlig eigenständigen Studienverlauf mit den gewünschten Inhalte zusammenzustellen (der Reiter „Sonstige Ressourcen“).

An dieser Stelle greift die Erweiterung zum „get your study plan“. Neben der Empfehlung von Inhalten wird der Content-Shop um die Empfehlung von Modulen erweitert:

Entscheidet ein Dozent, dass sein Modul M1 für Studierende relevant ist, die

1. Im zweiten Semester Informatik studieren oder
2. Mathematik studieren oder
3. das vorangegangene Modul M0 bereits absolviert haben,

ergibt sich folgende Situation:

Studentin A belegt in ihrem 3. Semester ihres Maschinenbau Studiums das Modul M0. Während sie das Modul noch nicht absolviert hat, findet sie Modul M1 ausschließlich im allgemeinen Bereich, in dem alle Module zur Verfügung stehen. Sobald sie jedoch erfolgreich das Modul M0 abgeschlossen hat, erfüllt sie die Voraussetzung 3, wodurch ihr Modul M1 auch als Empfehlung angezeigt wird.

5 Ausblick

Das vorgestellte System ist als solches noch nicht vollständig implementiert. Derzeit werden Empfehlungssystem sowie der Content-Shop und das Content-Management als Erweiterungen des bestehenden Identitymanagementsystems entwickelt.

Als Content Repositroy wird das seit Ende 2013 betriebene ownCloud der TU-berlin genutzt. Dabei wird die Kommunikation über Agenten realisiert. Das erzeugte Image steht aktuell sowohl als bootbarer USB-Stick zur Verfügung, als auch als in einer virtuellen Maschine ausführbaren Version. Als Betriebssystem dient derzeit ein Ubuntu Derivat.

Diese erste Version des Systems dient insbesondere durch die Integration in das bestehende IdM mit ca. 40.000 Nutzern zugleich als Basis zum Sammeln der Metadaten sowie zur Erhebung der gewünschten bzw. notwendigen Funktionalitäten und deren

Priorisierung. Um eine tatsächliche Umsetzung zu ermöglichen, muss sichergestellt werden, dass die Daten, auf denen aufgebaut wird, vorhanden und auch verlässlich sind. Ohne eine Übersicht über die Studienordnungen/Module sowie die bisherigen Leistungen der einzelnen Studierenden kann kein System geschaffen werden, welches eben diese Daten aufbereitet wiedergibt.

Des Weiteren müssen die Metadaten für die unterschiedlichen Herangehensweisen erfasst bzw. erhoben werden: Alle Modulverantwortlichen müssen festlegen, wofür ihre Module sich eignen und für wen sie empfohlen werden sollten.

Die Metadaten über das Studierendenverhalten müssen langfristig angelegt bzw. erfasst werden. Hierfür muss festgelegt werden, welche Daten erfasst werden sollen und wie diese ausgewertet werden.

Der weitere Ausbau des Recommender Systems selbst um die Studentengetriebene und die Metadatengetriebene Empfehlung muss implementiert werden.

Wurden diese Schritte durchgeführt, können die gewonnenen Informationen ins System integriert werden, um die Empfehlungen an die Studierenden noch weiter zu verbessern.

Durch Betrachtung der Planmodule und der belegten Module eines Studierenden kann der aktuelle Stand des Studirenden dargestellt werden. Dabei kann aufgezeigt werden, welche Bereiche bereits abgedeckt wurden, und welche noch offen sind. Dies ermöglicht es den Studierenden außerdem, ihren Studienfortschritt besser einzuschätzen.

6 Literaturverzeichnis

[FH09] Fischer, Holger; Hartau, Claudine: STiNE an der Universität Hamburg – Zur Einführung eines integrierten Campus-Management-Systems. 9. Internationale Tagung Wirtschaftsinformatik

[RHHL09] Rieger, Bodo; Haarmann, Thomas; Höckmann, Ellen and Lüttecke, Stiliana: Data Warehouse basierte Entscheidungsunterstützung für das Campus-Management von Hochschulen. 9. Internationale Tagung Wirtschaftsinformatik

[Ra09] Radenbach, Wolfgang: Integriertes Campus-Management durch Verknüpfung spezialisierter Standardsoftware. 9. Internationale Tagung Wirtschaftsinformatik

[RBK14] Ritter, Christopher; Bittner, Patrick; Kao, Odej: Der Weg von BYOD zum GYSE. 7. DFN Forum 2014: Kommunikationstechnologien

[HKR08] Hildmann, Thomas; Kao, Odej; Ritter, Christopher: Rollenbasierte Identitäts- und Autorisierungsverwaltung an der TU Berlin. 1. DFN-Forum Kommunikationstechnologien Verteilte Systeme im Wissenschaftsbereich, 2008.

[HR07] Hildmann, Thomas; Ritter, Christopher: TUBIS-Integration von Campusdiensten an der Technischen Universität Berlin. PIK-Praxis der Informationsverarbeitung und Kommunikation, 30(3):145–151, 2007.

[RHK10] Ritter, Christopher; Hildmann, Thomas; Kao, Odej: Erfahrungen und Perspektiven eines rollenbasierten IdM , 3. DFN-Forum Kommunikationstechnologien, 26. Mai 2010

[HW05] Hildmann, Thomas; Wilke, Thomas J.: Pseudonymous Authentication and Authorization enhancing ubiquitious Identity Management, Proceedings ISSE, September 26-29 2005

[GH00] Gebhardt, Thomas; Hildmann, Thomas: Rollen als Schlüssel für B2B-Anwendungen, DuD - Datenschutz und Datensicherheit, Nr. 24, 2000

Einsatz von MultiTouch-Tischen zur kooperativen Semesterorganisation

Raphael Zender, Vyacheslav Rogozhkin, Ulrike Lucke

Institut für Informatik
Universität Potsdam
August-Bebel-Str. 89
14482 Potsdam
vorname.nachname@uni-potsdam.de

Abstract: Die Fortschritte in der Hardware-Entwicklung sowie die durch die Smartphone/Tablet-Generation etablierte MultiTouch-Interaktionsform fördern die transparente Integration von IT-Systemen in Alltagsgegenstände und -szenarien. Systeme wie tischförmige MultiTouch-Displays werden erschwinglich, benutzbar und somit zunehmend tauglich für den Massenmarkt. Doch bieten sie auch einen Mehrwert für den Einsatz an Bildungseinrichtungen? Der Einsatz von MultiTouch-Tischen an Hochschulen wird in diesem Paper am Beispiel eines prototypischen Tisch-Terminals zur kooperativen studentischen Semesterplanung unter zielgerichteter Ausnutzung der besonderen Bedienparadigmen, Grenzen und Möglichkeiten dieser Technologie motiviert, ausgeführt und diskutiert.

1 Motivation

Menschenzentrierte IT-Systeme durchbrechen zunehmend das klassische Interaktionsparadigma grafischer Benutzungsoberflächen (GUI). Stattdessen verfolgt die Mensch-Computer-Interaktion natürlichere Ein- und Ausgabetechnologien. Vor allem berührungsempfindliche Oberflächen wie die PixelSense-Technologie gelten als Motoren der Entwicklung sogenannter Natural User Interfaces (NUI). In den 80er Jahren bereits thematisiert [Bux86], sind diese durch Fortschritte in der Hardware-Entwicklung und neuartige Interaktionskonzepte inzwischen sowohl erschwinglich als auch bedienbar geworden. Neben dem Smartphone/Tablet-Markt spielen vor allem im betrieblichen Umfeld auch größere Lösungen eine Rolle. Im Gegensatz zu ihren mobilen Verwandten können diese durch mehrere Personen gleichzeitig bedient und in Alltagsgegenstände (z. B. Wandbilder, Tische) integriert werden, so dass auch gemeinsam mit MultiTouch an Problemen gearbeitet werden kann. Durch so genannte *Tangibles* – vom System meist durch optische Codes erkannte physische Gegenstände – wird die Oberfläche noch enger mit der physischen Welt verknüpft [Sch13].

Vor allem im Ausstellungsbereich (z. B. Messen, Produktpräsentationen) sind MultiTouch-Displays heute eine feste Größe. Da die Kosten für derartige Systeme noch vergleichsweise hoch sind (z. B. ca. 7.000 € für einen Microsoft PixelSense-Tisch), ist der Heimmarkt

derzeit nicht relevant. Bildungseinrichtungen wie Hochschulen – die ein Schwerpunkt des studentischen Lebens sind und somit eine große Nutzerbasis haben – profitieren jedoch von derartigen Lösungen. Platziert an zentralen Anlaufpunkten (z. B. in der Aula, in der Mensa oder vor dem Audimax) können sie als Informations- und Interaktionsplattformen dienen. Eingesetzt in der Lehre sind sie vor allem in kooperativen Szenarien interessant, in denen die eigentliche IT in den Hintergrund treten soll.

In diesem Beitrag wird zunächst der Einsatz von MultiTouch-Systemen im Hochschulumfeld untersucht. Anhand eines konkreten Hochschul-Tools zur gemeinsamen Semesterplanung auf der Microsoft PixelSense-Hardware wird exemplarisch die Entwicklung eines MultiTouch-Systems von der Anforderungsanalyse über die prototypische Implementierung und die Evaluierung beleuchtet. Der Beitrag schließt mit einer Zusammenfassung der Ergebnisse und einem Ausblick auf weiterführende Arbeiten.

2 MultiTouch-Systeme im Hochschulumfeld

Großflächige MultiTouch-Systeme sind bisher eher selten im Umfeld deutscher Bildungseinrichtungen anzutreffen. Daher stammen viele der Einsatzszenarien in diesem Abschnitt aus internationalen Projekten oder Fallstudien.

Erweitert man das Spektrum der Bildungseinrichtungen (z. B. auf Museen), findet man vor allem Präsentationsszenarien, die dem Einsatzzweck in der Wirtschaft (Messen- und Produktpräsentationen) ähneln. Angereichert mit geeigneten didaktischen Strategien, ist beispielsweise die kooperative Interaktion mit Wissensnetzwerken im Kontext der Ausstellungsstücke möglich [KR08]. Über die Tangibles können zudem eigene Fortschritte bei der Begehung in einem physischen Objekt "gespeichert" und an geeigneter Stelle visualisiert werden, um das aufgenommene Wissen miteinander zu verknüpfen. Durch eine Feldstudie der University of Calgary in einem öffentlichen Aquarium wurde bereits eine gute Akzeptanz derartiger Displays in öffentlichen Ausstellungsräumen nachgewiesen [HC11].

In diesem Bereich sind auch innovative Möglichkeiten der Informationsvisualisierung angesiedelt. An der Nanyang Technological University in Singapur wurde beispielsweise mithilfe eines MultiTouch-Displays die Visualisierung von astrophysikalischen Zusammenhängen des Universums für Planetarien, Wissenschaftszentren und Schulen umgesetzt [FGN10]. Insbesondere bei der großflächigen Navigation durch derart komplexe naturwissenschaftliche Systeme zeigt sich die Stärke dieser Technologie.

Spätestens seit Apple mit *iTunes U*[1] und *iBooks Lehrbücher*[2] das Lernen mit Tablets propagierte und schlagartig populär machte, ist auch die Aus- und Weiterbildung mit derartigen MultiTouch-Systemen verbreitet. Im Vergleich zu diesen eher einfachen Digitalisierungen traditioneller Lehr- und Lernmethoden bieten größere MultiTouch-Displays weit innovativere Möglichkeiten des Lehrens und Lernens. Vor allem der wachsende Bedarf

[1] https://www.apple.com/de/education/ipad/itunes-u

[2] https://www.apple.com/de/education/ipad/ibooks-textbooks/

nach direktem Kontakt zwischen Lernenden und physikalischer Interaktion im Umfeld des rechnergestütztes Gruppenlernens wird durch diese Lösungen bedient [DE11].

In dem *Star Trek Classroom*-Projekt der University of Durham wurden beispielsweise Klassenräume mit mehreren MultiTouch-Tischen ausgestattet [MH13]. Diese ermöglichten den Schülern im Mathematik-Unterricht das kooperative Lernen am eigenen Tisch, aber auch in der Vernetzung mit anderen Tischen/Schülergruppen durch geeignete Lernanwendungen. Im Rahmen des Projektes wurden signifikant gesteigerte mathematische Fähigkeiten im Vergleich zur herkömmlichen Lernmethoden festgestellt. Zurückgeführt wurden diese auf die Möglichkeit, Lehrern in diesem Setting geeignete Tools für ein Mikromanagement der einzelnen Schüler anbieten zu können.

In einem anderen Projekt an der Universität BLINDED wurde mit dem *ProtoSense* [BLINDED] eine Paper Prototyping-Anwendung entwickelt, mit deren Hilfe angehende GUI-Entwickler diese Designmethode erlernen und im Rahmen ausführbarer Modelle ausprobieren können. Somit füllt diese Anwendung die Lücke zwischen nicht ausführbaren reinen Papier-Prototypen und eher aufwändig zu erstellenden, digitalen Wireframes.

Neben diesen Anwendungsbereichen sind große MultiTouch-Systeme Forschungsgegenstand an vielen Hochschulen und Forschungsinstituten. Diese befassen sich beispielsweise mit Tools zur medizinischen Diagnose und Operationsvorbereitung [RAIH+13], optimierten Interaktionswerkzeugen für die Überwachung von Produktionsabläufen und in der Logistik [RW10] sowie neuartigen Möglichkeiten in der Arbeitsplatzgestaltung durch verschiedene Display-Bauformen [WVSB10].

Das im folgenden vorgestellte MultiTouch-System für den Touch-Tisch Microsoft PixelSense zielt auf die Unterstützung von Studierenden bei der kooperativen Semesterplanung in öffentlichen Bereichen ihrer Hochschule.

3 PixelSense-Semesterplanung

Die zentralen Ziele der Semesterplanung auf einem tischgroßen MultiTouch-Display sind einerseits die Unterstützung von Studentengruppen bei der gemeinsamen Gestaltung eines Semesters (z. B. gemeinsame Veranstaltungen und Freizeiten) und andererseits die Untersuchung des Mehrwerts derartige IT-Systeme für Studierende in öffentlichen Bereichen einer Universität. Im folgenden wurde zunächst auf Grundlage des derzeit an der Universität BLINDED zur Semesterplanung eingesetzten Campus Managements Systems (CMS) eine Anforderungsanalyse unter den potentiellen Nutzern durchgeführt.

3.1 Anforderungen

Da großflächige MultiTouch-Displays für den vorgestellten Einsatzzweck noch nicht weit verbreitet sind, gibt es bisher nur grobe Designrichtlinien der jeweiligen Hersteller. Die aus diesen Richtlinien abgeleiteten Design-Anforderungen wurden im Rahmen dieses Pro-

Funktion	Notwendig	Wünschenswert	Unwichtig
Meine Leistungen	85,98%	8,54%	4,88%
Vorlesungsverzeichnis	82,32%	12,20%	4,88%
Meine Dokumente	70,12%	23,78%	5,49%
Prüfungsverwaltung	67,68%	17,07%	14,02%
Mein Stundenplan	65,85%	18,29%	15,24%
Meine Veranstaltungen	50,00%	35,37%	14,02%
Schnellzugriff auf Kurse	39,02%	46,34%	12,20%
Interaktiver Stundenplan	39,02%	44,51%	14,63%
Suche	36,59%	34,76%	28,05%
Fristen	35,37%	35,37%	28,05%
Adressverwaltung	32,93%	25,61%	40,85%
Synchronisation	25,61%	29,27%	43,29%
Datenschutz	25,00%	27,44%	45,73%
Ersatztermine	23,17%	62,80%	11,59%
Meine Einstufungstests	21,95%	24,39%	53,05%
Ausfälle	15,85%	25,61%	57,93%
Stundenplan per E-Mail	14,63%	44,51%	39,02%
Kollaborationswerkzeuge	14,63%	45,12%	37,80%
Zulassungen	13,41%	21,95%	62,60%
iCal-Export	12,20%	23,17%	62,20%
Vorschlagsystem für Kurse	11,59%	50,61%	34,76%
Heutige Veranstaltungen	10,98%	23,78%	64,02%
Notizzettel	9,76%	31,71%	56,10%
Kursbeschreibungen	5,49%	33,54%	59,15%
Soziale Netze	2,44%	7,93%	88,41%

Tabelle 1: Bewertung der Funktionen des CMS und möglicher neuer Features einer NUI-Variante: Umgesetzt wurden Features, die über ein Drittel der Befragten als notwendig erachtet (grau).

jektes nutzerzentriert um funktionale Anforderungen erweitert. Dafür wurde eine Online-Umfrage unter Studierenden – der zukünftigen Zielgruppe – durchgeführt. An der Umfrage nahmen 164 Studierende aus 11 verschiedenen Fachdisziplinen teil. Da das prototypische System für einen bestimmten Campus an der Universität BLINDED konzipiert wurde, waren vor allem Studierende der dort ansässigen Fachdisziplinen vertreten (Wirtschaftswissenschaften, Informatik, Politikwissenschaften).

Das bisher zur Semesterplanung eingesetzte Tool ist eine modifizierte Version des Campus Management Systems der HIS GmbH. Neben der reinen Semesterplanung werden die typischen CMS-Aufgaben rund um die Leistungsverwaltung, Veranstaltungsinformation und Dokumentorganisation unterstützt. Als Ausgangspunkt für die NUI-Semesterorganisation wurden die Studierenden nach den CMS-Funktionen gefragt, die sie im Hinblick auf eine MultiTouch-Variante für notwendig, wünschenswert und unwichtig halten. Tabelle 1 stellt die Ergebnisse dieser Umfrage dar.

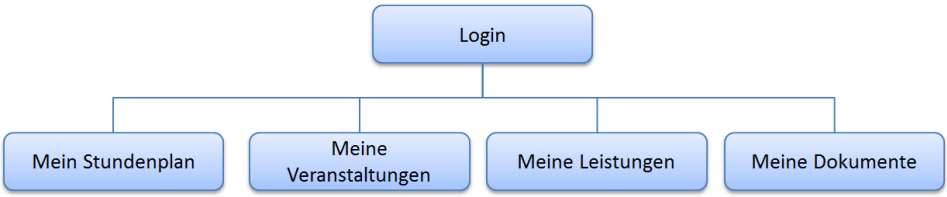


Abbildung 1: Sichtenstruktur der PixelSense-Anwendung

Ein zentrales Ergebnis der Umfrage war somit, dass eine reine Semesterplanung nur einen Bruchteil der wünschenswerten CMS-Funktionen auf einem Touch-Tisch abbilden würde. Die in Tabelle 1 im oberen Abschnitt hervorgehobenen Funktionen werden von mindestens zwei Dritteln der Befragten als notwendig erachtet. Aufgrund dessen wurde entschieden diese Funktionen in der prototypischen Implementierung zu berücksichtigen.

Zudem wurden die Aufgeschlossenheit der befragten Studierenden zur Semesterplanung mit einem MultiTouch-Tisch erfragt. 37% der Befragten äußerten sich klar positiv (*Würde ich machen.*) gegenüber dieser Möglichkeit, während nur 21% dieses Werkzeug klar ablehnten (*Würde ich nicht machen.*). Die restlichen Befragten gaben keine klare Einschätzung ab. Es ist anzunehmen, dass das Szenario derzeit noch recht schwer vorstellbar ist, da die wenigsten der Studierenden bisher Kontakt zu derartiger Hardware hatten.

3.2 Prototypische Implementierung

Die Implementierung wurde für die Microsoft PixelSense-Plattform entwickelt, die auf dem Tischcomputer SUR40 des Herstellers Samsung läuft [Mic11]. Der Touch-Tisch ist mit einem 1080p LCD HD Bildschirm auf einer Diagonalen von 40 Zoll ausgestattet. Pro acht Pixeln kommt ein optischer Sensor zum Einsatz, der Objekte auf der Tischoberfläche aufgrund ihrer Standform erkennt. Spezielle Objekte (z. B. Finger, PixelSense-Tags) sind zudem identifizierbar. Ein leistungsfähiger AMD Athlon II X2-Prozessor erzeugt in Zusammenarbeit mit der Radeon HD 6700M-Grafikkarte ein flüssiges Bediengefühl trotz hoher Auflösung und aufwendiger Touch-Sensorik. Als Betriebssystem kommt ein angepasstes Windows 7 zum Einsatz.

Als Programmmiersprache für die Anwendung wurde C# gewählt, da diese von dem PixelSense SDK direkt unterstützt wird. Das SDK fördert zudem die Berücksichtigung der von Microsoft empfohlenen Designrichtlinien in Form fertiger Kontrollelemente und Projektstrukturen. Die Benutzungsoberfläche wurde mithilfe der Windows Presentation Foundation (WPF) – einem grafischen Subsystem – realisiert und setzt auf die Deklaration von Benutzer-Sichten mithilfe der XML-basierten Extensible Application Markup Language (XAML). Die fünf wichtigsten Sichten für Studierende sind in Abbildung 1 im Zusammenhang benannt.

Mit der derzeitigen Version der entwickelten Software können bis zu vier Studierende gleichzeitig die Semesterplanung durchführen. Mit leichten Modifikationen in der Ansicht wäre dies auf sechs Nutzer erweiterbar. Größere Gruppen könnten nicht mehr sinnvoll um den eher kleinen Tisch platziert werden.

3.2.1 Login

Die erste Ansicht für Nutzer des Tisches ist der Login. Mithilfe ihres Nutzernamens und Passworts können sie sich authentifizieren, wobei die PixelSense-Anwendung diese ausschließlich an die Authentifizierungsfunktion des CMS weiterleitet und bei Erfolg eine grafische Übersicht über die vier weiteren Sichten darstellt. Zudem ist jederzeit der Logout möglich. Derzeit wird noch die Authentifizierung per Texteingabe eingesetzt. Für die Authentifizierung im laufenden Betrieb an einem öffentlichen Ort an der Hochschule ist dies keine Option. Eine 2-Wege-Authentifizierung ohne Eingabe von Passwörtern über die gut sichtbar eingeblendete Tastatur ist in der Entwicklung.

Per eingeblendetem Button können die Studierenden nun zwischen der Anzeige ihres Stundenplans, ihrer belegten Veranstaltungen, ihrer erbrachten Leistungen sowie der Dokumentenabfrage wählen.

3.2.2 Stundenplan

Die zentrale Funktionalität der Anwendung ist die kooperative Planung von Semester-Stundenplänen durch 1-4 Studierende. In Abbildung 2 ist diese für zwei angemeldete Studierende dargestellt. Für einen dritten und vierten Nutzer wird der Bildschirm weiter aufgeteilt. Aufgrund der hohen Auflösung und der großen Tischgröße ist die Bedienbarkeit weiterhin akzeptabel.

Neben der aktuellen Wochenansicht mit zuvor bereits gewählten Veranstaltungen werden am unteren Bildrand die noch zur Wahl stehenden weiteren Kurse dargestellt. Die grün hervorgehobenen Kurse stehen dem jeweiligen Studenten mit seinem aktuellen Studienfortschritt zur Belegung zur Verfügung. Durch eine einfache Wischbewegung kann der Kurs aus dem Auswahlfeld in den Stundenplan geschoben werden. Zudem werden im Stundenplan gelb hervorgehobene Lehrveranstaltungen zur Belegung empfohlen, sofern der Studienplan dies vorsieht, beispielsweise in Form von Pflichtveranstaltungen. Violett umrandete Veranstaltungen stellen gemeinsame Belegungen aller am Tisch angemeldeten Nutzer dar und ermöglichen die kooperative Semestergestaltung.

Zudem kann der Nutzer sich direkt für die Prüfung einer Lehrveranstaltung anmelden bzw. wird bei entsprechenden Anmeldefristen rechtzeitig vorher informiert. Weiterhin ist es möglich sich in einer Kartenansicht auf Basis von Bing-Maps das Gebäude auf dem Campus anzeigen zu lassen, in dem der Kurs stattfindet. Insbesondere bei Universitäten mit mehreren Campussen ist dies eine wichtige Information, da Fahrzeiten eingeplant werden müssen. Auch Treffen mit Kommilitonen können so leichter arrangiert werden.

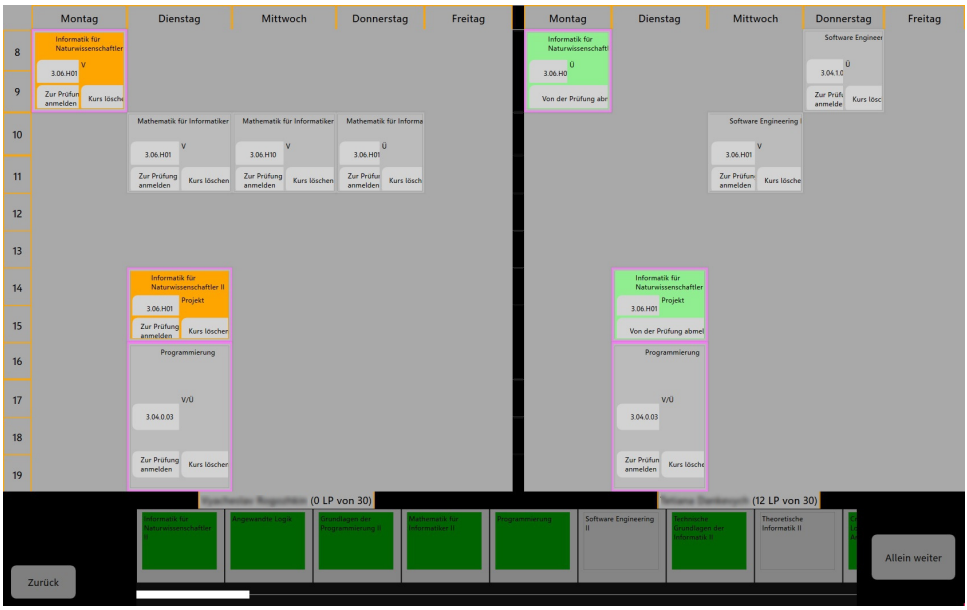


Abbildung 2: Planung eines Semesters durch zwei Studierende

3.2.3 Veranstaltungen

In der in Abbildung 3 dargestellten Veranstaltungsübersicht können eingeloggte Studierende alle bisher im Studium belegten Lehrveranstaltungen einsehen und nach Semestern filtern.

Jede einzelne Veranstaltung kann per Wischgeste aus dem Übersichtsfenster herausgelöst und in einem eigenen Fenster dargestellt werden. Dadurch können beispielweise ebenfalls am Tisch sitzende Kommilitonen diese mitsamt der dazugehörigen Metadaten separat betrachten – z. B. um zu entscheiden ob diese Lehrveranstaltung für sie ebenfalls interessant ist.

3.2.4 Leistungen

Die dritte Sicht dient der Darstellung der bisherigen Leistungen des Studierenden und wird mit Abbildung 4 beispielhaft illustriert. In dieser Ansicht wurde auf die Anzeige der Noten für die einzelnen Leistungen verzichtet, um der Privatsphäre des Nutzers in einem öffentlichen Raum gerecht zu werden. Auch hier kann wieder zwischen den einzelnen Semestern gefiltert werden. Die Gesamtübersicht hilft beispielsweise, um noch offene Leistungen zu identifizieren und in der Folge die Semesterplanung daraufhin auszurichten.

Abbildung 3: Veranstaltungsübersicht für das laufende Semester

3.2.5 Dokumente

Eine weitere Anforderung der Studierenden war die Möglichkeit relevante Dokumente anzeigen lassen zu können. In Kombination mit einem ebenfalls öffentlichen Drucker könnten diese direkt am Touch-Tisch als Druckauftrag abgesendet und über die Studierenden-Chipkarte abgerechnet werden. Folgende Dokumente werden derzeit angezeigt:

- Studienbescheinigung
- Übersicht über bisher belegte Kurse
- Notenübersicht

Alle Dokumente werden direkt als PDF dargestellt und könnten so neben der Druckoption beispielsweise auch direkt über einen USB-Stick oder eine E-Mail-Anwendung wiederverwendet werden. Zudem können alle Dokumente in deutsch oder englisch abgerufen werden.

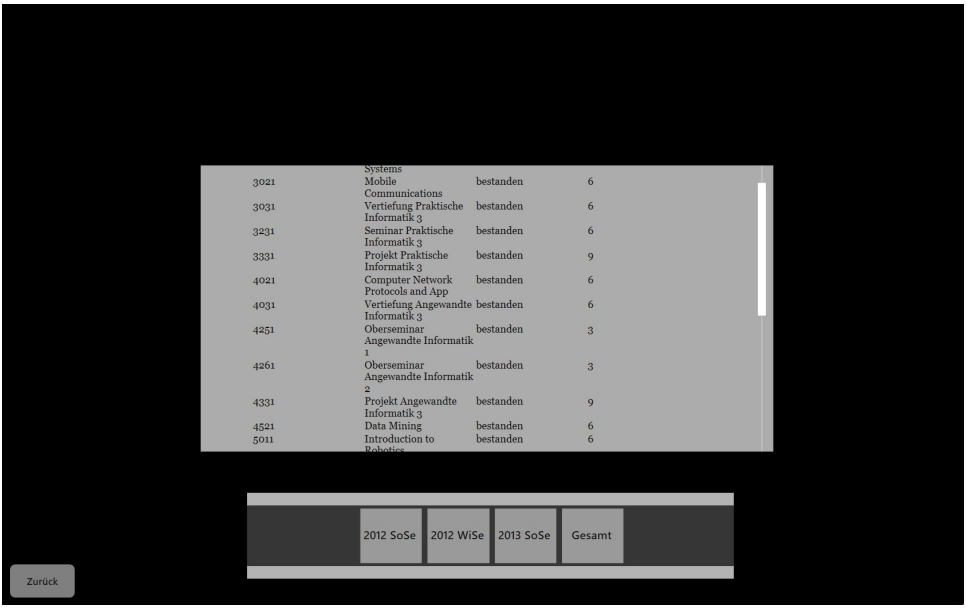


Abbildung 4: Leistungsübersicht für einen einzelnen Studenten

3.3 Evaluierung

Die Evaluierung der PixelSense-Anwendung an einem öffentlichen Ort auf dem Campus steht zum jetzigen Zeitpunkt noch aus. Zuvor müssen damit verbundene Fragestellungen bzgl. der Authentifizierung, der organisatorischen Verantwortlichkeit für die Hardware und die Auswahl des konkreten Standortes abschließend geklärt werden. Aus diesem Grund wurde zunächst eine tendenzielle Evaluierung der Nutzbarkeit des aktuellen Prototyps mit acht Studierenden aus den in der initialen Umfrage berücksichtigten Studienrichtungen durchgeführt.

Sie wurden aufgefordert konkrete Aufgaben mit dem System zu lösen und dabei gegenüber dem Testverantwortlichen mündlich über ihr Erlebnis mit dem System zu berichten (lautes Denken). Zu den Aufgaben gehörten Interaktionen mit allen Systemkomponenten (z. B. Kurse in den Stundenplan übertragen, für Prüfungen registrieren, konkrete Dokumente anzeigen). Die Aufgaben mussten sowohl in der Einzel- als auch der Dual-Ansicht bewältigt werden. Die Formulierung der Aufgaben beinhaltete dabei keinen konkreten Hinweis auf den Ort einer Funktion oder deren Namen sondern nur das Ziel der Aufgabe (z. B. *Lassen Sie sich Ihre aktuellen Leistungen anzeigen.*).

Die Erfahrungen der Testpersonen waren durchweg positiv. Sowohl das Login am System als auch das Verständnis der Funktionsübersicht funktionierten ohne nennenswerte Probleme. Bei der Stundenplanorganisation gab es vereinzelt Probleme durch Interaktionspattern, die von Tablets oder Smartphones bekannt waren, jedoch durch die Software

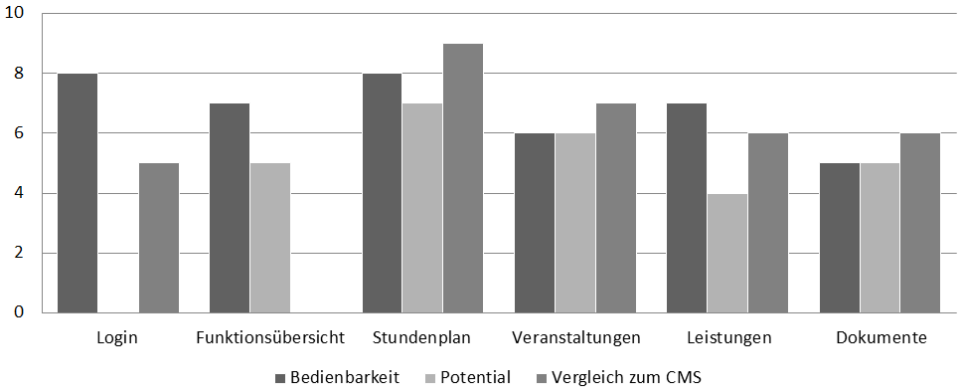


Abbildung 5: Bewertung der PixelSense-Anwendung durch die Testpersonen bzgl. der Bedienbarkeit, der Ausnutzung des Potentials und im Vergleich mit dem herkömmlichen CMS auf einer Skala von 0 bis 10

nicht oder in einem anderen Kontext genutzt wurden (z. B. Scrollbar anstelle einfacher Wischgeste beim Scrolling in der Wochenansicht). Die Übertragung der Kurse in den Wochenplan verlief problemlos, auch wenn zusätzliche Informationen gewünscht wurden (z. B. Anzeige der Credit Points für eine Lehrveranstaltung). Auch in der Übersicht der bisher absolvierten Lehrveranstaltungen und erreichten Leistungen gab es bis auf vereinzelte Missverständnisse bei den Touch-Gesten (z. B. Entfernung einer Kurs-Einzelansicht durch Zurückschieben auf die Kurs-Liste statt Close-Button) keine nennenswerten Probleme in der Bedienung. In der Dokumentenansicht bemängelten die Nutzer nur, dass die PDF-Ansichten sich nicht wie die anderen Fenster beliebig skalieren lassen. Leider liegt hier eine Limitierung des PixelSense-SDKs zur PDF-Anzeige vor.

Ein weiterer Wunsch der Nutzer war die Assistenz der PixelSense-Anwendung zur Studiumsorganisation zu erhöhen. Es wurde beispielsweise angeregt vom System automatisch die Kurse vorzuschlagen, die noch für die Erreichung der notwendigen ECTS-Punkte in einem Fächerblock besucht werden könnten. Dies könnte beispielsweise durch die Integration einer geeigneten Kursempfehlungskomponente realisiert werden [BR12].

Abbildung 5 fasst die tendenzielle Evaluierung mit den Testpersonen zusammen. Diese wurden aufgefordert auf einer Skala von 0 bis 10 die Bedienbarkeit zu bewerten (0:nicht intuitiv, 10:sehr intuitiv), die Ausnutzung des Potentials der Funktionalität durch die Software einzuschätzen (0:viele Features fehlen, 10:Funktionalität komplett vorhanden) und diese mit dem bereits bekannten CMS der Universität BLINDED zu vergleichen (weniger als 5:PixelSense-Version ist schlechter, mehr als 5:PixelSense-Version ist besser).

Im Ergebnis wird die PixelSense-Version der einzelnen Funktionen durchaus als verbesserungswürdig hinsichtlich ihrer Bedienbarkeit und Vollständigkeit angesehen. Dennoch wird sie in allen Bereichen mindestens ebenbürtig, oft sogar besser als das herkömmliche CMS bewertet.

4 Zusammenfassung und zukünftige Arbeiten

Großflächige MultiTouch-Displays werden erschwinglich, benutzbar und somit zunehmend tauglich für den Einsatz in Unternehmen und Hochschulen. In diesem Beitrag wurden die Einsatzmöglichkeiten derartiger Systeme für Hochschulen näher betrachtet und deren Potential aufgezeigt. Am Beispiel einer exemplarischen Anwendung zur kooperativen Semesterplanung auf dem Tischcomputer Microsoft PixelSense wurden zudem die Nutzbarkeit von Touch-Tischen im Studienalltag demonstriert und der Weg zu einer konkreten Umsetzung erläutert. Eine Gegenüberstellung mit einem herkömmlichen System für die Studienorganisation lieferte erste akzeptable Ergebnisse.

Neben der praxisnahen Evaluierung der vorgestellten Lösung im Studienalltag gibt es zudem vereinzelten Verbesserungsbedarf beim vorgestellten Prototypen. Allem voran muss ein geeignetes Authentifizierungskonzept integriert werden, dass eine geheime aber zuverlässige Authentifizierung im öffentlichen Raum erlaubt. Die entsprechende Integration eines 2-Wege-Authentifizierungsansatzes unter der Nutzung der Chipkarte jedes Studierenden (Besitz) und einer geheimen verdeckt einzugebenden Information (Wissen) ist derzeit in der Entwicklung.

Weiterhin werden die bisherigen Evaluierungsergebnisse für die Erhöhung der Assistenz der PixelSense-Anwendung sowie Überarbeitung einzelner Interaktionspattern im Hinblick auf die Tablet- und Smartphone-Erfahrungen der Nutzer herangezogen. Das PixelSense-SDK bietet bereits eine Reihe sinnvoller Ansätze für intuitive Bedienkonzepte (z. B. vorgefertigte Steuerelemente). Da allerdings keine der in der Breite bekannten Smartphones und Tablets diese Konzepte einsetzen, konnten diese erst bei der zweiten Begegnung durch die Testpersonen direkt und unmißverständlich genutzt werden.

Zusätzlich angeregte Funktionen (z. B. die dauerhafte Ausblendung nicht interessanter Kurse für einzelne Studierende) müssen hinsichtlich ihrer Wirkung und Umsetzbarkeit untersucht und ggf. in das Gesamtsystem integriert werden, bevor dieses in den praktischen Einsatz überführt wird.

Mittelfristig wird außerdem die Kopplung ausgewählter Funktionen bzw. die Wiederverwendung von einzelnen Modulen der PixelSense-Anwendung im Rahmen der universitätsweiten Studierenden-App angestrebt.

Literaturverzeichnis

[BR12] Annette Baumann und Lucas Reeh. Folgende Module könnten Sie auch interessieren? In *E-Learning Symposium 2012*, Seiten 67–72. Universitätsverlag Potsdam, 2012.

[Bux86] William Buxton. There's More to Interaction than Meets the Eye: Some Issues in Manual Input. In D. A. Norman und S. W. Draper, Hrsg., *User Centered System Design: New Perspectives on Human-Computer Interaction*, Seiten 319–337. Erlbaum, 1986.

[DE11] Pierre Dillenbourg und Michael Evans. Interactive tabletops in education. *International Journal of Computer-Supported Collaborative Learning*, 6(4):491–514, 2011.

[FGN10] Chi-Wing Fu, Wooi-Boon Goh und Junxiang Allen Ng. Multi-touch Techniques for Exploring Large-scale 3D Astrophysical Simulations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, Seiten 2213–2222. ACM, 2010.

[HC11] Uta Hinrichs und Sheelagh Carpendale. Gestures in the wild: studying multi-touch gesture sequences on interactive tabletop exhibits. In Desney S. Tan, Saleema Amershi, Bo Begole, Wendy A. Kellogg und Manas Tungare, Hrsg., *CHI*, Seiten 3023–3032. ACM, 2011.

[KR08] Daniel Klinkhammer und Harald Reiterer. Blended Museum - Perspektiven für eine vielfältige Besuchererfahrung (Enhancing Visitor Experience through Information Design and Interaction). *i-com*, 7(2):4–10, 2008.

[MH13] Emma M. Mercier und Steven E. Higgins. Collaborative learning with multi-touch technology: Developing adaptive expertise. *Learning and Instruction*, 25(0):13 – 23, 2013.

[Mic11] Microsoft. *Developing Applications for Microsoft Surface 2.0 (Whitepaper)*. Microsoft Corporation, 2011.

[RAIH+13] Felix Ritter, Jumana Al Issawi, Markus Harz, Simon Benten und Kathy J. Schilling. Combining Mobile Devices and Medical Workstations for Diagnostic Reading of Medical Images. *i-com*, 12(1):2–9, apr 2013.

[RW10] Kay Rogge und Thomas Wochinger. Multi-Touch in Logistik und Auftragsmanagement. *IT production*, (4):32–33, 2010.

[Sch13] Thomas Schlegel. Einleitung: Warum Multi-Touch? In Thomas Schlegel, Hrsg., *Multi-Touch*, Xpert.press, Seiten 3–12. Springer Berlin Heidelberg, 2013.

[WVSB10] Malte Weiss, Simon Voelker, Christine Sutter und Jan Borchers. BendDesk: Dragging Across the Curve. In *ACM International Conference on Interactive Tabletops and Surfaces*, ITS '10, Seiten 1–10. ACM, 2010.

Umweltinformatik zwischen Nachhaltigkeit und Wandel

(UINW)

Workshop: Umweltinformatik zwischen Nachhaltigkeit und Wandel (UINW 2014)[1]

1 Hintergrund und Motivation

Die Bedeutung der Informatik in den Naturwissenschaften, Lebenswissenschaften und vor allem in den Umweltwissenschaften ist evident. Umweltinformatik ist ein wichtiger Bereich der Angewandten Informatik. Dies wurde bereits in den achtziger Jahren erkannt und mit der Gründung des Fachausschuss Umweltinformatik in der Gesellschaft für Informatik 1986 umgesetzt. Die Mitglieder des Fachausschusses sind neben der Informatik auch anderen wissenschaftlichen Disziplinen angehörig, z. B. Chemie, Physik, Mathematik, Geographie, Geologie, Chemie etc. Somit ist der Fachausschuss sehr interdisziplinär aufgestellt. Der Fachausschuss Umweltinformatik gliedert sich in drei Fachgruppen und ist Teil des GI-Fachbereichs Informatik in den Lebenswissenschaften.

Ziel dieser Informatikdisziplin Umweltinformatik ist der Erfahrungsaustausch zur Analyse und Lösung von Umweltproblemen mit Hilfe aktuellster Methoden der Informations- und Kommunikationstechnologien (IKT). Ein zunehmend wichtiger Bereich ist dabei das mit Umweltthemen im Zusammenhang stehende Energiethema. Hier wäre beispielhaft die Steigerung von Energieeffizienz mittels Software, aber auch die damit in Verbindung stehende energiebewusste Softwareentwicklung zu nennen. Somit lässt sich feststellen, dass sich die Umweltinformatik aufgrund des dynamischen Anwendungsfeldes Umweltschutz und Nachhaltigkeit sowie der beständigen Weiterentwicklung der IKT erheblichen Herausforderungen gegen-übersteht. Der Workshop soll hierzu einen Beitrag leisten und aktuelle Trends diskutieren, aber auch den inter- und transdisziplinären Fokus schärfen und über den Tellerrand neuester Gadgets und Effizienzstrategien hinausblicken.

Dabei fallen gerade im Umwelt- bzw. Energiebereich erhebliche Datenmengen an, deren Komplexität es zu meistern gilt. In diesem Zusammenhang sind beispielhaft folgende Themenkreise zu nennen:

- Sehr große Datenmengen in Umweltanwendungen
- Geographische Informationssysteme (GIS)
- Open Data in Umweltinformationssystemen
- Betriebliche Umweltinformationssysteme (BUIS)
- Klimawandel und Adaptation
- Modellierung und Simulation von großen Datensätzen
- Umwelt- und Energieinformationssysteme
- Zusammenhänge zwischen Umwelt- und Gesundheitsdaten
- Smart Meter und Smart Grid
- Massendatenspeicherung in Energiesystemen einschließlich Sicherheitsaspekten

[1] http://www.enviroinfo.eu/uinw2014

Die Teilnehmerinnen und Teilnehmer des Workshops können sich über die aktuellen und zukünftigen Aktivitäten in der Umweltinformatik inklusive energierelevante Themen informieren.

Nach dem ersten Workshop „Umweltinformatik zwischen Nachhaltigkeit und Wandel 2013 “ in Koblenz setzt dieser zweite Workshop die begonnene Diskussionsreihe fort.

2 Workshop-Organisatoren

- Kristina Voigt, Institut for Computational Biology, Helmholtz Zentrum München
- Stefan Naumann, Institut für Softwaresysteme in Wirtschaft, Umwelt und Verwaltung, Hochschule Trier, Umwelt-Campus Birkenfeld

3 Programmkomitee

- Rainer Bruggemann, Leibniz-Institute of Fresh Water Ecology and Inland Fisheries
- Ulrike Freitag, Condat AG
- Werner Geiger, Karlsruhe Institute of Technology
- Werner Pillmann, International Society for Environmental Protection
- Wolf-Fritz Riekert, Stuttgart Media University
- Jochen Wittmann, HTW Berlin
- Volker Wohlgemuth, HTW Berlin

EnergieNetz – Eine webbasierte, erweiterbare Open-Source-Software für das Energiemanagement in lernenden Energieeffizienz-Netzwerken

Maximilian Schneider[1], Kurt Weissenbach[2], Volker Wohlgemuth[1]

[1]HTW Berlin
Studiengang Betriebliche Umweltinformatik
Wilhelminenhofstr. 75 A
12459 Berlin
Maximilian.Schneider@htw-berlin.de
Volker.Wohlgemuth@htw-berlin.de
[2]Modell Hohenlohe e.V.
Weststraße 37
74629 Pfedelbach
k.weissenbach@modell-hohenlohe.de

Abstract: Der vorliegende Beitrag beschreibt die Zielstellung und Konzeption einer webbasierten, erweiterbaren Open-Source-Software im Themenbereich Energiemanagement spezialisiert auf die Einsatzgebiete Experten-Netzwerke und betriebsinterne DIN EN ISO 50001. Unterstützt werden soll dies durch die Anbindung von mobilen Erfassungssystemen und der Etablierung einer auf Dauer angelegten Zusammenarbeit von Forschung, Lehre und betrieblicher Praxis zur Weiterentwicklung.

1 Motivation und Zielstellung

Steigende Energie- und Rohstoffpreise, ein verstärktes gesellschaftliches Interesse an Energiethemen und neue Kundenanforderungen führen dazu, dass der verantwortungsvolle Umgang mit Energie als Ressource immer wichtiger wird.

Im Jahr 2011 emittierte die deutsche Volkswirtschaft rund 917 Mio. t Treibhausgase [Um12]. In der effizienteren Nutzung der Energie sowie in der Substitution der fossilen Energieträger liegt ein erhebliches Minderungspotenzial der Emissionen. Die Bundesregierung hat sich das Ziel gesetzt, für die kommenden Jahre eine Verbesserung der Energieproduktivität von jährlich 2,1% zu erreichen. Des Weiteren soll eine CO_2-Minderung der Treibhausgasemissionen von 30 bis 40% gegenüber 1990 erreicht werden [BB10]. Während der letzten 10 Jahre lag die jährliche Energieproduktivitätssteigerung der deutschen Wirtschaft bei ca. 1,8% [BB12]. Um die angestrebten Ziele der

Bundesregierung zur CO_2-Minderung und Energieproduktivität zu erreichen, ist demnach der aktuelle Trend bei Weitem nicht ausreichend auf dem Weg zu einer wirklich nachhaltigen Industriegesellschaft.

Kleine und mittlere Unternehmen (im Folgenden KMU) erwirtschaften in Deutschland rund 36% der Umsätze aller Unternehmen und beschäftigen rund 60% aller Arbeiternehmer/innen. Betriebe mit weniger als 500 Mitarbeitern stellen rund 85% aller Ausbildungsplätze [In13]. KMU bieten aufgrund der Vielzahl an existierenden Unternehmen insgesamt ein enormes Potenzial zur Steigerung der Energieeffizienz in Deutschland [Jo11].

Mit dem Managementsystem DIN EN ISO 50001, dem Energieaudit nach DIN EN 16247-1, lernende Energieeffizienz-Netzwerke (LEEN) sowie Eco Management and Audit Scheme (EMAS) existieren Instrumente, die dazu geeignet sind, das betriebliche Energiemanagement zu unterstützen und die Energieeffizienz in Unternehmen zu erhöhen. Zwar existieren allgemeine Leitfäden (siehe [De06; Bu09; LBG12]) als Hilfestellung zur Umsetzung entsprechender Managementsysteme und Energieeffizienzmaßnahmen, allerdings fehlt den meisten KMU neben den personellen wie finanziellen Ressourcen, das notwendige Know-how diese umzusetzen [Ei11]. Daneben existieren generelle Einstiegshürden in KMU [IT10; FH11]. Häufig sind Energieverbrauchsdaten unbekannt, Energiekosten intransparent und Unternehmen nicht hinreichend für das Thema Energie sensibilisiert [HSR11]. Ein in diesem Zusammenhang vielversprechender Ansatz, die sog. lernenden Energieeffizienz-Netzwerke (im Folgenden lernende Netzwerke), wurde in der Schweiz entwickelt.[1] Dieser Ansatz basiert darauf, dass mehrere Unternehmen mit dem Ziel zusammenarbeiten, kosteneffektiv Energie einzusparen, indem sie voneinander lernen. Dazu schließen sich 10 bis 15 Unternehmen einer Region aus unterschiedlichen Branchen zusammen. Durch einen regelmäßig moderierten Austausch von Kenntnissen und Erfahrungen der Energieverantwortlichen der jeweiligen Unternehmen werden Synergien und Innovationen gefördert, die Such- und Entscheidungskosten für Energieeffizienzmaßnahmen wesentlich reduzieren. Das Modell Hohenlohe e.V. führt seit 2002 bundesweit lernende Netzwerke durch. Besonders fehlendes, kompetentes Personal, der Mangel an notwendiger Zeit und die Sorge, Mitarbeiter/innen „unproduktiv“ für die Aufgabe binden zu müssen, lassen Inhaber von KMU nach den Erfahrungen des Modell Hohenlohe e.V. häufig davor zurückschrecken, sich für entsprechende Projekte zu entscheiden.

Seit 2010 führt das Modell Hohenlohe e.V. auch speziell auf die Bedürfnisse von KMU zugeschnittene lernende Netzwerke durch [WS12]. Damit KMU in der Lage sind, nach einer ersten, extern begleiteten Anlaufphase eigenständig Energieeffizienzfragestellungen zu bearbeiten und diese nachhaltig in den betrieblichen Alltag zu integrieren, ist es notwendig, methodisches Wissen in den Unternehmen aufzubauen und vorzuhalten. Hierzu wird ein Vorgehen benötigt, das zum einen das Lernen fördert und auf der anderen Seite einfach erschließbare konkrete Lösungen offeriert. Dies soll mit einer datenbankgestützten Softwarelösung, die im Zusammenwirken einer dynamischen, auf ständige Verbesserung der Instrumente und auf Erfahrungsaustausch ausgerichteten

[1] Siehe http://www.energiemodell-zuerich.ch/www/index.php (letzter Zugriff 23.06.2014).

Community von Softwareentwicklern, KMU und Ingenieuren basiert, erfolgen. Häufig ist es aufgrund der nicht gegebenen Offenheit ansatzweise nutzbarer Softwaresysteme mit fachspezifischer Funktionalität (z.B. zum Energiecontrolling, Werkzeuge zur Maßnahmenverwaltung etc.) nicht oder nur mit sehr hohem Aufwand möglich, diese fachspezifische Funktionalität in einem anderen Kontext zur Verfügung zu stellen.

Kernidee des hier verfolgten Ansatzes ist daher die Nutzung von entsprechenden Komponenten, die typische „Fachfunktionen" in einem geschlossenen, medienbruchfreien System bereitstellen. Die Nutzung und Entwicklung eines Plug-In basierten Open Source-Rahmenwerks soll es dem Anwender ermöglichen, sich sein System bausteinorientiert aus bestehenden Plug-In-Komponenten zusammen zu stellen, um flexibler auf die sich immer wieder ändernden Anforderungen reagieren zu können.

Ein wesentliches Hemmnis ist heute, dass bei KMU für klimaschützende und energiesparende Lösungen notwendige Daten nicht digital vorliegen. KMU müssen ihre Daten häufig noch, mit Stift und Papier erfassen und diese im Anschluss zur weiteren Nutzung in Tabellenkalkulationsprogrammen digitalisieren [DLW08]. Dies wird durch die derzeit am Markt verfügbaren, elektronisch unterstützten Controlling-Systeme nur bedingt gelöst, da diese einerseits von den KMU als zu kompliziert und zu teuer abgelehnt werden und anderseits die mit den Systemen generierten Daten für eine systematische, elektronische Weiterverarbeitung nicht zur Verfügung stehen.

Bisher gibt es keine Lösung, die in angemessener Weise die Vorteile einer überbetrieblichen Zusammenarbeit in „gemanagten" Netzwerken mit den einzelbetrieblichen Anforderungen verbindet und so die Netzwerkvorteile in den Unternehmen ohne Medienbrüche zum Tragen bringt. Dies liegt zum einen daran, dass die Workflows der bestehenden Energiemanagementsoftware nicht einheitlich sind und zum anderen an den fehlenden Schnittstellen zur effizienten Datennutzung und Bewertung innerhalb eines Netzwerks von Unternehmen. Energiemanagementsoftware, die zumindest auf Teilaspekte adaptiert werden könnte, ist meist aufgrund ihres Fokus auf größere Unternehmen für den Einsatz in KMU funktional und damit finanziell überdimensioniert. Daher sind besondere Anforderungen in Bezug auf die Anpassbarkeit der Software zu beachten, welche es ermöglichen, die Funktionalität der Software unter Einbindung evtl. bereits vorhandener Insellösungen über Schnittstellen individuell bzw. flexibel an die unterschiedlichen betrieblichen Gegebenheiten eines KMUs anzupassen. Ebenso spielt die Gebrauchstauglichkeit der Anwendung eine Rolle. Künftig werden deshalb auch mobile Anwendungen für eine kostengünstige und flexible Datenerfassung an Bedeutung gewinnen und in ein Gesamtsystem eingebunden werden müssen. Mit mobilen Geräten (Smartphone/Tablet) konnte im Rahmen der Projekte MOEBIUS und OpenResKit bereits Erfahrungen zur Reduzierung des Aufwands bei der Datenerhebung sammeln und so die Bereitschaft in KMU, Daten für das Energie- und Stoffstrommanagement zu erfassen, erhöhen.

Durch den Einsatz einer flexiblen Open Source-Software und dem Aufbau einer, moderne Kommunikationsmittel nutzenden, Community versprechen sich die Autoren, die nachwachsende, für moderne IT-Werkzeuge aufgeschlossene Unternehmer-

generation für den Einstieg in einen systematischen, klimaschützenden Prozess zu gewinnen.

Der innovative Ansatz liegt darin, sowohl auf Seiten der methodischen Basis als auch auf der softwaretechnischen Umsetzungsseite je eine Community zu etablieren, welche sich der Idee der Unterstützung der Energieeffizienz in KMU verpflichtet fühlt. Das in dieser Abhandlung beschriebene Vorhaben nutzt hierzu die in den letzten Jahren aufgewachsenen Möglichkeiten von Social Media Ansätzen zu einer auf Dauer angelegten Zusammenarbeit von Forschung, Lehre und betrieblicher Praxis.

2 Konzeption

In diesem Vorhaben soll, in enger Zusammenarbeit mit praktischen Anwendern, ein auf KMU zugeschnittener Open Source basierter Softwarebaukasten entstehen, der die Reduzierung klimaschädlicher CO_2-Emissionen durch eine Steigerung der Energieeffizienz unterstützen soll. Die Zielgruppen des Projekts sind Träger von Energieeffizienz-Netzwerken, Ingenieure, Softwareentwickler und KMU. Bei dem Aufbau einer Community sollen darüber hinaus Organisationen der Wirtschaft und interessierte wissenschaftliche Institute als Multiplikatoren einbezogen werden. Ziel ist es, für KMU Methoden und Softwarewerkzeuge bereitzustellen, mit denen in einer ganzheitlichen Betrachtung des Unternehmens die Transparenz von Energiedaten erhöht, Energiedaten strukturiert erhoben und verarbeitet sowie Netzwerke gemanagt werden können. Eine zentrale Rolle spielen dabei einfache Softwarewerkzeuge, die den Nutzer bei Einzelproblemen unterstützen und zusätzlich dazu beitragen, eine strukturierte Datengrundlage für weitergehende Untersuchungen zu legen. Das Vorhaben baut auf die drei zentralen Säulen (siehe Abbildung 1) Wissen, Software und Schnittstellen auf, die miteinander verknüpft einen vielfältigen Fundus an Werkzeugen und leicht konsumierbares Wissen zum Thema Energieeffizienz bieten. Durch dieses Zusammenspiel sollen die Erfahrungen und das Wissen aus lernenden Netzwerken einer bundesweiten Gemeinschaft von KMU entsprechend des Open Source Ansatzes zur Verfügung gestellt werden, sodass auch deren Energieeffizienzpotenziale erschlossen werden können.

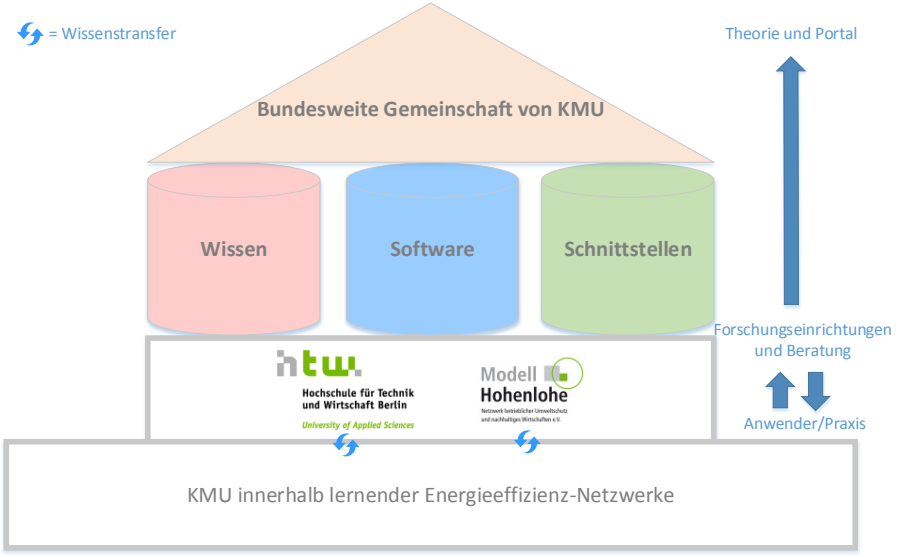


Abbildung 1: Geplanter Aufbau des Vorhabens

Software

Die Software ist die zentrale Säule, da sie die verschiedenen Akteure (Energiebeauftragter, Moderator, Netzwerkmanager und Energieberater) eines lernenden Netzwerks bei deren operativen Tätigkeiten wie Transparenzsteigerung, Planung und Datenerhebung in einem vollständigen digitalen Workflow unterstützen soll. Die zu erstellende Software setzt sich unter anderem aus einzelnen, einfachen, problemspezifischen Softwaretools zusammen, die je nach Anwendungsfall für einfache Analysen und zur Administration entwickelt werden. Diese Softwaretools machen die Methoden aus der Säule "Wissen" unmittelbar digital nutzbar. Da Energieeffizienzanalysen unternehmensspezifisch sind und auf vorhandene Strukturen aufbauen müssen, bedarf die zugrundeliegende Softwarearchitektur einer für alle Plattformen flexiblen Kombinier- und Erweiterbarkeit. Eine entsprechende Softwarearchitektur wurde bereits im Rahmen des an der HTW Berlin durchgeführten Projekts OpenResKit entwickelt [Wo14]. Diese soll die Basis für einen im Unternehmen zentral betriebenen Server bilden und in ihren einzelnen Bestandteilen zusammen mit den problemspezifischen Softwarewerkzeugen über eine Webplattform veröffentlicht und so beliebig kombiniert werden können. Dazu wird der Quellcode frei zur Verfügung gestellt, sodass interessierte Unternehmen diesen entsprechend ihren individuellen Bedürfnissen anpassen können. Die Webplattform dient daneben im Wesentlichen zur Vereinfachung und Unterstützung der Kommunikation innerhalb lernender Netzwerke (siehe Abbildung 2). Teilnehmende Unternehmen können dabei entsprechend der Gegebenheiten und Bedürfnisse energierelevante Daten, Maßnahmen etc. direkt über die Benutzungsoberfläche der Webplattform einpflegen oder diese automatisiert mit dem im Unternehmen optional betriebenen Server austauschen, welcher zusammen mit den problemspezifischen Softwarewerkzeugen eine genauere Analyse des Unternehmens erlaubt.

Somit kann je nach Erfordernis, Unternehmensgröße und vorhandener IT-Infrastruktur die passende Unterstützung für das Energiemanagement in einem lernenden Netzwerk angeboten werden.

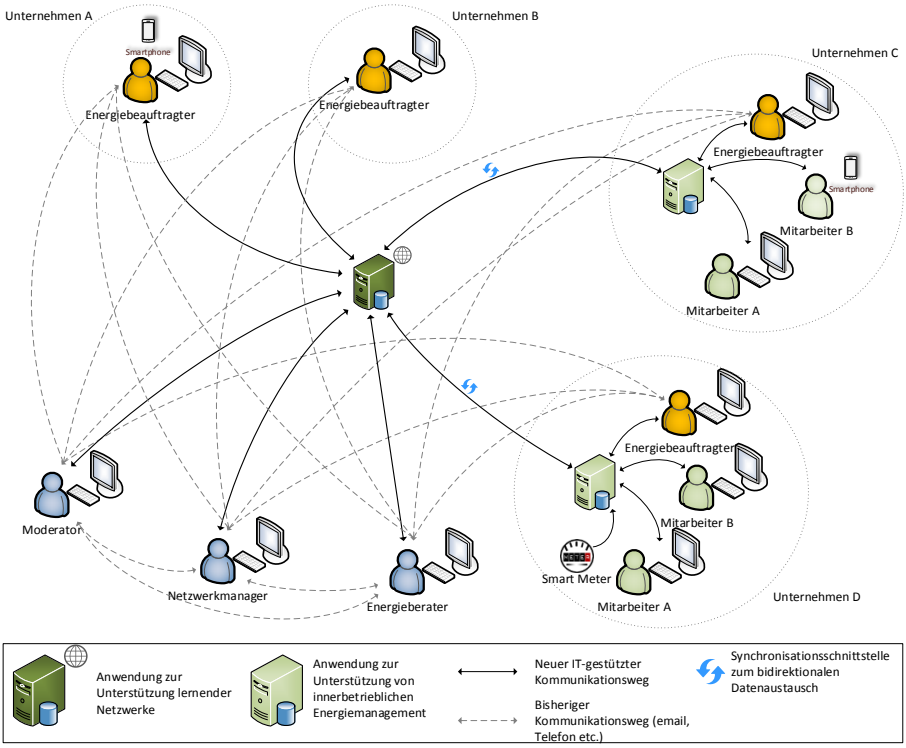


Abbildung 2: Schematische Darstellung EnergieNetz

Schnittstellen

Der zentral gehaltene Datenbestand, der durch die Verwaltung von Stammdaten auf einem datenbankgestützten Server und die einzelnen problembezogenen Softwaretools sowie die Anbindung gängiger Energiecontrolling-Systeme entstehen soll, ist intern untereinander verknüpft und bildet die Domäne des individuellen Betriebs ab. Um den bisherigen Kommunikationsaufwand innerhalb eines lernenden Netzwerks auf ein Minimum zu reduzieren, bietet die zentrale Anwendung standardisierte Webschnittstellen zum bidirektionalen Datenaustausch an und bildet im Zusammenspiel mit der dazugehörigen Webplattform ein mächtiges Instrument zur Optimierung und Unterstützung der Workflows innerhalb eines lernenden Netzwerks. Durch die Nutzung einer erweiterbaren Schnittstellentechnologie wird auch die Anbindung an weitere Systeme ermöglicht (BUIS/ERP).

Neben der genannten zentralen Anwendung existieren verschiedene mobile Anwendungen, die die Benutzer z.B. bei den für das Energiemanagement wichtigen Kontrollgängen oder Datenaufnahmen auf dem Firmengelände unterstützen sollen.

Wissen

Die Grundlage aller Tools sind wissenschaftlich belegte und erprobte Methoden, die nachvollziehbar präsentiert werden sollen. Ferner soll ein Vorgehensmodell entwickelt werden, das auf die unterschiedlichen Wissensstände der KMU reagieren kann. Gleichzeitig soll sich eine stetig wachsende Gemeinschaft etablieren, die einen regen Austausch von Ideen und Maßnahmen zur Energieeffizienz über das Wissensportal forciert. Neben den Ergebnissen aus diesem Vorhaben soll auch aus Forschungsprojekten mit ähnlicher Zielsetzung die Möglichkeit gegeben werden, ihre Ergebnisse in einem „Anwenderforum“ zu präsentieren. Dazu soll auf technischer Seite eine Webplattform in Form eines Forums für die wachsende, fachlich orientierte Community des Modell Hohenlohe e.V., bestehend aus Ingenieuren und Unternehmen, entstehen, um methodisches und technisches Wissen suchen, austauschen und einbringen zu können (siehe Abbildung 3). Daneben soll die Community des Partners HTW Berlin bestehend aus Softwareentwicklern und angehenden Softwareentwicklern (Studierende) ebenfalls in dem Forum integriert werden. Somit entsteht eine ganzheitliche Gemeinschaft aus Anwendern, Technikern und Softwareentwicklern, welche neben dem Gedankenaustausch im Forum zu Themen, auch speziellere Problemstellungen oder Anforderungen in Form von Arbeitskreisen oder direkten Aufträgen (Marketplace) bearbeitet und umsetzt. Somit soll die hier entstehende Softwarelösung auch über die Projektlaufzeit hinaus durch weitere Softwarekomponenten/Tools ergänzt werden.

Eine Einbeziehung neuer Gemeinschaftsmitglieder soll durch die Präsentation auf Konferenzen und bei Unternehmensnetzwerken (Handelskammern, regionale Verbände), sowie der Anbindung an Social Media Plattformen (XING, Twitter, Facebook, etc.) erfolgen, wodurch ein Wachstum der Gemeinschaft angestrebt wird.

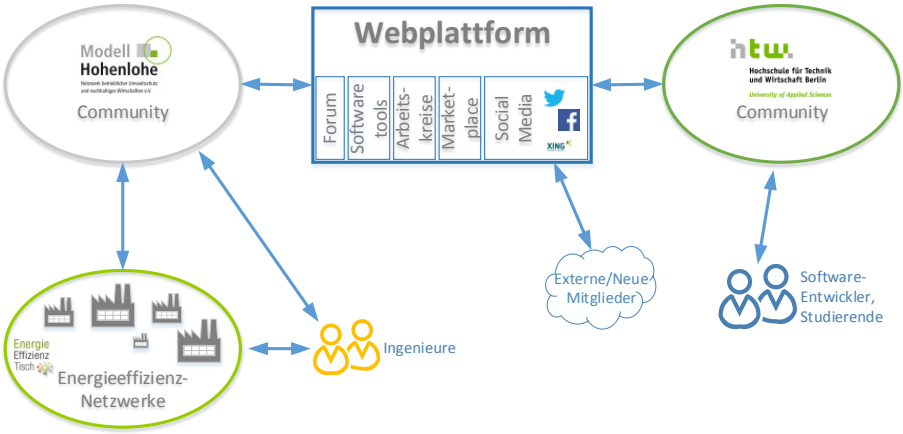


Abbildung 3: Webplattform als Bindeglied der Communities

3 Technische Umsetzung

Als Grundlage für die Entwicklung des firmeninternen Servers, sollen die Ergebnisse der im Projekt OpenResKit entwickelte Client-/Server-Architektur herangezogen werden. Die Architektur stellt dabei eine modular erweiterbare Domänenschicht zur Datenhaltung und -verteilung zur Verfügung.

Zur Entwicklung der Webplattform wird die ASP.NET Technologie herangezogen, welche auf dem Microsoft .NET Framework aufsetzt und es erlaubt, lernende Netzwerke mit den dazu gehörenden Unternehmen und deren Stammdaten (Energiebezüge, Energie-relevante Anlagen, Maßnahmen etc.) zu verwalten. In ihrer Gesamtheit bilden diese Daten, umgesetzt durch Plain Old CLR Objects (POCO), das Domänenmodell der zugrundeliegenden Anwendungsdomäne ab. Während das Domänenmodell dem objekt-orientierten Paradigma unterliegt, wird die Persistierung der Daten mit Hilfe eines relationalen Datenbankmodells realisiert. Um diesen Paradigmenwechsel zu vollziehen, wird ein objektrelationaler Mapper auf Basis von Microsoft Entity Framework (EF) verwendet, der eine Abstraktionsschicht zwischen der Domänenschicht und der Persistenz (Structured Query Language (SQL) Datenbank) bietet, die eine direkte Interaktion mit der Datenbank überflüssig macht. Da damit zu rechnen ist, dass die Webanwendung eher gelegentliche Anwendung im betrieblichen Alltag findet, stellen sich besondere Anforderungen an die Gebrauchstauglichkeit der Anwendung. Durch eine einfache Bedienung mit klar verständlichen Aufgaben, die einfach zu erlernen sind, soll so die Akzeptanz der Anwendung gegenüber der Nutzer erhöht werden. Realisiert werden soll dies durch gängige Bedienungskonventionen, Gestaltungsvorlagen sowie eine hohe Performanz mit einer vielfältigen Menge an Interaktionsmöglichkeiten über die Benutzungsoberfläche. Genutzt wird hierzu u.a. das Konzept der sog. Single-Page-Application (SPA), wonach die Anwendung nur aus einem einzigen Hypertext Markup Language (Html)-Dokument besteht, dessen Inhalte dynamisch mittels JavaScript und Asynchronous JavaScript and XML (Ajax)-Technologie nachgeladen werden. Dafür muss seitens des Servers ein entsprechender Webservice angeboten werden, der es gestattet, über eine standardisierte Schnittstelle Daten mit einem Webbrowser auszutauschen. Dazu implementiert der Server einen Representational State Transfer (REST)-Webservice mit Hilfe von Web API, der es erlaubt die auf dem Server persistierten Daten in Form von JavaScript Object Notation (JSON) bzw. Extensible Markup Language (XML) nicht nur einem Webbrowser sondern auch anderen Applikationen wie z.B. mobile Clients (Smartphone/Tablet) zur Verfügung zu stellen.

4 Webplattform

Das hier vorgestellte Projekt EnergieNetz befindet sich derzeit in der Phase der Implementierung der Webplattform, wobei ein iteratives Vorgehen verfolgt wird, d.h., dass Anforderungsanalyse, Konzeption und Umsetzung in einem sich wiederholenden Prozess ablaufen. Nach derzeitigem Stand und einem Top-down-Ansatz zugrunde liegend steht die Umsetzung eines mandantenfähigen Netzwerkmanagements[2] mit den

[2] Hiermit ist die Verwaltung von lernenden Netzwerken gemeint.

dazugehörenden Unternehmen und Benutzern im Fokus der Entwicklung. Das bedeutet, dass je nach Autorisierung des am System angemeldeten Anwenders, verschiedene Rechte gelten. Danach verfügen die Anwender - Moderator, Netzwerkmanager und Energieberater - über Lese- und Schreibrechte auf Netzwerkebene, während die Energiebeauftragten der Unternehmen lediglich Rechte auf ihr eigenes Unternehmen verfügen. Aktuell erfolgt die Umsetzung eines Dashboards zur Visualisierung von Energieverbrauchsdaten sowie CO_2-Emissionen auf Netzwerkebene. Die Benutzungsoberfläche, umgesetzt mittels responsive Webdesign[3], erlaubt über eine linksbündige Navigationsleiste zwischen den verschiedenen im System vorhandenen lernenden Netzwerken und deren Unternehmen zu navigieren (siehe Abbildung 4). Als Anwendungsbeispiel wurden die anonymisierten Daten eines in der Vergangenheit erfolgreich durchgeführten lernenden Netzwerks in das System eingepflegt. Das Dashboard soll dem Netzwerkmanager, Moderator oder Energieberater die Möglichkeit geben sich einen schnellen Überblick über den derzeitigen Stand der betreuten lernenden Netzwerke zu verschaffen.

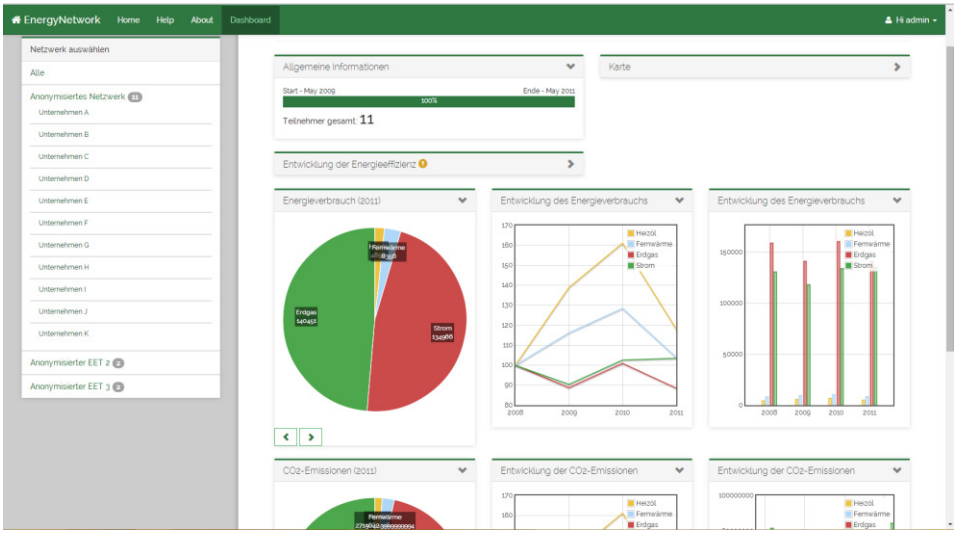


Abbildung 4: Screenshot Dashboard

5 Nächste Schritte und Ausblick

Bezüglich des Einsatzes der Software gibt es unter den Mitgliedsbetrieben des Modell Hohenlohe e.V. (KMU, Ingenieurbüros) bereits erhebliches Interesse. Bei KMU wird die günstige, einfache, modular auf die jeweiligen Bedürfnisse zugeschnittene Software in Verbindung mit einer über Netzwerke geführten Implementierung in den Betrieben zu einem Abbau von Hemmnissen führen, die bisher den Einstieg in Maßnahmen zur

[3] Responsive Webdesign ist ein gestalterischer und technischer Ansatz zur Erstellung von Webseiten. Dieser erlaubt die Anordnung und Darstellung der Webseite, entsprechend der Anforderungen und Möglichkeiten des jeweils verwendeten Endgeräts (Desktop, Smartphone und Tablet), anzupassen.

Energieeffizienzsteigerung und zum Abbau des klimaschädlichen CO_2 entgegenstehen. Eingebunden in eine bundesweite Community werden sie leichten Zugang zu vorhandenem Wissen bzgl. der Energieeffizienzsteigerung in KMU haben und mittels der entwickelten Software das verfügbare sowie neues Wissen in den Unternehmen transparent und leicht zugänglich über lange Zeiträume aufbauen und konservieren können. Die bisher in KMU wenig verbreitete Dokumentation von Energieverbrauchszahlen und Effizienzfortschritten wird einen Benchmark ermöglichen, der die Unternehmen zu weiteren Anstrengungen im Hinblick auf klimaschützende Maßnahmen motivieren wird. Außerdem aufbereitet und bündelt das Vorhaben dauerhaft in strukturierter Art und Weise bekanntes Methodenwissen aus Wirtschaft und Wissenschaft auf KMU und ermöglicht darüber hinaus direktes Feedback von den KMU in den Communities von Softwareentwicklern und Ingenieurbüros. Dadurch wird es ermöglicht auf KMU zugeschnittene Werkzeuge sich ändernden Bedürfnissen unter Berücksichtigung der technischen Weiterentwicklung fortlaufend anzupassen. Dabei werden der Open Source-Ansatz und die bis zum Ende des Projekts institutionalisierte Community dazu beitragen, dass vorhandener Sachverstand verschiedenster Disziplinen erschlossen werden kann und sich die klimaschützenden Wirkungen nicht nur verstetigen, sondern potenzieren werden.

Literaturverzeichnis

[Bu09] Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (Hrsg.). (2009). Energieeffizienz – Die intelligente Energiequelle – Tipps für Industrie und Gewerbe. Berlin.

[BB10] Bundesministerium für Wirtschaft und Technologie (BMWi), Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (Hrsg.). (2010). Das Energiekonzept der Bundesregierung 2010 und die Energiewende 2011. München.

[BB12] Bundesministerium für Wirtschaft und Technologie (BMWi), Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit (BMU) (Hrsg.). (2012). Erster Monitoring-Bericht „Energie der Zukunft". München.

[DLW08] Demir, S., Lotter, M. & Wohlgemuth, V., (2008). Durchführung einer Stoffstromanalyse als Ausgangspunkt für eine Potenzialanalyse mit den Schwerpunkten Material- und Energieeffizienz bei der PanTrac GmbH. In: Konzepte, Anwendungen, Realisierungen und Entwicklungstendenzen betrieblicher Umweltinformationssysteme (BUIS). Aachen: Shaker Verlag, S. 213-228.

[De06] Deutsche Bundesstiftung Umwelt (DBU) (Hrsg.). (2006). Energie effizient – Wie Industrie und Gewerbe Energie sparen können.

[Ei11] Eichhammer, W., Kohlhaas, M., Neuhoff, K., Rohde, C., Rosenberg, A. Schlomann, B. (2011). Untersuchung des Energieeinsparpotentials für ein Nachfolgemodell ab dem Jahr 2013ff zu Steuerbegünstigungen für Unternehmen des Produzierenden Gewerbes sowie der Land- und Forstwirtschaft bei der Energie- und Stromsteuer – Endbericht. Berlin: Bundesministerium der Finanzen (BMF).

[FH11] Fraunhofer ISI, IREES, Hassan, A. (Hrsg.). (2011). Möglichkeiten, Potenziale, Hemmnisse und Instrumente zur Senkung des Energieverbrauchs und der CO2-Emissionen von industriellen Branchentechnologien durch Prozessoptimierung und Einführung neuer Verfahrenstechniken - Schlussbericht. Karlsruhe u.a.

[HSR11] Hirzel, S., Sontag, B., Rohde, C. (2011). Kurzstudie - Betriebliches Energiemanagement in der industriellen Produktion. Fraunhofer-Institut für System- und Innovationsforschung (ISI). Karlsruhe.

[In13] Institut für Mittelstandsforschung (IfM) Bonn (2013): Abgerufen am 19. März 2014 von http://www.ifm-bonn.org/statistiken/mittelstand-im-ueberblick/#accordion=0&tab=1

[IT10] Irrek, W.; Thomas, S. (2010): Markttransformation und politische Instrumente. In: Pehnt M (Hrsg.): Energieeffizienz – Ein Lehr- und Handbuch. Heidelberg: Springer S. 35-86.

[Jo11] Jochem, E., Herbst, A., Mai, M., Reitze, F., Toro, F. A. (2011). Untersuchung des Energieeinsparpotentials für ein Nachfolge-Modell ab dem Jahr 2013ff zu Steuerbegünstigungen für Unternehmen des Produzierenden Gewerbes sowie der Land- und Forstwirtschaft bei der Energie- und Stromsteuer - Gutachten. Karlsruhe: Institut für Ressourceneffizienz und Energiestrategien (IREES GmbH).

[LBG12] Lieback, J.; Buser, J.; Gnebner, D. (2012): Leitfaden zum effizienten Energiemanagement - über drei Stufen in 18 Schritten zum effizienten Energiemanagement. GUTcert.

[Um12] UBA (2012): Umweltbundesamt. Nationale Trendtabellen für die deutsche Berichterstattung atmosphärischer Emissionen. Abgerufen am 19. März 2014 von http://www.umweltbundesamt.de/sites/default/files/medien/419/dokumente/2012_12_13_em_entwicklung_in_d_trendtabelle_thg_v1.2.0.zip

[WS12] Weissenbach, Kurt; Szilinski, Jürgen (2012): Abschlussbericht - Erster EnergieEffizienz-Tisch für KMU. Abgerufen am 19. März 2014 von http://www.modell-hohenlohe.de/download.php?file=af3f27e0791475e49d6975c3334c7e8e

[Wo14] Wohlgemuth, V., Krehahn, P., Ziep, T., Schiemann, L. (2014): Entwicklung eines Open-Source basierten Baukastens zur Unterstützung und Etablierung der Ressourceneffizienz in produzierenden KMU. In: Konzepte, Anwendungen, Realisierungen und Entwicklungstendenzen betrieblicher Umweltinformationssysteme (BUIS) – Konferenzband zu den 6. BUIS-Tagen. Aachen: Shaker Verlag.

Herausforderungen bei der Redigierung und Versionierung fortlaufender Massendaten

Thomas Gutzke, Frank Reußner

envi-systems GmbH
An der Eschollmühle 28
64297 Darmstadt
gutzke@envi-systems.com
frank@reussner.org

Abstract: In der Wasserver- und -entsorgung werden über Prozessleitsysteme in den meisten der jeweils über 10.000 Anlagen (Wasserwerken und Kläranlagen) mehrere hundert Datenpunkte aufgezeichnet – und das z. T. im Minuten- und Sekundentakt. Zur Prozessüberwachung, -steuerung und -optimierung werden von zahlreichen dieser Datenpunkte die gesamten Datenreihen kontinuierlich und vollständig aufgezeichnet und regelmäßig ausgewertet. Für eine nachhaltige Bewirtschaftung werden diese Daten nicht nur zur zeitnahen Überwachung und Steuerung benötigt. Für die Beantwortung zahlreicher fachlicher Fragestellungen, sind dieses Daten kontinuierlich und persistent aufzuzeichnen, so dass beispielsweise Entwicklungen über mehrere Jahrzehnte innerhalb einer Auswertung erfolgen können. Da es systembedingt vorkommen kann, dass in den Datenreihen Fehler auftreten (Datenlücken, Ausreißer, Drift) werden Redigierungen erforderlich. Da die Datenreihen sich einerseits kontinuierlich fortschreiben, aber z. T. auch mehrfach redigiert werden müssen, exisiteren hohe Anforderungen an die Redigierungsmöglichkeiten, die Datenarchivierungskonzepte, die Versionierbarkeit und nicht zuletzt an die Performance..

Dieser Beitrag stellt Möglichkeiten einer performanten und gleichzeitig diagrammgestützten Datenredigierung vor. Alle Redigierungsprozesse müssen aus Gründen der Nachvollziehbarkeit versioniert in einem Datenbanksystem abgelegt werden. Insbesondere die Tatsache, dass Datenreihen nie abgeschlossen werden, sondern sich kontinuierlich fortschreiben, stellt hohe Anforderungen an die Performance der Anwendung sowie der Datenbank auf der einen und an die Versionierung aller Redigierungsprozesse auf der anderen Seite. Beide Anforderungen beeinflussen sich dabei gegenseitig negativ. Dieser Beitrag beschreibt die Herausforderungen, die sich hieraus ergeben haben, stellt aber auch einen Ansatz vor, der im ersten Schritt in einer „einfachen“ aber funktionalen Variante erfolgreich umgesetzt wurde.

1 Einführung

In Deutschland wird die Wasserversorgung von ca. 6.500 Wasserversorgungsunternehmen sichergestellt. Im Bereich der Abwasserentsorgung sind über 10.000 Anlagen in Betrieb. In beiden Bereichen fallen zahlreiche umweltrelevante Daten an, die kontinuierlich erfasst, verwaltet und ausgewertet werden müssen. Die für den Betrieb relevanten Daten werden mittlerweile in den meisten Anlagen über Prozessleitsysteme erfasst. Für die Betreiber dieser Anlagen sind diese Daten wichtig, um einerseits im Rahmen der Berichtspflichten gegenüber den Behörden den fachgerechten Betrieb zu dokumentieren und anderseits einen wirtschaftlichen und ökologischen Betrieb der Anlage durch gezielte Steuerung zu erreichen. Die Daten werden dabei über Sensoren erfasst und meist über Steuerkabel (4-20mA) oder über eine Datenfernübertragung unredigiert an das interne Speichersystem übertragen. Leistungsstarke Prozessleitsysteme ermöglichen auf dieser Basis eine Adhoc-Überwachung aller angeschlossenen Anlagenteile. Viele Anlagen erlauben darüber hinaus eine aktive Anlagensteuerung. Über die letzten Jahrzehnte wurden die Theorien, Techniken und Verfahrensweisen im Betrieb dieser Anlagen so weit verfeinert, dass diese einen sehr hohen Wirkungsgrad aufweisen. Um dies zu erreichen, wurde u. a. ein immer dichteres Datenerfassungsnetz aufgebaut, welches zusätzlich die einzelnen Daten in immer kleiner werdenden Intervallen aufnimmt. Reichten früher Tageswerte aus, so werden heutzutage mehr und mehr Daten im Minuten- und z. T. Sekundentakt erfasst. Die Daten werden in der Regel in einem Ringspeicher abgelegt, wobei die Datenbestände auf Grund ihres Datenumfangs nach festlegbaren Zeiträumen ausgedünnt bzw. überschrieben werden. Die Archivierung erfolgt daher häufig in separaten Zusatzsystemen.

Hinsichtlich der Archivierung, Redigierung und Auswertung weisen diese Systeme jedoch große Schwächen auf. Werksmeister und Betriebsleiter stehen regelmäßig vor der Aufgabe, die Anlagenkomponenten in Bezug auf technische Leistungsdaten auszuwerten und effizient einzustellen. Auch die Erstellung von Berichten (z. B. im Rahmen der Eigenkontrollverordnung) erfordert Zugriff auf betriebsrelevante Datenbestände in jährlichen Zyklen. In beiden Fällen muss der Ringspeicher direkt angezapft werden, wobei die Daten i. d. R. vom Ringspeicher nach Excel kopiert und dort manuell weiterverarbeitet werden. Fehler in den Datenbeständen müssen hier z. T. sehr aufwändig und fehleranfällig korrigiert werden. Die diversen erforderlichen Auswertungen in Form von Diagrammen und Berichten erfordern zudem einen hohen manuellen Zeitaufwand.

2 Zielsetzung

Bei dem hier vorgestellten Projekt handelt es sich um eine ausgewählte Komponente des Softwaresystems „GW-Manager“ - einer Eigenentwicklung der Firma envi-systems GmbH (en14). Das im Jahr 2001 ursprünglich nur für den Bereich der Wasserversorgung entwickelte System kommt zunehmend auch in den Bereichen Abwasserentsorgung/Kläranlagenmanagement für Datenhaltungs- und Auswertungsaufgaben zum Einsatz. Das System setzt an der Schnittstelle zwischen Prozessleitsystem und den Optimierungsaufgaben der Fachanwender an und löst dabei das Problem des Handlings von Massendaten und deren Datenredigierung. Es verfolgt dabei folgende Ziele (siehe Abbildung 1):

- Automatisierte Datenübernahme in eine für Massendaten geeignete Datenbank
- Vollständige Langzeitarchivierung aller relevanten Daten
- Visuell unterstützte Datenredigierung
- Variantenbasierte Zeitreihenversionierung
- Bereitstellung komfortabler Analyse- und Auswertungsfunktionalitäten

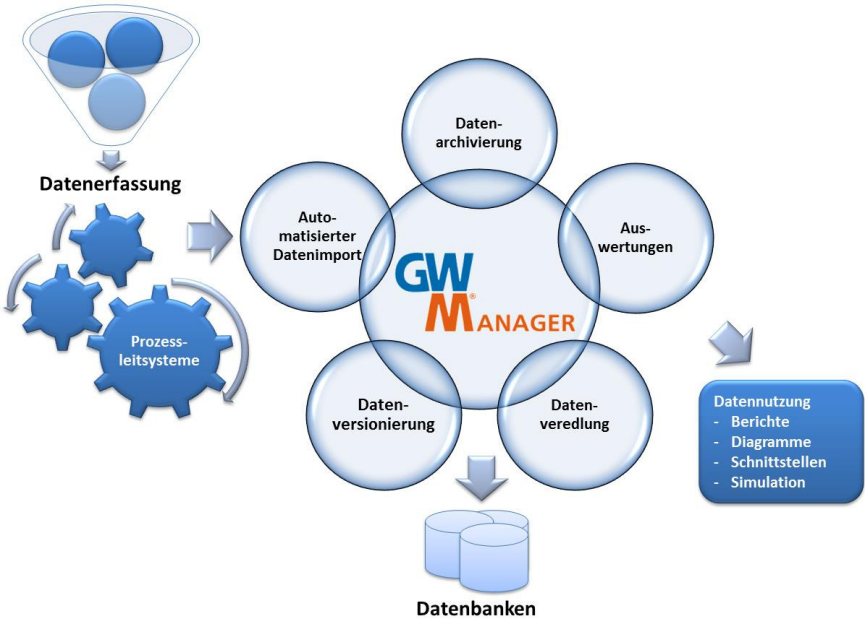


Abbildung 1: GW-Manager im Verbundsystem

3 Redigierung von Zeitreihen

Die Anforderungen, die an ein System gestellt werden, das Umweltmassendaten sowohl erfasst als auch bearbeitet, sind vielfältig. Nicht selten treten innerhalb einzelner Zeitreihen Fehler auf, die folgende Ursachen haben können:

- falsch kalibrierte Sensoren
- falsch zugeordnete Sensoren
- Stromausfall
- Stromschwankungen
- Alterungserscheinungen oder Verschmutzung von Sensoren (Drifts)

Für die Bereinigung dieser Fehler müssen Werkzeuge zur Bearbeitung von Ausreißern, Lückenschließung sowie Trend- bzw. Driftbereinigung bereitgestellt werden (siehe Abbildung 2). Die Bereinigung sollte zudem visuell unterstützt und performant umgesetzt werden.

Daraus folgen Anforderungen an ein System zur visuell unterstützten Datenredigierung, die meist gegenläufige Ziele haben. So wirkt sich z. B. eine Fokussierung auf die benutzerfreundliche Gestaltung der Datenredigierung unter Umständen negativ auf die Performanz des Gesamtsystems aus. Neben einer visuellen Datenkontrolle, die den Benutzer unterstützt, ist eine visuelle Datenbearbeitung (Drag ‘n Drop) eine weitere wichtige Anforderung. Hinsichtlich der Datenhaltung ist es zudem erforderlich, neben den erfassten Rohdaten auch eine Versionierung der redigierten Datenbestände zuzulassen.

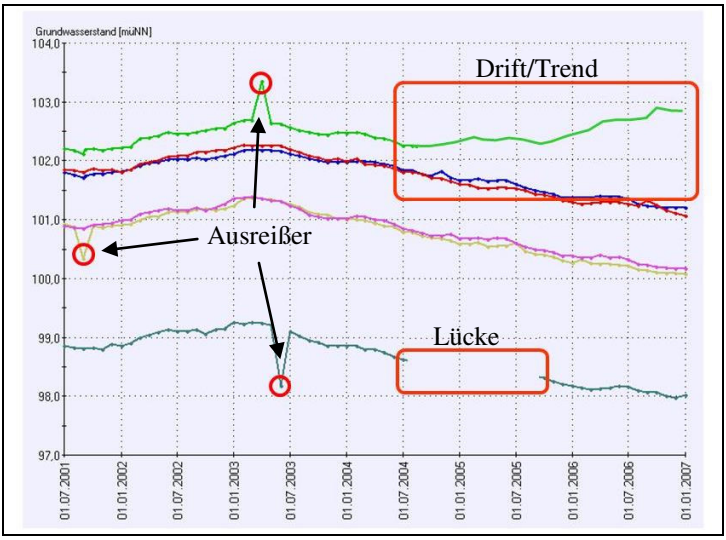


Abbildung 2: Erfordernisse der Datenredigierung

4 Versionierung fortlaufender Zeitreihen

Ziel einer Versionierung von Zeitreihen ist zum einen die Beurteilung des Qualitätszustands eines Datenpunkts. So kann beispielsweise zwischen Rohdaten oder verifizierten Daten unterschieden werden. Zum anderen ist das Ziel, mögliche Bearbeitungsschritte transparent und nachvollziehbar zu machen. Damit kann z. B. eine Bearbeitung von Datenpunkten bei Bedarf auch rückgängig gemacht werden. Denn häufig stellt sich erst im Nachhinein heraus, dass Annahmen sowie die vorgenommenen Redigierungen falsch waren.

Ein Beispiel für eine typische bearbeitete Zeitreihe ist in Abbildung 3 dargestellt. Auf der x-Achse liegen alle erfassten „Roh“-Daten. Insgesamt wurde die Zeitreihe in sieben einander folgenden Schritten bearbeitet. Alle Bearbeitungen sind in drei Zeiträumen vorgenommen worden. Je nach dem, ob man dabei alle sieben Schritte wieder nachvollziehen bzw. rückgängig machen möchte, folgen verschiedene Ansätze zur Versionierung dieser Zeitreihe.

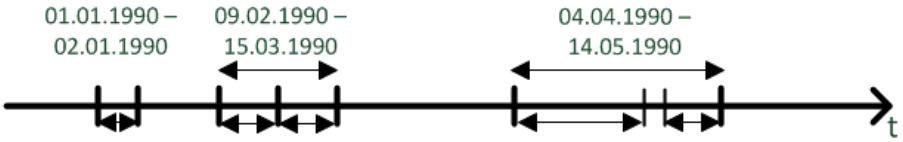


Abbildung 3: Änderungsabschnitte innerhalb einer Zeitreihe (Al12)

Die Herausforderungen bei der Versionierung von Zeitreihen liegen einerseits bei der Performanz und anderseits beim Speicherplatzbedarf, wobei sich beide Aspekte gegenseitig negativ beeinflussen. Bei der Entwicklung des GW-Managers sind für den Aufbau eines Versionierungssystems einige vielversprechende Arten der Versionierung analysiert worden, die diesen beiden Ansprüchen Rechnung tragen. Die klassischen Ansätze zur Datenversionierung beschäftigen sich u. a. mit der platzsparenden Datenhaltung sowie der performanten Datenrecherche (KSW86). Die gängigsten Muster sind:

- Vorwärtsorientiert,
    - Unterschiede von Version zu Version
    - Unterschiede von Ursprungsversion zu Version
- Rückwärtsorientiert
    - Unterschiede von Version zu Version
    - Unterschiede von Version zu Ursprungsversion
- Vorgang speichern statt Attributänderung in den Unterschieden

Eine vielversprechende Art ist z. B. die Erstellung einer Baumstruktur zu jedem Datenpunkt und damit zu einem graphentheoretischen Wald für jede Version der Zeitreihe (siehe Abbildung 4). Hier können dann Varianten aufeinander aufsetzen. So kann beispielsweise bei der Datenredigierung der Workflow abgebildet werden, bei dem als erster Schritt Ausreißer eliminiert werden, als zweiter Schritt Datenlöcher gefüllt und dann Trends/Drifts bereinigt werden. Dabei würde die Trendbereinigung als Variante auf die Daten der anderen beiden Varianten als Basis zurückgreifen.

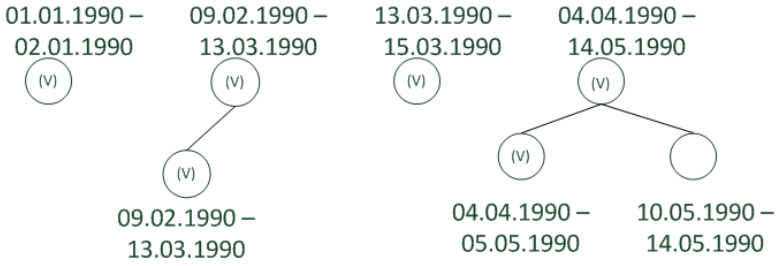


Abbildung 4: Hierarchische Datenstruktur von Zeitreihenabschnitten (Al12)

Jeder Ansatz hat seine Vor- und Nachteile in Bezug auf Redundanz und Komplexität der zu erstellenden Version. Außerdem haben die Arten auch unterschiedliche Anfälligkeiten bezüglich möglicher Konflikte durch Verklemmung (Deadlock). Allen gemeinsam ist aber, dass sie entwickelt wurden, um einen bestehenden Datenblock zu versionieren.

Die zu versionierenden Zeitreihen sind jedoch keine abgeschlossenen Datenblöcke, sondern werden fortlaufend erweitert, da die Sensoren und Logger auch fortlaufend neue Daten erheben. Durch diesen Umstand ist eine Zuordnung der neu hinzukommenden Daten zu einer bestimmten Version nicht eindeutig. Eine andere Art der Variantensysteme ist z. B. die Quellcodeverwaltung (z. B. SVN, GIT). Auch diese sind wenig geeignet fortlaufende Zeitreihen adäquat zu versionieren. Dies liegt daran, dass bei einer Zeitreihe die einzelnen Daten mehr gemeinsamen Bezug besitzen, als die Dokumente bzw. Textzeilen bei einer Quellcodeverwaltung. Beispielsweise beziehen sich Schritte zur Datenbereinigung auf dieselben Daten (Ausreißer- mit Trendbereinigung). Dies ist aber im Kontext einer Quellcodeverwaltung nicht abbildbar, da dort die zweite Änderung die Daten absolut verändert. Würde man in dem Beispiel erst die Trendbereinigung durchführen und dann die Ausreißerbereinigung, dann könnte man zwar die letzte Bereinigung zurücknehmen, nicht aber nur einzeln die Trendbereinigung. Die folgende Tabelle 1 zeigt eine Bewertungsmatrix der unterschiedlichen Variantensysteme in Bezug auf fortlaufende Zeitreihen.

Legende: **++: geeignet** **+ : bedingt geeignet** **- : eher ungeeignet** **--: ungeeignet**	**Speicherplatz-bedarf**	**Deadlocks**	**Abhängige Versionen**	**Zuordnung neu erhobener Daten**
Versionsabschnittsbildung an einer Zeitreihe	+	++	+	++
Autarke Variantenbildung	--	++	-	--
Hierarchische Variantenbildung	+	-	++	+
Variantenbildung mit Bezug auf n-Varianten	++	-	+	+
Quellcodeverwaltung	+	+	--	-

Tabelle 1: Vor- und Nachteile der Zeitreihen-Versionierungsstrategien

5 Projektstand

Bei der Umsetzung des Projekts wurde ein Schwerpunkt auf die Performanz gelegt, so dass der Benutzer effizient mit seinen Massendaten arbeiten kann. Hierfür wurden aus dem Bereich des Big Data die klassischen Werkzeuge analysiert. Durch die Anforderungen der Interaktion ist aber keines der gängigen Big Data Werkzeuge für dieses Projekt einsetzbar. Die Werkzeuge „Hadoop“, „Big Table“, „Map Reduce“ etc. sehen alle den Informationsfluss in eine Richtung vor, so dass zwar schnell Daten-Aggregationen aus-

geführt werden können (DG08), der Informationsfluss zurück in die Datenhaltung jedoch nicht vorgesehen ist bzw. keine Vorteile gegenüber herkömmlichen Werkzeugen (wie relationalen Datenbanksystemen) aufweist. Aus diesem Grund wurde das bestehende Datenhaltungssystem des GW-Managers (ein relationales Datenbanksystem auf Basis des Microsoft SQL-Servers oder alternativ Oracle) genutzt, das bereits alle wesentlichen Aufgaben erfüllt. Hinsichtlich der Verarbeitung von Massendaten, wurde die Datenbank sukzessive optimiert. Hierbei wurden insbesondere die Tabellen-Indexierung sowie die Auslagerung von SQL-Abfragen in Stored Procedures stark ausgebaut: Die Optimierungen, die vorgenommen wurden, sind zum einen die Überprüfung der Indizes einzelner Tabellen, vor allem jedoch die Optimierung aller SQL-Abfragen. Neben der Syntax der SQL-Abfragen konnte durch die Verlagerung der Abfrage aus dem Programmcode in die Datenbank selbst – als Stored Procedure – die Geschwindigkeit der Abfrage um den Faktor 15 erhöht werden.

Auch hat sich herausgestellt, dass im Programmcode selbst (.Net 4.5) noch eine Reihe von Optimierungsmöglichkeiten liegt. Trotz der in .Net 4.5 enthaltenen Garbage-Collection (GC) ist die Datenhaltung der Zeitreihen halbautomatisch implementiert worden. So wird zwar die GC normal von der Anwendung genutzt, wenn aber Zeitreihen entfernt werden, wird explizit die GC angesteuert, so dass der Speicher schnellstmöglich freigegeben wird. Dies ist erforderlich, da .Net-Anwendungen immer nur einen limitierten Hauptspeicher belegen dürfen. Bei 32-bit Betriebssystemen (die das Projekt auch bedienen soll) sind dies ca. 1,5 GByte maximale Speicherbelegung.

Ein weiterer wichtiger Schritt der Umsetzung lag in der Suche nach einem geeigneten Chart-Werkzeug. Hier hat sich gezeigt, dass die meisten Hersteller von sich aus nicht angeben, mit wie vielen Datenpunkten das Werkzeug noch effizient bedienbar ist. Lediglich einige Werkzeuge sind auf den Einsatz von Massendaten (> 500.000 Datenpunkten) ausgelegt. Die Wahl ist auf das Werkzeug TeeChart von Steema gefallen (St13). Hier sind mit den aktuellen Tests keine Performanzprobleme mit großen Zeitreihen aufgetreten. Auch eine Bearbeitung mittels Drag ‘n Drop kann durchgeführt werden.

6 Fazit und Ausblick

Das hier vorgestellte Projekt ist abgeschlossen und wird bereits produktiv eingesetzt. Die Massendaten können dabei direkt aus den Sensoren und Loggern in das relationale Datenbanksystem des GW-Managers importiert werden. Auch können Prozessleitsysteme angebunden werden, so dass alle wasser- sowie abwasser-relevanten Daten vollständig verfügbar sind. Mit dem Schritt der Datenerfassung in das bestehende Datenhaltungssystem, wurden alle Anforderungen bezüglich Datenarchivierung erfüllt. Dabei mussten jedoch zahlreiche Datenbankoptimierungen vorgenommen werden, um die Performance zu verbessern. Die Verbesserung von Indexierungen sowie der Einsatz von Stored Procedures führten hierbei zu Performance-Steigerungen um den Faktor 15.

Die Komponente zur Datenredigierung wurde fertiggestellt, so dass hier nun leistungsstarke Funktionalitäten zur Verfügung stehen (siehe Abbildung 5). Die Komponente verfügt über Werkzeuge zur Datenredigierung (z. B. Datenübernahme aus anderen

Messstellen, lineare Trendbereinigung), direkte grafische Datenbearbeitung (Drag ‘n Drop) und komfortable Datenvisualisierung (Zoom, Diagrammtyp Linie, Balken und Punkt etc.). Zur Datenanalyse sind verschiedene Statistiken (Mittelwert, Varianz, gleitendes Mittel usw.) oder automatisierte Berichte ad-hoc erstellbar.

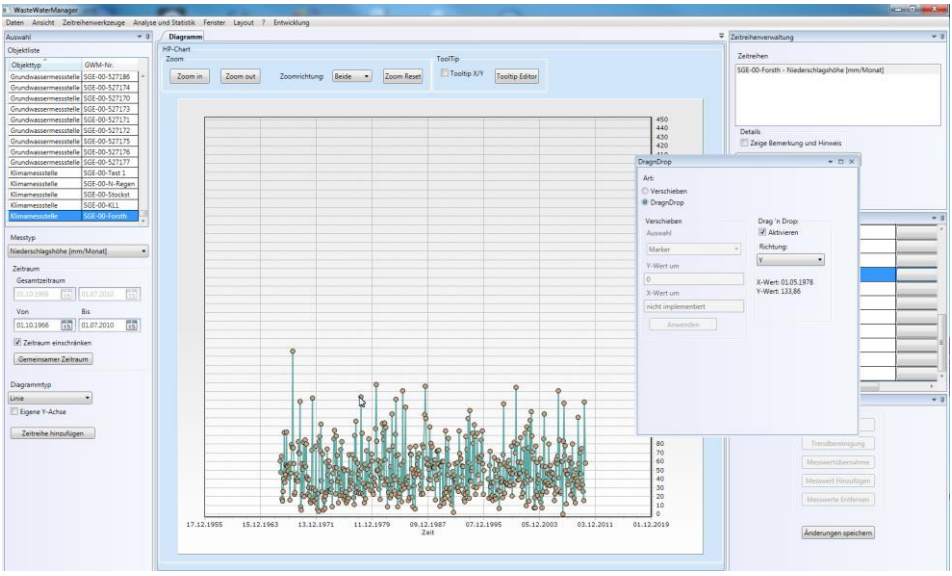


Abbildung 5: Datenredigierung mit dem GW-Manager

Die Versionierung redigierter Zeitreihen wurde in einer ersten „einfachen“ Variante realisiert. Die gewählte Variante unterstützt dabei lediglich eine Versionsebene. Die in diesem Beitrag vorgestellte „Baumstruktur“ (siehe Abbildung 4) soll Gegenstand zukünftiger Entwicklungen sein. Bei allen Weiterentwicklungen muss dabei die Wechselbeziehung zwischen einer performanten Datenbank-Anwendung und den hohen Anforderungen an eine fortlaufende Versionierung berücksichtigt werden.

Literaturverzeichnis

[Al12] Albrecht, Andreas: Entwicklung eines Prototyps zur hochperformanten grafischen Darstellung, Editierung und versionierbaren Speicherung von Massendaten. Abschlussarbeit Hochschule Darmstadt. Darmstadt, 2012.

[DG08] Dean, J.; Ghemawat, S.: MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), S.107-113. New York, 2008.

[en14] envi-systems GmbH: GW-Manager, Available at http://www.gw-manager.com/, 2014.

[KSW86] Klahold, P.; Schlageter, G.; Wilkes, W.: A general model for version management in databases. Gesamthochschule, FernUniversität Hagen, Hagen 1986. Available at: http://www.vldb.org/conf/1986/P319.PDF?origin=publication_detail [Zugegriffen am 21.April 2014].

[St13] Steema: Steema Support Central - Charting Components by Steema, Chart for .NET, VCL, Delphi, ActiveX, Java, PHP, Mobile Chart for Android and Phone 7. Available at: http://www.teechart.net/, 2013.

Environmental-Oriented Information Systems Design – The Concept of Life Cycle Impact Modelling and its Application to Cloud Computing

Florian Stiel, Frank Teuteberg

Fachgebiet Unternehmensrechnung und Wirtschaftsinformatik
Universität Osnabrück
Katharinenstraße 1
49074 Osnabrück
florian.stiel@uni-osnabrueck.de
frank.teuteberg@uni-osnabrueck.de

Abstract: Information technology is increasingly perceived not only as a driver of economic value creation, but also as a critical factor in the context of environmental sustainability. However, it is uncertain how the adoption of a new information technology within a specific organizational context can be evaluated from an environmental perspective. Consequently, there is a great need for methods relating the organizational use of IT with environmental assessments of different information system design alternatives. This work uses an outward-looking, problem solving and interdisciplinary research approach to address this problem. It aims to evaluate the environmental impact associated with the use of information technology within an organizational context, proposing an approach to compare different design alternatives for the implementation of an information system. Thus, existing information modelling approaches are extended by (life cycle) impact modelling; a concept that is grounded in the widely considered Life Cycle Assessment (LCA) approach. The methodological foundation for this work is therefore constructivist and design science oriented.

1 Introduction

Information technology (IT) is increasingly perceived not only as a driver of economic value creation, but also as a critical factor in the context of environmental sustainability. Accounting for approximately 2 percent of the global greenhouse gas emissions, IT is part of the problem; however, by addressing the other 98 percent with IT-enabled business transformations, IT can also be a possible part of the solution [GW08]. Today, the potential of IT to reduce environmental impact seems to be extensively covered in the literature and is qualitatively affirmed for a wide range of applications [Br12, Br13,

Me10, SRB13, Wa12, WBC10]. However, quantitative, outward-looking and problem solving research approaches to address the environmental outcome of a new information technology's roll out is still underrepresented [Br13, YT07]. It is uncertain how the adoption of a new information technology within a specific organizational context can be evaluated from an environmental perspective. Consequently, there is a great need for methods relating the organizational use of IT with environmental assessments of different information system design alternatives. This leads us to the following research question:

How can different information system design alternatives be compared from an environmental perspective?

This work uses an outward-looking, problem solving and interdisciplinary research approach to address this question [Br13]. It aims to evaluate the environmental impact associated with the use of information technology within an organizational context, proposing an approach to compare different design alternatives for the implementation of an information system. Thus, existing information modelling approaches [BB06, BSS09, FL04, Fr99] are extended by (life cycle) impact modelling; a concept that is grounded in the widely considered Life Cycle Assessment (LCA) approach. The methodological foundation for this work is therefore constructivist and design science oriented [He04, MS95].

A topical and still emerging example of new technological alternatives is cloud computing. Cloud computing provides the possibility to transform organizational processes by enabling a "ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources" [NI11]. Using cloud computing in business practices, as a way to reduce environmental impact, was discussed since about 5 years now [Be09, Ca13, CS13, Ga11, KMT12, Le12, WTM14]. The positive environmental effects result from improved load balancing or energetic scale effects within large-scale datacentres in particular. However, as for the application of IT in general, the question remains if the adoption of cloud computing is environmentally beneficial in specific organizational cases from the cloud users' perspective. Therefore, we will use the example of cloud migration to demonstrate the functionality of life cycle impact modelling.

Our work is structured as follows: In section 2, we review the terminology and scope of research as well as related works in the field of environmental-oriented information systems design involving cloud computing. In section 3 we introduce the concept of impact modelling. Section 4 covers the evaluation of the concept using cloud computing as an application example. Concluding remarks including implications and limitations as well as an outlook on future work is provided in section 5.

2 Research Foundation

Interdisciplinary research at the intersection of information systems, environmental and business science reveals several inconsistencies in denotation as well as terminology.

Environmental impact models and assessments show high sensitivity to inconsistent system boundary definitions often leading to incorrect conclusions and decisions [Su04]. Therefore, we will pay particular attention to the distinction between terminology and system definitions (Figure 1).

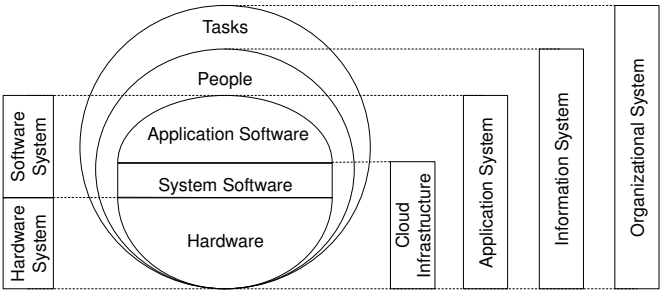


Figure 1 Systems Overview [BSS09, NI11, Te99, YBS08]

Organizational systems characterize the interaction of human, information and technology within an organization in order to fulfil certain tasks [BSS09]. Within this organizational system, *information systems* (IS) are used to fulfil the task to store, communicate and compute information [HL11]. IS involve human, information and technology, they thus constitute socio-technical systems [Te99]. The technological/IT part of the information system or *application system*, is composed of a non-physical *software system* (application and system software) and a physical *hardware system* [Te99]. The application system level is the technological level on which different technological alternatives can be compared.

A topical and recently discussed concept for an alternative implementation of application systems is cloud computing. From the perspective of IS design, cloud computing offers the opportunity to outsource parts of the application system, i.e. to reduce the amount of own physical hardware [Ak11, Le10, Nu11]. By applying one of the provided service models, the user gains access to the cloud infrastructure which is controlled and managed by the cloud provider. The application software is still located at the user, whereas the system software and hardware is located at the provider and applied via a permanent network access [NI11, YBS08]. The concept of cloud computing has thereby a significant impact on the design of information systems. Its importance for IS research is reflected by the broad interest in cloud computing at recent leading IS conferences (ICIS, AMCIS and ECIS). The relevance of cloud computing within the context of environmental sustainability is evaluated conducting a literature search to identify related work.

The literature search was conducted using a web based literature search [Br09]. The goal was to identify recent research outcomes as well as the application of cloud computing for environmental sustainability. The search was conducted querying the literature databases EBSCOhost, Sciencedirect and ISIwebofknowlege and Emerald with the search terms *sustainability*, *green* and *environment* combined with *cloud computing*. The statements within this chapter refer to the most representative and recent articles in this research area. Research on the positive environmental effects of cloud computing

commonly are found using the term “green cloud computing” [QM13]. Two major strategies for reducing environmental effects can be identified:

The majority of publications associate positive environmental effects of cloud computing with a higher energy efficiency of the datacentres, caused by highly integrated computer systems and networks, improved load balancing, energetic scale effects within large-scale datacentres, virtualization as well as energy and thermal aware cloud resource allocation [Be09, CS13]. [WTM14] estimate that the replacement of organizational on-side datacentres by cloud datacentres in this context leads to greenhouse gas reductions between 31-51% depending on the size of the replaced infrastructures. Moreover, the adjustments within cloud datacentres are estimated to result in a 25 %decrease in energy consumption by improved scheduling policies [Ga11] and an increase from 33 % to 74,8 % by large scale datacentre clustering [MNV12].

The second strategy to reduce environmental impact involves transferring computing capacities towards renewable energy sources [Le12]. Using simple migration techniques, around 30 % of the non-renewable energy consumption of US cloud infrastructure can be replaced by renewable energy, such as solar or wind power [Ma13]. [Ga14] performed a case study to estimate the reduction of greenhouse gas emissions by transferring energy-intensive computational work towards datacentres with renewable energy supplies. The case study suggests that up to 30 % CO_2 emissions caused by cloud datacentres may be saved by implementing environmental oriented routing strategies.

An attempt to consider environmental metrics, energy consumption models and cost models for the design of datacentre architectures from the provider perspective has been proposed on a case based level by [Ca13]. [Ca13] provide formal energy consumption models for cloud computing from the provider perspective. A formal CO_2 aware model was offered by [KMT12]. However, despite general progress in the field of green cloud computing, recent publications reveal that positive environmental effects are being not perceived by the potential cloud users; respondents “largely neglect ecological gains in particular” [GK13]. A consumer orientation in the field of green cloud computing has not yet taken place. [CWW13] underline the importance of an environmental aware modelling approach from the perspective of cloud computing consumers. This approach, however, does not exist so far. We contribute to this highly relevant research gap by introducing an approach for integrating environmental assessment into IS design.

3 Introducing the Life Cycle Impact Modelling Approach

3.1 Overview

LCA is used to counteract assess environmental effects within the earlier stages of the supply chain [Gu02, ISO06]. Traditionally, these effects do not occur within the scope of the system design. As an example, the environmental impact of an electrical car is not completely described by its electricity consumption but also by the environmental efficiency of its energy source, e.g. a coal-fired plant or a wind turbine. Assigning this

idea to the design of information systems led us to the development of a life cycle impact modelling approach for environmental-oriented information systems design.

The basic idea of the life cycle impact modelling is to help decision makers select the most environmental sustainable design alternative (represented by a To-Be-Model) for a current information system (represented by an As-Is-Model) considering not only internal factors like electricity consumption but also external factors like primary energy consumptions. Therefore, life cycle impact modelling constitutes a synthesis of process models (i.e. potentials models) [BSS09] and LCA models (i.e. inventory models) [ISO06]. Figure 2 provides an overview of the involved models as well as the numbered relations between them: (1) Initially, design alternatives for a current process (represented by an As-Is-Model) are specified using service oriented factual models. (2) & (3) Flow models (inventory models) are developed to represent physical flows associated with the design alternatives. (4) Environmental-impact models are derived from inventory models and provide a comprehensive overview of the environmental impact of the design alternatives. (5) Factual models can be further modified according to the environmental-impact model and inventory model or (6) a beneficial alternative is selected as a To-Be-Model for implementation.

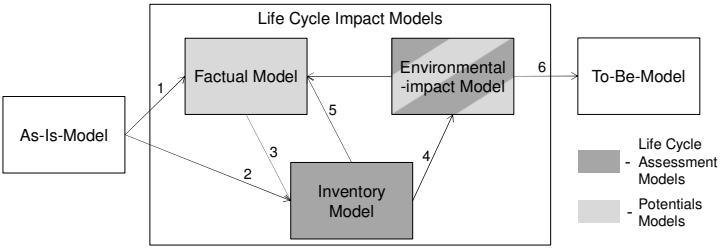


Figure 2 Life Cycle Impact Modelling Overview [ISO06][BSS09]

Proceeding from an As-Is-Model to the final implementation of the To-Be-Model involves knowledge of information modeling and LCA. Due to the type of this publication, we will restrict ourselves to the introduction of concepts that are directly relevant for the development of factual, inventory and environmental-impact models. For a deeper understanding of LCA and potentials modelling the relevant literature should be considered.

3.2 Potentials Models

Potentials models are used for the representation of different design alternatives [TB09]. According to [TB09] as well as [BSS09], potential design alternatives have to be investigated from two perspectives in order to provide a comprehensive overview of their financial and organizational consequences. Investigation should include: 1) a factual service oriented representation of the design alternatives depicting the configuration of activities, services and infrastructures using one or more factual models and 2) a value-based representation of the design alternatives depicting its economic consequences expressed by key figures within one or more value-based models.

The combination of both perspectives is required to consider the economic and business activity related consequences of a design decision. To create a factual representation, life cycle impact modelling adopts approaches advocated by [TB09] as well as [BSS09]. As the authors suggest, process modeling languages, e.g. EPC, Petri nets and UML-Diagrams, as well as the construction principles, i.e. explicit or implicit models, for the factual representation of design alternatives can be used in this regard.

A value-based representation of life cycle impact modelling focuses on environmental sustainability instead of economic sustainability. In order to avoid confusion, the environmental value-based representation of design alternatives will be called an *environmental impact model*. The environmental impact model depicts the ecological consequences of realizing the design alternative expressed by key figures. Whereas a value-based modeling considers process controlling methods, environmental-impact modeling considers LCA methods [Fi09, Gu02, ISO06] to evaluate the processes. Key figures to express the environmental impact of the design alternatives can be found among the various environmental impact categories such as [Gu02]: greenhouse gas emissions/climate change, primary energy consumptions, depletion of abiotic resources, depletion of biotic resources, stratospheric ozone depletion, impacts of land use/land competition, human toxicity, eco toxicity, photo-oxidant formation, acidification, eutrophication, impacts of ionizing radiation, odor, dust, noise, waste heat, casualties and desiccation. To derive these key figures, LCA methods involve a third perspective for investigating potential design alternatives.

3.3 Life Cycle Assessment

Life cycle assessment (LCA) is a method of compilation and evaluation of the potential environmental impacts of product systems "from raw material extraction, through energy and material production and manufacturing, to use and end of life treatment and final disposal" [ISO06]. A product in LCA can be any good or service categorized as software, hardware, service, processed material, activity performed, delivery of an intangible product or creation of ambience, performing one or more defined functions [ISO06]. Thus, from the perspective of LCA information systems, application systems, software systems and hardware systems can be interpreted as products because they offer a defined capacity to store, communicate and compute information [HL11]. Creation, use and disposal of the product involves a system of processes and physical flows called *product system* [ISO06]. In the case of information systems, the product systems involves all processes for hardware manufacturing, use and disposal, but also processes for creating electricity, creating and maintaining software, extracting and transporting raw materials that are used to create hardware components and transporting workers to their workplace. When conducting a LCA for IT/IS solutions, the complete product system needs to be represented by an inventory model, i.e. a flow diagram, consisting of energy, material and informational flows [Gu02, ST14]. Within life cycle inventory modelling, inventory models constitute a third perspective for investigating potential design alternatives (c.f. Figure 2). To ensure operability of the environmental impact modelling approach and due to the high complexity of product systems, software tools for inventory modeling as well as inventory databases containing predefined inventory

models for standard products (e.g. road transportation, power generation or semiconductor manufacturing) need to be applied. Recently, two types of LCA have been distinguished – attributional and consequential [Fi09]. Attributional LCA describes the related environmentally relevant physical flows based on the currents status-quo of a product system. Typically average data for existing processes is applied [NHD11]. Therefore, attributional LCA takes a more reactionary perspective, making it less suitable for the use in IS design. Consequential LCA provides an anticipatory view on environmentally relevant flows that changes in response to possible design decisions [Fi09]. It typically involves data from marginal technologies and processes [NHD11]. Thus, a consequential LCA better supports the life cycle inventory modelling of IS design alternatives.

4 Application to Cloud Computing and Evaluation

The quality of the life cycle impact modelling approach is evaluated using the fictional case of a medium-sized newspaper publishing company. Applying fictional cases for the evaluation of design artifacts are well-accepted within the literature [BSS09, Fi12, MGR12, ST13]. Following [Te09], we will use the fictional case to evaluate our approach from three perspectives: 1) The technical quality is evaluated to prove that our approach can be implemented by means of a software system. 2) The conceptual quality is evaluated to prove that widely established reference models, e.g. EPC, Petri nets and UML-Diagrams, can be applied within the scope of our approach. 3)The economic quality is evaluated to prove the efficacy of our approach. The term “economic” in the context of designing Green IS is not associated with the financial value but rather with the utility of our approach, c.f. [ST14].

The Case: The publishing company runs a medium-sized datacenter to provide internal services for employees such as user profile hosting, email hosting and terminal/remote desktop services. Due to load changes, high overcapacities have been accumulated to provide constant accessibility. Consequently, the company faces high expenditures on IT infrastructures and overhead as well as high energy consumption. The company aims to replace its own IT (data center) infrastructures with cloud computing services. The outsourcing risks of a cloud migration have been estimated not critical. Also a leading position in the field of CO_2 aware business processes is aspired. As a first test, the layout system (application system) used within the editorial process to automatically arrange articles on a newspaper page was outsourced to a cloud computing provider. The client and data transfer energy consumptions remain the same for all alternatives so they can be neglected when comparing the different cloud computing alternatives. Three providers were selected for investigation:

1 *Smart Cloud* offers Infrastructure as a Service using datacenters located in the US. The layout system is accessed via remote desktop applications. This datacenters run servers with a maximum energy consumption of 200 W and 200 Gigaflops maximum calculating performance. Measurements at the energy supply of the rack servers reveal an average consumption of 0.3 Wh/Terra-

floating point operations (TFLO) at a utilization of 60 %. The power usage effectiveness (PUE) of the Smart Cloud datacenters equals 1.4.

2 *Online Edit* offers Software as a Service using datacenters located in Norway. The layout system is accessed via internet browsers. This datacenters run servers with a maximum energy consumption of 250 W and 300 Gigaflops maximum calculating performance. Measurements at the energy supply of the rack servers reveal an average consumption of 0.23 Wh/TFLO at a utilization of 65 %. The PUE of the Online Edit datacenters equals 1.4.

3 *Jour Edit* offers Software as a Service using datacenters located in Poland. The layout system is accessed via internet browsers. This datacenters run servers with a maximum energy consumption of 500 W and 800 Gigaflops maximum calculating performance. Measurements at the energy supply of the rack servers reveal an average consumption of 0.23 Wh/TFLO at a utilization of 50 %. The PUE of the Jour Edit datacenters equals 1.3.

The implicit factual model including the three alternatives as well as the editorial process is depicted in Figure 3.

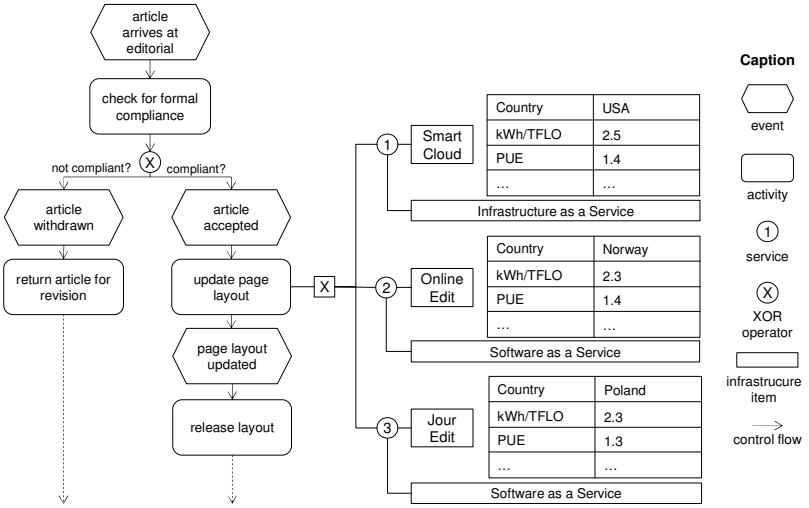


Figure 3 Factual model of different cloud computing service alternatives [BSS09]

To derive the environmental impact associated with the three design alternatives, inventory models are now developed. In this fictional case, we apply the commercial LCA software tool Umberto[1] for the implementation of petri-net based flow diagrams [Mu89]. Figure 4 depicts a simplified product system representing the cloud provider's use of electrical energy. On the right-hand side of the figure, the activity *update page layout* is depicted. The left-hand side shows the flows associated with the provided cloud function. The activity *cloud use* represents the use direct use of electricity (servers) as well as the use of facility functions (cooling, overhead). Direct energy consumption is

[1] www.umberto.de/en/

determined by the server's energy consumption (kWh/TFLO). Energy consumption caused by facility functions are represented by the activity *dc facility use* and are determined by the PUE of the datacenter. The activity *electricity, high voltage, production mix* represents consumptions and emissions associated with the production and distribution of the energy from raw material extraction to the datacenter. In our case we will use the life cycle database ecoinvent[2] to gather data for the physical flows associated with the production of electricity.

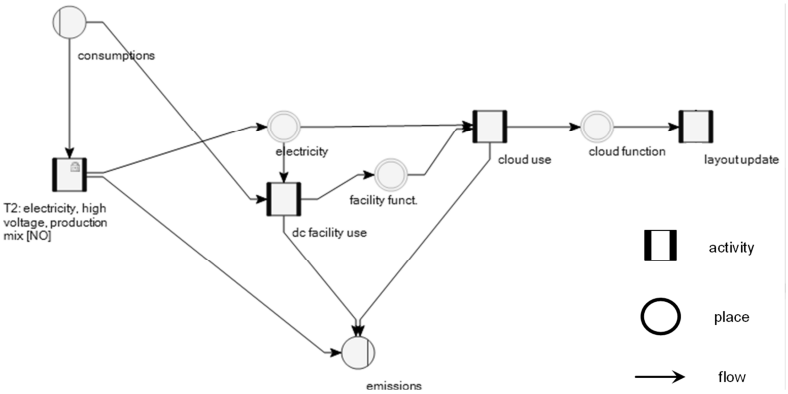


Figure 4 Simplified inventory model of cloud computing services

An impact model is calculated from the inventory model using the Umberto internal calculating engine. The impact model is depicted in Figure 5 in its short form, showing the environment impact of one hour service offered by the cloud provider *Online Edit*. CO_{2e}-emissions associated with the use of the service for one hour are depicted in line three (0.003 kg CO_{2e}). The services can now be compared based on the environmental impact model.

Product: layout post [A15 (layout update -> layout updated)] (1.00 unit)	
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **agricultural land occupation** w/o LT, ALOP w/o LT: **2.62E-03 m2a**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **climate change** w/o LT, GWP100 w/o LT: **0.03 kg CO2-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **fossil depletion** w/o LT, FDP w/o LT: **7.40E-03 kg oil-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **freshwater ecotoxicity** w/o LT, FETPinf w/o LT: **3.17E-04 kg 1,4-DCB-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **freshwater eutrophication** w/o LT, FEP w/o LT: **6.94E-06 kg P-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **human toxicity** w/o LT, HTPinf w/o LT: **0.01 kg 1,4-DCB-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **ionising radiation** w/o LT, IRP_HE w/o LT: **1.04E-03 kg U235-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **marine ecotoxicity** w/o LT, METPinf w/o LT: **3.18E-04 kg 1,4-DCB-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **marine eutrophication** w/o LT, MEP w/o LT: **6.00E-06 kg N-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **metal depletion** w/o LT, MDP w/o LT: **6.76E-03 kg Fe-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **natural land transformation** w/o LT, NLTP w/o LT: **3.45E-05 m2**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **ozone depletion** w/o LT, ODPinf w/o LT: **2.46E-09 kg CFC-11-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **particulate matter formation** w/o LT, PMFP w/o LT: **8.32E-05 kg PM10-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **photochemical oxidant formation** w/o LT, POFP w/o LT: **1.25E-04 kg NMVOC**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **terrestrial acidification** w/o LT, TAP100 w/o LT: **1.31E-04 kg SO2-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **terrestrial ecotoxicity** w/o LT, TETPinf w/o LT: **2.28E-05 kg 1,4-DCB-Eq**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **urban land occupation** w/o LT, ULOP w/o LT: **1.00 m2a**
	LCIA Method: ReCiPe Midpoint (H) w/o LT - **water depletion** w/o LT, WDP w/o LT: **1.08E-04 m3**

Figure 5 Shortened environmental-impact model of *Online Edit*

It can be stated that: 1) the software can be implemented by means of a software system, e.g. using LCA software tools, 2) widely established reference models, e.g. EPC or Petri nets, can be applied within the scope of our approach, 3) the approach fulfils its purpose to compare different design alternatives according to their environmental impact.

[2] www.ecoinvent.ch

5 Conclusions and Research Outlook

Developing an environmental sustainable society is one of the major challenges of our time. Researchers from various disciplines aim to develop sustainable products, services, behavior patterns and organizational systems. The IS discipline is now called on to contribute its share to this challenge by assessing the environmental impact of information systems and by designing new sustainable alternatives for current systems. One of the most promising concepts investigated by IS research is cloud computing. Despite being a young and still emerging field in IS research, recent publications have confirmed its potential to reduce the environmental impact of information systems. The life cycle impact modelling approach constitutes a first step towards an environmental oriented design of information systems. It offers IS designers the ability to forecast the environmental effects of their decisions. This ability will be indispensable to the design of more sustainable information systems. Due to our research approach, there are some specific limitations to this work. The evaluation builds on a single fictional case. This strategy was considered useful to demonstrate the fundamental features of our results. However, the question of whether other modelling languages and software tools besides the applied ones can be used within the approach remains unanswered. Due to the strict word limit, it was also not possible to give a deeper introduction into the development of inventory models. We suggest that interested readers consult related work on LCA [Fi09, Gu02]. The presented results may be thought of as a snapshot within the iterative design science research process. Further iterations and evaluations within real-world scenarios will performed to further refine the approach. The chosen metrics in section 4 should be recognized only as an example for demonstration purposes. Metrics like PUE or carbon usage power usage effectiveness should always be regarded in relation to performance-based metrics like TFLO. Acknowledging the limitations of this work, future research questions could include: Which modelling languages and software tools can be applied within the frame of the life cycle impact modelling approach? How can the approach be presented to business partners? Does the impact model contain enough information to support decision making? What are possible usage scenarios besides cloud computing? In the next step, business partners will be involved in expert interviews and case studies. Experiments on the behavior of decision makers will be conducted to investigate how environmental impact data is used for decision making.

Acknowledgement: *This work is part of the project IT-for-Green (Next Generation CEMIS for Environmental, Energy and Resource Management). The project is funded by the European regional development fund (grant no. W/A III 80119242). The authors are pleased to acknowledge the support by all involved project partners as well as the anonymous reviewers for their constructive comments.*

References

[Ak11] Ackermann, T.; Miede, A.; Buxmann, P.; Steinmetz, R.: Taxonomy of technological IT outsourcing risks: support for risk identification and quantification. In: Proc. ECIS (2011).

[BB06] vom Brocke, J.; Buddendick, C.: Reusable Conceptual Models-Requirements Based on the Design Science Research Paradigm. In: Proc. Conf. on Design Science Research (2006).

[Be09] Berl, A.; Gelenbe, E.; Di Girolamo, M.; Giuliani, G.; De Meer, H.; Dang, M. Q.; Pentikousis, K.: Energy-Efficient Cloud Computing. The Computer Journal 53 (2009) 7, S. 1045–1051.
[Br09] vom Brocke, J.; Simons, A.; Niehaven, B.; Riemer, K.; Plattfaut, R.; Cleven, A.: Reconstructing the giant: on the importance of rigour in documenting the literature search process. In: Proc. ECIS (2009).
[Br12] vom Brocke, J.; Loos, P.; Seidel, S.; Watson, R. T.: Green IS – Information Systems for Environmental Sustainability; BISE – Call for Papers Issue 5/2013. BISE 4 (2012) 1, S. 47–51.
[Br13] vom Brocke, J.; Watson, R. T.; Dwyer, C.; Elliot, S.; Melville, N.: Green information systems: directives for the is discipline. Communications of the AIS 33 (2013) 30, S. 509–520.
[BSS09] vom Brocke, J.; Sonnenberg, C.; Simons, A.: Value-oriented Information Systems Design: The Concept of Potentials Modeling and its Application to Service-oriented Architectures. BISE 1 (2009) 3, S. 223–233.
[Ca13] Castañé, G. G.; Núñez, A.; Llopis, P.; Carretero, J.: E-mc2: A formal framework for energy modelling in cloud computing. Simulation Modelling Practice and Theory 39 (2013), S. 56–75.
[Ca13] Callou, G.; Maciel, P.; Tutsch, D.; Ferreira, J.; Araújo, J.; Souza, R.: Estimating sustainability impact of high dependable data centers: a comparative study between Brazilian and US energy mixes. Computing 95 (2013) 12, S. 1137–1170.
[CS13] Chauhan, N. S.; Saxena, A.: A Green Software Development Life Cycle for Cloud Computing. IEEE Green IT (2013), S. 28–34.
[CWW13] Chang, V.; Walters, R. J.; Wills, G.: The development that leads to the Cloud Computing Business Framework. Int. Journal of Information Management 33 (2013) 3, S. 524–538.
[Fi09] Finnveden, G.; Hauschild, M. Z.; Ekvall, T.; Guinée, J.; Heijungs, R.; Hellweg, S.; Koehler, A.; Pennington, D.; Suh, S.: Recent developments in Life Cycle Assessment. Journal of Environmental Management 91 (2009) 1, S. 1–21.
[Fi12] Fill, H.-G.: On the Conceptualization of a Modeling Language for Semantic Model Annotations. In: Int. Conf. on Advanced IS Engineering 112 (2012), S. 134–148.
[FL04] Fettke, P.; Loos, P.: Referenzmodellierungsforschung. Wirtschaftsinformatik 46 (2004) 5, S. 331–340.
[Fr99] Frank, U.: Conceptual Modelling as the Core of the Information Systems Discipline - Perspectives and Epistemological Challenges. In: Proc. AMCIS (1999).
[Ga11] Garg, S. K.; Yeo, C. S.; Anandasivam, A.; Buyya, R.: Environment-conscious scheduling of HPC applications on distributed Cloud-oriented data centers. Journal of Parallel and Distributed Computing 71 (2011) 6, S. 732–749.
[Ga14] Gattulli, M.; Tornatore, M.; Fiandra, R.; Pattavina, A.: Low-Emissions Routing for Cloud Computing in IP-over-WDM Networks with Data Centers. IEEE on Selected Areas in Communications 32 (2014) 1, S. 28–38.
[GK13] Gottschalk, I.; Kirn, S.: Cloud Computing As a Tool for Enhancing Ecological Goals? BISE 5 (2013) 5, S. 299–313.
[Gu02] Guinée, J. B.; Gorrée, M.; Heijungs, R.; Huppes, G.; Kleijn, R.; Koning, A. de; Oers, L. van; Wegener Sleeswijk, A.; Suh, S.; Udo de Haes, H. A.; Bruijn, H. de; Duin, R. van; Huijbregts, M. A. J.: Handbook on Life Cycle Assessment - Operational Guide to the ISO Standards. Kluwer Academic Publishers, Dordrecht, 2002.
[GW08] Gartner; WWF: Assessment of Global Low- Carbon and Environmental Leadership in the ICT Sector. (2008).
[He04] Hevner, A. R.; March, S. T.; Park, J.; Ram, S.: Design science in information systems research. MIS Quarterly 28 (2004) 1, S. 75–105.
[HL11] Hilbert, M.; López, P.: The world's technological capacity to store, communicate, and compute information. Science 332 (2011) 6025, S. 60–65.
[ISO06] ISO: Environmental management - Life cycle assessment - Principles and framework (ISO 14040). International Organization for Standardization (2006).
[KMT12] Kessaci, Y.; Melab, N.; Talbi, E.-G.: A Pareto-based metaheuristic for scheduling HPC applications on a geographically distributed cloud federation. Cluster Computing 16 (2012) 3, S. 451–468.
[Le10] Leimeister, S.; Böhm, M.; Riedl, C.; Krcmar, H.: The business perspective of cloud computing: Actors, roles and value networks. In: Proc. ECIS (2010).

[Le12] Lemay, M.; Nguyen, K. K.; Arnaud, B. St.; Chariet, M.: Toward a Zero-Carbon Network. IEEE Green Computing (2012), S. 51–59.
[Ma13] Mandal, U.; Habib, M. F.; Zhang, S.; Mukherjee, B.; Tornatore, M.: Greening the cloud using renewable-energy-aware service migration. IEEE Network (2013), S. 36–43.
[Me10] Melville, N.: Information systems innovation for environmental sustainability. MISQ 34 (2010) 1, S. 1–21.
[MGR12] Mahmoud, T.; Gómez, J. M.; Rezgui, A.: Enhanced BI Systems with On-Demand Data Based on Semantic-Enabled Enterprise SOA. In: Proc. ECIS (2012).
[MNV12] Maheshwari, N.; Nanduri, R.; Varma, V.: Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework. Future Generation Computer Systems 28 (2012) 1, S. 119–127.
[MS95] March, S. T.; Smith, G. F.: Design and natural science research on information technology. Decision Support Systems 15 (1995) 4, S. 251–266.
[Mu89] Murata, T.: Petri nets: Properties, analysis and applications. In: Proc. of the IEEE 77 (1989) 4, S. 541–580.
[NHD11] Neupane, B.; Halog, A.; Dhungel, S.: Attributional life cycle assessment of woodchips for bioethanol production. Journal of Cleaner Production 19 (2011) 6–7, S. 733–741.
[NI11] NIST: The NIST Definition of Cloud Computing (Draft)–Recommendations of the National Institute of Standards and Technology. National Institute of Standards and Technology 145 (2011).
[Nu11] Nuseibeh, H.: Adoption of cloud computing in organizations. In: Proc. AMCIS (2011).
[QM13] Qiu, M.; Ma, C.: Guest Editorial: Green and Cloud Computing. Journal of Signal Processing Systems 72 (2013) 3, S. 149–150.
[SRB13] Seidel, S.; Recker, J.; Brocke, J. vom: Sensemaking and sustainable practicing: functional affordances of information systems in green transformations. MISQ (2013), [accepted].
[ST13] Stiel, F.; Teuteberg, F.: Towards a Conceptual Framwork for Life Cycle Assessment in Sustainable Information Systems Management. In: Proc. ECIS (2013).
[ST14] Stiel, F.; Teuteberg, F.: Measuring the environmental impact of IT/IS solutions – A life cycle impact modelling approach. Environmental Modelling & Software (2014), S. 1–24.
[Su04] Suh, S.; Lenzen, M.; Treloar, G. J.; Hondo, H.; Horvath, A.; Huppes, G.; Jolliet, O.; Klann, U.; Krewitt, W.; Moriguchi, Y.; Munksgaard, J.; Norris, G.: System Boundary Selection in Life-Cycle Inventories Using Hybrid Approaches. Environmental Science & Technology 38 (2004) 3, S. 657–664.
[TB09] Thomas, O.; vom Brocke, J.: A value-driven approach to the design of service-oriented information systems—making use of conceptual models. Information Systems and e-Business Management 8 (2009) 1, S. 67–97.
[Te09] Teuteberg, F.; Kluth, M.; Smolnik, S.; Ahlemann, F.: Semantic Benchmarking of Process Models-An Ontology-Based Approach. In: Proc. ICIS (2009).
[Te99] Teubner, R. A.: Organisations- und Informationssystemgestaltung. Theoretische Grundlagen und integrierte Methoden. Deutscher Universitätsverlag, 1999.
[Wa12] Watson, R. T.; Corbett, J.; Boudreau, M. C.; Webster, J.: An information strategy for environmental sustainability. Communications of the ACM 55 (2012) 7, S. 28.
[WBC10] Watson, R. T.; Bourdreau, M.-C.; Chen, A. J.: Information systems and environmentally sustainable development: energy informatics and new directions for the IS community. MISQ 34 (2010) 1, S. 23–38.
[WTM14] Williams, D. R.; Thomond, P.; Mackenzie, I.: The greenhouse gas abatement potential of enterprise cloud computing. Environmental Modelling & Software (2014), S. 1–7.
[YBS08] Youseff, L.; Butrico, M.; Da Silva, D.: Toward a Unified Ontology of Cloud Computing. In: Proc. Grid Computing Environments Workshop (2008).
[YT07] Yi, L.; Thomas, H. R.: A review of research on the environmental impact of e-business and ICT. Environment international 33 (2007) 6, S. 841–849.

Kriterien für Webportale zur Unterstützung nachhaltiger Regionalentwicklung am Fallbeispiel "vitaminBIR"

Eva Kern, Bianca Luther-Klee, Stefan Naumann, Andreas Filler

Hochschule Trier, Umwelt-Campus Birkenfeld
Institut für Softwaresysteme
Postfach 1380, 55761 Birkenfeld
{e.kern|b.luther|s.naumann|a.filler}@umwelt-campus.de

Abstract: Der ländliche Raum hat zumindest in den westlichen Industrieländern zunehmend mit dem demographischen Wandel zu kämpfen. Die Folgen des Bevölkerungsrückgangs schlagen sich in schlechterer Versorgung und mangelndem Arbeitsplatzangebot nieder. Umgekehrt finden Unternehmen schwieriger Fachpersonal, da die Region für hochqualifizierte Kräfte nicht attraktiv genug erscheint. Dies führt zu erhöhter Mobilität und geht generell zu Lasten einer nachhaltigen Entwicklung des ländlichen Raumes. In diesem Beitrag werden Kriterien für regionale Fachkräfteportale beschrieben, welche lokale Akteure in Sachen Arbeitsplatz-, Ferienjob- oder Praktikumssuche unterstützen und auch Unternehmen die Möglichkeit geben, die selbigen anzubieten. Die Stärkung des ländlichen Raumes entlastet auf diese Weise die Ballungsräume und führt auch zu reduziertem Pendelverkehr. An ein solch breit genutztes Portal, welches gleichzeitig die Stärken einer Region repräsentiert, sind hohe Anforderungen hinsichtlich Funktionalität, Usability und auch Interaktivität zu stellen. Dieser Beitrag beschreibt ein im Kontext der „Informations- und Kommunikationstechnologien für Nachhaltigkeit"-Debatte bestehendes Projekt und seinen Umsetzungsstand.

1 Einleitung und Motivation

Insbesondere in ländlichen Regionen herrscht durch steigende Qualitätsanforderungen und den demografischen Wandel der Wettbewerb der Regionen um gut ausgebildete Fachkräfte. Ursache sind rückläufige Zahlen bei den Erwerbsfähigen und der Abfluss von Wissen in andere Regionen [Hr10], [We10]. Einer der von diesen Bevölkerungsverlusten betroffenen Räume in Deutschland ist nach aktuellen Statistiken der Landkreis Birkenfeld in Rheinland-Pfalz als ländlich peripherer Kreis (Bevölkerungsrückgang 2009 zu 2012: -3,3 % [Pr13], prognostizierter Rückgang bis 2030: -15 %, bis 2060: -33 % [Bö12]). Daher wurde dieser Kreis als *LandZukunft* Modellregion (http://landzukunft-birkenfeld.de) ausgewählt [Ma11] und wird im Folgenden als Fallstudie herangezogen. Generell ist zu bedenken, dass ein Rückgang an Bevölkerung im Regelfall auch zu einem Rückgang von Versorgung vor Ort bzw. zu geringeren oder niedriger qualifizierten Arbeitsplatzangeboten führt. Entsprechend kommt es zu höheren Verkehrsaufkommen oder, bei Wegzug, zu größerer Belastung der Ballungsräume. Letztlich bedeutet Nachhaltigkeit den Erhalt von Lebensgrundlagen für die nachfolgenden Generationen (intergenerationelle Gerechtigkeit).

Ein Lösungsansatz um weiterem Bevölkerungsrückgang entgegenzuwirken, ist die Vernetzung und Kooperation zwischen Schule, Hochschule und Wirtschaft, um insbesondere der jungen Bevölkerungsgruppe die Potentiale ihrer Region aufzuzeigen [Ba03]. Allgemein kann die Kooperation als Instrument in der Regionalentwicklung einen Beitrag zur Erhöhung der Außenwirksamkeit dieser Region leisten sowie zur Zusammenführung verschiedener Akteure beitragen [Kn03]. Auf beides wird im vorliegenden Fallbeispiel abgezielt: Der Landkreis Birkenfeld bietet mit zwölf weiterführenden Schulen, zwei Standorten der Hochschule Trier und zahlreichen Unternehmen viel Potential für die Menschen vor Ort. Dieses ist jedoch bei den beteiligten Akteuren oft unbekannt. Es gilt, die Handelnden zusammenzuführen und zu vernetzen. Um diese Vernetzung voran zu treiben, wird derzeit (Stand 2014) das Internet-Portal www.vitamin-bir.de entwickelt. Dieses soll regionale Stellenangebote und Veranstaltungen auflisten, über Schulpatenschaften und Unternehmenskooperationen informieren und so eine Kommunikationsmöglichkeit für alle Zielgruppen, die eine nachhaltige Regionalentwicklung vorantreiben können, bieten.

Zunächst wurden hierzu anhand einer Analyse bestehender Websites Kriterien für regionale Portale festgelegt. In diesem Beitrag werden die Kriterien sowie die Analyseergebnisse vorgestellt. Zur Gewährleistung einer zielgruppenorientierten Entwicklung wurde außerdem eine quantitative Befragung im Landkreis Birkenfeld durchgeführt, deren Resultate im Folgenden im Ansatz vorgestellt werden.

2 Existierende Ansätze zur nachhaltigen Regionalentwicklung

Bereits zu Beginn der Durchdringung des WorldWideWeb in allen Bevölkerungsgruppen und deren Alltag wurden die Potentiale der Informations- und Kommunikationstechnologien (IKT) im „Wettbewerb der Regionen“ erkannt bzw. zumindest erahnt. So war und ist das Internet als Marketinginstrument auf regionaler Ebene nicht mehr wegzudenken. Bereits im Jahr 2000 wurde es insbesondere als Präsentationsmöglichkeit für Wirtschafts- und Tourismusstandorte aufgefasst, dem Potential zur Vernetzung und Zusammenführung in der Region wurde dagegen eher skeptisch entgegengesehen. [Se00] Spätestens seit Aufkommen der sogenannten „sozialen Netzwerke“ kann dieser These jedoch widersprochen werden.

Im Zusammenhang mit Regionalität ist auch immer das Ziel der Nachhaltigkeit zu sehen. Nach Lucas [Lu98] sind Konzeptionen nachhaltiger Regionalentwicklung durch die Berücksichtigung von Wechselwirkungen zwischen sozialen, ökologischen und ökonomischen Aspekten von der Problemdefinition bis hin zur Zielbildung gekennzeichnet. Darüber hinaus definiert Appel [Ap02] folgende Kennzeichen von Projekten der nachhaltigen Regionalentwicklung: (1) Bezug zur Nachhaltigkeit, (2) Einbindung der Akteure, (3) Nutzung von örtlichen Ressourcen, (4) Ziele sind nachhaltige Änderungen. Vor dem Hintergrund dieser Definition liegt die Schlussfolgerung nahe, dass regionale und kooperative Netzwerke zur nachhaltigen Regionalentwicklung beitragen können. Diese Unterstützung lokaler Kooperationen und Nachbarschaftseffekte kann mit Hilfe von IKT-Unterstützung geschehen [Pe10]. Es kann von einer „Nachhaltigkeit by IKT“ ge-

sprochen werden, bei der die Nutzer durch IKT in ihren Aktivitäten für eine nachhaltige Regionalentwicklung unterstützt werden.

Hinsichtlich der Gestaltung und Struktur regionaler Portale sind verschiedenste Möglichkeiten denkbar: Einerseits sind hier statische Seiten, die bspw. der bereits angesprochenen Präsentation von Wirtschaft und Tourismus dienen, zu nennen [Se00]. Andererseits sind vermehrt interaktive Angebote zu finden. Beispiele für regionale Wikis sind das Walgau Wiki zur Vernetzung von Organisation und Region mit dem Ziel ein „regionales Gedächtnis" zu schaffen [In10] und das MANO-Netzwerk zur Unterstützung von Kommunikations- und Kooperationsprozessen zwischen den beteiligen Akteuren an der Mikrosystemtechnik-Ausbildung in Nord-Ostdeutschland [Ej05]. Am Beispiel einer Community für Jugendliche erläutert Hahn [Ha09b] wie sich regionale Portale positiv auf die bestehenden Defizite ländlicher Gebiete auswirken können.

Ein ähnliches Prinzip wird mit dem Angebot regionaler Jobportale verfolgt. Jobsuchenden werden Stellenanzeigen aus ihrer näheren Umgebung angeboten. Sinnvoll erscheint dies, da laut einer Studie die überwiegende Anzahl der Stellensuchenden bereit ist, eine neue Arbeitsstelle innerhalb ihres eigenen Landkreises anzutreten (86,2 %), d. h. gezielt nach regionalen Angeboten gesucht wird [We11]. Internet-Stellenbörsen im Allgemeinen werden aktuell als aussichtsreichster Kanal einen Job zu finden angesehen (51,8 %) und auch von den meisten Suchenden (65,3 %) bei der Suche genutzt [We14].

Neben der Unterscheidung zwischen überregionalen (z. B. http://www.monster.de/) und regionalen Angeboten (z. B. http://www.fachkraefte-regional.de/) können die Portale hinsichtlich ihrer Zielgruppen und ihres Angebotes unterschieden werden. Einerseits existieren übergreifende Portale, andererseits sind diese auf Zielgruppen (z. B. nur für Studierende einer Hochschule: http://www.fh-bingen.de/info-services/proregio.html) oder bestimmte Jobarten fokussiert (z. B. Konzentration auf Ausbildungsstellen: http://www.ausbildungsstellen-pfalz.de/). Darüber hinaus finden sich regionale Portale, die über die Stellensuche hinausgehende Informationen zur Region und Unterstützung in weiteren Bereichen, z.B. Wohnungssuche, bieten. Hier ist beispielsweise das Portal der Ems-Achse zu nennen, welches in unsere Analyse einbezogen wurde.

3 Analyse regionaler Portale

Um eine Basis für die Festlegung der Anforderungen an ein neues Portal zu schaffen, bietet es sich an, existierende regionale Portale zu analysieren. Für eine solche Analyse wurden basierend auf einer explorativen Webanalyse das Fachkräfteportal „PFIFF - **P**ortal **f**ür **i**nteressierte **f**lexible **F**achkräfte" des Landes Sachsen-Anhalt (http://pfiff-sachsen-anhalt.de/), das Portal „Ems-Achse, Jobmotor Nordwest" des Vereins Wachstumsregion Ems-Achse e. V. (http://emsachse.de/) sowie das Portal des Vereins Münsterland e. V. (http://münsterland.com) mit der dazugehörigen Jobbörse „Jobwunderland" (http://jobwunderland.com), ausgewählt. Bei der Auswahl wurde nicht nach Region, sondern nach dem abgebildeten Angebot geschaut. Diese Portale haben die gleiche Intention - regionale Jobvermittlung - sind jedoch von der Struktur unterschiedlich aufgebaut. Während PFIFF als reines Jobportal angesehen werden kann, werden bei der

Ems-Achse ergänzende Services (Familien-Achse, Gründungs-Achse usw.) angeboten. Das Portal des Münsterlandes bezieht Aspekte wie Tourismus und Kultur ein und lagert die Jobsuche auf das „Jobwunderland“ aus. Diese Beispielportale decken die auftretenden Umsetzungsvarianten regionaler Portale ab. Sie wurden hinsichtlich ihrer Usability sowie der inhaltlichen und technischen Umsetzung betrachtet. Eine Auswahl der dazu herangezogenen Kriterien sowie resultierende Ergebnisse werden im Folgenden vorgestellt.

3.1 Usability

Unabhängig von der Zielgruppe sollte sich eine Website angenehm bedienen lassen [Ha09a], d. h. es sollte auf eine entsprechende Usability geachtet werden. Die Navigation und Übersichtlichkeit einer Seite können als Kernkonzepte der Usability angesehen werden und wurden daher im Rahmen der Analyse fokussiert.

Navigation: Für die Navigation einer Website definiert Krug [Kr06] die folgenden fünf wesentlichen Eigenschaften: Site-Kennung (Auf welcher Seite befinde ich mich?), Sektionen (Welche Inhalte enthält die Seite?), Utilities (Wo finde ich die wichtigsten Elemente, die nicht zur inhaltlichen Hierarchie zählen?), „Weg zurück“- und „Weg zur Suche“-Funktion.

Bei allen analysierten Portalen sind diese Navigationseigenschaften vorhanden, sodass eine gute Navigation gewährleistet ist.

Übersichtlichkeit: Dieses Kriterium bezieht sich auf Gestaltungsgrundsätze für das Layout einer Website. In diesem Zusammenhang spielen Symmetrie bzw. Asymmetrie und die Wahrnehmung einer Seite eine Rolle. [Ro13] Die Übersichtlichkeit für den Betrachter sollte trotz einer Fülle an Informationen nicht verloren gehen.

Auf den Portalen der Ems-Achse und des Münsterlandes stehen neben der Jobbörse zahlreiche weitere Informationen für den Nutzer zur Verfügung. Durch diese Masse an Angeboten wirken sie für neue Besucher zunächst erschlagend und die gebotenen Informationen können nicht gleich in ihrer Fülle verarbeitet werden.

3.2 Inhaltliche Umsetzung

Die Kriterien zur inhaltlichen Umsetzung zielen darauf ab, die Nutzererwartungen in die Entwicklung einzubeziehen und in der Nutzungsphase zu erfüllen. Sie berücksichtigen insbesondere Fragen, die den späteren Inhalt (z. B. Jobtypen) und die Wirkung der Seite (z. B. hinsichtlich Aktualität) bestimmen.

Aktualität: Gerade bei den Jobangeboten ist die Aktualität besonders wichtig, da veraltete Stellenanzeigen keinen Nutzen für die Stellensuchenden haben. Auf dieses Kriterium sollte daher besonderer Wert gelegt werden.

Bei allen analysierten Portalen ist die Aktualität gegeben. Dies betrifft sowohl die abgebildeten Stellenanzeigen als auch entsprechende Veranstaltungshinweise.

Art der Angebote: Die Frage nach der Art der Angebote bezieht sich auf die Jobtypen, für die Stellenanzeigen eingestellt werden können. Handelt es sich um ein Portal mit beschränktem Fokus hinsichtlich der Zielgruppe (z. B. reine Praktikumsbörse) legt dieser die Art der Angebote fest. Bei regionalen Portalen werden oft auch kleinere Tätigkeiten, z. B. Nachhilfeangebote, einbezogen, da sichergestellt ist, dass die örtliche Nähe zur Erbringung der angebotenen Leistung gegeben sein sollte.

Grundsätzlich kann bei allen betrachteten Portalen nach Stellenangeboten, Praktikums-, Ausbildungsplätzen, Traineetätigkeiten sowie nach Stellen für Abschlussarbeiten gesucht werden. Das Angebot fällt allerdings teilweise etwas gering aus.

Zielgruppe: Die Zielgruppe eines Portals bezieht alle Personen ein, die als potentielle Nutzer angesprochen werden sollen. Die Festlegung der Zielgruppe wirkt sich direkt auf die Gestaltung und die Inhalte der Seite aus. Bei der Analyse wurden die Zielgruppen „SchülerInnen", „Studierende" und „UnternehmerInnen" betrachtet, um einen direkten Bezug auf die für „vitaminBIR" identifizierten Gruppen herstellen zu können.

Bei den analysierten Portalen sind die Zielgruppen breit angelegt, d. h. Zielgruppe der Portale sind alle, die eine Stelle in der jeweiligen Region suchen. Auch Fachkräfte aus dem Ausland (PFIFF: Spanien im Rahmen eines Pilotprojektes mit der Autonomen Region Valencia; Ems-Achse: Niederlande, Spanien) sollen mit den Portalen adressiert werden. Während bei allgemeinen Portalen SchülerInnen, wenn überhaupt, eher indirekt mit Ausbildungsplätzen angesprochen werden, unterscheidet sich dies von den ausgewählten regionalen Portalen: Mit Projekten und Kampagnen wie bspw. dem Schüler-Workshop „Jugend denkt Münsterland" und „Junge Ems-Achse" sollen v. a. jungen Menschen die beruflichen Möglichkeiten in ihrer Region aufgezeigt werden.

Betreiber: Als Betreiber wird derjenige bezeichnet, der die Seite betreibt und für die Inhalte einsteht. Die Pflege der Inhalte muss nicht vom Betreiber direkt übernommen werden, sondern kann an die Nutzer übertragen werden, wenn bspw. UnternehmerInnen ihre Stellenanzeigen selbst einstellen und pflegen.

Alle untersuchten Portale werden von Vereinen betrieben. Im Fall von PFIFF treten der Verein „Bildungswerk der Wirtschaft Sachsen-Anhalt e.V." und die „isw Gesellschaft für wissenschaftliche Beratung und Dienstleistung mbH" als Projektträger auf. Die Ems-Achse wird von dem Verein „Wachstumsregion Ems-Achse e.V." getragen. Eine Ausnahme bildet die Gründungs-Achse, die von der „Emder Gründerinnen Zentrum GmbH" betrieben wird. Im Beispiel des Münsterlandes läuft die Vermarktung des gesamten Portals über den Verein „Münsterland e.V.", während das Angebot Jobwunderland von der „Aschendorff Media GmbH & Co. KG" betrieben wird.

Partner: Neben den Betreibern und Nutzern einer Seite treten bei regionalen Portalen häufig auch Partner auf. Damit sind diejenigen gemeint, die die Seite entweder direkt (bspw. finanziell) oder indirekt (bspw. durch weiterführende Angebote) unterstützen.

Partner der betrachteten Portale sind meist die Bundesagentur für Arbeit, Industrie- und Handelskammern sowie Handwerkskammern. Darüber hinaus treten bei PFIFF Städte, Landkreise, Vereine, Verbände, Unternehmen und wissenschaftliche Einrichtungen als

Partner auf. Das Portal des Münsterlandes wird von regionalen Wirtschaftsförderungseinrichtungen unterstützt.

Nutzungskosten: Unter dem Kriterium „Nutzungskosten“ wird analysiert, welche Kosten für die Nutzung als Stellensuchender bzw. als Stellenanbieter entstehen. Die Finanzierung des Angebotes an sich, d. h. die Umsetzung und der Betrieb des Portals werden hier nicht einbezogen.

Die Jobsuche ist bei allen Portalen für Stellensuchende kostenlos. Das Einstellen von Stellenangeboten ist bei zwei Portalen kostenlos (PFIFF, Ems-Achse), im Fall der Ems-Achse besteht die optionale Möglichkeit einer kostenpflichtigen Vereinsmitgliedschaft für Unternehmen und Organisationen. Diese bringt Vorteile wie eine Benachrichtigung bei passenden Bewerbern in der Job-Achse und Messepräsenz über Ems-Achse mobil mit sich. Beim Jobwunderland richtet sich der Preis für das Einstellen von Stellenangeboten nach der Anzeigengröße.

3.3 Technische Umsetzung

Die Aspekte der technischen Umsetzung beziehen sich insbesondere auf Entscheidungen, die für den Nutzer nicht direkt wahrnehmbar sind.

Social Media Anbindung: Bezüglich der Anbindung von Social Media Netzwerken werden insbesondere die Angebote mit hoher Relevanz für Nutzer (Facebook, Xing, Twitter [St14]) betrachtet. Eine Anbindung zu derartigen Portalen erscheint sowohl aus technischer Sicht (bspw. Google Ranking) als auch inhaltlicher Sicht (Erreichung der Zielgruppen über mehrere Kanäle) sinnvoll.

Alle geprüften Portale haben ihre aktivste Social Media Anbindung zu Facebook, d. h. es existieren entsprechende Fanseiten, die regelmäßig gepflegt werden und hohe Nutzerzahlen (je mind. 2.000 „Likes“) vorweisen. Zudem bestehen Aktivitäten bei XING, Google Plus und teilweise auf Twitter (PFIFF und Münsterland). PFIFF bietet außerdem einen RSS-Feed für Stellenangebote an.

Zugang und Funktionalität: Hinsichtlich des Zugangs zum Portal kann einerseits entschieden werden, ob dieser beschränkt ist oder nicht (bspw. erst nach Registrierung nutzbar), andererseits wie der Zugang selbst gestaltet ist (bspw. Login über bestehenden Account oder Anlegen eines neuen Accounts für das Portal). Zudem werden die angebotenen Funktionalitäten, z. B. Suchmöglichkeiten sowie Möglichkeit der Profil- und Anzeigenerstellung betrachtet.

Bei PFIFF und der Ems-Achse ist die Möglichkeit der Registrierung für Stellensuchende zum Anlegen eines Bewerber- bzw. Unternehmensprofils und von Stellenangeboten gegeben: Um ein Bewerberprofil anzulegen, muss bei der Ems-Achse nur die Benutzergruppe, der Name sowie E-Mailadresse und Passwort eingegeben werden. Weitere Angaben können später beliebig ergänzt und geändert werden. Im Fall von PFIFF müssen bei der Registrierung bereits einige Angaben zur Person sowie zum gewünschten Tätigkeitsfeld gemacht werden. Beim Jobwunderland wird keinerlei Möglichkeit zur Regi-

strierung geboten. Hier ist eine Anzeigenaufgabe für Stellenangebote nur per E-Mail möglich. Die Suchfunktion für Stellenanzeigen ist bei allen Portalen ohne Registrierung nutzbar.

Mobile Nutzung: Die Suche nach Stellen findet vermehrt über mobile Endgeräte statt [We14]. Das Portal sollte daher auch mobil ohne große Einschränkungen erreichbar und nutzbar sein.

Die Nutzung der Portale auf mobilen Endgeräten ist noch sehr eingeschränkt, lediglich die Job-Achse als Baustein der Ems-Achse bietet eine mobile Seite an. Eine App ist für keines der untersuchten Angebote bekannt.

Schnittstellen: Zur Unterstützung der Datenpflege und um insbesondere die Aktualität der Daten zu gewährleisten, erscheint es sinnvoll, geeignete Schnittstellen zu entsprechenden Datenlieferanten zu nutzen (bspw. Jobbörse der Bundesagentur für Arbeit). Das Kriterium betrachtet, ob derartige Möglichkeiten umgesetzt wurden.

Keines der Portale bietet eine Schnittstelle zur Bundesagentur für Arbeit. Schnittstellen gibt es mit anderen Jobportalen (PFIFF: Schnittstelle mit jobnetzwerk.de, Ems-Achse: EDR-Jobroboter) sowie mit dem Stellenmarkt einer Zeitung (Ems-Achse: Neue Osnabrücker Zeitung).

Software: Neben der Software, mit der das Portal selbst umgesetzt wurde, bieten sich Möglichkeiten der Unterstützung durch weitere Softwarepakete oder Dienste an, bspw. bei der Erfassung von Nutzerstatistiken. Hier soll verglichen werden, welche Software auf den verschiedenen Websites zum Einsatz kommt.

Für die Umsetzung der Portale wird insbesondere auf Open-Source gesetzt (Ems-Achse: TYPO 3, Münsterland: individuelle Software-Lösung mit Open-Source-Technik). Zur Analyse der Nutzerzahlen wird bei den Portalen die Software Piwik (PFIFF) oder Google Analytics (Münsterland) verwendet.

4 Ermittlung der Anforderungen: Zielgruppenorientierte Umfrage

Die nach Ejury [Ej05] notwendigen Schritte in der Entwicklung eines regionalen Internet-Portals sind: 1. Identifikation bisheriger Kommunikations- und Kooperationsmuster und 2. Herausfiltern spezifischer Nutzeranforderungen und Formulierungen von Anforderungen an eine neue Plattform. Basierend auf dieser Empfehlung wurde zur Analyse bisheriger Nutzergewohnheiten sowie zur Festlegung der Anforderungen an „vitaminBIR“ eine Umfrage mit Mitgliedern der Zielgruppen durchgeführt (Methode: Befragung durch Online- und Papierfragebogen, Grundgesamtheit: SchülerInnen, Studierende[1], UnternehmerInnen aus dem Landkreis Birkenfeld, Stichprobe: n = 304, davon 138 SchülerInnen mit μ = 16,64 Jahre, 150 Studierende mit μ = 26,23 Jahre, 16 UnternehmerInnen). Die Ergebnisse werden im Folgenden kurz vorgestellt.

[1] Die Gruppe „Studierende“ bezieht als eine der befragten Zielgruppen auch Interessierte aus der Region ein.

4.1 Analyse Ist-Zustand

Um bisherige Muster in der Vorgehensweise zur Stellensuche identifizieren zu können, wurden die Umfrage-Teilnehmer nach ihren Vorstellungen zum Vermittlungsprozess und den gewünschten Kontaktmöglichkeiten befragt. Insgesamt konnte festgestellt werden, dass alle Zielgruppen die aktive Kontaktaufnahme durch den Suchenden bevorzugen, der Anbietende bleibt während der Vermittlung – abgesehen von der Veröffentlichung der Stellenanzeige – zunächst passiv (vgl. Abbildung 1). Dabei werden elektronische Medien als Vermittlungskanal präferiert.

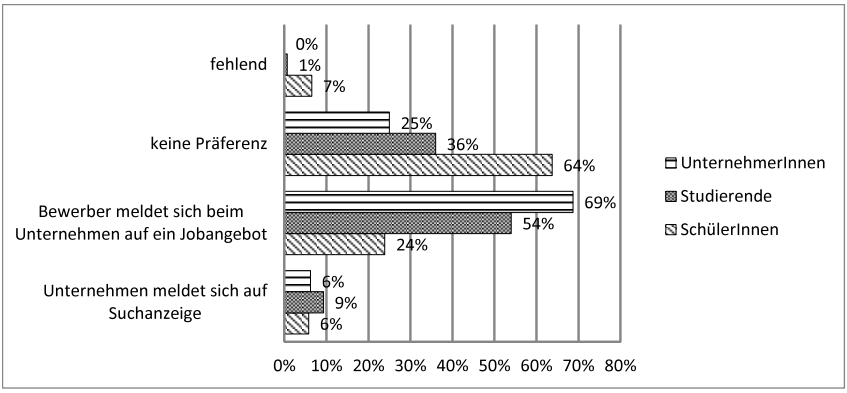


Abbildung 1: Ergebnisse der Abfrage nach der bevorzugten Kontaktaufnahme

Bisher sind im Rahmen der Job-Suche folgende Quellen für die Bewerber wichtig: Firmenwebsites (63,9 % bei SchülerInnen, 55,5 % bei Studierenden), Job-Plattform (54,5 % bei SchülerInnen, 59,6 % bei Studierenden) und bei SchülerInnen auch Freunde und Bekannte (54,9 %). Die UnternehmerInnen nutzen insbesondere ihre Firmenwebsite (78,5 %), die Zeitung (73,3 %) und die Job-Plattform der Bundesagentur für Arbeit (66,7 %) für die Bekanntmachung von Stellenanzeigen.

4.2 Anforderungsanalyse zur inhaltlichen Umsetzung

Um die technischen Kriterien für „vitaminBIR“ zu adressieren, wurden die Umfrage-Teilnehmenden hinsichtlich der Registrierung und der Suche befragt. Die Übereinstimmung mit den Vorschlägen konnte jeweils auf einer fünfstufigen Skala bewertet werden.

Von allen drei befragten Gruppen wird die Registrierung auf dem Portal über einen eigenen Account bevorzugt. Während die Anmeldung über Facebook bei den SchülerInnen überwiegend als neutral bewertet wird, wird die Nutzung der anderen vorgeschlagenen Accounts als unpraktisch angesehen (abgefragt wurden Facebook, Twitter, Google, Microsoft und XING). Auch Studierende halten eine Anmeldung mit bestehenden Accounts für unpraktisch, am ehesten ist für sie die Anmeldung mit dem bestehenden XING-Account denkbar (20,1 % für „(sehr) praktisch“). Ähnlich wird dies auch von UnternehmerInnen gesehen (57,2 % sehen einen neuen Account als (sehr) praktisch an). Aus Sicht der SchülerInnen wird v. a. ein Passwort (83,6 %) und ein Benutzername

(58,1 %) als notwendig bei einer Registrierung angesehen. Die Studierenden sind der Meinung, dass insbesondere E-Mailadresse (91,2 %) und Passwort (95,5 %) bei der Registrierung angeben werden müssen. Für UnternehmerInnen sind ein Passwort, Benutzername und Unternehmensname wichtig.

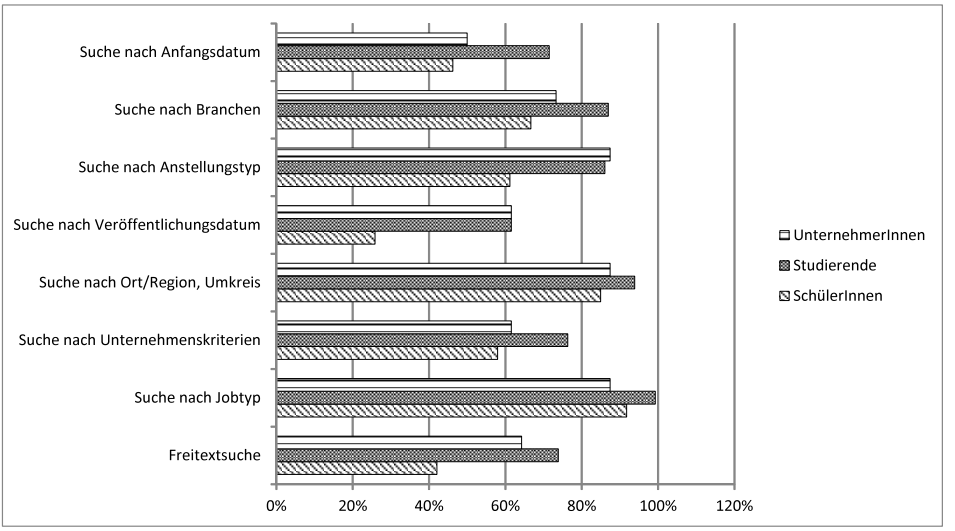


Abbildung 2: Ergebnisse der Abfrage nach der gewünschten Suchfunktion

Die befragten SchülerInnen sehen alle vorgeschlagen Arten der Suche als notwendig an, wobei insgesamt die Suche nach dem Anfangs-/Eintritts-/Veröffentlichungsdatum und eine Freitextsuche als am wenigsten wichtig bewertet wird (vgl. Abbildung 2). Für (unbedingt) notwendig sehen die befragten Studierenden eine Suche nach Jobtyp (99,3 %, davon 88,5 % für „unbedingt notwendig“) sowie nach Ort/Region und Umkreis (85 %) an.

4.3 Anforderungsanalyse zur technischen Umsetzung

Die SchülerInnen wollen v. a. nach Ausbildungsplätzen (84,3 %), Ferien- und Vollzeitjobs (je 75,4 %) sowie Praktika (72 %) suchen. Als weniger interessant werden Jobangebote für Trainees, Duales Studium, Werkstudenten, Doktorandenstellen und freie Mitarbeit gesehen. Die Studierenden möchten auf jeden Fall nach Vollzeitjob-Angeboten (93,1 %) und Jobs speziell für Hochschulabsolventen (86,9 %) suchen können. Nach allen weiteren Jobangeboten soll auch gesucht werden können, wobei eine Suche nach Ausbildungsplätzen und Dualem Studium, aber auch nach Doktorandenstellen und freier Mitarbeit weniger zutreffend bewertet wird. In erster Linie wollen die befragten Unternehmen Jobangebote wie Voll- (87,6 %) sowie, Teilzeitjobs (81,3 %), Ausbildung (84,6 %), und Praktika (71,4 %) ausschreiben können.

Die befragten Unternehmen sehen die Kontaktmöglichkeit (93,8 % „unbedingt notwendig“, 6,3 % „notwendig“) sowie Angaben zur Ausbildung / Studium (87,5 % „unbedingt

notwendig“) und Berufserfahrung, aber auch besondere Kenntnisse und einen Lebenslauf als „unbedingt notwendig“ an. Für sie notwendige Informationen wie Nachname, Vorname, Wohnort, Alter, Bewerbungsanschreiben, gewünschte Anstellung, Branche, möglicher Eintrittstermin, Sprachkenntnisse sowie Bewerberfoto, sollen von den Bewerbern bereitgestellt werden.

5 Schlussfolgerungen für das Portal „vitaminBIR“

Aufgrund der Analyse der anderen Portale sowie der Umfrage konnten gute Erkenntnisse für die Umsetzung des Portals „vitaminBIR“ gewonnen werden.

Usability: Um die Navigation, Gestaltung und Übersichtlichkeit des Portals „vitaminBIR“ zu überprüfen, wird das Portal in einem frühen Entwicklungsstadium von Personen, die der Zielgruppe entsprechen, getestet. Die aus den Usability-Tests gezogen Erkenntnisse werden anschließend umgesetzt und die Änderungen erneut getestet. Die Testphase findet insbesondere im Zeitraum von Mai bis Juli 2014 statt, da die Veröffentlichung des Portals für August 2014 geplant ist.

Inhaltliche Umsetzung: Um auch nach dem Ende der Projektlaufzeit den Nutzern ein attraktives und aktuelles Portal zur Verfügung stellen zu können, wird ein Pflegekonzept entwickelt. Wie aus der Umfrage hervorgeht, sind die Unternehmen bereit, ein breites Spektrum an Jobangeboten auf dem Portal anzubieten, was auch von den Nutzern sehr gewünscht ist. Während der Projektlaufzeit entstehen für den Nutzer keine Kosten. Ein entsprechendes Geschäftsmodell für die anschließende Realisierung wird im Projektrahmen erarbeitet. Die Umsetzung des Pflegkonzeptes, Geschäftsmodells und der Dauerbetrieb starten nach Projektende zum Januar 2015.

Technische Umsetzung: Für die Entwicklung des Portals „vitaminBIR“ wird TYPO3 verwendet. Die Anbindung von Social Media Portalen ist insbesondere für Facebook angedacht, da hier bereits Fanseiten des Umwelt-Campus und zugehöriger Organisationen bestehen. Der Zugang zum Portal wird über einen eigenen Account möglich sein, wie es von den meisten potenziellen Nutzern gewünscht wird. Für die Registrierung muss nur die E-Mailadresse eingegeben werden, vom System wird dann das erste Passwort erzeugt und zugesandt, welches vom Benutzer jederzeit geändert werden kann. Eine Schnittstelle mit der Bundesagentur für Arbeit wird integriert, um eine größere Vielfalt an Stellenangeboten auf dem Portal zu erhalten. Die Entwicklung einer App für das Portal wird zurzeit umgesetzt. Insgesamt findet die technische Umsetzung im August 2013 bis November 2014 (inkl. der Ausführung letzter Korrekturen) statt.

6 Fazit und Ausblick

Derzeit hat der ländliche Raum mit dem demographischen Wandel und einem zunehmenden Fachkräftemangel durch die Bevölkerungsabwanderung zu kämpfen. Um eine nachhaltige Entwicklung dieser Gebiete zu gewährleisten, gilt es Lösungsansätze zu finden, welche die Bedarfe adressieren und zu Verbesserungen vor Ort führen. Eine

Möglichkeit zur gemeinsamen Bewältigung dieser Herausforderungen wird in der Vernetzung der vielfältigen Akteurs- und Zielgruppen einer Region gesehen. Zur gebündelten Bereitstellung von Informationen und als Grundlage zur Schaffung von langfristigen Kooperationen wird auf den Einsatz von IKT gesetzt.

Im vorgestellten Fallbeispiel des Landkreis Birkenfeld wird zur Vernetzung regionaler Akteure aus Schule, Hochschule und Wirtschaft ein Internet-Portal entwickelt, das Angebot und Nachfrage bündeln soll und so eine schnelle und transparente Verteilung von Informationen ermöglicht. Auf dieser Basis können auf lange Sicht konkrete Maßnahmen zur nachhaltigen Entwicklung der Region realisiert werden. Das Beispiel erfüllt so die Kennzeichen von Projekten einer nachhaltigen Regionalentwicklung:

(1) Bezug zur Nachhaltigkeit: durch die Nutzung eines Online-Angebotes können sowohl aus ökologischer als auch aus ökonomischer Sicht Ressourcen gespart werden, da viele Prozesse in die virtuelle Welt ausgelagert werden können. Um gesellschaftlich soziale Aspekte für die beteiligten Akteure zu adressieren, darf die persönliche Kommunikation und Kooperation durch diese Auslagerung nicht komplett ersetzt werden. Insbesondere Regionen, die vom demografischen Wandel betroffen sind, brauchen zukunftsfähige Lösungen zur Adressierung der Folgen dieser Entwicklung. Hier kann IKT helfen, im Sinne einer nachhaltigen Entwicklung Kooperationen voranzutreiben, Akteure zu vernetzen sowie Potentiale zu erkennen und auszubauen.

(2) Einbindung der Akteure: SchülerInnen, Studierende, UnternehmerInnen sowie alle Interessierten aus der Region werden insbesondere in der Analysephase und der Testphase in die Entwicklung einbezogen.

(3) Nutzung von örtlichen Ressourcen: Die Entwicklung findet im Landkreis Birkenfeld statt und bildet die Potentiale von und für Menschen vor Ort ab.

(4) Ziele sind nachhaltige Änderungen: Der Aufbau des Portals stellt eine Innovation für die Region dar, mit der die Kooperation der verschiedenen Akteure langfristig geändert werden kann. Durch Transfer der Kooperationsmöglichkeiten ins Internet sind die Langfristigkeit und der orts- und zeitunabhängige Zugang gegeben.

Aus softwaretechnischer Sicht sind die herausgestellten Kriterien, insbesondere die Usability, zur Entwicklung eines regionalen Portals von besonderer Bedeutung. Des Weiteren konnten sowohl inhaltliche als auch technische Kriterien für derartige Portale herausgestellt werden. Die vorgestellte Fallstudie eines Portals zur Unterstützung regionaler Kooperationen im Birkenfelder Land macht die Funktion der IKT zur Unterstützung einer regionalen Entwicklung deutlich.

Danksagung

Das Projekt wird im Rahmen des Modellvorhabens *LandZukunft* des Bundesministeriums für Ernährung und Landwirtschaft gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren.

Literaturverzeichnis

[Ap02] Appel, E.: Konzeption und Durchführung von Projekten der nachhaltigen Regionalentwicklung. Projektevaluierung zur Ermittlung und Darstellung wesentlicher Faktoren bei der Umsetzung regionaler Entwicklungsinitiativen. Dissertation, Berlin, 2002.

[Ba03] Baier, A.; Bennholdt-Thomsen, V.: Der "Stoff", aus dem soziale Nähe ist. In (Kluge, T.; Schramm, E. Hrsg.): Aktivierung durch Nähe. Institut für sozial-ökologische Forschung (ISOE) GmbH, Frankfurt am Main, 2003; S. 12–21.

[Bö12] Böckmann, L. et al.: Statistische Analysen No. 25 2012, Rheinland-Pfalz 2060. Dritte regionalisierte Bevölkerungsvorausberechnung (Basisjahr 2010), Bad Ems, 2012.

[Ej05] Ejury, R.; Kalisch, C.: MANO goes Wiki? Chancen und Grenzen neuer Kommunikationswege in regionalen Netzwerken, 2005.

[Ha09a] Hauser, T.; Wenz, C.; Maurice, F.: Das Website-Handbuch. Programmierung und Design. Markt- & -Technik in Pearson Education Deutschland, München, 2009a.

[Ha09b] Hahn, M.: Regionale Nutzung virtueller Jugend-Communities im ländlichen Raum. Eine theoretische Studie am Beispiel der Jugend-Plattform "www.epos4.at". Magisterarbeit, Wien, 2009b.

[Hr10] Hradil, S.: Auswirkungen des Demographischen Wandels auf die Gesellschaft. In (Hradil, S.; Weingarten, J. Hrsg.): Alte Gesellschaft, Neue Gesellschaft? Zukunftsinitiative Rheinland-Pfalz (ZIRP), Mainz, 2010; S. 15–24.

[In10] Regionalentwicklung und Web 2.0. Theoretische Überlegungen und erste ergebnisse eines Fildversuchs der Regionalentwicklung im Walgau, 2010.

[Kn03] Knieling, J.: Kooperative Regionalplanung und Regional Governance: Praxisbeispiele, Theoriebezüge und Perspektiven. In Informationen zur Raumentwicklung, 2003, 8; S. 463–478.

[Kr06] Krug, S.: Don't make me think. Web usability - das intuitive Web ; [wie Webuser tatsächlich mit Websites umgehen ; verblüffend einfache Website-Strategien ; Layout und Navigation benutzerfreundlich gestalten]. Mitp-Verl., [Bonn], 2006.

[Lu98] Lucas, R.: Innovationsaufgabe Nachhaltige Regionalentwicklung. In Ökologisches Wirtschaften, 1998, 13.

[Ma11] Margarian, A.; Küpper, P.: Identifizierung peripherer Regionen mit strukturellen und wirtschaftlichen Problemen in DeutschlanIdentifizierung peripherer Regionen mit strukturellen und wirtschaftlichen Problemen in Deutschlan. Modellvorhaben LandZukunft - Auswahl der Regionen. http://www.land-zukunft.de/start-und-qualifizierungsphase/auswahl-der-regionen/.

[Pe10] Pellizzari, M.: Do Friends and Relatives Really Help in Getting a Good Job?: Industrial and labor relations review. JSTOR, 2010; S. 494–510.

[Pr13] Prognos Zukunftsatlas 2013 – Deutschlands Regionen im Zukunftswettbewerb, 2013.

[Ro13] Rohles, B.: Grundkurs Gutes Webdesign. Alles, was Sie über Gestaltung im Web wissen sollten. Galileo Press, Bonn, 2013.

[Se00] Seibel, M.: Regionalmarketing im Internet. In Standort, 2000, 24; S. 28–32.

[St14] Statista: Soziale Netzwerke - Besucherzahlen in Deutschland | 2013. http://de.statista.com/statistik/daten/studie/170467/umfrage/besucherzahlen-sozialer-netzwerke-in-deutschland/, 16.04.2014.

[We10] Weingarten, J.; Zahn, S.: Kultur, Wirtschaft und Demographie - Chancen der rheinland-pfälzischen Kultur- und Kreativwirtschaft im gesellschaftlichen Wandel. In (Rump, J.; Weingarten, J. Hrsg.): Neue Potentiale für Wirtschaft und Beschäftigung. Zukunftsinitiative Rheinland-Pfalz (ZIRP), 2010; S. 77–126.

[We11] Weitzel, T. et al.: Bewerbungspraxis 2011. Eine empirische Untersuchung mit über 10.000 Stellensuchenden und Karriereinteressierten, 2011.

[We14] Weitzel, T. et al.: Bewerbungspraxis 2014. Eine empirische Untersuchung mit über 10.000 Stellensuchenden und Karriereinteressierten im Internet, 2014.

Software Engineering und Integrative Nachhaltigkeit

Benno Schmidt, Andreas Wytzisk

Hochschule Bochum
Lennershofstr. 140
D-44801 Bochum
benno.schmidt@hs-bochum.de
andreas.wytzisk@hs-bochum.de

Abstract: Die Schaffung und Nutzung informationstechnischer Systeme geht einher mit positiven wie auch negativen Wirkungen auf Mensch, Gesellschaft und Umwelt. Ausgehend vom Konzept der Integrativen Nachhaltigkeit und einer Definition des Begriffs der Nachhaltigen Software werden die Anforderungen an nachhaltige Software-Engineering-Prozesse diskutiert. Auf Grundlage einer modellhaften Prozessbeschreibung werden mögliche Strategien identifiziert, welche eine zu den aufgestellten Zielen konforme Software-Entwicklung ermöglichen. Zuletzt werden mit Blick auf die dargestellte Problematik offene Forschungsfragen benannt.

1 Einleitung

Aspekte der Nachhaltigkeit werden heute intensiv mit Blick auf die Hardware-Komponenten informationstechnischer Systeme diskutiert (Stichwort "Green IT"). Für den Bereich der Software-Entwicklung finden in verschiedenen aktuellen Forschungsprojekten zwar verstärkt und mit Erfolg entsprechende Betrachtungen statt, jedoch ist der diesbezügliche Diskussionsprozess sicherlich noch nicht als abgeschlossen anzusehen [Ta11, Pe12, Pe13, Na13].

Zunächst lässt sich konstatieren: Die Entwicklung und Nutzung von Informations- und Kommunikationstechnik (ICT) ist grundsätzlich mit einem erheblichen Verbrauch von Ressourcen (Energie, Rohstoffe etc.) verbunden [Ga07]. Systemoptimierungen unter dem Effizienz-Aspekt sind zwar wünschenswert, lösen diese Problematik jedoch nicht abschließend: Insofern materielle Ressourcen irreversibel verloren gehen, ist ein energieeffizientes und Ressourcen-schonendes ICT-System zwar seiner ineffizienten Variante vorzuziehen, lässt sich aber sicherlich nicht allein auf Grund eines minimierten Bedarfs als nachhaltig bezeichnen. Insofern stellt sich die Frage, ob und ggf. wie sich Systeme schaffen lassen, deren umfassend betrachtete Auswirkung auf die ökologische und soziale Umwelt nicht-negativ oder sogar positiv ist. Explizit einzubeziehen in die Betrachtung sind dabei die Software-Komponenten und deren Herstellungsprozesse, die mit nennenswerten Bedarfen und Nebenwirkungen verbunden sein können.

Im Rahmen des im Aufbau befindlichen Forschungsschwerpunktes "Nachhaltige Entwicklung" an der Hochschule Bochum wird mit dem Nachhaltigkeitsziel auch eine gerechte Verteilung der Ressourcen verbunden. Eine zentrale Herausforderung stellt hierbei die globale Verbesserung der Lebensbedingungen für alle Menschen dar. Dies führt zu einer im Umfeld von Software-Entwicklungen zusätzlich zu beachtenden Anforderung.

Ausgehend vom Konzept der Integrativen Nachhaltigkeit (Kapitel 2) wird im Weiteren zunächst diskutiert, was sich unter nachhaltiger (im Gegensatz zu langlebiger oder effizienter) Software und unter nachhaltigen Software-Engineering-Prozessen verstehen lässt (Kapitel 3). Auf Grundlage einer modellhaften Prozessbeschreibung werden daraufhin Strategien abgeleitet, welche eine mit den Zielen vereinbare Software-Entwicklung erlauben, und praktische Beispiele genannt (Kapitel 4).

2 Nachhaltigkeitsbegriff

2.1 Integrative Nachhaltigkeit

Nachhaltigkeit lässt sich auf verschiedene Art und Weise definieren [Do14]. Häufig wird aus dem Leitbild der "Brundtland-Kommission" für Umwelt und Entwicklung der Vereinten Nationen zitiert [Wo87]: *"[Sustainable development] meets the needs of the present without compromising the ability of future generations to meet their own needs."* Im Rahmen der an der Hochschule Bochum durchgeführten Aktivitäten wird auf eine Definition Bezug genommen, die hierüber und auch über das "Drei-Säulen-Modell" mit seiner ökologischen, ökonomischen und sozialen Dimension hinaus geht. Es wird dabei an die vier Qualitäten der Pseudo-, schwachen, starken und Ultra-Nachhaltigkeit, die Martens und Schilder [MS12] aus ihrer politikwissenschaftlichen Perspektive darstellen, angeknüpft. Der dortige Begriff der Ultra-Nachhaltigkeit wurde hier leicht modifiziert und um psychologische und umweltwissenschaftliche Aspekte ergänzt, um schließlich zur Integrativen Nachhaltigkeit zu gelangen, welche sich durch folgende Merkmale auszeichnet [Sc13]:

- Verfolgt werden die Effizienz-, Konsistenz- und Suffizienz-Strategie (wie bei der starken Nachhaltigkeit). Die *Effizienz-Strategie* fokussiert auf die Minimierung des für eine Leistung zu erbringenden Aufwands, die *Konsistenz-Strategie* fordert eine Kreislaufführung der eingesetzten Stoffe und Energie. Die *Suffizienz-Strategie* zielt auf eine Änderung des Nutzungsverhaltens und der Nutzungsbedürfnisse ("nur das benutzen, was wir wirklich brauchen").
- Darüber hinaus werden eine Einheit von Mensch, Gesellschaft und Natur sowie der Gedanke der "fully-functioning society" zu Grunde gelegt. Somit wird davon ausgegangen, dass Menschen in der Lage sind, sich als Individuum zugehörig zu einer größeren Weltgemeinschaft zu empfinden.
- Naturerhalt und soziale Gerechtigkeit werden als die beiden wesentlichen Aspekte nachhaltiger Entwicklung angesehen.

2.2 Digitale Nachhaltigkeit

Im Umfeld der ICT häufig verwendet wird daneben der Begriff der Digitalen (oder Informationellen) Nachhaltigkeit, für den mehrere verschiedene Definitionen (bzw. verschiedene mit dem gleichen Begriff belegte Konzepte) existieren. Zumeist wird Wissen hier als immaterielle Ressource aufgefasst und der Nutzungsaspekt von Information inklusive offener Zugangsmöglichkeiten in den Vordergrund gestellt [Wi14, Bu08]. Insofern vernachlässigt diese Sicht (die stark die Open-Source-Philosophie stützt) mit Blick auf unsere Nachhaltigkeitsdefinition zunächst die ökologische Dimension, indem nicht explizit auf die Schonung natürlicher Ressourcen und nicht auf die oben geforderte Effizienz und Konsistenz abgezielt wird [Ma13]. Die weiteren Ausführungen werden zeigen, dass ungeachtet dessen das Software-gebundene Wissen (aus einer semiotischen Sichtweise heraus) positiv bilanzierbar sein kann. Weiterhin sind der offene Zugang zu Software-Artefakten und Wissen sowie die Möglichkeit der Partizipation an den Entwicklungsprozessen positiv hervorzuheben. Insofern vermag das Konzept der Digitalen Nachhaltigkeit verschiedene Teilaspekte Integrativ-nachhaltiger Software-Entwicklung anzusprechen.

3 Begriff der Nachhaltigen Software

In Anlehnung an den in der Literatur recht breit akzeptierten Vorschlag von [Di10] soll nachstehende Definition als erster Ausgangspunkt für die nachfolgenden Überlegungen dienen[1]: *(Integrativ-) Nachhaltige Software zeichnet sich dadurch aus, dass die direkten und indirekten negativen Auswirkungen auf Mensch, Gesellschaft und Umwelt, die sich aus der Entwicklung, dem Betrieb und der Verwendung der Software ergeben, minimal sind. Zudem sollen sich mit Blick auf eine (integrativ-) nachhaltige Entwicklung durch die Software langfristig positive Auswirkungen ergeben.*

Unter "langfristig positiven Wirkungen" lässt sich mit der Zielsetzung einer Integrativen Nachhaltigkeit eine Entwicklung verstehen, die den Einklang von Mensch, Gesellschaft und Natur im Sinne eines "bien vivir" unter Einhaltung natürlicher Stoffkreisläufe und frei von irreversiblen Änderungen anstrebt; siehe auch [Sc13]. Technologie im Allgemeinen und insbesondere auch ICT-Systeme werden dabei im Rahmen des genannten Forschungsschwerpunktes als treibender Faktor zur Entwicklung von Hilfsmitteln und Vorgehensmodellen zur Unterstützung und Förderung nachhaltiger Lebensweisen angesehen.[2]

4 Nachhaltige Software-Entwicklung

Oben stehende Definition für Nachhaltige Software beinhaltet bereits den Entwicklungsprozess. Trotzdem sei an dieser Stelle eine abstrakte Sicht auf unser Tun als

[1] Die gegenüber [Di10] vorgenommenen Änderungen sind in [Sc14] erläutert.
[2] An dieser Stelle wird marginal von der Ultra-Nachhaltigkeitssicht abgewichen; [MS12] spricht eine "radical critique of Western culture, including its anthropocentrism, belief in science and technology […]" an.

Software-Entwickler gewagt, welche als Grundlage für die Ableitung eines einfachen Prozessmodells dienen soll.

4.1 Software Engineering als Transformationsmodell

In einer naiven Anschauung lässt sich die Entwicklung und die sich anschließende Nutzung von Software prinzipiell als eine Transformation von (individuellem) menschlichem Wissen w und Daten d (worunter gemäß dem Von-Neumannschen Konzept auch Software-Artefakte, insbesondere Computer-Programme fallen) in neues Wissen w_t (u. a. Handlungswissen) und weitere Daten d_t, u. a. Software, betrachten. Stützen lässt sich diese Sichtweise z. B. durch die im Umfeld der semiotischen Forschung aufgebauten Modelle, die Software-Entwicklung als einen auf semiotischen Entitäten stattfindenden Prozess von Zeichen-Transformationen (d_t) sowie Bedeutungs- und Wissenszuordnungen verstehen [Na07].

Dieser sozio-technische Transformationsprozess, der mit einem Austausch von Mensch, Gesellschaft und natürlicher Umwelt einher geht, bedarf der Zufuhr geeigneter materieller (inkl. energetischer) Ressourcen *r,* die im Rahmen der Entwicklung und des Betriebs der Software ebenfalls transformiert werden [RB11, Na07, Ta11], z. B. für die Hardware benötigte Rohstoffe und elektrische Energie in CO_2, Abwärme, Computer-Müll usw. Teilweise werden diese materiellen Ressourcen irreversibel umgesetzt, d. h. es ergibt sich ein Verlustanteil ε.

Nun vermag Software innerhalb ihres Lebenszyklus als Mittel menschlichen Handelns zu fungieren [WF89] und bei den Nutzern Änderungen im Nachhaltigkeitsbewusstsein bewirken, was nachfolgend symbolisch als ψ-Komponente berücksichtigt ist:

$$w\,,\, d\,,\, \psi \;\rightarrow\; w_t\,,\, d_t\,,\, \psi_t$$

$$r \;\rightarrow\; r_t + \varepsilon$$

Die erste Transition [Sc14] erfasst immaterielle, prinzipiell erneuerbare "Wissensgüter". Die zweite Gleichung, in dessen Rahmen die Gegebenheiten der Erhaltungssätze aus der Physik zu beachten sind, bezieht sich auf die materiellen Ressourcen, für die seitens der Integrativen Nachhaltigkeit eine Einhaltung natürlicher Kreislauf-Eigenschaften gefordert wird (Konsistenz-Aspekt). Nun lässt sich formulieren: Der Wissens- und Bewusstseinsgewinn $\Delta w = w_t - w$, $\Delta\psi = \psi_t - \psi$ sollte mit Blick auf die Nachhaltigkeitsziele mindestens so groß sein wie der irreversibel transformierte Anteil der eingesetzten materiellen Ressourcen, also

$$\beta(\Delta w) + \beta(\Delta\psi) - \beta(\varepsilon) \;\geq\; 0,$$

wobei β eine geeignete Bewertungsfunktion bezeichnet. Das Potenzial einer pragmatischen Nutzung des "Wissensgutes Software", welches die Anwendungskompetenz des Nutzers voraussetzt, ist hierbei der Komponente w zugeschlagen (Daten und Software

sind ohne entsprechendes Wissen nicht nutzbar!).[3] Es lässt sich argumentieren, dass den generierten Software-Artefakten neu gewonnenes Wissen inhärent ist ("computation as semiosis" bei [Na07]), wobei zu gewährleisten ist, dass dieses positive Δw durch den Menschen erschließbar bleibt. Der Aspekt der Regenerativität des immateriellen Gutes "Software" wird insbesondere im Umfeld der bereits angesprochenen Digitalen Nachhaltigkeit hervorgehoben. Angesichts der Diskussion der gegenwärtigen Umweltprobleme sei an dieser Stelle allerdings nochmals die Notwendigkeit einer Dematerialisierung und Dekarbonisierung der Software-Entwicklungsprozesse konstatiert.

Der durch eine Software-Anwendung erreichbare Wissens- und Bewusstseinsgewinn, welcher der Erreichung der Nachhaltigkeitsziele dienlich sein soll, wird nur selten so offensichtlich sein, als dass kein Gegeneinanderabwägen der Gleichungsgrößen erforderlich wäre. (Die quantitative Erfassung der Größen mittels β ist dabei schwierig. Zudem sind letztlich "Äpfel und Birnen" abzuwägen.)

Ungeachtet der Validität des aufgebauten Modells eröffnet das Vorhandensein einer formalisierten Modell-Beschreibung die Möglichkeit einer sozialen Aneignung der dahinter liegenden Prozesse [Hü85] sowie die Ableitung grundsätzlicher Strategien zur nachhaltigen Software-Entwicklung, u. a.:

- *S1* Das gewonnene Wissen wird mit Blick auf die Ziele nachhaltiger Entwicklung maximiert (fokussiert das Δw).
- *S2* Die (irreversibel) eingesetzten Energiemengen und Ressourcen werden minimiert (minimales ε; Effizienzaspekt).
- *S2′* Es werden regenerative Energiequellen für Software-Entwicklung und -Betrieb genutzt (Energie-Anteil in ε soll gegen 0 streben).
- *S3* Es werden langlebige Software-Bausteine geschaffen (Reduzierung von ε durch längere Produkt-Lebenszyklen).
- *S4* Es wird ein breiter Zugang zu den geschaffenen Software-Artefakten und dem generierten Wissen gewährleistet (Vergrößern des Δw durch Multiplikatoren; Effizienz und soziale Gerechtigkeit).
- *S5* Es werden neue Entfaltungspotenziale des Integrativen Nachhaltigkeitswissens und von Nachhaltigkeitsbewusstsein ($\Delta\psi$) geschaffen.

Nicht explizit modelliert wurde hier der Aspekt der sozialen Gerechtigkeit. Diese *wesentliche Randbedingung* ist bei der Optimierung der oben angegebenen Zielgröße zu beachten, wobei "Gerechtigkeit" ein komplexer Begriff ist.[4]

[3] Hier wäre eine differenziertere Sicht auf Wissen möglich, z. B. durch eine Unterscheidung von explizitem und implizitem Wissen ("tacit knowledge").
[4] Wir tendieren zu dem Kriterium einer bedarfsbezogenen Verteilung.

4.2 Praktische Beispiele

4.2.1 "Software Engineering for the Planet"

Eine Software kann nachhaltig sein, insofern sie eine mit Blick auf die Nachhaltigkeitsziele sinnvolle Funktion hat, z. B. indem sie die praktische Durchführung einer Tätigkeit ermöglicht, welche einen direkten positiven Effekt auf die Umwelt hat (*"Software Engineering for the Planet"* [Ta11], Strategie *S1*). Genannt seien nur einige wenige Beispiele (die Liste ließe sich lang fortsetzen):

- eine E-Government-Anwendung (digitale Formulare per Internet statt Papier und Fahrt zum Amt),
- GIS-Werkzeuge zur Bearbeitung nachhaltigkeitsrelevanter georäumlicher Fragestellungen (z. B. die Optimierung des Standorts einer Windkraftanlage; für weitere Beispiele siehe [Sc13]),
- Umgebungen, welche die Partizipation an hinsichtlich der Nachhaltigkeitsziele relevanten gesellschaftlichen und politischen Prozessen oder an Forschungsaktivitäten im Sinne einer "Citizen Science" ermöglichen [JM13, Wy13].

Software vermag allerdings auch indirekt positiv zu wirken, beispielsweise indem sie dem Menschen die Handhabung komplexer Theoriegebilde ermöglicht oder der Entwicklung von Software-Komponenten dient, welche zukünftig direkte positive Effekte auf die Umwelt haben.

4.2.2 Organisatorische Aspekte

Gestaltungsspielraum besteht bei der Organisation der Arbeitsprozesse und des operationellen Betriebs. Nachstehende Beispiele tendieren in Richtung unseres Kriteriums, wobei die Strategie *S2* (bzw. *S2'*) verfolgt wird:

- In der Anwendung ist konzeptionell ein Mechanismus verankert, der zu einer *Kompensation negativer Umweltauswirkungen* führt. Beispiel: Durch die Nutzung der Anwendung werden automatisch Geldbeträge an nachhaltige Projekte gespendet.
- Unnötige Mobilität der Projektbeteiligten wird vermieden (Energieeffizienz); Telekommunikations- und Internet-Technologien werden weitestmöglich zur "Raumüberwindung" genutzt, insofern eine Face-to-face-Kommunikation verzichtbar ist.
- Im Rahmen der Software-Entwicklung benötigte Hardware-Ressourcen (z. B. Drucker, Server, Cloud-Computing-Systeme etc.) werden von mehreren Nutzern geteilt.
- Für Entwicklung und Betrieb der Anwendung ist ausschließlich eine *Nutzung regenerativer Energie* vorgesehen (unter Einhaltung geschlossener materieller Stoffkreisläufe). Beispiel: Nutzung erneuerbarer Energien im Softwarehaus und für Elektroautos der Mitarbeiter für die Fahrten zur Arbeitsstätte; intensive Nutzung von Green IT.

4.2.3 Unterstützung in den verschiedenen Entwicklungsphasen

Wesentlich schwieriger zu beantworten ist die Frage, wie sich die praktische Software-Entwicklung während der verschiedenen, ohnehin stark von Wirtschaftlichkeits- und Effizienz-Gedanken getriebenen Entwicklungsphasen (Anforderungsanalyse, Entwurf, Implementierung etc.), nachhaltig gestalten lässt. An dieser Stelle seien nur wenige mögliche Maßnahmen, deren Sinnhaftigkeit im Einzelfall durch die Software-Entwickler zu prüfen ist, beispielhaft aufgeführt. Die meisten (in Softwarehäusern häufig genannten) Beispiele der Auflistung tragen primär zu einer erhöhten *Effizienz* (Strategie *S2*) und größtenteils auch der *Langlebigkeit* (*S3*) von Software-Lösungen bei, scheinen also eher mit Blick auf eine schwache Nachhaltigkeit (Stichwort "Effizienz-Revolution" bei [MS12]) positiv bewertbar.

- Während der Anforderungsanalyse gibt es weitreichende Partizipationsmöglichkeiten für alle Interessensbeteiligten (Stakeholder) sowie für weitere potenzielle Akteure; siehe dazu auch [Mh13].
- Einzelne Systemteile (z. B. in Dienstketten einbindbare, generisch verwendbare Web-Dienste oder breit nutzbare Datenbank-Inhalte) lassen sich auch in Fremdsystemen oder späteren Neuentwicklungen direkt verwenden.
- Die wesentlichen konzeptuellen Elemente des Software-Entwurfs und der Implementierung werden kommuniziert (Wissenstransfer unter Software-Entwicklern statt "tacit developer knowledge"; erleichterte Erweiterbarkeit von Anwendungen um neue Funktionalität; Unterstützung weiterer kreativer Wertschöpfung etc.).
- Ein System bietet langlebige (offene) Interfaces an, wodurch nach Änderung der Implementierung (z. B. System-Update) die Applikationen, die diese APIs nutzen, weiterhin lauffähig sind.
- Web-Dienst-Schnittstellen werden nicht plötzlich (proprietär) geändert, so dass darauf aufbauende Anwendungen lauffähig bleiben.
- Eine von zahlreichen Anwendungen genutzte Software-Schnittstelle ist nicht unnötig komplex, so dass eine Inwertsetzung ohne zu großen Aufwand möglich ist (vgl. z. B. [Wy13] oder INSPIRE-Beispiel bei [JM13]).
- Eine millionenfach aufgerufene Funktion (z. B. innerhalb einer Massenanwendung mit sehr großer Nutzerzahl) arbeitet algorithmisch effizient, so dass CPU-Last und Energieverbrauch reduziert werden.

4.2.4 Software-Zugang

Strategie *S4* verfolgt die Gewährleistung eines möglichst breiten Zugangs zur Software und weitergehend zu den geschaffenen Software-technischen Artefakten und dem zugehörigen Entwickler- und Nutzungswissen. Beispiele:

- Der Benutzerschnittstellen-Entwurf berücksichtigt eine breite Software-Nutzung, u. a. durch technikferne Zielgruppen, für ältere Menschen, barrierefrei, in anderen Kulturkreisen etc.

- Die zu entwickelnde Anwendung stellt keine unangemessen hohe Hardware-Voraussetzungen und zwingt den (u. U. finanziell schlecht gestellten) Anwender nicht zum Kauf eines neuen Computers, obwohl der "alte" ansonsten noch gut laufen könnte.

Unter dem Aspekt der sozialen Gerechtigkeit ist u. a. die Gewährleistung gleichberechtigter Zugangsmöglichkeiten zu Informationsressourcen und Software von Allgemeininteresse positiv bewertbar. Hierunter fällt insbesondere folgende Maßnahme:

- Es wird ein freier Zugang zu den geschaffenen Software-Artefakten wie Quellcode ("Open Source"), Schnittstellen-Beschreibungen etc. gewährt.

Angemerkt sei an dieser Stelle, dass Open-Source-Software positive Beiträge zur Erreichung des Nachhaltigkeitsziels leisten kann, es aber nicht per se tut [Ma13]. Hingegen können auch kostenpflichtige Angebote oder proprietäre "Closed-Source"-Software (die häufig als Ideengeber und treibende Kraft für innovative Software-Anwendungen fungiert) durchaus sehr nachhaltig sein.

Kontrovers diskutiert wird die Rolle mobiler Web-Anwendungen, welche möglicherweise als wesentliches unterstützendes Element für den Übergang in nachhaltigere Lebensweisen fungieren können (siehe z. B. Diskussion zur Dekarbonisierung und Dematerialisierung durch Vernetzung auch in strukturschwachen Regionen der Welt, oder Beispiele aus dem ICT4D-Umfeld). Auf Grund des hohen infrastrukturellen Wertes mobiler Web-Technologie und der dadurch geschaffenen breiten Zugangsmöglichkeiten (und mit Blick auf die heutige Lebenswirklichkeit) sei die Auflistung daher um folgenden Punkt ergänzt:

- Die Anwendung ist lauffähig auf mobilen Endgeräten, wobei auch geringe Bandbreiten und/oder Offline-Betrieb unterstützt werden sollten.

5 Fragestellungen aus Sicht der Angewandten Forschung

Im Software-Engineering-Kontext bleiben zahlreiche Fragestellungen, die es im Rahmen einer (eher trans- als interdisziplinär ausrichtbaren) angewandten Forschung unter der Zielsetzung der (Integrativen) Nachhaltigkeit zu untersuchen gilt; vgl. auch [Pe12, JM13, Ta11, HL11, Na13].

- Wie groß ist der energetische und materielle Bedarf zur Durchführung von Software-Projekten inklusive sämtlicher "Nebenkosten" (Mobilität, Bürobedarf, sich anschließender produktiver Betrieb etc.) und wie lässt sich dieser Bedarf sinnvoll dem erzielbaren Software-Nutzen gegenüberstellen? Wie lässt sich der Grad der Nachhaltigkeit von Software und Software-Entwicklungsprozessen auf praktikable Art und Weise umfassend bewerten?
- Es fehlen praktisch nutzbare Entscheidungshilfen für die Auswahl alternativ einsetzbarer Software-Anwendungen und -Komponenten, Architekturen, Entwurfsmuster und Algorithmen mit Blick auf die Nachhaltigkeitsziele.

- Welche konkreten Inhalte könnte eine Zusammenstellung nachhaltiger Verhaltensweisen im Sinne selbstverpflichtender Gesten ("mit Δψ-Effekt") für Software-Entwickler haben?
- Wie ist ein Vorgehensmodell auszugestalten, das die projektbegleitende Bewertung der Nachhaltigkeit des Entwicklungsprozesses erlaubt und im Projektverlauf entscheidungsunterstützend wirken kann?
- In der Software-Entwicklung sind sehr häufig *Rebound-Effekte* beobachtbar (Zugewinne an Performanz, Speicher etc. werden sogleich ausgenutzt)! Welche Software-Entwurfsprinzipien lassen sich aus einem tieferen Verständnis dieses Zusammenhangs heraus gewinnen?
- Wie lassen sich Software-technische Artefakte durch Nicht-IT-Fachleute (in der Westlichen Kultur ebenso wie in Entwicklungsländern) einfacher nutzbringend verwenden?

Unabhängig vom Herstellprozess ergeben sich weitere Fragen, u. a.: Auf welche Weise können mobile Web-Anwendungen als wesentliche unterstützende Elemente für den Übergang in nachhaltigere Lebensweisen dienen (in allen Regionen dieser Welt)? Welche konkreten Anforderungen ergeben sich an nachhaltige Software mit Blick auf die Themen Privatsphäre, Datenschutz, Barrierefreiheit und soziale Gerechtigkeit? Welche Anforderungen ergeben sich (über den Software-Aspekt hinaus) für eine nachhaltige Fachdatenerfassung, -haltung und -bereitstellung?

Darüber hinaus führt der vorliegende Beitrag auch zu Fragestellungen, die primär aus sozialwissenschaftlicher Sicht zu untersuchen sind, z. B.: Wie lässt sich Nachhaltigkeitsbewusstein (ψ-Komponente) auf praktikable Art und Weise aussagekräftig messen?[5] Welche Abhängigkeiten bestehen zwischen Software-Herstellung, Software-Nutzung und global gerechter Ressourcen-Verteilung? Wie lässt sich die Kluft zwischen Software- und Wissensproduktion und gesellschaftlichem Handeln in Richtung nachhaltiger Lebensweisen überwinden?[6]

6 Fazit

Mit Blick auf das Ziel einer Integrativen Nachhaltigkeit ist der Begriff der Nachhaltigen Software nicht gleichbedeutend mit dem der effizienten, langlebigen oder frei zugänglichen Software. Vielmehr werden u. a. auch Aspekte der Konsistenz sowie die sich ergebenden sozialen Kontexte umfassend zu beachten sein.

Auf Grundlage eines vereinfachten, interdisziplinären Transformationsmodells, das Betrachtungselemente aus den Bereichen Semiotik, Informatik, Physik, Psychologie und dem sozialwissenschaftlichen Umfeld beinhaltet, wurden mögliche Strategien benannt, welche eine Software-Entwicklung konform zu den aufgestellten Nachhaltigkeitszielen

[5] Ein Instrument zur Durchführung derartiger Messungen wird derzeit im Labor für Nachhaltige Entwicklung (LaNE) der Hochschule Bochum entwickelt.
[6] Hier geht es um "action research" im Sinne von [SP12].

ermöglichen. Der vorgestellte Modellansatz kann hierbei lediglich als ergänzende Sichtweise auf die dargestellte vielschichtige Problematik dienen.

Bewertungen der Nachhaltigkeit einer Software oder eines Entwicklungsprozesses werden sich nicht auf einzelne Maßnahmen oder Software-Merkmale beziehen. Vielmehr muss die Gesamtbilanz positiv im Sinne der angestrebten nachhaltigen Entwicklung sein. In diesem Zusammenhang werden Methoden zur Bewertung der Nachhaltigkeit von Software und Software-Engineering-Prozessen ebenso wie praktisch nutzbare Entscheidungshilfen für den Software-Entwickler benötigt (mögliche Anhaltspunkte geben [In11, Ro12, EC13]). Zu berücksichtigen ist hierbei auch die zeitliche Komponente: So könnte beispielsweise eine heute als nachhaltig eingestufte Maßnahme fortschrittsbedingt morgen nicht mehr als nachhaltig gelten. Und auch Software-Artefakte unterliegen durchaus einem zeitlichen Werteverfall.

Eine effektive Verwendbarkeit der geschaffenen Software-Artefakte ist weiterhin anzustreben. Ziel weiterer Aktivitäten im Umfeld nachhaltiger Software-Entwicklung sollte u. a. eine Sensibilisierung der verschiedenen Akteure (Entwickler, Projektleiter, Kunden, Anwender, Entscheidungsträger etc.) für dieses global wichtige Thema sein.

Danksagung

Der vorliegende Text ist entstanden im Umfeld des Projekts "Erlebnisraum Nachhaltige Entwicklung (ENE)", das vom Ministerium für Innovation, Wissenschaft und Forschung (MIWF) des Landes Nordrhein-Westfalen im Rahmen des Förderprogramms "FH Struktur" gefördert wird. Projektziel ist u. a. die Entwicklung einer Software mit positivem $\Delta\psi$-Beitrag. Der Dank der Autoren gebührt Petra Schweizer-Ries, Oliver Stengel und vielen Nichtgenannten für die zahlreichen Anregungen und Anmerkungen.

Literaturverzeichnis

[Bu08] Busch, T.: Open Source und Nachhaltigkeit. Open Source Jahrbuch 2008, TU Berlin.

[Di10] Dick, M.; Naumann, S.; Kuhn, N.: A Model and Selected Instances of Green and Sustainable Software. Proceedings of the 9th IFIP TC 9 and 1st IFIP TC 11 International Conference, Brisbane, Australia, Sept. 2010; pp. 248-259.

[Do14] Dosch, K.; Sachsen, K. et al.: Lexikon der Nachhaltigkeit. Aachen: Aachener Stiftung Kathy Beys; http://www.nachhaltigkeit.info (Aufruf April 2014).

[EC13] European Commission (DG Comm. Networks, Content & Technology), eds.: ICT footprint : Pilot testing on methodologies for energy consumption and carbon footprint of the ICT-sector, final report, 2013; http://www.ecofys.com/en/publication/ict-footprint

[Ga07] Gartner, Inc.: Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO2 Emissions. Press Release, Stamford, CT, April 26, 2007; http://www.gartner.com/it/page.jsp?id=503867 (Aufruf April 2014).

[HL11] Hilty, L. M.; Lohmann, W.: The Five Most Neglected Issues in "Green IT". CEPIS Upgrade, XII(4), Oct. 2011; pp. 11–15.

[Hü85] Hülsmann, H.: Die technologische Formation - oder lasset uns Menschen machen. Berlin: Verlag Europäische Perspektiven, 1985.

[In11] International Organization for Standardization (ISO): ISO 14000 family (environmental management), 2004-11; http://www.iso.org/iso/iso14000

[JM13] Johann, T.; Maalej, W.: Position Paper: The Social Dimension of Sustainability in Requirements Engineering. Proceedings of the 2nd International Workshop on Requirements Engineering for Sustainable Systems, Rio, Brasil, July 15, 2013.

[Ma13] Martens, K.-U.: Digitale Nachhaltigkeit. In (Kegelmann, J.; Martens, K.-U., Hrsg.): Kommunale Nachhaltigkeit, Nomos-Verlag, 2013; S. 421-428.

[Mh13] Mahaux, M.: Could Participation Support Sustainability in Requirements Engineering? Proceedings of the 2nd International Workshop on Requirements Engineering for Sustainable Systems, Rio, Brasil, July 2013.

[MS12] Martens, J.; Schilder, K.: Sustainable Development. In (Krieger, J., ed.): The Oxford Companion to Comparative Politics, 2nd ed., Oxford, 2012; pp. 813-815.

[Na07] Nadin, M.: Semiotic Machine. Public Journal of Semiotics, I(1), Jan. 2007; pp. 57-75.

[Na13] Naumann, S.; Kern, E.; Dick, M.: Classifying Green Software Engineering - The GREENSOFT Model. 2. Workshop "Energy Aware Software-Engineering and Development" (EASED@BUIS), Oldenburg, Germany, April 2013.

[Pe12] Penzenstadler, B.; Bauer, V.; Calero, C.; Franch, X.: Sustainability in Software Engineering: A Systematic Literature Review. Evaluation & Assessment in Software Engineering (EASE 2012), Proceedings, Ciudad Real, Spain, May 2012; pp. 32-41.

[Pe13] Penzenstadler, B.: What does Sustainability mean in and for Software Engineering? Proceedings of the 1st International Conference on ICT for Sustainability (ICT4S), 2013.

[RB11] Radermacher, F. J.; Beyers, B.: Welt mit Zukunft : Die ökosoziale Perspektive. 7. Aufl., Hamburg: Murmann, 2011.

[Ro12] Rodriguez, A.: An Assessment Technique for Sustainability: Applying the IMAGINE Approach to Software Systems. Technical Report, Institut für Informatik der TU München. http://mediatum.ub.tum.de/doc/1126058/1126058.pdf

[Sc13] Schweizer-Ries, P.: Sustainability Science and its Contribution to IAPS: Seeking for Integrated Sustainability. IAPS Bulletin, 40, Autumn 2013; pp. 9-12.

[Sc13] Schmidt, B.: Geoinformatik-Ausbildung im Studiengang "Nachhaltige Entwicklung" der Hochschule Bochum. Hochschule Bochum, Geoinformatik-Bericht Nr. 14-01, 2014. Abrufbar unter http://www.geoinformatik-bochum.de/share/Bericht_14-01.pdf

[Sc14] Schmidt, B.: Strategien für eine integrativ-nachhaltige Software-Entwicklung. Hochschule Bochum, Geoinformatik-Bericht Nr. 14-02, 2014. Abrufbar unter http://www.geoinformatik-bochum.de/share/Bericht_14-02.pdf

[SP12] Schweizer-Ries, P.; Perkins, D.: Sustainability Science: Transdisciplinarity, Transepistemology, and Action Research. Umweltpsychologie, 16(1), 2012; pp. 6-10.

[Ta11] Taina, J.: Good, Bad, and Beautiful Software : In Search of Green Software Quality Factors. CEPIS Upgrade, XII(4), Oct. 2011; pp. 22–27.

[Wi14] Wikipedia, Artikel "Digitale Nachhaltigkeit", http://de.wikipedia.org/wiki/Digitale_Nachhaltigkeit (Aufruf April 2014).

[WF89] Winograd, T.; Flores, F.: Erkenntnis, Maschinen, Verstehen. Berlin: Rotbuch, 1989.

[Wo87] World Commission on Environment, eds.: Report of the World Commission on Environment and Development: Our Common Future. Oslo: UN, 1987. Abrufbar als PDF-Dokument unter http://www.un-documents.net/wced-ocf.htm (Aufruf April 2014).

[Wy13] Wytzisk, A.; Schmidt, B.; Nüst, D.: Echtzeitinformation und Kollaboration in Geodateninfrastrukturen. zfv (Zeitschrift für Geodäsie, Geoinformation und Landmanagement), Heft 5/2013; S. 333-338.

Big Data oder Grand Management Information Design?

Hans-Knud Arndt

Otto-von-Guericke Universität Magdeburg
Fakultät für Informatik
Arbeitsgruppe Wirtschaftsinformatik - Managementinformationssysteme
hans-knud.arndt@iti.cs.uni-magdeburg.de

Abstract: Big Data stellt Konzepte, Methoden, Technologien, IT-Architekturen sowie Werkzeuge zur Auswertung großer Volumina vielfältiger Informationen für Management-Entscheidungen bereit. Wichtiger Bestandteil heutiger Big Data-Architekturen stellen mobile Endgeräte wie z.B. Tablet-Computer dar. Am Beispiel des Tablet-Computers Microsoft Surface RT lässt sich aufzeigen, dass designbezogene Fehlentscheidungen gewaltige Auswirkungen auf die Nachhaltigkeit der Informations- und Kommunikationstechnik haben. Deshalb wird hier ein Denkansatz vorgeschlagen, der unter der Bezeichnung Grand Management Information Design durch eine Zusammenschau von Technik, Design und Absatzmarkt positiv auf die Nachhaltigkeit von Informations- und Kommunikationstechnik einwirken kann.

1 Big Data und Informations- und Kommunikationstechnik

In der heutigen Welt der Informations- und Kommunikationstechnik (IKT) spielt das Schlagwort Big Data eine große Rolle. Unter „Big Data“ wird im allgemeinen „ein Bündel neu entwickelter Methoden und Technologien, die die Erfassung, Speicherung und Analyse eines großen und beliebig erweiterbaren Volumens unterschiedlich strukturierter Daten ermöglicht“ [Hor13], verstanden. Für den deutschen Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM) ist Big Data eine Bezeichnung für „die wirtschaftlich sinnvolle Gewinnung und Nutzung entscheidungsrelevanter Erkenntnisse aus qualitativ vielfältigen und unterschiedlich strukturierten Informationen, die einem schnellen Wandel unterliegen und in bisher ungekanntem Umfang anfallen. Big Data stellt Konzepte, Methoden, Technologien, IT-Architekturen sowie Tools zur Verfügung, um die geradezu exponentiell steigenden Volumina vielfältiger Informationen in besser fundierte und zeitnahe Management-Entscheidungen umzusetzen und so die Innovations- und Wettbewerbsfähigkeit von Unternehmen zu verbessern“ [BTK14].

Umweltinformatik hat sich seit jeher als Big Data verstanden, denn gerade am Beginn der Umweltinformatik in den 1970er Jahren stand der Aufbau von Luftmeßnetzen, die sehr große Datenmengen produzierten, im Mittelpunkt des Interesses. Im Folgenden wurde im Rahmen der Umweltinformatik unter dem Stichwort „Umweltmonitoring“ die Datenerhebung und -speicherung von „Daten über Boden, Wasser, Luft und Lärm (..) [, die] an verschiedenen Orten regelmäßig erhoben“ werden [GRR94], verstanden. Gerade diese Datenerhebung im Umweltbereich führte zu „außerordentlich große[n] Mengen komplexer Rohdaten“ [GRR94]. Der Anspruch einer Speicherung und Verarbeitung äußerst großer Datenmengen hatte in der Umweltinformatik sogar zur Folge, dass in den 1990er Jahren noch intensiv an spezifischen Umweltdatenbanken geforscht wurde (siehe dazu z.B. [BaZ94]).

Aber auch Fragen der Umweltfreundlichkeit von Big Data und der passenden Infrastruktur als solches, z.B. Stromverbrauch, Energie sowie Fragen der Umweltfreundlichkeit von Informations- und Kommunikationssystemen als solches sind im Sinne einer Umweltinformatik bzw. vor dem Aspekt der Nachhaltigkeit zu diskutieren.

2 Informations- und Kommunikationstechnik und Nachhaltigkeit

Ein zentrales Betrachtungsobjekt der ökologischen und interdisziplinären Nachhaltigkeitsdebatte stellen die Ressourcen dar. Dies gilt genauso für die Informations- und Kommunikationstechnik (IKT). Denn die Annahmen der Ressourcentheorie besagen, dass sich die Art der Nutzung und Bereitstellung von Ressourcen unmittelbar auf den Erfolg von Informations- und Kommunikationssystemen (IKS) auswirken [WaH04].

IKS sind soziotechnische Systeme zur Unterstützung aller Aufgaben einer Organisation (also Unternehmen, Betriebe, Behörden, Krankenhäuser, Hochschulen etc.) mit Informationsbezug. Als zentrale Ressourcen eines IKS sind Information, Mensch und Technik anzusehen. Um die Verfügbarkeit dieser IKS-Ressourcen sicherzustellen, sind weitere Ressourcen wie Geld- und Betriebsmittel, Energie und Rohstoffe notwendig. Die Ressourcen des IKS haben unterschiedliche Eigenschaften. Auf der einen Seite sind diese Ressourcen entweder materiell, wie die Technik, oder immateriell, wie die Information. Auf der anderen Seite unterscheiden sich diese Ressourcen in ihrer Endlichkeit. Den meisten Untersuchungen zur Nachhaltigkeit liegt die Annahme knapper Ressourcen zugrunde [ABGK10].

Als knapp gilt einerseits, was weltweit in endlichem Maß vorhanden ist. Dies gilt im Hinblick auf IKT und Big Data vor allem auf die Energie. Werden konventionelle Energieformen verwendet, ist die eingesetzte Energie die endliche Ressource. Gerade die Diskussion um das Schlagwort „Green IT“ lenkt den Blick (möglicherweise einseitig) auf diese Ressource. Aber auch unter dem Begriff „Green IT“ wird weit mehr verstanden, denn Green IT „bezeichnet Bestrebungen, die Produktion und den Einsatz von Informations- und Kommunikationstechnologie (IKT) **umwelt- und ressourcenschonend** zu gestalten. IT-Einrichtungen – zu denen in diesem Zusammenhang auch Telekommunikationseinrichtungen gezählt werden – verursachen bei Unternehmen und Behörden inzwischen einen bedeutenden Anteil der **Energiekosten**. (...)Noch ungünstiger fällt die Umweltbilanz für IT-Produkte unter Einbeziehung des **Ressourcenbedarfs** (Verwendung einer Vielzahl seltener Rohstoffe) und der **Entsorgungsprobleme** (weltweit 5 Mio. Tonnen Elektroabfall mit giftigen Anteilen wie Blei, Quecksilber und Kadmium) aus“ [Hor10].

Knapp können Ressourcen aber auch sein, die – obwohl sie auf der Welt in großer Menge vorhanden oder regenerierbar sind – aber aus anderen Gründen, wie ökonomischer Leistungsfähigkeit, politischer Reglementierung oder lokaler Verteilung, nur in begrenztem Maß zur Verfügung stehen [UBA09].

Eine weitere Dimension der Nachhaltigkeit von IKT zeigt sich am Beispiel des Produkts „Surface RT“ (siehe Abbildung 1) des Unternehmens Microsoft. Bei dem Microsoft-Produkt Surface RT handelt es sich um einen 10,6 Zoll großen Tablet-Computer mit dem Betriebssystem Windows 8 (RT). Als Konkurrenzprodukt zu den iPad Tablet-Computern des Unternehmens Apple besteht das Gehäuse aus einer Magnesiumlegierung und ist mit einem Advanced RISC Machines (ARM)-Prozessor ausgerüstet. Das Surface RT verfügt über zwei eingebaute Kameras (Vorder- und Rückseite) sowie einem Wireless Local Area Network (WLAN)-Anschluss. Durch einen eingebauten, ausklappbaren Ständer auf der Rückseite des Surface RT kann der Tablet-Computer auf dem Tisch aufgestellt werden. Eine Besonderheit im Design stellt die mit einer Magnethalterung befestigte Abdeckung dar. Diese Abdeckung ist in zwei Varianten erhältlich. In der Variante des Surface Type Cover (5 mm stark) wird dem Nutzer eine Tastatur mit tatsächlich drückbaren Tasten und einem Trackpad bereitgestellt. Dagegen erhält der Nutzer in der Variante des Surface Touch Cover (3 mm stark) eine in unterschiedlichen Farben erhältliche Tastatur mit fühlbaren, aber nicht drückbaren (sensorgesteuerten) Tasten. Durch einen speziellen Lagesensor wird sichergestellt, dass die Möglichkeit von Eingaben deaktiviert wird, wenn die Tastatur das Surface RT im Sinne einer Abdeckung verschlossen hat oder nach hinten geklappt ist. Wird ein farbiges Surface Touch Cover verwendet, ändert sich der Hintergrund der Benutzeroberfläche von Windows 8 entsprechend der Farbe des Surface Touch Cover. Damit soll die Einheit aus Hardware und Software unterstrichen werden [DaI12].

Abbildung 1: Microsoft Surface RT (1. Generation)
(Bild: Microsoft, Quelle: [DaI12])

Als mobile Endgeräte stellen Tablet-Computer schon heute einen wichtigen Baustein in einer Big Data-Architektur dar. Auch aus diesem Grund werden Tablet-Computern im professionellen Umfeld weiter steigende Verbreitungszahlen vorausgesagt. Deshalb ist es auf den ersten Blick um so unverständlicher, dass ein für den Tablet-Markt als innovativ angekündigtes Produkt wie das Microsoft Surface RT bisher keine signifikante Marktdurchdringung erreichen konnte und für das Unternehmen zu grundlegenden Problemen, zunächst einmal in finanzieller Hinsicht, führte. So formuliert die Süddeutsche Zeitung: „Eigentlich wollte sich Microsoft mit seinem neuen Tablet-Computer in die Post-PC-Ära katapultieren. Stattdessen entpuppt sich das Surface als teurer Ladenhüter, der dem Unternehmen die Bilanz vermasselt“ [Süd14]. Übereinstimmend berichtet die Presse, dass Microsoft im Jahr 2013 900 Millionen US Dollar auf seinen Tablet-Computer Surface RT abschreiben musste, da sich dieser offenbar noch schlechter verkauft, als von Marktbeobachtern angenommen wurde [Süd14].

Aufgrund der 10-Q-Pflichtmitteilung vom 31. März 2014 an die US-Börsenaufsicht Securities and Exchange Commission (SEC) lässt sich feststellen, dass das Unternehmen Microsoft auch in den vergangenen neun Monaten mit seinen Surface Tablet-Computern einen Verlust von weiteren 300 Millionen US Dollar erwirtschaftet hat. Dabei ist den Experten allerdings unklar, ob der Grund für diesen Verlust in Werbeausgaben oder in hohen Produktionskosten zu suchen ist [Saw14], scheinbar aber nicht in einer weiteren Abschreibung nicht absetzbarer Geräte.

Vor dem Hintergrund der Nachhaltigkeit ist aber nicht nur der finanzielle Verlust von Interesse, sondern auch die physische Dimension des Beispiels Microsoft Surface RT. Experten haben im Internet Schätzungen aufgestellt, um wie viele Surface Geräte es sich bei einem Verlust von 900 Millionen US Dollar handeln muss: „Wie viele Surface Tablets Microsoft noch auf Halde liegen hat, kann man nur schätzen - dazu müsste man wissen, welchen Wert man bei der Abschreibung angesetzt hat. Selbst wenn man aber den aktuellen Verkaufspreis von 350 Dollar ansetzt, kommt man auf 2,5 Millionen Stück - das dürften weit mehr sein, als man bis jetzt insgesamt verkaufen konnte“ [Geu13]. Es handelt sich also entsprechend dieser Schätzung um den Ressourceneinsatz für 2,5 Millionen Stück. Bereits an dem Beispiel Microsoft Surface RT wird deutlich, dass Fehlentscheidungen von Unternehmen (in diesem Fall sogar von einem im IKT-Bereich äußerst etablierten Unternehmen) gewaltige Auswirkungen auf die Nachhaltigkeit der IKT haben.

Die Suche nach Gründen für diese beispielhafte Nicht-Erreichung einer Nachhaltigkeit von IKT führt aber zu dem Schluss, dass es

a) **nicht** am Gerätetypus gelegen haben kann, denn Tablet-Computer erfreuen sich einer gewaltigen und immer noch steigenden Beliebtheit bei den Konsumenten;

b) **nicht** an der Geräteausstattung gelegen haben, denn – auch wenn das Microsoft Surface RT nicht zur Spitzengruppe gehören mag – es sich doch um einen hinreichend gut ausgestatteten und hochwertigen Tablet-Computer handelt.

Deshalb ist das Design des Tablet-Computers (Hard- und Software) oder allgemein gesprochen, der Zusammenhang von Nachhaltigkeit und Design näher zu betrachten.

3 Nachhaltigkeit und Design

Das Microsoft Surface RT wird mit der für diesen Tablet-Computer speziellen Version von Windows 8 ausgeliefert. Microsoft hat bei der Entwicklung dieses Betriebssystems bewusst auf einen Design-getriebenen Ansatz gesetzt: „Wir reduzieren uns auf das Wesentliche, vermeiden den Ballast an Dekor und Funktionsüberfluss. Wir wollen reine, klare Produkte - und zwar unabhängig davon, welches Gerät ein Anwender gerade nutzt. Er soll immer dieselbe, geradlinige Erfahrung von Tiefe und Klarheit machen können“ (so zitiert in [Graf11]).

Das Betriebssystem Windows 8 mit dem Modern-User Interface („Metro“) wird deshalb auch von einigen Kritikern für ihr Design gelobt, aber dennoch kommt dieses Design bei den Kunden nicht so gut an: „Microsoft's Windows 8 Is The Perfect OS That Nobody Wants“ [Wil13].

Vielmehr wird in Veröffentlichungen das Betriebssystem „Microsoft Windows 8“ aus Anwendersicht als „Zwitterlösung“ oder sogar als „Frankensteins Monster“ bezeichnet, da Windows 8 (nicht ein Betriebssystem ist, sondern zwei (für den klassischen Personal Computer und für mobile Endgeräte mit berührungsempfindlichen Bildschirm, die auf unnatürliche Weise miteinander verschmolzen wurden („Windows isn't a single OS. It's two of them, mobile and desktop, fused together unnaturally like a Frankenstein's monster“ [Thu14]).

Diese Beobachtungen führen zu dem Schluss, dass Design zwar wichtig und entscheidend für den Erfolg und vor allen für die Nachhaltigkeit von IKT ist, aber für sich allein Design aber noch nicht ein hinreichend Kriterium darstellt. Deshalb soll auf das von der Firma Braun AG geprägte, vor dem Hintergrund des klassischen Industriedesign geprägten Grand Design im Sinne eines Grand Management Information Design zurückgegriffen werden, um so die Nachhaltigkeit von IKT sicherstellen zu können.

4 Design als Grand Management Information Design

Gerade die Zusammenarbeit der Hochschule für Gestaltung Ulm mit der Firma Max Braun OHG in Kronberg führte zu gestalterisch als technisch höchste bedeutsamen Elektro- und Elektronikprodukten: „Inzwischen arbeitete Erwin Braun im Zwiegespräch mit Hans Gugelot an einem Konzept, das er später ‚Grand Design’ nannte. Damit sollte insbesondere auch die technische Qualität noch einmal deutlich angehoben und Märkte, die man teilweise erst selbst schaffen mußte, erschlossen werden.
Ziel war die Entwicklung von hochklassigen Produkten, die ihre Qualität auch durch Eleganz signifikant ausdrücken“ [Pol05]

Als Grand Management Information Design soll hier ein Denkansatz für IKT verstanden werden, welcher den Benutzer bei seiner IKT-bezogenen Tätigkeit bestmöglich unterstützt und die Ausgestaltung an seinem nachhaltigen Bedarf und seinen Bedürfnissen ausrichtet. In diesem Sinne muss Design immer im Zusammenhang mit einer entsprechenden Technik und entsprechenden Märkten gesehen werden und darf niemals (wie es grundsätzlich bei einem Künstler der Fall ist) alleiniger Zweck sein.

Um den Denkansatz des Grand Management Information Design zu verdeutlichen, soll hier stellvertretend das Produkt „iPad“ (siehe Abbildung 2), welches das Unternehmen Apple im Jahr 2010 erstmals öffentlich vorstellte, betrachtet werden. Durch das iPad wurde eine neue Klasse mobiler Endgeräte, die Tablet-Computer, in die IKT-Welt eingeführt. Die Einführung von Tablet-Computern wurde anfänglich sogar vereinzelt skeptisch aufgenommen, so z.B. der Medienjournalist Jeff Jarvis: „I'm taking my iPad back to the store (…). It's really because I don't see a need for it. It's solving a problem, that I don't think exists (…)“ (so zitiert in [Höl10]). Aus heutiger Sicht ist aber vielmehr davon auszugehen, dass Tablet-Computer die klassischen Personal Computer (PC) immer stärker verdrängen werden. So stellt z.B. die Frankfurter Allgemeine Zeitung (FAZ) in ihrer Online-Ausgabe vom 03.03.2014 fest: „Tabletcomputer stellen den Markt auf den Kopf: 2013 wurden mehr als 195 Millionen dieser Geräte verkauft - so viele wie noch nie. Darunter leiden die Hersteller klassischer PCs“ [Fin14].

Abbildung 2: Apple iPad (1. Generation)
(Bild: Apple, Quelle: [CNT14])

Im Sinne des Grand Management Information Design zeigt das IKT-Produkt iPad in besonders deutlicher Weise, dass

(a) **Technik**: Durch das Produkt iPad wurde die technische Qualität mobiler Endgeräte mit berührungsempfindlichen Bildschirmen noch einmal deutlich angehoben.

(b) **Design**: Diese technische Qualität des iPads wird auch durch Eleganz sowohl von Hardware als auch durch die Benutzeroberfläche (Software) signifikant ausgedrückt.

(c) **Nachfrage**: Durch das iPad ist es - trotz anfänglicher Skepsis - gelungen, einen neuen Markt von außerordentlicher Bedeutung zu schaffen.

in einer Gesamtschau betrachtet werden müssen, um zu nachhaltig erfolgreichen IKT-Produkten bzw. IKT-Lösungen zu kommen.

Hilfestellung zur Operationalisierung des Grand Management Information Design gibt der langjährige und international sehr geschätzte Chefdesigner der Firma Braun AG, Dieter Rams, mit seinen Ideen des Entwerfens und Gestaltens von Elektro- und Elektronikprodukten, die von ihm in den „Zehn Thesen für Gutes Design“ zusammengefasst worden sind [Vit13]. Zur Ursachenanalyse für den bisherigen Misserfolg von Microsofts Tablet-Computer Surface RT können vor allem folgende drei der zehn Thesen für Gutes Design herangezogen werden [Vit13]:

- **Gutes Design macht ein Produkt brauchbar**
  Man kauft ein Produkt, um es zu benutzen. Es soll bestimmte Funktionen erfüllen – Primärfunktionen ebenso wie ergänzende psychologische und ästhetische Funktionen. Gutes Design optimiert die Brauchbarkeit und lässt alles unberücksichtigt, was nicht diesem Ziel dient oder ihm gar entgegensteht. (...)

- **Gutes Design ist ästhetisch**
  Die ästhetische Qualität eines Produktes ist integraler Aspekt seiner Brauchbarkeit. Denn Geräte, die man täglich benutzt, prägen das persönliche Umfeld und beeinflussen das Wohlbefinden. Schön sein kann aber nur, was gut gemacht ist. (...)

- **Gutes Design macht ein Produkt verständlich**
  Es verdeutlicht auf einleuchtende Weise die Struktur des Produkts. Mehr noch: Es kann das Produkt zum Sprechen bringen. Im besten Fall erklärt es sich dann selbst. (...)“

Daraus folgt, dass eine bestimmte Technologie (hier Tablet-Computer) mit einem bestimmten Benutzerkonzept (hier Microsoft Windows 8) im Sinne eines Guten Designs brauchbar, ästhetisch und verständlich gestaltet werden sollte. Für die Brauchbarkeit des Benutzerkonzepts Windows 8 von Microsoft ist aber festzustellen, dass dieses Benutzerkonzept aus Anwendersicht vielfach und heftig kritisiert wurde (siehe oben Stichwort „Zwitterlösung“ bzw. „Frankensteins Monster“ [Thu14]). Aber auch das Kritiker-/Designerlob für das gelungene Design von Windows 8 [Wil13] ist vor dem Hintergrund der These „Gutes Design ist ästhetisch“ als nicht zutreffend anzusehen, da es in dieser These heißt, dass: „Schön sein kann aber nur, was gut gemacht ist“ [Vit13]. Denn hier zeigen wiederum die Anwendererfahrungen, dass – im Vergleich zu anderen Tablet-Computern mit dem Benutzerkonzept von Apple (iOS) oder von Google (Android) – deutlich besser (im Sinne von besser gemacht) ausfallen. Und schließlich die Forderung eines Guten Designs nach Verständlichkeit ist ebenso kritisch im Hinblick auf das Benutzerkonzept Windows 8 zu sehen. Auch hier bevorzugt der Nutzer (derzeit) die Benutzerkonzepte iOS bzw. Android.

Deshalb wird durch den Denkansatz des Grand Management Information Design bewusst die Verbindung zwischen Technik, Design und Nachfrage hergestellt. In diesem Kontext kommt dem Design von IKT, also von Hardware (Technik) und insbesondere Softwaresystemen (Benutzerkonzepten) bzw. von ganzen IT-Ökosystemen eine immer größere Bedeutung zu [Arn13]. Und dies gilt sowohl für Endgeräte beim Nutzer (z.B. Tablet-Computer) als auch für große Big Data-Architekturen. Der Denkansatz des Grand Management Information Design kann unter konsequenter Berücksichtigung der „Zehn Thesen des Guten Designs“ von Dieter Rams zu einer deutlicheren Verbesserung der Nachhaltigkeit von IKT führen.

5 Ausblick

Gutes Design – immer in Verbindung mit der Betrachtung der entsprechenden Technik und unter Berücksichtigung der Nachfragesituation – bildet den Schlüssel zu mehr Umweltfreundlichkeit und Nachhaltigkeit von Informations- und Kommunikationssystemen bzw. Informations- und Kommunikationstechnik. Die Idee des Denkansatzes des Grand Management Information Design ist es deshalb, genau die drei Aspekte von Technik, Design und Nachfrage in Bezug auf IKS bzw. IKS zusammenzuführen. Deshalb stellt sich nicht die Frage nach einem Big Data oder einem Grand Management Information Design (wie der Titel des vorliegenden Beitrags lautet), sondern es gilt im Sinne der Umweltfreundlichkeit bzw. Nachhaltigkeit von Big Data die Ideen eines Grand Management Information Design anzuwenden.

Literaturverzeichnis

[ABGK10] Arndt, H.-K.; Blendinger, A.; Grabski, B.; Krüger, L.: Modell zur Nachhaltigkeit von Informations- und Kommunikationssystemen. In: Information Management und Consulting (IM), 25. Jg., Heft 3, 2010, S. 25-32.

[Arn13] Arndt, H.-K.: Umweltinformatik und Design - Eine relevante Fragestellung? In: Horbach, M. (Hrsg.): INFORMATIK 2013: Informatik angepasst an Mensch, Organisation und Umwelt (16.–20. September 2013, Koblenz, Germany), GI-Edition-Lecture Notes in Informatics (LNI), P-220, Gesellschaft für Informatik e.V., Bonn, 2013, S. 931-939.

[BaZ94] Baumewerd-Ahlmann, A.; Zink, L.: Umweltdatenbanken – Sachstand und Zukunftsperspektiven. In: Page, B.; Hilty L. M. (Hrsg.): Umweltinformatik: Informatikmethoden für Umweltschutz und Umweltforschung (Handbuch der Informatik; Bd. 13.3), Oldenbourg Verlag, München/Wien, 1994 S. 79-101.

[BTK14] Bitkom - Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (Hrsg.): Arbeitskreis Big Data,

[CNT14] CNET (Ed.): Remember this guy? The original iPad is still a mighty capable tablet. http://i.i.cbsi.com/cnwk.1d/i/tim2/2013/06/07/iPad_1st-gen_all_angles.jpg [2014-06-22].

[DoI12] Donath, A.; Ihlenfeld, J.: Microsoft Surface: Mit eigenen Tablets gegen Apple, http://www.golem.de/news/microsoft-surface-mit-eigenen-tablets-gegen-apple-1206-92606.html [2014-06-22].

[Fin14] Finsterbusch, S: iPad & Co.: Tabletrechner verdrängen PCs. In Frankfurter Allgemeine Zeitung (FAZ) Wirtschaft, Online-Ausgabe vom 03. März 2014, http://www.faz.net/aktuell/wirtschaft/netzwirtschaft/ipad-co-tabletrechner-verdraengen-pcs-12829779.html [2014-06-22].

[Geu13] Geuß, M.: 900 Millionen Dollar Verlust mit dem Surface RT - wie schafft man sowas? http://www.drwindows.de/content/2233-900-millionen-dollar-verlust-surface-rt-schafft.html [2014-04-22].

[Graf11] IT-Konzern im Umbruch: Plötzlich zählt das Bauhaus-Prinzip. In Süddeutsche Zeitung Digital vom 29.12.2011, http://www.sueddeutsche.de/digital/it-konzern-im-umbruch-microsoft-entdeckt-die-kraft-des-designs-1.1246227-2 [2013-06-30].

[GRR94] Günther, O.; Radermacher, F. J.; Riekert, W.-F..: Umweltmonitoring: Modelle, Methoden und Systeme. In: Page, B.; Hilty L. M. (Hrsg.): Umweltinformatik: Informatikmethoden für Umweltschutz und Umweltforschung (Handbuch der Informatik; Bd. 13.3), Oldenbourg Verlag, München/Wien, 1994 S. 51-77.

[Höl10] Höly, D.: Steve Jobs und das iPad: Kein Messias und nur eine Maschine (3), http://juiced.de/6005/steve-jobs-und-das-ipad-kein-messias-und-nur-einemaschine-3.htm - ftn1[2014-06-22]

[Hor10] Horvath, S.: Green IT. Aktueller Begriff, Deutscher Bundestag, Wissenschaftliche Dienste, Fachbereich 10 Kultur, Medien und Sport, 2010, http://www.bundestag.de/dokumente/analysen/2010/Green_IT.pdf [2014-04-22].

[Hor13] Horvath, S.: Big Data. Aktueller Begriff, Deutscher Bundestag, Wissenschaftliche Dienste, Fachbereich 10 Kultur, Medien und Sport, 2013, http://www.bundestag.de/dokumente/analysen/2013/Big_Data.pdf [2014-04-22].

[Pol05] Polster, B.: Braun: 50 Jahre Produktinnovation, Dumont Buchverlag, Köln, 2. Aufl., 2005, S. 45.

[Saw14] Sawall, A.: Tablets: Microsoft macht durch Surface 300 Millionen Dollar Verlust, http://www.golem.de/news/tablets-microsoft-macht-durch-surface-300-millionen-dollar-verlust-1404-106188.html[2014-06-22].

[Süd14] O.V.: Schwacher Tablet-Absatz- Surface zieht Microsoft nach unten. In Süddeutsche Zeitung Digital vom 19. Juli 2013, http://www.sueddeutsche.de/digital/schrumpfender-pc-markt-surface-vermasselt-microsoft-die-quartalszahlen-1.1725390 [2014-04-22].

[UBA09] UBA Umweltbundesamt: Green IT: Zukünftige Herausforderungen und Chancen. Hintergrundpapier für die BMU/UBA/BITKOM-Jahreskonferenz 2009, Dessau-Roßlau. 2009, http://www.umweltbundesamt.de/sites/default/files/medien/publikation/long/3726.pdf [2014-04-22].

[Vit13] Vitsoe Ltd (Hrsg.): Dieter Rams: Zehn Thesen für gutes Design, https://www.vitsoe.com/de/ueber-vitsoe/gutes-design [2014-04-22].

[WaH04] Wade M.; Hulland J: The resource-based view and information systems research. In: MIS Quarterly Vol. 28, No. 1, 2004, pp. 107-142.

[Wil13] Wilson, M.: Microsoft's Windows 8 Is The Perfect OS That Nobody Wants, http://www.fastcodesign.com/3023244/microsofts-windows-8-is-the-perfect-os-that-nobody-wants [2014-04-22].

The 4th International Workshop on Advanced Design Concepts and Practice

(ADCP2014)

The 4th International Workshop on Advanced Design Concepts and Practice (ADCP2014)[1]

Abstract: The ADCP (Advanced Design Concepts and Practice) series workshop is an important part of an International collaboration Innovation team project sponsored by Tsinghua University and another two grants including National Natural Science Foundation of China (51175284) and the Ministry of Science and Technology of the People´s Republic of China (2011ZX02403). The team involves professors from TU Delft, University of Stuttgart, Politecnico di Milano, Dublin City University, University College London, Tel-Aviv University, University of South Carolina and other universities in Europe and USA.

The ADCP workshops have been held in Beijing on July 6-8th of 2011, in Karlsruhe, Germany, on May 8th of 2012. ADCP Summer workshops have been held at Tsinghua University on August 16th 2013 and on June 24th 2014. ADCP web Conferences also were organized at TU Delft on August 17th, in Beijing on November 8th and on March 28th 2014. ADCP 2014 workshop will be held in Stuttgart, Germany, 26th September 2014, as one part of INFORMATIK 2014 - Big Data 44, Annual Meeting of the German Informatics Society in Cooperation with the research group of the German Informatics Society on "Graphics in design and Engineering - GI GRIB".

The ADCP project is now three years on its way. It is time to formulate our ambitions and the project outcome in more detail.

One of the aims is to deliver a design system that allows the development, re-use and extension of parametric models of the processing chambers. These models can be subsequently analyzed in a multi-physics code to help the designer to find theoretically feasible and maybe optimal values for the characteristic parameters of the design and processing parameters.

The system could consist of a dedicated user interface for the designer, a KBE code generation part, a multi-physics toolbox, an optimization toolbox, a postprocessor translating the analysis results into an evaluation of the systems performance in values and units understandable by the designer, a 3D printing module, an experimental set-up (usable for different shower head topologies) and any other module necessary to support the designers.

The system development is based on a number of unique design theories closely interacting with recent developments in mathematics and based on innovative developments in computer science. The theories can be tested with the design system.

Our contributions could/should be technically oriented but we could also help the project by motivating students to get on board and go beyond analysis using existing tools. Sketching a bigger picture of our ideas and showing that the application to the chamber design is just an example could help to get more students joining the team. The workshop is another opportunity to do this.

[1] http://adcp2012.com/ADCP2014.html

Chair: Professor Dieter Roller (University of Stuttgart)

Co-Chairs: Professor Michel van Tooren (Delft University of Technology), Professor Ji Linhong (Tsinghua University)

Academic Committee: Professor Imre Horvath (Delft University of Technology, the Netherlands), Professor Dieter Roller (University of Stuttgart, Germany), Professor Ashok K. Goel (Georgia Institute of Technology, USA), Professor Chen Li (Tsinghua University, China), Professor Chen Liping (Huazhong University of Science and Technology, China), Professor Giorgio Colombo (Politecnico di Milano, Italy), Professor Han Xu (Hunan University, China), Dr. Yuemin Hou (Tsinghua University, China), Professor Ji Linhong (Tsinghua University, China), Professor John Gero (University of North Carolina, USA), Professor Michel van Tooren (University of South Carolina, USA), Professor Ming Xinguo (Shanghai Jiaotong University, China), Professor Panos Y. Papalambros (University of Michigan, USA), Dr. Sean Hanna (University College London, UK), Dr. Stephen Daniels (Dublin City University, Ireland), Dr. Stephan Rudolph (University of Stuttgart, Germany), Professor Tetsuo Tomiyama (Cranfield University, UK), Dr. Zoltan Rusak (Delft University of Technology, the Netherlands), Prof. Offer Shai (Tel Aviv University, Israel)

Workshop Managers: MSc. Leila Zehtaban (University of Stuttgart), Dr. Yuemin Hou (Tsinghua University)

Optimum function for minimum weight evaluation of structural joints

L. T. Beilstein[1], S. Rudolph

Institute for Statics and Dynamics of Aerospace Structures
University of Stuttgart
Pfaffenwaldring 31
70569 Stuttgart

laura.beilstein@gmx.de
rudolph@isd.uni-stuttgart.de

Abstract: In this paper a general methodology for minimum weight selection of structural joints is reviewed. The evaluation and selection technique is based on analytical weight formulae for structural joints including different joining technologies for which through the use of Buckingham's Pi-Theorem a complete set of dimensionless similarity parameters is derived. This formal approach offers the possibility to evaluate and select the chosen design parameters in a scale-free form and ensures an optimal selection in terms of evaluation. The evaluation and selection approach represents a generic methodology for optimal choices in future knowledge-based design systems involving rule-based techniques.

1 Introduction

Accurate weight estimation in the early conceptual design stage in the aircraft industry is crucial since the performance characteristics of a design in subsequent design phases are highly affected by the weight. The ability to precisely predict weight in the early design phase will help to reduce costs and minimize design iterations.
Existing well established weight estimation formulae for structural components are often based on statistical methods [1, 2, 3]. The accuracy is, however, confined to such aircraft designs which are similar to designs of the underlying database. This apparent deficit clearly identifies the necessity for a physics-based augmentation. Hence, modern design approaches use both analytical and numerical calculations for weight prediction [4].
In the past, several strategies have been used either to optimize a structural concept, or to design a structure with regard to several design objectives (e. g. stiffness, strength, stability). The methodology of Ashby [5] for example focuses on material and process selection in mechanical design. The goal of the methodology reviewed in this paper is to provide a simple, fast and reasonably accurate weight forecast and evaluation method of structural joints at the conceptual design phase. Based on analytical weight formulae, a systematic approach is applied by using dimensional analysis in order to provide a generic evaluation method for minimum weight selection of different joining technologies.

[1] Formerly researcher at the Institute of Aircraft Design, University of Stuttgart

2 Analytical Weight Equations

Mass functions in symbolic form are derived from structural models (single lap joints under tensile load) and from analytical approaches for different joining technologies (e. g. riveting, adhesive bonding, welding) appropriate for conceptual design to optimize performance; which in this case is similar to weight minimization [6].

2.1 Riveted Joints

In the symbolic equation (1) a riveted single lap joint is exemplarily chosen and the dimensioning is based on performance of an ideal ductile material. Loads are transmitted through bearing stress [6]. The equation for minimum weight is as follows:

$$\frac{G_{\min}}{g b l^2} = \frac{j}{2 C_{LA} R_m R_{m_N}^2 m \eta^2} \left(C_{LD}\, j \left(4 C_{LA} C_R + 2 C_{LA}^2 (n-1) - n\pi \right) R_{m_N} \rho \right.$$
$$\left. + C_{LD}\, jn \left(\pi R_{m_N} + 2\sqrt{3} R_m C_L \beta \right) \rho_N \right) K^2 + \frac{j\rho}{R_m \eta} K \tag{1}$$

with $K = F/b\,l$ (structural index) [7].

3 Dimensional Analysis

The application of dimensional analysis to the relevance list of an equation results in a complete set of dimensionless products. These establish the groundwork for a systematic evaluation method suitable for weight comparison of different joining technologies.
For arbitrary complete physical equations $f(x_1,\ldots, x_n) = 0$ of n dimensioned variables the Pi-Theorem is applicable. The modern formulation of Pi-Theorem according to Görtler [8] is: ”From the existence of a complete and dimensionally homogeneous functional relationship $f(x_1, \ldots, x_n) = 0$ of n physical quantities $x_i \in R^+$, the existence of a functional relationship $F(\pi_1,\ldots, \pi_m) = 0$ with only $m = n - r$ dimensionless parameters π_j is always guaranteed”. The dimensionless parameters π_j (also called similarity variables or dimensionless products) take the form

$$\pi_j = x_{r+j} \prod_{i=0}^{r} x_i^{-\alpha_{ji}} \tag{2}$$

where $j = 1,\ldots,m \in N$ and $\alpha_{j1},\ldots, \alpha_{jr} \in R$ and with r as the rank of the dimensional matrix formed by means of the n quantities, see Rudolph [9, 10] for more details.

3.1 Characteristic number for structural joints

The intensive examination with derived pi-sets from (2) leads to a definition of a significant, “characteristic” dimensionless number. The **weight efficiency factor π_{Weight}** is definied as shown in equation (3). With the dimensionless parameter $\pi_A = 1/\pi_{Weight}$ only

the additional weight due to the extra volume of the joint is considered which results in the **optimum function** (refer to equation (4)).

$$\pi_{Weight} = \frac{F}{G}\frac{l}{L_R} \tag{3}$$

with $L_R = R_m / g\,\rho$ (tension length)

$$\pi_{\Delta Weight} = \pi_{A_{Jo\,int}} - \pi_{A_{Re\,ference}} \tag{4}$$

with $\pi_A = 1/\pi_{Weight} = (G\,L_R)/(F\,l)$

3.2 Riveted Joints

From the relevance list of equation (1) the dimensional matrix can be build up. By rank preserving operations (i. e. elimination of redundancy[2]) the matrix is brought into a diagonal form [9]. From the dimensional matrix consisting of $n = 15$ physical quantities and a rank with $r = 3$, $m = n - r = 12$ dimensionless products are obtained. The defined dimensionless characteristic numbers ("pi-set") for the description as an optimum function have been introduced and the result is shown in Figure 1.

$$\pi_A = \frac{G}{F}\frac{L_R}{l} \qquad \pi_B = \frac{F}{R_m l b} \qquad \pi_C = n \tag{5}$$

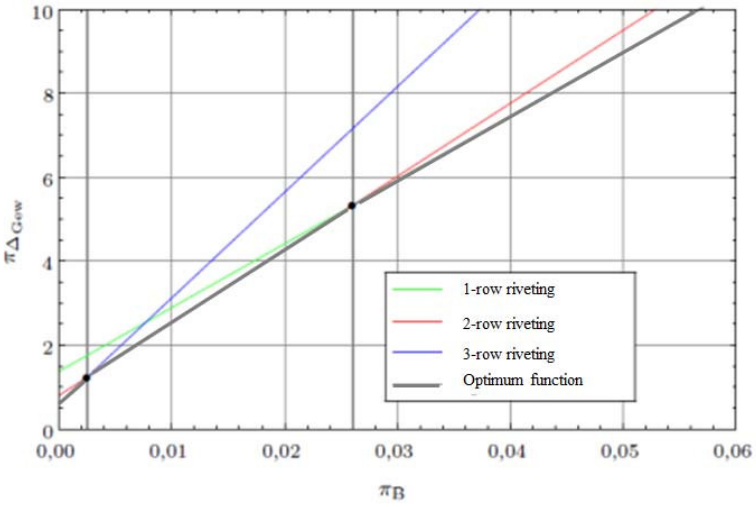


Figure 1 – Dimensionless additional weight of a single lap joint for riveted joints (1-row, 2-row, 3-row riveting) and the optimum function [6]

[2] redundancy-free representation is called "minimal" when the number of independent parameters cannot be further reduced [9].

Detailed analytical mass equations for structural joints (exemplarily selected) have been developed and, building on these, dimensionless parameters resulted from dimensionless analysis, make it possible to evaluate structural joints for minimum weight. For different joining technologies the dimensional analysis results in only three dimensionless products (refer to equation (5)). This pi-set enables a weight comparison by a so-called dimensionless optimum function [6] which can be used as a general evaluation method.

5 Conclusions

For aircraft manufacturers accurate weight prediction at early stages of preliminary design is vitally important to guarantee performance characteristics to customers. Current weight approximation formulae are often based on mathematical statistics which have been established in preliminary aircraft design due to their advantage of fewer input parameters and due to the lack of detailed CAD data. These known weight approximations result from data of existent aircraft and reflect the technology of the aircraft within that database. Therefore, alternative new approaches increasingly apply both analytical and numerical analyses. Besides, structural joints must be included in weight estimation processes to guarantee a more detailed and precise prediction result.
This paper presents and reviews a logical and scientifically stringent development of an evaluation strategy. The novel methodology has been used to determine analytical mass equations and to generate selection charts for different structural joints.
This concept of scale-free design knowledge enables an effective selection process for weight-minimal solutions for structural joints. Furthermore, this methodology represents a basic requirement for rule-based engineering and can be employed in implementing future applications of computer-aided aircraft design.

References

[1] D. P. Raymer, Aircraft Design: A Conceptual Approach, AIAA, Washington, D. C., 1989.

[2] J. Roskam, Airplane Design : Preliminary Sizing of Airplanes, Vol. I, DAR Corporation, Kansas, 2003.

[3] E. Torenbeek, Synthesis of Subsonic Airplane Design: an introduction to the preliminary design of subsonic general aviation and transport aircraft, with emphasis on layout, aerodynamic design, propulsion and performance, Kluwer Academic, Dordrecht, 1999.

[4] F. Hürlimann, Mass Estimation of Transport Aircraft Wingbox Structures with a CAD/CAE-Based Multidisciplinary Process, Dissertation, Eidgenössische Technische Hochschule, Zürich (2010).

[5] M. F. Ashby, Materials Selection in Mechanical Design, Elsevier, Amsterdam, 2005.

[6] L. Beilstein, Eine Methodik zur analytischen Gewichtsabschätzung und Bewertung von Strukturverbindungen im Flugzeugvorentwurf, Dissertation, Universität Stuttgart, Stuttgart (2012).

[7] J. Wiedemann, Leichtbau : Elemente und Konstruktion, Springer, Berlin, 2007.

[8] H. Görtler, Dimensionsanalyse: Theorie der physikalischen Dimensionen mit Anwendungen, Springer, Heidelberg, 1975.

[9] S. Rudolph, Übertragung von Ähnlichkeitsbegriffen, Habilitation, Universität Stuttgart, Stuttgart (2002).
[10] S. Rudolph, Eine Methodik zur systematischen Bewertung von Konstruktionen, Dissertation, Universität Stuttgart, Stuttgart (1994).

Intelligent requirement analyzis and detection for product development

Akram Chamakh, Dieter Roller

Institute of Computer-aided Product Development Systems
Universität Stuttgart
Universitätsstr. 38
D-70569 Stuttgart
Akram.Chamakh@informatik.uni-stuttgart.de
Dieter.Roller@informatik.uni-stuttgart.de

Abstract: The main goal of this paper is to present a software solution to analyze and to detect textual based requirements in the first stage of product development. After detecting the orders and requirements, the algorithm is supposed to generate automatically the related codes used for the identical requirements. This code will be used later in a related software or machine.

1 Introduction

Requirements management is gaining importance especially in design of complex products and systems. The subdivisions of requirements management are integrated from documentation processes, analyzis, tracking, prioritization and coordination of the requirements.

2 The proposed approaches

The textual based requirements are semantically detected, analyzed and categorized. The goal of syntactic and semantic analysis is to decompose a given text using the methods of computational linguistics in the smallest units (morphs) and then to provide them with both syntactic and semantic annotations [Ab83] [NE08]. In the following sections two different models to translate the requirements into formal models are proposed and explained.

2.1 Analysis and Evaluation of Requirements Documentation

In this model, the specified textual requirements are semantically analyzed and categorized into appropriate collections, as shown in Figure 1. The application domain at

this point plays essential role to connect the concepts with convertible information structure. In this context, a novel system was designed to analyze and evaluate the biased requirements. Consequently, the analyzed requirements are tested by a Requirements Extractor and Classifier system (REC) [1] whether they are classified correctly as a requirements or not. To accompolish this task, they are divided into classes using ontology [Sta]. These classes comprise weak requirements, optional requirements, candidate requirements and no requirements.

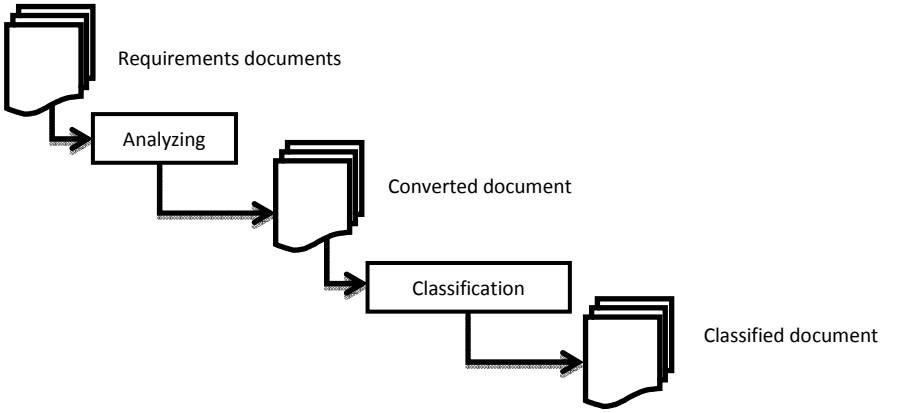


Figure 1: Workflow and technical steps of converting requirements documents into classified documents

The application domain typically provides very specific requirements of the technical system to be used for simulation or management of domain-specific tasks and problems [FIC08]. The arising requirements in this phase are incorporated particularly in the requirements analysis [Ge10]. The requirement analysis proceeds to a system development consequently to the modeling and to the generate the model for the realization. For further processing, the information obtained in the analysis phase is converted in a semantic model, which leads to an early detection of the requirements in the software-converting system [La10]. The static requirements analysis in the development time period and after that the dynamic requirements are especially time dependence. An example of a dynamic product requirement analysis is "to choose the most efficient solution". When the context dependence is rated high, the analysis will be more complex as it refers to specific needs and effective factors [RAC11].

A textual example for a context sensitive requirement is the following sentence: "the performance improvement of the previous versions". Considering the explicit and the implicit requirements, it becomes even more complex since the requirements are incorporated, in both current and other contexts. An implicit textual example of the requirements is presented by the following expression: "the system should be implemented with the latest object-oriented technology".

[1] REC uses the Stanford CoreNLP for all Natural Language Processing (NLP) analysis.

2.2 Detection and regenerating of requirements using a product configurator

Using this model, the requirements analyzis is performed for some other applications as well as for the product configurator. Presently, a configurator is one of the company's main and highly reliable software systems. Within the globalization scope, it is vital for almost all companies to be able to present their own individual product design of the national as well as the international market. In this regard, a product configurator is created to optimize and to extend the classical arrangement of functionalities of the configurator [EJO13]. The aim has been so defined to understand the requirements of the user in a more efficient way. Besides reaching the targeted solution, another aim is to fulfill and to recommend the complete requirements. In addition, the market feedback and the user satisfaction have crucial importance in this approach.

3 Conclusion

The requirement analyzis model should be created using intelligent procedure, like the detected structures, as a fundamental model or regenerating of requirements using a product configurator. Subsequently, this model is converted into the source code and is supplemented through adjustments of already available modules and is expanded into a completed software solution.

References

[Ab83] R. J. Abbott: Program Design by Informal English Descriptions.In: Communnications of the ACM 26 (1983), November, Nr. 11, S. 882–894. DOI 10.1145/182.358441. – ISSN 0001–0782.

[FIC08] S. A. Fahmi, A. Ibrahim, H.-J. Choi. An Approach to Support and Partially Automate Requirements Engineering Activities. 8th International Conferrence on Computer and Information Technology Workshops, 2008.

[Sta] Stanford Natural Language Processing Group. The Stanford NLP Group. URL http://nlp.stanford.edu/.

[La10] M. Landhaeusser. Automatische Auszeichnung von Semantik in natuerlichsprachlichen Spezifikationen. Diplomarbeit, Karlsruher Institut fuer Technologie, 2010.

[Ge10] T. Gelhausen, Modellextraktion aus natürlichen Sprachen – Eine Methode zur systematischen Erstellung von Domänenmodellen, KIT Scientific Publishing, ISBN 978-3-86644-547-5, Karlsruhe, 2010.

[NE08] N. Niu, S. Easterbrook. Extracting and Modeling Product Line Funcional Requirements. 16th IEEE International Requirements Engineering Conference, 2008.

[RAC11] R. Rauf, M. Antkiewicz, K. Czarnecki. Logical Structure Extraction from Software Requirements Documents. 19th International Requirements Engineering, Conference, 2011.

[EJO13] B. Erkus, D. Jakupovic, K. Oueslati. Configurator for project requirement, ProjINF, University of Stuttgart, 2013.

Optimizing patients' engagement by a cyber-physical rehabilitation system

Chong Li[1, 2], Zoltán Rusák[1], Imre Horváth[1], Yuemin Hou[2], Linhong Ji[2]

1 Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15,
2628 CE, Delft
C.Li-1@tudelft.nl
Z.Rusak@tudelft.nl
I.Horvath@tudelft.nl

2 Division of Intelligent and Bio-mimetic Machinery
The State Key Laboratory of Tribology
Tsinghua University
100084, Beijing
hym01@mails.tsinghua.edu.cn
jilh@tsinghua.edu.cn

Abstract: Successful rehabilitation of stroke patients is strongly dependent on the engagement of patients. During a whole rehabilitation program, mundane rehabilitation exercises can easily become routine for patients, leading to boredom and as a result to ineffective functional recovery. This has been taken into consideration just by few rehabilitation systems. Engagement in rehabilitation can be decomposed into long term engagement (LTE) and short term engagement (STE) for the reason that their indicators and stimulation strategies are different. This paper proposes various engaging strategies to optimize both LTE and STE of patients. Actually, proposes to combine the strategies concerning the whole rehabilitation program and the strategies to be applied during single rehabilitation exercises. Based on the proposed reasoning model, a cyber-physical computing-based solution for a personalized rehabilitation system is proposed. Various cyber-physical characteristics, such as function augmentation, adaptive and learning control, are being implemented in the system in order to realize the various strategies of engagement.

1 Introduction

Engagement in rehabilitation exercises has been defined as a construct that is driven by motivation and executed through active and effortful participation [LK10]. A positive rehabilitation outcome is strongly associated with high patient motivation and

engagement [La11]. Therefore, increased engagement during rehabilitation exercise is essential to achieve good outcomes in rehabilitation.

Researchers have studied long term participation in physical activities after rehabilitation, which is important for the survivors to sustain the benefits from rehabilitation [De06; MW09]. However, few researchers have studied engagement during rehabilitation program. Although stroke patients are required to participate actively during the rehabilitation program, it does not mean that they are mentally engaged. In fact, by definition, engagement extends participation beyond exercise execution and motivation [Ko07]. The difference between engagement and participation in this context is that engagement involves high levels of invested interest [LK10]. Several studies applied virtual reality [Zi13] and serious games [Bu09] in order to increase engagement during rehabilitation training exercise. In our previous study [Li14a], we found that, even if video games are used, the subjects' engagement level will decrease when they become familiar with the games. Therefore, a system which can adapt strategies to engage different patients is needed.

This paper aims at exploring how and why stroke patients engage in both short term and long term, and setting up a system model which can trigger and maintain engagement in short term and long term. The next section introduces the engagement model and the strategies for long term engagement (LTE) and short term engagement (STE). The third section proposes a system model with cyber-physical solutions in order to realize the goal of engaging the patients in both terms.

2 A reasoning model about the forms of engagement

Introduced above, we propose to consider engagement in rehabilitation as a combination of LTE and STE. LTE is defined as engagement throughout a whole rehabilitation program conducted in a hospital or a rehabilitation center, while STE is defined as engagement during a rehabilitation exercise. The reason of suggesting this classification is that their indicators are different, as well as the stimulation strategies that can be used to optimize the patients' LTE and STE. Indicators of engagement can be described in three aspects, which are cognitive, emotional and behavioral [Ai05]. In the context of stroke rehabilitation program, the indicators are striving to recover (cognitive), enjoyment of rehabilitation progress (emotional) and participation in everyday program (behavioral). Indicators of STE can be described as striving to exercise (cognitive), facial expressions (emotional) and intensity of muscle activities (behavioral) [Li14a].

The following figure shows a model about the forms of engagement. To optimize patients' engagement during rehabilitation program, strategies for LTE and STE should be combined. Before rehabilitation exercise, strategies for LTE will be applied. In this way, the patients can have an overview of the whole rehabilitation program. What the patients can expect from the whole program will be clear to them, which will contribute significantly to their willingness to comply with the whole treatment process. Then during single rehabilitation exercises, strategies for STE will be applied.

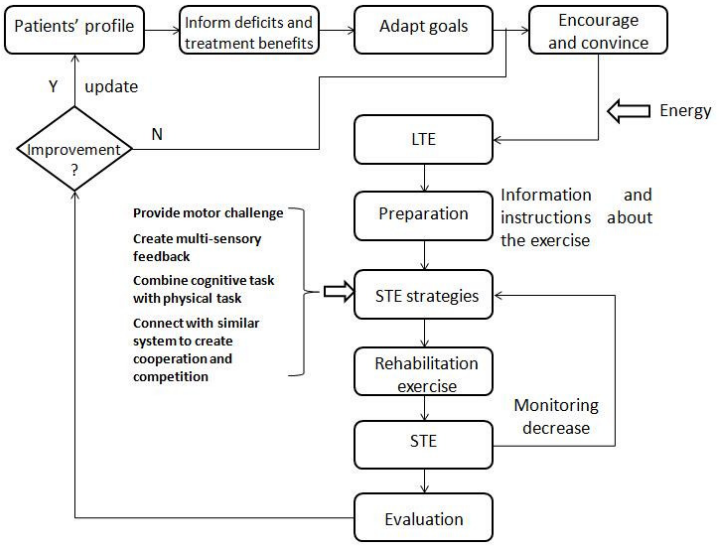


Fig.1 Model of strategies to optimize LTE and STE

LTE strategies can be explained based on the model proposed by Lequerica and Kortte for therapeutic engagement in rehabilitation [LK10]. The first strategy for LTE is informing the patients about their deficits and potential outcomes according to patients' profile. Then the patients can become aware of their own deficits, and this will contribute to their willingness to comply with the treatment and try to reach the goals. The potential benefits of the treatment will also be communicated to them. In this way the patients can know what they can expect and what they can benefit from the rehabilitation program so that they have the motivation to comply with the treatment. Next strategy is setting a proper goal for the patient. A goal, such as completing a task or reaching a Fugl Meyer score, is important because it gives the patients a clear goal and direction to make effort. Patients with different deficits or in different conditions should have different goals. The third strategy is to encourage and convince the patients. Patients' self-confidence or self-efficacy also contributes to engagement in rehabilitation program. To motivate the patients, physical therapists usually encourage them to give them confidence and convince them they have the ability to complete the tasks. Since the rehabilitation program is a long process, patients will be depressed inevitably. Therefore, it is important to encourage and convince the patients through the whole rehabilitation program. The last strategy for LTE is to update the goals. Updating the goals also plays an important role in maintaining engagement. This is because if the previous goal has been achieved, the patients opt to disengage according to the therapeutic engagement model. Therefore, it is of important significance to update the goals to engage the patients continuously along the rehabilitation program.

Strategies for STE have been identified in our previous study [Li14b]. It starts with a preparation stage, in which the patients will be informed about the objective of the exercise, the exercise contents and how to do the exercise. They will apply strategies for providing motor challenge, creating multi-sensory feedback, combining cognitive tasks

and physical tasks and providing cooperating and competing training. Engagement level will be monitored using the indicators of STE, such as intensity of muscle activities and facial expressions. Monitoring engagement level will provide information to the system about the engagement level of the patients, and the system can intervene in order to stimulate the patients. Then the maintenance of STE could contribute to LTE by driving and motivating the patients back to do the exercise next time.

3 Cyber-physical based rehabilitation system

3.1 System workflow

The objective of this system is to optimize LTE and STE in rehabilitation. Figure 2 shows the system workflow, which is designed to realize the strategies for LTE and STE. The big loop represents the LTE strategies, and the small loop represents the strategies for STE.

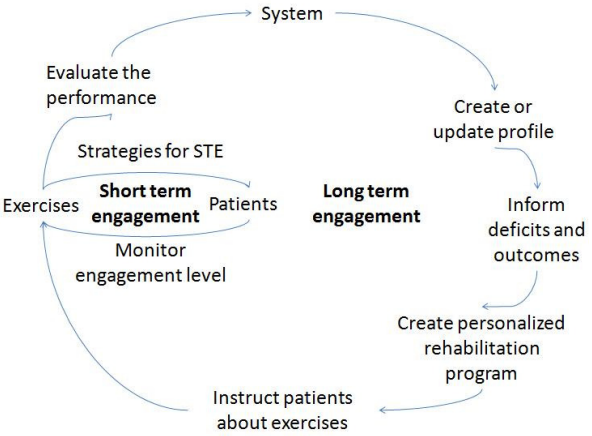


Fig.2 System workflow

3.2 System function

3.2.1 Personalizing rehabilitation program

(1) Create and update patient's profile: Patients' profiles will be the basis of personalized rehabilitation. The profile includes patient's information, name, gender, background, hobby, occupation, side of paresis, Fugl Meyer score for shoulder/elbow and wrist/hand, Motor Power score and Functional Independence Measure. (2) Inform the patient about deficits and potential outcomes: Before the personalized exercises start, the system can use a video or an avatar of the arm to show the patients their deficits and inform them their potential outcomes as well. (3) Generate personalized exercises for different patients: The physical therapist will decide which movements of their impaired limb or

joint should be trained. Then the system will generate personalized exercises for the patients, which are in different contexts. The system will list the exercises that could probably interest them according to their hobbies and backgrounds. (4) Inform and instruct the patient about the exercises: The system will encourage and convince the patients that they are able to do it so that the patients will be confident when they are doing the exercises. Then system will instruct them how to move to complete the tasks.

3.2.2 Monitoring engagement level

(1) EMG: According to our previous research, Root Mean Square (RMS) of EMG signal is found to be related to the subject's engagement level. The system can monitor and analyze the EMG signals of the muscles that are involved in the training exercise in order to be aware of the changes of the subject's engagement level. (2) Facial expression: Facial expression will also be used to indicate the engagement level. According to our previous research, when the subjects are smiling, usually their engagement level is higher [Li14a]. Therefore, the system will also monitor and analyze the facial expression.

3.2.3 Short term stimulation the patients

(1) Provide motor challenge: Every game will have different difficulty levels. If the patient completes the task in a certain level, the system will move to the next level and keep challenging the patientin order to prevent the patient will lose interest and get bored with the exercise. (2) Create multisensory feedback: The system will provide visual, auditory and tactile feedbacks to the patients in order to make them aware of the training exercises. Since some patients prefer certain feedback than others, the system could also provide personalized exercise that focuses on training that certain feedback. For example, if the patient's interest is listening to music, then the system could deliver a training which focuses on auditory feedback. (3) Combine cognitive tasks with physical tasks: The system can combine cognitive tasks and physical tasks in order to engage the patients. To realize this function, game exercises requiring cognitive actions can be delivered to the patients, and at the same time, the patients will be required to move a robotic device integrated with a human computer interface to complete the tasks. In this way the system can combine the two tasks. (4) Connect with similar systems to create cooperation and competition: Similar systems can be connected to create cooperation and competition training for the patients.

3.2.4 Training by learning

(1) Evaluate and memorize the results: After each exercise, the system will evaluate the stimulation strategies for the subjects and evaluate the performance of the patients in each game. The stimulation strategies will be evaluated by the changes of the engagement level after the stimulation strategy is used. If the engagement level increases, then it means this strategy works for this patient. Similarly, the performance of the patients will be evaluated by the game score of the patient. These results will be memorized in the system and updated in the patients' profiles. (2) Learn from the history: When another patient comes the next time, the system will apply a stimulation strategy which has been proved to be successful for a patient having similar background and the same hobby, so that the system can avoid the trial and error period in order to be more effective to engage the patient. The system could also make the game available with the proper difficulty level for the patients with similar FM scores.

4. Conclusion

This paper demonstrates that LTE of a whole rehabilitation program and STE during single rehabilitation exercises are both related with rehabilitation outcome and should be achieved in different ways for the reason that their indicators and stimulating strategies are different. Based on the reasoning model and the engaging strategies, framework for a cyber physical rehabilitation system is proposed with the intention to optimize the patients' engagement in both long term and short term. Although some systems aiming to engage the patients with game features have been introduced before, we found that the subjects' engagement level will decrease when they become familiar with the exercise integrated with video games. The proposed system may engage the patients to a higher level and for a longer time because it can provide personalized rehabilitation program and it can be adaptive based on patients' real time engagement level. Implementation and validation of the system is a task to be completed in the future.

References

[Ai05] Ainley, M., Frydenberg, E., & Russell, V. J. Student motivation and engagement. Schooling Issues Digest, 2005, 2, 1-11.

[Bu09] Burke, J. W., McNeill, M. D. J., Charles, D. K., Morrow, P. J., Crosbie, J. H., & McDonough, S. M. Optimising engagement for stroke rehabilitation using serious games. The Visual Computer, 2009, 25(12), 1085-1099

[De06] Desrosiers, J., Noreau, L., Rochette, A., Bourbonnais, D., Bravo, G., & Bourget, A. Predictors of long-term participation after stroke. Disability & Rehabilitation, 2006, 28(4), 221-230.

[Ko07] Kortte, K. B., Falk, L. D., Castillo, R. C., Johnson-Greene, D., & Wegener, S. T. The Hopkins rehabilitation engagement rating scale: development and psychometric properties. Arch Phys Med Rehabil, 2007, 88(7), 877-884.

[La11] Langhorne, P., Bernhardt, J., & Kwakkel, G. Stroke rehabilitation. Lancet, 2011, 377(9778), 1693-1702.

[Li14a] Li, C., Rusák, Z., Horváth, I., Ji L. The influence of complementing a robotic upper limb rehabilitation system with video games on the engagement of the subjects a study focusing on muscle activities. International Journal of Rehabilitation Research, in press, 2014.

[Li14b] Li, C., Rusák, Z., Horváth, I., Ji, L., Hou, Y. Current status of robotic stroke rehabilitation and opportunities for a cyber-physically assisted upper limb stroke rehabilitation. Proceedings of TMCE 2014, May 19-23, Budapest, Hungary.

[LK10] Lequerica, A. H., & Kortte, K. Therapeutic engagement: a proposed model of engagement in medical re habilitation. Am J Phys Med Rehabil, 2010, 89(5), 415-422.

[MW09] Morris, J. H., & Williams, B. Optimising long-term participation in physical activities after stroke: exploring new ways of working for physiotherapists. Physiotherapy, 2009, 95(3), 227-233.

[Zi13] Zimmerli, L., Jacky, M., Lünenburger, L., Riener, R., & Bolliger, M. Increasing patient engagement during virtual reality-based motor rehabilitation. Arch Phys Med Rehabil, 2013, 94(9), 1737-1746.

Approach to Learning Production Rules for Grammar-Based Functional Design

Julian R. Eichhoff, Dieter Roller

Institute of Computer-aided Product Development Systems
University of Stuttgart
Universitätsstr. 38
70569 Stuttgart, Germany
julian.eichhoff@informatik.uni-stuttgart.de
dieter.roller@informatik.uni-stuttgart.de

Abstract: This paper outlines the problem of efficiently formalizing functional design rationale in form of graph-rewriting production rules. It is argued that genetic programming provides a feasible machine learning approach to this.

1 Introduction

According to the VDI 2221 design guideline functional design is a crucial step in conceptual product design [VDI93]. Its goal is to define the purpose of the product in terms of input-output-relations on forms of energy, materials and signals. These so called functions are composed in a graph-based model termed function structure. Graph grammars and graph-rewriting systems in general have been used to automatically derive fully evolved function structures from simplified black-box models, e.g. [SC04]. These systems rely on a set of production rules for modifying function structures. Iterative application of rules results in step-wise model changes, where some sequences may lead to the desired results. However, since rules are formal representations of human design rationale their elicitation can become costly. E.g. [SC04] put a great effort in analyzing commonalities among function structures to acquire their rule set. In order to address this drawback we propose a machine learning (ML) approach to the acquisition of rules that grounds on the following hypothesis: Production rules can be efficiently learned from sample function structures using genetic programming (GP).

2 Counter Argument

One may argue that production rules cannot be acquired by ML since they resemble human design rationale that can be only accessed asking knowledgeable experts. Nonetheless, suppose sample data is accessible, i.e. there are pairs of black boxes and resulting function structures. Then hypotheses – here in the form of rule sets – can be formulated on how black boxes were transformed into function structures given that the

space of possible hypotheses is known. ML methods can be used to automatically search this space for hypotheses that best describe the sample data [Sa11].

3 Support for Learning Rules with Genetic Programming

Quite similar to our goal [Ma06] used GP to discover relationships among the set of design parameters and objectives. [YSA10] employed a genetic algorithm to find rules and rule sequences for growing shapes, whereas [OCB08] took a partial component analysis approach to discover shape grammar rules. Further support is also given by preliminary successes in the implementation of a prototype system. In this system a rule is represented by a tree consisting of elementary graph operations (find/add/link functions, etc.). The GP framework maintains a set of alternatives for the next best rule. This so called population is iteratively refined using the evolutionary principles of selection, mutation and recombination [Ko92]. To assess the fitness of rule candidates we consider the validity and applicability of rules as well as the validity of function structures produced by the current rule set and their similarity with sample function structures. First experiments show that simple rules can be learned with this prototype.

4 Conclusion

Using GP to learn production rules from sample function structures is feasible since: The suggested method has been used before in order to explicate design rationale from sample designs; other works specifically addressed the task of learning production rules; and proof of concept is given by a prototype implementation. In our future work we will extend the prototype to learn complex rules and complete rule sets.

References

[VDI93] Verein Deutscher Ingenieure: Methodik zum Entwickeln und Konstruieren technischer Systeme und Produkte (VDI 2221). Beuth, Berlin, 1993.

[SC04] Sridharan, P.; Campbell, M. I.: A Grammar for Function Structures. In Proc. of the ASME 2004 IDETC/CIE Conferences, Salt Lake City, 2004; pp. 41-55.

[Sa11] Sammut, C.: Learning as Search. In (Sammut, C.; Webb, G.I. Eds.): Encyclopedia of Machine Learning. Springer, New York, 2011; pp. 572-576.

[Ma06] Matthews, P.C.; Standingford, D.W.F.; Holden, C.M.E.; Wallace, K.M.: Learning inexpensive parametric design models using an augmented genetic programming technique, AIEDAM 20(1), 2006; pp. 1-18.

[YSA10] Yogev, O.; Shapiro, A.; Antonsson, E. K.: Computational Evolutionary Embryogeny, IEEE T. Evolut. Comput. 14(2), 2010; pp. 301-325.

[OCB08] Orsborn, S.; Cagan, J.; Boatwright, P.: A methodology for creating a statistically derived shape grammar composed of non-obvious shape chunks, Res. Eng. Des. 18(4), 2008; pp. 181-196.

[Ko92] Koza, J. R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection, MIT Press, Cambridge, 1992.

BPMN Process Design for Complex Product Development and Production

Dieter Roller, Erik Engesser

Institute of Computer-aided Product Development Systems
Universität Stuttgart
Universitätsstrasse 38
D-70569 Stuttgart, Germany
Dieter.Roller@informatik.uni-stuttgart.de
Erik.Engesser@informatik.uni-stuttgart.de

Abstract: In automotive development a lot of new power engine concepts were developed. Beneath the combustion engines hybrid drive systems and electric engines were developed. Consumers, public institutions and politics demand alternative drive systems to avoid air pollution, noise and the consumption of oil resources. Therefore electric drive products with innovative technologies were developed additionally. The problem for production plants is to today how these new technologies can be produced additionally to the existing cars with combustion engines. The objective of this paper is to design processes for complex product development with a lot of product variants an production processes which have to produce the complex product variants. The approach is to use BPMN 2.0 (Business Process Modelling Notation) to model the processes between development and production. BPMN 2.0 supports graphical process modelling and executable process modelling. The consequences of complex product development with high alternative drive systems to production lines were analyzed by BPMN models. A further approach is to use the research tool PROCAS (Process Optimization, Control, Analysis and Simulation) to analyze the consequences of high product variants in production. PROCAS use information of existing CAD (Computer Aided Design) and CAM (Computer Aided Manufacturing) systems. The concept of PROCAS is to determine how existing plants, production lines and inventories have to be extended.

1 Introduction

1.1 Problem and Challenge

The shortage of oil resources leads the development to find alternative drive systems. Rising oil prices causes direct cost of mobility [LI12]. The demand of alternative drives for vehicles is discussed by automotive producers, politicians and customers. Electric vehicles are one of the alternatives. Another motivation to produce electric vehicles is to reduce the atmospheric noise especially in big cities. The smoke in big cities like Beijing

damages the health and wellness of the population. The air pollution caused by the traffic could be eliminated with electric vehicles. In future the utilization of electric vehicles will be in cities and for short distances. New types of cars are already produced by several automotive manufactures. For example in Figure 1 is a vehicle with electric drive based on hydrogen fuel cell shown. Like in [CO12] described hydrogen fuel cells produce electric power with high efficiency. Essential advantage of hydrogen fuel cells compared to batteries is that currently the chemical energy storage in liquid or gaseous form has a higher energy density. Difficulty of hydrogen operation is the storage in the vehicle.

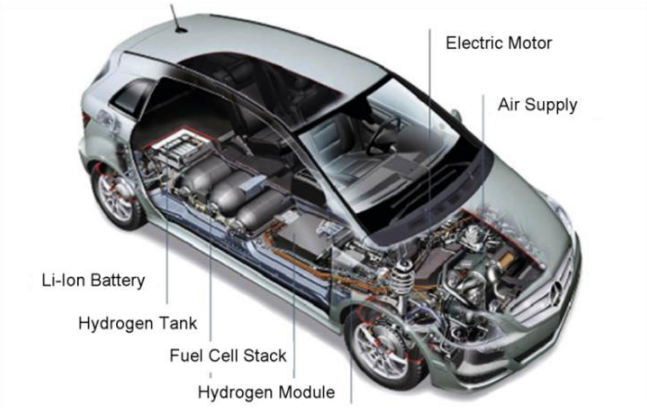


Figure 1: Vehicle with electric drive based on hydrogen fuel cell [CO12]

One research problem is how production of electric vehicles can be done. The existing production plants are manufacturing fuel powered vehicles. The demand of electric vehicles will not replace the fuel powered vehicles. So the variants of vehicle types will rising up. This means more complexity in production. Different kinds of product variants have to be built in the same production lines. In the automated production the robots have to do different kind of production workflows.

The research objective is to make an overview of the process workflow of product development, production and supply chain. The overview of the process workflows gives a possibility to illustrate how development and production planning and control departments work together. New process workflows for electric car production have to be integrated into the existing process workflows. Existing and new process models have to be simulated to examine if they work well and if there are any bottlenecks. After simulation the process workflows have to be optimized. One objective of the optimization is to reduce the complexity of the workflows. The objective of optimization is to avoid process wait times so that the production lead time can be reduced. A further optimization purpose is to reduce process steps so that the production can be done with less technical equipment like robots and with fewer employees to reduce costs. The development of PROCAS supports to model the development and production control processes in BPMN. Process workflows have to be designed graphically to describe and get a better understanding about the workflows. Additionally the BPMN process models visualize the interaction between the workflows of the development, suppliers, and

production plants. A further purpose is the development of an interface to CAD systems to get information about the parts of fuel powered and electrical devices. CAM systems deliver data about production steps and operating procedures.

In chapter 2 complex product development is explained. Chapter 3 introduces BPMN 2.0 [LE12] modelling and process models of product development and production. Chapter 4 shows process modelling and simulation with PROCAS. The conclusion is chapter 5.

2. Complex Product Development

2.1 Extension of Product Structure and Production Steps

Product development develops product structures where all product variants are organized in tree structures. In figure 2 the extension of product structure for alternative hybrid and electric drive cars is shown. The complexity of products increase with new variants. It is enlarged with new variants, modules, part groups and parts. Additionally there are more parallel production steps. In figure 2 production step 0 is the final product variant. Step 1 is setting modules together for the final product. Complex product variants have a lot of production steps with various part groups and parts. Manufacturing processes like forming, painting, assembling and quality checking become more complex. Robots have to do more different kinds of production steps and production tools. More part and part group variants have to be delivered and have to be available for assembling. Operators have to be trained for various assembling and testing techniques.

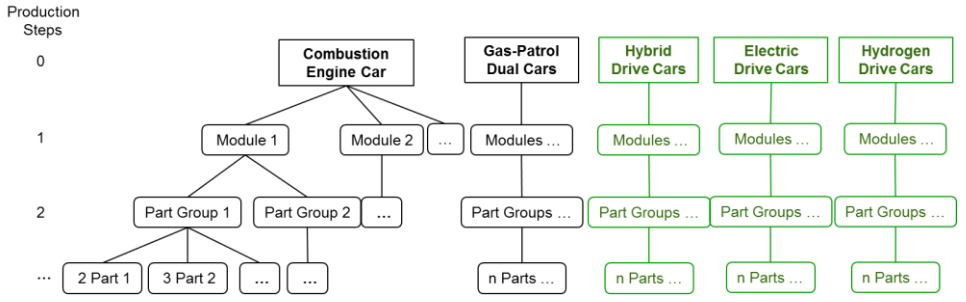


Figure 2: Extension of Product Structure and Production Steps

Thermal drives concepts remain the biggest market have to be produced with efficient technologies. Basic modules of a combustion engines are for example Combustion engine, fuel injection system, valve control system, water pump and charging system. A CNG (Compressed Natural Gas) car needs additional system modules for a dual system with patrol and gas.

As in [CO12] described, all concepts of electro motors are based on electric generated electromagnetic fields. The electromagnetic fields generate magnetic power by induction. The magnetic field remains in equal position for direct current motors. Three-

phase motors have a rotating magnetic field. Depending on the power, torque and efficiency factors different kind of electric motors are used for automotive drives. As in [RE11] described the operation of an electric hybrid vehicle is determined by the operation strategy. It depends on the optimization intention. Intentions can for example be emission or petrol reduction. The operation strategy controls the distribution of the required torque onto the combustion and the electric motor. Intention is to run the combustion motor in a favorable operation point. The classification can also be done by the absolute electric drive power P_{EM}. Therefore hybrid categories like micro hybrid ($P_{EM} < 6kW$), mild hybrid $P_{EM} < 6\text{-}20$ kW) and full hybrid $P_{EM} > 40kW$) exist.

3. Executable Process Modelling with BPMN 2.0

BPMN 2.0 Level 4 modelling is an XML language describing an executable process. Its focus is the XML details underneath [BR09]. In figure 3 a principle production process model is shown. The process model starts with receiving an order and order data as data object. First sub process is building the car body. If complex products are developed production processes have to be split into specific technologies. The exclusive gateway splits production processes with drive system variant information into the specific sub process. Therefore object data of drive system variant is needed. After installing the specific drive the sub processes are joined to the following sub process “assembly car”. The end state of the model is “car finished” output data as data object.

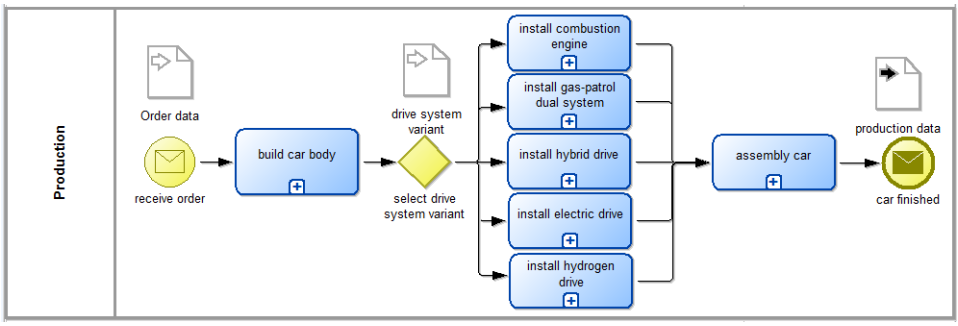


Figure 3: Example process diagram with data objects

A BPMN process model for product development is shown in figure 4. A process request message with the product specification is send. The first human task starts and makes the variant construction of components by CAD. The data object product specification is an input which describes for example product forms, sizes, functionalities, propositions. The designed component construction is the input for the human task construction of component assemblies and modules. This human task joins the components to assemblies and modules. The next human task integrates the modules into the product construction. A service task follows which analyses and simulates the product construction. If the result delivers any errors the previous tasks were informed and corrected. Finally the data objects product construction and analysis results are sent back as response message. The second pool of figure 4 shows the BPMN process model

of a production control. The component, assembly, module and product construction are sent to the production control workflow. First human task is the production planning. Production processes and production facilities like robots and assembly lines are planned. Therefore CAM systems are used. Order requests are received by the distributors. The human task order sequencing creates the production order sequence. Next service task calculate component demands and sends a message request to the suppliers. After receiving the components the next task produce the product. Next task checks the product quality. If there are any quality defects rework have to be done. Finally the produced product is returned to the distributor.

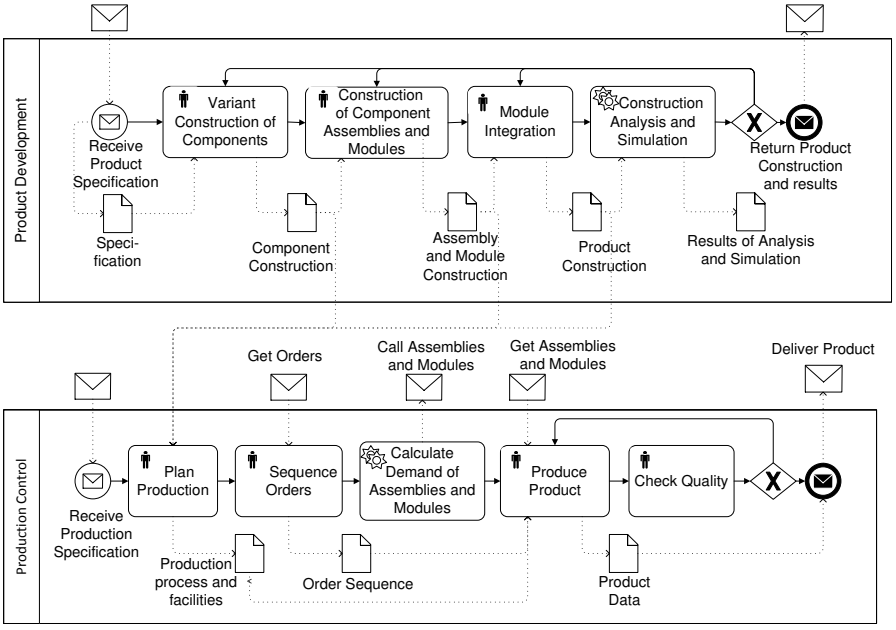


Figure 4: Process Modell of Product Development and Production Control

4 Process Modelling and Simulation with PROCAS

Figure 5 shows the concept of process modelling for the development and production of fuel powered and electric devices by PROCAS. The modelled BPMN process models are loaded into the database. In the shown left tree production steps of the product structure of figure 2 can be selected. A construction of an hybrid gear system is shown in the second window. Therefore CAD data is used as object data. The production workflow CAM data is also stored as object data. A robot assembly line is displayed in the right frame. The concept of PROCAS is to model and simulate BPMN process workflows [RE12]. PROCAS analyses the integration of new process models of electric drive production into existing production processes. Effects of variance increase of alternative drives and car bodies are analyzed. The complexity of production processes is evaluated. Consequences of construction changes on production processes are analyzed.

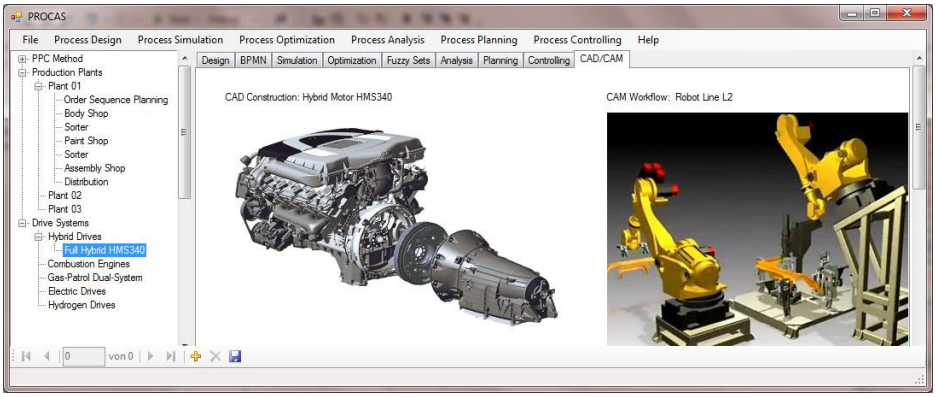


Figure 5: Process Modelling for Development and Production for Hybrid Drives with PROCAS

5. Conclusion

Products become more and more complex by new technologies and module variants. Product development develops complex product structures with various production steps. The rising complexity of product variants causes problems in production to produce all constructed variants. Executable BPMN 2.0 is an adequate approach to model processes of complex product development and production. The interfaces can be designed and executed. The concept of PROCAS is to integrate CAD and CAM information into the BPMN process model to simulate processes and interfaces. Further steps are to develop PROCAS to simulate existing plants. The simulation calculates how production lines and inventories have to be extended for more product variants in future. Input data therefore is BPMN process models of development and production, product structure, production steps, CAD and CAM data.

References

[LI12] Lienkamp, M.:Elektromobilität Hype oder Revolution?, Springer-Verlag, Berlin Heidelberg, 2012.

[CO12] Cornel, S.: Alternative Antriebe für Automobile, Hybridsysteme, Brennstoffzellen, alternative Energieträger, Springer-Verlag, Berlin Heidelberg, 2012.

[LE12] Frank Leymann, BPEL vs. BPMN 2.0: Should You Care, Second International Workshop, BPMN 2010, Potsdam, Germany, Springer-Verlag Berlin Heidelberg 2010.

[RE11] Reif, K.: Bosch Grundlagen Fahrzeug- und Motorentechnik Konventioneller Antrieb, Hybridantriebe, Bremsen, Elektronik, Vieweg+Teubner Verlag Springer Fachmedien, Wiesbaden GmbH 2011.

[BR09] Silver B.:, BPMN Method and Style (Code-Cassidy Press, 2009).

[RE12] Roller D., Engesser E: Modelling, Simulation and Fuzzy Decision Making of Distributed Production Control and Supply Chain Methodologies, IASTED, Crete, Greece, 2012, 781-044.

The Pipeline from CAD to Virtual Reality in VISIONAIR Project

Stefano Gagliardo[1], Franca Giannini[1], Marina Monti[1],
Stefano Mottura[2], Marios Pitikakis[1,3]

[1] IMATI-GE
CNR
Via De Marini, 6
16149 Genova, Italy
{gagliardo, giannini, monti}@ge.imati.cnr.it

[2] ITIA-MI
CNR
Via Bassini, 15
20133 Milano, Italy
stefano.mottura@itia.cnr.it

[3] Softeco Sismat S.r.l.
Via De Marini, 1
16149 Genova, Italy
marios.pitikakis@gmail.com

Abstract: The VISIONAIR project aims at federating and providing a unique access to advanced visualization resources available at research centers and universities from Israel and other 11 countries in Europe. The physical visualisation facilities are supported by the Virtual Visualisation Service Infrastructure providing repositories of shape models and archives of tools for their creation and processing. Here we present the newly developed service for the documentation of workflows aimed at supporting the preparation of Virtual Reality Environments from CAD /PLM data.

1 The VISIONAIR project and the VVS

Visualization tools and techniques provide important support to analyze and display large volumes of multidimensional data, often time varying, in such a way as to allow the user to extract significant features and results quickly and easily. Newly available technologies allow for greater realism in the visualization perception. Unfortunately, the use of such tools is still limited because they are still expensive, and their costs are not always justified with respect to the actual usage for most of researchers. To overcome these limitations and to enhance research and innovation capacities, the European

Commission is supporting the integration and the access to advanced visualization facilities around Europe through the VISIONAIR (VISION Advanced Infrastructure for Research) project financed in the frame of the FP7 Capacities program[1].

In the industrial manufacturing field, Virtual Reality environments offer good facilities for the assessment of product design solutions according to various points of view (e.g. productivity, working conditions, safety and ergonomics). Unfortunately, some limitations exist that prevent the full exploitation of such capabilities. In addition to the high costs, the direct use of design data is not possible because PLM/CAD systems provide detailed specification of the conceived product solution, which does not satisfy requirements of VR environment, both in terms of type and complexity of representation. Therefore, these data need to be prepared and adapted to the VR applications. Understanding the exact requirements to fulfill and the needed data transformation and processing sequence requires specific skill and experience that somehow limits a wider adoption of VR facilities to a larger audience.

The VISIONAIR project aims at federating and providing a unique access to advanced visualization resources available at research centers and universities from Israel and other 11 countries in Europe. Each site offers capacities in one or more specific rendering and visualization technologies including Virtual Reality, Augmented Reality, Holography, and High Quality Image Processing. The involved research teams and institutes have developed complex visualization environments for their own usage. By joining the VISIONAIR infrastructure, they open the access to these environments to the research community. The 29 facilities made available by VISIONAIR are clustered according to their main research and capabilities focus. The first cluster addresses tools and methods to navigate in huge data set issued from scientific simulations. The second cluster focuses on tools and methods that facilitate sharing and visualization of high quality images and videos over very high bandwidth networks. The cluster on Virtual Reality provides high-end immersive facilities that allow users to experience virtual environments and to interact with them. The last cluster involves platforms for synthetic environments mixing a large set of advanced facilities for remote collaborative works.

The access to these facilities is freely open during the project lifetime. The hosting institutions financially and technically support the usage of the physical facilities.

To aid the preparation of the experiments, a Virtual Visualization Service (VVS) facility is provided. It is based on an upgrade of the Digital Shape Workbench, initially developed by the AIM@SHAPE Network of Excellence[2,3] [At13]. The VVS acts as an e-science infrastructure enhanced with collaboration capabilities for the sharing and retrieval of shape data and related processing software tools. It is conceived as an ontology- and web-based resource repository, which involves the integration of the Shape Repository, the Tool Repository and the Ontology and Metadata Repository,

[1] VISIONAIR: VISION Advanced Infrastracture for research, http://www.infra-visionair.eu/.
[2] http://visionair.ge.imati.cnr.it/.
[3] AIM@SHAPE Advanced and Innovative Models and Tools for the Development of Semantic-based Systems for Handling, Acquiring, and Processing knowledge Embedded in multidimensional digital objects, ContractFP6 IST NoE 506766, http://www.aimatshape.net.

together with an advanced Search Framework [Va10]. Within VISIONAIR the VVS is aimed at supporting the experts in several activities either, to find processing tools thus avoiding to implement already existing components and to find and store shape data necessary to run their experiments and to create virtual environments.

2 From CAD to VR

The design and development of a VR application may involve a relevant amount of modules, such as 3D modelling, rendering, interaction, data exchange, behavior, sounds and physics. In industrial applications such as the simulation of manufacturing environments, the 3D models (that can represent very simple objects or complex assemblies or a working environment) are edited and managed so that their characteristic information are preserved as much as possible for being imported by the VR application.

Within the VISIONAIR project, a specific analysis of the elements involved in the CAD to VR pipeline has been carried out, in order to identify some key issues of the migration from CAD to a VR application: the objective was to verify the possibility of suggesting hints on how to overcome them more efficiently or of developing some specific services for their execution.

The first issue derives from the fact that a single specific file format for VR applications, describing and storing the information related to an object in terms of geometry, shape and structure does not exist; a set of various 3D formats are available, each one supporting certain properties and missing other ones. Moreover, software applications for VR are able of correctly loading certain 3D formats but does not cover all the included features. For example, in CAD objects are represented by precise surface models (e.g. NURBS) and in terms of functional features, such as holes, pockets and kinematic joints that are stored in the native file format. On the contrary, the 3D formats for the real-time graphics for visual simulation and virtual reality support a real-time rendering-oriented representation of the 3D model (i.e., tessellated geometry, textures).

Another identified key issue is that users of the CAD tools are, in general, not aware about how to manage the features of the 3D models into a VR application nor about the suitable tools for doing it; therefore, the CAD to VR pipeline is not limited to file conversion. VISIONAIR Partners have worked on the identification of a set of tasks that can be done when going from CAD-tool to VR application in order to preserve the original information or in order to apply some workarounds to the existing limits. This has been done by exploiting the partners' experience and by performing some tests of communication with the available VR facilities.

The categorization and formalization obtained from the analysis is represented with an ontology built on top of the Common Tool Ontology of the VVS. A software tool, exploiting such ontology, has be developed to support people in transforming 3D CAD models tools in a representation suitable for virtual reality environments applications.

Acknowledgments

This work has been partially supported by the European Commission under grant agreement 262044 VISIONAIR. The authors want to thank the VISIONAIR partners participating to JRA 9 and, in particular, Christian Weidig, Lionel Roucoules and Walter Terkaj.

Additional thanks to Prof. Bianca Falcidieno and all the participants to the AIM@SHAPE NoE for making possible the initial version of the Digital Shape Workbench.

References

[At13] Attene, M.; Giannini, F.; Pitikakis, M.; Spagnuolo, M.: The VISIONAIR Infrastructure Capabilities to Support Research. In: Computer-Aided Design and Applications, vol. 10 (5), 2013. Cad Solutions, Bangkok; pp. 851-862.

[Va10] Vasilakis, G. et al.: Knowledge-Based Representation of 3D Media. In: International Journal of Software Engineering, vol. 20(5), 2010. World Scientific Publishing Company, pp. 739-760. DOI: 10.1142/S0218194010004773.

Cooperative CAD System for Conceptual-Embodiment Design Phase of Product Design

Fariz Muharram Hasby, Dieter Roller

Insitut für Rechnergestützte Ingenieursysteme
Universität Stuttgart
Universitätsstr. 38
D-70569 Stuttgart
fariz.hasby@informatik.uni-stuttgart.de
Dieter.roller@informatik.uni-stuttgart.de

Abstract: In the modern globar market and manufacturing environment, industries have to put more focus on reducing time to market while reducing production costs on the development of new products. Collaborative product development is a growing trend that is being used to answer some of that problem. However, literature research has revealed that there are only a small number of researches that has talked about cooperative CAD systems. This paper will try to show that this type of system can be used in the conceptual-embodiment design phase while encouraging creative solutions from the designers involved. It will also list some functional requirements and areas of interest for these type of system that has been identified by this research at the time of writing.

1 Introduction

With the emergence of global market and the development of more sophisticated manufacturing technologies, customers now demand products with better quality while also having a low price. As a result, industries are trying to find ways to cut product development time and product development cost. One of the ways that these goals can be achieved is by applying collaborative techniques in product development, such as using a Collaborative Computer Aided Design (CAD) system.

These types of CAD system allow multiple designers to collaborate on certain projects, thereby enabling division of work and parallel work between designers. Collaborative CAD systems are also used to coordinate and spread important work informations among different functional areas, such as when applied in Concurrent Engineering methods between different functional divisions of a company.

In the future the development of CAD systems will take into account this collaborative aspect, along with the Cognitive, Conceptual and Creative aspects (the 4 C's of future CAD), as the base for subsequent CAD systems development [Go12]. Future CAD systems will aim to assist designers in the designing process itself, rather than just being a tool which is mainly used to document and spread design informations as the way it is being used now.

The obvious consequence of this development is that CAD systems will be used more in the conceptual-embodiment design phase of product design, as shown in Figure 1. It is in these phases that designers will need the most help to stimulate their creativities so that they can design cheap products that meet high quality specifications. [Wa02] has reviewed commercial and academical CAD systems that are available or being developed to support collaborative conceptual design and concluded that there are many opportunities for this field of research, as support for these phases of design from a CAD perspective is still relatively small.

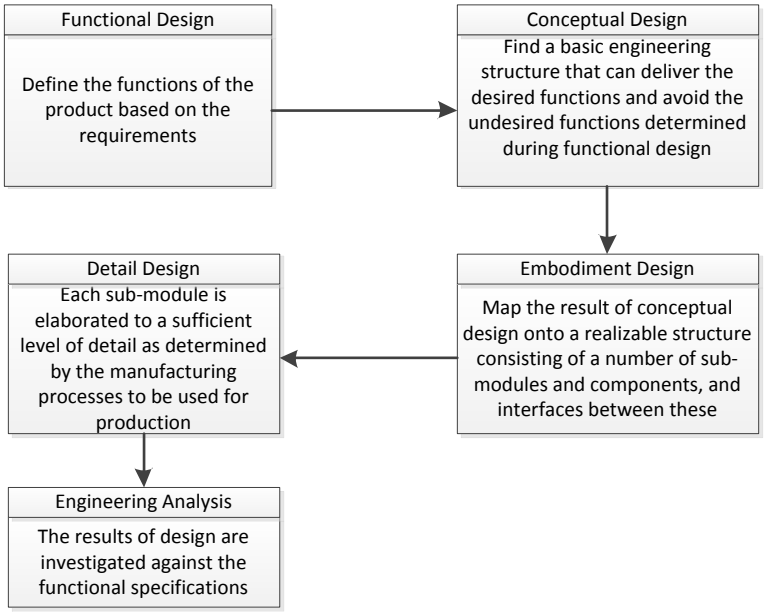


Figure 1 : Phases of Product Design, taken from [Ch12]

In accordance with the 4 C's of future CAD, this research will then try to develop a framework for a collaborative CAD system that will be used in the conceptual embodiment design phase of product design. Specifically, it will try to develop a CAD system that will enable multiple designers to cooperate on designing a product by allowing designers to exchange design ideas efficiently. The following sections of this paper will contain the results of literature research done during this research, the overview of the CAD system that will be developed, and the future development of this research.

2 Literature Research

In accordance with the topics of the research, further literature research has been done on the topics of general collaborative CAD systems, CAD systems for conceptual design

phase and creative CAD systems. The following section will detail some of the interesting findings from the reviewed literature.

2.1 Collaborative CAD Systems

There are many researches that have been done with the topic of collaborative CAD systems using many different technologies. Collaborative activites can be broadly organized into 2 different activities, Coordination and Cooperation activites [Ec00]. Coordination activites are activities where different designers work on parts of the same product and at the end those parts will be assembled into a complete product. On the other hand, Cooperation activities are activities done by different designers on a one product, as a whole complete product or a part of a product, at the same time.

Some researches, such as by [Wa09], [Mc12], [YM12], and [KK12] focuses on constraints settings by different designers. These constraints will then be propagated to all of the designers involved and every contraint violation will be broadcasted to each concerning designers. These designers must then negotiate to find a satisfying solution to those conflicts. These types of researches can be implicitly categorized as coordinative activities.

Other researches, such as done by [GLG10] and [LKB10] , try to solve the persistent naming problem, a significant problem in the collaborative CAD area that emerges from the interacting modeling operations of different designers. Traditionally, this problem is solved by file locking mechanisms, although these solutions increase the waiting time of designers during the design process. These types of researches usually work with the same file, so they can also be categorized into coordination activities.

Still others, such as [Bo10], [GM11], [Ch12], and [Ka13] focuses on how to share the modeled product with different individuals using various technologies, such as VR devices or smart devices. The goal is to facilitate reviewing and sharing among collaborators so as to increase the efficiency of collaborative design.

2.2 Conceptual CAD Systems

[Sa09] has developed a CAD system for Servive/ Product Engineering, while [KT12] and [No12] are examples of conceptual CAD systems for mechanical / electronical engineering. [KT12] focuses the collaborative aspect of their research on constraint control through parameter settings and the use of skeletal model, while [No12] focuses on reviewing activities by different collaborators.

Other researches focus on assisting concept generation, such as the computer aided TRIZ system developed by [Be12] or the computer aided collaborative shared design thinking platform developed by [DJL12]. In general it can be concluded that researches in this area are still relatively small, as concluded 10 years ago by [Wa02].

2.3 Creative CAD Systems

Creative CAD systems are CAD systems that are able to assist designers by giving creative alternative solutions for the problem at hand. Usually this is done by providing a creative database containing solutions, such as in [Li11], [CLX12], and [Go12] . When paired with a functional decomposition techniques, these database can then be matched to find creative solutions that may not be evident from the beginning.

In trying to create these kind of systems, researches, such as by [NZ12], [Mu12], [HCD08], are also being done to clarify what is meant by "creativity" and how can interacting with design tools affect a designer's creativity. These researches try to bridge the gap between cognitive creativity theory and the practical tools that designers use in the creative process itself.

3 Cooperative CAD System for Conceptual-Embodiment Design

Traditionally, collaborative CAD systems are used for coordination activities, so that multiple designers can be used to shorten the product development time. However, as mentioned in the earlier sections, the trend for the future is to design creative CAD systems. When extended as a collaborative system, it is clear that the cretive collaborative CAD system is more suited to be used in coordination activities, where several designers must work together to design new products that will conform to a set of specifications. This will enable designers to exchange and discuss creative ideas with each other, possibly leading to a better solution than from a single designer. Again, as mentioned before, these systems would be more suitable to be used in the Conceptual Design and Embodiment Design phase.

The proposed solution is by designing a CAD system that works in a hybrid top-down and bottom-up design approach. The designers using the system must work in phases, which are the functional design phase, the individual design phase, and the convergence phase. In the functional design phase, designers must model the product specifications into a functional decomposition of a complete product and the designers must agree on a particular part of a product that they want to design. After that, each designers can work on their individual workspace to find their preferred solution. On the convergence phase, remote meetings are held through the assistance of the CAD system to converge the probably dissimilar ideas into a convergent design. This process is then repeated until all of the product is finished. Figure 2 illustrates the workflow of the proposed solution.

4 Closing Remarks

In order for the system to be of use in a creative collaborative process, the system must also fulfill some requirements. This system should enable real time communication between designers, so that designers can easily understand each other's design intent and therefore decide the best solution. Furthermore it should also encourage the leveraging

of other designer's design . Lastly, as with all computer systems, it must be able to maintain data integrity and function with efficiency.

This research is currently still in its early phase, as such in the future more work will be done on the technical aspects so that the system as described here can be realized in its entirety.

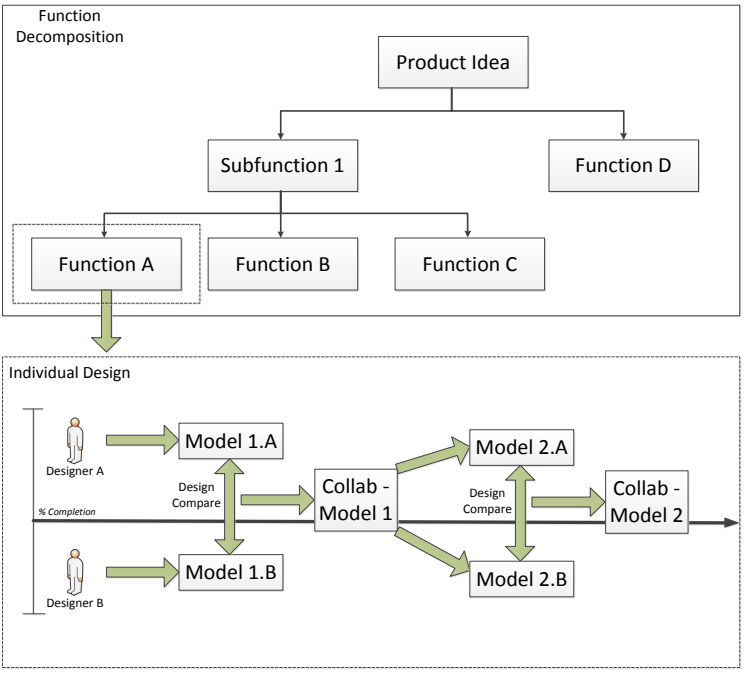


Figure 2 Workflow on the proposed CAD System

References

[Ap13] Aparajithan, S.; Lim, T.; James, R.; Raymond, S.; Zoe, K.; Ying, L.: The application of ubiquitous multimodal synchronous data capture in CAD. Computer Aided Design, 2013; http://dx.doi.org/10.1016/j.cad.2013.10.001

[Be12] Becattini, N.; Borgianni, Y.; Cascini, G.; Rotini, F.: Model and algorithm for computer-aided inventive problem analysis. Computer Aided Design, 2012; Vol.44, pp. 961-986

[Bo10] Bourdot, P.; Convard, T.; Picon, F.; Ammi, M.; Touraine, D.; Vezien, J.-M.: VR-CAD integration : Multimodal immersive interaction and advanced haptic paradigms for implicit edition of CAD models. Computer Aided Design, 2010; Vol. 42, pp. 445-461

[Ch12] Chen, X.; Gao, S.; Yang, Y.; Zhang, S.: Multi-level assembly model for top-down design of mechanical products. Computer Aided Design, 2012; Vol.44, pp. 1033-1048

[CLX12] Chen, Y.; Liu, Z.-L.; Xie, Y.-B.: A knowledge-based framework for creative conceptual design of multidisciplinary sytems. Computer Aided Design, 2012; Vol.44, pp.146-153

[DJL12] Du, J.; Jing, S.; Liu, J.: Creating shared design thinking process for collaborative design. Journal of network and Computer Applications, 2012; Vol.35, pp. 111-120

[Ec00] Eck, O.: Ein Kooperatives Transaktionskonzept fur CAD-Datenbanken. Institut fur Informatik der Universitat Stuttgart, 2000

[GLG10] Gao,L; Lu, T.; Gu, N.: CLAF : Solving intention violation of step-wise operations in CAD groupware. Advanced Engineering Informatics, 2010; Vol. 24, pp. 121-137

[GM11] Gujarathi, G.P.; Ma, Y.-S.: Parametric CAD/CAE integration using a common data model. Jounal of Manufacturing Systems, 2011; Vol. 30, pp. 118-132

[Go12] Goel, A. K.; Vattam, S.; Wiltgen, B.; Helms, M.: Cognitive, collaborative, conceptual and creative – Four characteristics of the next generation of knowledge-based CAD systems : A study in biologically inspired design. Computer Aided Design, 2012; Vol. 44, pp. 879-900.

[HCD08] Howard, T.J.; Culley, S.J.; Dekoninck, E.: Describing the creative design process by the integration of engineering design and cognitive psychology literature.Design Studies, 2008; Vol.29, pp.160-180

[Ka13] Kang, Y.: Kim, H.; Suzuki, H.; Han, S.: Editing 3D models on smart devices. Computer Aided Design, 2013; http://dx.doi.org/10.1016/j.cad.2013.08.011

[KK12] Kim, Y.S.; Kim, K.-Y.: DCR-based causal design knowledge evaluation method and system for future CAD applications. Computer Aided Design, 2012; Vol.44, 947-960

[KT12] Komoto, H.; Tomiyama, T.: A framework for computer-aided conceptual design and its application to system architecting of mechatronic products. Computer Aided Design, 2012; Vol.44, pp. 931-946

[Li11] Liu, X.Y.; Li, Y.; Pan, P.Y.; Li, W.Q.: Research on computer-aided creative design platform based on creativity model. Expert systems with applications, 2011; Vol.38, pp. 9973-9990

[LKB10] Lee, H.; Kim, J.; Banerjee, A.: Collaborative Intelligent CAD framework incorporating design history tracking algorithm. Computer Aided Design, 2010; Vol. 42, pp. 1125-1142

[Mc12] McIntosh, P.; Subic, A.; Lee, K.W.; Clifton, P.; Trivailo, P.; Leary, M.: An adaptable virtual engineering platform for distributed design based on open source game technology. Advances in Engineering Software, 2012; Vol. 43, pp. 71-86

[Mu12] Musta'amal, A.D.; Norman, E.; Jabor, M.K.; Buntat, Y.: Does CAD realy encourage creative behaviors among its users : A case study. Procedia - Social and Behavioral Sciences, 2012; Vol. 56, pp. 602-608

[No12] Noon, C.; Zhang, R.; Winer, E.; Oliver, J.; Gilmore, B.; Duncan, J.: A system for rapid creation and assessment of conceptual large vehicle designs using immersive virtual reality. Computers in Industry, 2012; Vol. 63, pp. 500-512

[NZ12] Nguyen, T.A.; Zeng, Y.: A physiological study of relationship between designer's mental effort and mental stress during conceptual design. Computer Aided Design, 2013; http://dx.doi.org/10.1016/j.cad.2013.10.002

[Sa09] Sakao, T.; Shimomura, Y.; Sundin, E.; Comstock, M.: Modeling design objects in CAD system for Service/Product Engineering. Computer Aided Design, 2009; Vol. 41, pp. 197-213

[Wa09] Wang, J.X.; Tang, M.X.; Song, L.N.; Jiang, S.Q.: Design and implementation of an agent-based collaborative product design system. Computers in Industry, 2009; Vol.60, pp. 520-535

[Wa02] Wang, L.; Shen, W.; Xie, H.; Neelamkavil, J.; Pardasani, A.: Collaborative conceptual design-state of the art and future trends. Computer Aided Design, 2002; Vol. 34, pp. 981-996

[YM12] Yin, C.-G.; Ma, Y.-S.: Parametric feature constraint modeling and mapping in product development. Advanced Engineering Informatics, 2012; Vol.26, pp. 539-552

On some theoretical issues of interaction with socialized and personalized cyber-physical systems

Imre Horváth[1], Zoltán Rusák[1], Yuemin Hou[2], Lihong Ji[2]

[1] Faculty of Industrial Design Engineering, Delft University of Technology
Landbergstraat 15, 2628CE Delft, the Netherlands

[2] Faculty of Mechanical Engineering, Tsinghua University, Qinqhuayuan,
Beijing 100084, China
Email-address of corresponding author: i.horvath@tudelft.nl

Abstract: The paradigm of cyber-physical systems (CPSs) is changing in our days. While a decade ago they were regarded as technical systems, they are now intellectualized as socially deeply embedded and behaviorally personalized systems. This has influence on the manifestations of CPSs and on the interaction with these systems. First, the paper casts light on the drivers of the development of social-cyber-physical systems. Then, it investigates the influence of socialization and personalization of cyber-physical systems on interaction. The last part of the paper looks into theoretical issues of interaction, such as coping with the interaction profile of cyber-physical systems, combining the intellectual domains of interaction, and interaction on various intentional levels.

1 Introduction: Seeing the big picture with the mind

Beginning at the end of the 1700s, the 'first industrial revolution' mechanized the industrial creation processes by using water and thermal energies. The 'second industrial revolution', which commenced at the beginning of the 1900s, exploited fossil and electric energies in production and transportation processes, and led to a widespread use of electromechanical systems in many application fields. Starting at the beginning of the 1950s, the 'third industrial revolution' introduced electronics in industrialized processes and fostered the proliferation of automation. The 'fourth industrial revolution' which was triggered by the emergence of digital computing and technologies at the midst of 1980s, introduced the use of computing in everyday working and living processes, and made the human societies information and service orientated. The change caused by the ever-growing large-scale application and sophistication of digital computing is often called 'intelligence revolution'.

Gradually replacing the work performed by human brain with work by machine brain and knowledge, the intelligence revolution has been made possible by the evolution of advanced computing and networking paradigms and technologies. As shown in Figure 1, there have been four significant shifts in the paradigm of digital computing in the last 60 years. The currently manifested paradigm, cyber-physical computing, is still differently interpreted by various research communities and has for this reason multiple different definitions. However, there seems to be an agreement on that cyber-physical systems

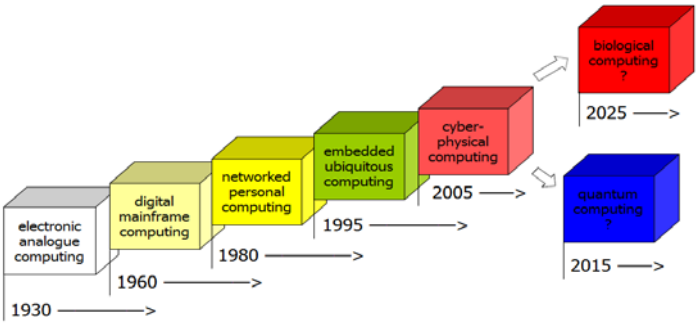


Figure 1: Shifting paradigms of computing

(CPSs) are confluences of knowledge and technologies of computing and informing, and knowledge and technologies of physical artefacts and engineered systems towards situated intelligent operation and servicing as actors in human and social contexts. There is a synergic relationship between the physical constituents and the cyber constituents of CPSs, and they are typically deeply embedded in their surrounding environments.

What the above overview indicates is that there is a continuous increase in the functional and structural complexities of engineered systems over time. The complexity of current large-scale CPSs (e.g., Internet) approaches the level of complexity of natural systems and thus they perform some sort of 'self-behavior' and evolution. Together with complexity, their heterogeneity is also increasing due to the number and interconnections of the involved material, energy and information flows and transformations [HE96]. The boundaries of complex systems are getting blurred, whereas the number and kinds of relationships that they have with other systems and their environments are non-linearly increasing. It has been argued by various complexity theories that complexity lends itself to emergence, dynamics, non-linearity and other forms of behavior that are not present in the constituents. In addition, CPSs are getting more socialized and personalized, therefore they raise new challenges for interoperation, interaction and interfacing, no matter if other systems, human individuals, or communities are concerned [BS14].

2 The influence of socialization and personalization of cyber-physical systems on interaction

The classical definitions of technical systems are not able to capture the essence of current complex engineered systems. The paradigm of CPSs goes beyond the paradigm of technical systems. The main constituents of CPSs are categorized as: (i) analogue hardware, (ii) digital hardware, (iii) control software, (iv) application software, (v) alphanumeric dataware, and (vi) semantic knowledgeware. Their major functions can be categorized as: (i) interfacing, (ii) sensing, (iii) searching, (iv) communication, (v) networking, (vi) computation, (vii) reasoning, (viii) conversion, (ix) control, (x) powering, and (xi) actuation. CPSs achieve a much higher integration and synergy of hardware, software and knowledgeware functionalities, technologies and components than any other type of systems ever before. Many CPSs are distributed and tightly

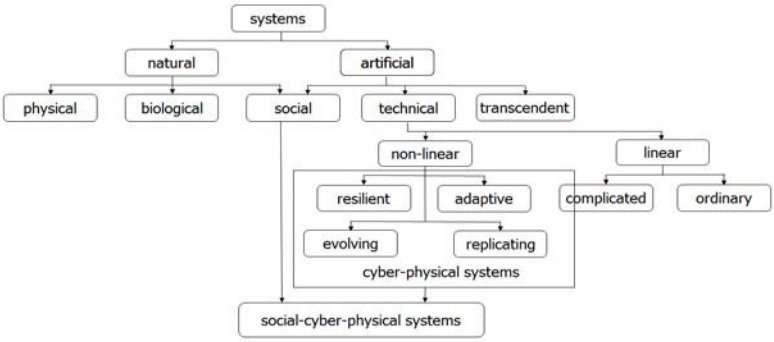


Figure 2: The position of SCPSs on the landscape of systems

interacting with the natural and engineered environments, but also penetrating into the cognitive domains of humans, and strongly influencing their social life and business.

It seems in these days that the recently emerged paradigm of CPSs is already going through a metamorphosis. Driven by the strive for system characteristics such as non-composite, openness, intelligence, autonomy, agency, resilience, adaptability, evolution and replication, the technology-driven system paradigm of CPSs is gradually transformed into a system paradigm that implies socially deeply embedded, cognitively aware, environment conscious, and partially autonomous systems. Therefore, we have proposed to consider future CPSs as social-cyber-physical systems (SCPSs) [HO13]. Figure 2 shows the position of SCPSs on the landscape of systems. Obviously, we still need to work out the knowledge that facilitates their proper intellectualization, design, implementation and application. We also need to develop those design and implementation principles that guarantee interoperation and interaction with these systems, which fulfil a kind of problem solving actor role, as opposed to the Internet of Things systems, which primarily play the role of infrastructural platforms.

When applied to living organisms, the term social refers to a set of characteristics of populations of humans or animals. It regards the interaction of organisms with other organisms and their collective co-existence, irrespective of whether they are aware of it or not, and irrespective of whether the interaction is voluntary or involuntary. In the context of cyber-physical systems, socialization means that they become able to establish

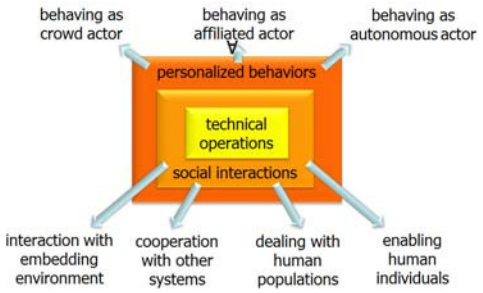
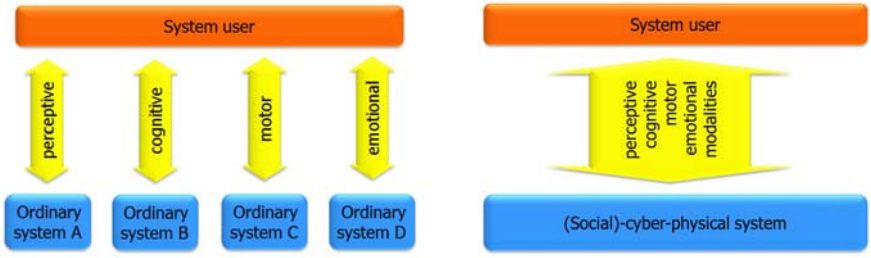


Figure 3: Dimensions of interactions and behaviors of current CPSs

an extensive range of non-technical relationships. As indicated in Figure 3, these relationships extend to: (i) the connections with natural and engineered environments, (ii) cooperation with other engineered systems, (iii) dealing with human populations in social contexts, and (iv) enabling humans by penetrating into their intellectual domains. Socialization also involves the increase of context awareness and the enhancement of social problem solving capabilities of SCPSs. Personalization means that the behavior of the system as a whole, or the constituents of it, will resemble the behavior of some natural agents [ZL96]. This natural agent-like operation can be achieved by designing system actors that are capable to behave as: (i) one of the actors in a crowd, (ii) affiliated individual actor, or (iii) fully autonomous actors. We are rapidly moving towards the situation when synthetic actors (agents) will be trusted near-peer teammates in SCPSs.

3 Some theoretical issues of interaction with socialized and personalized cyber-physical systems

The interaction profile of cyber-physical systems is rather complex. Not only human-system interactions, but also system-system interactions should be considered concurrently. The interaction can be looked at from perspectives such as the modality of interaction and the objective of interaction. It is well known that interaction with systems can in principle happen in four intellectual domains of humans, namely, in the motor, perceptive, cognitive and emotional domains. In case of interaction with conventional systems, we can typically identify a dominant modality (domain of interaction), which is accompanied by certain activities in the other domains (Figure 4.a). For instance, watching at a photo viewer (ordinary system A) is dominantly a perceptive interaction, but switching the device on or activating the software includes elements of both cognitive and motor interaction. Playing a computer game (ordinary system B) is based on a dominant cognitive interaction, but operating the game console needs motor interaction, viewing the game visuals needs perceptive interaction, and solving a Sudoku game needs cognitive interaction [GM05]. Steering and speeding a car (ordinary system C) is based on motor interaction that, however, is interwoven with perceptive and cognitive interactions. As a last example, being proud of the brand of a smartphone (ordinary system D) has effects on both the frequency and the way of using it (motor interaction), which in turn has effects on the perceptive and cognitive interaction.



a b

Figure 4: Interaction: (a) with ordinary systems, (b) social-cyber-physical systems

However, in the case of a social-cyber-physical system, interaction in the four domains is usually needed simultaneously and in a well-balanced manner. As an example, let us think of a cyber-physical stroke rehabilitation system. In this case, the physical rehabilitation involves motor interaction with various rehabilitation devices, such as assistive robotics-based exoskeletons. In combination with this, various cognitive exercises are also conducted in order to regenerate or enhance the mental capabilities of the patient based on dedicated cognitive interactions. In addition, various perceptive treatment exercises, including e.g. visual, audio, tactile interaction, are enforced in order to regain or enhance the lost perceptive capabilities of the patient. Simultaneously with these, the long term and short term engagement and motivation of the patient are triggered and maintained by emotion stimulating exercises and interaction. As indicated in Figure 4.b, these modalities blends into a hybrid interaction modality, the support of which is not easy and is still in its infancy.

Another issue is the level of interaction that is needed and can be realized between actors. Based on the information theory of Gitt, W., five levels of interaction can be identified [GI09]. As shown in Figure 5, low level interaction can be based on statistical or syntactic processing actions. On statistical level, physical signals and distribution patterns, while on syntactic level grammar and codes are the mediators of interaction. These make transportation and presentation of (physical) effects possible, but interpretation should be involved if comprehension of the meaning of the actions and communications plays a role in the interaction. It can be assumed that the actor is always pursuing a goal. While on semantic level it is achieving a shared meaning, on the pragmatic level of interaction the goal is the expected completion, and the concern is the way of executing the intended actions. Finally, on the level of apobetics, the intended purpose, the achievement of results, and the raised reflections are the concern of interaction. In other words, the latter means that if apobetic interaction is targeted, then the desired success and the objective of satisfaction should also be planned or designed.

We immediately face an inherent complexity if we combine the modalities/domains of interaction (motor, perceptive, cognitive or emotional) with the objectives of interaction (syntactic, semantic, pragmatic or apobetic), since each modality/domain can in principle be combined with each of the objectives shown on the right side of Figure 5. Current literature does not explain how these objectives can be considered and realized in the

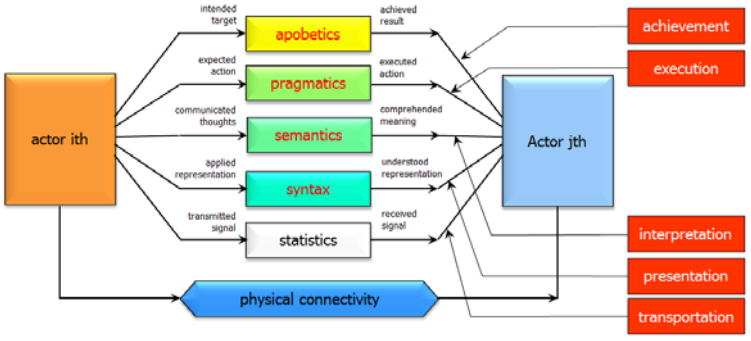


Figure 5: Levels of interaction between humans, agents and systems

case of the various modalities. Further research is needed in order to understand how the objectives can be realized in the case of a hybrid-modality-demanding interaction.

Interaction with behaviorally non-linear systems is another theoretical issue with significant implications on practical solutions. Resilient complex systems adjust and optimize their dominantly non-linear behavior to varying operational conditions without functional or structural adaptation. Adaptive complex systems change their structure, functionality and behavior at run-time as a response to changing objectives, environments and contexts that are unknown at design-time. In response to their perception of the environment and the systems themselves, evolving complex systems develop new configuration, functionality and behavior incrementally at run-time towards changed objectives or new opportunities based on learning. Replicating complex systems reproduce themselves in alternative manifestations based on system and environmental resources and according to non-evolutionary reproduction plans. The main challenge in the case of these systems arises from the fact that their long term operation cannot be exhaustively forecasted, and this causes indeterminism and uncertainties in designing the interactions. In addition, other practical issues can emerge due to the distributed and decentralized nature of SCPSs and from the fact that the zone of interaction between the system boundary and the environment boundary may be blurred and highly permeable.

Conclusions

In addition to addressing the role of complexity in the context of CPSs, the objective of this paper was to bring socialization and personalization of these systems in the focus, and to address some theoretical issues of interaction that need further attention in research and development. As theoretical issues: (i) the interaction profile, (ii) the multitude of interaction modalities, (iii) the intended objectives/levels of interaction, and (iv) the non-linear system behavioral were addressed with the goal of fostering further empirical research based on various SCPSs case studies and application contexts.

References

[HE96] Heylighen, F., 1996, The growth of structural and functional complexity during evolution, in The Evolution of Complexity, Kluwer Publishers, Dordrecht, 1-19.

[BS14] Broy, M. and Schmidt, A., 2014, Challenges in engineering cyber-physical systems, Computer, 47(2), 70-72.

[HO13] Horváth, I., 2014, What the design theory of social-cyber-physical systems must describe, explain and predict?, in An Anthology of Theories and Models of Design, ed. By Chakrabarti, A. and Blessing, L.T.M., Springer-Verlag, London, 99-120.

[GM05] Gersh, J.R., McKneely, J.A. and Remington, R.W., 2005, Cognitive engineering: Understanding human interaction with complex systems, Johns Hopkins APL Technical Digest, 26(4), 377-382.

[ZL96] Zachary, W., Le Mentec, J.C. and Ryder, J., 1996, Interface agents in complex systems, in Human Interaction with Complex Systems, Springer US, 35-52.

[GI09] Gitt, W., 2009, Scientific laws of information and their implications - Part 1, Journal of Creation, 23(2), 96-102.

First steps towards a mereo-operandi theory for a system feature-based architecting of cyber-physical systems

Shahab Pourtalebi, Imre Horváth, Eliab Z. Opiyo

Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
2628 CE Delft, the Netherlands
s.pourtalebihendehkhaleh@TUDelft.nl
{i.horvath;e.z.opiyo}@TUDelft.nl

Abstract: Cyber-physical systems (CPSs) are characterized by a high-level synergy among hardware, software and cyberware constituents. Though there has been intense research in various fields of complex engineering systems, still no advanced and comprehensive system design methodologies are available that would facilitate design and development of CPSs in various contexts and applications. This paper reports on the first steps and results of developing a comprehensive trans-disciplinary theory for conceptualization, modeling and realization of system architectures through a unified epistemological platform. The essence of the proposed theory is the fusion of mereotopological, modus operandi and CPS knowledge.

1 Design challenges of cyber-physical systems

As a manifestation of emerging highly complex systems, cyber-physical systems (CPSs) are rapidly proliferating. CPSs tightly integrate hardware, software and knowledgeware, are deeply embedded in the application environment, and typically, they are characterized by: (i) structural heterogeneity and complexity, (ii) operational synergy of components, (iii) nonlinear behavior, (iv) decentralized real-time operation capabilities, and (v) open architecture and ad-hoc functional connections between components. Although these characteristics provide completely new capabilities for CPSs, they also imply that the traditional methods and principles cannot realistically be utilized in design and development of CPSs [PHO14].

Our work concentrates on cyber-physical consumer durables (CPCDs). As a specific sub-category of CPSs, CPCDs are product-service combinations that reflect most of the functional and structural features of cyber-physical systems, while enabling a higher level of interaction with users and the embedding environments. The research reported here focuses on exploring the possibilities of designing customizable CPCDs and enabling this by some sort of methodological and/or computational means. Our previous investigation found that achieving the goals of mass customization (MC) in the case of

CPCDs is rather challenging [PHO13]. Additionally, it has been argued that due to their new system features, we need new design principles for MC of CPCDs.

The challenges of customization of CPCDs appear in combination with the challenges emerge due to the differences of approaches that are used for hardware, software and cyberware constituents, and due to the variety of aspects in which the heterogeneity should be addressed (i.e., control, sensing, actuating, exploration, informing, networking, adaptation, safety, security and reliability).

2 Why do we need a new methodology for architecting CPCDs?

The currently available approaches and methodologies of complex system development are typically mono-disciplinary, limited in terms of the scope of application, and do not guarantee compositionality of a CPCD as a whole. Nevertheless, components-based and platform-based architecting is becoming the conceptual and procedural standard for conceptualization, design and development of CPSs. The advantage of this approach is that it can be applied equally well to hardware, software and cyberware constituents. Considering these facts our intention is to develop a trans-disciplinary, widely applicable and system synergy-oriented approach for MC of CPCDs.

Our assumption has been that based on a common epistemological platform (i.e., underpinning theory) we may be able to bring the three domains in overlap and to develop a methodology that does not differentiate them. Obviously, the sought-after underpinning theory can only be based on new concepts. Having recognized the importance of this, we decided to extend the well-known concept of form and application features to the level of the constituents of the complex heterogeneous system.

In the system architecture context, product features are features of sub-systems and components. While, system features are features of the whole system. Product features can be application features or form features (Fig 1). Every component of a physical product is manifested in a certain form. These features are morphological definition of a component which can be distinguished by their material, surface quality, and everything that is related to shape of a part. The application features, by contrast, define realizations and applications of a component, i.e., design features, concept features, and manufacturing features. These definitions can be extended to software products as well. In this sense, codes, algorithms and programming language can be resembled to form features, and software development process features and software design features can be considered as application features.

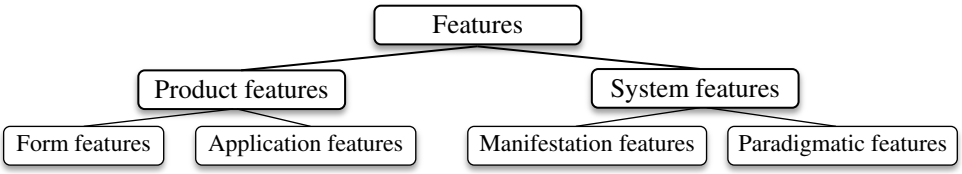


Figure 1: different kinds of features

System features are the result of components features as a composition. A system feature is a construct that include both paradigmatic features and manifestation features, which complement each other on different abstraction levels. We defined paradigmatic system feature (PSF) as a logically-based and physically-based abstraction of a system as whole that differentiates it from other comparable systems. These features are rooted in inherent characteristics of a system. A manifestation system feature (MSF) is related to a specific system in contrast to paradigmatic features that are about common features of a group of systems that share some similarities. They can be manipulated by designers and cannot violate overall system features.

Though system features are more related to the operational aspects of systems, product features are more associated by the structural architecture of components. Therefore, we need to address features composition by combining structural and operational architecture. In this sense, any existing artifact can be categorized in terms of these categories of features through operational and structural architecture.

3 Fundamentals of a new mereo-operandi theory

In order to develop an epistemological platform, it seemed to be necessary to blend various component theories that were able to capture the essential aspects of architecting CPCDs. In our research three originally distinct bodies of formal knowledge namely: (i) the theory of mereotolopology, (ii) the theory of engineering modus operandi, and (iii) the theory of CPSs have been combined in developing a unified theory.

3.1 Applied concept of mereotopology

Mereotopology was used to describe the part-whole relationships of operational domains of systems. Mereology literally means “science or theory of parts” [KYK08]. The name came from the formal theory of parts and associated concepts developed by Leśniewski [Ls82]. Topology, by contrast, implies ontological laws pertaining to the boundaries and interiors of wholes to relations of contact and connectedness, and so on [Sb96]. In the theory of mereotopology, mereology generally implies part-whole relation, while topology describes component connectedness. In our epistemological platform, mereotopology could play an important role in describing structural and operational architecture of CPSs. In this sense, relation of components-system and components together can be formalized and defined by mereotopological interpretations.

3.2 Applied concept of engineering modus operandi

By “engineering modus operandi” we refer to three things together: operation, functionality and behavior. If we consider operations as tangible outcomes of a system and the ways of delivering that output by sub-systems, we can describe its function at a higher level of generality and abstraction. In fact, operation explains how a function is accomplished. However, behavior defines the expected function of a system in its

context. Then, we need to consider operations of a system and its components in order to understand how the desired function is delivered, and the system behaves in its context.

In considering the operation of a system, we need to understand the operational architecture. "Operational architecture" is defined as the procedural architecture of a domain for realization of an operation. The architecture depends on the morphological characteristics of constituents actively involved in that operation. Additionally, the operational architecture and the morphology of constituents are supported by natural phenomena (e.g. gravity, friction, etc.). So, we need to consider three other levels in order to explain the operation of a system or a domain (Fig 2).

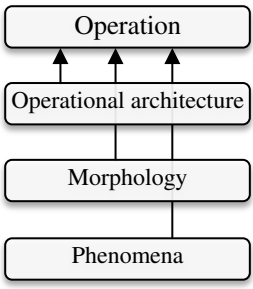


Figure 2: Elements of operation

3.4 Integrating the concepts

The theory of cyber-physical systems is used to take care of system compositionality and component composability issues. From the system compositionality view point, it can be noted that system output is the aim which could be realized by components. However, from the components composability viewpoint, emphasis is on the components which should work together in order to deliver the desired output of the system. Since both viewpoints can be followed in designing and developing CPCDs, compositionality and composability challenges should be considered for architecting of CPCDs. So, it is imperative to apply hierarchical decomposition of operational domains.

In order to describe manifestation system features of CPSs, a system can be represented based on several basic elements (as domains of operation) that define its operation. We named these basic elements as "operational building blocks" (OBBs). These OBBs should be (i) applicable for modeling of a large group of CPSs, (ii) able to be categorized operational-wise, (iii) based on simple rules for modeling, (iv) modular, (v) storable to form a repository, and (viii) open source (to be developed for different purpose). The operational and structural architecture of OBBs can be defined by mereotopological architecture when we are considering their relationship with their peers in the same level of granularity or their parental relationship, as well as engineering modus operandi when we want to explain procedural operation structure. We defined three main flows in domains of operation to consider operational processes, namely, material flow, energy flow and information flow. Material flow donates every kind of physical operation (physical input, output and manipulation of OBBs). Energy flow defines every kind of energy transmission, generation and consumption in OBBs. Information flow presents every kind of information processing, communication, sharing and storing in OBBs.

4 First results with validating the proposed mereo-operandi theory

A CPS as a system ($\sum$) consists of domains of operation (D_i), $\sum=\{ D_1, D_2,\ldots, D_n\}$. Domains of operation and their sub-domains can be represented by OBBs. Although hierarchical relations can be represented (through nesting) by OBBs, the degree of

granularity depends on the level that is required for modelling of a system. Features of an OBB can be determined by defining D={I,M,O} where I is the input of a specified domain, O is the output of that domain, and M is the manipulation which is done by the domain. In this sense, M(I)=O.

There are three kinds of input/output for representing material flow, information flow and energy flow. Different components of a system can be inferred by their inputs and outputs and/or combination of them. In light of this interpretation, for modeling a CPS, software, hardware or cyberware components can be modeled in a uniform manner, and based on the kind of flows they receive and send as input (I) and output (O).

A system and its components can be decomposed into OBBs. Then, we can introduce features of a system and its components by defining their OBBs' input, output and manipulation. Input and output of an OBB can be determined by its specifications (e.g. force (F), voltage (V), communication protocol (P), etc.). I={ F_i, V_i, P_i,…} , O={ F_o, V_o, P_o,…}. As well as manipulation of an OBB which could be defined by its specifications such as processing power (PP), produced noise (PN), energy consumption (EC), customization capability (CC), strength (S), durability (D), failure probability (FB), and any other aspects from designing, developing, customization, maintenance, and other perspectives. M={PP_m, PN_m, EC_m, CC_m, S_m, D_m, FB_m, …}

From composability viewpoint, manifestation features of a system can be defined by composing its OBBs' features. From compositionality viewpoint, manifestation features of a system can be decomposed for defining features of its OBBs. So, the proposed theory can satisfy both perspectives of design, development and architecting of CPSs. By modeling a system through OBBs, (i) specifications of each OBB in terms of input, output and manipulation can be formulated. These formulations could explain form feature and application feature mentioned in Section 2. (ii) Various operations of the system can be simulated based on different use scenarios. So, manifestation features of the system can be identified. (iii) Moreover, numerous open source algorithms can be developed and utilized in order to find opportunities for customizing the system, discovering probability of system failure, identifying bottlenecks and week points of the system, determining kind and level of user interaction and so on.

It is important to note that three libraries are needed to support our modeling approach. The first one is the product feature library. This library includes codes which define specifications of OBBs. It can be generated by producers of the components or other developers (i.e. translating product specification into codes). The second library is the manifestation system features library that could be provided by system designers or system architects. It shows various operation scenarios and considers system behavior in different context. The third library is the methodology and tool development library. This library contains design principles, development methods, maintenance principles, user-interaction regulations, etc., which could be developed by tool developers or knowledge engineers in form of some algorithms.

The major issue in our research is to show embedded customization capability in a system. It simply means having an OBB in our system with customizable M. This

customizable M means using different kinds of manipulations which can result in multiple operations: M={f_1, f_2, f_3, …} and $f_1(I)=O_1$, $f_2(I)=O_2$, $f_3(I)=O_3$, etc., and it is worth mentioning that different indexes of O might mean different channel of output, different values of same output or even using different physical output ports. Various manipulations of embedded customization might be stored in an open source library.

5 Conclusion

The main objective of our research was to develop a new theory, and methodology and/or tools for modeling CPSs. To achieve this objective, we needed to create an integrated epistemological foundation and a framework that support the development of a design methodology and a toolbox of enablers.

We have argued that, in addition to its mereotopological specification, MSFs are also characterized in terms of composability and compositionality, as well as of their operation. The physical operation is described in terms of the underpinning physical phenomena and principles, the morphology of the related domain, and the procedural structure of the operation. The OBBs can be used in modeling manifestation features, which carry mereo-operandi information, has specific meaning in particular operations, implementations and application contexts, and can be synthesized in alternative ways.

The developed theory, among others, can: (i) formally and symbolically present MSFs, (ii) represent hardware, software and cyberware constituents of CPSs in a uniform manner, (iii) allow specification and simulation of material, energy and information flow of CPSs, and support designing and implementation of CPSs. It is expected that the new theory will be able not only to integrate the currently disjointed hardware, software and cyberware knowledge, but will also bring the abstraction-based conceptualization and component/platform-based implementation of CPSs closer.

References

[PHO14] Pourtalebi, S.; Horváth, I.; Opiyo, E.: New features imply new principles? deriving design principles for mass customization of cyber-physical consumer durables, Proceedings of TMCE 2014, May 19-23, 2014, PP. 95-108, Budapest, Hungary.

[PHO13] Pourtalebi, S.; Horváth, I.; Opiyo, E.: Multi-aspect study of mass customization in the context of cyber-physical consumer durables, Proceedings of the ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, Portland, Oregon, USA.

[Hi98] Horváth, I.: Shifting paradigms of computer aided design, Delft University Press, Delft, The Netherlands, pp 1–36.

[KYK08] Kim, K.; Yang, H.; Kim, D.: Mereotopological assembly joint information representation for collaborative product design, Robotics and Computer-Integrated Manufacturing, Volume 24, Issue 6, December 2008, Pages 744-754.

[Ls82] Leśniewski, S.: On the foundations of mathematics, Topoi, 2 (1982), pp. 7–52.

[Sb96] Smith, B.: Mereotopology: A theory of parts and boundaries, Data & Knowledge Engineering, Volume 20, Issue 3, November 1996, Pages 287-303.

Towards affordance based human-system interaction based on cyber-physical systems

Zoltán Rusák[1], Imre Horváth[1], Yuemin Hou[2], Ji Lihong[2]

[1] Faculty of Industrial Design Engineering, Delft University of Technology
Landbergstraat 15, 2628CE Delft, the Netherlands

[2] Department of Precision Instrument, Tsinghua University
Beijing, P.R. China
email-address of corresponding author: z.rusak@tudelft.nl

Abstract: Human-computer and human-system interaction has already received extensive attention from HCI research communities, yet there are a number of issues that has not been addressed so far. In this paper we propose a novel vision on future human-system interaction that is governed by the principle of affordances and facilitated by cyber physical systems. The paper reflects on current theories and practices with the goal to adapt existing results in order to apply them in living and working environments operated by cyber-physical systems. It also proposes a number of research challenges that would pave the road to realize this vision.

1 Introduction

With the expected proliferation of cyber physical systems, conventional definition of product service systems needs to be refined. While traditional product service systems offer a given set of functions to serve specific user needs, Cyber Physical Systems (CPS)s are not only capable to fulfil these traditional functions, but they are also able to adapt themselves to emerging service demands. Cyber-physical systems consist of computational (i.e., hardware and software) and physical components, seamlessly integrated and closely interacting to sense the changing state of the real world. These systems involve a high degree of complexity at numerous spatial and temporal scales and highly networked communications integrating computational and physical components. It is expected that in the future, networked, cooperating, human-interactive systems will optimize human operations through high levels of situation awareness and adaptability [Le08]. CPSs will learn as they operate to maximize performance and resiliency, creating safe, secure, and reliable systems that can function as autonomously as desired by human systems designers. While humans will interact more seamlessly with the CPS of the future utilizing their emergent system behaviour, the ethical issues surrounding the human-machine interaction will be resolved prior to determining whether the human role will be as the operators of the machines ("human-in-the-loop") or as the partners of the machines ("human-in-the-mesh").

Building and further elaborating on this vision, we propose a novel scheme for affordance based human system interaction in Figure 1. In our interpretation, affordance

of a dynamically changing environment is the capability or quality of resources of the environment (e.g. technologies, products, components, artefacts, digital media) and their purposeful composition by a cyber-physical system that is recognized by end users in order to carry out a task in a given use context. This definition is partially in line with Norman's definition [No99] in the sense that it takes into account that: (1) the perceived properties of environment resources may not actually exist, (2) suggestions or clues as to how to use the properties of a resources may or may not be available, (3) perception/ recognition of an affordance can be dependent on the experience, and knowledge or culture of the end users, and (4) resource affordance can make a task performance difficult or easy. On the other hand, it goes beyond the perceptual definition, since it assumes that the actors are not only humans but also the system, and actors are not only perceiving affordances of the resources, but they also create, discover and manipulate them.

According to the technology affordance model of Gaver [Ga00], perceptible-, hidden-, false affordances and correct rejections can be distinguished based on the existence of an affordance and the available perceptual information about it. In our context, we interpret perceptible affordances as properties of individual environment resources or

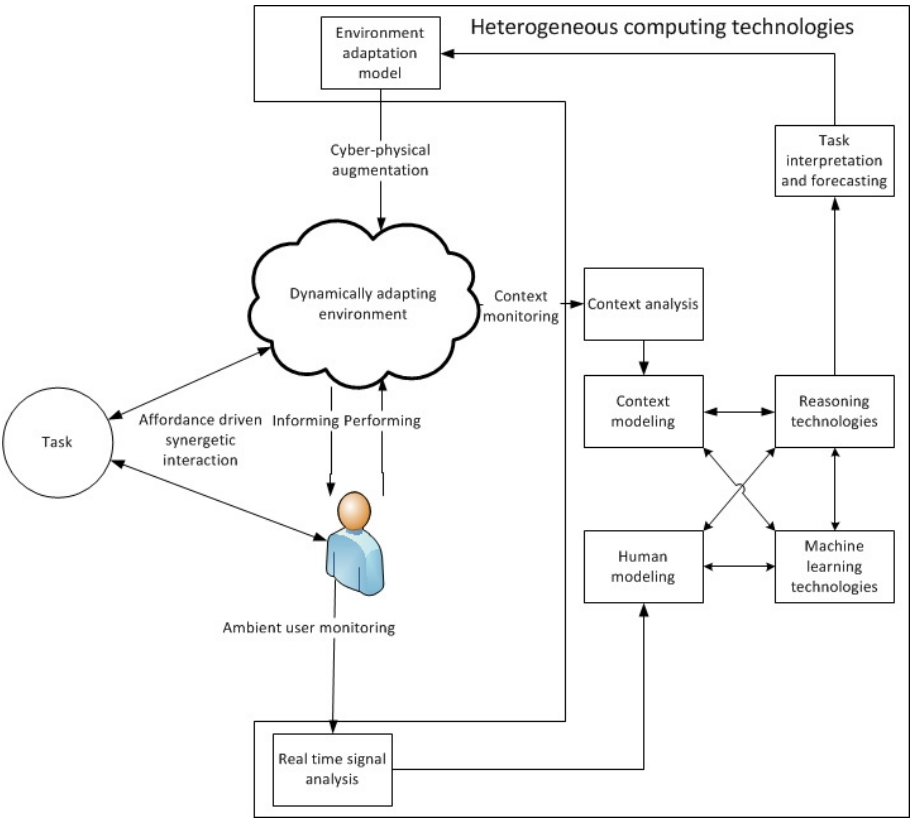


Figure 1: A model for Human System Interaction support by CPS principles

combination of them that is recognized by the actor to realize an action or quality of a product/service in a novel application context. We emphasize that the actor is not only the end user, but the cyber physical system itself that manipulates the environment resources and perceivable information to sufficiently inform actors.

In our vision, the cyber part of a CPS is represented by heterogeneous computing technologies, and the physical part of the system is envisioned as products and artefacts of natural and artificial environment that are together forming a service oriented cyber-physical system. In this scheme, it is envisioned that services are offered to the user through task oriented interaction, which is achieved based on available resources of the environment that are purposefully manipulated by heterogeneous computing technologies. Through this manipulation, the environment resources are presented to users considering users' capabilities and preferences, affordances of the environment and possible workflows for completing task(s) at hand. Task oriented purposeful manipulation of a dynamically adaptive environment, however, requires novel approaches capable to (i) monitor user characteristics and activities, (ii) inventorization and monitoring of artefact and products (iii) intelligent computing mechanisms to form models of the users, context and tasks, and (iv) synthesis and forecasting methods capable to model and simulate possible scenarios and expected outcomes. The paper presents a number of research challenges summarizing the essential issues of our concept.

2 Research Challenges

2.1 Ambient user monitoring and modelling

Ambient user monitoring and modelling enables cyber physical systems to characterize and profile users and to personalize services. Recently developed monitoring technologies are able to detect and interpret user activities, health and well-being, interest and preferences of users in given contexts, however, reliable monitoring of other characteristics among others emotions, motivation, engagement still require novel technological solutions [SS14]. When combined with advanced signal and pattern analysis technologies and machine learning techniques profiling of users and personalization of services can be achieved. To this end a number of research challenges should be addressed in the near future.

Research challenge 1: Systematic exploration of relationship among low level measurable indicators and high level user characteristics. For instance, while electromyography is primarily intended to measure muscle activities, the measured signal may also contain meaningful cues on the engagement level of a person. These secondary meanings of signals may be used to characterize user's engagement in motor rehabilitation.

Research challenge 2: Developing human modelling architecture that enables personalization and human characterization by compiling the distributed human information from separate sources.

2.2. Context awareness

Context information is important for applications that need to adapt to situations in which the user's context is rapidly changing. It is gathered from the environment surrounding the system to be utilized for enabling better adaptability of the applications. Dey and Abowd have introduced a general definition of context awareness of system [DA00]: *"A system is context-aware if it uses context to provide relevant information and/or service to the user, where relevancy depends on the user's task."* The problem is to find a solution for dynamic acquisition and representation of distributed context information and its efficient provisioning for the applications in order to enhance their adaptability and context-awareness.

Research challenge 3: Development of monitoring technologies for every day object usage and physical environment and interpreting their (potential) role in the actual context. Context awareness of a system highly depends on systems' capability to capture and interpret relevant information about resources and resource usage in the environment.

Research challenge 4: Development of new principles for virtualization, modelling and representation of a complex, heterogeneous physical world.

2.3 Reasoning

Machine and system intelligence is still limited to few links and rules in representing and interpreting information and knowledge over the physical space, physiological space, psychological space, socio space and mental space [HPS01]. As a result, these systems and machines are not capable to autonomously discover laws and solve emerging problems in these spaces. They are merely able to process pre-designed algorithms and data structures in the cyber space with a limited ability to go beyond the cyber space, to learn linking rules, to know the effect of linking, and to explain computing results according to physical, physiological, psychological and social laws [Zh11].

Research challenge 5: Interpretation of physical, physiological, psychological, social, and mental processes requires development of new theories, approaches and algorithms capable to establish new models of humans, tasks and context.

Research challenge 6: Overarching reasoning approach to establish relations between task, human and environment models.

2.3. Cyber-physical augmentation

Cyber-physical augmentation entails the concept that existing systems and environments may be dynamically extended with new functions by purposeful composition of elements and cyber augmentation of existing artefacts governed by user needs and task performance [HG13]. Synthesis of new composition, however, is not a straightforward process. It may be governed by examples and solutions from the past, analogies to similar context and workflows as well as by modelling and simulation of possible outcomes in the cyberspace. Cyber physical augmentation holds among others the following research challenges for my interaction scheme:

Research challenge 7: Exploration of principles for composition of dynamic service provisioning based on existing environment resources, personalized user needs and requirements for task completion.

Research challenge 8: Modelling and forecasting the influence of cyber-physical augmentation on usability of the environment.

2.5 Informing

Informing aims to offer the user means to perceive and to comprehend the necessary information about a dynamically changing environment. Dynamic adaptation of the environment to user's awareness and task execution process is of importance not only to avoid information overload, but also to provide essential clues according to the capabilities and knowledge of the user. In addition, it is also important that the physical and cyber cues offered to the user are timely synchronized and do not provide contradictory information.

Research challenge 9: Development of methods cyber-physical augmentation with the goal to inform users on semantic, pragmatic, and apobetic levels of communication [Gi06].

Research challenge 10: Synergistic presentation of environment affordances with consideration to physical, mental and emotional states of the user and task completion steps.

2.6 Performing

Performing covers activities aiming at completing tasks in a dynamically changing environment. Execution of these activities can be supported by implementing automatism, surrogating and complementing user actions as well as by assisting users in task execution. Highly adaptive and assistive systems are capable to adapt themselves to the level of assistance required from the user [MH14].

Research challenge 11: Development of principles and methods for dynamical adaptation of environment to user needs and task performance.

Research challenge 12: Development of artificial learning mechanism for real time assistance of users.

3 Conclusions

The paper presented a novel view on human-system interaction considering affordances of a dynamically changing environment, which are purposefully manipulated by cyber physical systems. The goal of this paper is to provide a position statement and to initiate a discussion in the scientific community of human computer interaction and cyber physical system developers. It identified a twelve research challenges that are needed to be addressed in order to realize affordance based human-system interaction. The authors strongly believe that these research challenges are addressing the most urgent issues related to human-system interaction as well as they are creating a roadmap for implementing a human-in-the-mesh concept.

4 References

[DA00] Dey, A. K., Abowd, G. D. Towards a Better Understanding of Context and Context-awareness. In: Proceedings of Computer-Human Interaction 2000 (CHI 2000), Workshop on The What, Who, Where, When and How of Context-Awareness, April 3, Hague, Netherlands. 12 p. 2000

[Ga00] Gaver, W. W.. Technology affordances. Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology. New Orleans, Louisiana, United States, ACM. 1991

[Gi06] Gitt, W., In the Beginning Was Information, First Master Book printing, Portland, Oregon, 2006

[HG13] Horváth, I., Gerritsen, B. H. M.. Outlining nine major design challenges of open, decentralized, adaptive cyber-physical systems ASME 2013 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference; 2013

[HPS01] Hippen, H.S., Pedreiw C.E., Soura, R.C. Neural networks for shortterm load forecasting: a review and evaluation, IEEE Transactions an Power Systems, 2001, 16 (1) pp. 44 -55,

[Le08] Lee, E.A. Cyber physical systems: Design challenges Proceedings - 11th IEEE Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing, ISORC 2008, pp. 363-369.

[MH14] W.S. Marras, P.A. Hancock, Putting mind and body back together: A human-systems approach to the integration of the physical and cognitive dimensions of task design and operations, Applied Ergonomics, Volume 45, Issue 1, January 2014, pp. 55-60

[No99] Norman D.A. Affordance, Conventions and Design. Interactions 6(3):38-43, ACM Press, 1999

[Ru14] Rusak, Z., Cimen, I., van der Helm, A., Horvath, I. Affordances for designing natural user interfaces for 3D modelling, Int. J. Computer Aided Engineering and Technology, accepted for publication, 2014

[SS14] Sánchez-Escribano, M.G. Sanz, R.. Emotions and the Engineering of Adaptiveness in Complex Systems, Procedia Computer Science, Volume 28, 2014, pp. 473-480,

[Zh11] Zhuge, H. Semantic linking through spaces for cyber-physical-socio intelligence: A methodology, Artificial Intelligence, Volume 175, Issues 5–6, April 2011, pp. 988-1019

Algorithms for Computational Mechanism Synthesis and Analysis

Hou Yuemin[1a], Michel van Tooren[2b], Ji Linhong[1c]

1 Institute of Design, Mechanical Department
Tsinghua University, Chengfu Rd, Beijing, 100084

2 Department of Mechanical Engineering,
University of South Carolina
1000 Catawba Street, Columbia, SC 29201 USA
a hym01@mails.tsinghua.edu.cn
b vantooren@cec.sc.edu
c jilh@tsinghua.edu.cn

Abstract: This paper proposes algorithms for mechanism synthesis and analysis, including Rotation, Center, Connection, Numeric Differentiation, Animation and Index. These algorithms are used to transfer behaviors into structures, and to transfer structures into kinematic and dynamic behaviors of the structures. The algorithms are valid for the synthesis and analysis of different mechnisms. These algorithms can be used iteratively to produce mechanisms that can fulfill simple and complex motions. The paper contributes simple algorithms for the design and analysis of different mechanisms.

1 Introduction

In design process, behaviors are bridges to link functions and structures. There are different definitions of the function, behavior and structure. Here, we use a simple definition. The structure is a physical entity, or object, represented as sketch and drawing in the design stage, e.g., a drawing of a washing machine or a clamp device. The function is the purpose for which the physical entity is used. The behavior is an ascription of the working state of the object, which can be described using symbols and measured quantitatively or qualitatively. Once the main function is decomposed into a group of functions, the next step is to determine the behaviors to fulfill the functions. What kinds of behaviors are selected depends on what working principles are used. When existing working structures are selected, the design and analysis procedures follow a routine steps: drawing a sketch of the structures, calculating the structural parameters and predicting the behaviors. Specific equations are required to calculate the structural parameters of specified types of mechanisms and specific rules are required to draw the structures, which do not cause lots of computing complexity but lack generic rules linking the behavior and the structure. This paper presents simple algorithms that can be used to generate structures and to predict behaviors of mechanisms. The paper analyzes

the design process of mechanisms and generalize simple algorithms in the second section, testifies the rules by using same rules for the syntheses and analyses of different types of mechanisms in the third section, and finally concludes the contribution of the paper.

2 Mechanism Synthesis: Case studies

In mechanical design, mechanisms are used to fulfill specific functions. A mechanism is a device designed to transform a set of input forces and movement into a desired set of output forces and movement. A mechanical device involves at least one mechanism. The synthesis of a mechanism means to find the dimensional parameters of the mechanisms. Existing methods of synthesis of mechanisms highly rely on the types of the mechanisms.

Current design methods of cam mechanisms include analytical methods and graphic methods. In the case of analytic methods, different equations are required for both design and analysis. In the case of graphic methods, the basic method is a kinematic inversion principle that adds an angular velocity $-\omega$ to the whole cam mechanism so that the cam stays stationary and that the follower and the frame rotate in the opposite direction than the cam does in reality, which method is termed "Inversion". In this case, the follower's motion is a combination of a relative turning motion around the cam center and a relative translating motion along a guide on the frame. The trace of the knife edge of the follower draws the profile of the cam. Different equations and design rules are used for different types of cam mechanisms. Also, different equations are used to descript the motions of the followers. However, a same rotation strategy is used for both analytical methods and graphic methods to establish equations in the analytical methods, and to generate the trace points of the knife edge during a motion period in the graphic methods. Different equations are required for computing the kinematic curves of displacements, velocity motions and acceleration motions. Current synthesis methods of linkage mechanisms also include analytical methods and graphic methods. The "Inversion" rotation strategy is also used. Besides, the center of the trace points of the motion path of the moved bars is produced to decide joint positions.

The equations used to calculate the dimensional parameters of a mechanism are different in the analytical methods, but the inversion method and the rotation strategy are used for both cam mechanisms and linkage mechanisms in the graphic methods. Graphic methods can produce both parameters and drawing, while analytical methods can only produce parameter values. However, the disadvantage of graphic methods is low dimensional precision. Thanks to computer science and technology, this problem can be solved by using computers.

Using computers to do drawing, the inversion methods are not necessary for the design of the cam profiles. The translating motion $s_i(t_i)$ of the knife edge of the follower can be directly transformed to the cam profile by rotating the points $s_i(t_i)$ with angles of $\delta_i(t_i)$. Furthermore, using software tools, the kinematic analysis can be generated by numerical differentiation, and the drawing can be executed by connecting points (drawing lines).

Analysis reports can be also produced easily by software. Hence, simple algorithms can be used to map behaviors onto structures, or vice versa.

3 Algorithms transforming behaviors to structures

Based on analysis of the design methods in the last section and our experience in engineering design, six simple algorithms are formulated. These algorithms are used iteratively to produce mechanisms.

(1) **Rotation for producing the profile of cams and linkages**

Rotation is a basic algorithm to form the structure. Any motion of a rigid body can be thought to rotate relative to another body and the center of the rotation is called the instant center. This is a principle of mechanics and it is the foundation of graphic methods in the synthesis of mechanisms. The rotation operation is used for the design of cam profiles and four bar mechanisms. This rule can be used to produce cam profiles without using the inversion method. By consequently rotating the discrete points $P_i(x,y)$ of the knife end of followers with step $\delta(i)$, cam profiles can be generated. Different kinds of cam profiles can be obtained by changing the working positions x_i and the rotational centre of the cams. Most important, this rule is also used to generate the profile of four bar mechanisms. The mathematical representation of the Rotation is

$$P(x,y)=\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \cdots & \cdots \\ x_n & y_n \end{bmatrix},\ T_R=\begin{bmatrix} \cos(\delta_i) & -\sin(\delta_i) \\ \sin(\delta_i) & \cos(\delta_i) \end{bmatrix},\ P'(x,y)=P(x,y)\cdot T_R \qquad (1)$$

(2) **Centre for joints of linkage mechanisms**

The algorithm Center is used to find the position of joints of linkage mechanisms as well as cam mechanisms. The relative motion of two bars is rotation and the joint is the center of the rotation. The Center algorithm is conducted after the Rotation operation. The relative displacement of two joints is obtained firstly through the Rotation operation and then the center of the rotation can be obtained. In the case of producing cam profiles, the center determines the type of a cam. A planar cam is produced when taking a point as the rotational center of the cam, while a cylindrical cam is produced when taking a line as the rotational center for same s(δ) curve. The mathematic representation of Center algorithm is $x^2+y^2=r^2$.

(3) Connection for the profile of mechanisms

The algorithm Connection means linking points to form curves or lines. Equations can produce parameter values but not the structure itself. Connection play a role of

producing forms. The profiles of the mechanisms are produced by linking points to form curves or lines.

(4) Numeric Differentiation for producing kinematic curves

The kinematic analysis is conducted based on "Numerical Differentiation" rather than different equations for different mechanisms. In this case, both cam mechanisms and planar linkage mechanisms can use same equations. The mathematic representation is

$$y' = \frac{y_{x+h} - y_{x-h}}{2h},\ y'' = \frac{y_{x+h} + y_{x-h} - 2y_x}{h^2} \tag{3}$$

where, y` and y`` are the first and second derivative, x is the value at which the derivative is calculated, h is a small number, y_{x+h} is function evaluated at x+h, and y_{x-h} is function evaluated at x – h.

(5) Index for generation of design and analysis report

Analysis reports are required in most cases. The structural parameters of the products, e.g. the length of bars and the profile of the cams in the context of mechanism synthesis, and kinematic and dynamic curves are listed on the reports. This is tedious work if to do it manually. Fortunately, quite a few of software provide the "report generator". To run "report" and establish a report file, user can establish an index to list required tables, figures and texts on a file with specified format (pdf, html, word, etc.) Then the report can be produced automatically.

(6) Animation for motion simulation

To simulate the motion of the structures, the Animation algorithm is required. Thanks to the computer technology, the animation of the mechanisms can be generated by playing the figures continuesly.

4 Examples: formation of structures for reciprocating motions

These six algorithms are tested by running examples on computers. Design examples are designing mechanisms to achieve reciprocating motions. The input data is working states of the followers or the bars, represented as a sequence of points. Other input data include the types of the mechanisms and basic parameters, such as base circle radius r_0 of cams. The whole process is automatically conducted.

Example 1: Mechanical synthesis. The result is shown in Figure 1. The rotation center is points in 1 (a), (b), (d) and (e), whihle the center is a line in (c). The input points in (a) , (c), (d) and (e) are on straight lines, while they are on an arc in (b). The outputs are slide cam mechanism, swing follower cam mechanism, cylindrical cam mechanism, crank rocker mechanism and swing follower cam mechanism in (a), (b), (c), (d) and (e)

respectivily. Rotation and Connection algorithms are used for all cases and Center algorithm is used for case (d) and (e).

Example 2: Analysis of swing follower cam mechanism. The Numerical Differentiation, Centre, Connection, Index and Animation algorithms are used. The results are shown in the Figure 2, including kinematic curves produced by Numerical Differentiation and Connection, static and dynamic analyses produced by equations and Connection, the animation of the mechanism produced by Animation, and other analysis and drawing. Figure 3 shows a report generated by the Index. The report includes tables listing parameters of the mechanism and figures generated in the design and analysis process.

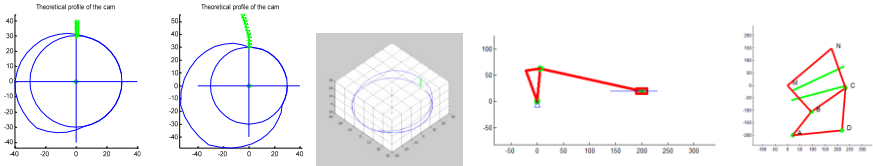


Figure 1: Profiles generated by the simple rules. a. planar cam with translating follower; b. planar cam with swing follower; c. cylindrical cam; d. crank slide mechanism; e. crank rocker mechanism

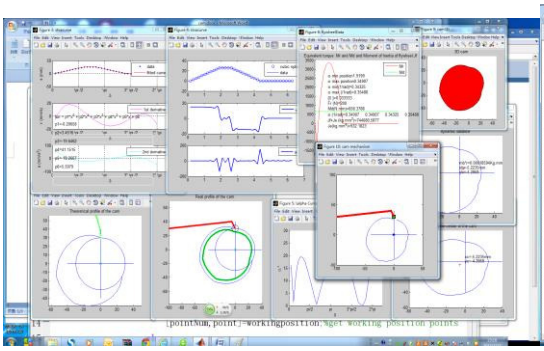


Figure 2: Output of analysis

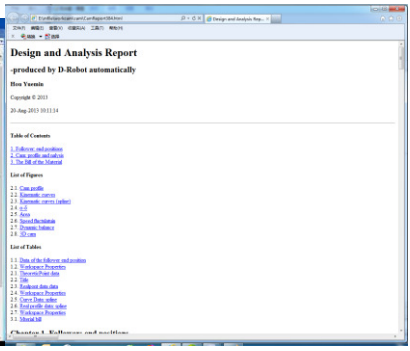


Figure 3: The report generated automatically

5 Related works

Many methods and algorithms have been used for mapping functions to structures. These methods include shape grammar [OM11], design grammar [Ru05], graph based representation and design rules [MHS13, ECSL13], first-order logic and Boolean operation [MHS13], Physics-Based Reasoning for grammar creation etc.. Some design methods combine many mathematic algorithms and AI and IT technology to automate the design process. These methods are used for creating structures step by step. Among these methods, KBE is a key technology to automate the design process of complex mechanical systems. The principle of this kind of design system is recording the design

process on computers. For example, the knowledge and design procedure of aircrafts, cars, small satellites and mold parts are recorded as software code so that repetitive design can be produced on computers [LV07]. The core at these methods is using software technology to build product templates including geometric description and the bill of material. The basic idea behind these methods is a predefined-template strategy, so it lacks of flexibility adapting to major structural modification and different design tasks. Current methods provide rules to assemble components into structures, while simple algorithms proposed here directly transform behaviours into structures step by step.

6 Conclusion and further works

Six simple algorithms are generated and presented. They are Rotation, Center, Connection, Differential, Animation and Index. They are used to transfer behaviors into structures, and to transfer structures into kinematic and dynamic behaviors of the structures. These algorithms can be used iteratively to produce mechanisms. Design experiments show that different mechanisms can be produced automatically from input behaviors with slight difference in working states based on these algorithms. The difference between the current mapping methods and the algorithms presented here is that they directly transfer behaviors into structures, while the current methods mainly focus on rules selecting components and assembling components into structures. The further work is using these algorithms to produce complex systems.

2 Acknowledgements

The presented work is jointly supported by National Natural Science Foundation of China (51175284), the Ministry of Science and Technology of the People´s Republic of China (2011ZX02403) .

References

[ECSL13] Egan, P. F.; Cagan, J.; Schunn, C.; LeDuc, P. R.: Using Multi-Agent Simulations and Structure–Behavior–Function Representations, Journal of Mechanical Design, 2013,135, 1-12.

[LV07] La Rocca, G.; van Tooren, M.J.L..L.: Enabling distributed multi-disciplinary design of complex products: a knowledge based engineering approach, Journal of Design Research, 2007; 5(3), 333-352.

[MHS13] Munzer, C.; Helms, B.; Shea, K.: Representations Into Boolean Satisfiability Problems for Computational Design Synthesis, Journal of Mechanical Design, 2013; 135, 1-13.

[OM11] Oster A.; McCormack J.: A Methodology for Creating Shape Rules During Product Design, Journal of Mechanical Design, (2011); 133(6), 1-12.

[Ru05] Rudolph, S.: Space station design rules, aerospace engineering, October, 2005; 43-45. Design, Journal of Mechanical Design, (2011); 133(6), 1-12.

Process Modeling of Dynamics-Driven Structural Design of High-Tech Semiconductor Fabrication Equipment

Luo Xin, Liang Chengyuan, Chen Xuedong

State Key Lab of Digital Manufacturing Equipment and Technology
Huazhong University of Science and Technology
1037 Luoyu Road
Wuhan, Hubei Province, China
mexinluo@mail.hust.edu.cn

Abstract: High-tech semiconductor fabrication equipment is dynamics-sensitive due to its ultra-precision requirements. Significant structural innovations have been made characterized by floating support and direct drive in order to eliminate effects of friction and vibration transfer. Nevertheless, complicate dynamic characteristics of multi-physics coupling, among electricity, magnetism, fluid, solid, heat, should be deliberately concerned in such kind of systems. Dynamics-driven design is essential to the success of development. However, lack of integrated design tool embarrasses the efficiency of design work and reuse of design knowledge. This paper proposes a process-driven dynamics design approach. Process control is involved to guide designers to accomplish design work. Dynamics model of components used in high-tech semiconductor fabrication equipment are abstracted and encapsulated as set of templates. In this paper, process meta-model based on XML are detailed, and a process test tool is developed to facilitate customization process modeling by simple graphical operations and attributes assignment.

1 Introduction

With the progress of semiconductor industry, the number of transistors that can be placed inexpensively into an integrated circuit chip (IC) doubles approximately every two years following the Moore's law. In front end process of IC production, high-tech IC fabrication equipments have achieved nano-precision level. It is well-recognized that the development of such kind of systems is strongly determined by dynamics-driven design. In the past two decades, significant structural innovations have made in IC equipments, characterized by floating support and direct drive, in order to eliminate effects of friction and vibration transfer [SB01][Sa02][KY03]. Nevertheless, these innovations also incurred multi-physics coupling problems among electrical, magnetic, fluid, solid and thermal variables [CC04], therefore, complicated, multi-physics coupled dynamics should be deliberately concerned during the whole design process of such kind of systems.

In this paper, a process-driven dynamics design approach is proposed, which is featured by template-based, process control and multi-disciplinary collaboration. In order to achieve process control, a process meta-model is detailed, which achieves the accumulation and reuse of process knowledge in the design process of IC equipment.

2 Process-driven dynamics design approach

The process-driven dynamics design approach is characterized by process control, template and multi-disciplinary collaboration, as illustrated in Figure 1. Designers can generate a design model of a component by instantiating templates of components, and then assemble the generated models of components into the multi-disciplinary model of the integrated system on a multi-disciplinary collaborative platform. The overall dynamic behaviors of the prototype of IC equipment affected by the multi-physics coupling can be analyzed, modified and optimized. All these operating processes are managed by the Process Controller. Designers can accomplish the whole design process of IC equipment step by step with the guidance of the Process Controller.

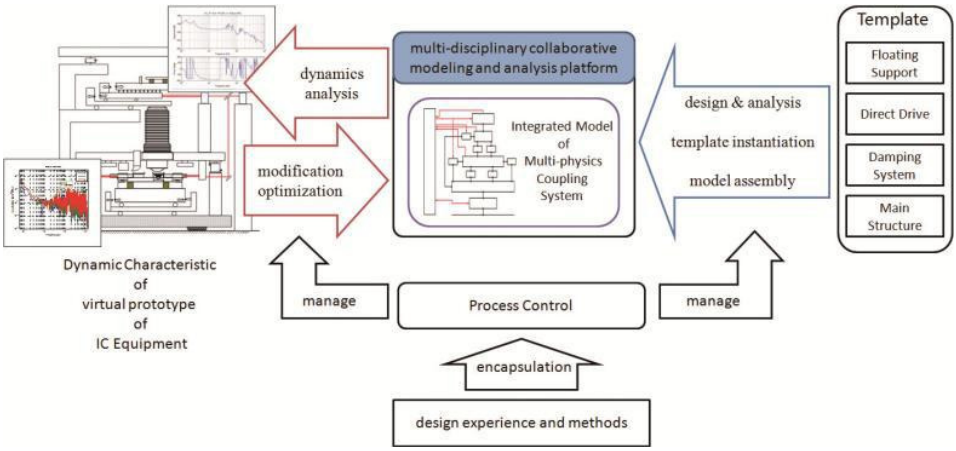


Figure 1: Process-driven dynamics design approach

In an IC equipment, components and subsystems, such as floating support, direct drive, damping system, present different structural characteristics and working principles. Corresponding templates are built for these components. A template of component or subsystem encapsulates all structural and functional characteristic variables, functions or equations of calculation and analysis [JL05]. A template also contains I/O interfaces, which exchange data with external modules and other templates, and call external softwares for analyzing of the multi-physics coupling effects in the component or subsystem. By assigning the characteristic variables and executing the intrinsic functions or external tools, the template is instantiated, and a parametric unified model of the specific subsystem or component is generated as the template instance. By linking all instances of components and subsystems according to structural connecting relationships, the integrated dynamic model is established on multi-disciplinary collaborative platform.

The Process Controller is able to accumulate and reuse of process knowledge in structural design of IC equipment. Design experience and methods which can be described as specific design or analysis process is modeled and stored to a process model file by Process Controller. And the specific process can be reproduced in the form of an executable process instance from the file to reuse the experience and methods. When a process instance is being executed, Process Controller manages the execution of tasks and data transference between tasks, controls the instantiation of templates in correct sequence, and leads the linking between template instances to construct the integrated system model. The Process Controller guides designers to accomplish the design, analysis and optimization processes with essential operating instructions step by step.

With assistance from the Process Controller, all template instances of components are assembled to the multi-disciplinary unified model of integrated system. Designers are enabled to modify and simulate the integrated model on the multi-disciplinary collaborative platform. Designers only need to modify values of characteristic variables in corresponding template instance or replace a new instance, for changing a component or subsystem in system. As soon as the simulation of integrated model is started, platform analyzes the linkage of template instances, checks and compiles the functions and equations in all instances to establish equations of whole system for solving dynamic characteristic of virtual prototype of IC equipment. Parametric models of all subsystems and components are calculated and updated via I/O interfaces of templates according to the linkage of instances, in order to achieve the multi-disciplinary collaborative analysis.

3 XML-based process model definition

Structural design process of IC equipment is featured by numerous node tasks, strict logic of executions of node tasks and complex relations of data transference.It is essential for process control to record the characteristic information of design process completely in an appropriate format, and maintain the process instance during execution. A process meta-model and corresponding maintaining rules are proposed, as illustrated in Figure 2.

Process meta-model consists of flow frame, nodes and arcs, with reference to workflow reference model [Ho06]. Flow frame generates the basis framework which describes the main features of specific process when process modeling, and manages the nodes and arcs contained in it. Nodes include action nodes and transition nodes which complete tasks and control executing routing in a process model. Action node calls template, third-party tool or external application to complete the assigned tasks when being executed. Transition node adjusts executing routing, and controls the executing sequence of nodes according to current executing state of process. Arcs which consist of control arcs and data arcs describe the relationship between the nodes they connect. Control arc represents the executing sequence between the two nodes linked to the arc. And data arc describes the data transference between nodes during executing of process. Specific process model is established by linking all nodes with arcs in the basic framework generated by flow frame according to the correct logical relationship.

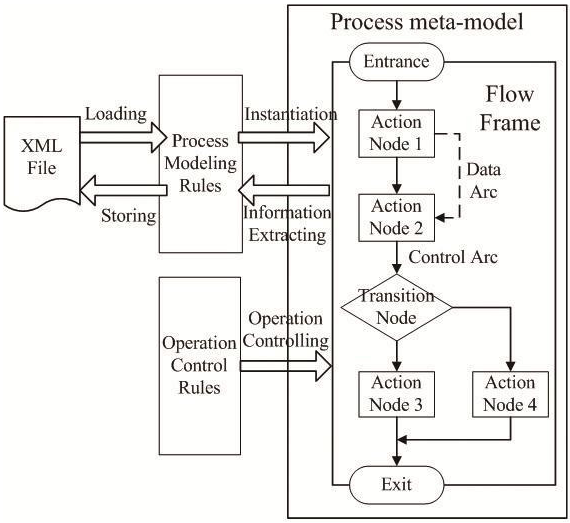


Figure 2: Process meta-model

In order to encapsulate process knowledge, it is essential of storing the structural characteristic information of specific process to files according to process meta-model. In process meta-model, the characteristic information is recorded in a classified and hierarchical data structure with several attributes in each object considering the difference of function and object type. With similar data structure to the tree structure in XML, the structural characteristic information of specific process can be saved to or loaded from XML file without inconvenient data conversion. And on the base that XML is communicated among modules in platform as the general data file, it is convenient that the process characteristic information is used by other modules on platform with XML file as medium. So it is feasible to save the structural characteristic information of specific process as XML file.

Process modeling rules manage the storage of the structural characteristic information of specific process to XML file, and control the generation of executable instance of specific process model from XML file. And operation control rules maintain the specific process instance when executing. during executing of process instance, nodes are executed one by one from the node declared as process entrance to the node declared as process exit, leaded by control arcs and controlled by transition nodes. And data is transferred from previous nodes to following nodes through data arcs. Correct executing permission is as essential as complete data input for executing a node. Referring to the concept of place/token in petri-net, similar receivers which ensure the completeness of data input and correct enable state of node are created [SW07]. Only when all receivers are filled with input data and executing permission, the node is permitted to be executed. When the execution of current active node is completed, the executing permission is shifted to the next node, and output data is also sent to the nodes connected with the current active node by data arcs.

A process test tool which is based on the process meta-model and maintaining rules, is developed to achieve customization and execution of process model. A simple analysis process of aerostatic bearing in IC equipment is tested. The corresponding process model is edited in the editing mode of process test tool by simple graphical operations and attributes assignment, as illustrated in Figure 3, and exported as a XML file in the prescribed format. In executing mode, Process test tool generates the executable instance by loading the XML file, as illustrated in Figure 4.

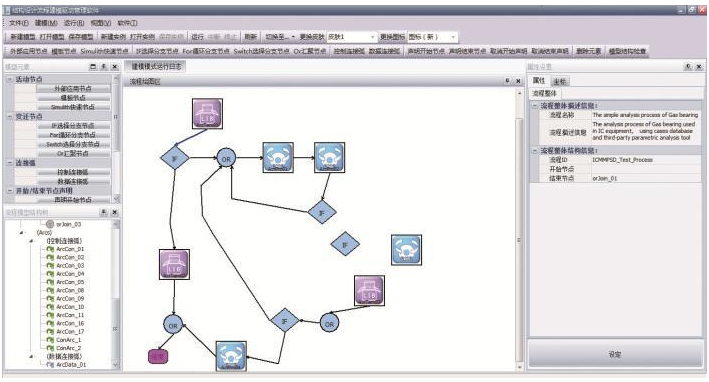


Figure 3: The test process model being edited

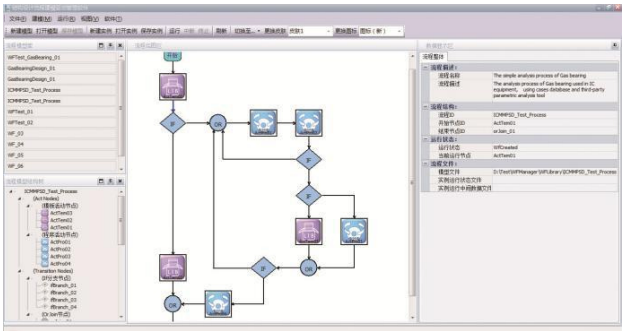


Figure 4: The executable process model instance

4 Conclusion

This paper proposes a process-driven dynamics design approach of high-tech IC fabrication equipment, with features of process control, template and multi-disciplinary collaboration, in order to facilitate design work of integrated system. By instantiating templates of components with assistance from process control, the unified dynamics model of integrated system is assembled and analyzed on the multi-disciplinary collaborative modeling and analysis platform quickly.

A XML-based process meta-model and corresponding control rules are also defined. With the process meta-model, a specific design or analysis process can be saved to or

loaded from a XML file, to achieve knowledge encapsulation and reuse of experience and methods in design of IC equipment. A process test tool is developed to facilitate customization process modeling by simple graphical operations and attributes assignment to adapt to the changing requirements of customers.

Currently, the work of this paper is at the beginning stage, with much further efforts needed for the establishment of a multi-disciplinary process-driven integrate design tool of IC equipment.

Acknowledgements

The work of this paper is partially supported by the National High Tech Research and Development Program of China (Grant No. 2009AA04Z148), and the National Science and Technology Major Project of China (Grant No.2011ZX02403-005).

References

[SB01] Stout, K. J.; Barrans, S. M.: Design of Aerostatic Bearings for Application to Nanometre Resolution Manufacturing Machine Systems, Tribology International, 2000; 33(12), 803-809.

[Sa02] Samir, M.: High Precision Linear Slide. Part I: Design and Construction, International Journal of Machine Tools and Manufacture, 2000; 40(7), 1039-1050.

[KY03] Kwang, S. J.; Yoon, S. B.: Precision Stage Using a Non-Contact Planar Actuator Based on Magnetic Suspension Technology, Mechatronics, 2003; 13(8-9 SPEC.), 981-999.

[CC04] Chen, X. D.; Chen, H.; Luo, X.; Ye, Y. X.; Hu, Y. T.: Air Vortices and Nano-Vibration of Aerostatic Bearings, Tribology Letters, 2011; 42(2), 179-183.

[JL05] Jia, W. C.; Luo, X.; Yuan, F.; Zhou, L.; Jiang, W.: Structural Feature-Based Model Templates Design for Ultra-Precision Positioning System: Proc. 2010 International Conference on Computer-Aided Manufacturing and Design (CMD2010), Hong Kong, China, 2010; S. 364-368.

[Ho06] Hollingsworth, D.: Workflow Management Coalition: The Workflow Reference Model, Workflow Management Coalition, Winchester, 1995.

[SW07] Salimfard, K.; Wright, M.: Petri Net-Based Modeling of Workflow Systems: An Overview, European Journal of Operational Research, 2001; 134, 664-676c.

Intelligent Control based on Location-Force Model

Jianfeng Sui, Linhong Ji

Department of Mechanical Engineering
Tsinghua University
Room 3403,Building 9003, No.1 Tsinghua Park, Haidian district, Beijing
100084
suijianfeng1986@sina.com
jilh@tsinghua.edu.cn

Abstract: This article introduces a multi-posture locomotor training device (MPLTD)with a closed-loop control scheme based on joint-angle feedback, which is able to overcome varies difficulties resulting from mechanical vibration and the weight of trainer to achieve higher accuracy trajectory. By introducing the force-field control scheme that used in the closed-loop control, the device can obtain the active-constrained mode including the passive one. The MPLTD is mainly composed of three systems: posture adjusting and weight support system, lower limb exoskeleton system, and control system, of which the lower limb exoskeleton system mainly includes the indifferent equilibrium mechanism with two degrees of freedom (DOF) and the driving torque is calculated by the Lagrangian function. In addition, a series of experiments-the weight support and the trajectory accuracy experiment demonstrate a good performance of mechanical structure and the closed-loop control.

1 Introduction of the MPLTD

The MPLTD consists of three systems: posture adjusting and weight support system, lower limb exoskeleton system, and control system. Lower leg exoskeleton length (from knee joint to pelma) is adjustable and the range is 390 ~570 mm, while upper leg exoskeleton length (from hip to knee) is also adjustable and the range was 325~505 mm [FGO12] ; The motion angle range of hip joint (angle between thigh longitudinal axis and neutral position) is 10° ~40° , while the knee joint range (angle between calf longitudinal axis and thigh longitudinal axis) is 0° ~70° [RB13], and the dorsiflexion and plantar flexion of ankle joint is in the normal range.

The lower limb exoskeleton system consists of indifferent equilibrium mechanism, linear motors, bandage, and pedal with gasbag. Inheriting the biological structure of human legs, the main structure of the indifferent equilibrium mechanism is designed based on serial-restraining and parallel-driving principle with two degrees of freedom, see Figure 2, whose structure size meets H/h=L/l. When the angle between the linear guideway and the bed frame is small, the direction of the combine strength of the mechanism's and

legs' weight is perpendicular to the linear guideway, so it almost do not need tractive force to make it move, which is the principle of indifferent equilibrium. This mechanism is very suitable for users under active-constrained mode, because users do not have to overcome the gravity of their own lower limb to achieve locomotion training. Linear actuator 2 drives hip joint and actuator 1 drives knee joint, interworking to simulate the gait for users' passive locomotion training.

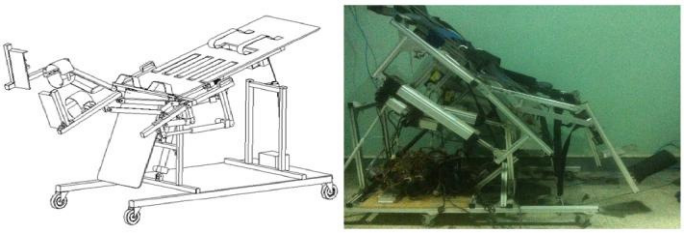


Figure 1: The multi-posture locomotor training device

Figure 2: The lower limb exoskeleton schematic diagram

2 Control System Design

In order to obtain the control algorithm using real-time feedback of joints and velocity control of the driven motors, a control system is developed, which includes industrial computer, controller, and sensors. Without feedback controllers design, the open-loop control mode of MPLTD based on the inverse kinematic modeling will have limited control accuracy due to external influences. Owing to the angle sensors, closed-loop control scheme for MPLTD based on feedback is prosed so that the motion control accuracy can be significantly improved. The passive mode and active-constrained mode are achieved by position, velocity and current control. In the passive mode, the control model of the MPLTD relates the joint angle velocities to the time derivative of gait simulation. The target of the control is for the joint angles measured by the encoders installed in the motor tails to approach their desired values; while in the active-constrained mode, torques of these two motors are controlled by current control, which is achieved according to the linear relationship between thrust and electric current within the operating range, and detailed interpretation will be mentioned later.

2.1 Location-Force Model

Active-constrained mode control algorithm is to determine the force $\vec{F}_{t+\Delta t}$ which should be applied to lower extremity (foot) by motors at the next time $t+\Delta t$, according the actual movement of the device end position $(\vec{P}_t, \dot{P}_t)$ at the instant time t, then the control function model is $\vec{F}_{t+\Delta t} = f(\vec{P}_t, \dot{P}_t)$. By changing the control function f, we can get all robotic exoskeletons end force for different positions at any time. Namely, that is, we can achieve different control effect and different force field in the plane. Active-constrained mode can be altered with a range of parameters; including constraint motion preset trajectory length, maximum auxiliary force size, etc., corresponding to $(x_t, y_t) \rightarrow F_{t+\Delta t}$ the different relationship between the position and the formation-force. The following figure (Figure 3) shows the location-force control parameters schematic, reflecting the output force of exoskeleton ends at different positions in a cross-section [Br72] [CBN11].

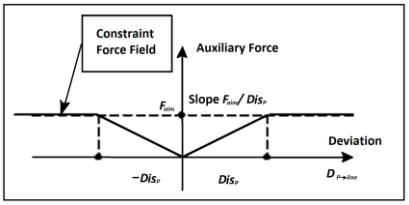


Figure 3: Force field controlling diagram

2.2 Intelligent Control—Active-constrained mode

In the active-constrained mode, if exoskeleton ends run within a set trajectory range, users can drive the lower extremity exoskeleton to do movement in the state of freedom; while exoskeleton ends deviate from the target trajectory, the constraint force field will pull limbs to the target trajectory from the wrong position by the most direct route. Due to users' weak control to muscles and joints, the variation of the constraint force should be soft and gentle. The degree of constraint force can be adjusted according to the actual physical conditions of users during training process, to achieve different training difficulty.

Taking the control algorithm of straight trajectory in the active-constrained mode as an example: AB is the desired straight trajectory, $A(x_1, y_1)$ is the starting point and $B(x_2, y_2)$ is the end point. The normal force of AB at the end point of the exoskeleton restrains the end point on the desired trajectory; and the tangential force provides auxiliary power or resistance from A to B, to achieve weight loss or gain.

The coordinate system is shown in Figure 5 $A(x_1, y_1) = (0,0)$. The auxiliary line l_L, l_A, l_B divide the motion space into five sub-regions Ⅰ, Ⅱ, Ⅲ, Ⅳ, Ⅴ, of which l_A, l_B are lines passing through points A, B respectively and perpendicular to the line AB, l_L is line segment whose distance is DP to line AB, and DP is the prescribed maximum error

beyond the trajectory line AB. $P(x_t, y_t)$ is the actual location of the end point of the exoskeleton. DPA , DPB , DPL are the distances from point P to lines l_L , l_A , l_B respectively, by which the sub-region where P is can be recognized. When P enters the sub-region Ⅲ, and still close to the point A ($|D_{PA}| \le D_P$), the force output satisfies the following relationship: $F_{t+\Delta t} = F_{aim} \times |D_{PA}| / D_{PL}$, $\theta_{t+\Delta t} = \pi + \arctan((y_t - y_1)/(x_t - _1))$, and the direction is from $P(x_t, y_t)$ to A. When P enters the sub-region Ⅲ, but far from the point A ($|D_{PA}| > D_P$),the force output satisfies the following relationship: $F_{t+\Delta t} = F_{aim}$, and the direction is from $P(x_t, y_t)$ to A. When P enters the sub-region Ⅳ, the principle is similar within sub-region Ⅲ . When P enters the sub-region Ⅰ ($D_{PL} \ge 0$, and $|D_{PL}| \ge D_P$), the force output satisfies the following relationship: $F_{t+\Delta t} = F_{aim}$, $\theta_{t+\Delta t} = a\tan 2(-x_2, y_2)$. When P enters the sub-region Ⅱ ($D_{PL} \le 0$, and $|D_{PL}| \ge D_P$), the force output satisfies the following relationship: $F_{t+\Delta t} = F_{aim}, \theta_{t+\Delta t} = \pi + a\tan 2(-x_2, y_2)$ .When P enters the sub-region Ⅴ , within the range of error band ($|D_{PL}| \le D_P$),the force output satisfies the following relationship: $F_{t+\Delta t} = F_{aim} \times |D_{PL}| / D_P$, $\theta_{t+\Delta t} = a\tan 2(-x_2, y_2)$

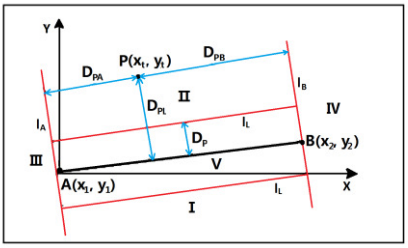


Figure 4: Control algorithm instruction

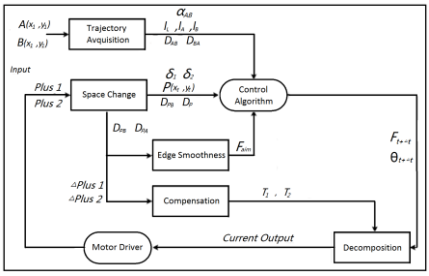


Figure 5: Active-constrained mode algorithm process structure

3 Experiment Results

To demonstrate the performance of the multi-posture locomotion training device, we have implemented three series of experiments: the repeatability and stability experiment of passive mode, and the accuracy experiment of active-constrained mode.

Ten healthy subjects are recruited to participate in these three series of experiments. All subjects have no history of serious diseases (6 men and 4 women, mean age: 24.5 years, range: 22-28). All subjects participate with informed consent and the approval of the local ethics committee. None of the subjects participate in similar experiments previously.

Figure 6: One subject's experimental process

3.1 The repeatability and stability experiment of passive mode.

The purpose of the experiment is to test the repeatability and stability of passive mode. In this experiment, step frequency is 0.2Hz, and the angular transducers fixed outside of the hip and knee joints of the exoskeleton collect angles changing with time as shown in Figure 7.

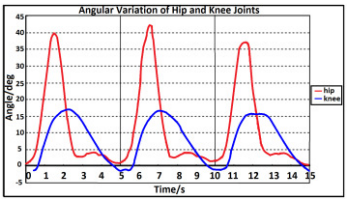


Figure 7: Angular variations of hip and knee joints

The motion parameters are accurate with a good reproducibility, because of the strong enough constraint force which is provided to lower lime by the exoskeleton in the passive mode. We calculate the maximum relative errors of hip and knee movement angles of the 10 subjects in the passive mode, and the results show that the relative error of the knee joint is slightly higher than the hip, but all are less than 6% as shown in Figure 8.

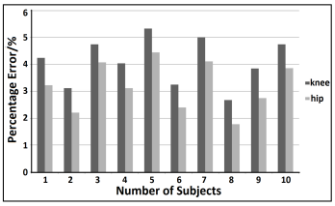


Figure 8: The results of the repeatability and stability experiment of passive mode.

3.2 The accuracy experiment of active-constrained mode.

Subjects do reciprocating linear motion without restraint, when the track is within the error range, the exoskeleton follows up; When the track is out of the error range, the

exoskeleton provides constraint force to reconnected lower limb within the error range. The experiment is to prove the accuracy of the straight trajectory which subjects participate in under the active-constrained mode.

Angles of hip and knee joints are collected in the same way as passive mode to calculate positions of the end of the exoskeleton by kinematics positive solution, and then the trajectories composed of these positions are drawn out in the sagittal plane in Figure 9. The results show all the trajectories are within the error range (block representing the error range in the Figure 9).

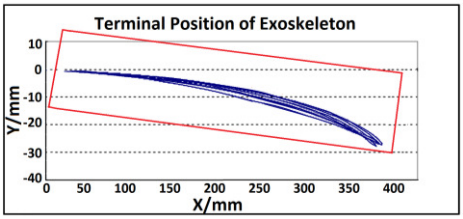


Figure 9: The results of the accuracy experiment of active-constrained mode.

4 Conclusions

This article provides the closed-loop control to accuracy control for the MPLTD. By introducing a force-field control, a new active-constrained mode is proposed for the indifferent equilibrium mechanism. The novel active-constrained mode is constructed by the matrix between trajectory and force field. A series of trajectory accuracy experiments in various angles demonstrate a satisfactory performance of the active-constrained mode. Considering the fatigue and aging of the indifferent equilibrium mechanism, we will focus on robust adaptive control for the MPLTD in the future work.

References

[FGO12] Fryar, D.; Gu, Q., Ogden, L.: Anthropometric Reference Data For Children And Adults: United States, 2007–2010. National Center for Health Statistics, Vital Health Stat, 2012; 1-40.

[RB13] Reese, B.; Bandy, D.: Joint Range Of Motion And Muscle Length Testing, Elsevier Health Sciences, 2013.

[Br72] Brunnstrom, S.: Movement Therapy in Hemiplegia: a Neurophysiologic Approach. J Gerontol, 1972; S.2904.

[CBN11] Candace, T.; Beth, P.; Nicholas, E.: Efficacy Of Rehabilitation Robotics For Walking Training In Neurological Disorders: A Review, J Rehabil Res Dev, 2011; 387-416.

A GPU Based Parallel Computing Mechanism Of an Etching Profile Evolution Model

Fan Wu, Yixu Song, Nan Wu, Hongjun Yang, Xiaomin Sun

State Key Laboratory on Intelligent Technology and Systems
Department of Computer Science and Technology
Tsinghua Universit
tChengfu Road, Tsinghua niversity
Beijing, 100084, China
lightning2-u.an@126.com

Abstract: Since the simulations for etching and deposition are increasingly time consuming, we explore the parallel computing mechanism of a typical profile evolution model on GPU platform in order to accelerate the process. It turns out the traditional parallelism will lead to either data unsafety or low efficiency, because of GPU's special architecture and a lack of safe protocols to protect a memory address from being simultaneously written by multi-threads. To overcome the dilemma, we propose an advanced parallel computing mechanism suitable for GPU platform by dividing each Thread's work into several parts. It is theoretically and experimentally proven that the multi-Threads writing the same address can be totally avoided. Experiment results show that the advanced parallel model is much faster than the traditional serial one.

1 Introduction

More and more efforts are made to explore the mechanism and principles of the plasma etching process during the manufacturing of microelectronic devices, and computer simulation is becoming an essential way to study, and even to predict this complicated process. Different models are used to simulate and to explore the process with some common species of gas and substrate. Since the Continuous Cellular Automata (CCA) was introduced[ZC00], many others extended and improved the model in different aspects by expanding the traditional 2D cellular model into a 3D model and introduced a time compensation algorithm [Ka99], combining complicated and specific physical and chemical principles into a 3D profile evolution model[GSV10], exploring more details about the electric field in the plasma etching process[MAC08]. To reduce the simulation time, we explore the potential mechanism of parallel computing for the traditional 2D profile evolution model on GPU (Graphic Processing Unit) by tracking movements of a swarm of particles simultaneously. However, the traditional way to translate a serial program into a parallel one fails on this model because of some special architectures and so called "warp" mechanism of GPU, which cannot provide a highly efficient way to protect a memory address being written by multi-Threads. So we propose an advanced

parallel computing mechanism that disintegrates the incidence-sputtering process into several subtasks to prevent multi-Threads from writing the same memory address and keep efficiency. The next Section briefly introduces the traditional profile evolution model; Section 3 discusses in detail about the problem with a common design of parallel computing using GPU, and then proposes an advanced mechanism; experiments and results of the advanced parallel model are reported in Section 4 with some analyses.

2 Traditional Profile Evolution Model

In this model, the substrate is assumed to be silicon just for convenience of latter introduction, and the particles are assumed to be the element inert to silicon to ensure only physical sputtering will occur so that more attention can be paid on the parallel computing mechanism. In order to represent the profile, the whole model is divided into lattice. Particles with energy and incidence angles from given distributions appear at the top-most of the lattice. Once a cell is hit by a particle, its value will decrease according to the energy and incidence angle of the particle. Once the value descends to 0, it will become a vacuity and some internal cells next to it will be exposed to other particles. A result of this traditional model shows that the middle bottom is higher than both side bottoms. That is because after being reflected from the side walls, particles are more likely to fall onto the side bottom of the silicon substrate than the center, resulting in the sputtering of more silicon atoms.

3 Parallel Mechanism for Profile Evolution and Detailed Discussion

A Graphic Processing Unit (GPU) is originally a specialized integrated circuit designed to rapidly build the images in a frame buffer intended to display on a screen in parallel. Now it is generalized to qualify some vastly repeating works and outperforms CPU. Owning independent Registers and Local Memory, every Thread can be seen as a little processor similar to CPU. The Threads from one Block can communicate and transfer data with each other through Shared Memory, to which a processor can access as fast as to Registers, but the Threads can only communicate with each other through Global Memory if they are from different Blocks. Each Thread runs a copy of the codes and all Threads run almost the same codes while they can run slightly different parts of the codes based on their unique orders. So GPU has great advantage in some highly repeated and little dependent works.

The first question to use parallel computing to accelerate a program is which parts of the program can separately run in parallel. According to the traditional serial model, an ordinary design is making different particles fall simultaneously, while there is always one particle in the serial model. As a consequence, it is possible that two particles simulated by two different Threads hit on the same cell nearly at the same time, which may cause data unsafety. In the serial model, particles fall one by one and the next one cannot fall until the effect of the previous one is finished. Conversely, in the multi-Threads design, two particles can affect a surface cell simultaneously, calculate the

results and write them to the same address regardless of other particles' effectiveness. Since there is no safe mechanism other than Atomic Operators for integers in GPU to protect the data, the eventual result simultaneously written by two Threads is uncertain and could be any one of the results from the two Threads. For instance, when a particle hit on a silicon cell, which should have become a vacuity cell after it has been sputtered by a previous particle but the sputtering process hasn't finished yet in that Thread, it could cause an invalid hit and sputtering or even some kind of unexpected fatal errors. In short, it is a typical Readers-Writers Problem. Even though it can be solved by using Spin-Lock or Monitor, the solutions cannot perform as well on GPU as on CPU because of GPU's special hardware architecture and controlling mechanism. In GPU, there is a concept of "warp", which is the most basic unit for instruction dispatching and controls 16 or 32 Threads' IPs (Instruction Pointer) to make sure these Threads are always synchronized. Whenever a Thread goes into a branch of the codes while others from the same warp don't, the others will wait for it to finish the branch and then run the same next instruction synchronously. It is this special mechanism that will cause deadlock in Spin-Lock or Monitor solution.

There is always only one Thread that can get the lock each time and if one Thread fails, it will keep trying and be blocked. Under GPU architecture, if two Threads from the same warp are getting the same lock, or if two particles simulated by the same warp are sputtering the same cell, only one of them can get authorized to access and write the effect to the cell while the other one has to wait for the lock to be released. On the other hand, even though one Thread can get the lock, it will still need to wait for the other one before it runs the critical codes since they belong to the same warp. As a consequence, when the one with the lock has no chance to release the lock and the other is waiting for the release of it, a deadlock appears. There is only one way to prevent this kind of deadlock: serializing the Threads from the same warp on the critical codes and preventing them from getting the same lock at the same time, which will inevitably lose a lot of efficiency. According to tests, the cost of the serializing Threads solution is unacceptably high, so an advanced design of parallel computing mechanism is required.

Whenever two Threads are about to write the same address at the same time, a lock mechanism must be added to protect the data, then that leads to Threads serialization and inevitably lose a lot of efficiency. So the dilemma can only be solved at the root, which is to say, to prevent any two Threads from writing the same address.

An applicable advanced mechanism is to separate the work of each Thread to several phases, where each phase ends whenever more than two Threads have the possibility of writing the same address. Before every phase ends, each Thread must store the data into a cache instead of writing it directly to the final address. Then at the beginning of the next phase each Thread will collect the data that is about to be written to a particular address from the cache, then it will calculate and write the final result to this address. For instance, in the advanced parallel computing mechanism, first phase of each Thread is to simulate the movement of a particle and to write the sputtering yield to cache after it hits on a silicon cell, and at the second phase each Thread will collect all the sputtering yields on a same cell and calculate the remains of the cell. In other words, each Thread simulates a particle in the first phase but simulates a cell in the second phase. In this

advanced mechanism, writing the same address by different Threads can be totally avoided by using cache, and only a few simple Atomic Operators can ensure that different Threads won't have access to the same address in cache. So the advanced mechanism for parallel computing can guarantee data safety. The process can be described as Figure 1. From the figure we can see the particles are coming swarm by swarm instead of one by one in traditional model.

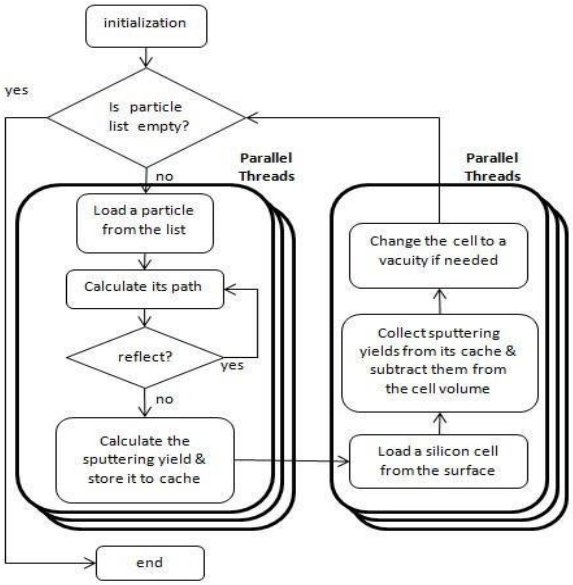


Figure 1 Flow chart of advanced parallel computing mechanism

4. Experiments and Results

We implement the traditional serial profile evolution model running on the Host with 8 Intel i7 CPU and 6GB RAM, and the advanced parallel computing model running on NVIDIA GeForce GTX 570 with 15 Blocks, each of which can run a maximum of 1024 Threads. In the first experiment, these two models run with the same particle list of 2M particles and the running times are recorded. The running time of advanced model involving 2~12 Blocks is shown in Figure 1. It is easy to understand that as the Block number increases, the number of particles that fall at the same time will increase so that the simulated process can be done within less time. And the running time is almost proportional to one divided by Block number. More specifically, no matter how many particles are in the last swarm of particles, the program should repeat one more time even though a lot of Threads are in leisure. So if PN and TN respectively indicate the total number of particles in the list and the total number of Threads in all Blocks, the running time should be proportional to [PN/TN]. When TN$\geqslant$PN, the running time will stop descending regardless of TN. And the fact is only when can the increase of the value of TN distinctively save the running time. The dashed line above all the data points in Figure 1 indicates the cost time of serial model running on CPU with the same particle

list. The result says that the advanced model can be approximately 6 times faster than the original model under the same input conditions.

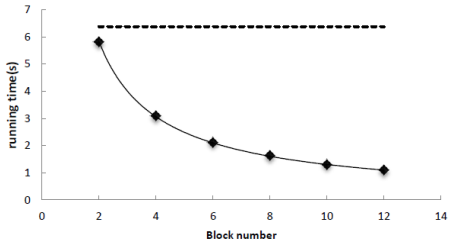


Figure 1 Running time of advanced model with increasing Block number, dashed line represents the running time of original model under the same conditions

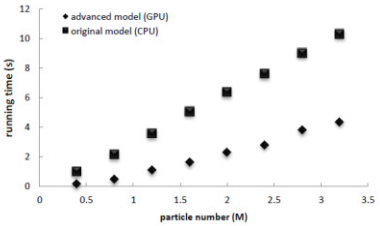


Figure 2 Increasing running time of two models with different particle numbers

According to the block diagrams of both original model and advanced model, the running times of these two models are expected to increase linearly as the particle number increases. Then it is confirmed by the test result shown in Figure 3. Since the time complexity of the original model is O(PN), there is not much prospective improvement in this aspect. But because of the increasing ratio being less than the serial model, the advanced model with only 6 Blocks can save at least half of the running time and the saved time will become more meaningful under a really large input particle list.

Another important result shown in Figure 3 indicates the difference between the final silicon profile in traditional serial model and that in advanced parallel model.Graph (a) in Figure 3 is the result from the serial model while (b) comes from the advanced parallel model with only one Block and under the same conditions with the serial model. In general, the etching depth of (b) is slightly less than that of (a). Before the cause can be analyzed, the sputtering yield compensation mechanism applied to the serial model must be introduced first. Because the silicon substrate is divided into lattice, and only one cell in the lattice can be sputtered each time. So it is unavoidable that some particles' sputtering yield is bigger than the remained volume of the cells they hit on, which will cause a little lost sputtering yield without the sputtering yield compensation mechanism. More specifically, after each sputtering that creates a yield exceeding the left volume of the cell, the remaining yield will be separated evenly to the adjacent silicon cell to make sure there is no lost sputtering yield. However, the mechanism applied in the traditional serial model cannot be inherited to the advanced parallel model because it naturally contains the possibility of writing the same address in different Threads, which is demonstrated that must be avoided in GPU programs. So the compensation has to be abandoned in the parallel model, which can cause some lost of sputtering yield that builds up the difference between (a) and (b). The more the total sputtering yield on a silicon cell within a swarm of particles is, the more yield will be lost when the remaining volume of the cell is nearly zero, and then the less deep the result will be. So we reckon that the advanced model, which runs on more Blocks or more Threads, may generate results that are less deep and probably less accurate.

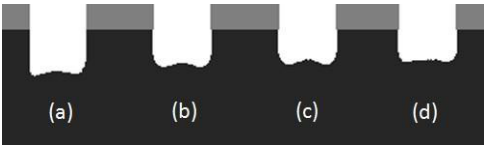


Figure 3 Results of traditional and advanced model (a) original model; (b) advanced model with 1 Block (c) advanced model with 8 Blocks; (d) advanced model with 12 Blocks

Graph (c) & (d) from Figure 3 provide an evidence supporting our theory while (c) is the result of advanced model running on 8 Blocks and (d) on 12 Blocks. This is the major reason why the Threads used by advanced model must be limited even though more Threads lead to more efficiency. However, it can be demonstrated that there is an instinct relationship between the lost yield and Threads number TN. So the compensation mechanism can somehow be interpreted into the function of sputtering yield $Y(E,\theta)$ to make it a GPU specialized function $Y(E,\theta,TN) = c(\sqrt{E} - \sqrt{E_{th}}) \cdot f(\theta) \cdot (1 + P(TN))$

where P(TN) represents a new compensation mechanism only depending on used Threads number TN in the GPU model. We are now exploring the expression of P(TN).

5 Conclusion

After thoroughly exploring the architecture and control mechanism of GPU and revising the whole process of the traditional serial model, we proposed an advanced parallel computing mechanism to overcome the dilemma caused by the "warp" mechanism of GPU. And the test results demonstrate that the advanced model outperforms the traditional one and has the potential to save more running time. It's quite important in today's increasingly complex models. In conclusion, GPU doesn't provide a safe protocol to protect the memory from being written by multi-Threads and the best solution is to totally prevent this kind of circumstance when designing the whole program. Exploring the sputtering yield compensation function P(TN) so that the results can be more consistent and convincing and complicating the model to test whether it can jeopardize the present parallel computing mechanism are interesting topics in the future works.

References

[ZC00] Zhu A.; Liu C.,: Micromachining Process Simulation Using a Continuous Cellular Automata Method, Journal of Microelectromechanical Sytems. 2000; 9(2), 252-261.

[Ka99] Karafyllidis I.: A three-dimensional photoresist etching simulator for TCAD, Modelling Simulation in Material Scicence and Engineering, 1999; 7, 157-168.

[GSV10] Guo W., Sawin H.H: Etching of Si O2 in C4 F8 /Ar plasmas. II. Simulation of surface roughening and local polymerization, Journal of Vacuum Science and Technology A: Vacuum, Surfaces and Films, 2010; 28(2): 259-270.

[MAC08] Madziwa-Nussinov T.G., Arnush D. and Chen F.F.: Ion orbits in plasma etching of semiconductors, Physics of Plasmas, 2008; 15, 013503 (2008); doi: 10.1063/1.2819681

Design of Processing Chamber Based on the Flow Uniformity

Huang Zun-di[1], Yang Tie-niu[1], Li Chang-ming[1], Huang Hui[2], Zhang Qi-li[1], Hou Shu-guang[1]

1. Department of Mechanical and Electrical Engineering, 2. Department of Information Engineering
WuYi University
Jiangmen, Guangdong Province, China
Jiangmen 529020
wyuhzd@aliyun.com
jmniuytn@126.com

Abstract: Orthogonal experimental design and flow field simulation are applied to the mechanical design reasonably in this paper. Based on the flow uniformity, design of processing chamber is completed. The results show that, Spalart-Allmaras one-equation turbulence model is effective for the shear flow simulation of wall in processing chamber with a low Reynolds number; design of showerhead is critical, which directly affects the distribution of flow field inside, as well as the etching effect; the location of exhaust port under the chamber scarcely affect the distribution of wind speed nearby the substrate, so gas exhaust unit is designed in the view of mechanical, in order to make the whole processing chamber achieve the dual effect of beauty and utility.

1 Introduction

With the rapid development of IC industry, chips become more and more integrated, which requires IC equipment manufacturers continue to introduce new product, invent new technologies, test new models, as much as possible to shorten product development cycles. Meanwhile, IC equipment and processes involve in the multidisciplinary field of knowledge that can be able to quickly and easily accept, analyze and use by researchers during the product development process. For most domestic enterprises and universities within the PVD, CVD, PECVD almost have the single structural design, and the uneven thickness during the coating, etching, and ionization deposition leads to the longer production cycle and some other problems as well. In this article, referenced to a lot of patent [PKK08] [NNK11] [SB210] [FCCPW09] [Du03] [SKD03]and paper about simulation [SSOO02] [Lu10] [CZDW08]of flow field inside about PVD、CVD and

PECVD equipment, design of processing chamber is completed, in order to realize the flow uniformity inside of the chamber.

2 Dimension parameter design

In this paper, the total height of cylindrical processing chamber is 200mm. The height change of processing chamber is the change of the distance between the substrate and the showerhead. Using the lifting mechanism control the movements of the substrate achieves the height variation between the substrate and the showerhead. The distance between the substrate and the showerhead sets 30mm. Consideration the distance variation between the chamber superior wall and the showerhead, the design adjusts to the distance between the first showerhead, the chamber superior wall and the second showerhead by bolt. The diameter change of processing chamber is achieved by increasing the lining in the chamber. According to the production design needs, the diameter of the substrate required is greater than 300mm, the distance between the outer edge of the substrate and the chamber inner wall is 50mm., so the chamber the minimum diameter sets 460mm; the maximum diameter sets 660mm.

3 Dimension parameter design

1) Turbulence model

The criteria of Reynolds number $\mathrm{Re} = \rho VL / \mu$ determines the laminar or turbulent. The density of air is 1.225kg/m3, viscosity coefficient is 1.7894e-5kg/ms, and the characteristic length is 8mm. So when V≤4.24m/s, the flow is laminar. In order to make the gas mixture more uniform, the flow should be turbulent. Setting the V=5 m/s, that is to say, the flux is 15000sccm. Spalart-Allmaras turbulence model is selected at last.

2) Mesh generation

The mesh model is generated respectively in different parts, which are inlet part, showerhead and inside of the processing chamber. In order to simulate the shear flow of the low-speed gas better, twenty boundary layers must be established round the wall which is discrete by a smaller dimension; the structure mesh is generated in all parts; local mesh refinement is necessary in showerhead part. At last, the total number of grid is 2 million, which meets the discrete requirement.

3) Boundary conditions

Inlet boundary conditions: diameter of gas inlet is 8mm, defined velocity-inlet with speed of 5m/s;

Outlet boundary conditions: the gas exhaust unit is defined the pressure-outlet in order to reflect the real flow field inside the processing chamber;

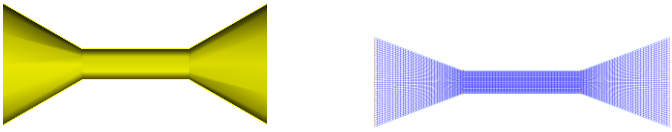
Wall boundary conditions: the shear friction between air and the wall of chamber when the air is flowing through it, so set the wall for no-slip wall boundary conditions.

4 Internal structure design

1) Showerhead design

Showerhead structure, which contains the structure of vent, the dimension, and the distribution, is a key factor. Design of showerhead is critical, which directly affects the distribution of flow field inside, as well as the etching effect.

There are two showerheads after the gas flow into the chamber. The first showerhead uses cylindrical holes. Holes on the second showerhead are shaped by tapered hole, cylindrical hole and the inverted tapered hole, of which the maximum diameter is 2mm, and the minimum diameter is 0.5mm. The specific structure is showed in Figure 1.



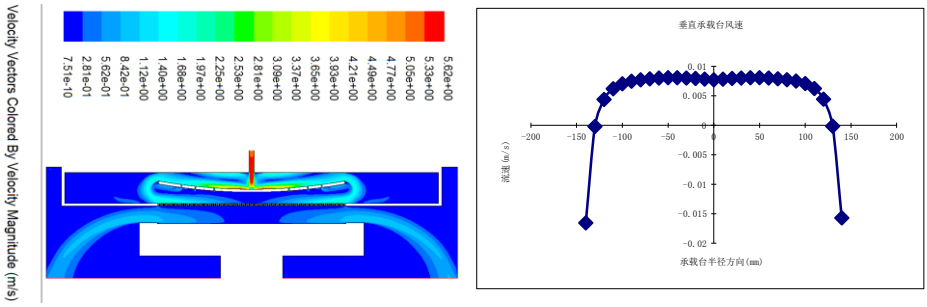
（a）Structure shape　　（b）mesh shape

Figure 1: Shape of hole on the second showerhead

After orthogonal experimental design and analysis of flow field simulation, the effective film range, the uniformity of flow near the substrate and the uniformity of flow near the showerhead of different objects are compared. Based on flow uniformity of design, the concave of the first showerhead faces the inlet. In the first showerhead, there are no holes in the center and three times holes distributed along the circumference. Radius of three times holes is 2.5mm, 1.5mm, 1.5mm respectively, as well as the distance between the holes center and the center of showerhead is 60mm, 90mm, 112.5mm respectively. Figure 2 shows velocity vector and the velocity distribution near substrate, simulated flow field inside of the processing chamber.

2) Gas exhaust unit design

The whole bottom is not completely the gas exhaust port. In order to keep the lifting mechanism which connects with the substrate work better and the air flows through the chamber quicker, the best setting place must be researched.

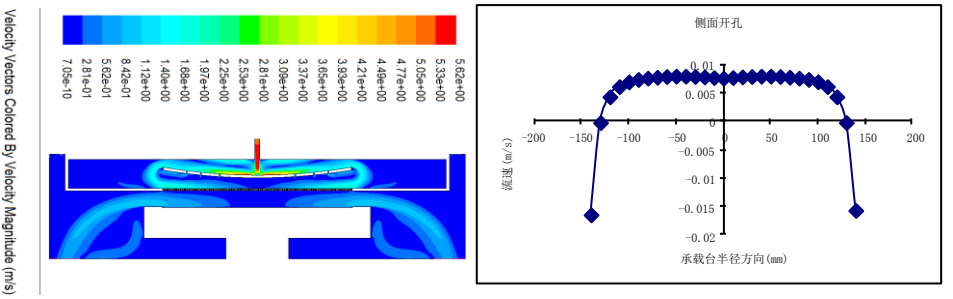


（a）Velocity vector （b）Velocity distribution near substrate

Figure 2: Flow field characteristics of processing chamber

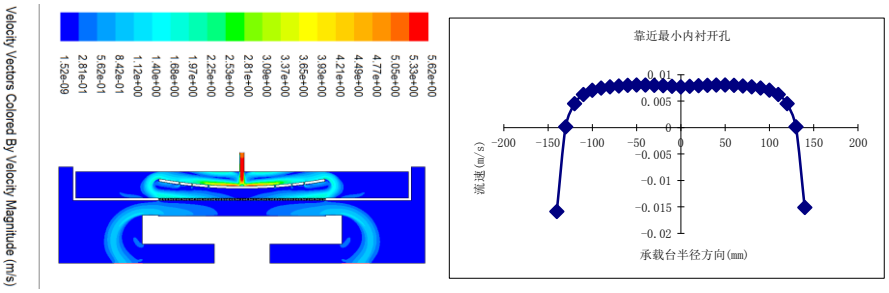
Figure 2 shows, after the design of showerhead, the effective film region is ± 90mm, in which velocity relative error is less than 5%.

On the bottom, there are three places to set the exhaust port. The first place project is near the side, close to the processing chamber wall; the second is near the minimum lining, between the minimum lining and substrate; the third is in the middle, close to the lifting mechanism. The diameter of the exhaust port is 50mm, and simulation results are shown in Figure 3 to Figure 5.



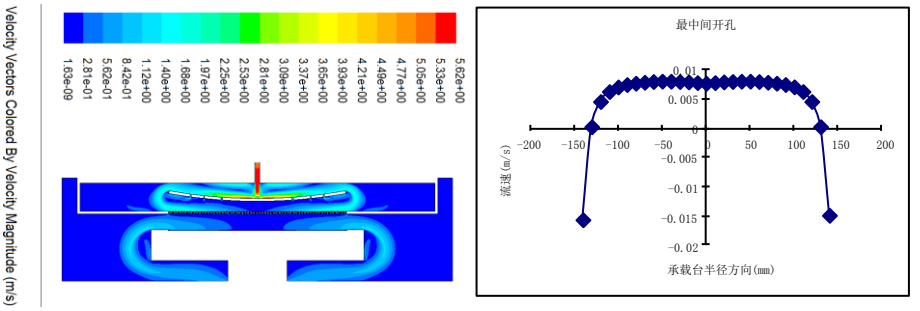
（a）velocity vector （b）velocity distribution near substrate

Figure 3 Flow field characteristics of processing chamber in project 1



（a）velocity vector （b）velocity distribution near substrate

Figure 4: Flow field characteristics of processing chamber in project 2



（a）velocity vector （b）velocity distribution near substrate

Figure 5: Flow field characteristics of processing chamber in project 3

The simulation results show, the location of exhaust port under the chamber scarcely affect the distribution of wind speed nearby the substrate, so three place projects are all well in principle. However, in Figure 3, diameter of the minimum lining is 460mm, and the first project will interfere its placement, so the first project is ruled out; in Figure 5, the flow field inside is very complex, and there will be a great vortex under the substrate. The lifting mechanism also greatly affects the gas exhaust, so the third project is ruled out; in Figure 4, when the port position is between the minimum lining and substrate, there is no interference among the lifting mechanism、the minimum lining and the gas exhaust, so the second project is the best.

5 Processing chamber structure

According to the research based on the flow uniformity, structure of processing chamber is designed in Figure 6. In figure 6, the lifting mechanism is used to complete the substrate's change in the height direction; linings are placed to achieve change in the diameter direction; showerhead is designed to achieve the airflow fully mixed, and ensure the flow uniformity near the substrate; gas exhaust unit is designed, without causing interference, to keep the chamber's flow balance.

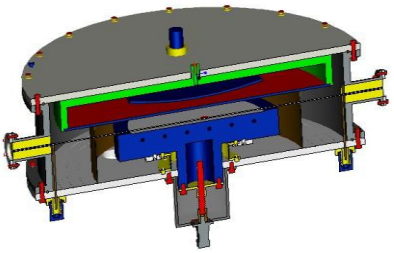


Figure 6: Structure of the processing chamber

6 Conclusions

The research shows that: Orthogonal experimental design and flow field simulation are applied to the mechanical design reasonably. Based on the flow uniformity, structure design of processing chamber is completed, greatly improving the efficiency and accuracy; Spalart-Allmaras one-equation turbulence model is effective for the shear flow simulation of wall in processing chamber with a low Reynolds number; Design of showerhead is critical, which directly affects the distribution of flow field inside, as well as the etching effect; the two showerheads make the gas mixture more uniform, and keep the flow uniformity nearby the substrate; the effective film region is ± 90mm, in which velocity relative error is less than 5%; The location of exhaust port under the chamber scarcely affect the distribution of wind speed nearby the substrate, so gas exhaust unit is designed in the view of mechanical, in order to make the whole processing chamber achieve the dual effect of beauty and utility.

Acknowledgments

The authors would like to thank the National Science & Technology Major Project (no. 02) of the Most (no. 2011ZX02403-004), china.

References

[PKK08] Park,Y.W.; Kim，K.B.; Kim，M.J.: Chemical Vapor Deposition Apparatus and Plasma Enhanced Chemical Vapor Deposition Apparatus, United States, US 2008/0295772 A1 [P], 2008.

[OT11] Nöll O.; Nöll O.; Kleyer T.; Gas Supply System and Method For Providing A Gaseous Deposition Medium, United States, US 2011/0033618, 2011.

[K10] Scott K., Butcher A.; Migration And Plasma Enhanced Chemical Vapor Deposition, United States, US 2010/0210067, 2010.

[GC09] Furuta G.; Choi Y.; Choi S.Y.;, Park B.S., White J.M., Anwar S., Tiner R.L.: Plasma Processing Apparatus And Method, United States, US 2009/0258162, 2009.

[S03] Dunham S.W.: Plasma Generator Assembly For Use in CVD And PECVD Processes, United States, US 6626998 B1, sep.30.2003.

[GJM03] Strauch G., Kaeppeler J., Dauelsberg M.: Gas-Admission Element for CVD Processes and Device, United States, US 2003/0177977 A1 [P], sep.25.2003.

[HMKK02] Setyawan H.; Shimada M.; Ohtsuka K.; Okuyama K.:Visualization And Numerical Simulation Of Fine Particle Transport In A Low-Pressure Parallel Plate Chemical Vapor Deposition Reactor, Chemical Engineering Science, 2002; 57, 497- 506.

[L10] Lu L.X.: Research on Computer Simulation of Growth Mechanism of a-Si:H Thin Film for PECVD, Ph.D Dissertation, Shanghai University, Shanghai, 2010.

[CZ08] Cheng J.; Zhu Y.; Duan G.; Wang C.: Analysis Of Processing Chamber Flow Field Characteristics For An ICP Etcher Based On Regression Orthogonal Design, Journal of Semiconductors, 2008, 29(4): 780-783.

Design and Analysis of a Lorentz Linear Motor for Precision Vibration-Isolation[1]

Yuan Fang, Wang Shibo, Zeng lizhan, Jin jianxin

State Key Lab of Digital Manufacturing Equipment & Technology
Huazhong University of Science &Technology
1037 Luoyu Road
Wuhan 430074, Hubei Province, China
zenglizhan@mail.hust.edu.cn

Abstract: This paper describes a design approach to Lorentz motors that serve as active damping component applicable in dynamic vibration isolators used for integrated circuit(IC) equipments. To satisfy the demand of high force density in constrained space and high current accuracy of motor drive for active damping, an analytical layer model with proper simplifications is employed to get preliminary dimensions, then a 3-D finite element(FE) model is built and analyzed to obtain a further optimal result. Finally a motor drive including current regulation circuit, H-bridge drive circuit, LC filter circuit and the current feedback circuit is developed. The performance of the motor is evaluated by experiments.

1 Introduction

Last decade witnesses a significant progress in semiconductor industry. The highly developed technology and high-precision manufacturing of IC industry make its machining precision and detection results ultra-high sensitive to the vibration, a small amount of which may introduce undesirable noise and affects the manufacturing accuracy. As a result, vibration isolation has been being an important research topic in IC industry [HL94]. The system and the excitation force of IC equipment often operates at varying frequencies, so ultra-precision vibration isolators adopting active isolation technique are the most common solution to vibration problems[CF05]. Air-cored Lorentz linear motors, are characterized by direct drive, high speed, high acceleration, high positioning precision, fast dynamics and low force pulsation, which make them easy to maintain, so they are well-recognized to be an appropriate choice as active dynamic actuators in precision vibration isolation systems [GP00][KJ04].

A kind of active dynamic isolator IC manufacturing using air-cored Lorentz motor for is introduced in this paper, shown in Figure 1. The dual-chamber pneumatic spring

[1] The work of this paper is partially supported by the National High Technology Research and Development Program of China (863, Grant No. 2009AA04Z148), and the National Science and Technology Major Project of the Ministry of Science and Technology of China (Grant No. 2011ZX02403).

provides high bearing capacity (>10000N) and isolate high frequency vibration. The motor works as a sky hook damper that introduces absolute damping into the system, providing small but fast damping force (<100N) responding to the velocity signal of low frequency vibration.

The active damping of the isolator calls for a near-zero stroke (≤ 20μm) but high force density in restricted space. In this paper, within the volume of 160×80×70mm^3, the force constant of object motor must reach the amount of 70N/A, which is almost double that of a traditional slotless Lorentz linear motor under the same condition, meanwhile the constant is strongly required with high linearity. As a result, the force generation capability of the object motor should be elaborate. In this case, an optimization design method for the object motor is essential[IS07][OK06][JC10]. The motor drive also must be proper designed to meet high bandwidth and current accuracy.

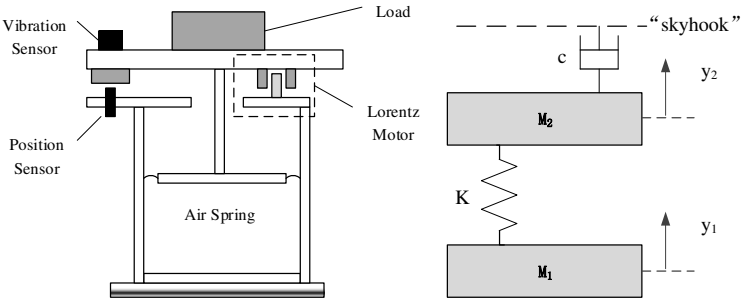


Figure 1: Structure of the active dynamic isolator and its schematic

2 Motor Design

2.1 Analytical Model and FE Analysis

The Lorentz motor in this paper consists of a primary located between two secondaries. Each secondary consists of rare earth permanent magnet (PM) poles mounted on a back iron. This arrangement provides two air gaps with small length. The primary is an air cored coil fixed in an aluminum alloy box.The lack of an iron core with teeth and slots and the makes it possible to employ a simple physical layer model to provide a basis for formulation of the motor [IS07], as shown in Figure 2, on the assumption that:

1. The air-gap magnetic field is 2D plane field.
2. The whole model is linear.
3. The permeability of iron layers is infinity.
4. No external magnetic field interference.

The model consists of two iron layers, two permanent magnet layers and an air layer; representing the back irons, PM poles and the air gaps plus the primary winding. The fundamental and harmonic component of the PM field can be obtained from Maxwell

equations in the different layers. The maximum value of the PM flux under each pole is related to the fundamental component, so by minimizing the total harmonic distortion of the PM field the force pulsations can be minimized and the utilization of PM fields can be maximized.

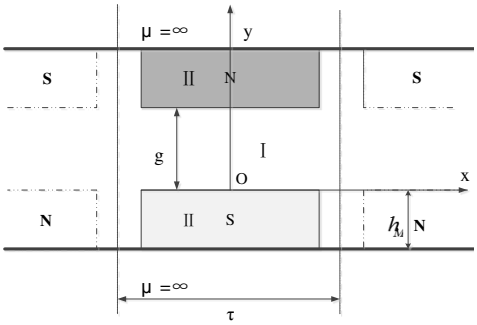


Figure 2: Layer model of air-cored Lorentz linear motor

The total harmonic distortion is given by

$$HD_{total} = \frac{\sqrt{\sum_{n=3,5,\ldots}^{\infty} \frac{n\pi}{\tau}\left(C_1 e^{\frac{n\pi g/2}{\tau}} + C_2 e^{\frac{-n\pi g/2}{\tau}}\right)^2}}{\frac{\pi}{\tau}\left(C_1 e^{\frac{\pi g/2}{\tau}} + C_2 e^{\frac{-\pi g/2}{\tau}}\right)} \tag{1}$$

where

$$\begin{cases} C_1 = \dfrac{\dfrac{4B_r\tau}{n^2\pi^2}\sin\left(\dfrac{\eta n\pi}{2}\right)e^{\frac{-2n\pi g/2}{\tau}}}{\left(e^{\frac{-2n\pi g/2}{\tau}}+1\right)+\left(\dfrac{\mu_M\left(-e^{\frac{-2n\pi g/2}{\tau}}+1\right)\left(e^{\frac{-2n\pi h_M}{\tau}}+1\right)}{\mu_0\left(e^{\frac{-2n\pi h_M}{\tau}}-1\right)}\right)} \\ C_2 = C_1 e^{\frac{2n\pi g/2}{\tau}} \end{cases} \tag{2}$$

Br is the remanence magnetic flux density, g is the air gap length h_M is the magnet height, τ is the pole pitch and η is the ratio of the magnet width to pole pitch.

As the stroke of the motor is near zero (≤ 20μm) and for control conveniently, the motor is designed to be a single-phase DC-PM.

From Eqs. (1), preliminary optimal motor dimensions are obtained with the help of Matlab with proper constraints. Then a 3-D FE model based on the preliminary result for further optimal design is built with the help of Ansoft, taking material nonlinearity, end effect and asymmetrical winding into account.

The 3D FE model uses volume current method and quadratic interpolation to approach field quantity in each element. The primary coil is equivalent to current density. Since there is flux leakage and end effect, a balloon boundary is adopted as solution region to simulate real magnetic fields. The whole model is divided into about 30,000 elements.

Figure 3 shows the analysis result of flux density distribution with the 3-D FE model. The flux density in the middle of air gap obtained from the analysis result well agrees with the analytical result, so the accuracy of the analytical model is validated.

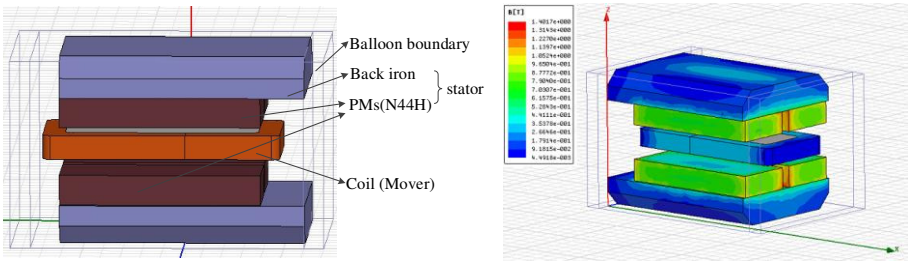


Figure 3: FE model of air-cored Lorentz linear motor and analysis results

The FE analysis results also help get the optimum working area of the primary, which is simply considered as an air layer with the air gap, adjust the PM dimensions and the air gap, optimize the inductance of coil and check whether the force constant attain the design target. With the 3-D FE analysis, further optimal motor dimensions are obtained.

2.2 Motor drive

The motor drive is designed based on PII control algorithm to provide the motor with fast following response and high control accuracy, the system diagram is shown in Figure 4.

In Figure 4, current regulator (CSR) output reference voltage $u_{CSR}(s)$ based on the voltage deviation $u_{\varepsilon}(s)$. Then $u_{CSR}(s)$ is sampled and converted into digital signals by digital converting unit ADC1 of microcontroller STM32. The signal is processed and turned into PWM wave, then output to H-Bridge power convertor. The PWM driving signals output to Lorentz motor through LC filter circuit. Voltage signals from Hall sensors are calculated by difference amplifier, then current feedback voltage u_{fs}) is outputted, making the drive a closed loop feedback system.

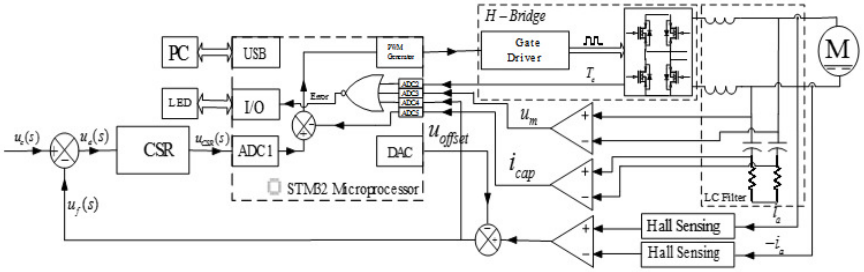


Figure 4: System diagram of motor drive

The zero-point bias voltage of Hall sensor is compensated by the analog signal u_{offset} from the digital-analog converting unit DAC of microcontroller STM32. Capacitance current of the LC filter is precision measured and sampled by ADC, the subtraction of which and $u_{CSR}(s)$ is used to improve the damping coefficient and to reduce second-order oscillation of the motor drive. There is also protection circuit in case of overvoltage, overcurrent and overheat.

3 Experiments

This paper develops an experimental prototype based on the design results, as shown in Figure 5. An air-cored Lorentz motor is built according to the design dimensions based on FE analysis results and the motor drive design. The driver module and the control module are developed as well. The experimental prototype also includes a test device to obtain the force constant and the dynamic response of the motor.

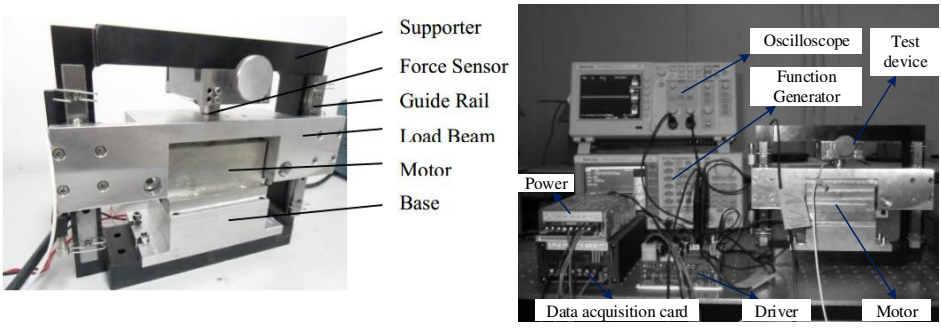


Figure 5: Experimental prototype

Test results are shown in Figure 6.The force constant of the designed motor is obtained as 72.16N/A (design target is 70N/A and the FE analysis result is 73.2N/A) with good linearity(the norm of residuals is 0.808). The following response of the motor force is fast (<0.5ms) within the working frequency (⩽200Hz), and the force pulsation of the motor is low.

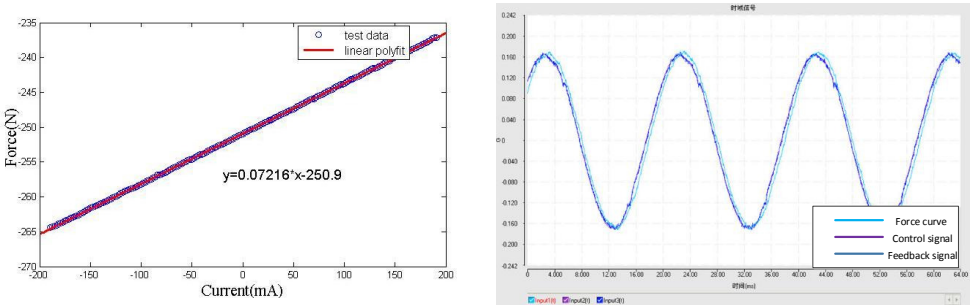


Figure 6: Force constant curve and dynamic response of the designed motor

4 Conclusion

This paper proposes a design method of Lorentz linear motor for active damping. An analytical layer model with proper simplifications is employed to quickly find preliminary optimal motor dimensions, and then the finite element (FE) model of the motor is built based on the analytical results. The analysis results of FE model help get further optimal motor dimensions. Then suitable controller is designed based on PII control algorithm. Finally, the performance of the motor is evaluated by experiments. The results of experiments show that the performance of the designed motor as the active damping component of a vibration isolation system is excellent. The proposed method is validated to be feasible.

References

[HL94] Huang, S. J.; Lian, R. J.: A dynamic isolator with active vibration control, J. of Sound and Vibration, 1994; 178(3), 323–335.

[CF05] Chen, Y. D.; Fuh, C. C.; Tung, P. C.;: Application of voice coil motors in active dynamic vibration isolators, IEEE Transactions on Magnetics, 2005; 41(3), 1149-1154.

[GP00] Gieras, J. F.; Piech, Z. J.: Linear synchronous motors, CRC Press, Boca Raton, 2000.

[KJ04] Kim, J. K.; Joo. S. W.; Hahn, S. C.: Static characteristics of linear BLDC motor using equivalent magnetic circuit and finite element method, IEEE Transactions on Magnetics, 2004; 40(2), 742–745.

[IS07] Isfahani, A. H.; Sadegh, V. Z.: Design optimization of a linear permanent magnet synchronous motor for extra low force pulsations, Energy Conversion and Management, 2007; 48, 443-449.

[OK06] Okonkwo, R. C.: Design and performance of permanent-magnet DC linear motors, IEEE Transactions on Magnetics, 2006; 42(9), 2179-2183.

[JC10] Jang, S. M.; Choi, J. H.; Park, J. H.: Optimum DC link voltage extraction based on dynamic performance analysis and design parameters of slotless double-sided permanent magnet linear synchronous motor, IEEE Transactions on Magnetics, 2010; 46(6), 2286-2289.

A Kind of Multi-disciplinary Simulation and Design Platform for IC Chamber based on Commercial Solver

Wu Xiaojing, Cheng Jia, Ji Linhong, Hou Yuemin, Lu Yijia

Department of Mechanical Engineering
Tsinghua University
100084 Beijing China
wuxiaojing218@mail.tsinghua.edu.cn
chengjia@mail.tsinghua.edu.cn

Abstract: In this paper, a kind of multi-disciplinary simulation and design platform for wafer manufacturing process with Chamber system is presented. This platform is developed as an in-house program, with different functional component for multi-disciplinary problems, which can drive the commercial FEM solver with code. There are also management function for user, products, and analysis or optimization process templates. This platform can be used for analyzing the multi-disciplinary process inside chamber system, DOE and optimization analysis for wafer manufacturing process, product and analyzing process template management.

1 Background

With the rapid development of manufacturing technology, the IC manufacturing process become a kind of very complex and high-accuracy process, which is composed with many different displinary problems, including fluid and thermal field, electric-magnetic field, and plasma field. Chamber is the main reaction part for wafer manufacturing process, such as PECVD, PVD, LPCVD, and etching[CY01]. There are plenty of factors in multi-field inside the chamber, which can affect the wafer fabricating performance. To improve quality of wafer fabrication, it should be considered in a multi-disciplinary interaction perspective[CY02][WT03].

In this paper, a kind of multi-disciplinary simulation and design platform for wafer manufacturing process with Chamber system is presented. This platform is developed as an in-house program, with different functional component for multi-disciplinary problems, which can drive the commercial FEM solver with code. There are also management function for user, products, and analysis or optimization process templates. This platform can be used for analyzing the multi-disciplinary process inside chamber system(such as thermal-fluid, magnetic, and plasma), DOE and optimization analysis for wafer manufacturing process, product and analyzing process template management[MS04].

2 Platform Introduction

2.1 System Structure

This platform system is composed with three functional group, which are User-Interaction module, Solver and Analysis module, and Data and Process management module.

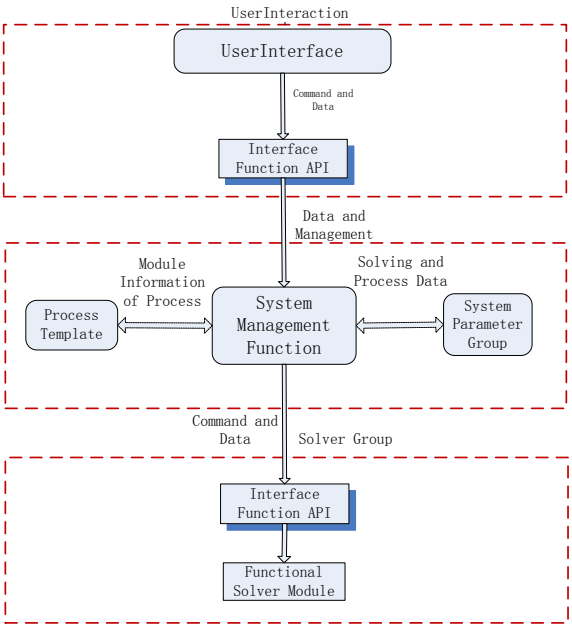


Figure 1: System structure

User-Interaction module is a connection module with user and platform. This module include user-interface and interface function API. The first one is the window for user and platform system, which can be used for receiving data and command from the user, and can provide analyzing results to user. The Interface function API can be used for delivering the data and command in and out between User-interface and platform system.

Data and Management module is the core management part for the system. This module includes system management function, system parameter, product, user database, and analyzing process template database. This part can be used for receiving the data and command from upstream part, store the parameters and products information into database, then deliver the command to downstream part from interface function API.

Solver module is the functional part for this platform system, which can execute simulation, data analyzing, DOE design, and process optimization. This module includes the API code, which can be used to drive the commercial FEM solver software, DOE

module, and optimization code. The different solver can be executed separately in the logic sequence of the analysis process templates to solve the multi-disciplinary problems inside chamber system. Thus the user can define any analysis problem on IC chamber as a solving process template into this platform system, which can be used for the similar problems with different working parameters, or do optimizations process for this issue.

2.2 Database Structure

This platform is not only a kind of simulation and analyzing tool with commercial software, but also a kind of management system for user, product, process template, and parameters. This platform include four group of data tables, which are user table, products table, simulation process template table, and simulation and analyzing parameters table.

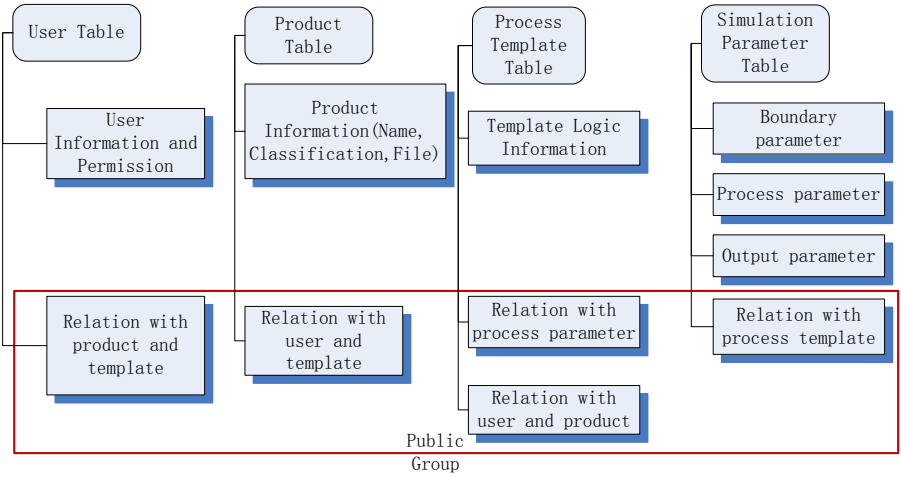


Figure 2: Database structure

User table is used for statistic and management of user data, including user information, permission, and the relation information with products and process templates

Product table is used for statistic and management of product data, which need analysis and optimization. These data include product information (such as name, classification, file), and relation information with user and analyzing process template.

Process template table is used for statistic and management of analyzing process template data, including template logic information, which show the solving sequence of each function component for one process, and relation information with process parameters, user and product.

Simulation parameter table is used for statistic and management of simulation process parameters, including model geometry parameters, simulation boundary parameters,

process control parameters and output parameters. Relation information with process template is also included in this table.

2.3 Working Process

This platform is a kind of process analyzing and data management system, which is based on user group and corresponding products. Each user and product possesses different process analyzing and related parameter records, which are focused on different physical field.

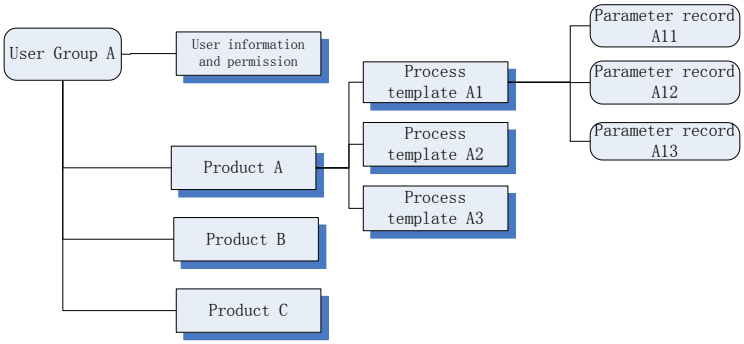


Figure 3:Working Process

Figure 3 shows one sample of working process of this platform.

3. Data Center and System Model

3.1 Experiment Platform for ICP system

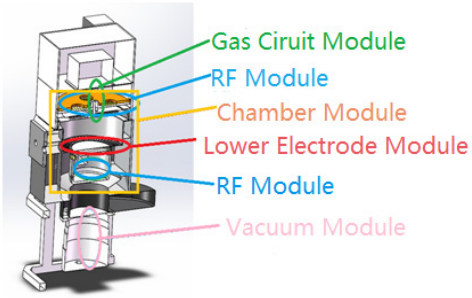


Figure 4: Experiment platform module

To verify the accuracy of simulation results, a experiment platform for ICP system has been building, which can obtain more process data of IC chamber in detail. This platform is combined with five parts, which are gas circuit module, RF module, chamber

module, lower electrode module, and vacuum module. The test data can be obtained from these detecting equipments, including plasma energy parameters, electron parameters and gas composition proportion.

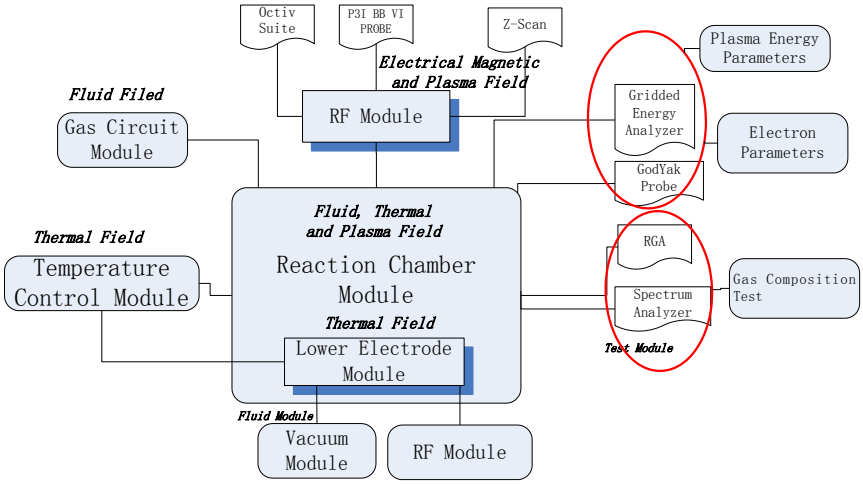


Figure 5: Experiment equipment system

3.2 Simulation Platform for ICP system

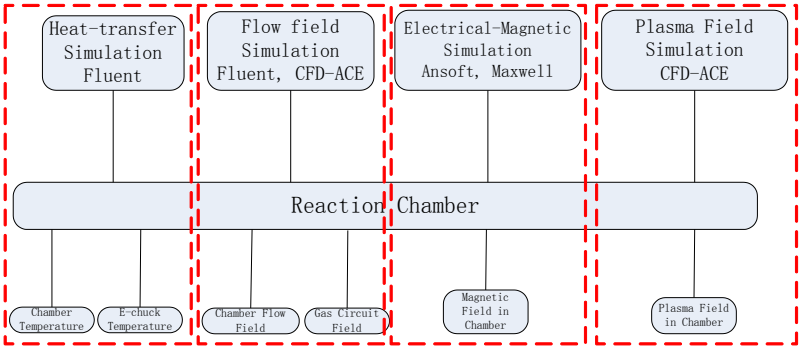


Figure 6: Simulation platform system

On account of limited data acquired from test system, simulation platform system is built to get more detailed results in multi-disciplinary fields. In the simulation system mentioned above, there are four categories of analyzing component for reacting chamber model, which are consisting of heat-transfer, flow field, electrical-magnetic and plasma field with different commercial FEM solvers, such as Fluent, CFD-ACE and Ansys.

With the simulation platform, detailed information of reaction chamber can be obtained, with multiple groups of operating parameters or geometry parameter sets. The results from this system can be used for comparison with data from test system to calibrating

simulation model. Furthermore, the simulation system after calibration with test data can be used to implement DOE process automatically, and create system research model for IC chamber model.

4.Conclusion

In this paper, simulation and experiment platform system has been built for the research on IC reaction chamber model. Both of the two systems can be used for launch multi-disciplinary FEM calculation and tests for chamber system. The calibration process of simulation platform can be implemented with data from experiment system, to improve the first one mentioned of these two systems[RME05].

Furthermore, the improved simulation platform can be used to do more intensive study for IC chamber, such as executing DOE process, optimization and creating system model to describe the relationship with different components inside complex IC chamber.

References

[CY01] Cheng, J.; Yu, Z.: Gas f low simulation resea rch on reaction chamber of ICP etcher . Semiconductor Technology , 2007; 32 (1).

[CY02] Cheng, J; Yu, Z.: Analysis of Processing Chamber Flow Field Characteristics for an ICP Etcher Based on Regression Orthogonal Design,Journal of Semiconductors, 2008; 29(4).

[WT03] Wu, Y.; Tian, L.; Integrated development platform of integrated circuit equipment, Computer Integrated Manufacturing System.2013; 19(9).

[MS04] Manish, C.; Sanjay, G.: Compensation of overlay errors due to mask bending and nonflatness for EUV masks, Proc. of SPIE; Vol. 7271 72710G-1.

[RME05] Ramaswamy, V.; Mikkelson, A.; Engelstad, R.: Predicting Wafer-Level IP Error due to Particle-Induced EUVL Reticle Distortion During Exposure Chucking, Proceedings of SPIE; Vol. 5992, p. 59923T, 2005.

Study of Design Method for Tangible User Interface

Xinyu Guan, Jingyan Qin, Yuemin Hou, Ji Linhong

Department of Mechanical Engineering
Tsinghua University
No. 1 Tsinghua Yuan, Haidian District, Beijing, China
100084 Beijing
guanxinyu1989@gmail.com
hym01@mails.tsinghua.edu.cn

Abstract: The third industrial revolution facilitated the rapid development of Tangible User Interface (TUI) in Internet of Things (IOT). The concept of TUI broke through the limitation of traditional Graphical User Interface (GUI). By giving objects digital information, the objects in our daily life can be transformed to the tangible interfaces between human and the digital world. Fully using daily life experience, the atoms and the bits can be combined perfectly, which gave the human-computer interaction (HCI) a new meaning. This paper involved the development of HCI, principles and evaluation criteria of human-computer interface design, the characteristics of TUI and in end proposed the method for TUI design.

1 Introduction

In the third industrial revolution, the cyber space realized the object transformation from physical to digital form, which promoted the social transformation from industrial to information. The fuse of physical and digital world made the bits exceed the physical boundaries and brought the atom revolution in the world of bits. This fuse and transformation led to TUI generation, which let us return our "global village" interweaving the reality and the digital. The concept of TUI broke through the limitation of traditional GUI. By giving the objects digital information, the objects in our daily life can be transformed to the tangible interfaces between human and the digital world. Fully using daily life experience, the atoms and the bits can be combined perfectly, which gave the HCI a new meaning. The fuse of digital and physical world provided wide space for the development of TUI.

2 Principles and evaluation criteria of human-computer interface design

According to the Human Interface by Jef Raskin [Ra00], the human-computer interface design should include two basic principles: 1) the computer cannot destroy the users' jobs, and 2) the computer cannot waste users' time or require users to do more jobs than necessary. Under the prerequisite of the two principles, seven specific criteria were proposed including visibility, feedback, affordance, simplicity, structure, consistency and tolerance to evaluate and guide the human-computer interface design. GUI, precisely, is the action part of visual experience and interaction of screen productions, and it emphasizes that human, computer and environment should be regarded as one system when designing the overall design [GRR06] . Additionally, GUI design gradually transforms from the task-centered to the user-centered design, and in the user-centered interaction design there are two phased objectives: 1) usability and 2) user experience.

3 Characteristics and method for TUI

TUI covers four main characteristics: 1) the objects and the bottom digital information are overlapped; 2) the objects themselves take the interactive control mechanism; 3) the objects can reflect the continuous changes of information carriers, and 4) the physical features of the objects can express the key features of digital systems. The analysis and design process is regarded as a reasonable and gradual refinement process from the context fitting the design to the design fitting the context. This process can be divided into five steps, i.e. context analysis, behavior analysis, mapping analysis, meaning analysis and rationality analysis, in which the mapping analysis is the most important. The implicit knowledge and the user cognition are the cores of the process. The objects, the information, the information carriers and the characteristics of user cognitive behavior should be considered comprehensively for the study on method for TUI. Besides, with the empathy the mental model of users, data structure and five senses of interface can achieve the same sentiment and assimilation, and in the end the interactive behavior.

References

[Ra00] Raskin, J.: The humane interface: new directions for designing interactive systems-Addison-Wesley Professional, 2000.

[GRR06] Giudice, F.; Rosa G.L.; Risitano A.: Product Design for the Environment: A Life Cycle Approach, Taylor & Francis, 2006.

The Prototype of Design Product Knowledge Representing System

Jia Lin, Jia Cheng, Hou Yuemin, Ji Linhong

Department of Mechanical Engineering,
Tsinghua University,
Beijing, China,
100084
linjia@tsinghua.edu.cn
chengjia@tsinghua.edu.cn
hym01@mails.tsinghua.edu.cn

Abstract: The purpose of this position paper is to introduce the prototype of a design product knowledge representing system which is theoretically guided by our recent work on extending Function-Behaviour-Structure ontology. The basic constructions and data-storing strategies are presented in the paper along with its major interface, thus to give an outline of the working pattern of this system.

1 Background

To represent those design products information, they must firstly be categorized into different classes within each of which same feature can be concluded. One of the most influential design ontologies is the Function-Behaviour-Structure (FBS) framework [GK04]. Despite of being widely adopted, the controversies about FBS theory were just as intense as its popularity in the latest decade, especially with the intricate definitions of key concepts [Ve13]. In our former research we introduced the object perception principles of cognitive psychology [SK07] to the design theory field. By taking this new approach, it is suggested that all the object systems are dominated by a common tree-like topology, and the Function, Behaviour and Structure of any object node in the treemap are hence interpreted as the relationships between it and its encompassing system, interacting objects and internal components, respectively. And the design system presented in this paper is thus developed to verify our theory. As design products information provide the knowledge framework for the design projects and link the variety of design activities, applying this system in design practise can visualize different design information and guide the progressing of design project. Furthermore, the classification of products information means that the design practice, which contains various procedures mapping from one kind of product description to another, are modularized. And with the standardized interfaces, the variety of design tools can be integrated as parts of the system.

2 The System

Based on our extending work on FBS ontology, a new design knowledge representing system is suggested. Within the system, a treemap of the design product is generated and the variety of design products information are represented as textual strings and allocated to each not of the tree. All data are stored in server which can be accessed by and the designers/engineers with webpage browser. The treemap will serve as a draft of the design product descriptions and guide the designers and engineers going through different design steps, like formulation, synthesis, evaluation, etc. Various design tools, such as geometric modellers, FEM calculators, optimizers, etc., can also connect to the system via web services, and take the design product describing information as the import or export data of their calculating cycle. The network-based construct allows parallel processing and real-time information exchanging among multiple human users and/or calculating units.

The major human interface of the system contains four functional areas. The Scenes area lists all the environments that the design product is involved. The Topology area lists all the objects in every granularity and organized them in the boundary-based treemap. The main screen displays all the entities and descriptions concerned in the design cases, include objects, interactions, roles, composing relationships, etc., and focus on the object which the user is currently working with. And the geometry model of that object is displayed in Model area.

All data are stored in relational database. Two most important classes in the system are Object and Interaction. Both involved in a designing scene are saved in the same table, with each instance occupying one row and attributes in columns. The operations of Object provide the data port for various design activities to be docked in, such as the design steps introduced in original and situated FBS frameworks including Formulation, Synthesis, Analysis and Evaluation, and also the Decomposition and Initialization which is suggested in our newly proposed nested FBS framework (in press). These activities can be either manual or software-centred. The classes diagram are depicted in Unified Modelling Language (UML).

References

[Ma14] Ma, Y.: Lifecycle Modelling, Analysis And Management Of Cyber-Physical Products, Advanced Design Concepts and Practice 2014 Summer Workshop, Beijing, 2014

[GK04] Gero, J. S.: Kannengiesser, U.; The Situated Function-Behaviour-Structure Framework. Design Studies, 2004, 25(4), 373-391.

[Ve13] Vermaas, P. E: The Coexistence Of Engineering Meanings Of Function: Four Responses And Their Methodological Implications. Artificial Intelligence For Engineering Design, Analysis and Manufacturing, 2013, 27(03), 191-202.

[SK07] Spelke, E. S.; Kinzler, K. D.: Core Knowledge, Developmental Science, 2007, 10(1), 89-96.

Cellular Genetic Algorithm Feature-based Modelling in Product Data Exchange

Bernadetta Kwintiana Ane

Institute of Computer-aided Product Development Systems
University of Stuttgart
Universitaetsstr. 38, D-70569 Stuttgart, Germany
ane@informatik.uni-stuttgart.de

Abstract: Currently, most of CAD/CAM systems are feature-based. Features are important to maintain geometric and non-geometric relations across product models. Often, during data translation some useful geometric and non-geometric information such as features and their associated semantics are stripped off. In this paper, we introduce a new method of feature-based modeling in product data exchange using cellular genetic algorithm (cGA). The cGA feature-based modelling is restructuring the B-rep of a product model based on cellular representation, to enable dynamic transmition of the cellular features into a B-rep at the manufacturing domain.

Keywords: Feature-based modelling, cellular genetic algorithm, CAD/CAM systems.

1 Background

Today, collaborative design is carried out not only amongst multidisciplinary product development teams within a company, but also across the boundaries of companies and time zones, with increased numbers of customers and suppliers involved in the process. During product development, information exists at every stage and stored in various forms. Those are produced by various computer-aided design and manufacturing (CAD/CAM) technologies which govern independent authoring tools and defined in proprietary formats [AR12]. This circumstance creates incompatibility whereas information sharing amongst multiple applications becoming bottleneck.

There are two methods for information sharing amongst heterogeneous CAD/CAM systems. One is direct translation and the other one is indirect translation, i.e., through a neutral intermediate format. In comparison to direct translation, the neutral format method is suitable for product data exchange amongst a large number of applications. In order to enable such interoperability, a commonly acceptable, comprehensive and well-defined information model is crucial [MTAC 09]. However, during data translation, some useful geometric and non-geometric information such as features and their associated semantics are usually stripped off. This paper introduces a new approach of feature-based design modelling using cellular genetic algorithm in product data exchange.

2 Feature-based Design

Currently, most of CAD/CAM systems are feature-based. Features are found very useful to encapsulate engineering intent in computer systems. Feature concept is defined for simple and commonly used design and machining geometry elements [Sh91]. In the field of engineering informatics, a feature is defined as a class of semantically endowed objects that accompanies product development from customer requests through to product release. A type of feature is a class of characteristic and semantic patterns of a set of interrelated domain-application entities involved for understanding, communication, reasoning and technical implementation amongst CAD/CAM systems [YM12].

In the field of product design and manufacturing, those existing feature definitions can be divided into two categories according to their granularity at different application levels, i.e. function features (FFs) and machining features (MFs). In the area of product design, a function feature is described as a spatial unit (element) representing a region of interest within a product which includes values and relationships, i.e., structure and constraints [JL07], [IKC08]. Whilst, in the area of manufacturing a machining feature is defined as the specification of a geometric region of a part, which includes shape, position, dimensions, tolerance, and surface finishing of that volumetric region [GZG04], [CG96], [CBZ08]. This definition is formed from the perspective of manufacturing convenience to minimize the total number of machining set-ups [CG96].

Features are domain-dependent. Features associate some generic solid shapes with engineering semantics to facilitate design and manufacturing processes. Function features contribute to the design construct basis of a part in terms of additive or subtractive volumes, and their semantics reflect the design intent and function. Machining features are normally derived from a manufacturing stock in terms of subtractive volumes, and they are conveniently associated with manufacturability analysis and process planning activities, such as fixture planning, machines and cutting tools selection, and machining operations planning [SM95].

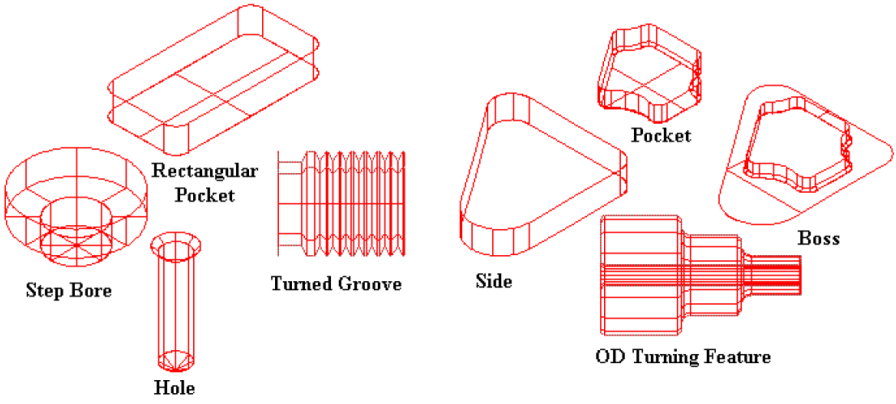


Figure 1: Common features of CAD/CAM applications [Image courtesy: FeatureCAM, 2008]

3 Product Data Exchange

Product design and manufacturing are semantically rich domains. Problems in these domains (e.g., reasoning, analysis, process planning, etc.) are difficult to solve as a large amount of data and information are available and complex decision-making processes are involved. In order to support feature-based applications to cross domains between design and manufacturing, there must be considerable similarity between the CAD/CAM systems. The systems must support the same level of information, same entities, and same semantics; only the representation and implementation can be different.

3.1 Feature Mapping

Feature models are domain dependent, i.e., when a part is designed by features, the resulting model is usually not in a form convenient for other applications, such as engineering analysis, process planning, and manufacturing. The two families of function features and machining features are genuinely different. To transform product information from a design-oriented feature model to a manufacturing-oriented model, it is necessary to transform the model from one viewpoint to another. This process is termed feature mapping. In this paper, the discussion of feature mapping is restricted to one-to-one mapping between spaces of the same dimensionality. Features such as holes, pockets, slots, and ribs might be similarly defined in two systems, even though the names given might be different.

There are several general-purpose methods for feature mapping. Here we use the cell-based mapping technique for mapping by cellular decomposition of solid geometry. The general idea behind cellular schemes for mapping is that in conjugate mapping, volumes are combined into different groups to yield features for different applications. Cellular decomposition, where the cells are solids or voids, might come directly from the original model or be obtained by various partitioning algorithms.

In this study, mapping from function features to machining features is performed by manufacturing feature search on the basis of a convex cellular decomposition [SSS94]. Each volumetric feature class is associated with a collection of topological entities (i.e., a set of characteristic faces) on its boundary, called definitional entities set. Definitional entities have a status attribute that is either positive or negative. Positive definitional entities are part of the model's boundary, whereas negative ones do not belong to that boundary. For instance, a 'slot' must have a positive floor and a negative roof in the model. A set of semantic constraints are imposed on each definitional entities set.

Feature models are represented in a cellular scheme which is a combination of CSG (constructive solid geometry) and B-rep (boundary representation). Feature interaction is detected based on feature definitional entities. Each feature is represented by either a positive cell or a negative one. Whenever two features interact, the respective cells are further decomposed. Parts of the resulting cells belonging to both features represent the interaction extent, while the remaining cells belong to either feature and carry some morphology of the original features.

Figure 2 describes the decomposed cells of a CAD model according to the cellular topology along with the feature modelling process. In comparison to the traditional boundary model, in cellular topology based non-manifold boundary representations, operations on volumes do not discard any input geometry. Part geometries are represented using cellular topologies [MCG09].

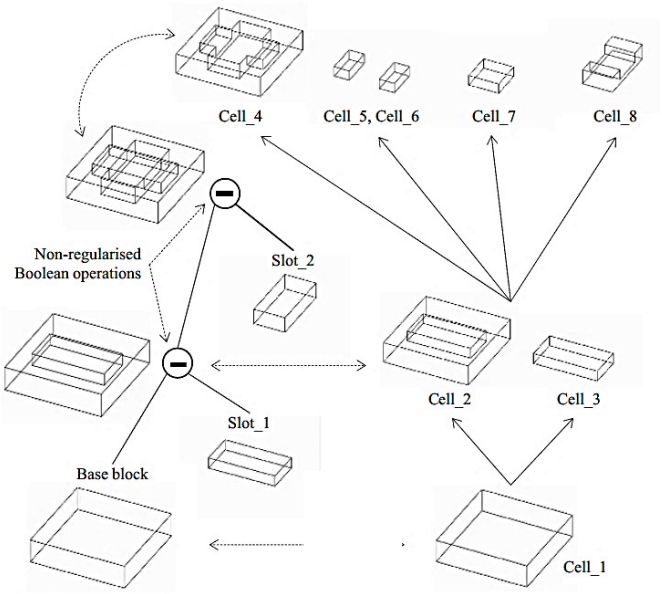


Figure 2: Feature model based on the cellular topology [MCG09]

3.2 Cellular Genetic Algorithm Features Model

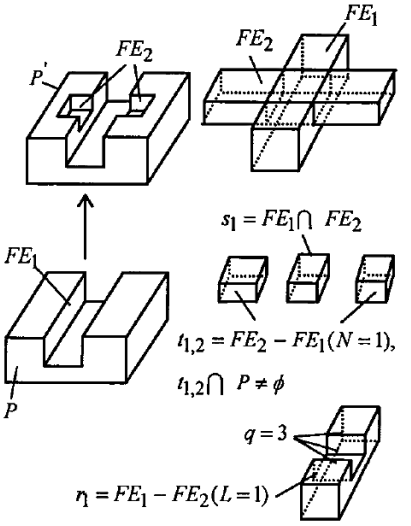
Features can be defined in various levels of abstractions and specifications, and different data exchange standards (e.g., IGES, DXF, PDES, or STEP) can be used to define different "schemas" for the features. To generate proper interpretations between function features and machining features, we propose an evolutionary method, called cellular genetic algorithm, to enable bi-directional communications and information conversions between various CAD/CAM application domains.

A cellular genetic algorithm (cGA) is a class of evolutionary algorithm (EA) with a decentralized population in which the tentative solutions evolve in overlapped neighborhoods, much in the same sense as cellular automata [AD08]. The essential idea of cellular model is to provide the EA population with a specific structure defined based on a connected feature-based graph, in which each node is to represent a single ‘gene’ (i.e., a feature) interacting with its nearest neighbors through the cGA operators (recombination, crossover, swap, mutation, insertion, and deletion).

The cGA encodes the decision variables of a search problem in finite length variable chains of an alphabet of a given cardinality. The resulting strings [0,1], which are candidate solutions to the search problem, are called chromosomes. A chromosome

represents a mechanical component of the product. It is composed of genes, representing features which construct the component. The different values of these genes are named 'aleles', storing information related to the feature attributes (i.e., dimensions, definitional entities, and orientation). An alele occupies a locus (pl.: loci), containing different values of feature vertex position. A gene is only allowed to recombine with close genes at the same chromosome. This leads us to a kind of locality known as isolation by distance.

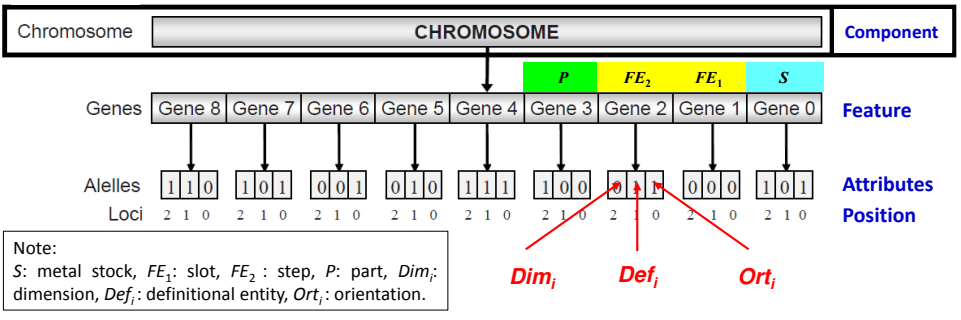
For illustration, Figure 3 depicts a general cGA model. This model analyses the interacting relationship between each function feature in a design feature-based graph with the machining features in an incrementally evolved intermediate manufacturing feature-based graph . In this model, the function feature being traversed is denoted as DF, machining feature being traversed is MF, and the manufacturing stock is S. The Boolean operation of S in volume is additive and the other features are subtractive.



FE_1 is adjacent:
$FE_2 \cap FE_1 \neq 0$,
$FE_2 \not\subset FE_1$,
$FE_1 \not\subset FE_2$,
r_1 and each S_i satisfies $p > 1$.

FE_2 is disconnected:
$FE_2 \cap FE_1 \neq 0$,
$FE_2 \not\subset FE_1$,
$FE_1 \not\subset FE_2$,
$N > 1$,
for each t_i: $t_i \cap P \neq 0$.

(a) The interacting relationship between features [LON06]



(b) Features coding

Figure 3: Cellular Genetic Algorithm features model

Conclusion

Features are important to maintain geometric and non-geometric relations across product models. The generic feature model provides flexibility and finer-level of granularity for an integrated feature-based modelling framework between CAD/CAM applications. Representing feature-level information uniformly is compulsory so that engineering meaning is completely shared amongst various application domains.

In this paper, we introduce a new method of feature-based modeling in product data exchange using cellular genetic algorithm (cGA). The cGA feature-based modelling is restructuring the B-rep of a product model based on cellular representation, and dynamically transmitting and embedding the cellular features into a B-rep at the manufacturing domain. The cGA algorithm applies six operators, i.e., recombination, crossover, swap, mutation, insertion, and deletion. Therefore, features and constraints can be maintained and propagated systematically during the data exchange.

References

[AD08] Alba, E.; Dorronsoro, B.: Cellular Genetic Algorithm. Springer Science+Business Media, LLC, New York, 2008.

[AR12] Ane, B.K.; Roller, D.: Collaborative Design: Using CAD Kernel in Knowledge-based Environment. In (Albers, A.; Behrendt, M.; Rusák, Z.): Proc. 9th International Symposium on Tools and Methods of Competitive Engineering (TMCE 2012), Karlsruhe, Germany, 2012.

[CBZ08] Chayoukhi, S.; Bouaziz, Z.; Zghal, A.: Cost estimation of joints preparation for GMAW welding process using feature model. Journal of Materials Processing Technology 199(1–3), 2008; p. 402–411.

[CG96] Chu, C.C.P.; Gadh, R.: Feature-based approach for set-up minimization of process design from product design. Computer-Aided Design 28(5), 1996; p. 321–332.

[GZG04] Gao, J.; Zheng, D.T.; Gindy, N.: Extraction of machining features for CAD/CAM integration. International Journal of Advanced Manufacturing Technology 24(7/8), 2004; p. 573–581.

[IKC08] Igwe, P.C.; Knopf, G.K.; Canas, R.: Developing alternative design concepts in VR environments using volumetric self-organizing feature maps. Journal of Intelligent Manufacturing 19(6), 2008; p. 661–675.

[JL07] Jin, Y.; Li, W.D.: Design concept generation: A hierarchical coevolutionary approach. Transactions of the ASME – Journal of Mechanical Design 129(10), 2007; p. 1012–1022.

[LON06] Li, W.D.; Ong, S.K.; Nee, A.Y.C.: Integrated ond Collaborative Product Development Environment: Technologies and Implementations. World Scientific Publishing, Singapore, 2006.

[MCG09] Ma, Y.-S.; Chen, G.; Thimm, G.: Fine Grain Feature Associations in Collaborative Design and Manufacturing – A Unified Approach. In (Wang, L.H. and Nee, A.Y.C.): Collaborative Design and Planning for Digital Manufacturing, Springer-Verlag, Heidelberg, 2009; p. 71-97.

[MTA 09] Maa, Y.-S.; Tang, S.-H.; Au; C.K.; Chen; J.-Y.: Collaborative feature-based design via operations with a fine-grain product database. Computers in Industry 60, 2009; p. 381–391.

[Sh91] Shah, J.: Assessment of feature technology, Computer Aided Design 23(5), 1991; p. 331–343.

[SI08] Saaksvuori, A.; Immonen, A.: Product Lifecycle Management., Springer, Heidelberg, 2008.

[SM95] Shah, J.J.; Mantyla, M.: Parametric and Feature-based CAD/CAM. John Wiley & Sons, Inc., New York, 1995.

[SSS94] Shah, J.J.; Shen, Y.; Shirur, A.: Determination of machining volumes from extensible sets of design features. In (Shah, J.J., Mantyla, M. and Nau, D.): Advances in Feature Based Manufacturing, Elsevier Science Publishers, 1994; p. 127-157.

[YM12] Yin, C.-G.; Ma, Y.-S.: Parametric feature constraint modeling and mapping in product development. Advanced Engineering Informatics 26, 2012; p. 539–552.

Risikokommunikation im Kontext von IT-Sicherheit und Safety

(RiskKom)

Workshop: Risikokommunikation im Kontext von IT-Sicherheit und Safety (RiskKom 2014)

Bernhard C. Witt

IT GRC Management / Datenschutz
it.sec GmbH & Co. KG
Einsteinstr. 55
89077 Ulm
bcwitt@it-sec.de

Abstract Die real vorhandene IT-Sicherheit einer Organisation hängt oft ab von der Risikokommunikation zwischen verschiedenen internen und externen Stellen, sowohl zur Awareness als auch bei Eintritt einer Datenpanne oder eines Krisenfalles. Behandelt werden aktuelle Forschungsergebnisse und Erfahrungsberichte aus der Praxis von Behörden und Unternehmen sowie im Kontext von Anwendern und der Gesellschaft.

1 Zum Thema des Workshops

Die IT-Sicherheit einer Organisation hängt nicht nur von den technischen Kenntnissen des IT-Sicherheitspersonals ab, sondern auch ganz entscheidend von der Kommunikation zwischen diesen Sicherheitsexperten, dem Management und den Benutzern bzw. Betroffenen. Dies betrifft sowohl die Kommunikation nach innen (z.B. hinsichtlich des Schaffens von Sicherheitsbewusstsein), als auch nach außen (z.B. bei Eintritt einer ggf. meldepflichtigen Datenpanne oder eines Krisenfalles).

Welches Sicherheitsniveau eine Institution tatsächlich umsetzt, ist zudem eng verbunden mit der persönlichen (subjektiven) Kosten-Nutzen-Einschätzung der vorhandenen Sicherheitsmaßnahmen. Dabei spielen sowohl die Einsicht in Notwendigkeiten als auch die Praktikabilität zu beachtender Regelungen eine Rolle.

Je komplexer eine Infrastruktur ist, je mehr Daten darüber abgewickelt werden, je mehr Personen betroffen sind und je stärker auf Daten oder Infrastrukturen vernetzt aus unterschiedlichen Kontexten zugegriffen wird, desto wichtiger ist es, einen angemessenen Schutz gewährleisten zu können. Allerdings werden dadurch auch die Zusammenhänge der Risiken komplexer und schwieriger zu kommunizieren.

In diesem Workshop der GI-FGn ECOM und SECMGT werden aktuelle Forschungsergebnisse und Erfahrungsberichte aus der Praxis im Bereich der Risikokommunikation diskutiert. Dabei wird auf den Ergebnissen des Vorjahresworkshops RiskKom zur Informatik 2013 aufgebaut und die dort fachlich begonnene Diskussion fortgesetzt.

2 Zum Ablauf des Workshops

10:30 – 12:00 Uhr: Begrüßung und Keynotes

Bernhard C. Witt (it.sec GmbH & Co. KG): Begrüßung, Einführung zum Workshop und Vorstellung der GI-FGn ECOM und SECMGT

Dirk Scheer (Uni Stuttgart / ZIRIUS): Ein Risiko ist ein Risiko ist kein Risiko! - Grundsätze der Risikowahrnehmung und Risikokommunikation

Kirsten Messer-Schmidt (excepture): IT-Risikokommunikation – ein Eingeständnis von Schwäche?

12:00 – 13:30 Uhr: Mittagspause

13:30 - 15:00 Uhr: Beiträge

Steffen Bartsch (TU Darmstadt / CASED), Carina Boos (Uni Kassel / provet), Gamze Canova (TU Darmstadt / CASED), Dominic Dyck (TU Darmstadt / CASED), Birgit Henhapl (usd AG), Michael Schultheis (TU Darmstadt / CASED) & Melanie Volkamer (TU Darmstadt / CASED): Interdisziplinäres Bewertungskonzept für Risiken

Jana Fruth (Uni Magdeburg), Sven Kuhlmann (Uni Magdeburg) & Jana Dittmann (Uni Magdeburg): Ein erster Prototyp: Sicherheitsguide für Grundschulkinder beim Umgang mit dem Internet

Thomas Pfeiffer (TU Darmstadt), Michaela Kauer (TU Darmstadt) & Julia Röth (TU Darmstadt): "A Bank Would Never Write That!" - A Qualitative Study on E-Mail Trust Decisions

Interdisziplinäres Bewertungskonzept für Risiken auf Webseiten

Steffen Bartsch, Carina Boos, Gamze Canova, Dominic Dyck,
Birgit Henhapl, Michael Schultheis, Melanie Volkamer

Universität Kassel/provet
Pfannkuchstr. 1, 34109 Kassel
Carina.Boos@uni-kassel.de

usd AG
Robert-Bosch-Straße 25 a, 63225 Langen
Birgit.Henhapl@usd.de

TU Darmstadt/CASED
Hochschulstraße 10, 64289 Darmstadt
Steffen.Bartsch@cased.de, Gamze.Canova@cased.de, D.Dyck@iad.tu-darmstadt.de,
Schultheis@iad.tu-darmstadt.de, Melanie.Volkamer@cased.de

Abstract: Dieser Beitrag beschreibt ein Konzept zur Verbesserung der Sicherheit von Nutzern im Internet: Angepasst auf die jeweilige IT-Sicherheitsexpertise sowie seiner Bereitschaft, Risiken einzugehen, werden dem Nutzer in risikoreichen Situationen unterschiedliche Interventionen geboten. Die Entscheidung, ob und welches Risiko existiert, wird auf rechtlicher und technischer Ebene getroffen: Indikatoren, ob Daten- und Verbraucherschutz eingehalten werden sowie ob grundlegende Maßnahmen der IT-Sicherheit umgesetzt sind, werden automatisiert erkannt und ausgewertet. Auf Grundlage der Risikoeinstufung des jeweiligen Szenarios sowie des antizipierten Risikoverhaltens und der IT-Sicherheitsexpertise des Nutzers wird über die Art der Intervention entschieden: Zusätzlich zu Warnmeldungen, die den Nutzer im Surfverhalten unterbrechen, existieren passive Interventionen, die den Nutzer nicht in seiner Handlung behindern, sowie eine permanente Anzeige über den Sicherheitsstatus einer Seite.

1 Einleitung

Der Einzug des Internets in den Alltag von immer mehr Menschen auch im Privatleben führt neben den großen Chancen für Verbraucher und Anbieter von Waren und Dienstleistungen auch seine Schattenseiten mit. Es entstehen nicht unerhebliche Risiken im Zusammenhang mit der IT-Sicherheit sowie dem Daten- und Verbraucherschutz: Beispiele sind das Ausspähen und Abgreifen von personenbezogenen Daten inklusive der Zugangs- und Kreditkartendaten sowie deren Missbrauch. Ebenso werden häufig verbraucherfeindliche oder unzumutbare Bedingungen nicht klar erkennbar in unübersichtlichen Verträgen formuliert oder sogar auf den Internetseiten versteckt.

Die zum Schutz der Nutzer angebotenen Mechanismen (TLS, Web of Trust, Gütesiegel, etc.) sind oft zeitaufwendig und setzen Expertenwissen voraus.

Das in diesem Beitrag vorgestellte Bewertungstool verspricht eine deutliche Verbesserung des Nutzerschutzes im Internet: Während der Aktionen des Nutzers auf einer Webseite (Surfen, Eingeben persönlicher Daten, Internetbanking, etc.) wird die Seite in Echtzeit auf Sicherheitsmängel sowie erkennbare Verstöße gegen und auf besondere Maßnahmen zum Verbraucher- und Datenschutz untersucht. Angepasst auf ein vorher einmalig erstelltes Risiko- und Expertise-Profil des Nutzers wird dieser bei Bedarf gewarnt und mit ihm verständlichen Handlungsempfehlungen oder -optionen dabei unterstützt zu entscheiden, ob er die Seite weiter verwenden oder verlassen möchte.

2 Analyse der Webseite

Das Bewertungstool untersucht die aktuelle Webseite auf bestimmte Indikatoren, die Aufschlüsse über die Einhaltung des Verbraucher- und des Datenschutzes (s. Abschnitt 2.1) sowie grundlegende Sicherheitsmechanismen (s. Abschnitt 2.2) geben. Ein wichtiger Indikator für beide Analysen ist der Webseitentyp (der mittels Machine Learning Ansätzen ermittelt wird [Ka14]). Das Ergebnis beider Untersuchungen wird anschließend zusammengeführt (s. Abschnitt 2.3). Basierend auf dem antizipierten Risiko und angepasst auf den individuellen Nutzer (s. Abschnitt 3) wird diesem eine Intervention (s. Abschnitt 4) angezeigt. Abbildung 2.1 illustriert das Bewertungskonzept im Überblick.

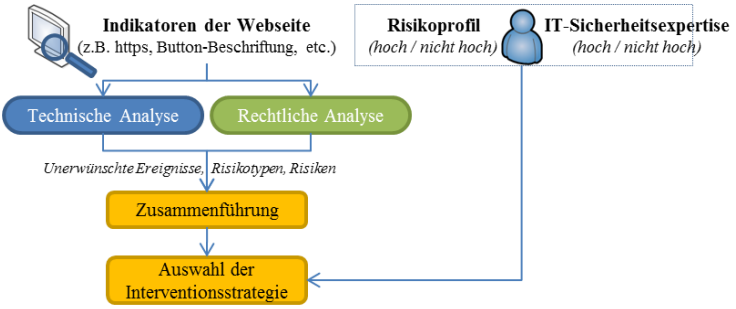


Abbildung 2.1: Bewertungstool im Überblick

2.1 Rechtliche Analyse

Um rechtliche Risiken bei der Nutzung einer Webseite automatisiert einzuschätzen, müssen aus Informationen Rückschlüsse gezogen werden, die automatisiert erhoben werden können. Daher kann und soll nicht die geprüft werden, ob eine Webseite rechtskonform, sondern ob sie vertrauenswürdig ist. Der Nutzer soll darüber informiert werden, ob er dem Anbieter vertrauen kann, nicht dagegen, ob er sein Recht notfalls auch vor Gericht durchsetzen kann.

Es wurde ein Risikomodell entwickelt, das eine Risikoabschätzung für jede Webseite ermöglicht. Dieses basiert auf deutschem Recht und schränkt die Aussagekraft der rechtlichen Risikoanalyse hauptsächlich auf deutsche Webseiten ein. Der Fokus liegt außerdem auf Web-Shops, für die eine Vielzahl einheitlicher Regelungen aus dem Verbraucher- und Datenschutzrecht gelten.

Um ein Schema zur Einschätzung rechtlicher Risiken zu erstellen, wurden nacheinander verschiedene Aspekte zueinander in Beziehung gesetzt: Zum einen wurden Indikatoren bestimmt, die automatisiert von einer Webseite erhoben werden können, um Aussagen über unerwünschte Ereignisse zu treffen, zum anderen abstrakte Risikowerte, die für jede denkbare Webseite eine Risikoabschätzung möglich machen.

Automatisiert auslesbare Indikatoren sind deshalb nötig, weil bei dem hier verfolgten Ansatz kein anbieterseitig eingebundenes System [LBB14] den Verbraucher- und Datenschutz unterstützen, sondern rechtliche Risiken automatisiert eingeschätzt werden. Solche Indikatoren sind bspw. das Vorhandensein, die Position und die Formatierung von Informationen, die rechtlich vorgeschrieben sind. Diese finden sich etwa oft in Datenschutzerklärungen. Aber etwa auch die Allgemeinen Geschäftsbedingungen enthalten automatisiert auslesbare Indikatoren [Bo14].

Anhand dieser Indikatoren können zu vielen unerwünschten Ereignissen bei der Nutzung von Webseiten Aussagen im Bezug auf den Verbraucher- und Datenschutz getroffen werden. Die unerwünschten Ereignisse ergeben sich aus den Gesetzesbegründungen, Rechtsprechung und Literatur. So sollen Verbraucher z. B. in einem Web-Shop davor geschützt werden, aus Versehen einen Vertrag abzuschließen („Abgabe einer übereilten Willenserklärung“). Verschiedene Normen versuchen diese unerwünschten Ereignisse zu verhindern.

Das Abschätzen der Wahrscheinlichkeiten von unerwünschten Ereignissen ist nur auf einem abstrakten Level möglich und muss Einzelfälle außer Acht lassen. Jedem unerwünschten Ereignis wurden deshalb im ersten Schritt alle Normen untergeordnet, die den Eintritt des Ereignisses verhindern sollen. Die Normen sind zueinander danach gewichtet, wie sehr sie vor dem unerwünschten Ereignis schützen.

In einem zweiten Schritt wurden jeder Norm automatisiert auslesbare Indikatoren zugeordnet, die Aufschluss darüber geben, ob die Normen eingehalten werden. Zu den einzelnen Indikatoren wird entsprechend ihrer Bedeutung für die jeweilige Norm wiederum eine Gewichtung, im Folgenden Indikatorwert genannt, festgelegt. Ein Beispiel: Verschiedene Normen – vor allem Informationspflichten – sollen davor schützen, dass ein Nutzer aus Versehen einen Vertrag abschließt. Besonders wichtig ist dabei die sogenannte Button-Lösung[1]: Um auf der Webseite eines Unternehmers einen Vertrag abzuschließen, durch den ein Verbraucher zur Zahlung verpflichtet wird, muss dies auf dem Button deutlich erkennbar werden, z.B. durch „jetzt kaufen“. Ob diese rechtliche Anforderung eingehalten wurde, kann aufgrund der strengen Vorgaben der Vorschrift weitgehend anhand der Beschriftung, der Formatierung und der Position des Buttons erkannt werden

[1] § 312g Abs. 3 BGB, ab dem 13.6.2014 § 312j Abs. 3 BGB n. F.

[BBV14]. Die Beschriftung ist hierbei am wichtigsten und erhält daher die höchste Gewichtung, d.h. den höchsten Indikatorwert.

Für jede Webseite kann anhand der Gewichtung der jeweiligen Norm sowie der Indikatorwerte die Wahrscheinlichkeit für jedes unerwünschte Ereignis eingeschätzt werden: Dazu werden die sich aus der Webseite ergebenden Indikatorwerte pro relevanter Norm addiert und mit der jeweiligen Normgewichtung multipliziert. Die gewichteten Gesamtwerte ergeben addiert das Risiko für das jeweilige unerwünschte Ereignis.

Als Ergebnis der Berechnung wird für eine konkrete Webseite die Liste der unerwünschten Ereignisse jeweils einem bestimmten quantitativen (z.B. „85 von 100 Punkten“) und qualitativen Wert („hohes Risiko“) zugeordnet.

2.2 Technische Analyse

Die technische Analyse basiert auf den Konzepten der qualitativen Risikoanalyse nach ISO 27005 [DIVSI12] und NIST 800-30 [SGF02] und besteht aus den zwei Komponenten, Risikomodell und Risikoabschätzung. Das Modell setzt sich wie folgt zusammen:

Ein Bedrohungsgraph stellt die Zusammenhänge zwischen Nutzeraktionen, technischen Bedrohungen und den daraus resultierenden unerwünschten Ereignissen dar. Zum Beispiel kann das „Senden von persönlichen Daten“ zu den Bedrohungen „Abhören“ und/oder „Phishing“ führen. Dies ermöglicht wiederum „Unautorisiertes Erfassen von Daten“. Dadurch ergeben sich die unerwünschten Ereignisse wie „Unautorisierte Zahlung“, „Unautorisierte Überweisung“ oder „Spam“, welche für den Benutzer von Bedeutung sind.

Aus dem Bedrohungsgraphen lässt sich in Form von Angriffspfaden ableiten, wie bestimmte unerwünschte Ereignisse eintreten können. Von einem unerwünschten Ereignis rückwärts gehend, kommt man zu verschiedenen Schwachstellen. Verschiedene Angriffspfade können daher zu denselben unerwünschten Ereignissen führen. Da sich die Wahrscheinlichkeit des Eintretens einer Bedrohung allerdings je nach Angriffspfad stark unterscheiden kann, werden diese durch sogenannte Szenarien unterschieden und die Wahrscheinlichkeiten für alle Angriffspfade getrennt ermittelt.

Im Beispiel von Abbildung 2.2 auf Seite 5 kann sowohl „Abhören“ als auch „Phishing“ zu den unerwünschten Ereignissen „Unautorisierte Zahlung“ und „Unautorisierte Überweisung“ führen. Während beim „Abhören“ das Szenario „Ungeschütztes Senden von Login-Daten“ schon ausreicht, eines der beiden unerwünschten Ereignisse hervorzurufen, ist beim „Phishing“ das „Ungeschützte Senden von Login-Daten an eine nicht vertrauensvolle Webseite“ erforderlich.

In der Risikoabschätzung folgt die Anwendung des Risikomodells auf eine konkrete Situation in Abhängigkeit von Indikatoren. Das Ergebnis liefert eine Liste von unerwünschten Ereignissen mit entsprechenden Risikowerten für die jeweilige Situation.

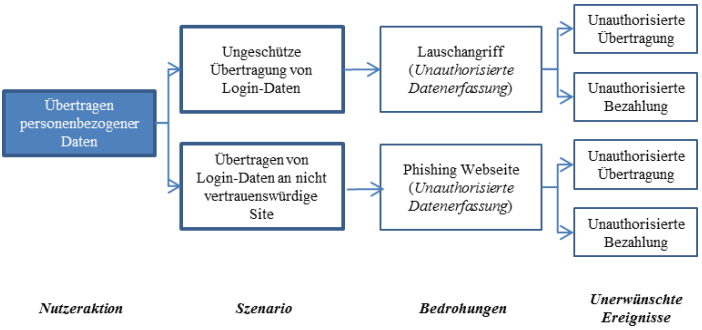


Abbildung 2.2: Beispiel eines Bedrohunggraphs

Zur Abschätzung des Risikos, oder zumindest der Wahrscheinlichkeit, einer bestimmten Situation, beschrieben über Indikatoren, werden Entscheidungsbäume aufgebaut. Die Entscheidungsknoten eines solchen Baumes stellen die Indikatoren dar, ausgehende Kanten jeweils mögliche Zustände der Indikatoren. Die Ergebnisse repräsentieren die Eintrittswahrscheinlichkeiten und je nach Ereignis auch eine zu erwartende Schadenshöhe.

In Abbildung 2.3 auf Seite 5 ist ein Entscheidungsbaum für die Eintrittswahrscheinlichkeit des unerwünschten Ereignisses „Unautorisierte Überweisung“ im Szenario „Ungeschütztes Senden von Login-Daten“ dargestellt.

Jedes Szenario und unerwünschte Ereignis muss einzeln betrachtet werden, da sich die Eintrittswahrscheinlichkeiten unterscheiden. Zusätzlich werden je Szenario die Wahrscheinlichkeiten der jeweiligen Aktion, die zu dem unerwünschten Ereignis führen, ebenfalls über einen Entscheidungsbaum bestimmt. Ist bspw. ein Passwort-Feld angeklickt, so ist die Aktionswahrscheinlichkeit, dass ein Passwort eingegeben wird „hoch“.

Das Risikomodell wird auf die jeweilige Situation angewendet, um das Risiko abzuschätzen. Für jedes einzelne Szenario werden Listen mit unerwünschten Ereignissen und deren Risikoabschätzung bestimmt.

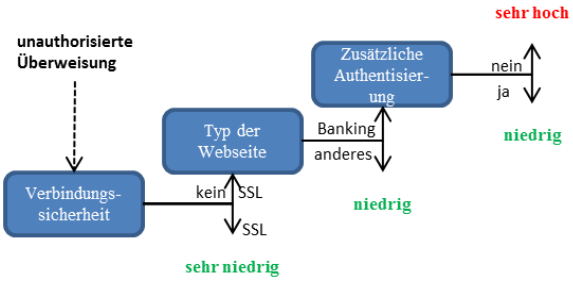


Abbildung 2.3: Beispiel eines Entscheidungsbaums (Ausschnitt)

Die Risikoabschätzung je Situation wird einmalig im Vorfeld in Anlehnung an die Prozess-FMEA [TC02] nach folgender Regel erstellt:

$$Risiko_{Ereignis} = Wahrscheinlichkeit_{Ereignis} * Wahrscheinlichkeit_{Aktion} * Schadenshöhe_{Ereignis}$$

Dabei werden die Wahrscheinlichkeiten sowie die antizipierten Schadenshöhen den jeweiligen Entscheidungsbäumen entnommen. Ergebnis ist eine Tabelle, der man je Situation und Aktion das antizipierte Risiko entnehmen kann.

Bei der Risikobewertung sind Fehleinschätzungen nicht auszuschließen. Eine vertrauenswürdige Webseite könnte vom Bewertungstool bspw. als nicht vertrauenswürdig eingestuft werden. Zudem ist es nicht möglich alle Risiken automatisiert zu messen. Negative Nutzerbewertungen über die betroffene Webseite auf einer anderen Webseite werden nicht in die Risikomessung einbezogen. Weitere nicht voll-automatisch und mit 100% iger Genauigkeit messbare Risikofaktoren sind z.B. Phishing oder Cross-Site-Scripting.

2.3 Zusammenführung der Risiken

Nachdem die rechtlichen und technischen Risiken abgeschätzt wurden, müssen die Ergebnisse zusammengeführt werden. Hierzu werden die beiden entstandenen Tabellen zu einer zusammengeführt. Hierbei ist es möglich, dass es zu doppelten Einträgen kommt, d.h. es gibt dasselbe unerwünschte Ereignis im technischen und im rechtlichen Bereich. In diesem Fall wird der Eintrag mit dem höheren Risikowert übernommen und der andere Eintrag wird entfernt.

Szenario	**unerwünschtes Ereignis**	**Risiko**
ungeschützte Übertragung von Login-Daten auf Banking-Seite	unautorisierte Bezahlung	sehr hoch
Übertragung von Login-Daten auf Phishing-Seite	unautorisierte Bezahlung	sehr hoch
Aufrufen einer Seite mit Drittanbietercookies	unerwünschte Profilbildung	mittel

Tabelle 1: Auszug der resultierenden Risikotabelle

3 Individualisierung auf den Nutzer

Um die Unterstützung der Nutzer auf ihr jeweiliges Verhalten im Internet sowie auf ihre IT-Sicherheitsexpertise abzustimmen, wird direkt während der Installation des Bewertungstools ein Risikoprofil des Nutzers erstellt sowie seine IT-Sicherheitsexpertise festgestellt.

3.1 Risikoprofil

Das Risikoprofil erfasst für den individuellen Nutzer die Wahrscheinlichkeit, dass sich diese Person risikoreich verhält. Risikoreiches Verhalten wurde indirekt per Fragebogen gemessen. Die abgefragten Szenarien stammten aus der DOSPERT-G (Johnson et al., 2004) und erfassten jeweils Wahrscheinlichkeit, Risikograd und Nutzen des potenziell risikoreichen Verhaltens. Die u.g. Items zur persönlichen Einstellung entstammen dem Bericht des Deutschen Instituts für Vertrauen und Sicherheit im Internet [DIVSI12] und wurden vom dort verwendeten Interview auf ein Fragebogendesign angepasst. Aufgrund einer Vorerhebung mit insgesamt 386 Probanden fließen folgende Merkmale ein, die mittels linearer Regression als signifikante Prädiktoren ermittelt wurden:

- Alter (16-100)
- Geschlecht (dichotom)
- Internet-Nutzungshäufigkeit (4-fach gestuft zw. „täglich“ und „seltener als monatlich“)
- Internet-Nutzungsdauer (4-fach gestuft zw. „bis zu 3 Jahren“ und „seit über 10 Jahren“)
- Besitz internetfähiger Geräte (0-5)
- Persönliche Einstellung zum Datenschutz (3 Fragen, 5-stufig Likert-skaliert von „stimme gar nicht zu“ bis „stimme vollkommen zu“):
    - *„Wir müssen uns an einen freieren Umgang mit Daten im Internet gewöhnen.“*
    - „Mir persönlich ist es egal, was mit meinen Daten im Internet geschieht.“
    - „Neue Angebote und Entwicklungen im Bereich Internet probiere ich immer sofort aus.“

Die jeweiligen Werte der Nutzer werden in die Regressionsgleichung eingesetzt und diese basierend darauf entweder in die Risikogruppe „hoch“ oder „nicht hoch“ eingeteilt. Die Einteilung in 2 Klassen erfolgte aufgrund projektinterner Überlegungen. Für eine weitere Verwendung kommt aber auch eine Einteilung in mehr Klassen, sowie eine gewichtete Verwertung anhand der Regressionsdaten in Frage. Anhand des Risikoprofils kann prognostiziert werden, ob sich eine Person potenziell eher risikoreich verhält. Es erfasst aber nicht, warum eine Person ein Risiko eingeht und ob sie sich dessen bewusst ist. Dazu wird die IT-Sicherheitsexpertise herangezogen.

3.2 IT-Sicherheitsexpertise

Die IT-Sicherheitsexpertise wird anhand eines weiteren Fragebogens erfasst, der u.a. folgende Fragen beinhaltet:

- Was ist die beste Definition für Cookie?
- Stellen Sie sich eine Internetseite vor, welche normalerweise nur von Ihnen und Ihren Freunden genutzt wird, aber grundsätzlich öffentlich zugänglich ist (z.B. ein Forum mit Fußballfreunden). Welche Gefahren ergeben sich, wenn Ihre E-Mail-Adresse zusammen mit Ihrem Namen auf dieser Seite im Internet veröffentlicht wird?

Die Antwortmöglichkeiten sind jeweils im Multiple-Choice-Format. Basierend auf den Antworten wird die IT-Sicherheitsexpertise des Nutzers bestimmt. Der Nutzer wird aufgrund dieses Wertes in die Gruppe „hoch“ oder „nicht hoch“ eingeteilt. Eine höherauflösende Einteilung wurde aus Aufwandsgründen vorerst abgelehnt, um die zweistufige Lösung evaluieren zu können.

4 Interventionsstrategien

Aufgrund der berechneten Werte (Gesamtrisiko, Risikoprofil, IT-Expertise) wird die Intervention angepasst an den einzelnen Nutzer selektiert.

4.1 Interventionsarten

Ziel ist es, Nutzer effektiv und gleichzeitig verständlich zu warnen. Dies wird mit einem Konzept aus aktiven und passiven Warnungen plus einer permanenten Statusanzeige für Risiken auf der momentan besuchten Seite verfolgt. Aktive Warnungen stoppen den Benutzer in seinem Surfverhalten. Er kann erst nach einem Klickdialog fortfahren. Passive Warnungen informieren dagegen den Nutzer über ein etwaiges Risiko (z.B. durch ein im unteren Bildschirmbereich erscheinendes Warnfenster), ohne ihn in seinem Surfverhalten zu stören. Die permanente Statusanzeige soll dauerhaft sichtbar sein und einen generellen Zustand der besuchten Seite mithilfe eines Icons (z.B. einer Ampel) darstellen, selbst wenn kein Anlass zu einer höherrangigen Intervention technischer oder rechtlicher Ebene besteht. Dieses Icon ist ortsfest und dient dem Nutzer als Möglichkeit ein schnelles Feedback über den Status der Seite zu erlangen.

Besonderes Augenmerk wird auf die Verständlichkeit der Nutzerhinweise gelegt. Der Nutzer wird nicht, wie bisher üblich, mit einem für ihn oft unverständlichen Hinweis auf die technischen Mängel (z.B. ungültiges Zertifikat) alleine gelassen. Er bekommt konkrete, von seiner Expertise abhängig formulierte Informationen über die Bedrohung(en), deren mögliche Konsequenzen und Empfehlungen zur weiteren Vorgehensweise. So wird bei Nicht-hoher Expertise z.B. bei Auffinden von Tracking-Cookies in der Warnmeldung darauf hingewiesen, dass auf dieser Site das Nutzerverhalten verfolgt wird. Bei

Nicht-hoher Expertise findet immer ein Trade-off zwischen Verständlichkeit und Informationsgehalt statt, die exakte Information (d.h. die, die auch der Nutzer mit hoher Expertise erhalten würde) ist aber jederzeit umschaltbar.

4.2 Entscheidungsalgorithmus zur Intervention

Die jeweilige Entscheidung zu den Interventionsarten Statuszeile, passive Intervention und aktive Intervention wird in der Abbildung 4.1 Nutzer mit nicht hoher und für Nutzer mit hoher IT-Sicherheitsexpertise dargestellt.

Ein Nutzer mit nicht hoher Expertise und hohem Wert im Risikoprofil erhält derzeit bei Benutzung des Tools selbst bei niedrigem Risiko aktive Warnungen. Nutzer mit hoher IT-Sicherheitsexpertise werden dagegen nur selten aktiv gewarnt und so seltener unnötig „behelligt“. Es ist jedoch in Planung diesen Algorithmus in Benutzerstudien noch weiter zu erproben und die gewonnenen Einsichten in den Algorithmus zu integrieren sowie die Ergebnisse zu publizieren.

a) bei nicht hoher IT-Sicherheitsexpertise b) bei hoher IT-Sicherheitsexpertise

Abbildung 4.1: Interventionswahl

Die Grenze, wann ein Nutzer eine aktive oder passive Warnung erhält, kann sich, abhängig von der auftretenden Dichte der Warnungen im Feldversuch, noch verschieben. Ebenfalls in Iteration befindet sich eine Variante des Algorithmus, bei der die Anzahl der Risiken, die selbst eine Intervention hervorrufen würden, in den Inhalt der gezeigten Intervention einfließt, sodass eine Website, die z.B. 5 niedrige und ein mittleres Risiko hat, einen höheren Risikowert erhält, als eine Seite mit nur einem mittlerem Risiko.

5 Anwendungsfallbeispiele

Im Folgenden werden zwei Fallbeispiele zur Verdeutlichung der Entstehung der verschiedenen Warnungen gegeben.

5.1 Aufrufen eines Webshops oder eines Forums ohne HTTPS

Situation: Ein Verbraucher klickt auf einen Link zu einem Webshop oder Forum

Bedrohungshintergrund: Technisches Problem mit der Verbindungsverschlüsselung, z.B. die Webseite verwendet kein HTTPS

Unerwünschtes Ereignis: Abhören des Passwortes (im Fall eines Logins), daraus resultierend: Bestellung (Webshop) bzw. Beiträge (Forum) im Namen des Verbrauchers

Indikatoren: Verbindungssicherheit: kein HTTPS, Typ der Webseite: Webshop/Forum

Resultierendes Risiko: Medium

Nutzer: hohes Risikoprofil, hohe IT-Sicherheitsexpertise

Resultierende Warnung: passiv (siehe Abbildung 5.1)

Abbildung 5.1: Anwendungsfall 1, Nutzer: hohe Expertise, hohes Risikoprofil

5.2 Aufrufen einer Seite mit Drittanbietercookies

Situation: Ein Verbraucher gibt eine URL zu einem Webshop ein. Der Benutzer führt auf der Webseite weitere Aktionen aus (bspw. Klicks)

Bedrohungshintergrund: Verwendung von Drittanbieter-Cookies mit langer Speicherzeit ohne den Verbraucher zu informieren

Unerwünschtes Ereignis: Profilbildung

Indikatoren: Typ der Webseite: Webshop, Web-of-Trust Bewertung: negativ (Datenschutz), Drittanbieter-Cookies: vorhanden (Speicherzeit: 5 Jahre), Datenschutzerklärung: vorhanden (Hinweis auf eigene Cookies, aber nicht auf die von Drittanbieter)

Resultierendes Risiko: **Mittel**

Nutzer: nicht hohes Risikoprofil, nicht hohe IT-Sicherheitsexpertise

Resultierende Warnung: aktiv (siehe Abbildung 5.2 unten)

Abbildung 5.2: Anwendungsfall 2, Nutzer: nicht hohe Expertise, nicht hohes Risikoprofil

6 Fazit und Ausblick

Das vorgestellte Bewertungstool[2] basiert zunächst auf einer Einschätzung des tatsächlichen Risikos für unerwünschte Ereignisse aus dem Bereich der IT-Sicherheit sowie des Verbraucher- und Datenschutzes auf einer bestimmten Webseite. Zusätzlich wird die Interventionsart an den einzelnen Nutzer angepasst, indem diese von seinem antizipierten Verhalten in risikoreichen Situationen (Risikoprofil) und seinem Verständnis von Risiken (IT-Sicherheitsexpertise) abhängig gemacht wird. Er wird genau soweit unterstützt, wie er es aufgrund seines Risikoprofils, seiner Einstellung zu Gefahren sowie seiner Expertise benötigt. Das hinlänglich bekannte „genervte Wegklicken“ von Sicherheitshinweisen, die zu häufig oder zu unverständlich sind, soll so vermieden werden.

Die interdisziplinäre Webseitenanalyse bezüglich IT-Sicherheit sowie Verbraucher- und Datenschutz (beschränkt auf den deutschen Raum) erweitert die bisherigen Angebote zur Nutzerunterstützung: Durch die Disziplinen übergreifend gleichförmige Bewertung sind die Ergebnisse der technischen und rechtlichen Analyse vergleichbar und der Nutzer wird über die technischen Problematiken hinaus auch im Daten- und Verbraucherschutz unterstützt. Da sowohl der rechtliche als auch technische Algorithmus auf Annahmen und Wahrscheinlichkeiten basiert, kann nicht sichergestellt werden, dass die Ergebnisse in jedem Einzelfall korrekt sind. Im weiteren Verlauf sind deshalb haftungsrechtliche Fragen und Möglichkeiten von Haftungsausschlüssen zu untersuchen.

Das Bewertungstool befindet sich derzeit sowohl in der Entwicklungs- als auch in der Evaluationsphase. Das Feedback von vielen Nutzern wird sicherlich sowohl weitere Verbesserungen in den Nutzerhinweisen als auch in den Einstufungen der Interventionskategorie bringen.

Weitere, hier nicht beschriebene, Arbeiten können auf der Projektseite www.secuso.informatik.tu-darmstadt.de/de/research/projects/projekt-inuse/[3] nachgelesen werden.

7 Acknowledgement

Der Beitrag entstand im Rahmen des Forschungsprojekts „Benutzerunterstützung zur Bewertung der Vertrauenswürdigkeit von Webseiten und Webshops (Internet Usage Support: InUse)“. Die Förderung des Vorhabens erfolgt aus Mitteln des BMEL auf Grund eines Beschlusses des Deutschen Bundestags. Die Projektträgerschaft erfolgt über die Bundesanstalt für Landwirtschaft und Ernährung (BLE).

[2] Das Bewertungstool kann auch auf einem abgesicherten Medium installiert werden. Dies wurde bereits in Zusammenarbeit mit der Firma KOBIL Systems GmbH prototypisch entwickelt. In diesem Paper steht das Bewertungstool im Vordergrund.
[3] Zugriffsdatum: 17.06.2014

8 Literaturverzeichnis

[BBV14] Boos, C.; Bartsch, S.; Volkamer, M.: Rechtliche und technische Nutzerunterstützung bei der Button-Lösung – Ein Lösungsvorschlag zur Erkennbarkeit von Kostenfallen als dem immer noch ungelösten Kernproblem. In: Computer und Recht (2014), S. 119-127

[Bo14] Boos, C., Nutzerunterstützung durch automatisierte Auswertung einzelner standardisierter Vertragsbedingungen, AGB: Was habe ich verpasst?. In: Verbraucher und Recht (2014), S. 47-52

[DIVSI12] Deutsches Institut für Vertrauen und Sicherheit im Internet. (2012). DIVSI Milieu-Studie zu Vertrauen und Sicherheit im Internet. DIVSI.

[II08] ISO/IEC 27005:2008: Information Technology – Security techniques – Information security risk management. ISO, Geneva, Switzerland (2008)

[Ka14] Karpf, A.-C.: Web Genre Classification for Context-Specific Security Warnings. Technische Universität Darmstadt, Master Thesis in Informatik (2014)

[LBB14] Luhn, S.; Bruns, I.; Böhme, R., Consumer Participation in Online Contracts, Exploring Cross-Out Clauses, Proceedings GI-SICHERHEIT (2014), S. 255-266

[SGF02] Stoneburger, G.; Goguen, A.; Feringa, A.: Risk Management Guide for Information Technology Systems – NIST Special Publication 800-30. Tech. Rep., National Institute of Standards and Technology (2002)

[TC02] Theden, P; Colsman, H.: Fehlermöglichkeits- und -einflussanalyse (FMEA). In: Qualitätstechniken. Werkzeuge zur Problemlösung und ständigen Verbesserung (Pocket Power). 3. Auflage, Hanser 2002. S. 78-89.

[WBB02] Weber, E. U., Blais, A. R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. Journal of behavioral decision making, 15(4), 263–290.

Ein erster Prototyp: Sicherheitsguide für Grundschulkinder beim Umgang mit dem Internet

Jana Fruth, Matthias Thimm, Sven Kuhlmann, Jana Dittmann
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
PO Box 4120, D-39016 Magdeburg
fruth@ovgu.de, matthias.thimm@st.ovgu.de, stuchsch@ovgu.de, jana.dittmann@ovgu.de

Abstract: Bereits Grundschulkinder nutzen das Internet regelmäßig. Dabei sind sie häufig Sicherheitsgefahren ausgesetzt, mit denen sie nicht umgehen können. In diesem Beitrag wird ein Konzept und Prototyp eines softwarebasierten Sicherheitsguides vorgestellt, der 6 bis 10 jährige Grundschulkinder für potentielle Sicherheitsgefahren im Internet sensibilisieren und ihnen Handlungskompetenzen für Sicherheitsmechanismen vermitteln soll. Der Prototyp wurde mit einer Nutzerstudie in einer Grundschule mit einer eigenen Methodik evaluiert. Die vermuteten Lerneffekte sind allerdings in Zukunft noch mit weiteren Tests zu belegen.

1 Einführung und Motivation

Das Internet nimmt in unserer Gesellschaft einen immer höheren Stellenwert ein. Internetnutzer sind in sämtlichen Altersgruppen zu finden. Diverse Studien belegen, dass vor allem der Anteil der Nutzer im Kindesalter stetig zunimmt. Laut [BR12] surfen ca. 40% der 6 bis 10-Jährigen mindestens einmal pro Woche im Internet. Dabei sind gerade Grundschulkinder wenig für Onlinegefahren sensibilisiert, wie verschiedene Studien [BR12][LHGO11][KHFD12] einstimmig belegen. Ursachen für das Eingehen von Sicherheitsrisiken im Internet von jungen Schulkindern sind u.a. das mangelnde Bewusstsein für Onlinegefahren und die mangelhafte Fähigkeit, kritische Situationen zu bewältigen oder abzuwehren [OEC11]. Beim Surfen im Internet können Kinder verschiedenen Sicherheitsgefahren ausgesetzt sein. Die häufigsten sicherheitskritischen Online-Aktivitäten von Kindern sind: das Veröffentlichen persönlicher Informationen, der Zugriff auf nicht vertrauenswürdige Webseiten oder Inhalte, das Herunterladen von Inhalten nicht vertrauenswürdiger Webseiten und das Chatten mit Fremden. Potentielle sicherheitskritische Konsequenzen können beispielsweise der Missbrauch persönlicher Informationen durch Fremde, Cybermobbing oder das Installieren von Schadsoftware sein [FSRD13].

Für die spätere Entwicklung im Umgang mit dem Internet ist es daher förderlich, Kindern schon im frühen Alter das Bewusstsein über die Gefahren des Internets zu vermitteln. Weiterhin sollten ihnen Hilfestellungen für die Handhabung von Sicherheitsmechanismen gegeben werden. Diese Arbeit ist nicht allein von den Eltern oder Lehrern zu bewerkstelligen, die sich oft im Hinblick auf die medienerzieherische Sicherheit im Internet, nicht

genügend informiert fühlen [KHFD12]. Derzeit gibt es nur wenige medienpädagogischen Angebote zur Sensibilisierung und dem Vermeiden von Onlinegefahren in Grundschulen. Lehrer und Medienpädagogen bestätigten hier einen hohen Bedarf, der bisher nicht durch vereinzelte Projekte gedeckt werden kann. In diesem Beitrag wird daher ein zum Unterricht alternativer Ansatz vorgestellt. Mit Hilfe eines so genannten softwarebasierten Sicherheitsguides[1], sollen Kinder im Grundschulalter für den sicheren Umgang mit dem Internet sensibilisiert werden und den Umgang mit Sicherheitsmechanismen erlernen. Der Sicherheitsguide soll hierbei Lehrer nicht ersetzen, sondern ist als zusätzliches Hilfsmittel für den Unterricht zu verstehen.

Folgende **Forschungsfragen** sollen dabei beantwortet werden: Welches neue Wissen und welche Handlungskompetenzen von Sicherheitsmechanismen bei Internettätigkeiten konnten Kinder mit Hilfe des Sicherheitsguides erwerben? Wie wird das multimediale Design des Sicherheitsguides von den Kindern bewertet? Eignet sich die kreierte Evaluationsmethodik, um Lerneffekte bei der Handhabung des Sicherheitsguides messen zu können?

2 State of the art: Sensibilisierung von Grundschulkindern für Onlinegefahren und Vermittlung sicherheitsbewussten Verhaltens

Im Folgenden werden verschiedene Konzepte und eigene Vorarbeiten vorgestellt, die Grundschulkinder für Onlinegefahren sensibilisieren sollen bzw. ihnen sicherheitsbewusstes Verhalten vermitteln helfen.

Es existieren zahlreiche Konzepte zur **Sicherheitssensibilisierung/-schulung von Grundschulkindern**[2]. Zum Einen wird ***Informationsmaterial*** von verschiedenen Initiativen zur Verfügung gestellt, z.B. vom deutschen Kinderhilfswerk [FO09]. Diese Ratgeber enthalten kindgerecht aufgearbeitete Informationen für Kinder zu Sicherheitsgefahren im Internet und Tipps für ein sicheres Verhalten im Netz. Weiterhin gibt es eine Vielzahl von ***Onlineangeboten***, die Kindern spielerisch Onlinegefahren und Sicherheitsmechanismen vermitteln, z.B. die Trickfilme des EU-Projekts "Sheeplive"[3]. Es gibt zum Thema auch vereinzelte ***Lernangebote*** in Grundschulen, die aber abhängig vom Bundesland bzw. Bildungsträger sind. Weiterhin bieten auch außerschulische Einrichtungen, wie die Polizei und Medienpädagogen Schulungen für Kinder an. Besonders die Polizei offeriert hier ein umfassendes Lehrangebot zur Internetsicherheit [Pol10]. Diese Angebote sind allerdings nicht für Grundschüler, sondern für ältere Kinder (ab ca. 12 Jahren) konzipiert. Die bisher vorgestellten Konzepte haben aus psychologischer Perspektive einen Nachteil: Die Informationen werden abstrakt, also getrennt von der eigentlichen Internetaktivität vermittelt. Es wird so genanntes Faktenwissen [And76] gelehrt. Unserer Ansicht können bessere Reaktionen erzielt werden, wenn Kindern in der aktuellen Situation im Internet, Hilfestellung zu aktuellen Sicherheitsbedrohungen gegeben und somit Handlungswissen vermittelt wird. Der in diesem Beitrag in Kap. 4 vorgestellte Sicherheitsguide ist ein erster

[1] Der in diesem Beitrag beschriebene softwarebasierte Sicherheitsguide basiert auf der Bachelorarbeit von Matthias Thimm [Thi13].

[2] Hinweis: Aufgrund des begrenzten Platzes können nur einige ausgewählte Konzepte erwähnt werden!

[3] de.sheeplive.eu, letzter Zugriff: 5. Mai 2014

Prototyp, der dieses Konzept umsetzt. Ein weiterer Ansatz ist das ***Konzept des sicheren Surfraums*** [Kal13]. Ziel ist es, Kinder erst gar nicht Onlinegefahren auszusetzen. Dieser Ansatz verhindert den Aufbau von Handlungswissen. Technische Lösungen zur Umsetzung des Surfraums sind kindgerechte Suchmaschinen im Internet, wie fragFINN[4], und Jugendschutzsoftware, wie JusProg[5]. Die Verwendung eines freien Ansatzes könnte Kindern die Möglichkeit geben, sich mit Onlinegefahren auseinanderzusetzen und so direkt am Beispiel nach dem Ursache-Wirkungs-Prinzip zu lernen und Handlungskompetenzen aufzubauen. Der in Kap. 4 vorgestellte Prototyp setzt diesen Lösungsansatz des Sicherheitsguides um.

Eigene Vorarbeiten: In [KHFD12] wird eine eigene *Studie zum Sicherheitsbewusstsein von Kindern* bezüglich ausgewählter Webseiten beschrieben. Evaluationsziele sind die Thematisierung ausgewählter Sicherheitsthemen auf den untersuchten Webseiten und die kinderfreundliche Umsetzung der Inhalte. Die Studie ergab, dass Kinder ein gemindertes Sicherheitsbewusstsein bezüglich der von ihnen oft genutzten Webseiten haben und computererfahrene Kinder Webseiten hinsichtlich ihrer Gefährlichkeit positiver bewerten, als weniger computererfahrene Kinder. [MTF+12] beschreibt ein Konzept *multimedialer Sicherheitswarnungen für Smartphones für Grundschulkinder*, welches exemplarisch für ausgewählte Angriffsszenarien umgesetzt, prototypisch auf einem Smartphone implementiert und in einer Nutzerstudie getestet und evaluiert wurde. Der multimediale kindgerechte Designansatz dient als Basis für die softwareseitige Unterstützung für Kinder beim sicheren Surfen im Internet, wie sie in diesem Beitrag beschrieben wird.

3 Designkonzept: Sicherheitsguide für Grundschulkinder

Im folgenden Kapitel wird das Designkonzept des Sicherheitsguides für Grundschulkinder beschrieben. Zunächst wird auf verschiedene Informationsstufen und Einsatzmodi des Sicherheitsguides eingegangen.

Informationsstufen und Einsatzmodi: Die Bewertung der auf einer Webseite getätigten Aktion des Kindes, erfolgt über die Einstufung in verschiedene ***Informationsstufen***, die zugehörige Ereignisse auslösen. Diese Stufen sind an den Bedrohungsstufen des BSI-Standards 100-2 angelehnt [Bun08]. Diese ereignisbasierten Aktionen werden je nach Informationsstufe vom Sicherheitsguide dem Nutzer auf unterschiedliche Art und Weise zurückgemeldet, zum Beispiel als Information (Hinweise, positives Feedback) oder Warnung (Bedrohungen). Kindern soll dadurch eine schnelle Einschätzung der aktuellen Bedrohungslage ermöglicht werden. Folgende vier Informationsstufen wurden umgesetzt: Die *Hinweisstufe* hat rein informative und unterstützende Funktion. Den Kindern werden nähere Informationen zur Funktionsweise einer Webseite und Sicherheitsmechanismen gegen bestimmte Onlinegefahren erläutert. Die *Schutzstufe* erzeugt positives Feedback, z.B. bei Eingabe eines als sicher eingestuften Passwortes. Die *Bedrohungsstufe mittel* meldet dem Nutzer mittels Warnmeldungen eine Verletzung von einzelnen oder mehreren Sicherheitsaspekten zurück, z.B. im Fall des Betretens eines Anmeldebereiches einer

[4] www.fragfinn.de, letzter Zugriff: 5. Mai 2014
[5] www.jugendschutzprogramm.de, letzter Zugriff: 5. Mai 2014

Webseite eines sozialen Netzwerks, welche als sicherheitsgefährdend gewertet wird. Die *Bedrohungsstufe hoch* weist auf ein erhöhtes Sicherheitsrisiko hin, z.B. bei Offenlegung persönlicher Informationen.
Die Aktivierung der vier genannten Informationsstufen basiert auf häufig durchgeführten Internettätigkeiten [BR12] (s. Tab. 1). Die ersten fünf in Tab. 1 angeführten Aktionen, be-

Tabelle 1: Oft getätigte Aktionen im Internet [BR12].

Tätigkeit	Prozentzahl der Befragten	Altersrahmen der Befragten
Nutzung einer Community-Seite	44%	6-13 Jahre
Nutzung einer Online-Anmeldung	>44%	6-13 Jahre
Nutzung einer E-Mail-Seite	40%	10-11 Jahre
Nutzung einer Chat-Seite	30%	10-11 Jahre
Nutzung einer Suchmaschine	99%	6-13 Jahre
Eingabe eines Passwortes	>44%	6-13 Jahre
Herunterladen einer Datei	36%	10-11 Jahre

schreiben das Betreten einer Internetseite. Der Sicherheitsguide soll die jeweilige Seite erkennen und für den identifizierten Webseitentyp neue eventbasierte Aktivierungen für die verschiedenen Informationsstufen zur Verfügung stellen. Eine Ausnahme bilden die letzten beiden in Tab. 1 aufgelisteten Internetaktionen. Über diese Ereignisse wird permanent, also unabhängig von der aufgerufenen Webseite, gewacht.
Die Informationsstufen ermöglichen es, an die Tätigkeiten des Nutzers angepasste Rückmeldung zu geben. Allerdings sind wiederholt eingeblendete Warnhinweisen zu meiden, um bei den Kindern Frustration und Ablehnung des Sicherheitsguides zu verhindern [Jür03]. Eine Lösung der Problematik wäre die Verwendung verschiedener ***Einsatzmodi***. Eine Aufteilung in einen Betriebsmodus und einen Lernmodus wäre denkbar. Der *Betriebsmodus* soll auf eine weitestgehend unterbrechungsfreie Nutzung des Internets ausgelegt sein. Es sollen nur Warnmeldungen der beiden Bedrohungsstufen angezeigt werden. Zur Gewährleistung eines unterbrechungsfreien Betriebs sollte weiterhin in diesem Modus die Wiederholung einer Nachricht zeitweise blockiert werden können. Der *Lernmodus* hingegen, ist dazu konzipiert, Kindern ihre Handlungen zu erklären und ihnen diese besser zu verdeutlichen. Daher sollten in diesem Modus Meldungen aller Informationsstufen angezeigt werden.

Multimediales Designkonzept: Der in diesem Beitrag beschriebene Sicherheitsguide für Grundschulkinder ist mittels multimedialer Mittel durch visuelle und akustische Designelemente kindgerecht gestaltet worden (s. Abb. 1). Das ***visuelle Design*** ist das primäre Designkriterium, da Informationen auf Rechnern im Allgemeinen visuell vorliegen und im Gegensatz zu akustischen und haptischen Designelementen, permanent verfügbar sind. Das visuelle Feedback des Sicherheitsguides besteht vor allem aus der Einblendung von (Warn-)Hinweisen und einer Helferfigur, die das Kind unterstützend begleitet, sowie einer Informationsansicht, die die derzeitige Gefahrensituation dem Kind rückmeldet. Eine *Helferfigur* wurde gewählt, da Kinder häufig dazu tendieren, Formen eine Seele zuzuordnen. Diese abstrahierte, comichafte Art der Darstellung ist Kindern auch bereits aus Computerspielen und Trickfilmen bekannt und wird dort gut angenommen [MTF+12]. Die Mimik des Helfers eignet sich sehr gut, um dem Kind eine schnelle visuelle Rückmeldung über den aktuellen Bedrohungsgrad zu geben. Die Farbe des Gesichts soll zudem eine von drei

verschiedenen Informationsstufen (s. Abb. 1(a), Helferfigur nach [6]) ausdrücken. Hier bieten sich die Farben der Ampel an, die jedes Grundschulkind bereits kennt und denen es somit automatisch eine Bedeutung zuordnen kann [MTF+12].
Neben der Änderung der Mimik und Hintergrundfarbe soll eine so genannte *Informationsansicht* die aktuelle Sicherheitsgefahrenlage und die Art der Bedrohung Kindern leicht verständlich visualisieren. Falls der Nutzer nähere Informationen wünscht, wird zur Informationsansicht gewechselt. Diese ist als Baum gestaltet (s. Abb. 1(b)). Sie soll Kindern vermitteln, dass analog zum Lebewesen Baum auch der Umgang mit dem Internet eine gewisse Sorgfalt bedarf. Die Blätter des Baumes stellen die sechs Sicherheitsaspekte (Vertraulichkeit, Integrität, Authentizität, Verfügbarkeit, Nichtabstreitbarkeit und Privatspäre) und deren derzeitigen Gefährdungszustand dar. Um diese Aspekte für Kinder verständlicher zu halten, werden hierbei abstrahierte Symbole genutzt, wie zum Beispiel ein Briefumschlag für die Vertraulichkeit. Ist ein Sicherheitsaspekt gefährdet, wird dies durch Anfärbung des entsprechenden Blattes des Baums visualisiert. Durch Klick auf das Symbol des jeweiligen Sicherheitsaspekts werden dem Nutzer nähere Informationen zum derzeitigen Zustand dieses Bereiches angezeigt. Die *Textgestaltung* ist an die Fähigkeiten

(a) Rückmeldung der Informationsstufe mittels Mimik und Farben (links oben: Bedrohungsstufe hoch (rot), rechts oben: Bedrohungsstufe mittel (gelb), unten: Schutzstufe (grün))

(b) Informationsansicht

Abbildung 1: Visuelles Design des Sicherheitsguides

von Grundschülern angepasst. Dazu gehört eine leicht verständliche Sprache und die Verwendung kinderfreundlicher Synonyme und Metaphern [FSRD13]. Es werden lineare, serifenlose Schriftarten verwendet, da sie auf digitalen Ausgabegeräten einfacher lesbar und besser verständlicher sind [Jür03]. Weiterhin wird eine ausreichend große Schriftgröße verwendet, um die Lesbarkeit zu gewährleisten, da Kinder im Grundschulalter noch nicht so viel Routine im Lesen haben wie ältere Kinder.
Das ***akustische Design*** soll das visuelle Design unterstützen. Die Stimmung der Helferfigur spiegelt eine *abstrahierte Sprache aus verschiedenen Tonfolgen* wider, welche aus Videospielen für Kinder bekannt sind und dort gut angenommen werden. Weiterhin werden diverse *Warnsignale* verwendet, die sich je nach der von einer Aktion ausgelösten

[6] http://www.deviantart.com/art/Expression-Chart-Pandora-258457532, letzter Zugriff: 17. Juni 2014

Informationsstufe, in ihrem Klang, ihrer Länge und der Anzahl ihrer Wiederholungen unterscheiden. In der Schutzstufe wird ein belohnendes akustisches Feedback durch eine höhere, kurze, nicht wiederholte Tonreihenfolge zurückgegeben. Das akustische Feedback der beiden Bedrohungsstufen ist im Gegensatz dazu um einiges tiefer und in die Länge gezogen, welches in der Bedrohungsstufe hoch zusätzlich noch wiederholt wird.

4 Prototypische Implementierung des Sicherheitsguides

Im Folgendem werden der Sicherheitsguide für Grundschulkinder und die Testumgebung beschrieben.

Aufbau des Sicherheitsguides: Der Prototyp wurde als browserbasiertes Firefox-Addon realisiert, welcher von den Probanden innerhalb einer geschlossenen Umgebung, in Form eines Offline-HTML-Szenarios, genutzt wird. Die geschlossene Testumgebung war aufgrund datenschutzrechtlicher Anforderungen notwendig geworden. Diese fasst verschiedene Testszenarien auf drei HTML-Seiten zusammen. Während die Probanden in dieser Testumgebung bestimmte Aufgaben erledigen, überwacht der Prototyp die getätigten Aktionen und gibt diesen Feedback. Bei sicherheitskritischen oder sicherheitskonformen Aktionen in der Testumgebung gibt der Sicherheitsguide Rückmeldung in Form von (Warn-) Meldungen. Die Funktionsfähigkeit des Addons ist dennoch außerhalb dieser Testumgebung gewährleistet.

Für die Implementierung des Sicherheitsguides wurden die in Kap. 3 gewählten Designkriterien realisiert. Umgesetzt wurde nur der *Lernmodus*, um Hinweise zu Lerneffekte zu evaluieren. Die Programmstruktur des Sicherheitsguides unterliegt einem modularen Aufbau, welcher eine schnelle Erweiterung und kompletten Austausch von Textinformationen, Validationslisten und Helferfiguren innerhalb weniger Minuten ermöglicht. Innerhalb der Startform werden separat hinterlegte *Validationslisten* geladen. Diese können entweder ganze Internetadressen oder Teilzeichenketten enthalten und können aktuell besuchte Seiten mit Hilfe regulärer Ausdrücke klassifizieren. Es wird validiert, welche Rückmeldung auf der besuchten Seite zu geben ist und welche seitenspezifischen Inhaltsskripte aktiviert werden müssen. Je nach Informationsstufe ändern sich der Text der angezeigten Hinweis- und Warnmeldungen und das optische Erscheinungsbild der Helferfigur. Die Meldungen bieten dem Nutzer zwei Auswahlmöglichkeiten an. Mit einem Klick auf “Okay” schließt sich das Fenster. Mit dem Klick auf “weitere Informationen” öffnet sich die Informationsansicht (s. Abb. 1(b)), welche die derzeitige Sicherheitsgefahrenlage beschreibt.

Die virtuelle **Testumgebung** ist in drei verschiedene, selbst erstellte Webseiten aufgeteilt, um datenschutzrechtlichen Anforderungen zu genügen. Den Probanden werden ausgewählte Abbildungen realer Szenarien präsentiert[7]. Die Aufgabe der Probanden ist es, die auf der Seite geforderten Eingaben zu tätigen. Dabei wird das Kind vom Sicherheitsguide je nach Einhaltung der Sicherheitsanforderungen ein positives oder negatives Feedback auf seine getätigten Eingaben erhalten. Generell gibt der Sicherheitsguides zunächst Ratschläge über das Verhalten auf einer bestimmten Seite. Der Aufbau der Testseiten ge-

[7] Hinweis: Der Sicherheitsguide funktioniert auch für reale Webseiten.

staltet sich folgendermaßen:

Die Testseite der *Registrierung* dient der Vermittlung der Erstellung eines sicheren Passwortes und dem Umgang mit privaten Daten. Die Webseite enthält mit Sternchen gekennzeichnete Pflichtfelder, die ausgefüllt werden sollen, Zusatzfelder und die Verlinkung auf AGBs. Zudem soll der Proband ein Passwort eintragen. Je nach Stärke des Passwortes wird ein passendes Feedback mit Ratschlägen für Verbesserungen angezeigt (s. Abb. 1(a)). Weiterhin erhält das Kind vor dem Absenden der eingegebenen Daten einen Hinweis, seine Eingaben noch einmal zu kontrollieren.

Die *E-Mail*-Testumgebung soll adäquates Verhalten gegenüber unbekannten Absendern von E-Mails vermitteln sowie über die nicht immer gegebene Unverfälschtheit des Inhalts der E-Mail aufklären. Der Proband hat zuerst die Aufgabe eine E-Mail zu verfassen und diese zu versenden. Diese landet dann im eigenen Postfach. Beim Öffnen des Postfachs wird das Kind noch einmal ermahnt nur E-Mails von bekannten Absendern anzunehmen und darauf hingewiesen, dass empfangene E-Mails nicht immer so ankommen müssen, wie sie abgeschickt wurden. So ist unter seinem abgeschickten Text der Spruch: “Hier könnte auch Werbung stehen” zu lesen.

Die *Chat*-Testseite soll den sensiblen Umgang mit persönlichen Informationen vermitteln und über die Anonymität des Chatpartners aufklären. Der Sicherheitsguide kontrolliert während des Programmablaufs die Eingaben der Probanden und die Ausgabe des Gegenübers. Im Chat werden zuerst zwei typische Chatnachrichten gesendet. Die dritte Nachricht fragt nach dem Alter des Probanden. Der Sicherheitsguide warnt hierbei den Probanden, keine allzu persönlichen Informationen (z.B. Adresse, Telefonnummer) an Unbekannte preiszugeben.

5 Evaluationsmethodik

Der softwarebasierte Prototyp (s. Kap. 4) des Sicherheitsguides wird in einem Testszenario von Probanden getestet und evaluiert. Dazu stehen verschiedene **Werkzeuge** zur Auswahl [LE11]. Eine Möglichkeit ist die ***Beobachtung*** der Probanden während der Nutzung des Prototyps. Anhand der Reaktionen der Probanden in bestimmten Situationen, können Rückschlüsse zur *Usability* des Prototyps gewonnen werden. Diese Reaktionen werden vom zuständigen Versuchsleiter in ein Protokoll schriftlich aufgenommen und im Nachhinein ausgewertet. Bei der ***Methode des lauten Denkens*** [LE11] verbalisiert die Testperson ihre Gedankengänge während sie den Prototypen testet. Sie kann in kleinen Gruppen und für ältere Kinder verwendet werden, um das Textverständnis und die Akzeptanz des Designkonzeptes zu ermitteln.

Um einen Wissenszuwachs und letztendlich *Lerneffekte* nachzuweisen zu können, wurde eine ***Fragebogenbefragung*** in Form eines Online-Quizzes durchgeführt [EW12]. Dazu werden mindestens drei Befragungen benötigt. Während der ersten Befragung, wird das technische Vorwissen der Probanden gemessen. Die zweite Befragung erfolgt unmittelbar nach dem Testen des Prototyps, um den direkten Lerneffekt ableiten zu können. Eine dritte Befragung sollte um einige Tage zeitversetzt erfolgen und evaluieren, über welche Zeit-

spanne das gelernte Wissen im Gedächtnis bleibt. Da keine standardisierten Fragebögen für die Evaluation geeignet waren, wurde ein eigener Fragebogen entwickelt. Er setzt sich aus einführenden Fragen zur Person des Testers und vier großen Fragebereichen mit je vier Fragen zu den Themen aus dem Bereich der Internetsicherheit zusammen (s. Abb. 2). Während es bei den persönlichen Fragen verschiedene Antwortarten gibt, hat der zweite Fragebogenteil eine festgelegte Anzahl an Antwortmöglichkeiten in Form einer Ratingskala. Diese wurden mit einem Psychologen erarbeitet. Wenn sich Probanden unsicher bei der Beantwortung einer Frage sind, tendieren sie dazu eine mittlere Antwort zu wählen. Um dies zu umgehen ist es besser, eine gerade Anzahl an Antwortmöglichkeiten vorzugeben (hier: völlig richtig, meistens richtig, meistens falsch, völlig falsch)[8].

Tabelle 2: Inhalte des Fragebogens

Bereich	**Inhalte**
Persönliches	Geschlecht, Alter, Einschätzung der Internetkenntnisse
Allgemein	Vertrauen in Suchmaschinenergebnisse, Gefahreneinschätzung von Downloads
Anmeldung	Vollständigkeit, Korrektheit und Vertraulichkeit der Eingaben, Handhabung der AGBs
Passwort	Länge, Zusammensetzung des Passwortes
E-Mail	Vertraulichkeit, Integrität der E-Mail-Nachrichten, technische Details
Chat	Authentizität des Chatpartners, Vertraulichkeit und Handhabung von Chatinhalten

6 Nutzerstudie: Durchführung und Testergebnisse

Im Folgendem werden die durchgeführte Nutzerstudie in einer Grundschule zur Evaluation des Sicherheitsguide-Prototypen beschrieben und die Testergebnisse erläutert und interpretiert. Die **Nutzerstudie** fand in der Dreisprachigen Internationalen Grundschule Magdeburg mit den Schülern einer 4. Klasse statt. Die Kinder hatten bereits gute technische Vorkenntnisse, da sie seit der 2. Klasse eigene Laptops im Unterricht nutzen. Die Nutzertests wurden auf vier verschiedene Tage und in drei verschiedene Phasen aufgeteilt. Es wurde ein Vortest (Fragebogen), ein Haupttest (Test Prototyp, Fragebogen) und ein Nachtest (Fragebogen) durchgeführt. Nach dem Einholen der Einverständniserklärung war es 12 Schüler/innen möglich, an den Testdurchläufen teilzunehmen. Beim *Vortest* wurde ein Online-Fragebogen von den Kindern in ca. 30 Minuten in ihrem Klassenzimmer selbständig an ihren persönlichen Rechnern ausgefüllt. Ziel war es, das technische Vorwissen der Kinder zu ermitteln, um einen Vergleichswert für spätere Befragungen zu bekommen. Der *Haupttest* wurde ca. eine Woche später an zwei Tagen nachmittags in Gruppen mit vier Kindern durchgeführt. Diese wurden im Werkraum platziert und jeweils in Zweiergruppen von einem Versuchsleiter betreut. Es herrschten für die Kinder bekannte Bedingungen zur Durchführung der Testläufe, was dem Angstabbau diente und eine entspannte Atmosphäre schuf. Schwerpunkt des Haupttests war der Test des Prototyps und die Fragebogenbefragung. Den Probanden wurden, die von ihnen zu bewältigenden Aufgaben

[8]Die mittleren Werte erhielten den Präfix “meistens”, um größere Verwirrungen bei den Kindern zu vermeiden, da die Distanz zwischen “meistens falsch” und “meistens richtig” größer wirkt und somit weniger verwirrt als die einfache Nutzung von richtig und falsch.

schriftlich ausgehändigt. Weiterhin wurde den ihnen erklärt, dass es schwerpunktmäßig nicht auf die Lösung der gestellten Aufgabe, sondern auf die Beobachtung des Helfers ankam. Der Test dauerte ca. 30 Minuten. Es folgte die Beantwortung des bereits bekannten Fragebogens, was ca. 10 Minuten beanspruchte. Anschließend wurden Fragen zur Usability des Sicherheitsguides und der Helferfigur gestellt. Abschließend fand ca. eine Woche später ein viertelstündiger *Nachtest* statt, bei welchem evaluiert wurde, inwiefern sich das erlangte Wissen eingeprägt hat. Hierzu wurde der Fragebogen ein letztes Mal von den Grundschulkindern ausgefüllt.

Testergebnisse: Die in der Grundschule durchgeführte Nutzerstudie sollte, Daten zur Evaluation von Lerneffekten und der Usability des Prototypen erheben.

Der eingetretene *Lerneffekt* wurde mit einem Online-Fragebogen (s. Tab. 2) evaluiert. Um die Antworten beurteilen zu können, war es nötig den Fragen eine gewisse Wertigkeit zu geben. In Zusammenarbeit mit IT-Security-Experten wurden den einzelnen Antwortmöglichkeiten die Werte von 1 (ganz falsch) bis 4 (absolut korrekt) zugewiesen. Den dazwischen liegenden Antworten wurde je nach ihrer Tendenz zur Korrektheit der Antwort, der Wert 2 oder 3 zugewiesen. Es konnten maximal 192 Punkte pro Fragenkomplex und 960 Punkte insgesamt erreicht werden. Den Antworten der anonymisierten Probanden wurde die zugehörige Punktzahl für jede Testphase zugewiesen. Anhand der erreichten Punktzahlen war es möglich das Feedback des Sicherheitsguides zu evaluieren. Eine steigende Punktzahl zwischen dem Vor- und dem Haupttest könnte auf einen Wissenszuwachs hindeuteten. Eine höhere Punktzahl des Nachtests im Vergleich zum Vortest könnte ein Hinweis auf einen eingetretenen, beständigen Lerneffekt sein. Um einen Überblick über die Bewertung des Sicherheitsguides als Ganzes zu bekommen, wurden die einzelnen Fragen innerhalb einer Testphase aufsummiert (s. Abb. 2). Die Zahlen deuten auf einen Wissenszuwachs vom Vor- zum Haupttest hin. Weiterhin weisen die vom Helfer abgedeckten Fragebereiche einen höheren Punktestand zwischen Haupt- und Nachtest auf, was auf Lerneffekte hinweisen könnte. Die statistische Untersuchung der Testergebnisse mittels multivariater Varianzanalyse konnte allerdings keine signifikante Steigerung des Wissens- und Lernzuwachses belegen. Daher wurde die Effektstärke nach Cohen [Bor05] jeweils für die Ausprägungen innerhalb der fünf Kategorien zwischen Vortest und Nachbzw. Langzeittest ermittelt als Indikator für die praktische Relevanz der Ergebnisse bei nicht signifikanten Unterschieden in kleinen Stichprobengrößen. Der Konvention [Bor05] folgend werden diese als: 0,2 kleiner, 0,4 bis 0,5 mittlerer und 0,8 starker Effekt definiert. Hierbei wurden zwei mittlere Effekte gefunden. Der “Passwortschutz” (d=0,46) und in der Kategorie “Chat” (d=0,41). Demnach profitieren die Kinder vor allem in diesen beiden Kategorien am stärksten von den Funktionen des Sicherheitsguides, was hierbei auf ein hohes Lernpotential schließen lässt.

Die erzielten Ergebnisse sind allerdings kritisch zu sehen. Der Erfolg der Wissensvermittlung des Sicherheitsguides hätte durch eine Nutzerstudie mit mehreren Kontrollgruppen besser eingeordnet werden können. Während die Probanden der ersten Kontrollgruppe in keinerlei Hinsicht für Onlinegefahren sensibilisiert worden wären, hätte einer zweiten Kontrollgruppe klassisches Informationsmaterial zu IT-Securitygefahren im Internet ausgehändigt werden können.

Aussagen zur *Usability* des Sicherheitsguides wurden mit der Methode des lauten Denkens, der Beobachtung sowie der Befragung der Grundschulkinder erhoben. Nach Aus-

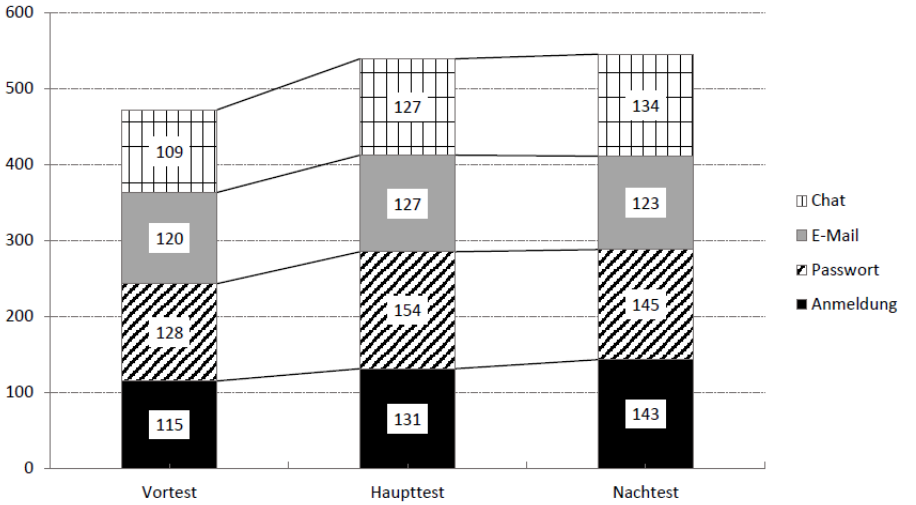


Abbildung 2: Punkteverteilung der vom Helfer behandelten Themengebiete über den Testzeitraum

sagen der Kinder war der Sicherheitsguide problemlos bedienbar. Gute Bewertungen erhielten auch das optische Design, einschließlich der Helferfiguren und der Informationsansicht. Wobei letztere von den Kindern selten angeklickt wurde. Als Lösungsalternative wäre eine Verschmelzung von Helfer und Informationsansicht denkbar.

7 Zusammenfassung und Ausblick

Ziel dieser Arbeit war es, Kinder im Grundschulalter von 6 bis 10 Jahren für den sicheren Umgang mit dem Internet zu sensibilisieren und ihnen Handlungskompetenzen für Sicherheitsmechanismen vermitteln. Restriktive Ansätze bieten hier jedoch keine Lösungen. In dieser Arbeit wurde ein alternativer Ansatz vorgestellt, der diese Ziele mittels eines softwarebasierten Sicherheitsguides umsetzt. Der Prototyp wurde mit einer Nutzerstudie in einer Grundschule evaluiert. Aus Datenschutzgründen wurde dazu ein virtuelles Testszenario mit realen Anwendungsfeldern aus dem Internet realisiert, in welchem die Kinder verschiedene Aufgaben zu lösen hatten (z.B. Anmeldung in einem fiktiven Portal). Die Evaluierung ergab, dass Grundschulkinder für Internetgefahren sensibilisiert werden sollten. Die Testergebnisse lassen vermuten, dass der Einsatz eines softwarebasierten Sicherheitsguides als Ergänzung zum Unterricht sinnvoll sein könnte, was allerdings mit weiteren Studien zu belegen ist.

In **Zukunft** ist das Design des Sicherheitsguides noch zu erweitern bzw. anzupassen. Beispiele sind die eindeutigere Formulierung von Informationstexten, die Nutzung mehrerer Helferfiguren und die Optimierung der Interaktion zwischen der Helferfigur und der Informationsansicht. Weiterhin sollten die Sicherheitsbegriffe, wie Vertraulichkeit, expliziter

vermittelt werden. Nähere Erläuterungen der Begriffe sind bisher nur in der Informationsansicht durch Anklicken der Symbole abrufbar. Die bisherige Nutzerstudie fand nur in einem kleinen Rahmen mit 12 Probanden an einer eher technisch-orientierten Schule statt. Um die Evaluationsergebnisse aussagekräftiger zu gestalten, wären Studien mit einer größeren Anzahl an Personen und verschiedenen Bildungshintergründen empfehlenswert. Der Prototyp ist bisher im Lernmodus implementiert und könnte um einen Betriebsmodus erweitert werden. Weiterhin soll der Sicherheitsguide auf andere Geräte (z.B. Mobiltelefone, Roboter) portiert werden. So könnte er Nutzer nicht nur für potentielle Security-Gefahren, sondern auch für Wechselwirkungen zwischen Vorfällen der Security und Safety sensibilisieren und ihnen Handlungskompetenzen für Sicherheitsmechanismen vermitteln.

Wie in anderen Warndomänen (z.B. Automobil, Produktion) spielt auch im Bereich der IT Sicherheit die Resilienz[9] des Adressaten eine vermittelnde Rolle in Bezug auf die Umsetzung der gelernten Inhalte. Hierfür sollten in folgenden Untersuchungen gezielt die Resilienzfaktoren: Emotionale Intelligenz (Problemfixierung vs. Problemlösungsorientierung) oder Kontrollüberzeugungen betrachtet werden. Je nach Ausprägung führen diese besonders auf der Verhaltensebene zur stabilen oder instabilen Anwendung von gelernten Inhalten. Darüber hinaus lassen sich diese psychologischen Prädispositionen aus den Reaktionen auf Warnungen ableiten und somit in die Gestaltung der Warnungen und Hilfestellungen integrieren. Auf diese Weise ließen sich sowohl besonders reaktive Anwendertypen identifizieren und entsprechend eine eher restriktive Umsetzung des Sicherheitsguides realisieren. Dagegen sind besonders vulnerable Anwender durch massive Warnungen schnell verunsichert. Diese sollten durch möglichst sensible Hilfe unterstützt werden. Darüber hinaus sollten gezielt Ressourcen vermittelt werden, die der Problemlösung hilfreich sind. Das bedeutet in erster Instanz eine Ableitung der Kontrollüberzeugung des Kindes in den jeweiligen Bedrohungsszenarien und eine aktive Vermittlung an soziale Ressourcen für Kinder mit sehr geringen Kontrollüberzeugungen. Für Kinder mit (zu) hohen Kontrollüberzeugungen sollten gezielt technische Ressourcen angeboten werden, die in der jeweiligen Situation Hilfestellung beinhalten. Für diese situations- und individuumsadaptive Gestaltung der Unterstützung sollte im Lernmodus eine entsprechende Reaktions- und Bewertungserfassung realisiert werden, aus der dann die oben genannten Informationen erfasst werden.

Danksagung

Wir danken der Dreisprachigen Internationalen Grundschule Magdeburg und Dr. Michael Knuth. Teile der Arbeit von Jana Fruth sind während des ViERforES[10]-Projektes entstanden, welches vom Deutschen Ministerium für Bildung und Forschung (Projektnummer 01IM10002A) finanziert wurde. Die vorliegende Arbeit wurde in Teilen durch die Deutsche Forschungsgemeinschaft unterstützt (Projekt ORCHideas, DFG GZ: 863/4-1).

[9] Unter Resilienz wird die psychische Widerstandskraft verstanden, die es ermöglicht schwierige Lebenssituationen unbeschadet zu überstehen [WW13].

[10] http://www.vierfores.de/, letzter Zugriff: 5. Mai 2014

Literatur

[And76] J. R. Anderson. *Language, memory, and thought*. The Experimental psychology series. L. Erlbaum Associates and Distributed by the Halsted Press Division of Wiley, Hillsdale and N.J and New York, 1976.

[Bor05] J. Bortz. *Statistik für Human- und Sozialwissenschaftler*. Heidelberg : Springer Medizin, 6. Auflage, 2005.

[BR12] P. Behrens und T. Rathgeb. *KIM-Studie 2012: Kinder + Medien, Computer + Internet: Basisuntersuchung zum Medienumgang 6- bis 13-Jähriger in Deutschland*. Medienpädagogischer Forschungverbund Südwest, 2012.

[Bun08] Bundesamt für Sicherheit in der Informationstechnik (BSI). Standard 100-2: IT-Grundschutz Methodology, 2008.

[EW12] W. Edelmann und S. Wittmann. *Lernspychologie*. Weinheim, Beltz, 2012.

[FO09] S. Frank und S. Ostermann. *Kindersache: Der Internet-Guide für Kids*. Dt. Kinderhilfswerk, Berlin, 3. Auflage, 2009.

[FSRD13] J. Fruth, C. Schulze, M. Rohde und J. Dittmann. E-Learning of IT Security Threats: A Game Prototype for Children. In *Communications and Multimedia Security*, Lecture Notes in Computer Science, Seiten 162–172. Springer, 2013.

[Jür03] S. Jürgens. *Diplomarbeit: Evaluation von world-wide-web basierten Benutzungsschnittstellen für Kinder*. Universitat Hamburg, Fachbereich Informatik, 2003.

[Kal13] A. Kallweit. *Ein Netz für Kinder - Surfen ohne Risiko? Ein praktischer Leitfaden für Eltern und Pädagogen*. Service Kinder und Jugend. Bundesministerium für Familie Senioren Frauen und Jugend, Berlin, 10. Auflage, 2013.

[KHFD12] S. Kuhlmann, T. Hoppe, J. Fruth und J. Dittmann. Voruntersuchungen und erste Ergebnisse zur Webseitengestaltung für die Situationsbewusste Unterstützung von Kindern in IT-Sicherheitsfragen. In *Informatik 2012*, Braunschweig, 2012.

[LE11] J. Liebal und M. Exner. *Usability fuer Kids*. Vieweg+Teubner Verlag, Springer Fachmedien Wiesbaden GmbH, 2011.

[LHGO11] S. Livingstone, L. Haddon, A. Görzig und K. Ólafsson. EU Kids Online II - Final Report. 2011.

[MTF+12] W. Menzel, S. Tuchscheerer, J. Fruth, C. Krätzer und J. Dittmann. Design and evaluation of security multimedia warnings for children's smartphones. In *SPIE, Conference Multimedia on Mobile Devices*, Burlingame, Calif., 2012.

[OEC11] OECD. *The Protection of Children Online: Risks Faced by Children Online and Policies to Protect Them*. 2011.

[Pol10] *Im Netz der neuen Medien: Internet, Handy und Computerspiele - Chancen und Risiken für Kinder und Jugendliche*. Polizeiliche Kriminalprävention, Stuttgart, 2010.

[Thi13] M. Thimm. *Bachelorarbeit: Konzeption und prototypische Implementierung eines Sicherheitsguides für Kinder für den Umgang mit dem Internet*. Uni. Magdeburg, 2013.

[WW13] A. Wieland und C.M. Wallenburg. The influence of relational competencies on supply chain resilience: a relational view. In *International Journal of Physical Distribution & Logistics Management*, Jgg. 43, Seiten 300–320, 2013.

"A Bank Would Never Write That!" A Qualitative Study on E-Mail Trust Decisions

Thomas Pfeiffer, Michaela Kauer and Julia Röth
Institute of Ergonomics
Technische Universität Darmstadt, Otto-Berndt-Straße 2
64287, Darmstadt

t.pfeiffer@iad.tu-darmstadt.de; kauer@iad.tu-darmstadt.de; julia.roeth@gmx.de

Abstract: In order to communicate the risk of fraudulent e-mails to users properly, it is important to know which aspects they focus on when evaluating the trustworthiness of an e-mail. To that end, a study was conducted to test predictions derived from a decision model by asking participants how they would react to each of eight e-mails and why. The study confirms results from previous research showing that content as well as visual and linguistic aspects, but also technical aspects such as sender address and link URL are considered by recipients. It also adds new findings like the fact that through experience and education, users form rules such as "A bank will never ask you for account details via e-mail" or the fact that attachments in HTML format or implausible sending times raise suspicions in users. These findings can be used to inform the design of anti-fraud education and user interfaces of e-mail clients.

1 Introduction

One of the most prominent risks of e-mail communication is the exposure to malicious e-mails which aim to trick recipients into giving away sensitive information (phishing), transferring money (e.g. the so-called "Nigerian" or "419" scam [SW11]) or opening malicious attachments. In order to support users in protecting themselves from these risks, they have to be communicated to them either through preventive education or through the user interface of (web, desktop or mobile) e-mail clients. To optimally target these communication efforts, it is important to first learn which aspects users focus on when deciding how to react to an e-mail, in order to identify in which aspects their behavior has to be improved. This paper covers a study which aims to both confirm results from previous research which on this topic and to uncover aspects relevant to users which have not been documented yet.

The paper is structured as follows: Section 2 describes the model we created and which this study is based on, and the research questions derived from it. Section 3 documents the methods used in the study, the results of which are detailed in section 4 and discussed in section 5. Section 6 concludes the paper and offers an outlook on the application of the results and on further research.

2 Background and Research Questions

The study presented in this paper is based on a model to predict recipients' reactions to calls to action contained in online messages, which was presented by Pfeiffer, Theuerling and Kauer [PTK13]. That model is based on results from previous research on user reactions to e-mails and websites from both the social engineering and e-commerce fields (see aforementioned paper for details, as they do not fit into this paper's space limit). It contains several sets of variables that influence recipients' reactions: attributes of the message, attributes of the purported sender, attributes of context, knowledge/experience of recipient, demographic attributes of recipient, and personality traits of the recipient. Using interviews, the study presented in this paper focuses on the attributes of the message and purported sender as well as contextual factors which influence recipients' reactions.

The attributes of the message predicted to influence reactions are URLs of links (including whether they start with https), the subject line, design aspects such as company/brand logos and colors, trust seals, language (including spelling and grammar errors), personalization (such as inclusion of the recipient's name or account number), the presence of a footer with company and legal information, the overall narrative strength of the e-mail's content, the quality of the (additional) information and content provided in the e-mail, a perceived persuasion attempt, and perceived privacy and security protection communicated through the e-mail. The aspects of the e-mail's (purported) sender which are predicted to influence recipients' reactions are the name and address in the "from" field, familiarity with (receiving e-mails from) the sender, as well as the (purported) sender's reputation. Contextual aspects are the contextual plausibility of the e-mail (for example whether an order the message is referring to was actually placed by the recipient). Another contextual aspect, recipient's computer's perceived vulnerability to attacks, is not expected to be of relevance to participants of this study because it presents a fictitious scenario where participants cannot assess the computer's vulnerability.

3 Methods

The study took place in 2013 in a lab at the Technische Universität Darmstadt and consisted of a questionnaire before and interviews after participants were shown examples of authentic and fraudulent e-mails. It was one part of a larger setup which also included eye tracking and the presentation of e-commerce websites (all of which were authentic). Due to space limitations, this paper only reports the results from the interviews. After an introduction, participants filled in an online questionnaire. It contains questions on socio-demographic information (age, gender, educational level and monthly household income) as well as experience with and usage of the Internet and electronic devices, and the familiarity with software used in the study.

After completing the questionnaire, participants were shown eight different e-mails in Portable Document Format in randomized order, each of them twice. PDFs had to be used due to restrictions in the eye tracking software. The set of e-mails consist of four

authentic commercial e-mails and four real-world phishing examples. If they contain links whose URL is not directly visible, it is included at the bottom of the PDF since it is not possible to use mouse-over to see the URL in the PDF. Except for the fraudulent e-mail purportedly from Facebook (written in English), the e-mails are written in German. In the following descriptions, text is translated.

Table 1 and 2 detail the characteristics of the e-mails with regard to the features mentioned in section 2, tables 3 and 4 do the same for the fraudulent ones. URLs are shortened to the domain names (or other domain names contained elsewhere in the URL).

Purp. sender	**comdirect bank**	**Amazon Marketplace**
"From" address	comdirect.postbox@noreply.comdirect.de	Solution 21 - Amazon Marketplace <chp00fxqpq4225@ marketplace.amazon.de>
Subject	PostBox: You have received a new financial report	Receipt-201216235
Link URLs	www.comdirect.de	www.amazon.de/..., www.adobe.com/...
Design/layout	Plain text	Plain text
Language	Professional, no errors	Professional, no errors
Form of address	Impersonal	Impersonal
Signature	comdirect bank AG	Kind regards
Other personaliz.	Last 3 digits of account-no.	None
Footer	Company info with link to website	None
Story	Information about available financial report	Information about invoice incl. number and date (no need for payment implied)
Call to action	Go to personal area on website (no link)	None (implicit: read invoice)
Additional info	Best bank award	Info on Adobe Reader required to read PDF incl. download link
Priv./Sec. info	Info can be unsubscribed from	Info on secure online purchasing, security and anonymity features of Amazon Marketplace e-mail system
Attachment	None	Invoice-201216235-16307.pdf

Table 1: Features of Authentic E-mails

When seeing the e-mails for the first time, participants were instructed to look at them carefully and form an opinion about them (while there eye movement was being tracked). After that, the same e-mail was shown a second time and the experimenter asked them how they would have reacted to the e-mail and for which reasons, as well as which features of it they noticed and whether they have influenced their opinion. The e-mails were shown twice in order to separate the eye-tracking from the interview part but still allow participants to point out the aspects they looked at in the e-mails. Their answers were

Purp. sender	**Zalando**	**Facebook**
"From" address	Zalando Team <team@info.zalando.de>	Facebook <update+kjdmwwjmvi5m@facebookmail.com>
Subject	Your new password at Zalando	Juli, you have unread notifications
Link URLs	www.zalando.de/...	www.facebook.com/login.php/..., www.facebook.com/o.php/...
Design/layout	Logo, fonts, colors, two trust seals	Facebook logo and other icons
Language	Professional, no errors	Informal, no errors
Form of address	Personal	Personal
Signature	Kind regards, your Zalando team	None
Other personaliz.	None	E-mail addr. contained in login link
Footer	Detailed comp. and contact info	Postal address
Story	Link to reset password at user's request	Info about unread notifications
Call to action	Click link, enter new password	Click link to read messages
Additional info	Link only valid for 24h, advertisements	None
Priv./Sec. info	None	Info for unsubscribing from news service
Attachment	None	None

Table 2: Features of Authentic E-mails Part 2

noted down in shortened form by the experimenter. Afterwards, participants were shown the e-commerce websites and filled out a second questionnaire. These parts, however, are not relevant to the results presented in this paper, since they happened after these results were gathered.

Participants were recruited by Julia Roeth from her private and academic social networks. There was no compensation for participating in the study, participation was completely voluntary.

4 Results

4.1 Sample

The study had a total of 34 participants, eight of which were female. The age ranged from 20 to 66, with a median of 25 years. One participant had an intermediate school-leaving certificate, two had a qualification for entrance to universities of applied science. Fifteen

Purp. sender	**Facebook**	**Bank**
"From" address	Facebook administration <heedonghong@ftp3.scmedia.com.hk>	Bob Daslamm <bobdaslamm@gmx.de>
Subject	'You have 1 personal notification from Facebook Administration	Account blocked, assistance required
Link URLs	http://cp990.perso.sfr.fr/...	http://wwwbanksq.biz/?meine.deutsche-bank.de/...
Design/layout	Facebook logo and other icons	Plain
Language	Informal, no errors	Simple, no errors, missing spaces
Form of address	None	Impersonal
Signature	None	MariaWolf
Other personaliz.	Part of e-mail address in link	None
Footer	Postal address	None
Story	Info: Message from Facebook administr. waiting for response	Bank account blocked due to suspicious actions
Call to action	Click link to read message	Click link to activate update and restore access
Priv./Sec. info	None	None
Attachment	None	None

Table 3: Features of Fraudulent E-mails

participants had a general university entrance qualification, 16 had a degree from a general university or a university of applied sciences.

All participants used the internet daily. 24 participants had been using the internet for more than 10 years, nine for 7 to 10 years and one participant for 3 to 7 years.

4.2 Methods of Analysis

The reasons given for the participants' reactions to each e-mail (from the experimenter's notes) were categorized and for each participant and each e-mail it was recorded whether a category was not mentioned at all, mentioned as supporting the decision or as contradicting the decision. Then the total number of mentions of each category and the number of participants which mentioned a category at least once were counted.

4.3 Predicted Features

At first, the aspects predicted by the model to influence the decision (see section 2) are analyzed. Table 5 illustrates the number of mentions of these features as well as how many

Purp. sender	**PayPal**	**Visa/Mastercard**
"From" address	PayPal.de <Service@paypal-europa.biz>	Visa und Mastercard <service@visa.de>
Subject	PayPal - account	Important message from Visa / Mastercard
Link URLs	http://support.paypal-service-europa.biz/ .../verify_account	None
Design/layout	PayPal logo + colors, icons (all look cut out)	Plain
Language	Professional, only one capit. error	Low linguistic quality, errors, encoding problem
Form of address	Impersonal	Impersonal
Signature	None	Regards
Other personaliz.	Reference number	None
Footer	Legal and company info	None
Story	Access to acc. restricted due to abnormal activities during last transaction	Credit card suspended due to unauthorized use
Call to action	Click link and enter account info to verify ownership	Fill credit card info into attached form and send back for verification
Additional info	Info about transaction	None
Priv./Sec. info	None	None
Attachment	None	Begrenzte_Kreditkarten_bilden.html (literal translation of restricted_creditcard_form)

Table 4: Features of Fraudulent E-mails Part 2

participants mentioned it at least once, and whether these results support the prediction. To control for possible learning effects, it shows the number of participants mentioning an aspect in the first e-mail and in the last e-mail they saw, respectively.

The "from" address was mentioned at least once by almost all participants. Most participants simply stated that certain "from" addresses appeared either authentic or dubious to them. Mostly the addresses in the fraudulent e-mails appeared dubious to participants because they were inconsistent with the signatory in the content, because they appeared unprofessional or came from an unexpected top-level domain. However, the cryptic appearance of the local parts of both the authentic Amazon and Facebook e-mails were mentioned as negative aspects by several participants as well. Link URL was also mentioned by a vast majority of participants. Most participants correctly identified the URLs in the fraudulent e-mail as suspicious because of domains which did not match the purported sender or lack of https and the URLs in authentic e-mails as trustworthy because of their domain names were known or at least fit the purported sender. However, the long cryptic link URL in the authentic Facebook e-mail was perceived as suspicious by some participants as well. Most of the mentions of subject lines were negative comments on the

Feature	Mentions	Part. Mentioning[a]	Supported?[b]
"From" Address	116	32 (12/17)	+++
Link URL	64	28 (7/10)	+++
Subject	7	6 (3/1)	+
Design / Layout	73	29 (6/6)	+++
Seal	12	10 (1/2)	+
Language	38	24 (4/5)	+++
Personalization	24	15 (2/2)	++
Signature / Comp. Info	17	12 (2/1)	+
Narrative Strength	24	17 (0/3)	++
Information / Cont. Quality	22	13 (4/4)	++
Context	40	28 (4/4)	+++
Persuasion Attempt	23	17 (1/3)	++
Privacy/Security Prot.	3	3 (0/0)	+
Reputation	1	1 (0/0)	-
Familiarity	31	21 (4/0)	++

[a] In brackets: Only mentions for the first/last email a participant has seen

[b] +++ : mentioned by most (24-34) participants, ++ : many (13-23) participants, + : few (2-12) participants, - : at most one participant

Table 5: Mentions of Factors Predicted to Influence Decision

subject of the e-mail about a blocked account, which was very simplistic and contained a capitalization error.

Another aspect mentioned by the vast majority of participants was design/layout. Missing design elements or bad layout were mostly mentioned as negative aspects of most of the fraudulent e-mails. The design of the Zalando e-mail with strong branding elements was mentioned as a positive aspect by many participants, but the design of the fraudulent "PayPal" e-mail was mentioned as a positive aspect by six participants as well, although three were suspicious about the authenticity of the logo. Mentions of trust seals were mostly concerning the presence of e-commerce trust seals as a positive aspect of the Zalando e-mail.

Language was mostly mentioned as negative aspects of most of the fraudulent e-mails due to their spelling and grammar mistakes and poor language in general, and as positive aspects of the authentic e-mails. The "personalization" category contains positive mentions of personal form of address or negative mentions of a lack thereof, as well positive mentions of the invoice number in the Amazon e-mail or the last digits of the account number in the comdirect e-mail. Mentions of the company information in the comdirect and Zalando e-mails, the lack of a personal signature and company information in the Visa/Mastercard e-mail and the mismatch between signatory name and sender address in the blocked account e-mail were categorized as "Signature / Company Information". The category "Narrative Strength" contains descriptions of e-mail content as (im)plausible, authentic/dubious, (not) making sense, but also explicit mentions that Visa and Mastercard are two different companies and therefore there cannot be an e-mail from both combined.

Comments about the amount and quality of general/supplemental information provided in e-mails was categorized as "Information / Content Quality". Especially the extensive information in the Zalando e-mail was praised by five participants, but also criticized by two. Three participants praised the information given in the purported PayPal e-mail.

Context was another aspect that the vast majority of participants mentioned at least once. That category consists mostly of participants mentioning that they would click the link in the Zalando e-mail or open the PDF attached to the Amazon e-mail if they had requested a password reset or placed the order mentioned, respectively. Four participants mentioned they would only go to the comdirect website if they had an account with them. The majority of the comments in the "Perceived Persuasion Attempt" category came from participants who either rejected the purported PayPal e-mail because it urged them to click a link and enter their login data (five participants), or who praised the comdirect or Amazon e-mails for being only informative in nature.

Of the three mentions of privacy/security protection, two are concerning the explanation of the privacy-preserving mechanisms in Amazon Marketplace's e-mail system, the third concerns the link to unsubscribe from notifications in the authentic Facebook e-mail. The only mention of reputation was one participant mentioning that she trusts Amazon. The last aspect which was predicted to influence decisions is familiarity with (receiving e-mails from) the sender. Comments in this category were from people who were cautious due to being unfamiliar with receiving that kind of e-mails from that sender (Amazon or Facebook), who trusted the Zalando e-mail because they recognized the logo from commercials, or who ignored or deleted e-mails because they have often received similar spam/phishing e-mails.

4.4 Exploratory Analysis

The analysis of the qualitative data found additional categories of aspects mentioned as affecting the decisions. Table 6 illustrates those additional aspects each with its total number of mentions and the number of participants mentioning it.

Aspect	**No. of Mentions Total**	**No. of Part. Mentioning**[a]
Encoding/Spaces	21	18 (3/3)
Time sent	7	6 (1/0)
Attachment	25	20 (1/4)
Advertisements	7	7 (2/0)
Limitation on Link	3	3 (0/0)
Curiosity	9	6 (1/0)
General Rule/Caution	81	31 (10/9)

[a] In brackets: Only mentions for the first/last email a participant has seen

Table 6: Frequencies of Additional Aspects Influencing Decisions

"Encoding/Spaces" refers to the encoding errors in the Visa/Mastercard e-mail and to the

missing spaces between some words in the account suspension e-mail, which participants mentioned as affecting credibility negatively. "Time sent" refers to participants finding the time at which the Visa/Mastercard e-mail was sent (midnight) suspicious. "Attachments" refers to participants mentioning either that they found the HTML file attached to the purported Visa/Mastercard mail to be suspicious or that the PDF receipt attached to the Amazon e-mail appeared either trustworthy or suspicious to them. Advertisements mostly refers to participants mentioning the advertisements in the password reset e-mail from Zalando as annoying and uncalled-for in this context, whereas the fact that the link for resetting the password was only valid for 24 hours ("Limitation on Link") was mentioned as a positive aspect by three participants. If participants asked themselves questions like "Where will I get when I click the link?" and said they would click the link to find out, or if they indicated they would click the link and then investigate the site closely, or that they might click the link in the Facebook e-mail to see the news, those reasons were categorized as "Curiosity". The category "General Rule/Caution" contains reasons like "I never read e-mails from Facebook", "Banking is a delicate matter" or "A bank adviser would call me on the phone". With 81 overall mentions and 31 out of 34 participants mentioning it at least once, General Rule/Caution turned out to be the second-most prominent category of reasons overall.

5 Discussion

All of the attributes of the message, sender or context that were predicted to influence the intention to follow a call to action contained in an e-mail were indeed mentioned by participants. The "from" address, design, link URL, context and language were the most prominent ones. The purported sender's reputation, however, was mentioned only once, perceived privacy/security protection was mentioned only by three participants.

The prediction that the purported sender's reputation affects the recipient's trust and intention to follow a call to action was mostly derived from e-commerce literature, where reputation and brand have a large impact on willingness to purchase from a website. It could be argued that since the e-mails shown in this experiment were directed at existing customers, participants imagined they were existing customers of the purported senders. According to the two-stage model by McKnight et al. [MCK00], reputation only affects consumer's trust in the early, exploratory phase and is displaced by actual usage experience afterwards. However, the results from this study still contradict previous studies in the phishing context [KAC06], where participants did indicate that the purported sender's reputation/brand is important for their decision to trust them.

The fact that perceived privacy/security protection was mentioned only by three participants can be explained by the fact that the e-mails used here do not mention those protections in direct relation to the calls to action. They only contain links to unsubscribing from notifications (Facebook) or information about privacy and security protection in the e-mail system in general (Amazon). In this light, it is interesting to see that three participants still mentioned the general privacy-related information as contributing to their trust in the e-mail and/or its sender. The subject line was mentioned by only six participants.

This could be due to the fact that participants were directly presented with the full e-mail, whereas in reality, the subject line is more relevant before an e-mail has been opened, at least in e-mail clients which allow deleting an e-mail without viewing its body.

There were also reasons mentioned which were not predicted. While the encoding problems in one e-mail which were mentioned by many participants were specific to this particular e-mail, other reasons apply to other e-mails as well. The advertisements mentioned by seven participants as a negative aspect of the Zalando e-mail indicate that they did not expect an e-mail with a link to reset the account password to contain ads. The time when it was sent is another attribute which was not found to influence credibility of an e-mail before. Curiosity, which was interpreted as the reason behind the decision to click a link by six participants, was not found as a reason for following calls to action in previous studies, but Wainer et al. [WDK11] found it as an important factor for reading e-mails in general.

The most important unpredicted reason, however, was a general rule or caution. 31 of our 34 participants gave reasons in this category. That this was not found in previous studies can be explained by the fact that previous studies either focused on experimentally manipulated attributes of the e-mail, measured personality traits or asked participants for the reason why they thought an e-mail was authentic or not (which would not reveal general rules such as "I never read e-mails from Facebook" which relate to behavior, though they could have revealed rules such as "A bank adviser would call me on the phone"). These general rules may have been acquired through general anti-phishing education or from banks' explicit information to their customers (or, in the case of Facebook, result from the fact that users do not find these messages useful because they regularly log into Facebook anyway). Another reason mentioned by the majority of participants were attachments. Although previous quantitative phishing research has used e-mails with attachments as stimuli [DJCF07], we know of no study were they were mentioned by participants of qualitative studies.

Before discussing the implications of this study's results, its limitations of course have to be considered. First of all, this being a laboratory experiment, the situation participants were in differed from real-life e-mail reading situations in several aspects. Participants were observed both by the eye tracker camera and the experimenter while reading the e-mails, which may have put them under pressure to be extra careful (although they were not told that their task was to distinguish between authentic and fraudulent e-mails). The fact that even under these conditions, three out of 34 participants would have clicked the link in the e-mail purportedly from PayPal, however, indicates that participants were not overly cautious. Since participants were able to see the e-mail while being asked about what had led them to their reaction, they may have noticed features of it which they had not noticed before. Therefore, the importance of some aspects for user's decisions may be overestimated. This is particularly true for the link URL, which was always visible at the bottom of the e-mail, whereas in reality, it is only seen when the mouse cursor hovers over the link. To make the situation more realistic, actual HTML e-mails viewed in web browsers or e-mail clients will be used in future studies.

Seeing eight e-mails in succession could theoretically have introduced learning effects. As shown in tables 5 and 6, however, the number of mentions remained relatively stable be-

tween the first and last e-mail seen for most aspects. The only aspects being mentioned by disproportionately few participants for the first e-mail were narrative strength, persuasion attempt and attachment. Familiarity, on the other hand, was mentioned by disproportionately few participants for the last e-mail. Whether these results reflect actual learning effects has to be investigated in further studies. Another limitation to generalization is the composition of the sample. Although we were able to sample a relatively wide range of age and education groups, a majority of participants were still university students. Therefore, the sample is not representative of all e-mail users.

For these reasons, the actual proportion of successful attacks in this study should not be generalized to real-world proportions.

6 Conclusion and Outlook

Overall, the results from this study are largely in line with the model and thus with previous research. They show that recipients base their decision whether to follow a call to action contained in an e-mail on aspects of design, content, language, but also on more technical aspects like the "from" address, link URL or the time it was sent. The fact that the phishing e-mail that purported to come from PayPal tricked three of our participants into believing it was authentic and clicking the link to enter their account data demonstrates that even in 2013, a phishing e-mail which is well written, contains graphical elements from the original brand and presents a convincing narrative can still be dangerous.

Providing users with general rules such as "A bank will never ask you for account details via e-mail" is confirmed to be effective in protecting them from at least some phishing attacks. Apparently, banks have been more effective than PayPal in teaching their customers such rules, and the results from this study emphasize the importance of improving their risk communication efforts.

Apart from general rules, another important result is that either users should be taught to pay less attention to design aspects or that the design elements should be hidden by default when displaying e-mails from unknown addresses or with other automatically identifiable suspicious attributes (which some e-mail clients already do) to prevent them from distracting users from more reliable indicators of authenticity.

The fact that context played a major role in participants' trust decisions emphasizes the danger of spear phishing [WHC+12], which abuses this reliance on context by targeting recipients based on information gathered on the web (for example finding out the e-mail address of an ebay user who is bidding on an auction which is about to end and sending him or her an e-mail about that auction). It therefore has to be communicated to e-mail users that a fitting context does not guarantee that an e-mail is authentic.

The central contribution of this study to research in this area is demonstrating that the question participants are asked can limit or enhance the insights gained from their answers. General rules or curiosity as reasons for why users do or don't follow calls to action in e-mails cannot be found when asking users to decide whether an e-mail is authentic or fraudulent and give reasons for that decision. This should be kept in mind whenever open

questions seek to find new reasons instead of just confirming known ones.

With interviews as its main method, this study was well-suited to focus on the aspects of the message which influence the recipient's decision. Other aspects of the model such as personality traits or knowledge and experience have to be evaluated in quantitative studies. A further iteration of the study presented in this paper could aim for improved generalizability by putting participants in a situation which is closer to real life. However, a potential bias caused by participants' reflection on their own thought processes – which is necessary for qualitative studies – can never be completely avoided.

Acknowledgements We thank the anonymous reviewers for their helpful comments, although unfortunately some of them could not be implemented due to space limitations.

References

[DJCF07] Ronald C Dodge Jr, Curtis Carver, and Aaron J Ferguson. Phishing for user security awareness. *Computers & Security*, 26(1):73–80, 2007.

[KAC06] Ponnurangam Kumaraguru, Alessandro Acquisti, and Lorrie Faith Cranor. Trust modelling for online transactions: a phishing scenario. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*, PST '06, pages 11:1–11:9, New York, NY, USA, 2006. ACM.

[MCK00] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Trust in e-commerce vendors: a two-stage model. In *Proceedings of the twenty first international conference on Information systems*, ICIS '00, page 532–536, Atlanta, GA, USA, 2000. Association for Information Systems.

[PTK13] Thomas Pfeiffer, Heike Theuerling, and Michaela Kauer. Click Me If You Can! – How do users decide whether to follow a call to action in an online message? In Louis Marinos and Ioannis Askoxylakis, editors, *Human Aspects of Information Security, Privacy, and Trust*, number 8030 in Lecture Notes in Computer Science, pages 155–166. Springer Berlin Heidelberg, January 2013.

[SW11] Frank Stajano and Paul Wilson. Understanding scam victims: seven principles for systems security. *Communications of the ACM*, 54(3):70–75, 2011.

[WDK11] J. Wainer, L. Dabbish, and R. Kraut. Should I open this email?: inbox-level cues, curiosity and attention to email. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, page 3439–3448, 2011.

[WHC+12] Jingguo Wang, T. Herath, Rui Chen, A. Vishwanath, and H.R. Rao. Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email. *IEEE Transactions on Professional Communication*, 55(4):345 –362, December 2012.

Automotive Software Engineering

(ASE)

Automotive Software Engineering (ASE)[1]

Abstract: Der Trend, immer komplexerer und umfangreicherer Softwareumfänge im Fahrzeug, aber auch in den unterstützenden Prozessen, ist ungebrochen. Mobilität und Logistik stellen mittlerweile Software-Ökosysteme dar, die noch viele Herausforderungen - wie auch das des Tagungsmottos „Big Data – Komplexität meistern“ - bergen.

Der 12. Workshop Automotive Software Engineering beschäftigt sich mit Methoden, Techniken und Werkzeugen in der Software Entwicklung für Fahrzeuge, d.h. PKW und Nutzfahrzeuge. Aktuelle und zukünftige Fahrzeuge sind zunehmend untereinander und mit der IT-Infrastruktur vernetzt. Die Zuverlässigkeit, Sicherheit (Security und Safety), Bedienbarkeit aber auch der Datenschutz (Privacy) stellen in diesem Zusammenhang essentielle Qualitätsaspekte dar.

Der Austausch und die Diskussion darüber, wie diese auch in Zukunft vor dem Hintergrund zukünftiger Software Ökosysteme erreicht und garantiert werden können, stellen Inhalt und Ziele des diesjährigen Workshops dar.

Inhalte und Zielgruppe

In der Workshopreihe Automotive Software Engineering der gleichnamigen GI-Fachgruppe werden Herausforderungen und Lösungsansätze des Automotive Software Engineering vorgestellt und diskutiert. In der diesjährigen 12. Ausgabe des Workshops stehen die Themen „Connected Services und Big Data“, „Architektur” und „Validierung” mit insgesamt acht Beiträgen auf dem Programm.

Es konnten wieder zwei Keynote-Sprecher gewonnen werden, die das Themenfeld einmal aus Sicht der Industrie und einmal aus Sicht der Forschung beleuchten.

Der Workshop richtet sich an Forscher, Entwickler und Anwender bei Fahrzeugherstellern und -zulieferern, sowie an Mitarbeiter/innen aus Forschungsinstituten, Universitäten und Fachhochschulen, die im Gebiet Automotive Software Engineering arbeiten. Im Fokus stehen traditionell weniger theoretische, sondern vielmehr praxisnahe Arbeiten.

[1] http://www.iese.fraunhofer.de/ase2014

Keynotes

Next Generation Online Services in the Automotive Industry: Opportunities with Big Data

Vortragender: Gerhard Stanzl, Leiter Vorentwicklung Online- / Car2X-Systeme, Audi Automotive Venture GmbH

Abstract: Automotive services are more and more based on collected and aggregated data of “car status”, “infrastructure”, “user profiling” and “environment information”. The aggregation of data from these categories is offering next generation surprising and differentiating services to improve traffic safety, reduce emissions, and enhance comfort for the customer. In addition, the integration of data and services will result in an essential building block for Piloted Driving.

Roadmap Automotive Software Engineering 2030

Vortragender: Stefan Kowalewski, RWTH Aachen

Abstract: Viele Anwendungstrends, aber auch gesellschaftliche Rahmenbedingungen, werden die Softwareentwicklung in den nächsten Jahren rund um Mobilitätslösungen vor große Herausforderungen stellen. Im Rahmen des Vortrags werden Beispiele dieser Herausforderungen, sowie technische und methodische Lösungsansätze gegeben. Schließlich werden Handlungsempfehlungen für die Umsetzung erläutert.

Organisatoren: Dr.-Ing. Mario Trapp (Fraunhofer IESE), Ralf Kalmar (Fraunhofer IESE), Markus Schurius (Audi Electronics Venture GmbH)

Programmkomitee: Dr. Christian Allmann (Audi AG), Prof. Dr. Dr. h.c. Manfred Broy (Technische Universität München), Dr. Stefan Bunzel (Continental AG), Dr. Mirko Conrad (samoconsult GmbH), Prof. Dr. Werner Damm (Universität Oldenburg), Dr. Peter Dencker (ETAS GmbH), Dr. Heiko Dörr (Model Engineering Solutions GmbH), Bernd Frielingsdorf (Ford Werke GmbH), Prof. Dr. Bernhard Hohlfeld (ICS AG u. Technische Universität Dresden), Prof. Dr. Stefan Jähnichen (Technische Universität Berlin), Ralf Kalmar (Fraunhofer IESE), Prof. Dr. Stefan Kowalewski (RWTH Aachen), Prof. Dr. Reiner Kriesten (Hochschule Karlsruhe), Prof. Dr. Thomas Kropf (Robert Bosch GmbH), Peter Manhart (Daimler AG), Prof. Dr. Peter Pickel (John Deere ETIC), Prof. Dr. Klaus Pohl (Universität Duisburg-Essen), Prof. Dr. Wolfgang Pree (Universität Salzburg), Dr. Alexandre Saad (BMW Group), Prof. Dr. Ina Schäfer (Technische Universität Braunschweig), Prof. Dr. Jörn Schneider (Hochschule Trier), Dr. Sebastian Siegl (Audi Electronics Venture GmbH), Dr. Reinhard Stolle (BMW Car IT GmbH), Claus Stellwag (Elektrobit Automotive GmbH), Prof. Dr. Michael Uelschen (Fachhochschule Osnabrück), Dr. Marcel Wille (Volkswagen AG), Dr. Fabian Wolf (Volkswagen AG), Dr. Dirk Ziegenbein (Robert Bosch GmbH)

Putting the car on the map: A scalable map matching system for the Open Source Community

Sebastian Mattheis,
Kazi Khaled Al-Zahid, Birgit Engelmann, Andreas Hildisch,
Stefan Holder, Olexiy Lazarevych, Daniel Mohr, Felix Sedlmeier,
Richard Zinck

BMW Car IT GmbH - Munich Lab
Petuelring 116
80809 Munich
sebastian.mattheis@bmw-carit.de

Abstract: Recent years have seen a proliferation of mobile devices connected to the internet, including connected cars. These systems send a stream of position annotated messages when requesting location based services. The position information in turn can be used to improve those services. Here, we focus on online map matching of the most recent position of a connected vehicle on a road map. This information, aggregated and privacy aware, can serve as a basis e.g. for machine learning algorithms used to improve traffic prediction. We describe a system for online map matching in the backend that implements a state of the art algorithm based on a Hidden Markov Model. This system uses only open source software and open data. The development of the map matcher was motivated by a perceived lack of a scalable system in the open source realm. We discuss its role as part of a scalable backend system designed to provide spatially aware services.

1 Introduction

Connected mobile devices have become commonplace, reaching from smartphones and tablets to cars and watches. These devices use their connection to the internet to provide various services such as navigation, traffic alerts, shopping, restaurant reviews and other points of interest (POI) information. Many of these applications need to know the location of the user in to provide the best service. Hence, several studies have been published that evaluate movement patterns in cities from pedestrians to cyclists and taxis ([RWFP06, LLC+14, KSR13, LZL+12]). Here, our focus is on cars as connected devices. In contrast to pedestrians and cyclists, cars are bound in their movement to road networks. In many cases, it is not necessary to know the exact location of a car in order to provide a service, e.g. pharmacies close by can be listed with only a rough position estimate of the car. Nevertheless, in order to receive relevant traffic information, it is essential to know a drivers exact location including the position on a road segment. Crucially, such information can be used to assess the current traffic situation at a specific location. In

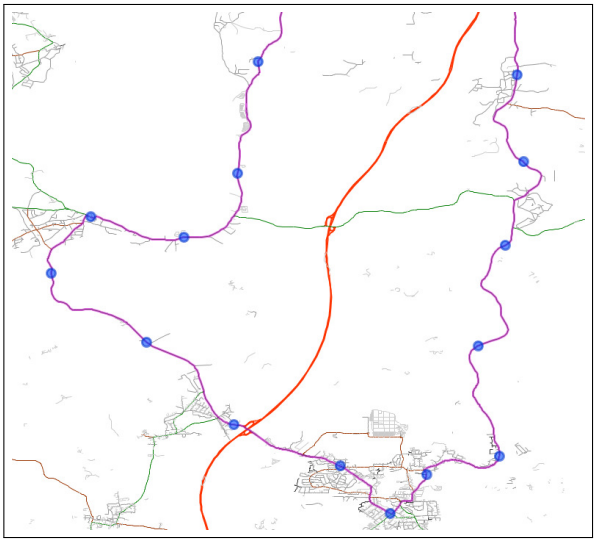


Figure 1: Shows the map matched path of a vehicle (blue dots connected by purple line). The road map is compiled from OSM data. Red color indicates a primary road, green color secondary roads.

such examples, noisy location information, provided e.g. by GPS measurement, must be mapped onto the most likely position on a road network. This problem is referred to as map matching.

Here, we describe a system for online map matching in the backend. It is part of a scalable backend system designed to provide spatially aware online services. The development of this system was motivated by a perceived lack of a scalable map matcher in the Open Source realm. The main contribution is the map matcher[1] that is to be released to the Open Source community. Section 2 gives an introduction to the algorithm we used for map matching. Section 3 presents the architecture of the map matcher and discusses options for scaling. Finally, we discuss future steps.

2 Foundations

2.1 Map matching

A trajectory is defined here as the path of a vehicle on a road network. The objective of trajectory map matching is to estimate this path from noisy position data (e.g. from GPS). The road map represents the topology of the road network and, since every point is localized, it also provides a geometric representation. The vertices of the road topology correspond to intersections. Edges represent end-to-end connections between intersections, also referred

[1] https://github.com/bmwcarit/barefoot

to as road segments. Since we define edges to be unidirectional, a bidirectional road segment is represented by two edges with opposite directions. The vehicle's path is sampled in a sequence of position measurements $(z_0, \ldots, z_T)$ during a time interval $[0, 1, \ldots, T]$. Each position measurement z_t with $t \in [0, T]$ corresponds to a position s_t on the map, i.e. a position on an edge of the road topology. Since position measurements z_t are subject to measurement errors, they can be matched to more than one map position. The set of possible map positions for each measurement z_t is denoted as the set of position candidates $\mathcal{S}_t$.

Map matching can be modeled as a probabilistic problem. It corresponds to finding the most likely sequence of system states in a Markov chain [RN99].

Definition 1. Trajectory map matching is defined as finding the most likely sequence of position candidates $\tilde{P} = (s_0, \ldots, s_T)$ with

$$\tilde{\mathcal{P}} = \underset{s_0,\ldots,s_T}{\arg\max}\, p(s_0, \ldots, s_T | z_0, \ldots, z_T), \tag{1}$$

where $s_t \in \mathcal{S}_t$ with $t \in [0, T]$. The probability that the sequence $(s_0, \ldots, s_T)$ corresponds to the actual trajectory of the vehicle, given the measurements $(z_0, \ldots, z_T)$, is denoted as $p(s_0, \ldots, s_T | z_0, \ldots, z_T)$.

Definition 1 restricts trajectory map matching to finding a sequence of map positions only. It ignores the vehicle's path between map positions. Nevertheless, the path gives evidence on the likeliness of a position sequence e.g. a shorter path is more likely than a longer one. This is considered in the probability $p(s_0, \ldots, s_T | z_0, \ldots, z_T)$. However, identifying a path requires route assumptions for interpolating the vehicle's trajectory. This is left to a router that can be configured to use an arbitrary routing cost function. The router takes a tuple of map positions (s_{t-1}, s_t) with $t \in [1, T]$ and returns the path from position s_{t-1} to position s_t, denoted as $\langle s_{t-1}, s_t \rangle$, with minimum costs according to the chosen cost function. A path is a sequence of edges of the road topology. The obtained paths are then assessed with a probability and can be considered in the overall solution of the map matching problem.

Trajectory map matching is computationally expensive, especially if the number of position measurements is large. However, many online services require a position estimate s_t only for the most recent measurement z_t. In a Markov chain, this corresponds to system state filtering [RN99].

Definition 2. Online map matching is defined as finding the most likely map position $\tilde{s}_t$ with

$$\tilde{s}_t = \underset{s_t}{\arg\max}\, p(s_t | z_0, \ldots, z_t), \tag{2}$$

where $s_t \in \mathcal{S}_t$ and $t \in [0, T]$. The probability that the position candidate s_t corresponds to the actual position of the vehicle on the road network, provided the measurements up to the most recent $(z_0, \ldots, z_t)$, is denoted as $p(s_t | z_0, \ldots, z_t)$.

Our focus lies on online map matching as it is required by many online services.

2.2 Hidden Markov Model map matching

Hidden Markov Model map matching has been established as state of the art method for both offline [NK09] and online map matching [GDM+12]. The first-order Hidden Markov Model (HMM) defines a system's behavior over time as a sequence of system states $(s_0, \ldots, s_T)$. They are referred to as hidden since the actual system states can only be observed by a sequence of emissions $(z_0, \ldots, z_T)$ [RN99]. As illustrated in Figure 2, each emission z_t is the result of an observation (wavy arrows) of the system's state s_t at time $t \in [0, T]$. The state s_t at time t is one of many possible states in a finite set of state candidates $\mathcal{S}_t$. Since observations are subject to measurement noise, their uncertainty can be modeled by an emission probability $p(z_t|s_t)$. The state transitions, depicted as straight arrows, are part of a stochastic process with transition probability $p(s_t|s_{t-1})$.

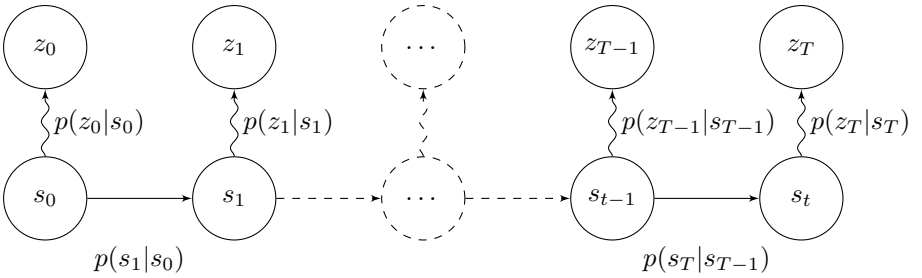


Figure 2: First-order Hidden Markov Model for trajectory map matching. Each emission z_t is the result of an observation (wavy arrows) of the system's state s_t at time $t \in [0, T]$. The state transitions (straight arrows) are part of a stochastic process with transition probability $p(s_t|s_{t-1})$. A state refers to a map position of a vehicle and emissions refer to position measurements. The transition between system states corresponds to the path between map positions.

The first-order HMM makes the following assumptions about state transitions and emissions [RN99]:

- Markov state assumption: The system state s_t at time $t \in [1, T]$ depends only on the previous state s_{t-1} such that $p(s_t|s_0, \ldots, s_{t-1}) = p(s_t|s_{t-1})$.
- Markov sensor assumption: The probability of emission z_t at time $t \in [0, T]$ depends only on the current state s_t such that $p(z_t|z_0, \ldots, z_{t-1}, s_0, \ldots, s_t) = p(z_t|s_t)$.

In the context of map matching, a state refers to a map position of a vehicle and emissions refer to position measurements. The transition between system states corresponds to the path between map positions. The solution to the online map matching problem follows a recursive approach of state filtering [RN99]:

Theorem 1. The most likely map position $\tilde{s}_t$ corresponding to a position measurement z_t can be deduced from Definition 2 given that

$$p(s_t|z_0 \ldots z_t) = \alpha \cdot p(z_t|s_t) \cdot \sum_{s_{t-1}}^{\mathcal{S}_{t-1}} p(s_t|s_{t-1}) \cdot p(s_{t-1}|z_0 \ldots z_{t-1}), \tag{3}$$

where $p(s_t|z_0 \dots z_t)$ is the probability of position candidate $s_t \in \mathcal{S}_t$, given the position measurements $(z_0, \dots, z_t)$. Further, $p(s_{t-1}|z_0, \dots, z_{t-1})$ is the probability of position candidate $s_{t-1} \in \mathcal{S}_{t-1}$, given the position measurements $(z_0, \dots, z_{t-1})$. The emission probability of position candidate s_t is $p(z_t|s_t)$ and transition probability from position s_{t-1} to position s_t is $p(s_t|s_{t-1})$, and α is a normalizing constant. The probability of initial position candidates $s_0 \in \mathcal{S}_0$ is

$$p(s_0|z_0) = \alpha \cdot p(z_0|s_0). \tag{4}$$

Both equations (3 and 4) are derived using Bayes' rule.

2.3 Online map matching algorithm

Theorem 1 provides a recursive solution to the online map matching problem (Definition 2). It can be implemented as an online algorithm that iteratively determines the best estimate of the vehicle's map position $\tilde{s}_t$. Each position measurement z_t triggers an iteration of the algorithm and updates the position estimate. The implementation of the algorithm is straight-forward. It remains to provide robust strategies for the selection of position candidates and the determination of their emission and transition probabilities. The strategies of our implementation are mainly adopted from [NK09] which are similar to those used by [LZZ$^+$09, GDM$^+$12]. It identifies map position candidates $\mathcal{S}_t$ as points on a road segment that have minimum geodesic distance to the measurement z_t. The road segments must overlap with a geodesic circle around the position measurement z_t. The emission probabilities $p(z_t|s_t)$ are defined as a Gaussian distribution

$$p(z_t|s_t) \sim \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left\{-\frac{||z_t - s_t||^2}{2\sigma_z^2}\right\}, \tag{5}$$

where σ_z is the standard deviation of GPS measurements and $||z_t - s_t||$ is the geodesic distance between map location s_t and measurement z_t. The determination of transition probabilities requires routing from position candidate s_{t-1} to s_t. The obtained path is denoted as $\langle s_{t-1}, s_t \rangle$ and has length $|\langle s_{t-1}, s_t \rangle|$. In [NK09], transition probabilities have been experimentally determined to fit a negative exponential distribution

$$p(s_t|s_{t-1}) \sim \lambda \exp\Big\{\lambda \left(||z_t - z_{t-1}|| - |\langle s_t - s_{t-1} \rangle|\right)\Big\}, \tag{6}$$

where it remains to find the best parameterization (estimate of λ) for a specific sampling set, i.e. the sequence of position measurements.

3 The system architecture

3.1 The online map-matching system

The online map-matching system processes a stream of position measurements $(z_0, \ldots, z_t)$ in order to determine a vehicle's most likely position $\tilde{s}_t$ on the map that corresponds to the most recent position measurement z_t. Figure 3 depicts the high level architecture of our system. It consists of an online map matching component (Filter), a geometrical road map (Map), a routing unit with a separate road map topology (Router) and a memory unit that provides access to state information of the tracked vehicle (State).

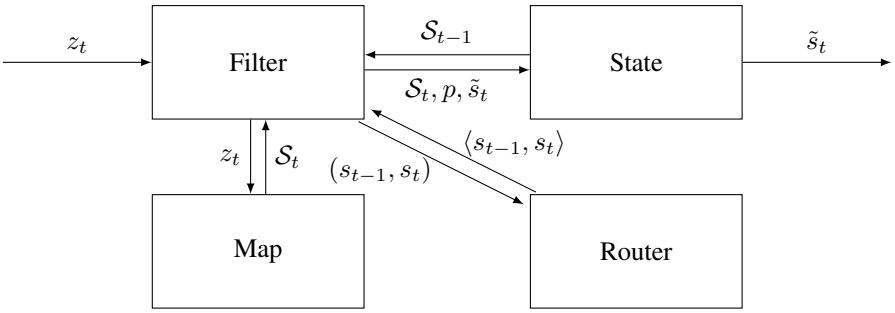


Figure 3: Architecture of a HMM map matching system consisting of an online map matching component (Filter), a geometrical road map (Map), a routing unit with a separate road map topology (Router) and a memory unit that provides access to state information of tracked vehicle (State).

The filter implements the online map matching algorithm as described in Section 2.3 and is executed when it receives a position measurement z_t. In the first step, the filter selects map position candidates $\mathcal{S}_t$ from the map, i.e. road segments near measurement z_t. The map is compiled from OpenStreetMap[2] data and represents road geometries using geometry data structures provided with ESRI's Java Geometry API[3]. This library comes with a spatial index data structure (Quadtree) that enables efficient spatial range search for nearby road segments [Sam90]. Spatial operations such as geodesic distance calculations or point-to-line projections in WGS-84 map projection are only partially supported. To provide this functionality, we used and extended the Java implementation of GeographicLib[4] maintained by Charles Karney which provides even more exact geodesic distance calculations [Kar13].

In the second step, the filter requests the path $\langle s_{t-1}, s_t \rangle$ for each pair of position candidates (s_{t-1}, s_t) with $s_{t-1} \in \mathcal{S}_{t-1}$ and $s_t \in \mathcal{S}_t$. The path is provided by the router component. We evaluated several Open Source routers that rely on OpenStreetMap data. The pgRouting[5] project provides routing extensions for the PostgreSQL database management system

[2] http://openstreetmap.org
[3] http://github.com/Esri/geometry-api-java
[4] http://geographiclib.sourceforge.net
[5] http://pgrouting.org

that can be executed within a SQL query. It has a significant overhead since the routing algorithms are implemented in C functions that load data from database into memory on each execution. GraphHopper[6] is a Java-based router that is highly optimized for long distance routing, e.g. using graph contraction hierarchies. However, its programming interface does not support routing between pre-defined road segments. Routino[7] is another router implementation for OpenStreetMap data and was discarded because it is licensed under the AGPL (GNU Affero General Public License) that we assessed as too restrictive.

In the final step, the filter determines emission probabilities $p(z_t|s_t)$, transition probabilities $p(s_t|s_{t-1})$ and posterior probabilities $p(s_t|z_0, \ldots, z_t)$ for each pair of position candidates (s_{t-1}, s_t) with $s_{t-1} \in \mathcal{S}_{t-1}$ and $s_t \in \mathcal{S}_t$. The vehicle's state information, i.e. map position candidates $\mathcal{S}_t$, probabilities $p(s_t|z_0, \ldots, z_t)$ for each $s_t \in \mathcal{S}_t$ (in Figure 3 denoted as p) and the most likely map position $\tilde{s}_t$, is saved to memory. To enable online map matching of multiple vehicles, it is necessary to associate incoming position measurements z_t and state information saved in memory with object identifiers. To provide efficient memory access, it is necessary to maintain a search index on those identifiers. Furthermore, services typically query for all objects that are within a certain range. Thus, we also use a spatial search index on position estimates $\tilde{s}_t$ to increase search performance [Sam90].

3.2 Strategies for scalability

Our online map-matching system must process and organize information timely without being affected in its performance by the number of mobile objects that it tracks. We refer to the number of messages per second as the system's load. A system is scalable if it guarantees a fixed response time independently of its load by increasing computing, storage and communication elements. There are two approaches to system scaling: The first is to scale up a computer system by increasing the capacity of existing resources, such as by using faster processors or bigger memory. This is referred to as vertical scaling which is the focus of research in multicore systems and storage media. The second approach is referred to as horizontal scaling in which a computer system is scaled out by adding more machines (cluster nodes). Here, we focus on horizontal scaling.

Figure 4 shows a generic architecture for a scalable online map-matching system. Components can be implemented using different Open Source software solutions. The architecture scales up the online map-matching system of Figure 3. It uses two major clusters, i.e compute cluster (C), with n nodes, and storage cluster (S), with m nodes. Each node of the compute cluster runs an instance of the filter, map and router. The state memory is distributed over nodes $s_1, \ldots, s_m$ of the storage cluster (S). A load-balancer (L) distributes incoming messages over compute nodes. Each compute node performs map matching as described for a single-node system, while it reads and writes state information from or to the storage cluster. The filters of compute nodes access state information directly by object identifiers. In contrast, service applications (A) query state information mostly by spatial search through a query layer (Q) provided by the storage cluster.

[6] http://graphhopper.com
[7] http://routino.org

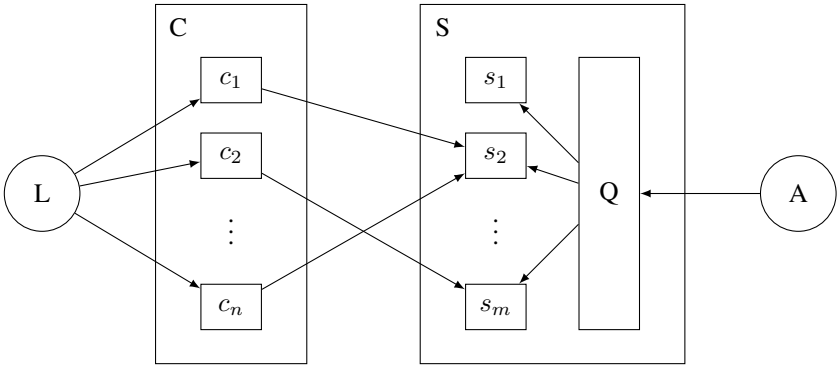


Figure 4: Scalable system architecture for online map matching. A load balancer (L) distributes message streams to a compute cluster (C). The compute cluster uses the storage cluster (S) that itself is queried from applications (A) via a query layer (Q). We refer to two physically separated clusters; however, advanced approaches may separate these clusters only logically to enable physical locality of computing and storage resources.

The major design decisions made with this architecture are:

- The compute and storage cluster can be physically separated. Hence, the size of each cluster (m and n) can be scaled independently such that only those resources have to be added that are running out, i.e. either storage or computing resources.
- The compute cluster consists of interchangeable nodes meaning that each message can be processed by any of those nodes. This simplifies load-balancing.
- The storage cluster is accessed by compute nodes and services in a different manner, allowing for the independent optimization of each type of state access, i.e. by identifier search or spatial search.

The components of this architecture (Figure 4) can be implemented completely with Open Source software. Apache Kafka[8] is a distributed messaging system that can be used as multi-node load-balancer. A broker distributes incoming messages over a set of queues, referred to as partitions, randomly. Each compute node can fetch messages from a dedicated partition exclusively in a work-stealing fashion. In each partition, ordering of messages is strictly conserved; however, distribution over partitions violates strict ordering of messages among partitions. This is a relaxation of fairness that is necessary to gain scalability. [HKLP12] In our system, this is acceptable if the ordering of messages sent from a participant is conserved with high probability, which is satisfied if no partition has significantly higher throughput than any other. Messages that are processed out-of-order are discarded. This is necessary to prevent overwriting of state information with older information. Hence, the service quality corresponds to that of a soft real-time system.

The compute cluster requires a runtime environment for the map matcher components, i.e. filter, map and router. This also requires capabilities for scalable real-time stream

[8] http://kafka.apache.org

processing as it is available with Apache Storm[9]. It provides a framework for implementing directed acyclic computation graphs which are automatically deployed to compute nodes. Message streams are processed under soft real-time constraints. It is fail-safe and scalable to an arbitrary number of nodes. Apache Spark Streaming[10], a recent extension of Apache Spark for stream processing, provides an alternative with similar capabilities. Synchronization between compute nodes can be implemented with Apache Zookeeper[11], a scalable distributed directory service for maintaining configuration information and providing distributed synchronization. It is also used by various other software components as e.g. Apache Kafka and Apache Storm.

The storage cluster can be realized by Apache Cassandra[12], a database management system that provides key-value data access and linear scalability. During message processing, compute nodes can access state information by object identifiers with low latency by using respective index structures. The index structure is organized in such a way that the range of identifiers is partitioned over storage nodes where each storage node holds state information for a well-defined range of identifiers. This way, accessing storage by object identifiers can be forwarded directly to the respective storage node. (This is also referred to as direct mapping.)

Apache Cassandra supports map reduce queries with Apache Hadoop[13] by implementing the interface of the Hadoop Distributed File System (HDFS). This is sufficient to implement a query layer e.g. by using the following approach: To enable efficient search for objects by spatial properties, i.e. the object's last-known position, one must define a secondary index. This index can be created by hashing the position, e.g. using a geohash as implemented in one of various Open Source libraries[14]. A general problem with secondary indices, however, is that the values associated with a key can be located on any storage node. This is referred to as fully-associativity which is contrary to direct mapping. As a consequence, secondary index keys must be queried on each storage node which decreases performance. Nevertheless, map reduce with Apache Hadoop can query each storage node in parallel avoiding high latencies. Another performance problem arises with geohashes that require a range query to be split to all possible geohashes within this range. Querying larger areas may then be subject to massive performance drops. The problem is similar with querying nearest neighbors and could be solved in future with spatial index data structures such as Quadtree [Sam90], R-Tree [Gut84] or R*-Tree [BHPSS90].

4 Discussion

In this paper we discussed the role of map matching as a first and central step in providing location based services. We describe a map matcher that is both scalable and uses a state

[9] http://storm.incubator.apache.org
[10] http://spark.incubator.apache.org/docs/0.9.0/streaming-programming-guide.html
[11] http://zookeeper.apache.org
[12] http://cassandra.apache.org
[13] http://hadoop.apache.org
[14] https://github.com/davidmoten/geo

of the art map matching algorithm (Hidden Markov Model) that we believe is missing in the Open Source domain. This software is important yet non-differentiating from the OEM (Original Equipment Manufacturer) point of view and can hence be published and developed in a cooperative Open Source community.

The quality of trajectory map matching depends on the selection of state candidates and determining emission and transition probabilities. These problems are tightly coupled since candidates can be selected by maximizing the observation and transition probabilities. This is not yet fully utilized as the transition probability depends on local characteristics of the road network that are ignored. Further, in-vehicle systems could exploit information from the car (vehicle turns, velocity, computer vision, etc.). Backend systems, in turn, can exploit statistical information including:

- Turn probabilities at intersections can serve as base rate if other measurements are inclusive or contradictory.
- The velocity distribution of road segments can provide information for map matching, e.g. if parallel roads are close to each other and one of the two is a secondary road with usually lower speeds.

As coverage of technologies like LTE increases we anticipate a steady rise in communication between connected devices and backend systems, including cars. Nonetheless, since many services require the use of multiple data sources, we expect that the backend will continue to need the ability to localize all information on a common map and hence will require a scalable map matcher component.

We envision our map matcher as a component of a scalable, Open Source based backend ecosystem optimized for spatial-temporal data and service management. Such an ecosystem should enable scalable stream processing and data management with low latency service access and the ability to process large data sets for statistical information that can be used to optimize services. The quality of these services, in turn, is a differentiating factor where competition between market participants will take place. We believe that a tight integration to projects from established players in the Open Source community, such as the OpenStreetMap and the Apache Software Foundation, will be the key to success.

References

[BHPSS90] N. Beckmann, Kriegel H.-P, R. Schneider, and B. Seeger. The R*-tree: an efficient and robust access method for points and rectangles. In *International conference on management of data*, 1990.

[GDM+12] C.Y. Goh, J. Dauwels, N. Mitrovic, M.T. Asif, A. Oran, and P. Jaillet. Online map-matching based on Hidden Markov model for real-time traffic sensing applications. In *International IEEE Conference on Intelligent Transportation Systems*, 2012.

[Gut84] Antonin Guttman. R-trees: a dynamic index structure for spatial searching. In *International conference on management of data*, 1984.

[HKLP12] A. Haas, C. Kirsch, M. Lippautz, and H. Payer. How FIFO is your concurrent FIFO queue? In *Workshop on Relaxing synchronization for multicore and manycore scalability*, 2012.

[Kar13] C. Karney. Algorithms for geodesics. *Journal of Geodesy*, 87(1):43–55, 2013.

[KSR13] Kevin S. Kung, Stanislav Sobolevsky, and Carlo Ratti. Exploring universal patterns in human home/work commuting from mobile phone data. *CoRR*, abs/1311.2911, 2013.

[LLC+14] Thomas Louail, Maxime Lenormand, Oliva García Cantú, Miguel Picornell, Ricardo Herranz, Enrique Frias-Martinez, José J Ramasco, and Marc Barthelemy. From mobile phone data to the spatial structure of cities. *arXiv preprint arXiv:1401.4540*, 2014.

[LZL+12] Xiao Liang, Xudong Zheng, Weifeng Lv, Tongyu Zhu, and Ke Xu. The scaling of human mobility by taxis is exponential. *Physica A: Statistical Mechanics and its Applications*, 391(5):2135–2144, 2012.

[LZZ+09] Y. Lou, C. Zhang, Y. Zheng, X . Xie, W. Wang, and Y. Huang. Map-matching for Low-sampling-rate GPS Trajectories. In *Proceedings of the International Conference on Advances in Geographic Information Systems*, 2009.

[NK09] P. Newson and J. Krumm. Hidden Markov Map Matching Through Noise and Sparseness. In *Proceedings of International Conference on Advances in Geographic Information Systems*, 2009.

[RN99] S. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1999.

[RWFP06] Carlo Ratti, S Williams, D Frenchman, and RM Pulselli. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B Planning and Design*, 33(5):727, 2006.

[Sam90] H. Samet. *The Design and Analysis of Spatial Data Structures*. Addison-Wesley, 1990.

Location based learning of user behavior for proactive recommender systems in car comfort functions

Thomas Stone, Olga Birth, André Gensler, Andreas Huber, Martin Jänicke, Bernhard Sick

Universität Kassel
Intelligent Embedded Systems
{gensler,jaenicke,bsick}@uni-kassel.de

BMW Group
Forschung und Technik
Olga.Birth@bmw.de

BMW Group
Intelligente Komfortfunktionen
{Andreas.AH.Huber,Thomas.Stone}@bmw.de

Abstract:
In-car comfort functions decrease driving stress and therefore increase safety. Drivers typically do not use all available comfort functions optimally, if at all, in every situation they would offer a substantial increase in comfort. Automating such functions as proactive recommender systems would exploit the full potential for decreasing driver stress. Because comfort functions are highly dependent on the driver's habits, learning the individual user behavior is necessary. We propose a probabilistic method for modelling and predicting location dependent user behavior of comfort function activations. The model applies second-order uncertainty to evaluate the certainty about inferred parameter values and it deals with novelty and decaying observations explicitly. The results of this study show that the use of probabilistic models for learning location based user behavior in car comfort functions is a promising technique and gives reason to further investigate this area of studies.

1 Introduction

The primary research efforts in the area of vehicular automation is concentrated on context-aware ADAS. For automotive automation, context-awareness is an important issue, because many complex use cases become more dynamic and ambigous with a higher level of automation. Comfort functions also have potential for automation. Automating these functions eliminates or simplifies the user interaction, decreasing driver stress. As for now,

most comfort functions' system parameters are set to fit the average user. The parameters are determined by user studies and expert knowledge, incorporating little to no online, user-specific contextual information. The driver's information, like calendar data, frequent routes, the vehicle's current location or the individual pattern of handling assistance and comfort systems, can improve the precision of automation and enable personalized features, therefore increasing the acceptance of such systems.

For the development of future automated comfort functions, learning the user's individual behavior is a vital part and makes personalized, proactive recommendations possible. As an example, the class of climatic comfort functions is used dependent on climatic conditions, personal preferences and environmental circumstances. For non-climatic functions, the vehicle's current geopraphic or road type based location may have a strong effect on the probability of the activation by the user, e.g., parking assistance systems may only be used close to free parking spaces. Systems which automate functions, depending on such locations of interest, are called location based learning systems. We will focus on learning locations of interest and dependent user function activations. Further on, using this knowledge for proactive recommendations will be called location based automation and recommendation and is the target application of the proposed algorithms.

There are several major requirements that any procative vehicular recommentation system must satisfy. The model must be embedded, soft real-time capable, operating economically with memory and processing time consumption. Also, a mechanism for self-reflection is necessary to detect when the certainty about inferred parameters is high enough to make proactive recommendations. A very important property is the concept of online and lifelong learning, which means that a system can cope with novel situations as well as already observed ones fast (a precise definition is given in [FH06]). To give the user the possibility to control the system directly up to a certain degree, implicit and explicit feedback must be taken into account by the system. Explicit feedback is a user action (e.g., activating an automated function or counteracting to an automation), while implicit feedback is the lack of user action at a certain location which can also mean to accept the actions of such a system. At last, the user must have an understanding about the system's mispredictions. In a highly subjective field such as comfort functions, it is barely possible to construct a never-failing prediction system.

In the following proposal, we will describe a system that will meet all of these goals, whether implicitly or through explicit modeling.

2 Related work

There has been a lot of research in the area of recommendation systems for increasing comfort in cars. Such systems include context-aware and proactive recommendations, such as optimal gas refueling recommendations [WBE09] or models for recognizing driver fatigue for trip pause suggestions [Yan07].

The systems presented in [GKFC11], [KGT$^+$08] and [APRR07] incorporate contextual information. Some of the information is retrieved through the internet, while others, e.g.,

the user's preferences, must be specified explicitly by the user.

Complex contextual models for calculating recommendations are presented in [BNWP11], [WY10] and [Yan07]. The incorporated model in [BNWP11] is based on Multi-Criteria Decision Making and Fuzzy Logic, yielding complex algorithms for a pseudo-probabilistic system. These systems do not have a self-reflection mechanism like second-order uncertainty.

There are systems that incorporate an explicit user model. [FPFA11] and [FMA12] define three main information sources for combining these into a recommendation, i.e., input information, context information and profile information. [WE07] defines a contextual recommendation system as a function $U \times I \times C \rightarrow R$, where U is the user and his information, I is a possible recommendation item, C is the contextual information and R is the rating for the recommendation. The recommendation system then chooses the best rated items for the recommendation. [WY10] and [Yan07] use Dynamic Bayesian Networks to model different states of a driver, e.g., the driver fatigue or driver attention.

In all these systems, important parameters are mostly set through expert knowledge, i.e., surveys of experts, and field studies. These systems do not adapt their parameters online to the user's habits, though such techniques exist in machine learning. Some of the presented algorithms use probabilistic, mostly pseudo-probabilistic, methods to overcome the problem of uncertain or missing knowledge. Also, some algorithms do not use any learning algorithms at all. Altogether, most of these systems use simple techniques for inference and data representation, that do not implement a self-reflection mechanism, which is a major issue in automotive recommender systems. This work introduces probabilistic algorithms with self-reflection capabilities and the ability to adapt parameters for an explicit user modelling. Also, our proposed algorithm does not need any prior information from the user and does not require the driver to interact more with the system than activating or deactivating the automated function.

3 Location based Learning

A location based learning system must implement several key concepts, such as observation of the user's actions, user feedback evaluation and the adaption of system parameters for user specific system behavior. A "learning system" in general can be defined as one that makes decisions upon previously gathered knowledge inferred from observations of the user's interaction.When having made enough observations, the system can start working autonomously or at least suggest actions to the user based on his past behavior.

In a vehicular environment, messages from a car data bus, e.g., CAN, serve as observations B. These messages may contain information about sensors, actuators, media data or other information distributed across different electronic control units. Beside location information, these observations also include activations A of functions triggered by the user that we want to automate. We use a probablistic approach to model the likeliness of activations in the system. The overall probability of an activation at an abstract location O (which can be translated to a more general definition of locations such as "at a certain

intersection", "at home", etc.) is then built with

$$p(A|B) = \sum_O p(A, O|B), \tag{1}$$

$$= \sum_O p(A|B, O) \cdot p(O|B). \tag{2}$$

Depending on the value $p(A|B)$, the result can be interpreted as the certainty of the system for an activation A given an observation B. To specify our problem, we make the following assumptions:

1. The probability of an activation A given observation B and location O corresponds to the probability of an activation A given a location O, i. e., the information of an observation does not yield any further information if the location is known.

$$p(A|B, O) = p(A|O) \tag{3}$$

2. If an observation B is very likely at location O_j, all other locations are very unlikely.

$$\forall O_i : p(O_i|B) \approx 0, i \neq j \tag{4}$$

$$p(O_j|B) \approx 1 \tag{5}$$

Taking into account these assumptions, the overall probability for a location O_j can be formulated as the joint probability

$$p(A|B) = p(A|O_j) \cdot p(O_j|B). \tag{6}$$

The following sections describe the elements of this equation in more detail. The term $p(O|B)$, the *Location Distribution*, is described in Section 3.2. The term $p(A|O)$, the *Activation Distribution*, is detailed in Section 3.3. The adaption of these distributions is performed with user *feedback* (see Section 3.1) and is described first as the two models are based on some of the feedback assumptions.

3.1 Feedback

In general, a learning system has to combine two fundamentally different goals:

1. Stability: The system has to reliably recognize the user's wish and increase the security of the recognition over time with additional knowledge.

2. Plasticity: The system has to be able to react reasonably fast to a change in the user's usage pattern.

Consequently, a mechanism which increases its certainty over time (respectively with more observations) has to be used. Furthermore, it has to be able to adapt to changes in the users behavior, even after the system got sure with respect to its stability.

Feedback is the main mechanism to enable learning in our system and it applies to probablistic zones (each modeled with a distribution $p(A|B)$) at a location O which we will call *Activation Zones (AZ)*. The different forms of possible feedback depend on the (re-)action of the user and the current system state. Table 1 gives insight on the possible forms of interaction.

Case Nr.	AZ exists?	System Activation	Driver-(re-)action	Remarks	Adaption $p(A\|O)$	Adaption $p(O\|B)$
1	✗	✗	$\circ$	no learning request	$\circ$	✗
2	✗	✗	$+$	automation request (new AZ)	$++$	✓✓
3	✓	✗	$\circ$	no automation desired	$-$	✗
4	✓	✗	$+$	automation request	$++$	✓
5	✓	✓	$-$	counteraction	$--$	✗
6	✓	✓	$\circ$	acceptance	$+$	✗

Table 1: Feedback matrix: $\circ$ represents no action, $+$ represents positive (re)action, $-$ represents negative (re)action, double symbols represent stronger reactions and represent a feedback strength hierarchy.

If no information about the location O is present in the system, the corresponding cases of the feedback matrix are the cases 1 and 2. The default value of the system is case 1 (no activation, no known AZ), whereas case 2 represents a first activation of the observed activation A, which should lead to the creation of a new AZ to be observed. If the system enters an already observed AZ, cases $3-6$ are evaluated. The value of the activation distribution $p(A|O)$ (see Section 3.3 for details) decides whether the autonomous system activation is triggered or not (✓/ ✗). So, if the system is in case 3 (Activation: ✗, no driver action), the resulting activation probability of $p(A|B)$ will be lowered while the position distribution is not adapted. The same scheme applies to the cases $4-6$.

3.2 Location Distribution

The location distribution describes the probability of the car to be at a certain location O when making an observation B, i. e., the probability distribution $p(O|B)$. In the context of location based learning, locations where an activation A of a function to be automated is observed, are relevant. In general, the location O does not only have to incorporate geographic coordinates, but can be extended by other dimensions in order to define the location more precisely. O therefore can be seen as a state variable of the system. For the sake of explanation, we will stick to just the geographic coordinates here. Observations always occur at a singular position, by itself they do not have any spatial size. Still, they have to tolerate spatial neighborhood in order to be incorporated into a model. To model

abstract locations O, spatial neighborhood is considered in an Activation Zone (AZ) by evaluating the distribution $p(O|B)$.

The distribution $p(O|B)$ depends on the distance between the observed point and the AZ. The further away the observed point from the AZ, the lower is the probability to be inside this zone. Taking into account the above assumptions Eq. 4 and Eq. 5, it makes sense to model each AZ independently. An approximation of the location distribution can be made by using a normal distribution in the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}. \tag{7}$$

In Eq. 7, $\mathbf{x}$ is the coordinate value of the current observation B, μ is the center of the AZ and Σ the covariance matrix describing the spatial form and size of the distribution. Here, $\mathbf{x}$ and μ are assumed to be two-dimensional observations (latitude, longitude), consequently the covariance matrix happens to be a symmetrical 2x2 matrix. For the sake of parameter reduction we simplify $\boldsymbol{\Sigma}$ to be an isotropic covariance matrix with only one degree of freedom in the form

$$\boldsymbol{\Sigma}_{iso} = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}. \tag{8}$$

The distance between the observation and the center of the AZ can then be determined via the Mahalanobis distance $\Delta(\mathbf{x})$ which is determined by

$$\Delta(\mathbf{x}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{9}$$

$\Delta(\mathbf{x})$ can be compared with a threshold ρ in order to determine the belonging of an observation $\mathbf{x}$ to the AZ. An advantage of the normal distribution is that its parameters μ and Σ can be adapted using update formulas, so we do not have to store each observation separately.

The parameter ρ can be chosen by evaluating the χ^2 distribution as the Mahalanobis distance $\Delta(\mathbf{x})$ of a D-dimensional random variable $\mathbf{x}$ ($\mathbf{x} \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \mathbf{x} \in \mathbb{R}^D$) is χ^2_k distributed. k hereby represents the degrees of freedom for the distribution, so in our case $k = D = 2$ when considering 2D coordinates. We can then set ρ to a value which guarantees that a certain percentage of the values for $\mathbf{x}$ lies within the AZ. ρ therefore acts as an upper limit for the Mahalanobis distance.

For our location model we introduce an iterative location refinement which improves its accuracy with an increasing number of observations. If an activation of an observed variable occures and is not assigned to an already existing AZ, a new zone is created (see case 2 of Table 1). The first activation of this AZ (belonging to an abstract location O) does not necessarily have to be at the ideal zone center. Figure 1 (a), (b) visualizes the zone creation. The size of the zone (parameter $\boldsymbol{\Sigma}$) has to be initialized according to the requirements of the application.

If the system performs an automation request (case 4 of Table 1), the parameters of the location will be adjusted. As we can see from the table, only in case 4 of the feedback matrix an adaption of the location is performed. This is done by means of the current

(a) Initial situation

(b) Initialized Activation Zone

(c) Adaption Phase (I)

(d) Adaption Phase (II)

Figure 1: Creation of a new AZ by a single activation of the system. The real zone center (abstract location O) is marked by the black cross, the initialized zone center and size by the red cross and circle. Note that the new AZ (initialized using observation B) is not located at the location of O. Over time, the location center ought to increase its accuracy with the true location O in both center and size of the AZ.

observation $\mathbf{x}$ via a formula incorporating the the previous (two-dimensional) center $\mu_\mathbf{i}$ to calculate the new center $\mu_{\mathbf{i+1}}$. When applying the formula, the center of the AZ is shifted with each iteration.

Likewise, the values a of the covariance matrix $\mathbf{\Sigma}$ are assumed to be $a = \sigma_i^2$, the updated parameter σ_{i+1}^2 is then calculated via an update formula taking into account the previous variance σ_i^2 and the deviation of the current observation $\mathbf{x}$ from the center $\mu_\mathbf{i}$. Note that in both update formulas, only the current values of μ and σ^2 have to be stored, not every single observation of $\mathbf{x}$. It should be remarked that a limit for the minimum size of the AZ can be introduced to avoid overfitting to the data. An example of the iterative adaption of an AZ is shown in Figure 1 (b), (c), (d).

3.3 Activation Distribution

Every Activation Zone (AZ) has the associated position distribution $p(O|B)$ (see Section 3.2) and the activation distribution $p(A|O)$, which is described here. The type of activation distribution chosen is the Bernoulli distribution, which describes a random experiment that only has the outcomes *success* or *no success*. These outcomes can be translated to activation or no activation of the observed function when passing through an AZ. The probability distribution should not only return a probability p in the range $[0, 1]$ for the activation probability, but also be aware of the certainty the distribution has regarding the estimation of p. To additionally take these requirements into account, we use a distribution which is conjugated to the Bernoulli Distribution, the Beta Distribution [Bis06].

The Beta Distribution has two input parameters a and b, which again represent activation respectively no activation. The idea of this second-order technique is to treat the probabilistic parameter p of the Bernoulli distribution as a random variable itself and determine its mean $\mathbb{E}[\mathrm{p}]$ with

$$\mathbb{E}[\mathrm{p}] = \frac{a}{a+b}. \tag{10}$$

Furthermore, the variance of the Beta Distribution is defined by

$$\mathrm{Var}[\mathrm{p}] = \frac{ab}{(a+b+1)(a+b)^2}. \tag{11}$$

$\mathrm{Var}[\mathrm{p}]$ can be used as a measure for the certainty of the expected mean $\mathbb{E}[\mathrm{p}]$.

To recall, learning algorithms should take into account the goals stability and plasticity (Section 3.1). By just summing up (multiples of) a and b depending on the users action when passing through an AZ, we can guarantee stability, but the algorithm falls short on plasticity, i.e., the algorithm cannot react fast on sudden changes in the user's behavior.

To also incorporate plasticity, we introduce an adapted form of feedback handling which weights each observation $P(l)$ (where $l = 1$ is the most current observation) according to their order of occurrence (e.g., using an exponentially decreasing weighting function) with the weighting term $G(l)$ and sums them up to get the respective a and b. Hence, old values will get obsolete, consequently only a fixed number of N observations has to be stored in memory.

4 Case study

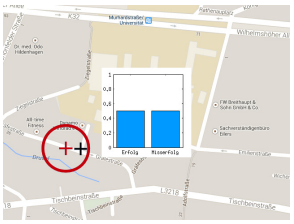
In a case study, the implemented algorithms were evaluated for a comfort function that is typically activated at specific geographic locations. The automated comfort function is a high-class optional equipment mostly sold in packages. For the feature selection, parameter determination and algorithm evaluation, test data had to be recorded in real-world trips. About ten BMW owners, regularly using the selected comfort function, participated in the data recordings, over a hundred trips from every participant were recorded on average. The car owners were instructed about the general goal of the recommendation system and

to use the automated comfort function as naturally as possible. Some subjects not having any technical background, ensured results close to a real-world application of the system. For reasons of data privacy, only the necessary data was filtered on the vehicle bus and then recorded with the designated logging hardware. The data recorded contained the activations and deactivations of the comfort function, filtered and mapped GPS data, date and time information and possibly function related environmental data.

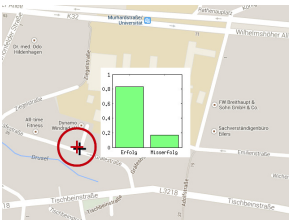
Step	$\mathbb{E}$[p]	Var[p]	Case of Table 1 in buffer $P(l)$, $l = 1, ..., N$									
			1			·	·	·	·			N
1	0.5	0.0227	2	2								
2	0.75	0.0089	4	2	2							
3	0.83	0.0045	4	4	2	2						
4	0.90	0.0018	4	4	4	2	2					
5	0.91	0.0015	6	4	4	4	2	2				
6	0.92	0.0013	6	6	4	4	4	2	2			
7	0.79	0.0024	5	6	6	4	4	4	2	2		
8	0.71	0.0026	3	5	6	6	4	4	4	2	2	
9	0.65	0.0027	3	3	5	6	6	4	4	4	2	2
10	0.64	0.0026	3	3	3	5	6	6	4	4	4	2
11	0.57	0.0028	3	3	3	3	5	6	6	4	4	4
12	0.47	0.0029	3	3	3	3	3	5	6	6	4	4

Table 2: The contents of an observations buffer of an AZ with 10 elements and 12 updates. The buffer is a FIFO buffer with new observations added on the left side.

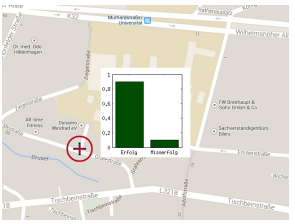
The following case study illustrates how the proposed system updates the activation distribution $p(A|O)$. For the sake of simplicity, we assume our weighting term $G(l)$ to be equally weighted, i.e., $G = [1, ..., 1]$, for the values of our feedback matrix cases we assume values following the feedback strength hierarchy (see Table 1). The example given in Table 2 and Figure 2 both showing the same experiment describes how the system updates the parameters a and b. At initialization, both parameters are set equally, i.e., $\mathbb{E}[\text{p}] = 0.5$, which is realized by adding two different forms of equally strong feedback to the buffer $P(l)$ (see Table 2 (1), Figure 2 (a)). If an activation request (case 4 of Table 1) is observed, the value of a rises while b stays constant (Table 2 (2)-(4), Figures 2 (b), (c)). At the same time, the position distribution $p(O|B)$ is also refined (see Figures 2 (b), (c)). It is worth noting, that the variance, which is given to express parameter uncertainty, decreases the more observations are made by the system. In our case study, we request the activation threshold for $\mathbb{E}[\text{p}]$ to be greater or equal a threshold $\tau = 0.9$, which is achieved in Table 2 (4). The system now automatically activates the comfort function when the driver passes the AZ. If the user accepts the systems behavior, i.e., he does not counteract to it (case 6 of Table 1), the system continues to improve its mean $\mathbb{E}[p]$ (Table 2 (5), (6), Figure 2 (d)). In our case study, the user now changes his usage pattern, counteracting the automatic system activation (case 5 of Table 1). This leads to a decrease of the estimated mean $\mathbb{E}[p]$ below the threshold τ. (Table 2 (7), Figure 2 (e)). When the behavior of the driver changes, $Var[\text{p}]$ increases slightly, expressing increased uncertainty about the estimated mean $\mathbb{E}[p]$. The system will now stop to activate the automated function, if the user continues to not activate the system at the specified location (Table 2 (8) - (12)), $\mathbb{E}[p]$ will continue decreasing (case 3 of Table 1). Over time, the estimated mean $\mathbb{E}[p]$ may get as low as a predefined obsolescence threshold so that the AZ will get deleted if not relevant (Figure 2 (f)).



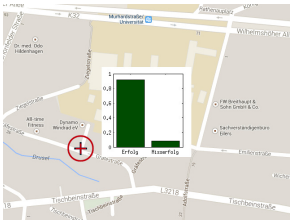
(a) AZ initialization, $\mathbb{E}[p] = 0{,}5$



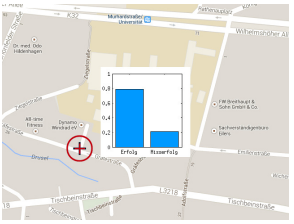
(b) after two automation requests, $\mathbb{E}[p] \approx 0{,}83$, $\mathbb{E}[p] < \tau$



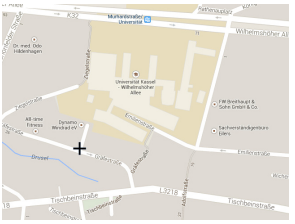
(c) AZ after three automation requests (system now active), $\mathbb{E}[p] = 0{,}90$, $\mathbb{E}[p] \geq \tau$



(d) AZ after two acceptances, $\mathbb{E}[p] \approx 0{,}92$



(e) AZ after counter action, $\mathbb{E}[p] = 0{,}79$, $\mathbb{E}[p] < \tau$



(f) AZ deleted (under obsolescence threshold)

Figure 2: Example of activation updates of an activation zone (AZ). In the histogram of (c), the activation threshold τ is exceeded. In figures (a) – (c) the adaption of the location density distribution is shown.

The processing time is reasonably small, as most of the time only the position distribution $p(O|B)$ has to be calculated (which mainly consists of the computation of the Mahalanobis distance $\Delta(x)$). The computing time scales linearly with the number of AZ. The memory footprint is small enough for embedded use, as only the buffer of activations $P(l)$, the spatial coordinates μ, the AZ size σ and the number of position refinements i have to be stored.

5 Conclusion

The introduced probabilistic system predicts a user activation request for a specified comfort function depending on previously observed activations at points of interest. The probability distribution of the activation and the locations of interest are modelled explicitly. With the use of second-order uncertainty, it implements the necessary self-reflecting mechanism useful for automotive needs. Also, there is no need for the user to explicitly give any information for initialization to the fully automated system. Therefore, this system is appropriate for decision making in a proactive recommender system and can be adapted to any kind of comfort function.

The given evaluations show that the system works efficiently, is numerically stable and has

enough plasticity to work as an embedded, online and life-long learning algorithm. It is also possible to change the model's plasticity and stability further through adjustments to the employed system parameters. Before field operation, the appropriate system parameters have to be discovered through a long-term study with a high number of subjects.

6 Outlook

In future work we will enhance the presented model's targeted set of functions. A focus of this enhancement will be the correlation and dependency of function activations, e.g., the usage of certain heating functions may correlate to a certain temperature or the usage of other comfort functions. As a consequence, the proposed approach for mobile, context based systems in [SBG98] will be applied.

Until now, the model distinguishes only between binary activation states. In future versions of the presented model, specific levels of activation, that in a sense have an ordering, will be handled to automate even more comfort functions. We also aim to extend our model to gradual output levels.

Also, we will enhance the model's observation B with more features. Part of the features incorporated will be the day time and the drive purpose of frequent trips, e.g., the daily trip to work. This should give the model information about the user's control behavior, e.g., the purpose of a trip or even environmental conditions. In future work, more general location models will be examined.

As for now, for complex recommendation models, some system parameters, e.g., decision thresholds, must be set by experts or through knowledge gained from field studies. This sometimes is a vague and expensive process. So the process of choosing the right parameters must be simplified or partially transferred into the field. In general, a wide field of research possibilities are models that change the system parameters online to better accommodate to individual users behavior. One example would be a collaborative filtering approach for online parameter refinement in proactive, automotive recommendation systems.

References

[APRR07] M. Ablaßmeier, T. Poitschke, S. Reifinger, and G. Rigoll. Context-aware information agents for the automotive domain using Bayesian networks. In *Proceedings of the Symposium on Human Interface: Part I*, pages 561–570, 2007.

[Bis06] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Secaucus, USA, 2006.

[BNWP11] R. Bader, E. Neufeld, W. Woerndl, and V. Prinz. Context-Aware POI Recommendations in an Automotive Scenario using Multi-Criteria Decision Making Methods. In *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation, Palo Alto, USA*, pages 23–30, 2011.

[FH06] S. Furao and O. Hasegawa. An incremental network for on-line unsupervised classification and topology learning. *Neural Networks*, 19(1):90–106, 2006.

[FMA12] J. Ferreira, V. Monteiro, and J. Afonso. Electric vehicle route recommender system. In *International Conference on Smart Grids and IT Systems, Chengdu, CN*, volume 1, pages 1–10, 2012.

[FPFA11] J. Ferreira, P. Pereira, P. Filipe, and J. Afonso. Recommender system for drivers of electric vehicles. In *Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT'11), Lisbon ,PT*, 2011.

[GKFC11] O. Gusikhin, E. Klampfl, D. Filev, and Y. Chen. Emotive Driver Advisor System (EDAS). In *Informatics in Control, Automation and Robotics*, volume 89 of *Lecture Notes in Electrical Engineering (LNEE)*, pages 21–36. Springer Berlin Heidelberg, 2011.

[KGT+08] E. Klampfi, O. Gusikhin, K. Theisen, Y. Liu, and Giuli T.J. Intelligent Refueling Advisory System. In *Proceedings of the second International Workshop in Intelligent Vehicle, Madeira, POR*, volume 2, pages 60–72, 2008.

[SBG98] A. Schmidt, M. Beigl, and H. Gellersen. There is more to Context than Location. *Computers and Graphics*, 23:893–901, 1998.

[WBE09] W. Woerndl, M. Brocco, and R. Eigner. Context-Aware Recommender Systems in Mobile Scenarios. *International Journal of Information Technology and Web Engineering*, 4(1):67–85, 2009.

[WE07] W. Woerndl and R. Eigner. Context-Aware, Collaborative Applications for Inter-Networked Cars. In *16th IEEE International Workshops on Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE'07), Paris, FR*, volume 16, pages 180–185, 2007.

[WY10] Q. Wu and W. Yu. A Driver Abnormality Recondition Model Based On Dynamic Bayesian Network for Ubiquitous Computing. In *Third International Conference on Advanced Computer Theory and Engineering (ICACTE'10), Chengdu, CN*, volume 3, pages 320–324, 2010.

[Yan07] J. H. Yang. *Analysis and Detection of Driver Fatigue Caused by Sleep Deprivation*. PhD thesis, Massachusetts Institute of Technology, 2007.

Interface Definition and Code Generation in heterogeneous Development Environments from a Single-Source

A Domain Specific Language Approach for Automotive Predevelopment Environments

Dominik Bauch, David Lenz, Pascal Minnerup and Andrei Avram

fortiss GmbH
An-Institut Technische Universität München
Guerickestr. 25, 80805 Munich, Germany
{bauch,dlenz,minnerup}@fortiss.org, andrei.avram@tum.de

Abstract: Typical automotive software development involves a broad spectrum of development environments and this leads to friction loss, unnecessary efforts and inconsistencies discovered later in the development process. In this paper we advocate that language engineering technologies can be used for the interface definition and code generation in order to glue artifacts and to support the work within such heterogeneous development environments. For our implementation we used the domain specific language stack *mbeddr* based on the *Meta Programming System* (MPS) from JetBrains. We show our preliminary experience with our method gained during the work on a component-based, distributed advanced driver assistant system. The system has to be realized for both the automotive specific prototyping framework ADTF in combination with MATLAB/Simulink and the AUTOSAR RTE middleware. With the help of the proposed approach, we were able to generate large redundant parts of the software system including different interface representations with only small modifications of the existing architecture.

1 Introduction

For the efficient and fast specification as well as implementation of novel software-intensive and highly integrated advanced driver assistance systems the utilization of a broad spectrum of specialized development tools is essential. Virtually all today's advanced driver assistance systems are built on top of a component-based software architecture and hence well-defined interfaces describing the individual's component communication flow are vital. But as an inherent consequence of the in general heterogeneous development environment, most of the interfaces have to be specified in various manifestations to fulfill the specific syntactic and semantic requirements of the respective tools. For example, MATLAB/Simulink[1] represents its interface with so-called bus objects whereas in a C programming environment typically simple records are used (cf. Figure 1). Although, these interfaces will contain semantically identical data they have to be maintained simultaneously

[1] `www.mathworks.com`

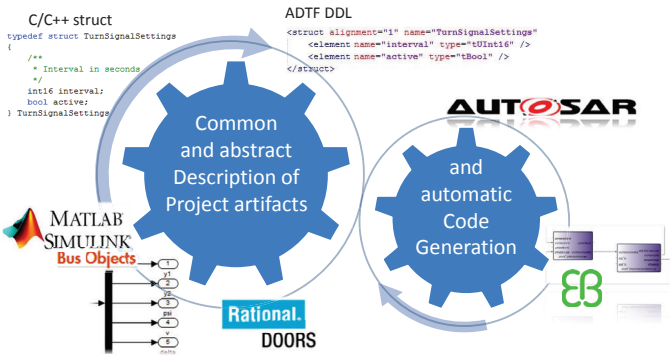


Figure 1: Centralized interface description in affiliation with a code generation framework for ADTF, MATLAB and AUTOSAR

with different tools. With the project's progress the individual interface representations will drift apart and an inconsistent specification is the inevitable consequence.

In addition, typically there are also two parallel and just loosely coupled lines of development which will intensify the above mentioned situation. There is on the one hand a lightweight and agile software engineering process with very short release cycles for the development of the actual functional code of the new advanced driver assistance system. This development branch uses an automotive-specific rapid prototyping environment that allows for an integration of various tools like the *Automotive Data and Time-Triggered Framework* (ADTF). On the other hand there is a more heavy weight systems engineering process fitted for the development of the software's target architecture and ECU itself. Nowadays, the target architecture is typically based on a specific implementation of the well-known *Automotive Open System Architecture* (AUTOSAR) standard. At certain points in the project's timeline, both development branches with their independently evolved interface specifications will be merged together. The better the functional code follows the interface specifications of the target architecture at this moment the less effort is expected.

In our understanding, to reduce costs and risks for maintenance and integration it is indispensable to introduce a centralized and common specification method, which is able to ensure high interface integrity even in such complex, distributed and particularly heterogeneous development environments. We believe also that it is advantageous for the quality of the software system if its specification is decoupled from the specific capabilities of the individual development tools. Within an automotive predevelopment project, there are also some recurring, error-prone and especially time consuming tasks which should be automated as far as practicable. Especially with the help of a suitable code generation framework the mostly trivial source code for the wrappers of the functional software, needed for its embedding within the target architecture as well as within the rapid prototyping environment, can be derived from an interface specification[2]. Figure 1 outlines

[2]Of course, assuming an implementation architecture for the functional code is given

our notion that abstract interface descriptions should serve as a unified representation of concrete interface definitions as well as a source for platform-dependent wrappers.

Contribution of this paper In the following, we will present an approach based on our domain specific language *Rapid Automotive Predevelopment Interface Definition* (RAPID). It is a possible solution to the problem of inconsistencies with interface specifications within heterogeneous and distributed development environments along with component based architectures. Using this domain specific language (DSL) as a kind of glue-code, we are able to establish reliable links between different development and documentation tools whereupon only one common and centralized description of the relevant project artifacts exists. Furthermore, we leverage the information content of RAPID, in conjunction with the knowledge about a project-wide established implementation architecture for the functional code, to generate ready-to-use software component wrappers. This code generation allows the seamless integration of our functional code in the automotive-specific rapid prototyping environment ADTF as well as in an AUTOSAR target architecture.

In the following sections of the paper we introduce the sketched approach in detail and discuss its integration into an automotive predevelopment process. Within Section 2, related work is discussed. Section 3 gives a brief overview of the tools and techniques used for the implementation of our DSL approach. Subsequently, Section 4 describes our domain specific language and illustrates its usage with an example. Furthermore, we present our preliminary experiences and outline the integration into an automotive development process. Finally, the last section summarizes our work and highlights ideas for future tasks.

2 Related Work

Amongst others, we see our observations and hypotheses stated in Section 1 as well as our drawn conclusions in general confirmed by the author of [1]. In this paper, Broy identified and described the most important key challenges future automotive software engineering projects will have to cope with. Some of the fields which are addressed by our paper in a pragmatic way, namely an improved development tool integration and a sophisticated interface specification method together with a fitting requirements engineering, are explicitly mentioned by Broy.

The authors of [2] point out that domain specific abstractions in requirement models can help by the integration of continuous and discrete systems. Whereas models of continuous systems can be used for code generation, the models of discrete systems are very helpful for an unambiguous interface specification and the description of relationships between components (especially when the models are enriched by a behavior specification) but with the main drawback that two artifacts, model and code, have to be maintained and synchronized. The approach proposed in our paper addresses exactly this issue and ensures at all times a coherent and consistent view on the interfaces. Hence, all information about the interfaces' specifications is deposited in one centralized source.

In [3], the authors present a vision on how an integrated model-based tool environment should look like in order to ensure a seamless development process. Especially in the automotive domain, the process is mostly imposed by the tools available and thus many proprietary software and file formats dominate the domain. As each tool is specialized for one development step, models have to be repeatedly (mostly manually) transformed or rebuilt throughout the process. This leads to information-loss, redundancy, inconsistency and thus to inefficiency. Although the authors show how tools have to be designed to overcome these drawbacks, the description is on an abstract level and more a guideline and roadmap towards an improved model-based process. In contrast, our work aims at focusing on a concrete aspect of this process, i.e. the interface definitions, and shows how model-based design can overcome inconsistencies between different branches of development.

An overview about Domain Specific Languages is presented in [4]. As in our approach, the paper identifies maintainability and portability as some of the main advantages of domain specific languages. Additionally they point out that DSLs allow optimization improve the testability. The main challenges of DSLs are seen in implementing a DSL and the availability of DSLs. Our research adds some practical experience about designing and using DSLs.

Such practical experience has also been produced by other research groups. In [5] a DSL is presented that provides real time and validation properties that are particularly useful for embedded computing. The research focuses on designing a language that allows high confidence in the developed software. The approach described in [6] includes generating middleware specific code out of a domain specific XML document. The goal of that paper is to manage the complexity and heterogeneity of distributes online games.

There are several approaches of creating interface definition languages. Well known examples are the AUTOSAR Textual Language Framework [7] or the OMG interface specification language [8]. These interface specification languages can express very complex interfaces, but are either not able to store all information necessary for the code generation presented in this paper or require a large overhead to express it. Furthermore we want to exploit the advantages of language engineering for making a compact and extendible language that can evolve with the predevelopment needs.

3 Language Engineering Tools

For realizing the domain specific language RAPID defined in Section 4 we rely on the existing *Meta-Programming-System* (MPS) from JetBrains [9] and on the language stack *mbeddr* [10]. In this section we briefly introduce these tools and the concepts behind them.

3.1 Meta Programming System

MPS is a language workbench [11], i.e. a framework that allows the definition and extension of domain specific languages and provides tools to work with them.

After defining the syntax and semantics of a language in form of a meta model in MPS, one can also use MPS to write models in this language. All the information of a model is stored in a so-called abstract syntax tree (AST) representing the elements of the syntax of the defined DSL. In contrast to plain text editors in traditional IDEs for programming languages, MPS implements a projectional editor for the language syntax. That means that one modifies the AST directly instead of relying on parsers to generate this tree from plain text. Parsers can be very complicated to build and to maintain. Projectional editing has the advantage that the written code is correct by construction. In other words, only syntactically and semantically correct input is allowed. This is ensured by the type system derived from the meta model. The prevention of incorrect input makes the use of a DSL straight forward and makes it less error-prone and thus usable for domain experts that are not familiar with general purpose programming languages like C.

Within the MPS framework, it is possible to define generators that transform the specified model into arbitrary representations. One of those representations can be conventional type definitions, e.g. C-structs or Matlab bus objects, but also more sophisticated functional (wrapper) code. This approach allows generating different views of the same information for different development environments. This ensures that changes have to be made only for the model and all representations can be regenerated in a coherent manner. This concept of MPS with projectional editing and generation of artifacts can be seen in Figure 2.

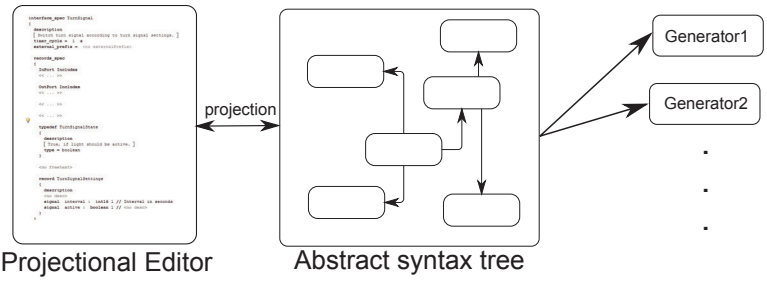


Figure 2: MPS projectional editing concept. The editor changes the AST directly. Generators build different representations of the model.

3.2 mbeddr

mbeddr provides an environment within MPS for embedded software development in C. It offers higher-level extensions for the C-language like state machines, unit tests, etc. as well as static checks and verification mechanisms. It aims at making C extendable without adding additional runtime-costs. For our approach, we use the ability of MPS to extend a DSL in order to use the already defined semantics of datatypes and structures of mbeddr. Also parts of the existing code generation for C can be re-used and extended for our purposes.

4 RAPID — The proposed DSL

The tools described in Section 3 are the basis for defining our DSL RAPID. This section presents the meta model of the interface description. Next, it describes the generation of wrapper code and other development artifacts based on an instance of this meta model. An example will illustrate the whole process.

On the one hand, the meta model of the interface description has to support all information that is necessary for generating the wrapper code and other needed development artifacts of the automotive project. On the other hand, it should remain simple in order to allow fast adaptation to new requirements of the specific project and simple creation of projectional editors. In contrast to general purpose interface definition languages like OMG's IDL [8], the meta model presented in this section is optimized to fulfill these requirements.

Figure 3 shows the meta model of the interface description. The root class is the system specification (*system_spec*). It is the starting point from which a user can reach all other artifacts and it is, in principle, a collection of all component interfaces of the whole software system. The system specification can contain an arbitrary number of these interface specifications (*interface_spec* in Figure 3). Each instance of an *interface_spec* includes the data types used (*records_spec*), the incoming connection ports (*inputports_spec*) and the outgoing connection ports (*outputport_spec*). Finally it contains a mapping specification (*mappings_spec*) that the generator code can use to map the results of triggered input port functions to output ports, handle timer events and store status information.

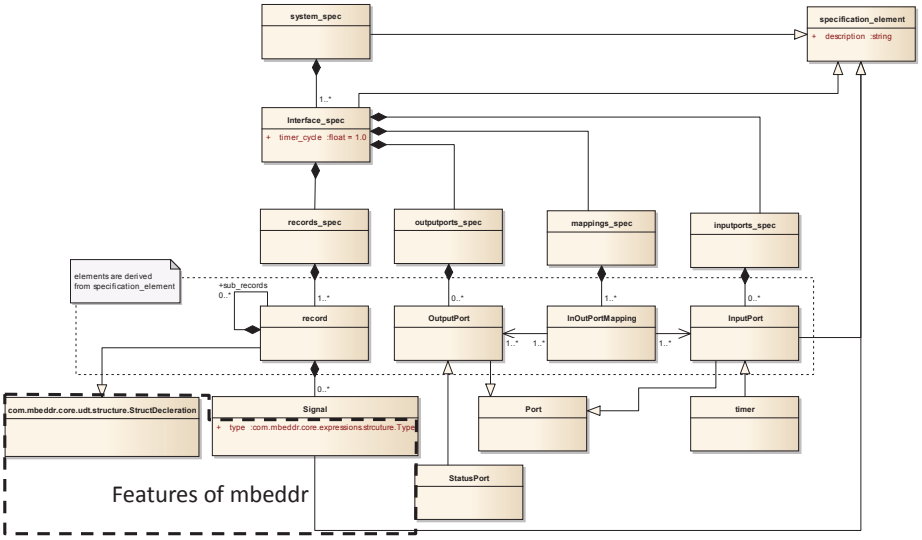


Figure 3: RAPID meta model

The members of the *records_spec*, *outputports_spec*, *mappings_spec* and *inputports_spec* are derived from a common specification element that offers documentation capabilities. A record, which is based on the mbeddr's declaration of structs, contains *Signal*s and

nested records (*sub_records*). Signals refer to some of the built-in data types defined in mbeddr. For the ports used in the *InOutPortMapping*, there are two kinds of special values. The *StatusPort* is a virtual outport that defines the wrapper code should not forward the mapped input directly to the functional code, but store it to be available at any place in the component. The timer is a special inport that is not triggered on receiving data from another component, but triggered periodically.

```
/* Switch turnsignal */
interface_spec TurnSignal
  timer_cycle = 0.1 s
  records_spec
    record TurnSignalSettings
      signal interval : uint16 /* Interval in seconds */
      signal active : bool /* Activate the turnsignal */
  inputports_spec
    port turnSignalSettingsPort: TurnSignalSettings
  outputports_spec
    port turnSignalStatePort: TurnSignalState
  mappings_spec
    mapping timer => turnSignalStatePort
```

Listing 1: Example of an interface specification for a turn signal controller. But in principle due to the concepts of projectional editing the formatting of the DSL can be arbitrary.

Listing 1 shows an example of an interface specification for a turn signal controller. The turn signal controller is a simple component that expects a turn signal setting as an input and switches the turn signal on and off according to this setting. For example the signal setting can specify that the controller should switch the turn signal on and off every half second. The corresponding software component includes an input port for setting the turn signal behavior and an output port stating, whether the light is currently active. Listing 1 defines the types *TurnSignalState* and *TurnSignalSettings* for these ports. Additionally there is a mapping (*turnSignalStateMapping*) stating that at every timer event, the wrapper code should write a new value to the *turnSignalStatePort*. Finally, the interface definition specifies the cycle time of this timer event as 0.1 seconds. Hence, the actual user can command the turn signal controller to switch the turn signal on and off at any interval that is a multiple of 0.1 seconds.

This short and simple RAPID specification suffices for generating the ADTF and potentially AUTOSAR wrappers.[3] In ADTF the software engineer can configure the software system by connecting software components in a filter graph. Each software component is represented as an ADTF filter. As described in the introduction we use this filter as a wrapper for the actual functional code. Figure 4 shows an excerpt of our architecture allowing such a separation of functional code and wrapper code. The *MessagePassingInterface* contains a sending method for each *OutPort* defined in RAPID. Plus, for each *StatusPort*, it offers a getter method for reading the stored value. The ADTF filter as well as the AUTOSAR wrapper has to implement these methods. Additionally, the filters con-

[3] with wrapper we refer to code that embeds the functional code to an AUTOSAR conform environment

Figure 4: Excerpt of our implementation architecture for functional code and corresponding wrappers

tain code for initialization, creation of the pins, processing incoming data and converting it to the target C structs. Code generators can generate all this code using the information of the RAPID files.

The remaining part is the *Functional Component* (compare Figure 4). For this component, a framework is generated including initialization and deinitialization functions plus one receiving function for each input port defined in RAPID. These port functions are also defined in an interface of the functional component. This way, any change of the generated interface is immediately recognized by the compiler and the software developer is forced to update the functional code, too.

Furthermore and as depicted in Figure 4, the *Functional Component* has only a reference to the abstract *MessagePassingInterface*. Hence, it does not know, whether it is using ADTF or AUTOSAR for communicating with other components. Thus, the developer does not have to cope with the implementation details of the specific middleware and they can reuse the same functional code for ADTF and AUTOSAR. Only for the final serial deployment the software engineers should remove the abstraction layers introduced for platform independence and optimize the functional code. In the example used for this paper, the output port *turnSignalStatePort* is mapped to a timer. Therefore, the generated interface for the functional code also includes a method with *TurnSignalState* as return type that the wrapper code will trigger periodically.

All in all, the code generation replaces the error-prone activity of writing more than 400 lines of repeating code by hand. These 400 lines of code do not include the completely identical functions that a software engineer can implement in a common base class for all ADTF filters.

Besides the in detail explained generation of platform dependent wrapper code, there exist also several artifacts in a heterogeneous development environment describing the same interface:

- Software systems implemented in C/C++ need C interfaces,
- components implemented in MATLAB/Simulink need Simulink Bus Objects,
- debugging and connection tools in ADTF need ADTF data definition language (DDL) files and
- the project management needs to include the interfaces in the requirements specification, amongst others, to make contracts with suppliers for a later series development.

```
<adtf:ddl xmlns:adtf="adtf">
  <header><description>Switch turnsignal</description></header>
  <structs><struct name="TurnSignalSettings">
      <element ... name="interval" type="tUInt16"/>
      <element ... name="active" type="tBool"/>
  </struct></structs>
</adtf:ddl>
```

(a) ADTF description of the record TurnSignalSettings

```
typedef struct TurnSignalSettings
{
  int16 interval; // Interval in seconds
  boolean active; // Activate the turnsignal
}TurnSignalSettings;
```

(b) C description of the struct TurnSignalSettings

	A	B	C
1	**Name**	**Type**	**descritpion**
2	TurnSignalSettings.interval	uint16	Interval in seconds
3	TurnSignalSettings.active	boolean	Activate the turnsignal

(c) Tabular representation of TurnSignalSettings

Figure 5: Different representations of the TurnSignal interface

Without code generation, developers have to maintain all these semantically equal descriptions manually and inconsistencies are unavoidable. The interface description language presented in this paper solves this problem by creating a single source of information as we are able to generate all the above listed artifacts with the help of RAPID. Listing 5a shows how the struct *TurnSignalSettings* is represented in ADTF with its DDL format, the C/C++ version is depicted in Listing 5b and Figure 5c shows a tabular representation that a project leader can use for importing it to a requirements management tool. The implementation of the corresponding generation files is straight forward, as each of these representations mainly contains the same information. If a new representation is necessary that contains additional information, developers can add this information to the RAPID meta model. This way, RAPID offers the tools necessary for keeping many development artifacts consistent.

4.1 Experience with the code generation

We tested the code generation on our sub-project, a part of a distributed advanced driver assistant system. The sub-project consists of a total of 17 components with 119 ports. For developing the code generation we found several places in the source code that we had to change, because we had implemented slightly different architectures for similar wrapper code. The code generation enforced to implement a common architecture. Some of these architectural inconsistencies were even undetected defects in the source code. Further-

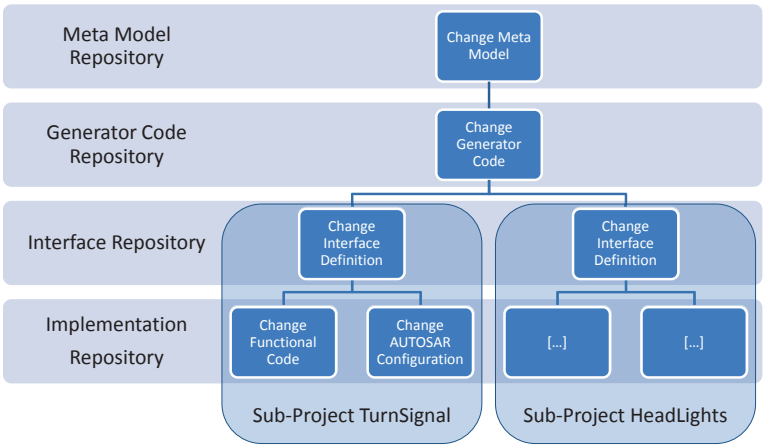


Figure 6: Hierarchy of RAPID artifacts. The higher a modified artifact is located in the hierarchy, the higher the impact on other teams.

more, we had to rename variables and some namespaces to a common naming pattern that code generators can describe and update our C include hierarchy to a systematical pattern. As we had already strictly separated functional code and wrapper code no other changes were necessary.

4.2 Integration into the Development Process

As described in the introduction, there are typically two or more independent development lines and many independent teams using different development tools like MATLAB/Simulink, ADTF or an AUTOSAR tool chain. The previous section described how RAPID code generation can generate the different artifacts needed by these teams. The process engineers should integrate this code generation into the development process. The following section describes this integration.

Figure 6 visualizes the general principle of integrating the interface definition and code generation framework into the development process. The engineers working on a specific development line (i.e. the functional prototype of the sub-project *TurnSignal* depicted in Figure 6) should not directly change the generated interfaces or other generated code. If possible, developers should make the changes in the non-generated functional code: the implementation repository level shown in Figure 6. If the change requires modifying the interfaces, they should change the RAPID files and regenerate the language specific generated interface code. Automatic build scripts should include the regeneration and the developers should check in RAPID files to a common version control repository. As soon as the other teams update and rebuild their projects, the automatic build script will generate their language and domain specific interfaces, too (compare common box "'Change

Interface Definition"' in Figure 6). If the interface change causes a problem for one of the projects, this is recognized at this early stage and not postponed to the integration process. Generating the code with mbeddr and build scripts is possible by using ANT scripts[4]. In order to ensure that no developer uses old generated code files, version control should exclude the generated code.

In some cases changing the RAPID description will not suffice, because the information that the developer needs to change is contained in the code generators. The code generation framework should be a separate version controlled project that is included into other projects using for example git sub modules[5]. As the code generation is included in an automatic build script, changes to the generators will have an effect on all other teams depending on it very quickly. As indicated in Figure 6, a large number of projects might depend on these generators.

The same applies to the project management information specified in a requirements management tool. If a bad specification is found in the specification file, requirements engineer should not make the change directly in the requirements management tool. Instead, they should make the change in the RAPID files and regenerate the requirements specification document. This ensures consistency between the functional software and the requirements and hence the project team can use the working prototype as a proof for a good requirements specification.

5 Conclusions and Future Work

In this paper we presented the domain specific language RAPID together with a powerful code generation framework that has the capability to better integrate tools of a heterogeneous development environment. We showed that RAPID can act as a single-source for required artifacts within a software-intensive industrial automotive pre-development setting. As a result, no more information about interface specifications is duplicated and hence modifications have only to be done at one dedicated place. Consistency in the resulting interface specification for the individual tools is built in. Furthermore, language engineering is able to embrace the concepts of the developer's domain. Therefore, in our context, using language engineering for the system specification is a powerful and at the same time very intuitively usable to instrument.

As shown in Section 4.2, such an approach can be seamlessly embedded in an already established distributed development process. One reason for this is the very compact and natural (in terms of our domain) design of RAPID that allows easy and pretty fast migration of existing code. Furthermore, the possibility of code generation out of the domain specific language is particularly easy with the leveraged framework because there is no need to parse complicated syntactical constructs thus the design of the language can be chosen arbitrarily.

Up to our measurements, we have reduced the amount hand-written code, which is nec-

[4] http://confluence.jetbrains.com/display/MPSD30/HowTo+--+MPS+and+ant
[5] http://git-scm.com/book/en/Git-Tools-Submodules

essary for the interface specification and the implementation for a wrapper (filter) for the ADTF environment in our setting, by a factor of eight. We assume that the factor in terms of working hours is even higher due to fact that developers can focus on the actual important things of their work.

We are currently extending our code generation framework so we will be able to generate AUTOSAR *RunnableEntity*s from a RAPID interface specification. This will be done in the same fashion as we did for the ADTF filters. With this extension we can embed our functional code within the rapid prototyping environment of the agile line of development as well as within the systems engineering line of development. This is achieved fully transparently and without any additional effort. Additionally, an implementation of the proposed connection to requirements documentation tools, e.g. DOORS from IBM Rational, as well as to the tools used for the ECU system specification is a valuable goal which will increase productivity and in particular quality of the overall process and, by implication, the product's quality more and more. Finally, we will assure our position about saving working hours with the help of a case study done within our project about a highly integrated advanced driver assistance system.

References

[1] M. Broy, "Challenges in automotive software engineering," in *Proceedings of the 28th international conference on Software engineering*, pp. 33–42, 2006.

[2] A. Pretschner, M. Broy, I. H. Kruger, and T. Stauner, "Software engineering for automotive systems: A roadmap," in *2007 Future of Software Engineering*, pp. 55–71, 2007.

[3] M. Broy, M. Feilkas, M. Herrmannsdoerfer, S. Merenda, and D. Ratiu, "Seamless model-based development: From isolated tools to integrated model engineering environments," *Proceedings of the IEEE*, vol. 98, no. 4, pp. 526–545, 2010.

[4] A. Van Deursen, P. Klint, and J. Visser, "Domain-specific languages: An annotated bibliography.," *Sigplan Notices*, vol. 35, no. 6, pp. 26–36, 2000.

[5] K. Hammond and G. Michaelson, "Hume: a domain-specific language for real-time embedded systems," in *Generative Programming and Component Engineering*, pp. 37–56, 2003.

[6] T.-Y. Hsiao and S.-M. Yuan, "Practical middleware for massively multiplayer online games," *Internet Computing, IEEE*, vol. 9, no. 5, pp. 47–54, 2005.

[7] Artop User Group, "ARText — An AUTOSAR Textual Language Framework." `https://www.artop.org/`.

[8] Object Management Group, "OMG formal specifications." `http://www.omg.org/spec/`.

[9] Jetbrains, "Meta programming system (MPS)." `http://www.jetbrains.com/mps/`.

[10] M. Voelter, D. Ratiu, B. Schaetz, and B. Kolb, "mbeddr: an extensible c-based programming language and IDE for embedded systems," in *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pp. 121–140, 2012.

[11] M. Fowler, "Language workbenches: The killer-app for domain specific languages," 2005.

Automated Generation of AUTOSAR Description File for Safety-Critical Software Architectures

Georg Macher
Institute for Technical Informatics
Graz University of Technology
georg.macher@tugraz.at

Eric Armengaud
AVL List GmbH
eric.armengaud@avl.com

Christian Kreiner
Institute for Technical Informatics
Graz University of Technology
christian.kreiner@tugraz.at

Abstract: Automotive embedded systems have become very complex, are strongly integrated, and the safety-criticality of these systems pose new challenges. Distributed system development, short time-to-market intervals, and automotive safety standards (such as ISO 26262) require efficient and consistent product development along the entire development lifecycle. The de-facto industry standard AUTOSAR aims to standardize an open automotive software architecture and framework to facilitate the exchange of information across company boundaries for the software development process. However, providing consistency of the safety concept during the entire product life cycle is a tedious task. The aim of this paper is to enhance a model-driven system and safety-engineering framework with AUTOSAR aligned software-architecture design. This approach is part of a tool-chain solution enabling the seamless description of safety-critical systems, from requirements at the system level down to software component implementation. To that aim a tool bridge is proposed in order to seamlessly transfer artifacts from system development level to software development level based on AUTOSAR exchange format files.

1 Introduction

Embedded electronic control systems are strong innovation drivers for the automotive domain. The introduction of these systems enables the deployment of more advanced control strategies, such as better driveability and driver assistance systems. With the replacement of well-established mechanical systems by electronic systems, beginning with throttle-by-wire in early 1990s, electronic control systems became the main driver of innovation in the automotive domain. At the same time, the higher degree of integration and the safety-criticality of the control application poses new challenges. The independence of different applications (with different criticality levels) running on the same platform must be made evident. In parallel, new computing architectures with services integrated in hardware enable the development of new software architectures and safety concepts.

Safety standards such as ISO 26262 [ISO11] for electrical and electronical systems for road vehicles have been established to provide guidance during the development of safety-critical systems. They provide a well-defined safety lifecycle based on hazard identification and mitigation, and they define a long list of work-products to be generated.

One important challenge in this context is to provide evidence of consistency during product development among the different work-products.

The contribution of this paper is to bridge the existing gap between model-driven system development tools and software engineering tools. More specifically, the approach relies on the automated generation of software architecture description files based on the existing high level control system description (e.g., based on SysML format). This description includes specific system level information such as ASIL or trace to related (safety) requirement. The approach relies on the AUTOSAR [AUT09] aligned integration of software architectures into the system development tools and the standardized information exchange format of AUTOSAR. The goal is to support a consistent and traceable refinement from the early concept phase to single safety-critical software component implementation.

The document is organized as follows: Section 2 presents an introduction to AUTOSAR. Then, model-based development and integrated tool chains are presented in Section 3. In Section 4 a description of the proposed approach for the generation of software architecture description file according to AUTOSAR standard is provided. An application of the approach is presented in Section 5. Finally, this work is concluded in Section 6 with an overview of the presented approach.

2 AUTOSAR Overview

Several approaches deal with model-based system development and AUTOSAR software development. Different tool vendors provide AUTOSAR tool chains, which support different development stages of AUTOSAR aligned development.

Nevertheless, in terms of safety-critical development, artifact traceability, and support of ISO 26262 safety features there is still room for improvement and ongoing development, e.g. [Eis11, SS12]. Several configurations of AUTOSAR basic software modules and ECU description files still need to be done manually with several non-interacting tools.

The deployment of the AUTOSAR paradigm in projects is usually resource and cost intensive due to (a) the paradigm change to component-based design and related new activities such as interface description in XML format, (b) the complexity of the standard and of the tools with the large number of components and configuration parameters, and (c) the high costs for configuration tools and basic SW layers proposed by the suppliers. In order to support the AUTOSAR deployment, three Implementation Conformance Classes (ICC) have been proposed. These Implementation Conformance Classes have been introduced to smooth the changing process from conventional software development to AUTOSAR software development and are fully in line with the AUTOSAR standard [AUT09].

The first class, ICC1, introduces (a) software component development according to the AUTOSAR standard, a feature supported by almost all common software development tools, and (b) the AUTOSAR runtime environment (RTE). This feature already supports two of the main benefits of AUTOSAR aligned development, which are hardware independent software development, and allocation of dedicated processing cores at a very late development state. Basic software drivers and operating systems features remain unchanged and can be reused as they stand. Solely an interface wrapper for RTE standard interface needs to be adopted. This approach is frequently used when getting started with AUTOSAR, when porting applications to an AUTOSAR environment, or when the project

budget does not support full AUTOSAR tooling. Additionally, this approach is also perfectly suitable for introducing new technologies (e.g. new multi-core technologies) or if special hardware features (e.g. specific safety and security hardware modules) are not yet available in standard AUTOSAR implementation.

On the other hand, AUTOSAR Implementation Conformance Classes 3 (ICC3) implies full AUTOSAR conformity, implementation of the AUTOSAR interfaces, and standards for all structures of the software architecture. Furthermore ICC3 requires a higher level of granularity of the software architecture. This approach requires complete AUTOSAR tooling and vendor support. Nevertheless, ICC3 gives maximum flexibility and highest accuracy of the interface specifications.

ICC2 is an intermediate step between ICC1 and ICC3. Vertical subdivisions are already introduced and monolithic blocks of the basic software are already separated. The clustering on ICC2 level is not restricted by AUTOSAR and could lead to different optimizations. This ensures optimization strategies in terms of reduced execution time and memory footprint. ICC2 clustering can be used anytime when ICC3 structuring leads to non-negligible overhead and flexibility of ICC3 approach is not required.

The main benefit of the ICC1 approach clearly relies on the time-saving in terms of no additional familiarization with usually very complex and time-consuming AUTOSAR tools. Nevertheless, this approach does not take advantage of the AUTOSAR benefits of standardized component interfaces for exchange of components, supporting tools for RTE configuration, and communication interfaces. But in terms of safety-critical software development special attention has to be paid to the manual interface description (such as Excel files), manually coded interfaces, and manual configurations (see the parts with 'manual rework' marking in Figure 1). Manual document changes are difficult to trace and therefore require additional alertness to be paid in terms of safety-critical system development.

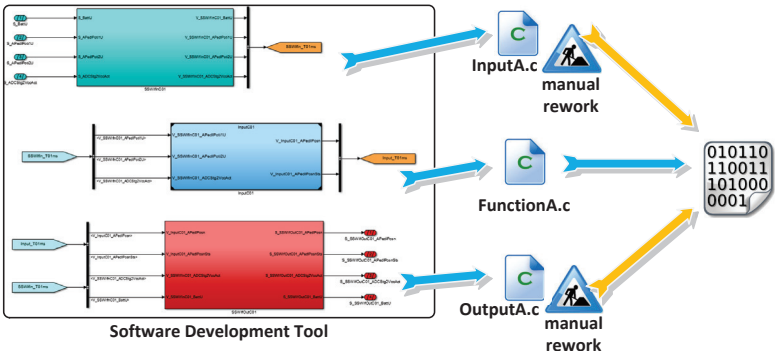


Figure 1: ICC1 AUTOSAR Approach Methodology with Required Manual Intervention

ICC1 mainly focuses on SW engineering and more specifically on the integration of ASW into a given SW architecture. However, the aspects related to control systems engineering (including HW/SW co-design) are not integrated and aspects such as HW/SW interface definition must be performed manually. The proposed approach in this work enhances this aspect and provides a framework for the visualization of interface wrapper configuration

and automated generation of the interfacing C files. Furthermore, the approach enhances the standard AUTOSAR software component description with links to involved development artifacts of prior development processes (e.g. traces to requirements, ASIL level, additional constraints).

3 Model-based Software Development and Integrated Toolchains

Model-based Systems and Software development as well as tool integration are engineering domains and research topics aimed at moving the development steps closer together and thus improving the consistency of the system over the expertise and domain boundaries. Broy et al. [BFH+08] mention concepts and theories for model-based development of embedded software systems. The authors also claim model-based development the best approach to manage the large amount of information and complexity of modern embedded systems with safety constraints. The paper illustrates why seamless solutions have not been achieved so far, they mention commonly used solutions, and arising problems by using an inadequate tool-chain (e.g. redundancy, inconsistency and lack of automation).

Chen et. al. [CJL+08] presents an approach that bridges the gap between model-based systems engineering and the safety process of automotive embedded systems. The systematic approach uses the EAST-ADL2 architecture description language to develop safety cases in close relation to the system model and analysis of malfunctions. Although the work provides a system model for keeping various engineering information across multiple levels of abstraction and concerns consistent, their approach ends just before the definition of software architecture design. More recently the MAENAD Project [1] is focusing on design methodologies for electric vehicles based on EAST-ADL2 language.

The work of Holtmann et al. [HMM11] highlights process and tooling gaps between different modeling aspects of a model-based development process. Often, different specialized models for specific aspects are used at different development stages with varying abstraction levels. Traceability between these different models is commonly established via manual linking. The authors claim that there is a lack of automation for those linking tasks and missing guidance which model should to be used at which specific development stage. The proposed tool-chain mentions two important gaps: First, missing links between system level tools and software development tools. Second, several very specific and non-interacting tools which require manual synchronization, are therefore often inconsistent, rely on redundant information, and due to a lack of automation require redundant manual work.

This issue is also addressed by Giese et al. [GHN10]. System design models have to be correctly transferred to the software engineering model, and later changes must be kept consistent. The authors propose a model synchronization approach consisting of tool adapters between SysML models and software engineering models in AUTOSAR representation. One drawback of this approach stems from the bidirectional model transformation, each transformation step implies potential sources for ambiguous mapping and model mismatching.

An important topic to deal with is the gap between system architecture and software architecture - especially while considering component-based approaches such as UML and

[1] http://maenad.eu/

SysML for system architecture description and AUTOSAR for SW architecture description.

Pagel et al. [PB06] mention the benefit of generating XML schema files directly from a platform-independent model (PIM) for data exchange via different tools. Performing extra transformation steps would only add potential sources for error and ambiguous mappings could result in unwanted side-effects. We also spotted a potential drawback for the previously mentioned approach of Giese et. al. [GHN10].

Boldt [Bol09] proposed the use of a tailored Unified Modeling Language (UML) or System Modeling Language (SysML) profile as the most powerful and extensible way to integrate an AUTOSAR method in company process flows. The author highlights the option of UML to link requirements to ECU, SWC or runnables.

An automotive tool-chain for AUTOSAR is also presented by Voget [Vog14]. The work focuses on ARTOP, a common platform for innovations which provides common base functionality for development of AUTOSAR compliant tools. Unfortunately the Eclipse based ARTOP platform serving only as a common base for AUTOSAR tool development, is not a tool solution, and also requires time-consuming initial training to even get started to develop a desired tool.

Among these, we evaluated several commercial available tools. The following paragraph provides a brief overview of tools supporting the AUTOSAR approach, this list is not intended to be exhaustive. The tools Matlab/Simulink and Embedded Coder are widely used for automotive software development. It is possible to import and export AUTOSAR software component descriptions and generate AUTOSAR software code in an integrated environment with both tools. However, this tool focuses solely on software development. Dassault System AUTOSAR Builder is a software platform for system and ECU (Electronic Control Unit) design, based on the ARTOP tool environment. The tool suite imports model-based design (MBD) descriptions and generates AUTOSAR compliant C code. The embedded software platform Arctic Core includes a real-time operating system, communication services, memory services, and drivers for different microcontroller devices. Continental's CESSAR-CT Tool-Box integrates the AUTOSAR methodology in an already existing process landscapes. The modular tool includes several editors, an AUTOSAR conformance validation, and code generation framework. Elektobit offers a complete product line, called EB tresos. One part of this product line, EB tresos AutoCore, is an AUTOSAR compliant basic software, supporting AUTOSAR compliant OS for single and multicore ECUs. ETAS AUTOSAR Solutions are designed to cooperate with ETAS's development tools, such as ASCET, INCA, and others. Vector's AUTOSAR tool chain consists of PREEvision, DaVinci Developer, and DaVinci Configurator Pro. These tools provide tool support for model-based development, AUTOSAR basic software, and RTE configuration.

Nevertheless, these tools mainly target the AUTOSAR ICC3 approach and usually focus on the configuration and analysis of the AUTOSAR stack and SW architecture. The proposed contribution in this paper is (a) the formalization of control system description (including BSW and HW) especially interesting for ICC1 and (b) the import / export to AUTOSAR XML files for the traceability of architecture artifacts such as ASIL or timing constraints from system description down to SW implementation, which are interesting for all ICCs.

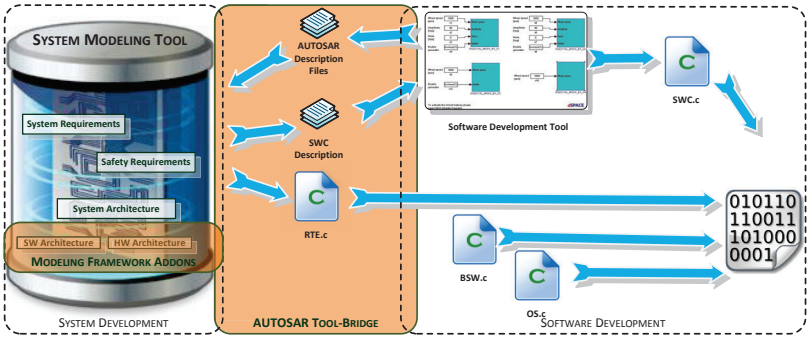


Figure 2: Portrayal of the Bridging Approach Transferring System Development Artifacts to SW Development Phase

4 AUTOSAR Tool-Bridge for Model-based Software Development Approach

The basis of our approach stems from the CESAR Project [RW12] and was further improved by a feasibility study [Mad12] to support continuous safety related system development according to ISO 26262 at concept phase and system development level. For a more detailed overview of the tool-chain as a whole see [MAK14] and [Mad12].

The contribution proposed in this work is an extension of this framework towards software development in the context of AUTOSAR. More specifically, our contribution consists of the following parts:

- *AUTOSAR UML modeling framework*: Enhancement of an UML profile for the definition of AUTOSAR specific artifacts, more precisely, for the definition of the components interfaces (based on the virtual function bus). This is required for consistent SW system description, see Figure 2 – modeling framework addon.
- *BSW and HW module modeling framework*: Enhancement of an UML profile to describe BSW components and HW components. This is required to ensure consistency of the specification and implementation for the entire control system, see Figure 2 – modeling framework addon.
- *SW architecture exporter*: Exporter to generate the resulting SW architecture as AUTOSAR XML files for import in third party tools for further detailed development, see Figure 2 – AUTOSAR tool bridge.
- *AUTOSAR file importer*: Importer to integrate refined SW architecture as AUTOSAR XML file (e.g., as a result of round-trip engineering), see Figure 2 – AUTOSAR tool bridge.

This proposed approach closes the gap, also mentioned by Giese et al. [GHN10], Holtmann et al. [HMM11], and Sandmann and Seibt [SS12], between system-level development at abstract UML-like representations and software-level development modeling tools (e.g.

Matlab Simulink/Targetlink). This bridging supports consistency of information transfer between system engineering tools and software engineering tools. Further, the approach minimizes redundant manual information exchange between these tools and contributes to simplify seamless safety argumentation according to ISO 26262 for the developed system. Benefits of this development approach are highly noticeable in terms of re-engineering cycles, tool changes, and reworking of development artifacts with alternating dependencies, as also mentioned by Broy et al. [BFH+08]. Closing this gap creates a seamless tool-chain from initial design, to software architectures in a model-based development environment and final decisions in code implementation in compliance with ISO 26262. Our approach supports the ICC1 AUTOSAR approach and relies on the AUTOSAR specification [AUT09](currently implementing AUTOSAR $R3.2$ due to compatibility with Matlab 2011 based SW development tools) for architectural approach, definition of application software interfaces, and exchange formats.

4.1 AUTOSAR UML Modeling Framework

The first contribution is the development of a specific UML modeling framework enabling software architecture design in AUTOSAR like representation within the system development tool (Enterprise Architect). This EA profile takes advantage of the AUTOSAR virtual function bus (VFB) abstraction layer and enables an explicit definition of AUTOSAR components, component interfaces, and connections between interfaces. This provides the possibility to define software architecture and ensures proper definition of the communication between the architecture artifacts, including interface specifications (e.g. upper limits, initial values, formulas). In addition to standard VFB AUTOSAR artifacts our profile supports graphical representation of ASIL, assignment to dedicated signals and modules, and supports specification of runnables with respect to timing constraints (such as WCET), ASIL, priority, and required stack sizes. This meta information enables mapping of tasks to a specific core and establishment of a valid scheduling in a later development phase.

The proposed approach thus supports the ISO 26262 requirements of traceability along the development process, even for ICC1 AUTOSAR development. Hence, the AUTOSAR-aligned representation can be linked to system development artifacts and traces to requirements can be easily established. These explicit links can be further used for constraints checking, traceability of development decisions (e.g. for safety case generation), and reuse. Figure 3 shows an example of a safety-relevant software module (AUTOSAR Composition) and its ASIL decomposition in two components with lower ASIL levels, represented in Enterprise Architect. This integrated definition of system artifacts and software module in one tool furthermore supports the work of safety engineers by adding values and visual labels for safety-relevant software modules.

4.2 BSW and HW Module Modeling Framework

Special basic software (BSW) and hardware module representations are assigned to establish links to the underlying basic software and hardware layers. The AUTOSAR architectural approach ensures hardware-independent development of application software modules until a very late development phase and therefore enables application software developers and basic software developers to work in parallel. Nevertheless, safety-critical sys-

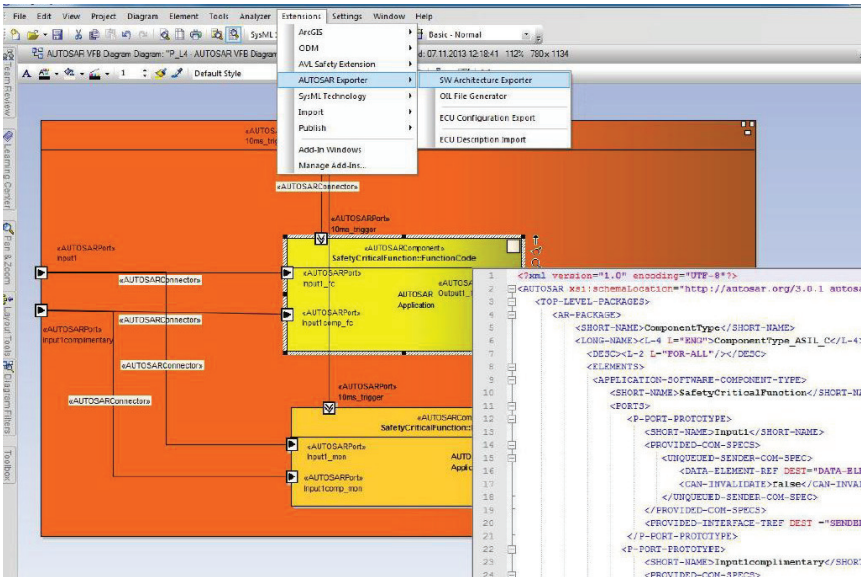


Figure 3: Screenshot of the SW Architecture Representation within the System Development Tool and Extension of Bridging Approach

tem development concerns also hardware development and support for hardware-software co-design is absolutely essential. The hardware profile allows a graphical representation of hardware resources (such as ADC, CAN), calculation engines (core), and connected peripherals which interact with the safety-critical software. This additional information enables the mapping of tasks to a specific core and establishment of a valid scheduling in a later development phase. Furthermore, the profile enables an intuitive graphical way of establishing software and hardware dependencies and a hardware-software interface (HSI), as required by ISO 26262. In combination with the ICC1 AUTOSAR development approach this profile enables the possibility of a traceable automatic RTE configuration generation instead of the typical manual software component interface definition.

4.3 SW Architecture Exporter

The third part of the approach is an exporter which is able to export the software architecture, component containers, and their interconnections designed in SysML to a software development tool (e.g. Matlab/Simulink) via AUTOSAR XML files (see Figure 3). Most of the state of the art software development tools are able to generate AUTOSAR-conform code and software component description files or support the import of AUTOSAR modules. This ensures flexibility of the approach in terms of the preferred software development tool in use (e.g. Matlab/Simulink or ETAS ASCET) and ensures tool-independence of the presented approach.

The exporter generates an AUTOSAR conform software component description files (currently AUTOSAR $R3.2$ to be compatible with Matlab 2011) enriched with system and

safety development artifact traces. Information that is not importable by default AUTOSAR import functions of third-party tools is transferred via description and long-name values of individual models and is therefore still available for the user of this particular tool. Using this exporter, the (safety) context can be efficiently exported and communicated to the software experts in their native development tools, thus improving the consistency of the product development.

4.4 Import of AUTOSAR Files

The fourth part of the approach is the import functionality add-on for the system development tool. This functionality, in combination with the export function, enables bidirectional update of software architecture representation in the system development tool and the software modules under development. The importer also re-imports additional information from the ARXML file stemming from software development level. On the one hand, this provides input for the previously mentioned timing estimations of task, and on the other hand it ensures consistency between system development artifacts and changes done in the software development tool. Finally, the import functionality enables reuse of available AUTOSAR software modules, guarantees consistency of information, and shares information more precisely and less ambiguously.

5 Application of the Proposed Approach

This section demonstrates the benefits of the introduced approach in terms of ISO 26262 aligned development. As a first step of ISO 26262 related safety-critical system development, the boundaries of the system and its interacting environment must be specified. For each system on each level of abstraction, system targets (requirements and use cases) and a system structure can be refined. The system design is completed by the definition of the hardware-software interface (HSI). This mapping provides a basis for concurrent development of software and hardware.

The definition of the software architecture is usually done by a software system architect within the software development tool (such as Matlab/Simulink). With our approach this work package is included in the system development tool in an AUTOSAR-aligned representation (already depicted in Figure 3). On one hand, this does not hamper the work of the software system architect and the AUTOSAR format based exporter ensures consistent transferring to the software development tool, in a way that the software unit design remains unaffected. On the other hand this change offers a significant benefit for development of safety-critical software in terms of traceability and replicability of design decisions.

The presented approach bridges the existing lack of tool support for transferring SW architectures between system design and software implementation tools. Due to the basis of information transfer via standardized AUTOSAR exchange formats this tool-independent approach can also be used to link additional tools to the development tool-chain. Furthermore, the approach guarantees traceable links of safety concept considerations throughout the entire development cycle, due to the single source of information principle (all relevant information and design decisions within the system development tool), as well as improved re-useability of software modules by adding information and safety constraints to

the AUTOSAR software component description file. Figure 4 depicts the add-on of traceable artifact links of the approach for a specific subsystem. Required ASIL and related requirements can be additionally stored within the ARXML file, as well as information about other dependencies (e.g. required HW resources or required core allocation diversity). The figure illustrates the transfer of artifacts between the separated development phases of SW development and system development, and indicates how traceability can be supported.

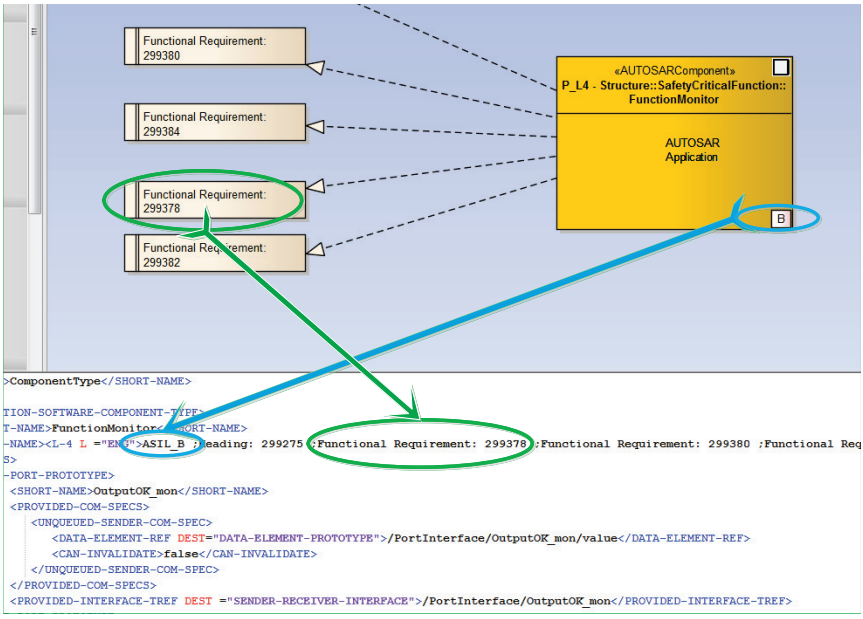


Figure 4: Traceability of Artifacts between System Engineering and SW Development

These AUTOSAR software module description files can be imported by software development tools (such as Matlab/Simulink), thus minimizing the effort of error-prone manual work without adequate tool support. The possibility of re-importation of information from changed AUTOSAR software component description files into the system model ensures round-trip engineering and consistency of the implementation and the system model. The previously mentioned definition of hardware-software interface enables the automatic generation of interface wrapper files (*.c* and *.h* files), thus also reducing the amount of manual changes of source code files without adequate tracing of changes.

To provide a comparison of the improvements of our approach we defined the three layer E-Gas monitoring functionality accordingly [ZS07] as use-case. This elementary use-case is well-known in the automotive domain, but is nevertheless representative in our opinion, due to the fact that several safety-critical systems base on this approach. Table 1 gives an overview of the improvements indicated compared to AUTOSAR ICC1 approach.

In terms of getting started with AUTOSAR aligned development our approach does not rely on full AUTOSAR tooling support, but rather features the smooth first step approach of the ICC1 AUTOSAR approach. In terms of safety-critical development the presented

approach supports round-trip engineering by tool-supported information transfer between separated tools and links to supporting safety-relevant information. Furthermore, the approach eliminates the need of manual interface source code rework and ensures reproducibility and traceability arguments for this task. These indicators infer that the presented approach surmounts the main drawbacks of the ICC1 AUTOSAR approach.

Evaluation Criteria	**ICC1 AUTOSAR Approach (reference)**	**AUTOSAR Tool-Bridge Approach**
11x Interface definitions (consisting of limits, data type, scaling, unit, default value)	usually done with SW engineering tool, no representation within system development tool	graphically with option list support
Interface consistency evaluation	needs to be done manually at review without tool support	automated consistency checks possible for point-to-point connections
SW Architecture definition (3 levels, 7 modules, 30 connections)	usually done with SW engineering tool, no representation within system development tool	graphically establish-able within system development tool, automatic export to ARXML
SW Architecture update	usually no representation within system development tool	import and export functionality within system development tool
ASIL assignment	not supported	for each module and each BSW mapping possible
RTE configuration	manually	mapping automatically generated

Table 1: Approach Improvement Indicators

6 Conclusion

An important challenge for the development of safety-critical automotive systems is to ensure consistency of the safety relevant artifacts (e.g., safety goals, concepts, requirements and mechanisms) over the development cycle. This is especially challenging due to the large number of skills, tools, teams and institutions involved in the development. This work presents an approach to bridge tool gaps between an existing model-driven system and safety engineering framework and software engineering tools, based on domain standard AUTOSAR. The implemented tool extension transfers artifacts from system development tools to software development tools, thereby creating traceable links across tool boundaries, and relying on standardized AUTOSAR exchange files. The main benefits of this enhancement are: improved consistency and traceability from the initial design at the system level down to the single software components, as well as a reduction of cumbersome and error-prone manual rework along the system development path. Further improvements of the approach include the introduction of AUTOSAR without relying on full AUTOSAR tooling support, and progress in terms of reproducibility and traceability of safety-critical arguments for the software development.

Acknowledgments

The authors would like to acknowledge the financial support of the "COMET K2 - Competence Centers for Excellent Technologies Programme" of the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT), the Austrian Federal Ministry of Economy, Family and Youth (BMWFJ), the Austrian Research Promotion Agency (FFG), the Province of Styria, and the Styrian Business Promotion Agency (SFG).

Furthermore, we would like to express our thanks to our supporting project partners, AVL List GmbH, Virtual Vehicle Research Center, and Graz University of Technology.

References

[AUT09] AUTOSAR development cooperation. AUTOSAR AUTomotive Open System ARchitecture, 2009.

[BFH+08] Manfred Broy, Martin Feilkas, Markus Herrmannsdoerfer, Stefano Merenda, and Daniel Ratiu. Seamless Model-based Development: from Isolated Tool to Integrated Model Engineering Environments. *IEEE Magazin*, 2008.

[Bol09] Richard Boldt. Modeling AUTOSAR systems with a UML/SysML profile. Technical report, IBM Software Group, July 2009.

[CJL+08] DeJiu Chen, Rolf Johansson, Henrik Loenn, Yiannis Papadopoulos, Anders Sandberg, Fredrik Toerner, and Martin Toerngren. Modelling Support for Design of Safety-Critical Automotive Embedded Systems. In *SAFECOMP 2008*, pages 72 – 85, 2008.

[Eis11] Ulrich Eisemann. Modellbasierte Entwicklung in einer AUTOSAR-Werkzeugkette. www.elektroniknet.de/automotive/sonstige/artikel/74849/, Jannuary 2011.

[GHN10] Holger Giese, Stephan Hildebrandt, and Stefan Neumann. Model Synchronization at Work: Keeping SysML and AUTOSAR Models Consistent. *LNCS 5765*, pages pp. 555 –579, 2010.

[HMM11] Joerg Holtmann, Jan Meyer, and Matthias Meyer. A Seamless Model-Based Development Process for Automotive Systems, 2011.

[ISO11] ISO - International Organization for Standardization. ISO 26262 Road vehicles Functional Safety Part 1-10, 2011.

[Mad12] Roland Mader. *Computer-Aided Model-Based Safety Engineering of Automotive Systems*. PhD thesis, Graz University of Technology, 2012.

[MAK14] Georg Macher, Eric Armengaud, and Christian Kreiner. Bridging Automotive Systems, Safety and Software Engineering by a Seamless Tool Chain. In *7th European Congress Embedded Real Time Software and Systems Proceedings*, pages 256 –271, 2014.

[PB06] Mike Pagel and Mark Broerkens. Definition and Generation of Data Exchange Formats in AUTOSAR, process independent model. *lNCS 4066*, pages pp. 52–65, 2006.

[RW12] Ajitha Rajan and Thomas Wahl. *CESAR Project Book*. Springer Verlag, 2012.

[SS12] Guido Sandmann and Michael Seibt. AUTOSAR-Compliant Development Workflows: From Architecture to Implementation - Tool Interoperability for Round-Trip Engineering and Verification & Validation. *SAE World Congress & Exhibition 2012*, (SAE 2012-01-0962), 2012.

[Vog14] Stefan Voget. SAFE RTP: An open source reference tool platform for the safety modeling and analysis. In *Embedded Real Time Software and Systems Conference Proceedings*, 2014.

[ZS07] Thomas Zurawka and Joerg Schaeuffele. Method for checking the safety and reliability of a software-based electronic system, January 2007.

Comparison of Architectural Design Decisions for Resource-Constrained Self-Driving Cars – A Multiple Case-Study

Christian Berger[1] and Michael Dukaczewski[2]

[1]Department of Computer Science and Engineering
Chalmers | University of Gothenburg
Sweden
christian.berger@chalmers.se

[2]Software Engineering and Automotive Informatics
Technische Universität Braunschweig
Germany
m.dukaczewski@tu-bs.de

Abstract: *Context:* Self-Driving cars are getting more and more attention with public demonstration from all important automotive OEMs but also from companies, which do not have a long history in the automotive industry. Fostered by large international competitions in the last decade, several automotive OEMs have already announced to bring this technology to the market around 2020.
Objective: International competitions like the 2007 DARPA Urban Challenge did not focus on efficient usage of resources to realize the self-driving vehicular functionality. Since the automotive industry is very cost-sensitive, realizing reliable and robust self-driving functionality is challenging when expensive and sophisticated sensors mounted very visibly on the vehicle's roof for example cannot be used. Therefore, the goal for this study is to investigate how architectural design decisions of recent self-driving vehicular technology consider resource-efficiency.
Method: In a multiple case study, the architectural design decisions derived for resource-constrained self-driving miniature cars for the international competition CaroloCup are compared with architectural designs from recent real-scale self-driving cars.
Results: Scaling down available resources for realizing self-driving vehicular technology puts additional constraints on the architectural design; especially reusability of software components in platform-independent algorithmic concepts are prevailing.
Conclusion: Software frameworks like the robotic operating system (ROS) enable fast prototypical solutions; however, architectural support for resource-constrained devices is limited. Here, architectural design drivers as realized in AUTOSAR are more suitable.

1 Introduction

Competitions like the 2007 DARPA Urban Challenge [RBL+08], the Grand Cooperative Driving Challenge in 2011 and its announced successor in 2016 [AEE+11, Hig11], as well as successful large-scale test runs on public roads [Thr10] led to the announcement from all important automotive OEMs that self-driving vehicular technology is said to be available for customers around 2020 [Hir13]. Since the technology has proven its feasibility on urban roads like in Braunschweig, Germany as reported by [NHO+11] or in

the VisLab Intercontinental Challenge (VIAC) [BBFZ11], the next technological challenges are robustness, reliability, and foremost resource-efficiency.

1.1 Motivation and Problem Domain

Today's prototypical cars use either a broad variety of different sensors or depend on expensive and sophisticated sensors that can perceive a huge data volume from the surroundings so that algorithmic solutions have much information available to derive reasonable and reliable driving decisions. However, such large, clearly visible, and predominantly expensive sensors are rather unrealistic being part of a self-driving car for the mass-market.

Therefore, today's technological solutions need to become more resource-efficient in terms of realizing (at least) the same robustness and reliability as the existing self-driving vehicles with their broad and expensive selection of sensors by using less, cheaper, and simpler sensors [NP14]. Thus, limitations for the available hardware resources need to be carefully addressed also in the software architecture for self-driving cars.

1.2 Research Goal and Research Questions

Therefore, the research goal for this study is to identify and elaborate on architectural design drivers for self-driving vehicular technology that is subject to resource constraints. Thus, the following research questions were derived:

RQ-1: *Which design drivers affect the software architecture of self-driving cars when resource constraints need to be obeyed?*

RQ-2: *Which hardware- and software-related architectural design decisions can be found in successful self-driving cars?*

RQ-3: *Which experiences can be drawn from a competition of self-driving miniature cars regarding resource constraints and their impact on the software architecture for self-driving cars?*

1.3 Contributions of the Article

The contributions of this work include (a) an elaboration of the impact of design decisions originating from resource-constrained environments on the software architecture of self-driving vehicular technology by outlining results from a systematic literature review, (b) a selected overview of architectural design decisions for the hardware and software architecture of recent successful self-driving cars, and (c) transferring the conclusions drawn from the software architecture of self-driving miniature cars to real-scale cars regarding the software design for resource-constrained environments.

1.4 Structure of the Article

The rest of the article is structured as described in the following: In Sec. 2, an overview of related work is presented with a specific focus on resource-constrained self-driving vehicular technology. Sec. 3 describes the results from a multiple case study that we conducted using four self-driving cars projects including two projects for the development of self-driving miniature cars under resource-constraints. Sec. 4 discusses the outcome of the multiple case study, before the article concludes in Sec. 5.

2 Related Work

To address *RQ-1*, we conducted a systematic literature review (SLR) according to the guidelines provided by Kitchenham and Charters [KC07]. The results are described in the following.

2.1 Data Sources and Search Strategy

We used the following digital libraries as data sources for the SLR: ACM Digital Library (ACM DL), IEEE Xplore Digital Library (IEEE Xplore), and SpringerLink. Additionally, we used Google Scholar to complement the aforementioned data bases to also cover data sources, which are not represented in the aforementioned libraries. Within these databases, the following search strategy was applied:

Publication date: "2004 or newer" to ensure that related work might be influenced by the outcome for the 2004 DARPA Grand Challenge.

Search phrase A: "resource-constrained" AND "self-driving" and "car"

Search phrase B: "resource-constrained" AND "self-driving" AND "vehicular" AND "technology"

2.2 Study Selection

After the results were retrieved from the databases mentioned before, the following *exclusion criteria* were applied: (a) The publication was not inside the desired time frame (E_1), (b) the work was a duplicate (E_2), or (c) the title and abstract were referring to an unrelated domain (E_3). The remaining publications that are related to this article according to title and abstract are listed in the last column in Tab. 1.

Search phrase	Library	#	E_1	E_2	E_3	$\surd$	References
A	ACM DL	1	-	-	1	-	
A	IEEE Xplore	25	-	1	19	5	[JKK+14] [KKLR13] [KBRJ12] [CQB+12] [UW08]
A	SpringerLink	3	-	-	2	1	[FGRS14]
A	Google Scholar	20	1	5[1,2,3,4]	11	3	[Kre09] [WLM13] [ME10]
B	ACM DL	0	-	-	-	-	
B	IEEE Xplore	0	-	-	-	-	
B	SpringerLink	0	-	-	-	-	
B	Google Scholar	4	-	1[5]	1	2	[GLPL14] [Chu13]
	Total	53	1	7	34	11	

Table 1: Results from the systematic literature review.

2.3 Study Quality Assessment

Kitchenham and Charters state different risks for bias that might occur when conducting an SLR. To address publication bias, which refers to the risk that only papers are considered that report about positive results, we complement the well-known libraries by including Google Scholar as an additional data source.

The risk of performance bias refers to the dependency of the results on the researcher who is carrying out the SLR. Since the amount of results that we retrieved from the databases was surprisingly small, no specific means were applied to address this risk. Instead, all papers that are relevant to this study are summarized in the following.

2.4 Data Extraction and Synthesis

The remaining eleven papers are summarized and their relevance to this study is discussed.

Jo et al. [JKK+14] describe in their work a distributed system architecture for self-driving cars that relies on AUTOSAR and FlexRay. The motivation for the former decision were "reusability, reliability, transferability, and maintainability of the software.", while for the latter, predictability in such a safety-critical system is required. However, the authors also conclude that the existing AUTOSAR architecture contains many constraints that limit the flexibility, which is required to successfully realize a self-driving car. Thus, they omitted "many standard AUTOSAR components related to product reliability."

Kim et al. [KKLR13] proposed a thread-based parallelization to realize the motion planning in self-driving cars to evaluate potential paths that can be considered as drivable. However, they do no outline design constraints for a resource-efficient realization of a self-driving car.

In [KBRJ12], Kim et al. present the middleware SAFER, which they applied to the winning self-driving car "Boss"; Urmson and Whittaker [UW08] give a brief overview of this vehicle. While the middleware aims for fault-tolerance, no design considerations are presented and discussed regarding resource-constraints.

Chong et al. [CQB+12] present briefly their self-driving vehicle. Due to the very short article, no design considerations are presented or discussed.

In the doctoral thesis from Krenn [Kre09], the system architecture for an adaptive system for intelligent systems is discussed. The main focus is not on resource-constrained software architectures for self-driving cars.

Wang et al. [WLM13] describe an architecture for the ontology-based integration of components in the context of the " Internet of the Things" (IoT). In the main focus of the work is the handling of restrictions in terms of heavy communication layers for resource-constrained devices like RFID tags. However, specific considerations for the software architecture of self-driving cars are not presented or discussed.

[1] One article is a duplicate to the one found with the second search query in Google Scholar; article was not accessible.

[2] Two articles are duplicates to the ones found with the first search query in SpringerLink.

[3] One article is a duplicate to the one found with the first search query in ACM DL.

[4] Includes a duplicate to [Chu13].

[5] The article is the duplicate to the one found with the first search query; article was not accessible.

Fortino et al. [FGRS14] discuss requirements for smart environments as needed by the IoT. The authors have identified the following requirements, which they evaluated for different IoT-capable middlewares like the robot operating system (ROS):

1. "Abstraction over heterogeneous input and output hardware devices,"
2. "Abstraction over hardware and software interfaces,"
3. "Abstraction over data streams (continuous or discrete data or events) and data types,"
4. "Abstraction over physicality (location, context),"
5. "Abstraction over the development process."

Medvidovic and Edwards present a roadmap for "software architecture and mobility" [ME10]. In their work, they also cover some of the frameworks presented by Fortino et al. for mobile devices like smart phone but also for mobile robotic platforms. In their research roadmap, they confirm that the software engineering is specifically challenged by resource-constrained devices. While they suggest the use of generative techniques to create systems that are "correct by construction", some restrictions in terms deployment costs need to be kept in mind when dealing with such devices. Their observation is in line with the one from Jo et al. [JKK+14] as described before. These design considerations can also be found in AUTOSAR. However, they do not discuss specific requirements that originate from resource-constrained devices.

Gerla et al. [GLPL14] report the "vehicular cloud" that combines independently acting autonomous vehicles with a cloud-based infrastructure. While they are envisioning various use cases in combination with vehicle-to-X communication, specific design aspects from resource-constrained devices are not discussed.

Chua focuses in his thesis [Chu13] mainly on unmanned aerial systems that are used in urban search and rescue environments. While specific aspects are discussed regarding resource-constraints, no reflections are presented how this should be considered in the design of a software architecture.

3 Multiple Case Study

For addressing *RQ-2*, a multiple case study was conducted covering four projects for self-driving cars. We are reporting this study according to Runeson and Höst [RH08].

3.1 Method

The objective for this multiple case study is to evaluate design considerations of the software architecture from different self-driving cars. In this regard, a specific focus is given on how resource-constraints are considered in the architectural design.

These architectural designs for self-driving cars are evaluated according to the criteria that have been identified from the SLR as provided by Fortino et al. [FGRS14]:

1. "Abstraction over heterogeneous input and output hardware devices,"
2. "Abstraction over hardware and software interfaces,"
3. "Abstraction over data streams (continuous or discrete data or events) and data types,"
4. "Abstraction over physicality (location, context),"
5. "Abstraction over the development process."

The following cases are presented and compared in the multiple case study:

1. A ROS-powered platform for autonomous driving evaluated by a company as described by Ainhauser et al. [ABH+13],
2. a ROS-powered platform for autonomous driving evaluated by a research laboratory as outlined by Fernandes et al. [FSP+14],
3. results from a Java-based design of a self-driving miniature car that participated in the international competition for self-driving miniature vehicles CaroloCup[6],
4. results from a C++-based design of a self-driving miniature car that participated also in this competition [BAH13].

3.2 Case Description

The first architecture in our case study is described by Ainhauser et al. [ABH+13], where ROS as a platform for future autonomous driving functions is discussed. The authors report that for autonomous driving two environments are needed: A hardware-oriented environment for sensors and actuators, as well as an environment for functions like connectivity and route planning. For this, the two frameworks AUTOSAR and ROS are used: AUTOSAR as the standard for the embedded hardware and ROS for high-level functions like communication, to build the environment model, and planning the route. AUTOSAR components are integrated by ROS stubs. They suppose that ROS is currently the most suitable existing middleware for the needs of autonomous driving. They propose the architecture in Fig. 1 on the left hand side.

The second case study is taken from a very recent publication in the Journal of Systems Architecture published by Fernandes et al. [FSP+14], where two outdoor intelligent vehicles platforms named CaRINA I and CaRINA II were described. Similarly as in the first case, an architecture based on ROS is proposed. Since the authors have developed the architecture for the first vehicle and could transfer it to the second one, they suggest it as a reference architecture. Fig. 1 shows the architecture on the right hand side.

The third case study is taken from team "piCar" of the Technische Universitaet Braunschweig, that won the Junior Edition at CaroloCup 2014 with their car. The vehicle uses the Eleanor Framework, which was developed at the Institute of Software Engineering and

[6] www.carolocup.de

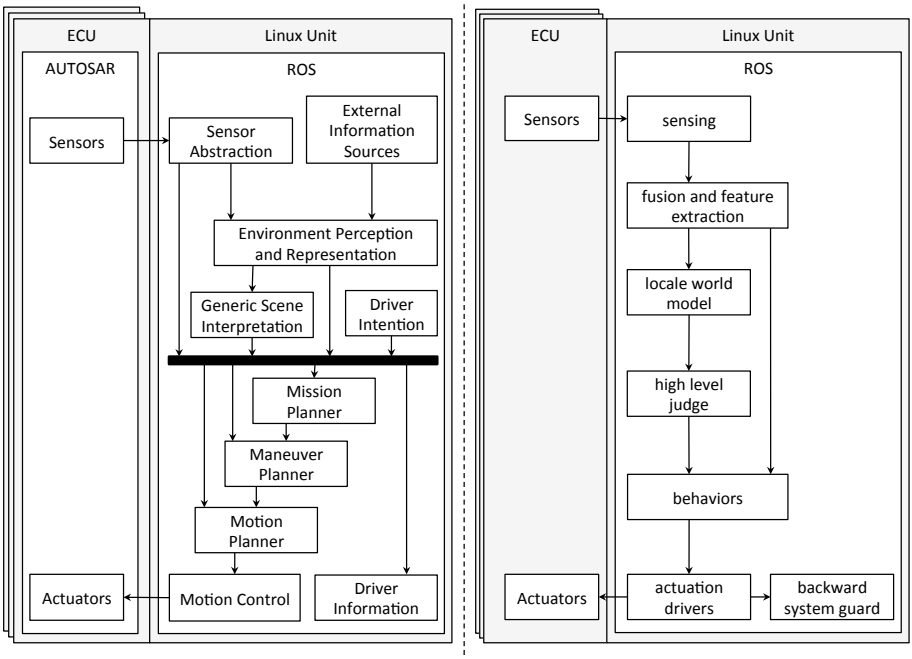


Figure 1: Comparison of the architecture that uses ROS as software platform for the self-driving car: On the left hand side, the industrial study is shown in comparison with the research laboratory realization as depicted on the right hand side.

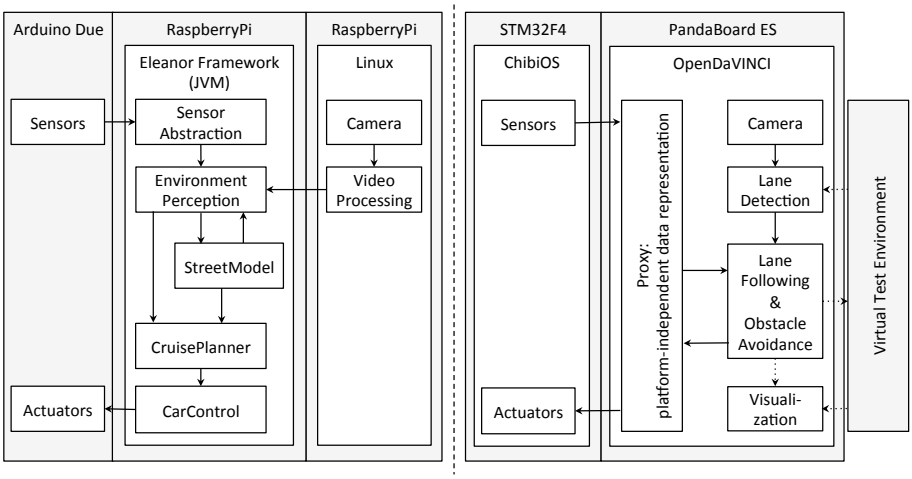


Figure 2: Comparing the architectural design of two resource-constrained self-driving miniature cars: On the left hand side, the architecture from team from Technische Universitaet Braunschweig is shown and on the right hand side, the design of the team from Chalmers | University of Gothenburg is shown. Both teams participated in the international competition CaroloCup.

Automotive Informatics. The architecture of the vehicle is shown in Figure 2 on the left hand side.

The vehicle uses an Arduino microcontroller to control sensors and actuators as well as two RaspberryPis. One of them acts as a central processing unit for route planning and the other is used to connect the camera and for video processing. The Eleanor Framework is a Java-based framework designed specifically to meet the needs of limited hardware, however, yet to offer hardware independence and extensibility.

The framework provides a service and extension registry inspired by OSGi[7], so all functions are implemented as plugins that can have dependencies among each other. It also provides useful basic functionality such as the abstraction of sensors and actuators as well as a thread manager to schedule the plugins. In contrast to the other cases, this system uses inter-process-communication (IPC) between the plugins.

The last case study is also taken from a team that participated twice in the CaroloCup and that won the Junior Edition at CaroloCup 2013. The architecture used in cars of Chalmers | University of Gothenburg is shown in Fig. 2 on the right hand side. The team used the simulation and development middleware OpenDaVINCI in combination with the real-time operating system ChibiOS/RT[8] as hardware abstraction layer (HAL) on the embedded system, which is realized by using an STM32F4 board. In this regard, this case shows similarities with the first case, where AUTOSAR is used as HAL. On the high-level data processing where the self-driving functionality is realized with C++, platform-independent data structures are used to enable reusability in the virtual test environment [BCH+13].

4 Analysis and Discussion

To address *RQ-3* we compared the four case studies, to extract differences and commonalities. We have found the following results:

All four cases have a strict separation between low-level functions to control the hardware and a high-level layer for environment perception and behavioral planning. The separation is achieved by a distribution of functions to different hardware. Hardware-related functions were deployed to microcontrollers, where the first and fourth case use a hardware abstraction layer (AUTOSAR and ChibiOS/RT) and the third case implements the functions directly on the Arduino; no information was available for the second case.

On the high-level, all cases uses an abstraction layer over the hardware devices, which is in line with the criteria identified by Fortino et al. [FGRS14]: Case 1 uses ROS stubs for AUTOSAR components, case 2 uses modules that encapsulate the low-level protocols for each hardware component, case 3 uses a plugins that communicates with the hardware part and provides high-level services for other plugins, and case 4 uses a proxy that transparently maps platform-independent data structures to platform-specific representations that depend on the concrete hardware environment.

[7] www.osgi.org
[8] www.chibios.org

Due to the abstraction of the concrete hardware, (a) high-level algorithms such as route planning can be developed independently of the hardware, and (b) the hardware may be replaced without the need to change to high-level software. This is an important aspect since the configuration of commercial-of-the-shelf components, which are used as software/hardware-interface in resource-constrained environments like the CaroloCup competition, might change. Thus, even the low-level software that handles the data from sensors and to actuators can be reused, when an appropriate HAL like AUTOSAR or ChibiOS/RT is used.

The abstraction over hardware and software interfaces in the first two cases is realized with the infrastructure and facilities provided by ROS, where the modules are loosely coupled through topic publishing/subscription realized via TCP or UDP communication. The third case uses a service registry, where plugins can register their interfaces to exchange data. Case 4 uses a UDP-multicast session for a loosely coupled communication. As all cases uses a concept of a platform-independent data representation, there is an abstraction over data streams and data types as suggested by Fortino et al. [FGRS14].

As the first two cases use ROS that supports a distributed architecture, there is an abstraction over physicality; a similar concept is realized for case 4. Case 3 has a high level of abstraction over the context by using a Java virtual machine but there is no direct support for IPC.

As the development environment usually does not match the hardware on the vehicle, the compilation of the code has to be done on the limited hardware of the vehicle, which can be very time consuming, or by using a cross compiling tool chain. So, the development process of the low-level part depends on the hardware in-use. It can be simplified by using a HAL like AUTOSAR in case 1 or ChibiOS/RT in case 4. The development process of the high-level part in case 1 and 2 is given by the tool chain of ROS and the one of case 4 is set by OpenDaVINCI. As case 3 uses Java, any Java-based development process can be used. Here the usage of Java has the advantage that the code can be compiled on fast development machines and deployed on any hardware supported by Java.

We found that all four cases are very similar. Each uses a central processing unit for the high-level part with the responsibility for communication, creating an environment model, planning the driving route, and decide which maneuver to drive. Sensors and actuators are always connected through an ECU, which is abstracted by a low-level HAL in two cases.

Furthermore, the cases show that the low-level data processing can be realized with resource- and energy-efficient and hardware environments to handle the data from sensors and to actuators. A prevailing design consideration is the HAL which not only allows reusability of software components on that layer; moreover, it also enables code-generation of the software components due to standardized APIs.

For the higher levels, more processing power is reported to be required – from the results of the SLR but also to handle mass-data from cameras or to interpret and derive driving decisions. On this layer, reusing the self-driving functionality realized by the software components is of interest – for instance by embedding them in virtual test environments for example. A specific design aspect that originates from the resource-constrained low-level hardware environment is the mapping of platform-independent data structures to the

	Case 1	Case 2	Case 3	Case 4
separation between low/high-level functions	+	+	+	+
hardware abstraction layer	+	?	-	+
high-level abstraction layer	+	+	+	+
loosely coupled components	+	+	+	+
platform-independent data representation	(+)	(+)	+	+
platform-independent development process	-	-	+	-
energy-efficient low-level data processing	+	+	+	+

Table 2: Comparison of architectural design drivers.

concrete platform-specific representation. Our findings are summarized in Tab. 2.

In the following, we are reporting about threats to validity according to Runeson and Höst [RH08]. Considering construct validity, the setup of the study by using a systematic literature review first to investigate the field followed by a multiple case study is appropriate to investigate design criteria for resource-constrained self-driving cars. The use of the SLR also reduces the risk of reliability bias since the evaluation criteria originate from the findings of the SLR.

Regarding external validity, we analyzed four different cases by using criteria that are reported by literature. Due to the recent publication dates of the investigated first two cases as well as the involvement in the design and realization in the last two cases, the findings can be generalized to design criteria for self-driving cars with a focus on resource-constraints.

Additionally, the specific design of a multiple case study reduces the bias regarding internal validity because four different cases are investigated that were driven by four independently working research groups.

5 Conclusion and Outlook

Self-driving cars are said to be part of our future mobility. In the last decade, the research in this field has facilitated the development and state-of-the-art so that this technology is evaluated nowadays in large-scale on public roads. However, to bring the technology to the mass-market, the hardware costs especially for sophisticated sensors that are mounted on the vehicle's roof need to be scaled down.

In this study, results from a systematic literature review are presented that evaluate how resource-constraints are considered in the design of the software architecture of self-driving cars. This literature study unveiled that common guidelines could not be identified on how to handle such constraints accordingly. Therefore, a multiple case study was conducted to evaluate four projects for self-driving cars regarding design considerations for resource-constraints.

Common design criteria include abstractions from concrete hardware environments. While such a design does not only enable the independence from a concrete hardware, it also

allows the transparent software reuse in virtual test environments for example. As future work, specific design criteria for platform-independent data models in combination with automated scenario generation for such test environments are of interest. However to operate autonomous vehicles on public roads, the legal framework in different countries needs to be evolved as well.

References

[ABH+13] Christoph Ainhauser, Lukas Bulwahn, Andreas Hildisch, Stefan Holder, Olexiy Lazarevych, Daniel Mohr, Tilmann Ochs, Michael Rudorfer, Oliver Scheickl, Tillmann Schumm, and Felix Sedlmeier. Autonomous Driving Needs ROS (ROS as Platform for Autonomous Driving Functions), May 2013.

[AEE+11] Bruno Augusto, Alireza Ebadighajari, Cristofer Englund, Usman Hakeem, Naga V Irukulapati, Josef Nilsson, Ali Raza, and Reza Sadeghitabar. Technical Aspects on Team Chalmers Solution to Cooperative Driving. Technical report, Chalmers University of Technology, Göteborg, 2011.

[BAH13] Christian Berger, Md. Abdullah Al Mamun, and Jörgen Hansson. COTS-Architecture with a Real-Time OS for a Self-Driving Miniature Vehicle. In Elad M. Schiller and Henrik Lönn, editors, *Proceedings of the 2nd Workshop on Architecting Safety in Collaborative Mobile Systems (ASCoMS)*, pages 1–12, Toulouse, France, September 2013.

[BBFZ11] Alberto Broggi, Michele Buzzoni, Mirko Felisa, and Paolo Zani. Stereo obstacle detection in challenging environments: the VIAC experience. In *IROS*, pages 1599–1604, 2011.

[BCH+13] Christian Berger, Michel Chaudron, Rogardt Heldal, Olaf Landsiedel, and Elad M. Schiller. Model-based, Composable Simulation for the Development of Autonomous Miniature Vehicles. In *Proceedings of the SCS/IEEE Symposium on Theory of Modeling and Simulation*, San Diego, CA, USA, April 2013.

[Chu13] Noon Heng Chua. *Integration of Multiple Unmanned Systems in an Urban Search and Rescue Environment*. Master thesis, Naval Postgraduate School, March 2013.

[CQB+12] Z J Chong, B Qin, T Bandyopadhyay, T Wongpiromsarn, B Rebsamen, P Dai, S Kim, M H Ang, D Hsu, D Rus, and E Frazzoli. Autonomy for Mobility on Demand. In *Proceedings of the IEEE International Conference on Intelligent Robotics and Systems*, pages 4235–4236, Vilamoura, Algarve, Portugal, October 2012.

[FGRS14] Giancarlo Fortino, Antonio Guerrieri, Wilma Russo, and Claudio Savaglio. *Internet of Things Based on Smart Objects*. Springer International Publishing, Cham, 2014.

[FSP+14] Leandro C. Fernandes, Jefferson R. Souza, Gustavo Pessin, Patrick Y. Shinzato, Daniel Sales, Caio Mendes, Marcos Prado, Rafael Klaser, André Chaves Magalhães, Alberto Hata, Daniel Pigatto, Kalinka Castelo Branco, Valdir Grassi, Fernando S. Osorio, and Denis F. Wolf. CaRINA Intelligent Robotic Car: Architectural design and applications. *Journal of Systems Architecture*, 60(4):372–392, April 2014.

[GLPL14] Mario Gerla, Eun-Kyu Lee, Giovanni Pau, and Uichin Lee. Internet of Vehicles: From Intelligent Grid to Autonomous Cars and Vehicular Clouds. In *Proceedings of the IEEE World Forum on Internet of the Things*, pages 241–246, March 2014.

[Hig11] High Tech Automotive Systems. Grand Cooperative Driving Challenge, January 2011.

[Hir13] Jerry Hirsch. Self-driving cars inch closer to mainstream availability, October 2013.

[JKK+14] Kichun Jo, Junsoo Kim, Dongchul Kim, Chulhoon Jang, and Myoungho Sunwoo. Development of Autonomous Car â Part I: Distributed System Architecture and Development Process. *IEEE Transactions on Industrial Electronics*, 0046:1–10, April 2014.

[KBRJ12] Junsung Kim, Gaurav Bhatia, Ragunathan Rajkumar, and Markus Jochim. SAFER: System-level Architecture for Failure Evasion in Real-time Applications. In *Proceedings of the 33rd IEEE 33rd Real-Time Systems Symposium*, pages 227–236. Ieee, December 2012.

[KC07] Barbara Kitchenham and Stuart Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical report, Keele University and Durham University Joint Report, Durham, UK, July 2007.

[KKLR13] Junsung Kim, Hyoseung Kim, Karthik Lakshmanan, and Ragunathan Rajkumar. Parallel Scheduling for Cyber-Physical Systems: Analysis and Case Study on a Self-Driving Car. In *Proceedings of the 4th International Conference on Cyber-Physical Systems*, pages 31–40, Philadelphia, PA, USA, April 2013.

[Kre09] Willibald Karl Krenn. *Self Reasoning In Resource-Contrained Autonomous Systems*. Phd thesis, Graz University of Technology, 2009.

[ME10] Nenad Medvidovic and George Edwards. Software architecture and mobility: A roadmap. *Journal of Systems and Software*, 83(6):885–898, June 2010.

[NHO+11] Tobias Nothdurft, Peter Hecker, Sebastian Ohl, Falko Saust, Markus Maurer, Andreas Reschka, and Jürgen Rüdiger Böhmer. Stadtpilot: First Fully Autonomous Test Drives in Urban Traffic. In *Proceedings of the International IEEE Conference on Intelligent Transportation Systems*, pages 919–924, Washington, DC, USA, October 2011.

[NP14] Paul Newman and Ingmar Posner. Robotics Science For Smarter Cars, May 2014.

[RBL+08] Fred W. Rauskolb, Kai Berger, Christian Lipski, Marcus Magnor, Karsten Cornelsen, Jan Effertz, Thomas Form, Fabian Graefe, Sebastian Ohl, Walter Schumacher, Jörn-Marten Wille, Peter Hecker, Tobias Nothdurft, Michael Doering, Kai Homeier, Johannes Morgenroth, Lars Wolf, Christian Basarke, Christian Berger, Tim Gülke, Felix Klose, and Bernhard Rumpe. Caroline: An Autonomously Driving Vehicle for Urban Environments. *Journal of Field Robotics*, 25(9):674–724, September 2008.

[RH08] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, 14(2):131–164, December 2008.

[Thr10] Sebastian Thrun. What we're driving at, April 2010.

[UW08] Chris Urmson and William Whittaker. Self-Driving Cars and the Urban Challenge. *IEEE Intelligent Transportation Systems*, 23(2):66–68, March 2008.

[WLM13] Wei Wang, Kevin Lee, and David Murray. Building a Generic Architecture for the Internet of Things. In *Proceedings of the 8th IEEE International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, pages 333–338, April 2013.

Considering Architectural Properties in Real-time Play-out

Jörg Holtmann
Software Engineering, Project Group Mechatronic Systems Design,
Fraunhofer Institute for Production Technology IPT
Zukunftsmeile 1, 33102 Paderborn, Germany
joerg.holtmann@ipt.fraunhofer.de

Dimitar Shipchanov
Department of Computer Science, University of Paderborn
Warburger Str. 100, 33098 Paderborn, Germany
mitko@mail.upb.de

Abstract: Real-time embedded systems (RTES), as in the automotive domain, provide their functionality by executing software operations on hardware with restricted resources and by communicating via buses. The properties of the underlying architecture, i.e., execution times of software operations and bus latencies, cause delays during the provision of the functionality. At the same time, RTES have to fulfill strict real-time requirements. The fulfillment of such real-time requirements under consideration of delays induced by architectural properties should be taken into account already during requirements engineering (RE) to avoid costly iterations in subsequent development phases. In previous work, we developed a formal RE approach based on a recent Live Sequence Chart (LSC) variant, so-called Modal Sequence Diagrams (MSDs). This scenario-based RE approach allows to validate the requirements by means of simulation, i.e., the play-out algorithm originally conceived for LSCs. Our MSD play-out approach considers assumptions on the environment as well as real-time requirements and is applicable to hierarchical component architectures, which makes it well suited for automotive systems. However, delays induced by architectural properties are not considered. In order to consider this important aspect, we introduce in this paper an approach enabling the annotation of software operation execution times and connector latencies to hierarchical component architectures by means of the MARTE profile. These assumptions about the architectural properties can be verified against the real-time requirements specified in the MSDs by means of simulation. We illustrate the approach by means of an example of an automotive RTES.

1 Introduction

The growing functionality of real-time embedded systems (RTES) in modern vehicles has led to thousands of software operations distributed over 100 electronic control units (ECUs) that communicate via multiple bus systems [VPBK10]. The ECUs executing the software operations have restricted resources leading to best and worst case execution times, and the buses delivering messages for the purpose of communication have latencies. These *architectural properties* cause *delays* during the provision of the RTES functionality. At the same time, RTES have to fulfill strict real-time requirements. The fulfillment of

such real-time requirements under consideration of delays induced by architectural properties should be taken into account as early as possible in the development process, particularly in the requirements engineering (RE) phase. This is due to the fact that the detection and fixing of defects in the system under development (SUD) in subsequent development phases cause costly development iterations (e.g., [Poh10]).

In previous work, we developed a formal RE approach based on a recent Live Sequence Chart (LSC) [DH01] variant compliant to the Unified Modeling Language 2 (UML2) [Obj11a], so-called *Modal Sequence Diagrams (MSDs)* [HM08]. This scenario-based RE approach allows to validate the requirements by means of simulation, i.e., the play-out algorithm originally conceived for LSCs [HM03]. Play-out enables to execute MSD/LSC specifications, which allows the requirements engineers to understand the behavior emerging from the interplay of the scenarios. Our MSD play-out approach implemented in the SCENARIOTOOLS[1] tool suite considers assumptions on the environment [BGPLM13] as well as real-time requirements [BGH+14] and is applicable to hierarchical component architectures [HM13], which makes it well suited for automotive systems.

However, delays induced by architectural properties as mentioned above are not considered. In order to consider this important aspect, we introduce in this paper an extended real-time play-out approach. Our contribution is twofold. First, we enable the annotation of architectural properties to hierarchical component architectures that serve as structural basis for the MSDs [HM13]. To do so, we annotate execution times to software operations and latencies to communication connectors by means of the UML2 profile Modeling and Analysis of Real-Time and Embedded Systems (MARTE) [Obj11b]. Second, we extend our MSD real-time play-out approach [BGH+14] to investigate delays induced by these architectural properties in the simulation of scenarios. This enables to simulatively verify the assumptions about the architectural properties against the real-time requirements specified in the MSDs.

We illustrate the approach by means of an automotive ECU that controls body and comfort functions of a car, a so-called Body Control Module (BCM). A BCM centrally controls distributed vehicle functions like central locking, turn signals, and brake light and has to interact with several other ECUs to accomplish this task. We focus on one scenario related to the central locking functionality. We assume that the BCM has locked all doors after the car has been started and a certain velocity threshold has initially been reached. In the hazard situation that a crash is detected, the BCM has to unlock all doors within a time limit after a crash was detected, so that all passengers can escape or can be rescued from outside.

This paper is structured as follows. Sect. 2 introduces the fundamentals of MSDs, their structural basis, their real-time requirements, and our real-time play-out approach. Sect. 3 presents the extended modeling and simulation approach. Sect. 4 covers related work. Finally, Sect. 5 summarizes this paper and provides an outlook on future work.

[1] http://scenariotools.org/

2 Foundations

In this section, we introduce component-based MSDs (Sect. 2.1), their basic semantics (Sect. 2.2), and their extensions to specify real-time requirements (Sect. 2.3). In Sect. 2.4, we introduce real-time play-out for the simulative validation of timed MSDs. For illustration purposes, we use the running example of the BCM, which has to unlock all doors in a hazard situation. The particular models are shown in Fig. 1, which we explain step by step in the following subsections. In general, the figure describes the requirement that when a crash is detected by the CrashSensor, the BCM has to send an `unlock` command to the DoorLock and receive a corresponding acknowledgement message within 70 time units.

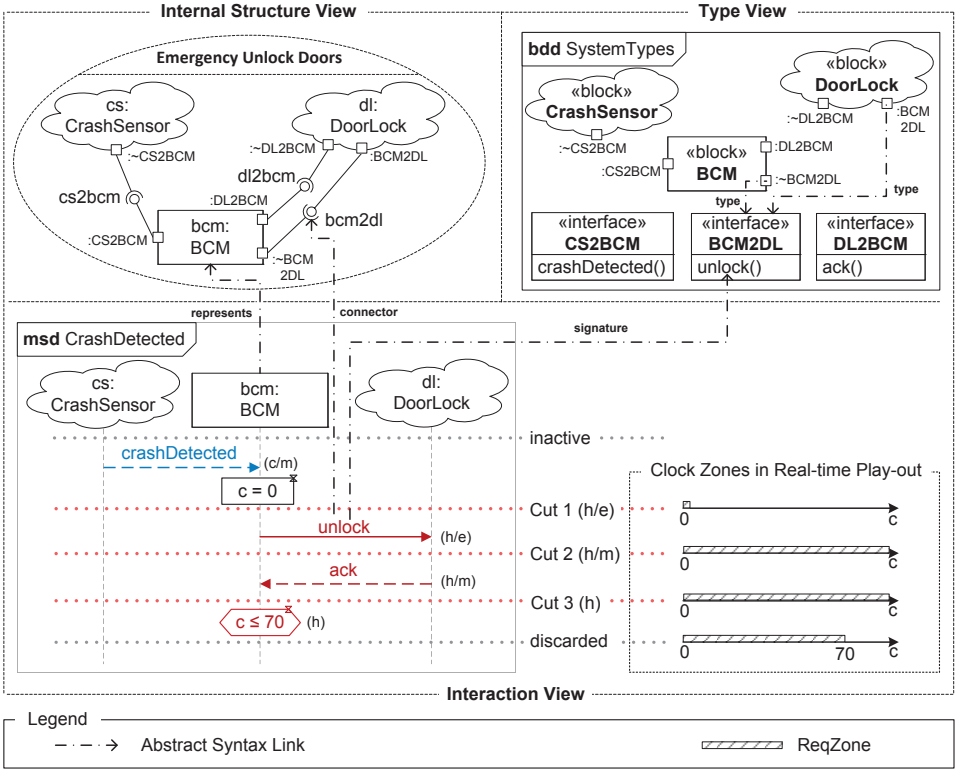


Figure 1: Component-based MSD with clock zones in real-time play-out

2.1 Hierarchical Component Architectures as Structural Basis for MSDs

In [HM13], we introduced hierarchical component architectures as structural basis for MSDs. We use a subset of the structural models of the Systems Modeling Language (SysML) [Obj10] and the UML2 to specify components, ports, and connectors as structural basis of the SUD.

In the Type View, we define component types and interfaces using a SysML Block Definition Diagram System Types, as shown in the upper right corner of Fig. 1. We distinguish between *system (controllable) components* and *environment (uncontrollable) components*. The set of all system components represents (subsystems of) the SUD and the set of all environment components stands for the environment. In Fig. 1, the cloud-shaped elements CrashSensor and DoorLock represent environment component types and BCM represents a system component type. The particular ports of the component types have provided and required interfaces to specify directed message exchange. For example, the port :BCM2DL of DoorLock provides the interface BCM2DL that is required by the port :~BCM2DL of BCM (the ~ specifies that the direction of the interface typing the port is reversed, therefore it is a required interface) [Obj11a]. Thus, BCM and DoorLock have a compatible required and provided interface for sending and receiving the message `unlock`, respectively.

In the Internal Structure View, we use a UML2 Collaboration Diagram Emergency Unlock Doors to specify the actual top-level component-based architecture, which is shown in the upper left corner of Fig. 1. We specify the directed assembly connector cs2bcm from the component cs:CrashSensor to bcm:BCM, and the two assembly connectors dl2bcm and bcm2dl between the components bcm and dl:DoorLock for a bidirectional communication. The latter one connects the aforementioned ports to allow the sending of the message `unlock` from bcm to dl. Note that we do not consider multiple hierarchy levels of component architectures in this paper as in [HM13] due to space restrictions.

2.2 Basic MSD Semantics

Based on the component architecture described in the last subsection, the MSD CrashDetected can be specified in the Interaction View (lower left part of Fig. 1). An MSD basically consists of *lifelines* and *messages*. A lifeline represents a participant in the component architecture (e.g., the lifeline/component bcm:BCM). Messages, represented by arrows between lifelines, define requirements on the communication between the components. Messages have a *temperature* and an *execution kind*. The temperature is used to specify provisional (*cold*) or mandatory (*hot*) behavior. For a cold message, a different message specified by the MSD may occur at this point in time.Cold messages are visualized with blue color and are annotated with (c), e.g., `crashDetected`. The semantics of a hot message is that other messages specified by the MSD must not occur at this point in time. Hot messages are shown in red and are annotated with (h), e.g., `unlock` and `ack`.The execution kind determines whether a message is *monitored*, visualized with a dashed arrow and annotated with (m), or *executed*, represented by a solid arrow and annotated with (e). The semantics is that monitored messages can but do not need to occur, while executed messages must occur eventually. In Fig. 1, the messages `crashDetected` and `ack` are monitored and the message `unlock` is executed.

The actual sending or receiving of a message at runtime or during play-out is called an *event*. A message is *unifiable* with an event iff the sending and the receiving lifelines of the message represent the same sending and receiving components of the event, and the event name is the same as the message operation name. Messages that have no preceding

message (i.e., crashDetected) are called *minimal messages*. An MSD proceeds when events occur that are unifiable with its messages and becomes *active* if its minimal message is unifiable with the event currently being executed. A *cut* is the current state of the lifeline locations in an active MSD. A location is a point on a lifeline where a message is sent/received. If the cut is immediately before a message, the message is called *enabled*. If any of the enabled messages in the active MSDs is hot, then the cut is hot. Otherwise, it is cold. If at least one enabled message is executed, the cut is executed, otherwise it is monitored. In Fig. 1, Cut 1 is hot and executed, as indicated by the annotation (h/e). Cut 2 is hot and monitored (h/m), and Cut 3 is hot (h). If the cut reaches the end of the diagram, the active copy of the MSD is discarded.

A violation occurs if the currently executed event is unifiable with a message that is specified in an active MSD but is not enabled. There are two outcomes depending on the temperature of the cut. If the cut is cold, a cold violation occurs. This is a legal trace, and active MSD copy is discarded. If the cut is hot, this is a *safety violation*, which must never happen. An executed cut must eventually progress, otherwise it is considered a *liveness violation*.

2.3 Real-time Requirements in MSDs

In [BGH+14], we introduced modeling constructs for the specification of real-time requirements to MSDs. In Fig. 1, these modeling constructs have an additional hour-glass icon. We base on the concept of *clock variables* [AD94], i.e., real-value variables that increase synchronously and linearly with time. A *clock reset* sets the value of a clock to zero, as shown by the c=0 assignment in the rectangle after message crashDetected in Fig. 1. *Time conditions* are expressions over a clock variable that evaluate to a Boolean value and are represented by hexagons covering one or more lifelines. They have the form $x \bowtie expr$, where x is a clock variable, $expr$ is an expression evaluating to an integer value, and $\bowtie$ can be any operator from $<, \leq, \geq, >$. Time conditions also have a temperature. Enabled cold time conditions are evaluated immediately. If they evaluate to false, a cold violation occurs. Otherwise, the cut progresses. In contrast, a cut can only progress beyond an enabled hot time condition if it evaluates to true. If an event occurs that would let the cut progress beyond a hot time condition that evaluates to false, a safety violation occurs. If it never evaluates to true, a liveness violation occurs. In the MSD shown in Fig. 1 the hot time condition $c \leq 70$ is depicted, which is also called a *maximal delay*. Thus, the real-time requirement of the MSD specifies that bcm has to send the unlock command to dl and receive the corresponding acknowledgment within 70 time units after it received a crashDetected message.

2.4 Real-time Play-out

The play-out algorithm allows the behavior of the system as specified in MSDs to be simulated, thus enabling the validation of the reaction of the system on environment or user inputs. At the beginning of the simulation, the algorithm waits for an environment

event to occur. When it occurs, MSDs are activated whose initial event is unifiable with the environment event. Next, play-out chooses non-deterministically one of the enabled system events and executes it. This operation is called a step. After the event has been executed, the cut progresses and a new list of enabled messages is created. The process is repeated until there are no active MSDs with executed cut left. The system then waits for the next environment event. In case of a hot violation the algorithms terminates.

In [BGH+14], we introduced a real-time play-out algorithm, which is based on *clock zones* [Dil90, BY04]. A clock zone is the solution set of a clock constraint, where the clock is a real-valued variable [BY04]. The clock zone can be interpreted as a symbolic state, which represents an infinite number of clock values that satisfy the conditions applied over one or more clock variables.

Real-time play-out uses a clock zone to keep track of the time progress, implicitly specified by clock resets and time conditions, during its execution. In this paper, we call this clock zone the *requirements clock zone (ReqZone)*. The ReqZone is associated with each MSD cut, as shown in the lower right corner of Fig. 1. The valuations of clock c are shown in the time lines of the corresponding cut. In Cut 1, the MSD is activated and clock c is reset to zero. Once the `unlock` message is sent, time starts to progress and the clock can have arbitrary values, as indicated by the time line corresponding to Cut 2. After message `ack` in Cut 3, there are no further constraints for clock c, so its value can still be arbitrary. Finally, we assume that `ack` occurs within 70 time units and hence the time condition is fulfilled. Therefore, after evaluating the time condition $c \leq 70$ to true, the upper bound of the clock is modified accordingly. If `ack` would occur later than 70 time units, a liveness violation would occur.

3 Verifying Architectural Properties Against Scenario-based Real-time Requirements

Distributed RTES provide their functionality by executing software operations on distributed hardware with restricted resources and by communicating via buses. For example, in a car, software components are deployed on different ECUs, which need time to process requests and communicate via different kinds of bus systems with differing communication latencies. These architectural properties cause delays during the actual provision of the SUD's functionality and have to be verified w.r.t. real-time requirements.

Therefore, we consider two sets of timing constraints in our extended real-time play-out approach: Real-time requirements specified by means of timed MSDs and assumptions about architectural properties affecting the timing behavior annotated to the component architecture. Based on this information, we compare and analyze the two sets of timing constraints and check whether the proposed solution, i.e., the assumptions about the architectural properties, solves the problem, i.e., the real-time requirements. We postulate that the assumptions about the architectural properties stem from experiences with earlier projects and are quite coarse-grained. Therefore, our approach is not designed to replace state-of-the-art timing analysis tools in subsequent development phases. Instead, it aims at

enabling to identify and resolve possible requirements violations or defects already in the RE phase.

First, in Sect. 3.1, we show how the assumptions about architectural properties are specified in the component architecture. In Sect. 3.2, we present how delays stemming from the architectural properties influence the time progress in real-time play-out. Finally, in Sect. 3.3, we show how to compare this time progress in the simulation with the real-time requirements in MSDs and introduce rules for this comparison. Fig. 2 revisits the models of Fig. 1 and sketches our extended approach.

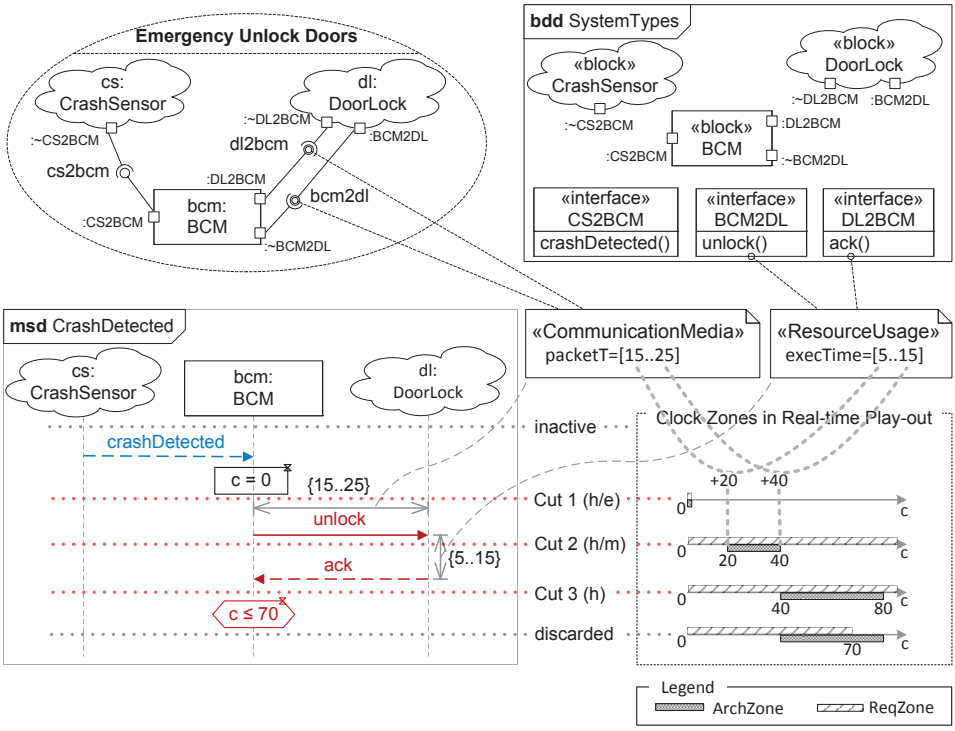


Figure 2: Annotated architectural properties considered in real-time play-out

3.1 Annotating Assumptions about Architectural Properties to Hierarchical Component Architectures

We use a subset of the MARTE profile [Obj11b] to annotate assumptions about architectural properties affecting the SUD's timing behavior to the component architectures that serve as structural basis for MSDs. More specifically, we annotate *communication latencies* to assembly connectors and *software operation execution times* to operations specified in the component interfaces. Both kinds of architectural properties are specified by bounded or half-bounded intervals. This avoids to awkwardly investigate particular con-

crete values in the simulation and enables to consider the whole range between a minimal and maximal value (cf. Sect. 3.3).

In our approach, assembly connectors represent communication media with various physical characteristics. For the analysis of distributed RTES, we are interested in the communication delay between two interacting components caused by the latency of the communication media. For this purpose, we annotate assembly connectors with the communication latency in the architectural model. Fig. 2 shows how the latency is specified using the «CommunicationMedia» stereotype with the attribute packetT from MARTE. It specifies the time needed to transmit a request via this connector. For example, the message `unlock` is sent from bcm:BCM to dl:DoorLock via the connector bcm2dl. The communication latency is between 15 and 25 time units, as specified by the annotation and indicated by the horizontal duration constraint of the message `unlock` in the MSD.

The other cause for delays that we consider are the software operation execution times of RTES. We apply the execTime attribute from the «ResourceUsage» stereotype in MARTE for the annotation of this property. As defined in [Obj11b], it specifies the time a component, whose port implements this interface, will be in use due to usage. Components interact using ports and exchange messages as specified in interfaces. We annotate the software operations defined in an interface and consider their execution time if the interface is implemented as a required one by a component port. For example, the message `unlock` in Fig. 2 is associated with an operation from the BCM2DL interface. The execution time of this operation specifies that it will take between 5 and 15 time units for dl to process the request. This is also indicated by the vertical duration constraint from the receiving of the `unlock` message to the subsequent sending of the `ack` message by dl.

3.2 Considering Architectural Properties in Real-time Play-out

The communication between the components and the execution of software operations of a distributed RTES consume time, expressed using communication latencies and operation execution times. For each message exchanged between two entities there are two possible delay intervals, i.e., the *communication delay interval* and the *operation execution delay interval*. As shown in Fig. 2, there are two duration constraints for the message `unlock` indicating these delays. The communication delay interval is induced by the communication latency and the operation execution delay interval by the operation execution time. The communication delay interval specifies all possible message delays from a message sending event until its receiving event. The operation execution delay interval determines all possible time delays due to processing a message from the receiving event until the sending event of the subsequent message.

During play-out, the delay intervals are considered when an event occurs based on the annotated structural model. To capture and investigate the delay intervals related with the architectural properties we introduce an additional clock zone, called *architectural clock zone (ArchZone)*. Whenever a message with at least one delay interval is sent in real-time play-out, the ArchZone is modified accordingly. We increment the lower-bound constraints for all clock variables in the zone with the lower bounds from the intervals.

For example, in Fig. 2, when the message `unlock` is sent, we sum the lower bounds of the communication and operation execution intervals, i.e., $15 + 5 = 20$. Then, we increment the lower bound of the clock variable c in the ArchZone using that sum, as shown in the time line corresponding to Cut 2. Similarly, we increment the upper-bound constraints for all clocks using the upper bounds from the intervals, i.e., we add the upper bounds $25 + 15 = 40$ to the clock variable c in the ArchZone in Cut 2. The clock valuations for Cut 3 are computed analogously by summing up the communication delay interval of dl2bcm and the operation execution delay of `ack()`.

3.3 Comparing Architectural Time Progress with Real-time Requirements

The ReqZone and ArchZone are modified independently at the occurrence of an event and a time condition, respectively. After each such event in the simulation, we compare the two clock zones to verify whether the real-time requirements are violated by the architectural time progress. A clock zone can be interpreted as an infinite set of clock values that satisfies certain time constraints. Based on this set-oriented interpretation, we identify three possible outcomes for the comparison of the two zones.

1. The ArchZone is a subset of ReqZone: In this case, no matter what the clock valuations in ArchZone are, they will always be within the desired real-time requirements specified by the ReqZone. Therefore, we say that the zones are *consistent* and no violations can occur. This situation holds for the cuts 1 to 3 in Fig. 2.

2. The two zones are disjoint: Clearly, they will never be consistent, i.e., all clock values defined by ArchZone are not defined by ReqZone, and there will always be a violation of the real-time requirements when we consider the delays induced by the architectural properties. In this case, the semantics of the time condition temperature as presented in Sect. 2.3 hold.

3. The two zones intersect: In this case, both zones have clock values in common, but there are also values in the ArchZone that are not defined by the ReqZone. The intersection indicates that there could be an inconsistency between the zones that might lead to a violation of the real-time requirements. This is the case for the cut terminated in Fig. 2, where the ArchZone has the clock valuation $40 \leq c \leq 80$ and the ReqZone has the valuation $0 \leq c \leq 70$. In this case, it is up to the play-out user to decide whether to follow a pessimistic procedure (i.e., resolve the inconsistency since there could be a possible violation of a real-time requirement) or to judge whether the situation can occur in reality (i.e., possibly not to resolve the inconsistency).

Using this approach, we are able to detect inconsistencies between the assumptions about the architectural properties affecting the SUD's timing behavior and real-time requirements. In the context of the software development process, if the zones are inconsistent, this means that the real-time requirements in the specification are not satisfied by the architectural properties. There are two possible approaches to resolve the discrepancy. One is

to redesign the architecture of the system, so that the real-time requirements are satisfied. The other approach is to renegotiate the real-time requirements with the customer.

4 Related Work

Harel and Marelly [HM02] introduced the first version of the play-out algorithm for timed LSCs. Unfortunately, this timed play-out bases on a discrete time model, which is inconvenient for the specification of real-time systems since time has to be discretized a priori and time progress is influenced by discrete ticks (i.e., 71 particular ticks have to be performed to violate the time condition in our example). In contrast, we use a continuous time model, which aggregates infinitely many states only differing in concrete clock values to symbolic states (i.e., $c \leq 70$ and $c > 70$). Goknil et al. [GDPFM13] present an approach for the simulation of real-time requirements that are typical for automotive systems. Both approaches only consider requirements but neglect assumptions about architectural properties that affect the timing behavior. Therefore, they can only identify contradictory or under-specified real-time requirements but cannot verify whether the assumed architecture would fulfill the requirements.

Larsen et al. [LLNP09] present an approach for the formal verification of behavioral models in terms of timed automata w.r.t. real-time requirements specified in LSC scenarios. The approach bases on the UPPAAL model checker that also features simulation. The authors do not consider the architecture of the SUD. Maoz and Harel [MH11] visualize traces of reactive systems and thereby enable the user of their approach to inspect these traces in terms of high abstraction level requirements specified by means of MSDs. They do not focus on embedded systems and hence cover real-time requirements only in a very restricted manner. Lettrari and Klose [LK01] propose an approach to monitor and test real-time systems against real-time requirements specified by means of Message Sequence Charts. As our previous work on verifying traces of gear shift sequences against real-time MSD specifications [BGH+14], all of these approaches require behavioral design models or the fully implemented system as input for the verification against the requirements. Therefore, these approaches are only applicable in later development phases but not in RE.

5 Conclusion and Outlook

This paper presents an extension of our requirements modeling and simulation approach based on the scenario-based formalism of MSDs and implemented in the tool suite SCENARIOTOOLS. The extension enables to consider assumptions about architectural properties of the SUD affecting timing behavior and to verify whether they fulfill real-time requirements in real-time play-out. To achieve this objective, we extended on the one hand the modeling approach with the possibility to annotate architectural properties (i.e., operation execution times and connector latencies) to the hierarchical component architectures serving as structural basis for MSDs. We use a subset of the MARTE profile [Obj11b] for

these annotations. On the other hand, we extended real-time play-out to consider message delays (i.e., operation execution and communication delays) induced by the architectural properties and to verify them w.r.t. real-time requirements specified in MSDs. For the verification, we compare the time progress influenced by the message delays with the allowed time intervals specified by the real-time requirements.

Our extended approach enables to take assumptions about architectural properties affecting the timing behavior into account already in the RE phase. The simulation helps to identify possible violations of real-time requirements due to message delays induced by architectural properties. Therefore, possible requirements violations or defects can be identified early in the development process so that a different architecture can be chosen or the real-time requirements can be renegotiated with the customers. Thus, the approach helps to avoid costly iterations in subsequent development phases.

In the future, we plan to extend our approach into two different directions. On the one hand, our RE approach also covers the formal verification for the consistency of MSD specifications [GF12]. We plan to combine this formal verification approach with the extension presented in this paper in order to prove that the SUD, considering its architectural properties, does not violate real-time requirements instead of manually simulating all possible execution paths. On the other hand, we want to annotate assumptions about probabilistic architectural properties (e.g., quality of service aspects of bus systems) to the hierarchical component architecture and consider them in the requirements simulation.

Acknowledgments. This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster "Intelligent Technical Systems OstWestfalenLippe" (it's OWL) and managed by the Project Management Agency Karlsruhe (PTKA). The authors thank Thorsten Koch for helpful comments for the camera ready version of this paper.

References

[AD94] R. Alur and D. L. Dill. A theory of timed automata. *Theoretical computer science*, 126(2):183–235, 1994.

[BGH+14] C. Brenner, J. Greenyer, J. Holtmann, G. Liebel, G. Stieglbauer, and M. Tichy. ScenarioTools Real-Time Play-Out for Test Sequence Validation in an Automotive Case Study. In *Proc.* 13^{th} *Int. Workshop on Graph Transformation and Visual Modeling Techniques (GT-VMT 2014)*, 2014.

[BGPLM13] C. Brenner, J. Greenyer, and V. Panzica La Manna. The ScenarioTools Play-Out of Modal Sequence Diagram Specifications with Environment Assumptions. In *Proc.* 12^{th} *Int. Workshop on Graph Transformation and Visual Modeling Techniques (GT-VMT 2013)*, 2013.

[BY04] J. Bengtsson and W. Yi. Timed Automata: Semantics, Algorithms and Tools. In *Lectures on Concurrency and Petri Nets*, volume 3098 of *LNCS*, pages 87–124. Springer, 2004.

[DH01] W. Damm and D. Harel. LSCs: Breathing Life into Message Sequence Charts. *Formal Methods in System Design*, 19:45–80, 2001.

[Dil90] D. L. Dill. Timing Assumptions and Verification of Finite-State Concurrent Systems. In *Automatic Verification Methods for Finite State Systems*, volume 407 of *LNCS*, pages 197–212. Springer, 1990.

[GDPFM13] A. Goknil, J. DeAntoni, M.-A. Peraldi-Frati, and F. Mallet. Tool Support for the Analysis of TADL2 Timing Constraints Using TimeSquare. In *Proc.* 18^{th} *Int. Conf. on Engineering of Complex Computer Systems (ICECCS)*, pages 145–154, 2013.

[GF12] J. Greenyer and J. Frieben. Consistency Checking Scenario-Based Specifications of Dynamic Systems by Combining Simulation and Synthesis. In *Proc.* 4^{th} *Workshop on Behaviour Modelling - Foundations and Applications*, BM-FA '12, pages 2:1–2:9. ACM, 2012.

[HM02] D. Harel and R. Marelly. Playing with Time: On the Specification and Execution of Time-Enriched LSCs. In *Proc.* 10^{th} *IEEE Int. Symposium on Modeling, Analysis and Simulation of Computer and Telecommunications Systems 2002 (MASCOTS 2002)*, pages 193–202. IEEE Computer Society, 2002.

[HM03] D. Harel and R. Marelly. *Come, let's play: Scenario-based programming using LSCs and the play-engine*. Springer, 2003.

[HM08] D. Harel and S. Maoz. Assert and negate revisited: Modal semantics for UML sequence diagrams. *SoSyM*, 7:237–252, 2008.

[HM13] J. Holtmann and M. Meyer. Play-out for Hierarchical Component Architectures. In *Proc.* 11^{th} *Workshop Automotive Software Engineering*, volume P-220 of *LNI*, pages 2458–2472. Bonner Köllen Verlag, 2013.

[LK01] M. Lettrari and J. Klose. Scenario-Based Monitoring and Testing of Real-Time UML Models. In *≪UML≫ 2001 — The Unified Modeling Language. Modeling Languages, Concepts, and Tools*, volume 2185 of *LNCS*, pages 317–328. Springer, 2001.

[LLNP09] K. G. Larsen, S. Li, B. Nielsen, and S. Pusinskas. Verifying Real-Time Systems against Scenario-Based Requirements. In *FM 2009: Formal Methods*, volume 5850 of *LNCS*, pages 676–691. Springer, 2009.

[MH11] S. Maoz and D. Harel. On tracing reactive systems. *SoSyM*, 10(4):447–468, 2011.

[Obj10] Object Management Group. OMG Systems Modeling Language (OMG SysML): Version 1.2, OMG Document Number: formal/2010-06-01, 2010.

[Obj11a] Object Management Group. OMG Unified Modeling Language (OMG UML) Superstructure: V2.4, OMG Document Number: ptc/2010-11-14, 2011.

[Obj11b] Object Management Group. UML Profile for MARTE: Modeling and Analysis of Real-Time Embedded Systems: Version 1.1, OMG Document Number: formal/2011-06-02, 2011.

[Poh10] K. Pohl. *Requirements Engineering: Fundamentals, Principles, and Techniques*. Springer, 2010.

[VPBK10] K. Venkatesh P., M. Broy, and I. Krüger. Scanning Advances in Aerospace & Automobile Software Technology. *Proceedings of the IEEE*, 98(4):510–514, 2010.

Towards Efficient and Effective Testing in Automotive Software Development

Remo Lachmann and Ina Schaefer

Institute of Software Engineering and Automotive Informatics
Technische Universität Braunschweig
38106 Braunschweig
Email: {r.lachmann, i.schaefer}@tu-bs.de

Abstract: Software systems become more and more complex and control safety-critical applications. Hence, efficient and effective testing procedures are required in order to ensure software and system quality. Automotive software engineering has particular challenges with respect to efficient and effective testing, such as black-box test scenarios, natural language specifications and highly manual testing activities. In this paper, we present six automotive testing challenges and suitable approaches to solve them. We cover uniform test case specification to remove redundancies in test cases, automatic test case selection and prioritization, test case combination, and machine learning techniques for regression test selection. The presented approaches are illustrated by a case study from the automotive domain.

1 Introduction

Software systems become more and more complex and operate in safety-critical environments, e.g., in form of driver assistance systems for automobiles. Therefore, software quality has become an important property of automotive software systems, which has to be ensured for the resulting products. Software testing tackles this issue and is one of the most important and expensive parts of modern automotive software development. Failures, which have not been found by software testing, can cause immense financial downsides and risk for human lives.

The general testing process according to Spillner and Linz [SL12] is split into five phases as shown in Figure 1. It begins with the initial planning phase. Test concept and test plan are defined in this phase. Resources have to be allocated and test goals are defined. After the initial planning, the analysis and design phase follows. Here, test cases are described in an abstract way and defined according to the specification of the software. Subsequently, the test cases and test scripts are implemented. Then, the test cases to be executed are selected and run on the system-under-test, either automatically or manually. After test case execution, the results have to be documented and analyzed. The project lead has to decide whether further testing is necessary. This can be done by analyzing the rate of discovered and fixed failures over time or by other metrics. If testing is not finished, a new iteration of the testing process can be initiated.

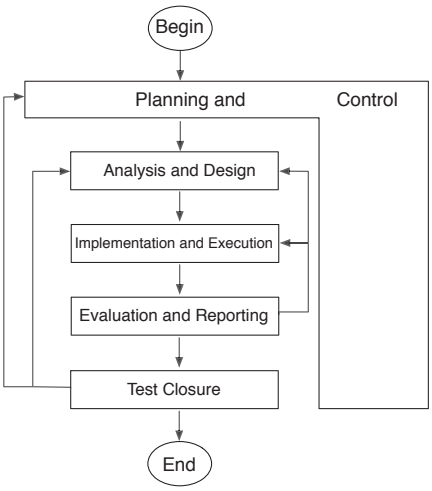


Figure 1: The general testing process [SL12]

Testing in general is heavily restricted by resources which leads to the goal to make testing more efficient and at the same time more effective. In automotive software testing, there are particular challenges. First, in the automotive industry there is a special relationship between OEM, supplier and testers. As software components are often developed by suppliers and tested locally, the integration step often happens by OEM or is performed by another supplier. Due to contractual restrictions, source code often is not available for integration and system testing. This leads to back-box testing scenarios where testing is based on the systems specification. Second, test artifacts are often specified in natural language [LS13]. This increases the difficulty of making assumptions about testing progress and coverage of the system-under-test and to automate test case selection and execution. Third, in most cases, not all specified test cases can be executed due to resource constraints. Therefore, a subset of test cases has to be selected. This is a very critical task, as it is not trivial to select the right test cases, which will probably detect failures. In black-box testing, which is the default for automotive software and systems, no source code is available. This leads to a lack of popular code-coverage metrics. Instead, other heuristics have to be introduced in the automotive domain for test case selection. Fourth, in automotive software development, many sub-processes of the overall testing process are performed manually. Automation of testing processes, including test case selection and prioritization, is a desirable goal, as it can reduce the testing effort compared to manual execution [KTS10].

In order to make automotive software testing more efficient, in this paper, we present six different testing challenges and according solutions within automotive software development taking the specific constraints and challenges in this domain into account. These approach cover test cases specified in natural language, automation of test case selection and prioritization and machine learning for regression test selection. We illustrate the proposed concepts using a running example, the *Body Comfort System* case study [LLLS13].

The BCS consists of different automotive comfort features, which are designed as software product line. Some of the included features are, for example, exterior mirror heating, automatic power window with finger protection, remote control etc. Certain features require other features, e.g., remote control requires a central locking system. Others exclude each other, e.g., manual and automatic power window.

The rest of the paper is structured as follows: Following to the general testing process, Section 2 describes two concepts to improve the design phase. Section 3 proposes approaches which support automated test case selection and prioritization in black-box testing. In Section 4, concepts to reuse knowledge from previous test executions for regression testing are introduced. Section 5 summarizes related work. Finally, in Section 6, we conclude the paper with possible future work.

2 Analysis and Design

Creating test cases in the analysis and design phase is a fundamental task, as it lays the foundation for the testing process. This phase is also important for any form of test automation, as it defines which information is available for testing and how it is accessible. However, occurring problems can get very expensive in later testing phases, e.g., missing test cases, insufficient test case descriptions, test case redundancy etc. The earlier such problems are detected, the cheaper they are to solve.

We describe two approaches to improve testing within this testing phase. First, we describe *systematical test case specification* for natural language test cases, which lays the foundation for test automation. Second, we propose an concept for *removal of test case redundancies*.

2.1 Systematic specification of black-box test cases

In the analysis phase, test cases are often formulated in natural language on system level. Usually, test cases consist of a precondition, action and expected result. Actions are divided into different steps of actions which the tester has to perform to execute the test case. Test cases are usually connected to certain requirements, which are often also specified in natural language. Both requirements and test cases are usually stored in test case management tools, e.g., HP Quality Center, IBM DOORS. Such tools allow for a traceability between test cases and requirements. One goal in black-box testing is to achieve at least requirements coverage. For example, a test case regarding the finger protection feature of the BCS case study could be written in natural language. The specification of the feature defines, that, if enabled, the finger protection detects objects (fingers) in the open window and blocks the window closure to prevent any harm to the finger. A test case for this feature could be defined by two independent testers as shown in Figure 2. A problem, which occurs in the definition of test cases is their actual description. As natural language is still predominant in industrial context, it depends on the tester how a test case is described.

Name:	01_powerWindow
Precondition:	Car has a power window, which is open. Finger protection is activated.
Action:	An object is held into upper third of the window. Then, button "window close" is pressed.
Expected Result:	Window closes until the object is detected. Then, it stops. The Finger Protection LED blinks.

Name:	tc_1_fingerProtection
Precondition:	PW is build in. PW is open. FP is activated
Action:	Hold hand in PW. Press BU.
Expected Result:	1) Window moves up 2) Window stops before hand 3) LED_FP blinks

Figure 2: Test cases for feature Finger Protection

For example, in the given test cases in Figure 2, one tester decided to write full sentences, while the other short bullets and abbreviations. Abbreviations are problematic, as they need to be defined. Otherwise, the test case description is opaque for other persons involved. There are many possibilities, where test case definitions can be different, but still have the same intention.

If test cases are analyzed the following aspects can be considered: if certain components or features are covered more than others, if there are redundancies between test cases, if test cases are too complex, if certain test cases can be combined etc. If a tool would have to analyze test cases as shown in Figure 2, it would be a very difficult and resource consuming task, as natural language processing is still a tough problem, especially if the descriptions are inconsistent as in the example. A solution to this problem is the definition of a systematic test case specification. This requires the test case designer to write test cases using a certain vocabulary, containing allowed abbreviations and terms, in a predefined structure. For example, in the given BCS scenario, the test designer always has to write *PowerWindow* and define precondition, action and expected result in this order. It can also be defined, how many test steps per test case are allowed. A thesaurus can be used to predefine all necessary buttons, fields, parameters etc. and how they are referenced within the test case description. This leads to a uniform test case specification. For example, one could argue that in the BCS case study, the test case variant shown right-hand side in Figure 2 is written according to the defined presets. Finally, such a test case specification has to be supported by the used tools. The tools can suggest to the test case designer, which parameters to reference, where to write them etc.

2.2 Removal of redundancies

Managing large sets of test cases as usual for automotive software development is a challenge. The number of test cases is growing over time, as new projects within a company are often based on previous software versions and reuse artifacts created for other projects [LLL+14]. Such artifacts can be requirements, software components, test cases

etc. But reuse can also introduce redundancy between artifacts. For example, assume a new version of the body comfort system if a new car model is developed. The previous project had several thousands of test cases, of which a lot can be reused. On the tooling side, the test cases are copied into the new project structure. But there is also new functionality to be tested. This leads to newly designed test cases, which are added into the projects structure. As a tester can not check thousands of old test cases for their exact description and test goal, it is possible that a newly described test case is covered by an already existing one. This can also easily happen if more than one test case designer works on the same subproject or functionality. If one now assumes the test cases are written in natural language as in Figure 2, redundancy is hard to detect. Within the BCS case study, there have been 97 requirements defined, for which we have defined 128 test cases in natural language [Beh13]. In interviews with industry experts, we found that automotive projects often comprise several hundred thousand requirements and test cases.

Redundancy can lead to highly inefficient testing as the *same* test case is performed several times without change leading to no novel results. One solution to tackle redundancy would be abandonment of reuse of test artifacts. This would force testers to write every test case from scratch. Thus, this is not an efficient approach due to the high number of test cases. We need a more intelligent approach, which can detect and reduce redundancy while still allowing for high reuse of artifacts. Therefore, test cases are processed for any reoccurring parts. This requires the test cases to be comparable. The approach shown in Section 2.1 builds a necessary foundation for the removal of redundancies in natural language test cases. If the test case specification is conform to a unified test specification, one can find similar parts within the test case structure over all test cases and identify redundant test cases. After the processing of the test cases, redundant test cases can either be marked and, if desired, removed. When writing a new test case, the test designer could be directly warned when his test cases is very similar to already existing ones. Another usage scenario is the merging of different projects. Here, redundancy could also be automatically detected and the respective test cases would be represented by one test case.

3 Implementation and Execution

After all test cases have been described, they have to be executed. It is also possible that the high-level test cases have to be implemented or additional test scripts have to be written. Even if the test cases are systematically described and free of redundancy, one has to find a trade-off between the potentially high number of designed test cases versus the very limited amount of available testing resources. The main challenges are the selection of test cases to be executed, and the ordering of these test cases. Test case prioritization [RUCH01, ZHZ+13] is necessary, as usually not all selected test cases can be executed because test cases have to be performed manually. For example, in the BCS case study, the finger protection has to be tested with human input, as an object has to be placed within the window. In test case prioritization, more important test cases are tested before less important ones. Therefore, some testing goals can be met earlier, which leads to benefits, especially, if not all selected test cases can be executed. Test case selection

and prioritization becomes even more challenging in the automotive domain since the source code of the system under test is unknown. Such black-box test scenarios offer a limited amount of information, e.g., about coverage metrics, than available in white-box techniques.

3.1 Effective test case selection and prioritization in black-box testing

A repeating effort has to be made to select the *right* test cases for test case execution. The first question, which needs to be answered, is what a good test case actually is. Kaner [Kan03] investigated this question and describes the quality of test case using several factors. In general testing, a good test case can be defined according to several distinct criteria, e.g., a test has a high likelihood to find a bug, it tests functions that are more likely to be used in reality, etc. Therefore, for creating test suites it is important to select those test cases, which fit the current test goal. The selected set of test cases defines a *test suite*, which then will be executed by the testers.

Assuming high-level test cases in a black-box testing scenario, the creation of a test suite usually is a manual task, which involves a lot of expert knowledge and experience regarding the system under test. This leads to two main problems. First, the knowledge of the test suite designer is often lost when the person leaves the project or company. Thus, the test selection is not exactly repeatable by others. It can be a very time consuming task to develop the necessary skills to create ”good” test suites. Second, manual selection is inefficient regarding the distribution of test cases, the time to select them and the failures they will find. As the selection procedure is not formally defined, results are not predictable, but only by the personal estimation of the test suite designer. Tool support could help to address test suite creation in a more structured and systematic way. Inexperienced testers could be guided to select and order test cases in test suites with the goal to test those first, which are more likely to expose failures.

To enable this, we introduce a greedy test case prioritization and selection technique, which is suitable for black-box testing. It requires a defined set of test cases and a set of selection/prioritization criteria as input. The result is a ordered list of test cases. A schematic description for this approach is shown in Fig. 3. The approach uses a given test suite, for which the test cases are ordered in some arbitrary way, e.g., alphabetical or after the date of creation. A set of predefined criteria is selected, such as the execution time, priority, tested functionality or history of the test cases. For example, the execution time of test cases could be used, to select those, which are executed fastest. This would be useful in a very time restricted environment. Another possible criterion is the test case history, which encompasses the number of failures found by a test case and their priority. Such a criteria is useful in a regression testing scenario, where test cases have to be executed again after some software update. Previously failed functionality should be tested again. A greedy ordering algorithm computes distinct ordered test suites for every given selection criterion. The best test case with respect to the selected criterion is always put first. To get a resulting test suite, the orderings are combined. To make this step more flexible for different project domains and project phases, the test suites can be combined based

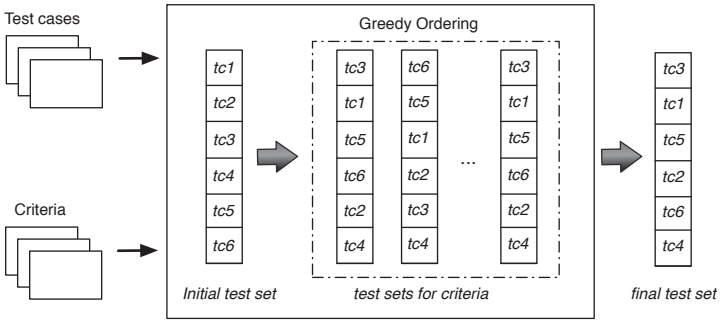


Figure 3: A greedy regression selection strategy

on some weighting function between the criteria. The weights for every criterion is either defined by the tester or can be automatically adjusted. For example, if a new software release within the BCS case study is to be released, where test case history is important, but the test resources are limited, the test case execution time and test history are used as criteria. The test history could be weighted with 75% and the execution time with 25% to impact the result. The resulting test suite is ordered in a form, that the first test case is the most important one, followed by the second most important one etc. Restricted in time and resources, the tester tries to test as many tests as possible in the given ordering.

This approach is very suitable for black-box testing in the automotive domain, as the defined criteria do not require any knowledge about the software and the source code itself, but only depend on the test cases. This solves the issue of missing code metrics, which would be primarily used in white-box testing. It is not guaranteed to find the optimum of selected test cases, but this approach could be used as a fast way to find a rather good result. A key factor for the success of this approach are the defined criteria and their weights.

3.2 Test case combination

One of the most common restrictions in testing is the available time. One possibility to increase the number of executable test cases under time restrictions is a combination of test cases. Similar to the approach described in Section 3.1, the ordering of test cases can lead to increased testing efficiency. Executing important test cases first is only a good guess, as failures are hidden and unknown both to tester and his tool. Thus, reducing the test execution time could lead to improved results compared to manual test selection. This can be achieved by ordering test cases according to their given pre- and post-conditions. For the BCS case study, there are two test cases for the feature *power window control*. This feature allows the user to move the power window using buttons on the remote control key. The test cases are specified as shown in Figure 4. One can see that these two test cases can be connected to a logical sequence, where first the window is moved up and then moved down. The post-condition of the first test case is the pre-condition of the second.

Name:	03_RemCon_Up
Precondition:	Central Locking System and automatic power window is build in. Power window is closed.
Action:	Press button "window up" on remote control.
Expected Result:	Power window moves up, until top position is reached (closed).

Name:	04_RemCon_down
Precondition:	Central Locking System and automatic power window is build in. Power window is closed.
Action:	Press button " window down" on remote control.
Expected Result:	Power window moves down, until bottom position is reached (open).

Figure 4: Two example test cases for control power windows

While the given example is rather trivial, finding such structures and coherences in several thousand test cases, which are written in natural language, is not. One possibility would be to investigate natural language terms and sentences to find possible combinations of test cases into test sequences. This requires a uniform test case specification as described in Section 2.1. Another possibility is to use model-based testing approaches [UL07] in order to generate test sequences from several test cases. State machines have proven to be a suitable method to represent executable system specifications in form of test models and to derive test cases for the modeled system. As benefit, transitions between states are clearly labeled and can be compared easily. Graph-based algorithms allow for fast traversal and comparisons within large sets of test models. Although model-based testing is an widely discussed and explored topic, one familiar problem is the lack of such models in industrial context. This has led to different approaches for *learning* test models based on given test cases [SM12, RSM08]. One can analyze these models and see what parts of the model are covered by which test cases. Therefore, it is possible to combine test cases,where the final state of one test case is the initial state of the other. Furthermore, thelearned models can be used to derive new test cases for the system. The better the model, the more detailed and distinct test cases one can generate.

4 Evaluation and Reporting of Testing results

After implementing and executing test cases, the test results have to be evaluated. Without evaluation, no knowledge can be extracted, and no new information is gained. Besides, the adaption of the testing process and the test cases is important. This incorporates the change of test case selection, test case prioritization, resource allocation etc. In regression testing, Kaner [Kan03] defines the goal to write *reusable test cases*. A good regression test case should be likely to fail, if changes induce errors in the tested area of the program under test. In regression testing, those test cases are selected for execution that are most likely to find failures introduced by change to an already tested system. This process is currently highly based on expert knowledge and the ability of testers to have insights into current test status. In the following, we show two possible approaches based on machine learning (ML) to introduce automation into this process.

4.1 Machine learning in regression testing

High-level test cases are often annotated with meta-data, e.g., the priority of test cases, tested functionality, test execution history. The meta-data can be used for test case selection and prioritization, as described in Section 3.1. For example, within the BCS case study one could argue that a test case regarding the basic functionality of an alarm system has a higher priority than a test case which tests the respective LED for this feature. The meta-data is defined by the test designer and is expensive in its creation, as it is not a trivial task to decide, e.g., what the priority of a test case is. This problem increases with every test iteration as associated meta-data should change over time, as the system under test does heavily change (e.g., bug fixes, new features, etc.). This leads to the question how meta-data can be changed after a test campaign without relying on manual intervention and expert knowledge.

One possible solution to avoid manual intervention when adapting the meta-data after test campaigns is to apply machine learning techniques (ML). ML has been already introduced to software testing [Bri08, LPV13]. ML approaches identify patterns in data, which can be used to derive new knowledge, and can be applied for decision-making. ML techniques can analyze the test case specification, requirements, test results and available testing history. Based on the findings, an ML approach can adapt the meta-data associated to the test cases for the next test iteration and selection process. For instance, one could envision a classifier, which decides if a test case is likely to reveal a failure or not and, hence, assign it a higher priority for a next test execution.

Figure 5 illustrates an example of the BCS study. The represented test case has been described in Figure 2. Let's assume, that the test case has been selected for a test suite and was executed. The test case revealed an important failure. The automated ML approach decides, that the priority of the test case has to be changed from the previous value *2* to *4* as the failure bug fix has to be verified. The execution time has been updated accordingly to the new result.

Figure 5: Adaption of black-box meta-data after change

4.2 Automatic regression test criteria selection

Test case selection for regression testing is a recurring task, which is often done manually. As described in Section 3.1, one can introduce a concept, where specific selection criteria are used to select an ordered list of test cases. The described method still needs human input, as the meta-data assigned to the test-cases for selection, the selection criteria and

the corresponding weights have to be provided by the tester. As shown in the previous section, machine learning techniques can be used to adapt meta-data in a black-box testing context. After each test run, however, the tester still would have to decide, how he has to alter the previously used criteria and their weights.

Therefore, we propose an extension to the approach of Section 3.1 in order to make it more automated. Instead of manually defining selection criteria, the extracted data from previous test runs can be used to adapt the selection criteria and weights for the next test campaign. For example, assume there is a defined testing criterion that a certain feature, e.g., the finger protection, is always tested. Within the latest test campaign, no failures have been found for this feature. Subsequently, the test system will automatically deselect this selection criterion or assign it a lower weight in the next run. This leads to a very flexible testing approach, where the tester is guided in his or her decision making, but can still make manual adjustments.

5 Related Work

In this section, we review related work in the areas covered in this paper, i.e., black-box testing in general and the specification of natural language test cases, test case selection and prioritization and machine learning-based testing.

Unsystematically specified test cases can lead to a high effort in introducing automation. Barros et al. present controlled natural language for use cases [BNHT11]. Strobbe et al. defined the *Test Case Description Language* in form of an XML-structure to describe test case meta-data [SHVV06]. This restricted languages help to create comparable and processable test case specifications. They also lay the foundation for possible tooling.

Test case selection and prioritization has been covered in the literature in the context of regression test selection and efficient testing of software product lines. Evolutionary algorithms have shown to be very convenient in prioritizing test cases [FKCA12]. Regression testing leads problems in selection and prioritization of test cases. Engström et al. [ERS10] have conducted a systematical review of empirical evaluations of regression test selection approaches. They notice that a comparison of the reviewed 28 techniques is rather hard, as they depend on many factors, which also influence the respective evaluations. Some examples for such factors are the granularity of the techniques, how frequent the regression cycles are or the products under test.

Software product lines are very relevant in the automotive context, as cars can be configured by the customer in many ways. Engström and Runeson [ER11] surveyed testing techniques for software product lines. They have identified three main research challenges in this field: (i) A (too) large number of tests, (ii) the trade-off between reusable components and the concrete products and (iii) treatment of variability. Cmyrev and Reißig [CR13] introduced a test case reduction technique for variant-rich systems, where a subset of product variants are selected while guaranteeing a complete requirements coverage. This approach is based on a greedy selection strategy and the connection between features of the product line and its requirements. The selection tries to find a minimal set of product configu-

rations while still covering 100% of the defined requirements. Johansen et al. present an algorithm to generate t-wise covering arrays for large feature models [JHF12]. This allows for combinatorial interaction testing of software product lines.

Briand [Bri08] has given an overview of machine learning approaches in software testing. He defines the role of ML in software testing and states that it can help to solve some of the long-standing software testing problems. One of the presented approaches is called MELBA (*MachinE Learning based refinement of BlAck-box test specification*) which is introduced as an ML based approach, which helps to refine test suites in black-box testing [BLB08]. It uses an incremental process, which is semi-automated. The approach helps to understand limits of test suites and to find redundancies. It does not, however, sort test cases in any order or make predictions about test case outcomes.

While the above works focus only on a single approach, this paper gives an overview of different techniques over three phases of the testing process to make testing in the automotive domain more effective and efficient.

6 Conclusion and Future Work

In this paper, we gave an overview of some challenges in automotive software testing. We focused on black-box testing scenarios, as source code often is not available in an OEM-supplier scenario in the automotive domain. Following the general testing process, we have presented six approaches to improve test efficiency considering the constraints of automotive software development. Some of the proposed concepts are already realizable, e.g., the systematic test case description. Others need some future work to be of practical use, e.g., machine learning approaches in black-box testing.

References

[Beh13] M. Behr. Evaluation eines anforderungsbasierten, delta-orientierten Testverfahrens für Softwareproduktlinien. Bachelor thesis, TU Braunschweig, 2013.

[BLB08] L.C. Briand, Y. Labiche, and Z. Bawar. Using Machine Learning to Refine Black-Box Test Specifications and Test Suites. In *Quality Software, 2008. QSIC '08. The Eighth International Conference on*, pages 135–144, Aug 2008.

[BNHT11] F. A. Barros, L. Neves, E. Hori, and D. Torres. The ucsCNL: A Controlled Natural Language for Use Case Specifications. In *SEKE*, pages 250–253. Knowledge Systems Institute Graduate School, 2011.

[Bri08] L.C. Briand. Novel Applications of Machine Learning in Software Testing. In *Quality Software, 2008. QSIC '08. The Eighth International Conference on*, pages 3–10, Aug 2008.

[CR13] A. Cmyrev and R. Reißig. Optimierte Varianten- und Anforderungsabdeckung im Test. In *Automotive Software Engineering Workshop, 43. GI Jahrestagung*, 2013.

[ER11] E. Engström and P. Runeson. Software product line testing – A systematic mapping study. *Information and Software Technology*, 53:2–13, 2011.

[ERS10] E. Engström, P. Runeson, and M. Skoglund. A systematic review on regression test selection techniques. *Information and Software Technology*, 52:14–30, 2010.

[FKCA12] J. Ferrer, P. M. Kruse, F. Chicano, and E. Alba. Evolutionary Algorithm for Prioritized Pairwise Test Data Generation. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation*, GECCO '12, pages 1213–1220. ACM, 2012.

[JHF12] M. F. Johansen, Ø. Haugen, and F. Fleurey. An algorithm for generating t-wise covering arrays from large feature models. In *SPLC*, pages 46–55, 2012.

[Kan03] C. Kaner. What is a good test case? In *Software Testing Analysis & Review Conference (STAR) East*, 2003.

[KTS10] J. Kasurinen, O. Taipale, and K. Smolander. Software Test Automation in Practice: Empirical Observations. *Advances in Software Engineering*, 2010:571–579, 2010.

[LLL+14] M. Lochau, S. Lity, R. Lachmann, I. Schaefer, and U. Goltz. Delta-oriented model-based integration testing of large-scale systems. *The Journal of Systems and Software*, 91:63–84, 2014.

[LLLS13] S. Lity, R. Lachmann, M. Lochau, and I. Schaefer. Delta-oriented Software Product Line Test Models - The Body Comfort System Case Study. Technical report, TU Braunschweig, 2013.

[LPV13] A. R. Lenz, A. Pozo, and S. R. Vergilio. Linking software testing results with a machine learning approach. *Engineering Applications of Artificial Intelligence*, 26(5–6):1631 – 1640, 2013.

[LS13] R. Lachmann and I. Schaefer. Herausforderungen beim Testen von Fahrerassistenzsystemen. In *11. Workshop Automotive Software Engineering (ASE)*, 2013.

[RSM08] H. Raffelt, B. Steffen, and T. Margaria. Dynamic Testing via Automata Learning. In *Proceedings of the 3rd International Haifa Verification Conference on Hardware and Software: Verification and Testing*, HVC'07, pages 136–152. Springer-Verlag, 2008.

[RUCH01] G. Rothermel, R. H. Untch, C. Chu, and M. J. Harrold. Prioritizing Test Cases For Regression Testing. *IEEE Transactions on software engineering*, Vol.27 No.10:929–948, 2001.

[SHVV06] C. Strobbe, S. Herramhof, E. Vlachogiannis, and C. A. Velasco. Test Case Description Language (TCDL): Test Case Metadata for Conformance Evaluation. In *ICCHP*, pages 164–171, 2006.

[SL12] A. Spillner and T. Linz. *Basiswissen Softwaretest - Aus-und Weiterbildung zum Certified Tester*. dpunkt.verlag, 2012.

[SM12] M. A. Sindhu and K. Meinke. IDS: An Incremental Learning Algorithm for Finite Automata. *CoRR*, abs/1206.2691, 2012.

[UL07] M. Utting and B. Legeard. *Practical Model-based Testing*. Morgan Kaufmann, 2007.

[ZHZ+13] L. Zhang, D. Hao, L. Zhang, G. Rothermel, and H. Mei. Bridging the Gap between the Total and Additional Test-Case Prioritization Strategies. In *International Conference on Software Engineering, ICSE 2013*, 2013.

Frühzeitiges Absichern von Entwurfsentscheidungen durch Virtuelle Deployments

Thomas Kuhn

Embedded Software Development
Fraunhofer IESE
Fraunhofer-Platz 1
67663 Kaiserslautern
thomas.kuhn@iese.fraunhofer.de

Abstract: Multi-Core ECUs ermöglichen eine Vielzahl neuartiger Software- und Systemarchitekturen, fordern Architekten jedoch auch viele grundlegende Entscheidungen ab, die oft weichenstellend für die spätere Implementierung sind. Diese Entscheidungen sind oft schwierig zu revidieren, müssen jedoch oft basierend auf Entscheidungen getroffen werden, da es nicht möglich ist deren Auswirkungen frühzeitig zu untersuchen. In diesem Paper stellen wir das Simulationsframework FERAL vor, dass es ermöglicht bestehende oder prototypische Verhaltenskomponenten in einer simulierten Umgebung auszuführen, die die relevanten Eigenschaften der geplanten Softwarearchitekturen widerspiegelt. Die Verhaltenskomponenten müssen dafür nicht modifiziert werden. Auswirkungen von Design- und Deploymententscheidungen können so frühzeitig vorhergesagt und gegebenenfalls korrigiert werden.

1 Einleitung

Multi-Core Prozessoren und die damit mögliche Konsolidierung von unabhängigen Steuergerätefunktionen auf einer gemeinsamen Hardware stellen Software und Funktionsentwickler vor Herausforderungen. Bisher wurde Software in der Regel auf föderierten Architekturen betrieben; jede Funktion wird auf einem unabhängigen Steuergerät ausgeführt. Daten werden über Netzwerke ausgetauscht. Dies vereinfacht die Entwicklung von Funktionen, da die netzwerkbasierte Kommunikation klare Schnittstellen vorschreibt. Interferenzen zwischen Funktionen treten in der Regel nur dann auf wenn mehrere Funktionen den gleichen Aktuator nutzen. Diesem Aspekt wird dadurch Rechnung getragen, dass Funktionen, die die gleichen Aktuatoren nutzen, normalerweise auf demselben Steuergerät ausgeführt werden, und daher während der Steuergeräteentwicklung frühzeitig miteinander integriert werden können.

Verglichen mit den für die Steuergeräteentwicklung bisher genutzten Prozessoren bieten Multi-Core ECU[1]s deutlich mehr Ressourcen. Diese können dazu genutzt werden um

[1] ECU = Electronic Control Unit, diese Abkürzung bezeichnet eine Steuergerätehardware für Automotive-Anwendungen

eine Vielzahl von Funktionen auf einem einzigen Steuergerät zu konsolidieren. Ebenfalls denkt man darüber nach, Regelkreise auf mehrere Steuergeräte zu verteilen, um zum Beispiel eine effizientere Datenverarbeitung in der Nähe von Sensoren zu ermöglichen und freie Prozessor- und Speicherkapazitäten besser nutzen zu können.

Software-Plattformen wie zum Beispiel AUTOSAR [KF09] unterstützen die Entwickler teilweise bei diesen Aktivitäten. AUTOSAR trennt Plattformdienste und Applikationsfunktionen, mehrere Funktionen können auf einer Steuergerätehardware ausgeführt werden und auch über Steuergerätegrenzen hinaus miteinander kommunizieren. Diese Kommunikation wird durch AUTOSAR automatisch auf das genutzte Kommunikationsnetzwerk, zum Beispiel CAN, abgebildet. Ebenso bildet AUTOSAR Anwendungsfunktionen auf Tasks ab, die im Fall einer Multi-Core ECU gleichzeitig ausgeführt werden können. Auch bei der Nutzung von AUTOSAR müssen Entwickler sicherstellen, dass bei der parallelen Ausführung ihrer Applikationen keine Dateninkonsistenzen auftreten und geteilte Ressourcen konfliktfrei genutzt werden können. Ebenfalls müssen auf mehrere Steuergeräte verteilte Algorithmen robust sein gegenüber Kommunikationsfehlern, sowie absichtlichem oder unabsichtlichem Fehlverhalten anderer Busteilnehmer.

Daher ist es erforderlich, dass sowohl Software- als auch Plattformkonzepte frühzeitig und unabhängig voneinander getestet werden, um die notwendige Betriebssicherheit zu gewährleisten und teuren Fehlentscheidungen vorzubeugen. In dieser Arbeit illustrieren wir, wie Parallelisierungs- und Deployment-Entscheidungen frühzeitig mittels des Simulationsframeworks Fraunhofer FERAL (FramEwork for Simulator Coupling on Requirements and Architecture Level) durch virtuelle Deployments getestet werden können. Wir zeigen, wie Entwickler die Auswirkungen ihrer Entscheidungen durch Simulationstechnologien frühzeitig evaluieren und bewerten können. Im Gegensatz zu häufig genutzten Hardware-in-the-Loop Lösungen können Architekturkonzepte daher bereits getestet werden, bevor diese für die Zielplattform umgesetzt wurden. Ebenfalls illustrieren wir wie diese Evaluation ohne eine Modifikation des evaluierten Applikationscodes nur durch umgebende Simulationskomponenten durchgeführt werden kann um Designfehler frühzeitig zu erkennen und den Aufwand für Simulationen signifikant zu senken.

2 Software im Automobilbereich

Im Fokus dieser Arbeit steht Software im Automobilbereich, die steuernde und regelnde Aufgaben wahrnimmt. Diese Software realisiert verschiedene Funktionen, wobei eine Funktion eine größere Einheit darstellt. Funktionen sind in der Regel weiter in Unterfunktionen strukturiert, und werden schließlich durch Runnables implementiert. Häufig wird eine Funktion daher durch mehrere kommunizierende Runnables realisiert. Dieses kommunizierende Netzwerk bezeichnet man als Funktionsnetz.

2.1 Automotive Software

Diese Kommunikation wird Beispielsweise mittels globaler Variablen realisiert. Basierend auf den von einem Runnable gelesenen und geschriebenen Daten lassen sich Funktionsnetze konstruieren, die die Abhängigkeiten zwischen Funktionen darstellen.

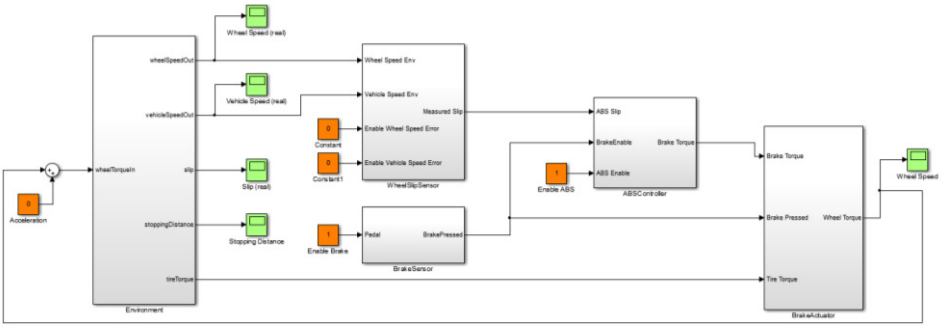


Abbildung 1: Beispielsystem in Simulink

Abbildung 1 illustriert ein in Simulink erstelltes Funktionsnetz das ein vereinfachtes ABS System realisiert. Es werden vier Funktionen und die Umgebung simuliert:

- Der Wheel Slip Sensor überwacht ein Rad des Autos und prüft ob dieses blockiert und rutscht.
- Der Brake Sensor überwacht das Bremspedal des Brake-By-Wire Systems und signalisiert dem ABS Controller elektrisch einen Bremswunsch des Fahrers.
- Der ABS Controller steuert das ABS System und entscheidet ob mit der vollen Bremskraft gebremst werden soll, oder ob diese aufgrund blockierender Räder reduziert werden muss.
- Der Brake Actuator nimmt die Befehle des ABS Controllers und des Fahrers entgegen und steuert die Bremse. Der ABS Controller kann den Bremsdruck abschwächen, jedoch keinen eigenen Bremsdruck aufbauen.

2.2 Simulation auf Funktionsebene mit FERAL

Während der Funktionsentwicklung muss es möglich sein ein System zu testen. Wird die Entwicklung wie bei dem in Abbildung 1 gezeigten Beispielsystem vollständig in einem Werkzeug durchgeführt, so ist es natürlich möglich das Verhalten der Algorithmen innerhalb dieser Entwicklungsumgebung ohne zusätzliche Werkzeuge zu testen. Immer häufiger ist dies jedoch nicht möglich, da entweder Applikationskomponenten die das Verhalten von Runnables realisieren nicht im Quelltest vorliegen, oder diese mit anderen Werkzeugen entwickelt wurden. Das Testen eines Funktionsnetzes erfordert in diesem Fall eine semantische Integration von verschiedenen Runnables. Diese Integration sollte

ermöglich werden ohne dass Applikationskomponenten verändert werden müssen. Dies ermöglicht die Integration von Komponenten, deren Quelltext oder Designmodell nicht verfügbar ist.

Das Fraunhofer FERAL Simulationsframework ermöglicht die integrierte Simulation von verschiedenen Runnables in einem Kontext. Dieser wird komponentenbasiert realisiert. Das Kernkonzept von FERAL ist dabei die Trennung von Simulationsmodellen und dem simuliertem Verhalten. Simulationsmodelle, d.h. Berechnungs- und Kommunikationsmodelle (Models of Computation and Communication - MOCC) die zur Ausführung von Funktionen und Simulationsmodellen erforderlich sind werden wiederverwendbar von Direktoren implementiert. Berechnungsmodelle definieren wann eine bestimmte Komponente ausgeführt wird, und wie lange diese Ausführung in Simulationszeit dauert und dauern darf. Kommunikationsmodelle definieren wann welche Daten wie zwischen Simulationskomponenten ausgetauscht werden. Das konkrete simulierte Verhalten einer Funktion wird durch einen Worker bereitgestellt. Worker erfordern bestimmte Verhaltensmodelle, daher können diese nur zusammen mit kompatiblen Direktoren eingesetzt werden. Diese Trennung ermöglicht eine effiziente Simulatorkopplung: diese erfordert eine Integration der unterschiedlichen Ausführungs- und Kommunikationsmodelle, die wiederverwendbar durch Direktoren bereitgestellt wird. Die Integration eines konkreten Simulators wird mittels eines Workers realisiert, der lediglich eine technische Schnittstelle zum Simulationsmodell implementieren muss.

Angelehnt an Ptolemy [Ek03] definieren sowohl Direktoren als auch Worker eine gemeinsame Schnittstelle. Dies ermöglicht die Schachtelung von Direktoren und Workern in hierarchische Szenarien. Jede Komponente bis auf die Wurzel hat exakt einen übergeordneten Direktor, der deren Ausführung kontrolliert. Worker treten nur als Blätter auf. Auf diese Weise werden Simulationsszenarien realisiert, die unterschiedliche Simulationsmodelle vereinigen. Die vollständige Semantik des FERAL Frameworks ist in [Ku13] dokumentiert. Für die Ausführung von Komponenten im Kontext von virtuellen Deployments mit FERAL sind insbesondere die folgenden Komponenten relevant:

- TimeTriggeredDirector: Dieser Direktor führt die kontrollierten Komponenten in festen zeitschritten aus. Optional kann eine Reihenfolge vergeben werden.

- EventTriggeredDirector: Dieser Direktor führt enthaltene Komponenten dann aus, wenn ein Ereignis für eine Komponente aufgetreten ist. Dies kann ein Timer oder ein Nachrichtenempfang sein.

- SimulinkWorker: Integriert ein Simulink System als Verhaltensmodell. Dieser Worker benötigt eine Ausführsemantik, die zu der des TimeTriggeredDirector kompatibel ist.

- FMUWorker: Analog zu dem SimulinkWorker, für Functional Mockup Units.

- TimeTriggeredWorker: Worker dessen Ausführung zu festen Zeitpunkten durch dessen Direktor angestoßen wird. Ein- und Ausgangsports verhalten sich wie

Variablen die während der Ausführung beliebig gelesen und geschrieben werden. Dies ist die Basisklasse des Simulink- und FMU Workers.

- EventTriggeredWorker: Ereignisbasierter Worker, der von seinem Direktor bei Nachrichtenempfang oder Ablauf eines selbst gesetzten Timers ausgeführt wird. Eingabeports verhalten sich wie Nachrichtenqueues, Nachrichten an Ausgangsports werden vom Direktor zu definierten Zeitpunkten an verbundene Eingabeports anderer Komponenten weitergeleitet.

- Constant: Spezieller TimeTriggeredWorker der lediglich eine Konstante an einem Ausgangsport bereitstellt.

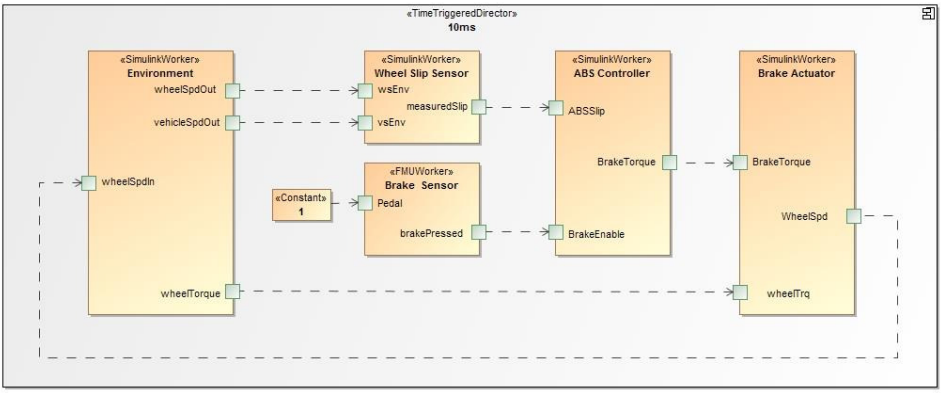


Abbildung 2: Modell eines Simulationsszenarios des FERAL Simulationsframeworks

Abbildung 2 illustriert den Aufbau solches Szenario mit Komponenten des FERAL Frameworks. Genutzt wird ein TimeTriggeredDirector, die alle seine kontrollierten Komponenten in äquidistanten Zeitschritten aufruft und den Datenaustausch über Ports und Links kontrolliert. Verschiedene Worker binden Simulink und FMU Subsysteme ein und ermöglichen die gemeinsame Ausführung des Systems in einem integrierten Szenario.

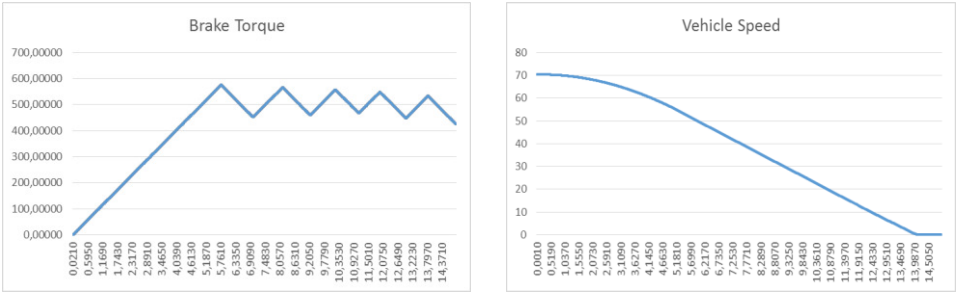


Abbildung 3: Verhalten des Beispielsystems in einem simulierten Szenario

Abbildung 3 illustriert das Verhalten des simulierten Systems. Man erkennt deutlich das Verhalten das ABS Systems der die Räder, sobald sie aufgrund einer zu hohen Bremskraft zu blockieren beginnen wieder freigibt und so das Bremsverhalten des Automobils optimiert.

3 Virtuelle Deployments mit Fraunhofer FERAL

Runnables werden auf einer realen Zielplattform durch ein Netzwerk von Steuergeräten (Electronic Control Unit - ECU) ausgeführt. Dabei spielen Deployment Entscheidungen eine wesentliche Rolle. Diese legen fest auf welcher ECU ein Runnable ausgeführt wird, welcher Task diese Ausführung übernimmt, und welche Daten aufgrund des Deployments über Kommunikationsnetze ausgetauscht werden müssen. Das Deployment hat einen wesentlichen Einfluss auf die Ausführung eines Algorithmus. Wartezeiten (Delays) und deren Varianzen (Jitter) die z.B. durch Multitasking und die Nutzung von Kommunikationsnetzwerken bei der Ausführung von Runnables entstehen, beeinflussen das Verhalten eines Algorithmus signifikant. Um die Auswirkungen dieser Entscheidungen vorherzusagen muss das in Abbildung 2 gezeigte Szenario daher um die Simulation von Deployments erweitert werden. Ziel ist hierbei, die Komponenten des FERAL Frameworks zu nutzen um Eigenschaften der Zielplattform in die Simulation einzubringen, ohne hierfür die Applikationskomponenten anzupassen.

3.1 Virtuelle Deployments auf ECUs

Um ein Deployment zu simulieren müssen in einem ersten Schritt die Tasks der Zielplattform mit FERAL nachgebildet werden. Hierbei sind zwei Arten von Tasks besonders relevant:

- Zeitgesteuerte Tasks wechseln zu bestimmten Zeitpunkten in den Zustand „Bereit“. Abhängig von ihrer Priorität werden diese danach mehr oder weniger zeitnah ausgeführt. Diese Tasks werden für die meisten Runnables genutzt, die steuernde und regelnde Aufgaben übernehmen.

- Ereignisgesteuerte Tasks werden durch ein externes Ereignis in den Zustand „Bereit“ versetzt. Dies ist häufig ein Interrupt, kann aber auch durch den Empfang einer Nachricht über ein Netzwerk oder den Aufruf einer bereitgestellten Funktion durch einen anderen Task ausgelöst werden. Ein Beispiel für einen solchen Task ist die Steuerung eines Einspritzvorgangs, der mit der Kurbelwellenposition und Geschwindigkeit synchronisiert werden muss, und für den daher keine feste Periodendauer definiert werden kann.

Mit FERAL werden Deployments zum Beispiel mit Hilfe des HybridScheduler Direktors realisiert. Es handelt sich hierbei um eine spezialisierte Form des EventTriggeredDirector. Der HybridScheduler realisiert das prioritätsbasierte Scheduling eines OSEK konformen Betriebssystems und kann entweder präemptiv oder nicht präemptiv konfiguriert werden. Runnables werden weiterhin über Worker

eingebunden, die Ausführung der Worker wird bei Verwendung dieses Direktors durch Tasks gesteuert, auf die die Worker allokiert wurden.

Abbildung 4 zeigt ein virtuelles Deployment mit FERAL, dass eine AUTOSAR-konforme Ausführplattform simuliert. Die Kommunikation zwischen dem Worker und der Plattform findet über Signale statt. Die Ausführung von Workern wird bei Nutzung dieses Direktors über Tasks gesteuert. Es sind zwei Tasks definiert: ein periodischer Task führt den ABSController alle 10ms aus. Ein ereignisbasierter Task führt die RTE aus und reagiert dabei auf Scheduling Events des simulierten Betriebssystems. Die RTE ermöglicht den Datenaustausch zwischen Tasks und steuert die Netzwerkkommunikation. Ausführzeiten die für Taskwechsel oder den atomaren Zugriff auf Speicheradressen benötigt wird, wird durch die Ausführzeiten der RTE simuliert.

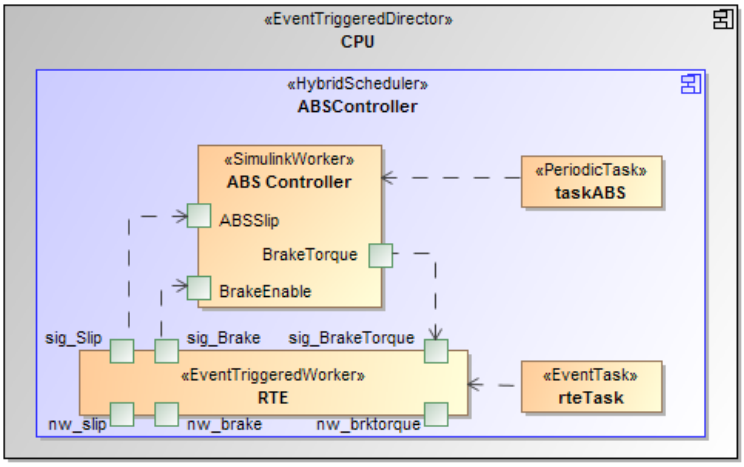


Abbildung 4: Modell eines Simulationsszenarios für virtuelles Deployment auf ECU

Die Runnables des simulierten Szenarios werden hier bereits mit einer Ressourcenbeschränkung ausgeführt. Im Gegensatz zu dem vorher genutzten TimeTriggeredDirector implementiert der hier genutzte Scheduler ein Ausführmodell das Tasks auf einer festen Anzahl Rechenkernen ausführt. Jeder Task benötigt eine bestimmte Menge Rechenzeit, für die er den Kern blockiert. Die Reihenfolge in der bereite Tasks ausgeführt werden wird durch deren Priorität bestimmt.

AUTOSAR nutzt ein preemptives Scheduling, daher können höherpriore Tasks solche mit niedriger Priorität verdrängen. Dies kann die Ausführzeit eines Tasks signifikant verlängern. FERAL bildet dieses Verhalten nach, indem die Ausführung eines Tasks mehrere definierte Zeitintervalle dauern kann. Nach jedem Zeitintervall wird geprüft ob der Task noch weiter rechnen darf, oder von einem Task mit höherer Priorität verdrängt wurde. Abbildung 5 zeigt dafür ein Beispiel. Hier wird ein Task mit einer Gesamtlaufzeit von 800µs in Zeitschlitzen von 100µs ausgeführt und während der Ausführung von einem Task mit einer höheren Priorität verdrängt. Die Länge der Zeitschlitze kann für jeden Task abhängig von der geforderten Genauigkeit individuell

definiert werden, ebenfalls müssen diese nicht äquidistant sein. FERAL führt die Berechnung eines Tasks in diesen Zeitschlitzen aus. Da bestehende Runnables unmodifiziert ausgeführt werden sollen ist es möglich, dass diese nur einen atomaren Zeitschritt realisieren. Dies ist zum Beispiel bei Code der aus Simulink generiert wurde der Fall. In diesem Fall muss der Task festlegen, in welchem Zeitschlitz das Runnable tatsächlich ausgeführt werden soll. Ferner muss definiert werden in welchen Zeitschlitzen Eingangsdaten gelesen- und Ausgangsdaten geschrieben werden. Das Standardverhalten des Direktors definiert alle lesenden Operationen und die atomare Berechnungen während des ersten Zeitschlitzes, die schreibenden Operationen während des letzten Zeitschlitzes.

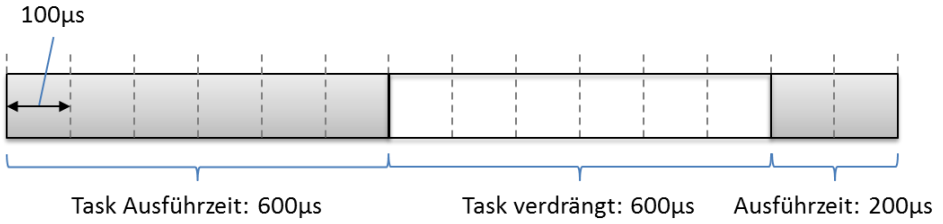


Abbildung 5: Simulationsszenario des FERAL Simulationsframeworks

Mit dem gezeigten Ansatz wird ein Deployment auf eine ECU simuliert, ohne das die eigentlichen Runnables dafür modifiziert werden müssen. Einfache Parallelisierungsfehler können so bereits erkannt werden: greift ein Task schreibend auf Variablen zu, die durch einen verdrängten Task ebenfalls genutzt werden, so handelt es sich dabei in der Regel um ein ungewolltes Verhalten. Außerdem ist es möglich den Einfluss eines Schedulings auf das Verhalten eines Systems zu bewerten.

3.2 Virtuelle Deployments auf Multi-Core ECUs

Die parallele Ausführung von Funktionen wird aufgrund der sich immer weiter verbreitenden Multi-Core Prozessoren immer wichtiger. FERAL simuliert diese Deployments wie in Abbildung 6 illustriert durch zusätzliche Kerne, die simuliert parallel ausgeführt werden. Jeder dieser Kerne wird durch einen eigenen Direktor vom Typ HybridScheduler repräsentiert. Zusätzlich zu den beiden bereits allokierten Funktionen WheelSlipSensor und BrakeActuator wird nun eine weitere Funktion PowerSteeringActuator auf einen zweiten Kern der ECU allokiert. Abbildung 6 zeigt diese Multi-Core ECU als Erweiterung des Deployments aus Abbildung 4. Aus Gründen der Übersicht wurden nur die Worker ohne Tasks modelliert.

Um Dateninkonsistenzen bei der gleichzeitigen Ausführung von Runnables auf Multi-Core Prozessoren zu vermeiden werden zum Beispiel Softwaretransaktionen genutzt (siehe [ST97]). Diese kopieren Eingabedaten zu einem definierten Zeitpunkt in einen lokalen Puffer. Das ausgeführte Runnable arbeitet nur auf diesen Daten, die zu einem weiteren Zeitpunkt mit den globalen Daten wieder synchronisiert werden. So muss das Runnable nicht geändert werden und die Synchronisationskonzepte können unabhängig

von der Implementierung der Runnables realisiert werden. Finden die Lese/Schreib-Transaktionen in definierten Zeitschlitzen statt in denen nur eine Transaktion aktiv ist, kann ein konfliktfreier Zugriff garantiert werden. Aufgrund dieser Eigenschaften ermöglichen Softwaretransaktionen sehr elegante Parallelisierungskonzepte.

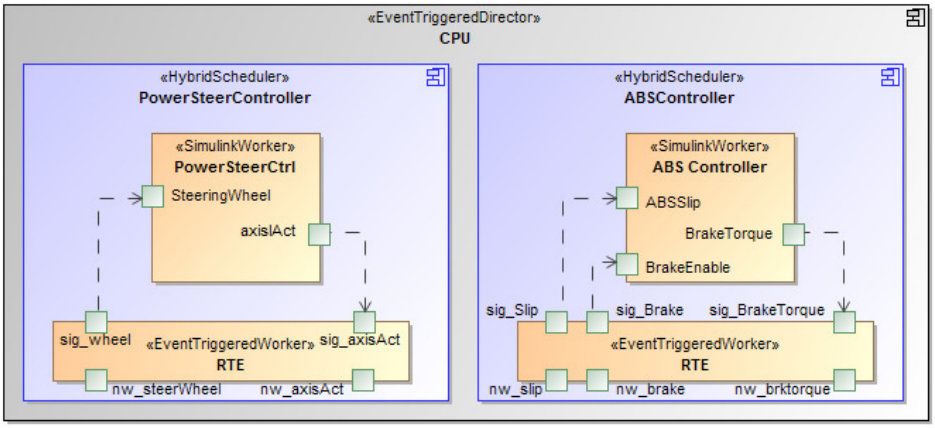


Abbildung 6: Simulationsszenario für eine Multi-Core ECU

Mittels FERAL können diese Konzepte validiert werden. Abbildung 7 illustriert die Modellierung einer Softwaretransaktion mit dem FERAL Simulationsframework. Auf diese Weise ist es möglich, frühzeitig den Einfluss von Softwaretransaktionen zu überprüfen, sowie zu überprüfen ob Transaktionen basierend auf Tasklaufzeiten korrekt konfiguriert wurden und konfliktfrei ablaufen.

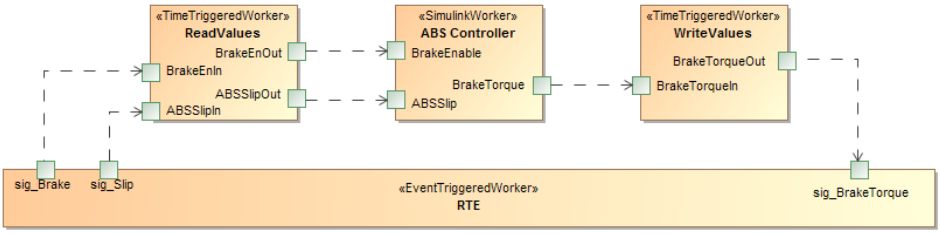


Abbildung 7: Modell der Realisierung einer Softwaretransaktion in FERAL

3.3 Virtuelle Deployments auf Steuergerätenetzwerke

Neben dem Task-Scheduling ist auch die Verteilung von Funktionen auf Steuergerätenetzwerke relevant. Abbildung 8 zeigt das Modell eines solches komplexes Simulationsszenario. Dieses besteht aus drei Single-Core ECUs, sowie einem CAN Bus Netzwerk das die ECUs miteinander verbindet. Das Umgebungsmodell wird von einem TimeTriggeredDirector ohne Ressourcenconstraints ausgeführt. Die Softwarefunktionen

werden erneut durch den HybridScheduler ausgeführt um das Scheduling der ECU zu simulieren. Das Kommunikationsnetz wird durch ereignisbasierte Worker realisiert. Die RTE wurde nicht dargestellt um die Grafik zu vereinfachen.

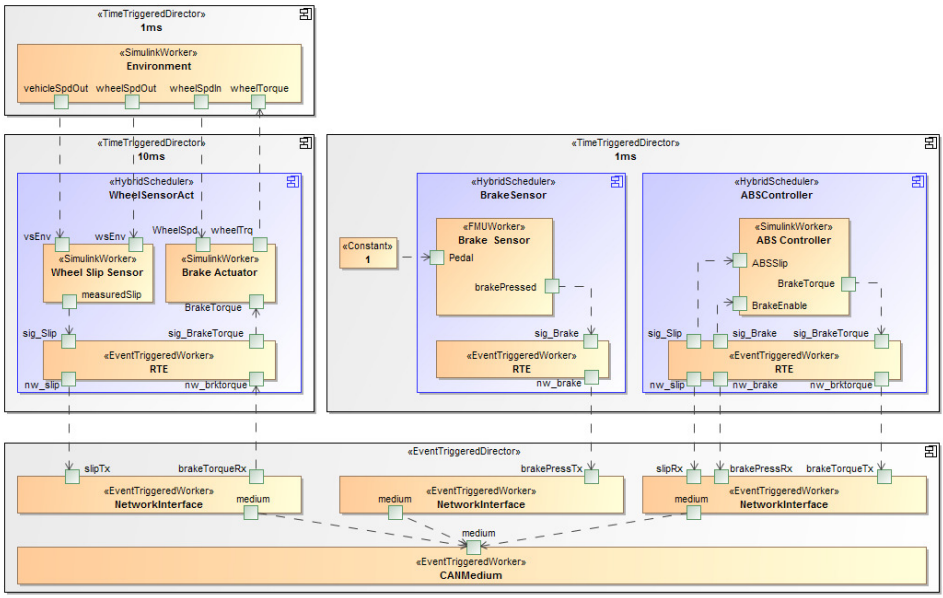
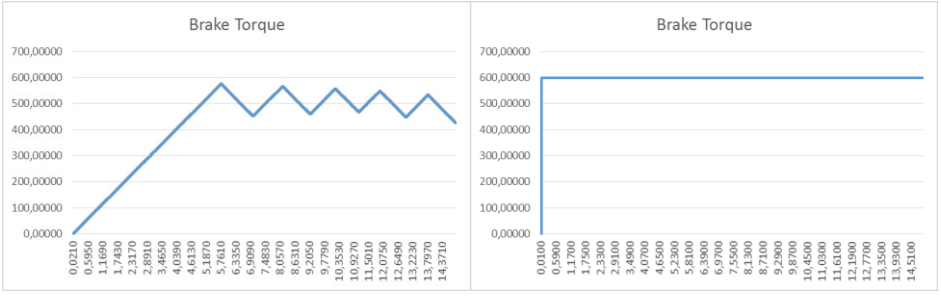


Abbildung 8: Modell eines FERAL Simulationsszenarios für ein Steuergerätenetzwerk

Wie man in Abbildung 9.a sehen kann beeinflusst dieses Deployment die Ausführung der Komponenten nur minimal. Dies liegt daran, dass ausreichend Ressourcen in diesem Szenario zur Verfügung stehen, und daher sowohl die Ausführzeit als auch die Kommunikation nur minimalen Verzögerungen unterliegt, die in dieser Größenordnung für den Algorithmus nicht kritisch sind.



(a) ohne Powersteering-Komponenten (b) mit Powersteering-Komponenten

Abbildung 9: Resultate der Simulation

Abbildung 9.b zeigt das Verhalten dieses Szenario mit zwei zusätzlich verteilten Funktionen: Um Kosten zu sparen wurden Funktionen der Servolenkung auf die ECUs BrakeSensor und WheelSensorAct verteilt. Hierdurch wird ein weiterer Regelkreis in das System eingebracht, der zusätzliche Ressourcen benötigt. Dadurch wird der CAN Bus überlastet und Nachrichten mit niedrigerer Priorität verworfen. Im betrachteten Beispiel sind dies die Nachrichten der Bremskraftsteuerung des ABS, während die Brake-by-Wire und Steer-by-Wire Signale der Bremse und der Servolenkung übertragen werden. Aufgrund der durch die Simulationskomponenten über ihre Schnittstellen publizierten Daten lässt sie die Quelle der Störungen leicht finden und beheben, zum Beispiel in dem Nachrichten zusammengefasst werden, oder mit niedrigerer Frequenz übertragen werden. Dies zeigt, dass eine Fehlersuche in einem virtuellen Deployment schnell möglich ist und eine sinnvolle Indikation bezüglich des späteren Systemverhaltens liefert.

4 Verwandte Arbeiten

Die Kopplung von Simulatoren ist ein aktuelles Forschungsthema das bereits in anderen Arbeiten betrachtet wurde. Hierbei wurden in der Regel spezialisierte Lösungen für einen Anwendungsfall entwickelt: Ein konkrete Lösung ist in [Wü97] beschrieben. Sie wurde bereits 1997 publiziert und beschreibt die Kopplung der Simulationswerkzeuge ANSYS und SABER um die Entwicklung mikroelektronischer Schaltungen zu vereinfachen. ANSYS ist ein Werkzeug das lineare und nichtlineare kontinuierliche Systeme simuliert, SABER simuliert elektronische Schaltungen und Systeme. Die publizierte Simulatorkopplung fokussiert auf die Simulation thermischer Eigenschaften. Die Autoren der als [FW07] referenzierten Veröffentlichung fokussieren im Gegensatz dazu auf die funktionale Evaluation ihres Systems. Durch die Kopplung der Simulatoren SLSim und SMASH wurde eine integrierte Simulation von VHDL Code und kontinuierlichen SPICE-Modellen ermöglicht.

Modelisar [Bl11] ist ein Standard, der die Kapselung von Simulationsmodellen in Functional Mockup Units (FMU) definiert. Hierdurch kann Funktionscode zwischen Stakeholdern mit einer einheitlichen Simulationsschnittstelle weitergegeben werden. FERAL unterstützt diese Schnittstelle um Applikationscode in virtuelle Deployments einzubinden. Das Ptolemy II Framework [Ek03] wird genutzt um die Kombination von unterschiedlichen Ausführmodellen zu erforschen. Es teilt Systeme in unterschiedliche, verschachtelbare semantische Domänen auf, die jeweils die direkt enthaltenen Komponenten kontrollieren. Dieses Framework ist ebenfalls die Basis für den in FERAL genutzten Kopplungsansatz, wurde jedoch an einigen Stellen adaptiert (siehe [Ku13]) um die speziellen Erfordernisse einer Simulation zu realisieren.

Bestehende Lösungen zur Systemsimulation basierend auf SystemC Modellen (z.B. Open Virtual Platforms – www.ovpworld.org) nutzen Instruction Set Simulatoren und Hardwaremodelle zur Erstellung von Simulationsmodellen. Diese ermöglichen es fertige Implementierungen auf einer simulierten Zielhardware zu testen. Im Gegensatz zu FERAL erfordert dies eine fertige Implementierung für die Zielplattform und adressiert daher eher Funktionstests als die frühe Validation von Architekturentscheidungen.

5 Zusammenfassung und Ausblick

Zukünftige Architekturen für eingebettete Systeme werden wesentlich komplexer sein als heute genutzte Architekturen. Um eine sinnvolle Entscheidungsgrundlage für Software- und Systemarchitekten zu bieten sind daher auch neue Ansätze zur Entscheidungsunterstützung erforderlich. Mit FERAL ermöglichen wir die frühzeitige Simulationen auf Systemebene. Dabei wird der Systemkontext durch Simulationskomponenten realisiert, eine Anpassung des Funktionscodes ist nicht notwendig. Verhaltenscode und Verhaltensmodelle können über Worker unverändert in Simulationen eingebunden werden. Dies ermöglicht die schnelle Konstruktion von integrierten Simulationsszenarien und erlaubt die effiziente Evaluation einer Vielzahl von Entscheidungsmöglichkeiten basierend auf Fakten.

Das in diesem Artikel vorgestellte Vorgehen wird im Prognostics Center des Fraunhofer IESE angewandt. Die Ansätze sind praxiserprobt und zeigen, dass eine Verteilung und Parallelisierung bestehender Software auch im Bereich sicherheitskritischer eingebetteter Systeme möglich ist. Um den Aufwand hierfür gering zu halten und die zu treffenden Architekturentscheidungen abzusichern verwenden wir am Fraunhofer IESE Analysewerkzeuge und Softwaremodelle, um zum Beispiel die Auswirkungen der Aufteilung von Runnables in Gruppen frühzeitig vorherzusagen. So ist gewährleistet, dass eine gefundene Parallelisierungsarchitektur robust gegen Änderungen ist und konzeptionelle Entscheidungen Bestand haben, zum Beispiel dann wenn zukünftig leistungsfähigere Prozessoren mit noch mehr Kernen genutzt werden sollen.

Literaturverzeichnis

[KF09] Kindel, O., Friedrich, M.: Softwareentwicklung mit AUTOSAR. Grundlagen, Engineering, Management für die Praxis. dpunkt.verlag, 2009, ISBN 978-3-89864-563-8.

[Ek03] Eker, J., Janneck, J., Lee, E., Liu, J., Liu, X., Ludvig, J., Sachs, S., Xiong, Y.: Taming heterogeneity - the Ptolemy approach. Proceedings of the IEEE, 91(1):127-144, January 2003.

[Ku13] Kuhn, T., Forster, T., Braun, T., Gotzhein, R. FERAL – Framework for Simulator Coupling on Requirements and Architecture Level. Proceedings of the 11th ACM-IEEE International Conference on Formal Methods and Models for System Design (MEMOCODE'13), Portland (Oregon), USA, 2013.

[ST97] Shavit, N., Touitou, D. Software transactional memory. Distributed Computing. Volume 10, Number 2. February 1997.

[Wü97] Wünsche, S., Clauss, C., Schwarz, P., Winkler, F. Microsystem Design Using Simulator Coupling. European Design and Test Conference, 1997. ED&TC 97, IEEE (1997)

[FW07] Frank, F., Weigel, R. Co-Simulation of SPICE Netlists and VHDL-AMS Models via a Simulator Interface. International Symposium on Signals, Systems and Electronics, 2007. ISSSE '07. (2007)

[Bl11] Blochwitz, T., Otter, M., Arnold, M., Bausch, C., Clauß, C., Elmqvist, H., Junghanns, A., Mauss, J., Monteiro, M., Neidhold, T., Neumerkel, D., Olsson, H., Peetz, J., Wolf, S. The Functional Mockup Interface for Tool independent Exchange of Simulation Models. Proceedings of the 8th International Modelica Conference. 8th International Modelica Conference, 20.-22 März 2011, Dresden.

Didaktik und Grundlagen der Informatik, Festveranstaltung zum 70. Geburtstag von Volker Claus

Workshop "Didaktik und Grundlagen der Informatik"[1]

Festveranstaltung zum 70. Geburtstag von GI-Fellow Prof. Dr. Dr. h.c. Volker Claus

Universität Stuttgart

Abstract: Der Workshop behandelt Gegenstände und Verbindungen zwischen der Didaktik der Informatik und den theoretischen Grundlagen der Informatik, zwei Bereiche, in denen Volker Claus, der im Juni 2014 seinen 70. Geburtstag feierte, hervorragende Beiträge für die Informatik geleistet hat.

Der Workshop, der im Rahmen der GI-Jahrestagung 2014 stattfindet, ist damit auch seinem Wirken für die deutsche Informatiklandschaft gewidmet.

Ernst-Rüdiger Olderog, langjähriger Kollege von Volker Claus in der Theorie-Gruppe an der Universität Oldenburg, behandelt Entscheidbarkeitsfragen aus dem Wechselfeld von Petri-Netzen und Automaten.

Karsten Weicker, Doktorand und langjähriger Mitarbeiter von Volker Claus an der Universität Stuttgart, beteiligt sich mit einer Arbeit zum Einsatz evolutionärer Algorithmen bei der Konstruktion von belastbaren LEGO-Bauwerken.

Oliver Vornberger, der seinerzeit an der Universität Dortmund Informatik studiert hat und sein Interesse für didaktische Fragestellungen auch auf die damaligen Vorlesungen von Volker Claus zurückfährt, berichtet über seine Erfahrungen bei der Durchführung eines MOOCs (Massive Open Online Course) zum Thema "Algorithmen und Datenstrukturen".

Ralf Romeike, einer der wissenschaftlichen Enkel von Volker Claus, widmet sich der informatischen Bildung von Kindern und Jugendlichen: Wie können sie zur Gestaltung der digitalen Gesellschaft befähigt werden?

[1] http://informatikdidaktik.de/FestKolloqVC

Von Petri-Spielen zu endlichen Automaten

Ernst-Rüdiger Olderog

Universität Oldenburg
olderog@informatik.uni-oldenburg.de

Abstract: Gemeinsam mit Bernd Finkbeiner, Universität des Saarlandes, führen wir Petri-Spiele als einen neuartigen Spiele-Typ mit unvollständiger Information ein. Die einzelnen Spieler sind in System- und Umgebungsspieler unterteilt und werden durch Token im Petri-Netz dargestellt. Die Spieler kennen die Zustände der anderen Spieler nicht, es sei denn, sie kommunizieren miteinander. Dann wird maximal die gesamte kausale Vergangenheit ausgetauscht. Das Ziel der Systemspieler ist es, unabhängig von der Umgebung gewisse unsichere Zustände zu vermeiden.

Wir untersuchen die Frage, ob es entscheidbar ist, dass es eine gewinnende Strategie für die Systemspieler gibt. Für beschränkte Petri-Spiele mit nur einem Umgebungsspieler, aber beliebig vielen Systemspielern ist der Entscheidbarkeitsbeweis gelungen. Er basiert auf Konzepten der Petri-Netz-Theorie. Insbesondere haben wir eine neuartige Definition von maximalem Schnitt in den Entfaltungen von Petri-Netzen gefunden, mit denen sich im Spielablauf Wiederholungen mit äquivalentem Wissenstand der Systemspieler erkennen lassen. Es konnte sogar gezeigt werden, dass die Komplexität des Entscheidungsproblems EXPTIME-vollständig ist. Aus einer mit dem Entscheidungsalgorithmus konstruierten endlichen Gewinnstrategie lässt sich ein verteilter Controller synthetisieren, bestehend aus der parallelen Komposition lokaler endlicher Automaten für jeden einzelnen Systemspieler.

Für beschränkte Petri-Spiele mit mehr als einem Umgebungsspieler ist die Entscheidbarkeitsfrage jedoch weiterhin ungelöst. Für unbeschränkte Petri-Spiele konnte durch eine Reduktion gezeigt werden, dass die Frage nach der Existenz einer gewinnenden Strategie unentscheidbar ist.

Evolution von LEGO-Bauwerken

Karsten Weicker, Martin Waßmann

HTWK Leipzig
karsten.weicker@htwk-leipzig.de

Abstract: Der Einsatz evolutionärer Algorithmen für die Erzeugung von LEGO-Bauwerken, die auf eine vorgegebene Art belastbar sind, geht zurück auf Funes & Pollack. Dieser Ansatz ist aufgrund verschiedener konzeptioneller Entscheidungen in seiner Anwendung eingeschränkt: So können z.B. nur Strukturen mit genau einem Lager betrachtet werden, wodurch zwar Kranausleger, aber keine Brücken erzeugbar sind. Bei der Suche nach einem konzeptionell neuen Ansatz hat sich als Kernproblem die zuverlässige Berechnung herauskristallisiert, wann eine LEGO-Struktur stabil ist. Die bisher eingesetzte Modellierung der Kräfte erwartet die Lösung (oder zumindest heuristische Näherungslösung) des NP-vollständigen Multi-Commodity-Flow-Problems. Durch die sukzessive Betrachtung der Zugkräfte und der Drehmomente als Single-Commodity-Flow-Probleme wird im hier präsentierten Ansatz eine verbesserte Simulation der Stabilität eines Bauwerks ermöglicht. Mittels dieser Modellierung und der Berechnung des maximalen Flusses werden durch einen evolutionären Algorithmus erzeugte LEGO-Strukturen bewertet. Der evolutionäre Algorithmus arbeitet direkt auf LEGO-Bauwerken (als Graph mit speziellen Randbedingungen) - ein neues Konzept, das die bisherige Konstruktion von Bauwerken aus baumartigen Konstruktionsanleitungen ersetzt. Resultierende, zunächst zweidimensional konstruierte Kranausleger übertreffen die Ergebnisse aus der Literatur und sind z.B. 66,4 cm lang und mit 236 g am Ende belastbar bzw. 29,6 cm lang und mit 410 g belastbar. Es werden auch Ergebnisse für die Evolution von Brücken präsentiert, wobei hier insbesondere die veränderte Konfiguration des evolutionären Algorithmus interessant ist. Auch eine Erweiterung zu dreidimensionalen LEGO-Bauwerken wird vorgestellt.

So mooc wi dat - Von der Vorlesung zum Massive Open Online Course

Oliver Vornberger

Universität Osnabrück
oliver@uos.de

Abstract: Zentraler Bestandteil der universitären Lehre bildet seit Jahrhunderten die Vorlesung: Einer redet, Hunderte hören zu. Ermöglicht durch schnelle Netze und multimediale Webportale bahnt sich seit etwa drei Jahren eine ernstzunehmende Alternative den Weg: der MOOC (Massive Open Online Course). Hier wühlen sich Tausende von Teilnehmern nach eigener Zeiteinteilung durch interaktives Lehrmaterial, stellen Fragen in Diskussionsforen und benoten sich gegenseitig im Peer Reviewing. Die Teilnahme ist typischerweise kostenlos, nur für die Prüfung fällt eine moderate Gebühr an.

Der Vortrag berichtet von unserem MOOC "Algorithmen und Datenstrukturen", der im WS 2013/14 an der Universität Osnabrück produziert und im SS 2014 auf der Internetplattform Iversity ausgestrahlt wurde. Er führt in die Grundlagen des Programmierens anhand der Programmiersprache Java ein. Der MOOC besteht aus 237 Videosequenzen von jeweils 2 bis 3 Minuten Dauer, 132 Multiple Choice Quizzes und 13 umfangreichen Hausaufgaben. Von den 6909 eingeschriebenen Teilnehmern schauten sich 2080 das erste Video an, 508 das letzte. Zur Abschlussklausur im August reisten Z Studierende nach Osnabrück, Hamburg und Berlin; U davon erwarben ein Abschlusszertifikat über 6 ECTS Credits.

Ob sich MOOCs durchsetzen werden, ist ungewiss; in jedem Falle wird sowohl vom Lehrenden als auch vom Lernenden überdurchschnittliches Engagement erwartet.

Hinweis: Z und U waren bei Redaktionsschluss noch nicht bekannt.

Wie Kinder und Jugendliche lernen, die digitale Gesellschaft mitzugestalten

Ralf Romeike

Universität Erlangen
ralf.romeike@fau.de

Abstract: Die "digitale Gesellschaft" ist davon geprägt, dass (nicht nur) Kinder und Jugendliche online Spiele spielen, surfen, chatten und liken. Aber nur wenige sind dazu in der Lage, selbst Websites, Spiele und Apps zu erstellen und damit das kreative Potenzial digitaler Medien für sich zu nutzen. Der Vortrag zeigt Perspektiven auf, wie informatische Bildung zu einem gestalterischen Umgang mit Informatiksystemen befähigt und somit dazu beiträgt, dass Schülerinnen und Schüler die digitale Gesellschaft verstehen und sich selbstbestimmt an ihrer Entwicklung beteiligen können.

Doktoranden-Programm

Doktorandenworkshop 2014

Markus Krötzsch[1], Dirk Pflüger[2], Ansgar Scherp[3]

[1]TU Dresden, Fakultät für Informatik
markus.kroetzsch@tu-dresden.de
[2]Universität Stuttgart, Institut für Parallele und Verteilte Systeme
Dirk.Pflueger@ipvs.uni-stuttgart.de
[3]Christian-Albrechts-Universität zu Kiel, Technische Fakultät
asc@informatik.uni-kiel.de

Das Doktorandenprogramm der INFORMATIK 2014, das zusammen mit dem Doktorandenprogramm der KI 2014 stattfindet, erlaubt jungen Forscherinnen und Forschern ihr Thema in einem Vortrag oder als Poster einem breiten Publikum vorzustellen.

Wir freuen uns über das rege Interesse und alle Einreichungen. Die gesammelten Beiträge, fünf Full Paper und neun Short Paper der Proceedings stellen einen sehr schönen Streifzug durch die Breite der Informatik dar und behandeln Themen von der Datensicherheit über Mobilfunknetze bis hin zum Reverse Engineering. Sie sind alle in Alleinautorenschaft der Doktorandinnen und Doktoranden entstanden.

Wir danken allen Gutachtern und allen Autoren für Ihren Beitrag zum Gelingen des Doktorandenworkshops

Markus Krötzsch, Dirk Pflüger, Ansgar Scherp

Programmkomitee

- Klaus-Dieter Althoff (Universität Hildesheim)
- Collin Atkinson (Universität Mannheim)
- Wolf-Tilo Balke (Universität Braunschweig)
- Stefan Conrad (Universität Düsseldorf)
- Andreas Dengel (Technische Universität Kaiserslautern)
- Rolf Drechsler (Universität Bremen)
- Wolfgang Effelsberg (Universität Mannheim)
- Michael Gertz (Universität Heidelberg)
- Wilhelm Hasselbring (Universität zu Kiel)
- Andreas Henrich (Universität Bamberg)
- Melanie Herschel (Universite Paris Sud)
- Harald Kosch (Universität Passau)
- Antonio Krüger (Saarland University)
- Michael Langhammer (Universität Karlsruhe)

- Miriam Mehl (Universität Stuttgart)
- Bernd Neumann (Universität Hamburg)
- Thomas Neumann (TU München)
- Dirk Nowotka (Universität zu Kiel)
- Isabella Peters (Universität Kiel)
- Thomas Plagemann (Universität Oslo)
- Ralf Reussner (Universität Karlsruhe)
- Raúl Rojas (FU Berlin)
- Sebastian Rudolph (TU Dresden)
- Ulrike Sattler (University of Manchester)
- Lars Schmidt-Thieme (Universität Hildesheim)
- Ingo Schmitt (Universität Cottbus)
- Sven Schwarz (Technische Universität Kaiserslautern)
- Philipp Slusallek (Universität des Saarlandes)
- Mathias Soeken (Universität Bremen)
- Viorica Sofronie-Stokkermans (Universität Koblenz-Landau)
- Frank Steinmann (Technische Universität Kaiserslautern)
- Heiner Stuckenschmidt (Universität Mannheim)
- Volker Turau (TU Hamburg-Harburg)
- Stefan Wagner (Universität Stuttgart)
- Andreas Winter (Universität Oldenburg)

Delta Operation Language for Model Difference Representation

Dilshodbek Kuryazov
Software Engineering Group
University of Oldenburg, Germany
kuryazov@se.uni-oldenburg.de

Abstract: Software models evolve over time undergoing various changes and resulting in several versions during their lifetime. Models are differentiated during the evolution process and the differences between subsequent model versions are represented in differences documents for further analysis and manipulations as history information. Software models have rich data structures which differ from the code of software systems. That is why, a representation approach for the model differences has to provide effective handling and management of difference information. Furthermore, the model differences represented in the differences documents have to be easy to access and reuse.

This paper introduces the *Delta Operations Language (DOL)*, a meta-model independent and operation-based approach to model difference representations. The approach represents the model differences in terms of DOL and provides several *DOL Services* to access and reuse the DOL-based differences for further analysis and manipulations. To explain ideas behind the approach, it is applied to UML activity diagrams as a running example.

1 Motivation

Software models are designed using software modeling languages (cf. Unified Modeling Language (UML) [UML]). Models are widely used as abstractions of systems structural and behavioural artefacts. Abstraction of any software system is quite useful and practical to understand and trace system aspects from different viewpoints.

Software models follow varying concepts than the code of a software because of the paradigm shift between code and design levels. Software models have rich data structures with different syntax and semantics. Though, like the code of software projects, software models evolve over time undergoing various changes such extensions, corrections and improvements. These changes are applied to software models by a team of modellers using *Collaborative Modeling* and *Model Version Control* tools.

Several version control systems exist for code-based software systems (e.g. Subversion [CFP04], Git [Loe09] etc.). Software models can also be represented in the XMI (XML Metadata Interchange) serialization [Obj]. But, it is commonly agreed that code-centric version control systems can not completely handle the differences of software models [Cic08], [EMF].

Subsequently applying changes to software models results in several versions of the same model artefact. Since a model has several versions during its life time, the model

histories has to be handled and the differences between subsequent model versions have to be represented in appropriate ways.

Representing the model differences allows to store the histories of software models. Therefore, finding an appropriate approach for model difference representation is the best aid to build up various applications and tools such as model versioning, model history analysis and collaborative modeling. Considering aforementioned challenges, this thesis addresses the problem of model difference representation.

This paper introduces the general ***Delta Operations Language (DOL)***, meta-model generic and operation-based approach for model difference representations. Conceptually, DOL is a set of domain specific languages for model difference representation in terms of *operations*. A specific DOL for a specific modeling language is derived from the *meta-model* of a modeling language. A specific DOL is fully capable of representing the differences of models conforming the given meta-model in terms of DOL operations. Only changed elements between model versions are calculated and represented in difference documents referred to as *Modeling Deltas*. The operations in a modeling delta are called *Delta operations*.

Additionally, the DOL approach aims at providing several ***DOL Services*** which can access and reuse delta operations. These operative DOL services improve applicability and re-usability of the DOL-based modeling deltas in different application areas. The operation-based DOL approach addresses several use cases in this PhD proposal. These are *Generic Model Versioning System (GMoVerS)*, *Model History Analysis* and *Collaborative Modeling* that are discussed in Section 5.

The paper is structured as follows: The related approaches are discussed in Section 2. Section 3 explains a simplified difference representation example of the DOL approach. In Section 4, the main ideas behind the thesis are portrayed. Section 5 presents application areas of the approach. Current and ongoing works are expressed in Section 6. The paper ends up by defining expected contributions of the thesis in Section 7.

2 Related Work

There are several approaches to model difference representations providing some additional services. This section discusses some of related difference representation approaches which provide a support to deal with some aspects and principles of model difference representation.

An operation-based difference representation is introduced by Alanen and Porres [AP03] which is also meta-model independent. It detects differences and union of models and represents the differences as a sequence of difference operations. The approach supplies conflict resolver service that is extended by the combination of the difference operations. Another operation-based approach is *EMF Store* framework [HK13] introduced by Helming and Koegel. EMF Store is a data model repository for EMF (Eclipse Modeling Framework) models [EMF]. The framework supports collaborative modeling services such as semantic versioning, branching and conflict detection services. The EMF Store platform is extended by Krusche and Bruegge with Model-based Real-time Synchronization [KB14]. The EMF Store provides change tracking feature for atomic changes.

A meta-model independent and model-based difference representation is introduced by Cicchetti et al. [Cic08]. The approach uses a *difference model* for representing the dif-

ferences. The differences model conforms to the differences meta-model derived from the base meta-model by *automatic transformations*. The approach provides a *difference application* service which requires *model matching* between the differences and base models. Cicchetti et. al. has *conflict resolution* support for differences merging. Another model-based difference representation approach, EMF compare and merge [EMF] was introduced for differencing EMF models.

SMOVER (Semantically enhanced Model Version Control System) [ABSK07] uses *standard SQL database approach* to store the model differences and provides several standard versioning services such as *add*, *checkout*, *commit* and *update*. SMOVER mostly addresses flexible difference merging technique and requires to have an adapter which lies between external modeling tools and SMOVER.

DeltaEcore – A Model-Based Delta Language Generation Framework [SSA14] addresses engineering delta modeling languages for software product lines. The approach aims at automatically derive delta operations for software architectures of software product lines.

In contrast to the existing approaches, a number of additional contributions of the approach in this paper can be indicated. Deficiency of additional services to reuse and exploit representation information diminishes the applicability of an approach in a wide range of applications. Thus, the DOL-based difference representation approach aims at providing a number of the DOL-services which can access and reuse difference representation information. Meanwhile, the existing approaches provide only few services for exploitation and manipulation of model difference information. The DOL-based difference representation uses simple text-based notations instead of complex models to storage history by means of delta operations i.e. text-based difference representation for the graphical software models. Moreover, DOL provides the specific orchestration of the DOL-services extending application areas.

3 Example

This section describes a simple example of DOL-based model difference representation. To explain the idea of the DOL approach, a simplified UML activity diagram [UML] is employed in this section. The meta-model of a modeling language is required to represent the model differences in terms of DOL. The DOL operations are derived from the simplified meta-model of UML activity diagram depicted in Figure 1. A specific DOL is derived for this specific meta-model in this paper, but the DOL Generator is completely independent from meta-models.

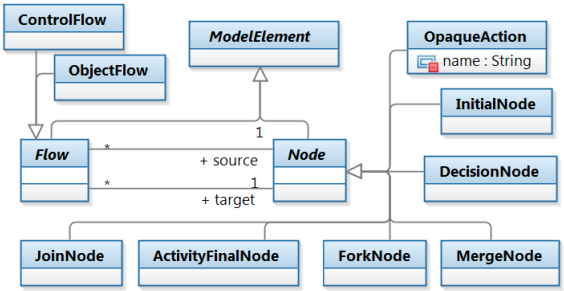


Figure 1: Simplified UML Activity Diagram meta-model

All meta-classes are of type *Node* or *Flow*. Each *Flow* has the target and source attributes and *Actions* have the *name* attribute whereas the other classes have no attributes.

Figure 2 portrays three subsequent versions of the same UML activity diagram performing an *Ordering System* example. All model versions conform to the same simplified meta-model shown in Figure 1.

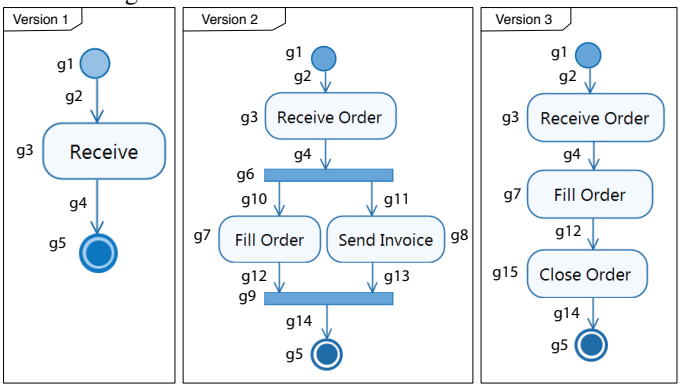


Figure 2: Example activity diagram in three concurrent versions

Each model element is assigned to a *persistent identifier* (g_x). The first version has an *Action* named *Receive*, an *Initial* and a *Final* node and *(Control) Flows* connecting these nodes. In the second version, the model has *Fork, Join nodes*, two *Actions* named *Fill Order* and *Send Invoice*. The target end of *Control Flow* `g4` is reconnected to the fork node, the name of the *Receive* action `g3` is changed and several control flows are created connecting these nodes. Finally, the model reaches into the third version. A new action with name *Close Order* is created, the target ends of `g4`, `g12`, and the source end of `g14` are reconnected. *Fork* and *Join*-nodes, the *Send Invoice* action, and the control flows `g10`, `g11` and `g13` are deleted.

In order to derive the specific DOL for UML activity diagrams, the meta-model in Figure 1 is used and three atomic operations `create`, `delete` and `change` are applied to the concepts of the meta-model (generation of a specific DOL is explained in Section 4.1). A delta-operation creates or deletes a model element or changes its attributes. The DOL approach considers that three basic operations are sufficient for deriving the complete set of the DOL operations for difference representations and the specific DOL is capable of representing all differences by these three operations. Other change operations like *moving* a group of elements from one place to another in a model can be achieved by changing one (or several) association(s) between a part that should be moved and the rest of a model.

Like classical version management systems for source code (cf. Subversion [CFP04], Git [Loe09]), the DOL approach follows a *backward delta* approach, where the most recent version of a model and several modeling deltas are stored directly. The last version (the third version in this example) is also represented by the DOL-operations which consists only of the creation operations.

```
1 g1=createInitialNode();
2 g3=createOpaqueAction("Receive Order");
3 g7=createOpaqueAction("Fill Order");
4 g15=createOpaqueAction("Close Order");
5 g5=createActivityFinalNode();
6 g2=createControlFlow(g1,g3);
7 g4=createControlFlow(g3,g7);
8 g12=createControlFlow(g7,g15);
9 g14=createControlFlow(g15,g5);
```

Figure 3: Active delta

Figure 3 depicts the modeling delta named *active delta* which gives directly the last model version. Each delta operation has a ***Do part*** (cf. *g1=createInitialNode();*) which describes the *kind of change* by means of *operations* (one of create, change, delete) and an ***Object part*** (with attributes if required) (cf. *g1=createInitialNode();*) which refers to the modeling concept. To refer to model elements from the delta operations, *unique persistent identifiers* are used as *references*.

Likewise, the differences deltas are directed in reverse order i.e. each change leads from the later version to the previous. Producing delta operations in reverse order is quite practical for implementation of applications.

```
1  g6 = createForkNode();
2  g7 = createOpaqueAction("Send Invoice");
3  g9 = createJoinNode();
4  g4.changeTarget(g6);
5  g12.changeTarget(g9);
6  g14.changeSource(g9);
7  g10 = createControlFlow(g6,g7);
8  g11 = createControlFlow(g6,g8);
9  g13 = createControlFlow(g8,g9);
10 g15.delete();
```

Figure 4: Delta between active and the second versions

```
1  g3.changeName("Receive");
2  g4.changeTarget(g5);
3  g6.delete();
4  g7.delete();
5  g8.delete();
6  g9.delete();
7  g10.delete();
8  g11.delete();
9  g12.delete();
10 g13.delete();
11 g14.delete();
```

Figure 5: Delta between the second and first versions

The differences delta in Figure 4 depicts the differences between the third and second versions. Finally, Figure 5 illustrates the differences delta between the second and first versions.

The modeling deltas in these figures are executable descriptions of the differences. Each of the difference deltas reverts the model to the previous versions. The modeling delta in Figure 4 reverts the model to the second version from the third and the modeling delta in Figure 5 reverts the model to the first version from the second.

4 Approach

This section discusses a general architecture of the proposed approach as depicted in Figure 6. It has three main levels such as ***DOL Generation*** (discussed in Section 4.1), ***DOL Services*** (Section 4.2) and ***DOL Applications*** (Section 5).

DOL Generation defines generating a specific DOL by importing the meta-model of a modeling language. The resulting DOL is produced in form of Java Interface. The DOL approach also provides several services that lie at the *DOL Services* level. These services serve to manage, manipulate and reuse the DOL-based modeling deltas stored in the repository. The third level depicts *DOL Applications* developed by the specific orchestrations of the DOL-services.

Each DOL-service has a particular task and is involved in constructing the specific orchestrations in the framework of the DOL applications. For instance, models are designed in external modeling tools and parsed into the internal representations by the *Adapter*. The difference *calculator* is used to detect the differences between the compared models and produce the differences and active deltas by implementing the DOL interface. The difference calculator uses an *Optimizer* to produce the optimized modeling deltas and to write them into the repository. After all, other DOL-services such as a *Patcher* and a Change

Tracer utilize the DOL operations. The patcher reverts older model versions. The change tracer tracks a specific model element and reports the change history information of that element. Then, these change reports can be browsed to see and analyse the change history.

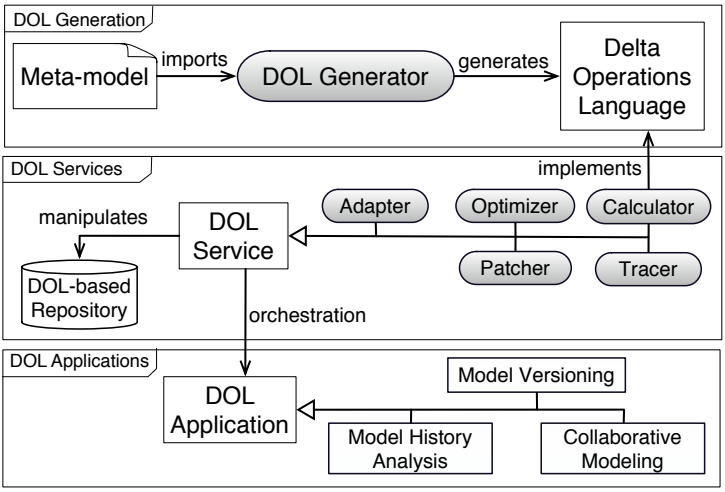


Figure 6: Overall architecture of the approach

These DOL-services are employed in developing the DOL applications such as *Model Versioning*, *Collaborative Modeling* and *Model History Analysis* (discussed in Section 5).

4.1 DOL Generation

The Delta Operations Language (DOL) is a family of the operation-based languages for model difference representations. This section shows how a specific DOL is generated for a specific modeling language. As mentioned above, the meta-model of a modeling language is required to derive a specific DOL. The DOL Generator imports the meta-model depicted in Figure 1 as the example and applies three basic operations to each concept of that meta-model.

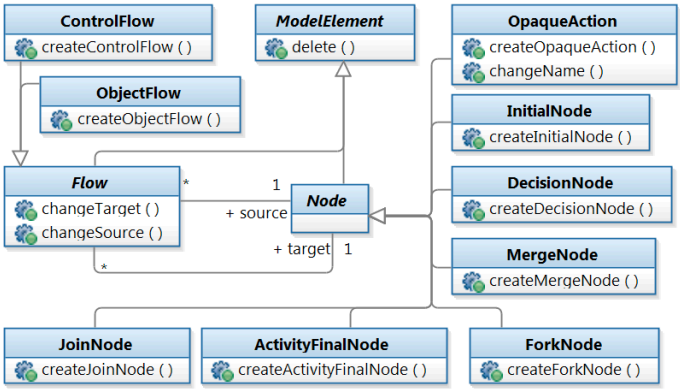


Figure 7: Meta-model of the specific DOL operations

The specific meta-model in Figure 7 depicts an abstraction of the specific DOL operations shown in Figure 8.

Each model element can be deleted and created with relevant parameters if they have attributes. Only attributes can be changed and all associations are associated with attributes. The specific DOL is generated as an Java Interface. The *methods* of the resulting interface are parametrized with the meta-model concepts including one of the *create*, *delete* and *change* operations (Figure 8). Implementations of these methods result in an analogous operation with relevant parameters.

```
1  //------ Creations ------
2  InitialNode createInitialNode();
3  OpaqueAction createOpaqueAction(String name);
4  ForkNode createForkNode();
5  JoinNode createJoinNode();
6  DecisionNode createDecisionNode();
7  MergeNode createMergeNode();
8  ActivityFinalNode createActivityFinalNode();
9  ControlFlow createControlFlow(Node source, Node Target);
10 ObjectFlow createObjectFlow(Node source, Node Target);
11 //------ Changes ------
12 void changeName(String newName);
13 void changeSource(Node newSource);
14 void changeTarget(Node newTarget);
15 //------ Deletion ------
16 void delete();
```

Figure 8: Interface generated from UML Activity Diagram meta-model

The DOL Generator is completely independent from modeling languages. Eventually, each interface method has the same structure: a *Do part* and an *Object part*. The delta operations are produced by implementations of that interface by assigning persistent identifiers.

4.2 DOL Services

In order to reuse and exploit the DOL-based modeling deltas, the difference representation approach introduces a set of reasonable *DOL-services* as depicted in Figure 6. The DOL-services consist of an *Adapter*, a difference *Calculator*, a delta *Optimizer*, a *Patcher* and a change *Tracer* which are discussed in detail, below.

Adapter. Software models are designed in model designing tools. Integrating version management or collaborative modeling tools with model designing tools is a challenge. But, these tools provide export/import feature for software models by XML Metadata Interchange (XMI) [Obj] format without layout information. Therefore, the *Adapter* is developed to exchange models between external modeling tools and the DOL applications.

In the DOL applications, models are represented as TGraphs internally [ERW08] to make them generally processable. The *Adapter* can parse models in the XMI serializations to TGraphs and vice versa.

Calculator. The difference calculator detects the differences between the differentiated model versions using state-based comparison and produces modeling deltas in terms of the DOL operations. Several approaches already exist for difference calculation. Therefore, Küpker [Kü13] investigated existing approaches such as UMLDiff [XS05] and SiDiff [SG08] [TBWK07] and combined them to *gDiff* (generic differentiating).

The difference calculator calculates the model similarities using the similarity metrics such as a name [XS05] and a structural similarity [TBWK07]. Created, deleted or changed elements are detected for each candidate element with the highest similarity.

The calculator produces two modeling deltas in the operation-based format by implementing a specific DOL Interface: the *active* and the *differences delta*. If the first version of a model is given, the calculator produces only the active delta without the differences delta. The persistent identifiers are assigned to delta operations by the difference calculator while matching model elements.

Optimizer. After the difference calculator produces the modeling deltas, they are optimized to improve their efficiency. For instance, if a particular model element is created and later deleted in the same delta, these two operations (create and delete) are registered for one element where both have no affect. Another example might be changing one element several times in one modeling delta. In this case, it is optimal to save only the last change instead of several change operations for that model element.

Moreover, the order of delta operations in modeling deltas is also important to avoid the lost and fuzziness of change information. The creation operations are placed on the first place in a modeling delta, changes come second and deletions are dropped to the end.

Patcher. The *patcher* reverts a model to earlier versions by *applying* (sequences of) modeling deltas to the last version. Modellers may need to revert a model to an older versions in the case of lost or damage of data. The inputs for the patcher are a *model* and several *modeling deltas*. An input delta is applied to an input model resulting in previous version of a model.

Since the delta operations are the executable descriptions, they are implemented by *in-place* model transformations [Kah06].

Tracer. To comprehend and analyse the histories of evolving models, the DOL approach provides a Change *Tracer* service which contributes to detect necessary information about each change history.

The change tracer receives the set of modeling deltas from the repository as input and searches for change information about a considered model element based on its persistent identifier. Detected change information is used by history analysis application (Section 5).

For instance, Figure 9 illustrates all history information of the control flow `g4` in the example in Section 3. The traced element was referred to in three versions. Information about these states is tracked by slicing three modeling deltas, the *active delta* in Figure 3 and two differences deltas in Figures 4 and 5.

```
1  g4 = createControlFlow(g3,g7);
2  g4.changeTarget(g6);
3  g4.changeTarget(g5);
```

Figure 9: History information of Control Flow `g4`.

5 DOL Applications

As a proof of concept, the DOL-approach is being prototypically implemented. To demonstrate its applicability, it is planned to use the approach in the context of three use cases namely *Modeling Versioning*, *Model History Analysis* and *Collaborative Modeling*. The architectures of these applications are built by the specific orchestrations of the DOL-

services.

Model Versioning. Model versioning aims at managing and manipulating models as well as storing and reusing the model differences. Model versioning systems are essential on handling models and their histories. To this end, the DOL difference representation approach is applied to develop ***G**eneric **M**odel **Ver**sioning **S**ystem (GMoVerS)*. GMoVerS uses the DOL operations to represent the model differences.

Current implementations of GMoVerS support *adding* models to the versioning system, *committing* changes and *reverting* older versions. In each of these activities, the relevant DOL-services are involved in a certain order based on data-flow. Figure 10 displays a screen-shot of the GMoVerS development environment.

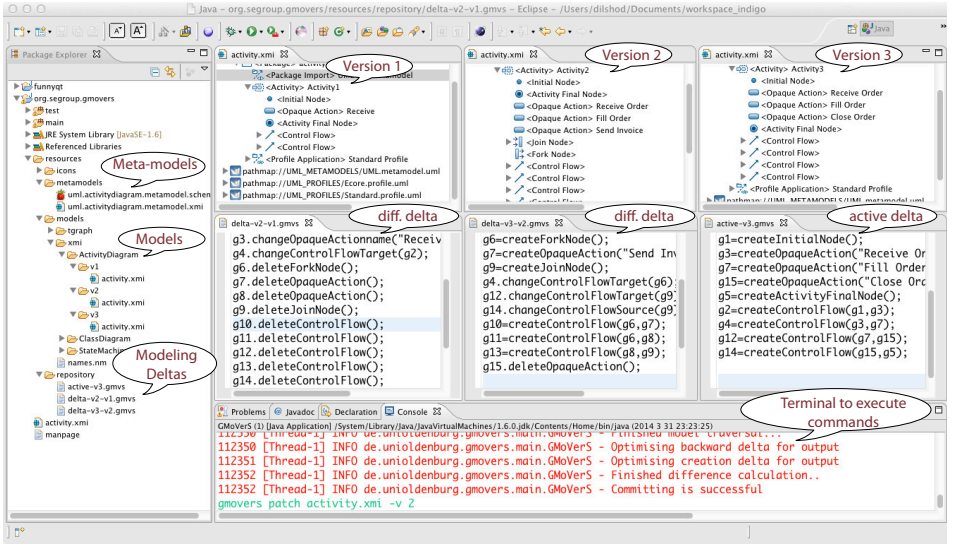


Figure 10: A screen-shot of GMoVerS

First of all, a model has to be ***added*** under version control to start model versioning. The *Adapter* is required to parse a model in the exchange format to internal representation, then the *calculator* is involved to produce the *active delta* for the new model. Similarly, ***committing*** changes requires the *adapter* to parse new versions to internal representation, the *patcher* to revert the previous model version, again the *calculator* to calculate the differences between the previous and committed versions and produce the differences and active deltas. ***Reverting*** needs the *patcher* firstly to revert the recent model version, then it is again called to revert requested version.

The explorer on the left side of Figure 10 displays the workspace including the metamodel in Figure 1, the models and modeling deltas in Section 3. Here, all DOL-based modeling deltas belong to the GMoVerS repository. On the upper row of the right side, Figure 10 displays three subsequent versions of the example model in Figure 2 in exchange formats. The central part of the right shows two differences deltas and the active delta. Finally, the most bottom of the screen-shot is a terminal to type and execute aforementioned version control commands such as add, commit and patch.

Model History Analysis. Analyzing the model histories is crucial to understand what

changes are made to models and to know how a model evolves during its life-time. History analysis is built on the top of the change history *tracer*. The change tracer fetches a set of modeling deltas from the repository and detects history information from these deltas based on persistent identifiers. After detecting necessary history information, it is browsed as depicted in Figure 11.

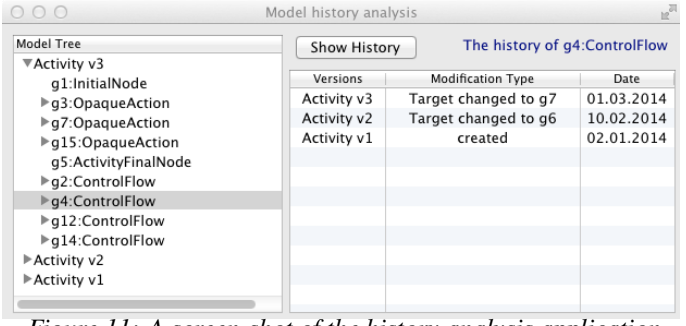


Figure 11: A screen-shot of the history analysis application

The screen-shot in Figure 11 displays the example model in Section 3. The change tracer firstly builds a tree for each model version running throughout all modeling deltas and outlines them in *Model Tree* as shown on the left side of Figure 11 including all model elements. If a specific element is selected and show history button is clicked, the tracer detects history information of the selected model element based on its persisted identifier and shows it in *Tabular View*. For example, the table on the right side shows the history of the *g4:Control Flow* in Figure 9.

Collaborative Modeling. Another prominent application of DOL is collaborative modeling which facilitates a teamwork of several designers on a shared model repository at runtime.

The implementation of the application is planned to rely on representing changes in terms of the DOL operations and exchanging these changes by the appropriate modeling deltas. To this end, it is planned to develop extra DOL-services named *runtime operation recorder* to detect operations embodying changes. The operation recorder will be integrated into end-user model designing tools and it detects changes made by modellers. While manipulating a model, the DOL-based change operations are recorded in modeling deltas. An evolving model has several development branches and *synchronization* is required among various development branches. The synchronization of changes basically is synchronization of the DOL-based modeling deltas among various branches.

6 Current and Ongoing Work

Extensive literature research has been done to become familiar with the existing difference representation techniques and additional services related to this thesis. A first sketch of the DOL approach was published in 2012 [KJW12].

The DOL-based difference representation approach is applied to versioning Sustainability Reports at companies in [KSW13]. As sustainability reports at companies are also subject to constant changes and evolution, companies intend to report and analyse sustainability information to provide the sustainable future. Thus, reports have to be versioned to

analyse the histories of sustainability reports.

The DOL generator and the DOL services such as the *adapter*, the difference *calculator*, the delta *optimizer*, the delta *patcher* and the change *tracer* (depicted in Figure 6) are implemented so far. *GMoVerS* and *Model History Analysis* applications are elaborated by the specific orchestrations of these DOL-services.

Building the specific orchestration of *Collaborative Modeling* is remaining as ongoing work. Further services are requested to build up the collaborative modeling application. Therefore, the DOL-services will be extended with *a runtime operation recorder* and a *synchronizer* for collaborative modeling. A runtime operation recorder facilitates the end-users with runtime operation registration. These modeling deltas will be synchronized by a *DOL-synchronizer*.

Finalizing extension and evaluation of the model versioning and history analysis applications on the large scaled software models is expected till August 2014. Implementation of further DOL-services and writing actual thesis is planned within the year 2014. Proof read and submission of the thesis is scheduled for the beginning of 2015.

7 Expected Contributions

This paper proposed the PhD thesis focusing on development of an approach towards model difference representation. The DOL approach is designed using existing software engineering technologies. The operation-based difference representation approach facilitates several additional DOL services to employ and reuse the DOL-based modeling deltas. The approach satisfies several reasonable principles and requirements: (1) *Meta-model Generic* – applicable to various modeling languages with respect to their meta-models, (2) *Tool Independent* – has own *Adapter*, (3) *Operation-based* – embodies all necessary information about a change using a meaningful syntax and completely follows compound modeling concepts, (4) *Delta-based* – only changed model objects are referred to in modeling deltas, (5) *Complete* – carries precise information about each change, (6) *Reusable* – provides several DOL-services which can directly access and manage the DOL-based modeling deltas making modeling deltas straightforward and accessible.

With these principles, the DOL approach has the following contributions for its applications:

GMoVerS. The model differences are represented in terms of DOL and stored in small modeling deltas. Only changes are referred to in modeling deltas. The delta operations are directly executable descriptions of the model differences allowing model manipulations when needed.

Model history analysis. The delta operations assigned to persistent identifiers allow the change tracer to rapidly detect all necessary information about the change histories by verifying small modeling deltas.

Collaborative modeling. Synchronization of changes will be eased by exchanging small DOL-based deltas. The various model designing tools which are running on different platforms can communicate with each other in terms of DOL.

References

[ABSK07] K. Altmanninger, A. Bergmayr, W. Schwinger, and G. Kotsis. Semantically Enhanced Conflict Detection between Model Versions in SMoVer by example. In *Proceedings of the Int. Workshop on Semantic-Based Software Development at OOPSLA*, 2007.

[AP03] M. Alanen and I. Porres. Difference and union of models. In *UML 2003. LNCS*, pages 2–17. Springer, 2003.

[CFP04] B. Collins-Sussman, B. W. Fitzpatrick, and C. M. Pilato. *Version Control with Subversion.* O'Reilly Media, 2004.

[Cic08] A. Cicchetti. *Difference Representation and Conflict*. PhD thesis, University of L'Aquila, (Italy), April 2008.

[EMF] EMF: Eclipse Modeling Framework (Compare). http://www.eclipse.org/emf/compare.

[ERW08] J. Ebert, V. Riediger, and A. Winter. Graph Technology in Reverse Engineering, The TGraph Approach. In *10th Workshop Software Reengineering (WSR 2008)*, volume 126, pages 67–81. GI (LNI), 2008.

[HK13] J. Helming and M. Koegel. EMFStore., 2013. http://eclipse.org/emfstore.

[Kah06] S. Kahle. JGraLab: Konzeption. *Entwurf und Implementierung einer Java-Klassenbibliothek für TGraphen*, 2006.

[KB14] S. Krusche and B. Bruegge. Model-based Real-time Synchronization. In *Inter. Workshop on Comparison and Versioning of Software Models (CVSM'14)*, Feb. 2014.

[KJW12] D. Kuryazov, J. Jelschen, and A. Winter. Describing Modeling Delta By Model Transformation. In *Softwaretechnik Trends (Issue on International Workshop on Comparison and Versioning of Software Models (CVSM 2012))*, no. Band 32 Heft 4. Gesellschaft für Informatik, November 2012.

[KSW13] D. Kuryazov, A. Solsbach, and A. Winter. Versioning Sustainability Reports. In *5.BUIS-Tage: IT-gestütztes Ressourcen- und Energiemanagement*, pages 409–419. Springer-Verlag, 2013.

[Kü13] C. Küpker. General Model Difference Calculation. Bachelor Thesis, Carl von Ossietzky University of Oldenburg, June 2013.

[Loe09] J. Loeliger. *Version Control with Git: Powerful Tools and Techniques for Collaborative Software Development*. O'Reilly Media, 2009.

[Obj] Object Management Group. XMI Specification, v1. 2.

[SG08] M. Schmidt and T. Gloetzner. Constructing Difference Tools for Models Using the SiDiff Framework. *ICSE 2008*, pages 947–948, May 10-18 2008.

[SSA14] C. Seidl, I. Schaefer, and U. Aßmann. DeltaEcore - A Model-Based Delta Language Generation Framework. In *Modellierung*, pages 81–96, 2014.

[TBWK07] C. Treude, S. Berlik, S. Wenzel, and U. Kelter. Difference Computation of Large Models. In *Proceedings of the the 6th Joint Meeting of the European Software Engineering Conference*, pages 295–304. ACM Press, 2007.

[UML] UML: Unified Modeling Language. http://www.uml.org.

[XS05] Z. Xing and E. Stroulia. *UMLDiff: An Algorithm for Object-Oriented Design Differencing.*, pages 54–65. 6. ACM, 2005.

Towards an Automated Reverse Engineering of Design Models from Trace Recordings

Andreas Sailer
Laboratory of Safe and Secure Systems (LaS^3)
Ostbayerische Technische Hochschule (OTH) Regensburg
Seybothstraße 2, 93053 Regensburg, Germany
andreas.sailer@oth-regensburg.de

Abstract: Recording traces from an embedded system is a common procedure in the process of developing real-time systems. While this technique is generally used for debugging and performance analysis, the recordings contain much more information. These pieces of information however can be useful to generate or maintain a system model in an automated manner. Therefore, the goal is to develop an approach that creates a model of a real-time system, especially of its observed temporal behaviour, from events logged in a trace recording. This research aims to extract, analyse and deduce information about the system under observation from a limited view of the proceedings in the system. Also as a consequence of this intention a detailed specification on events necessary for specific use cases emerges. To achieve all this, an iterative approach consisting of three consecutive steps is proposed. At first the necessary algorithms will be developed on a model-based foundation utilising a discrete event simulation. In a next step, the implementation will be aligned to take also specifics of the execution on real hardware into account. Finally, the approach can be evaluated on hardware with examples from real life.

1 Motivation

Contrary to popular belief, the development of software for embedded and non-embedded systems are fundamentally different [GLT03]. Processes and technologies have to address additional requirements such as timing constraints. This means the correctness of such a system depends not only on the functional correctness but also on its temporal accuracy. Consequently, traditional debuggers cannot be used to observe the runtime behaviour of software since they require halting a system's execution, which would result in a significant alteration of the timing behaviour. Instead, tracing is used during the development of embedded software systems. 'Concretely, trace recording implies detection and storage of relevant events during runtime, for later off-line analysis' [KWK10, p.1].
Nowadays, execution traces are used in most parts of the industry for dynamic timing analysis [WEE$^+$08]. From the recorded events in a trace, the event-to-event execution time is calculated. That way the minimal and maximal observed execution times are determined, which help to estimate the actual execution time bounds. Metrics that are calculated this way are for example the Execution Time (ET) of functions [SSD$^+$13], and the Response Time (RT) [QHE12] or the mean Normalised Blocking Time (mNBT) [AMS$^+$13] of tasks.

However, much more information can be extracted from execution traces. This is due to the fact that events from different abstraction levels of a system can be recorded [Hus07]. On the one hand a trace can contain events from the system level, where system events such as task switches or the creation and termination of tasks can be registered as well as calls to functions. On the other hand also proceedings on a process level can be monitored. Those are for example events because of accesses to resources such as data signals or semaphore queues. Therefore, the main goal of this research is to define an approach how a temporally accurate design model of a real-time system can be automatically created based on trace recordings.
Such a generated model has many advantages over a manually maintained one and can be used for multiple purposes. In the simplest way it can tremendously help to understand the behaviour of a software system in general [HLL05]. Although model-driven software development (MDSD) is standard in the embedded industry, it is also common to modify the generated code afterwards by hand. Thus, an approach, which can easily keep a model up-to-date, would be a huge benefit not only for simple documentation purposes during development but also for the verification of software models [HA05] or the maintenance of legacy systems [TVD12]. But the currently probably most important reason for such a generated model is simulation-based timing analysis [KKM11]. With the tendency towards a commitment to multi-core technology, software that has run on a single core for years has now to be allocated on different cores and optimised regarding certain metrics. Thus, the use of a discrete event simulation helps to handle the resulting complexity by imitating the key characteristics of a system on the basis of a model. That way it is possible to estimate the effect of design decisions without the need for physical access to the system.
Although trace recordings can give a valuable in-depth look into a system's execution behaviour, collecting all this data is expensive. In the area of embedded systems, where only very limited resources are available, this might result in an unintended alteration of the system behaviour [Gai86] or an unmanageably large growth in data traffic [RS13]. Due to the fact that not every event contributes to the model extraction in the same way, an additional motivation is to develop a detailed specification on all the events necessary for specific use cases. That way a work flow is targeted, which can make a precise statement about the needed quality of a trace recording in order to achieve specific goals.
The remainder of this paper is organised in the following way: In the next section, we discuss the presented problem in more detail. After that, an overview of related research and state-of-the-art is given. In Section 4, we present the approach to implement the aforementioned research goals. We conclude with an outline of the current state of research and the achieved results so far.

2 Problem

As mentioned above, the main goal of this research is to define an approach that creates a design model of an embedded system, which is coherent and consistent in timing, based on the information contained in execution traces. Another prerequisite is that this model

has to be compliant to AUTOSAR [aut13], the Automotive Open System Architecture, in order to guarantee applicability in the industry for this research.
The mentioned model must therefore cover elements from hardware to software as well as operating system characteristics. The former has to include an abstract description of the used processing unit with the number of cores and their frequencies.
Main focus of the research however will be on the software model. This must describe how the system is composed in the matter of tasks, functions and used resources. Each of these components have again a number of characteristics that have to be determined. For tasks these are, for example, the priority, preemptability, maximum number of concurrent activations, period and offset. In case of functions, especially the exact modelling of accesses to resources provided by the operating system, such as data signals or semaphores, is of importance. That way the dependency between functions in the system is comprehensible, which is crucial for deploying software on multiple processing cores. Finally, the operating system has to be described by determining the used scheduling algorithm and modelling its characteristics.
Based on these requirements this work is organised around the following four research problems:

1. **Method for model extraction**
   First of all, a method for automatically building an AUTOSAR compliant model based on the observed behaviour recorded in a trace has to be developed. Such a general applicable method itself already states a huge innovation for the industry due to the fact that currently only very limited research has been done in the area of reverse engineering for embedded systems [KKM11].
   The method can be developed based on different technical approaches. On the one hand it can be implemented purely algorithmically, where all known is explicitly programmed. On the other hand knowledge can also be learned inductively.
   Therefore, a method has to be developed, which not only solves this problem best but which also performs best regarding accompanied challenges such as creating the most confidence in the resulting model or providing the best robustness regarding abnormalities in the trace.

2. **A heuristic for tracing**
   The information contained in a trace depends on many facts such as, for example, the length of time recorded in real time. This means that recording for a longer time period can result in more details about the system. Similarly, the levels of abstractions considered as well as the granularity of the system proceedings observed are crucial. Hence, the quality of a trace can vary greatly. However, the accuracy of the model is again dependent on the quality of the measured data.
   Unfortunately, recording every detail of the proceedings in a system all the time is also not possible. This is due to the fact that every measurement affects the timing behaviour of the system in a little, known as the so called probe effect [Gai86]. Thus, extensive recording can lead to erroneous runtime behaviour of a system. Furthermore, tracing has to cope with the limited system resources and connectivity available for embedded systems [TVD12]. Those lead to the fact that not enough trace data can be stored and transferred to obtain all arising events.

For those reasons, we think a heuristic for tracing is necessary, which tackles the question how to get the most usable information into a trace recording for building a coherent and consistent model regarding a specific use case.

3. **Model refinement**
   Another fact regarding tracing is that each recording only contains information about the system's behaviour that has actually been observed. So each time just a fraction of the complete system behaviour is received [HA05]. The quality of the targeted model however is dependent on those observations. If there is no recording of a certain component of the system, the approach and consequently the model cannot get knowledge about this. Yet, the more detailed and complete the created model is, the more valuable and significant it gets for further purposes.
   Hence, additional thoughts have to be given on how new, informative recordings can be explored in order to gradually refine the quality of an existing model.

4. **Quality of the approach**
   Finally, the quality of the approach has to be assessed in a convincing and confident way. One of the goals of the reverse engineering is to get a model that can be used for timing simulation and further analysis. As a consequence, the resulting model is only of value, if and only if a clear statement about its quality can be made. Quality in this case means, to measure to what extent the model behaviour conforms to that of the actual system. When considering the targeted field of application for such an automatic reverse engineering approach, for example the automotive industry, where functionality is bought from suppliers in form of black boxes, it gets obvious that trace recordings are the only available information sources about the actual system. Hence, in order to get such a quality measure, a precise way to compare a model with trace recordings has to be defined.

3 Related Work

In general 'surprisingly little research in reverse engineering targets embedded systems' [KKM11, p.8]. In this work a broad overview on the research activities in the area of software reverse engineering for embedded systems is given.
One of the latest publications about architecture recovery mentioned there is [MW10]. Marburger and Westfechtel present an approach to analyse the structure as well as the behaviour of a system. For that, a state machine is extracted from the available source code. Our research however seeks applicability in the automotive industry. There the availability of source code cannot be ensured, since it is common to purchase functionality from external suppliers, which then have to be considered as black boxes.
Huselius [HA05, AHNW06, HAHP06, Hus07] provides probably the most related work. In the context of his Ph.D. thesis [Hus07] he researched reverse engineering of legacy real-time systems in order to develop an automated approach based on execution-time recording. For this, he creates a probabilistic state-machine model expressed in the ART-ML language [HA05]. Nevertheless, his work provides some shortcomings, which are

crucial for our purpose to use the model for simulation-based timing analysis. ART-ML, for example, provides a very high-level view on the system. This means resource accesses are associated with tasks instead of functions. To be able to allocate software to different cores and to optimise it regarding certain metrics however, the knowledge of functions and their execution order is significant. He moreover makes certain assumptions for his research, such as the knowledge of the priority or the triggering method of the tasks, which are not acceptable for our intended applicability.
One of the latest research activities in this area comes from Ciccozzi [CSCS13]. There, a round-trip approach for component-based embedded systems is proposed. The goal is to provide an automated support for deployment decisions at modelling level. It however focusses only on the determination of extra-functional properties and doesn't consider architectural or behavioural characteristics for back-propagation to the design model.
In addition to the aforementioned related works, which cover existing ambitions in the area of embedded systems, relevant approaches can also be found in the area of process mining. The goal there is to extract information from event logs in order to capture the underlying business process. An overview on the main issues around this topic can be found in [VdAW04]. As the name already implies, process mining focusses on the application of data mining techniques in order to achieve the desired goal. Attention however should be paid to the fact, that models for real-time systems have to satisfy much stricter requirements then business processes.
The confidence in the quality of the resulting model plays a crucial role. Therefore, also a vast variety of literature on model validation can be found. Balci proposes in [Bal90] different techniques for model validation in simulation studies. This work is amended by Sargent [Sar04]. Law presents in [Law09] general techniques for validating a simulation model. In the end, all of the above list just a set of statistical techniques, which can be used for general validation. Huselius [HKHP07, Hus07] developed a metric, which is used to evaluate the quality of models from embedded real-time systems by analysing the response time of the available tasks. Kraft discusses in [Kra10] a possibility for the validation of simulation models by comparing simulation traces with traces from the modelled system. Although different steps are proposed in this work, no particular properties are presented, but just a vague approach how to find those properties. Furthermore, most steps proposed target obviously a manual validation and are thus not feasible for an automatic approach.

4 Approach

In order to achieve the projected goals, this research is divided into three consecutive phases. However, the first and second phase are meant to be executed iteratively. For example in cases, where the evaluation at the end of a phase reveals insufficient quality, an iteration of the phase or even the previous one is necessary to improve the algorithms. In the following subsections, the individual phases are presented in detail.

1. Simulation-based Development & Validation Phase

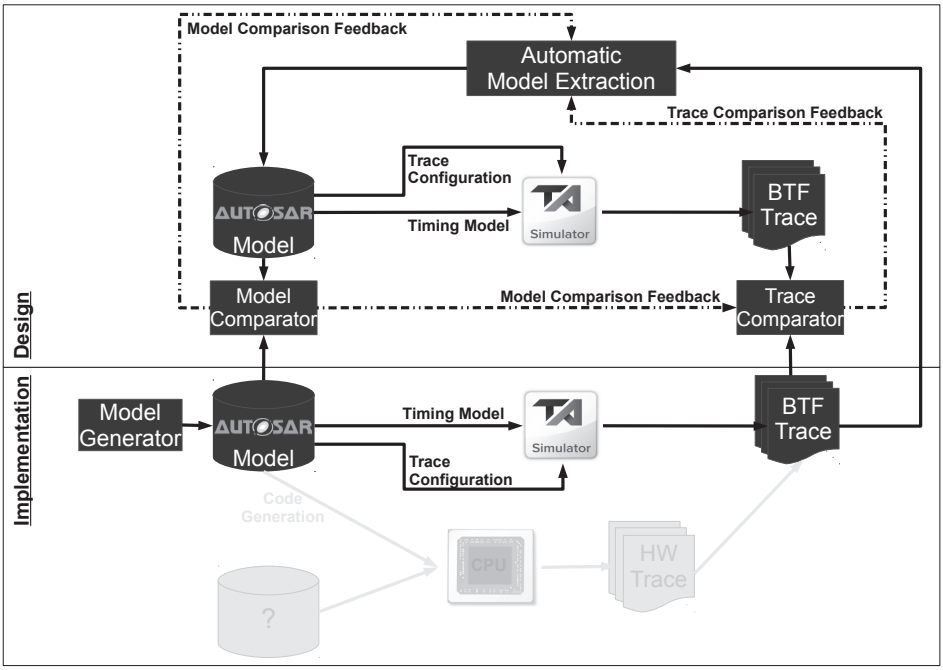


Figure 1: Simulation-based Development & Validation Phase

In the first phase as depicted in Figure 1, all necessary development is done utilizing a discrete event simulation. For this, the simulation tool from the Timing Architects Toolsuite [TA] is used. The advantage thereby is that trace recordings can be generated much faster and more comfortable than working with real hardware. The proposed approach starts with an implementation model, which is pictured in the lower part of the figure. Based on this AUTOSAR compliant model, the simulator creates trace recordings corresponding the configured granularity and length. That way the resulting information contained within a single trace file can be set electively. This functionality is designated to investigate the problem of specifying a heuristic for tracing.

The generated trace recordings serve then as input for the automatic model extraction. As the name already implies, this automatic model extraction builds then again an AUTOSAR compliant model based on the information contained in trace recordings. Thus, the quality of the resulting model can easily be determined by comparing it with the implementation model. That way the development of a method for model extraction can be done model-based, which makes it possible to make a clear statement about the quality of the resulting model and as a consequence also about the quality of the method itself.

However, later during the operation of the automatic reverse engineering approach, there might not be an implementation model available to compare to. As a consequence, a pos-

sibility to assess the created model with regards to trace recordings from the origin source has to be available. This is done by simulating the gained model and afterwards comparing the resulting trace with other trace recordings from the original source. This procedure makes it also possible to test the potential of supervised learning techniques in the automatic model extraction as well as in the trace comparison.
The development proceeds then until the automatic model extraction continually produces models of the desired quality. This is evaluated by using variations of major architectural system designs for the implementation model.

2. Operational Adoption & Validation Phase

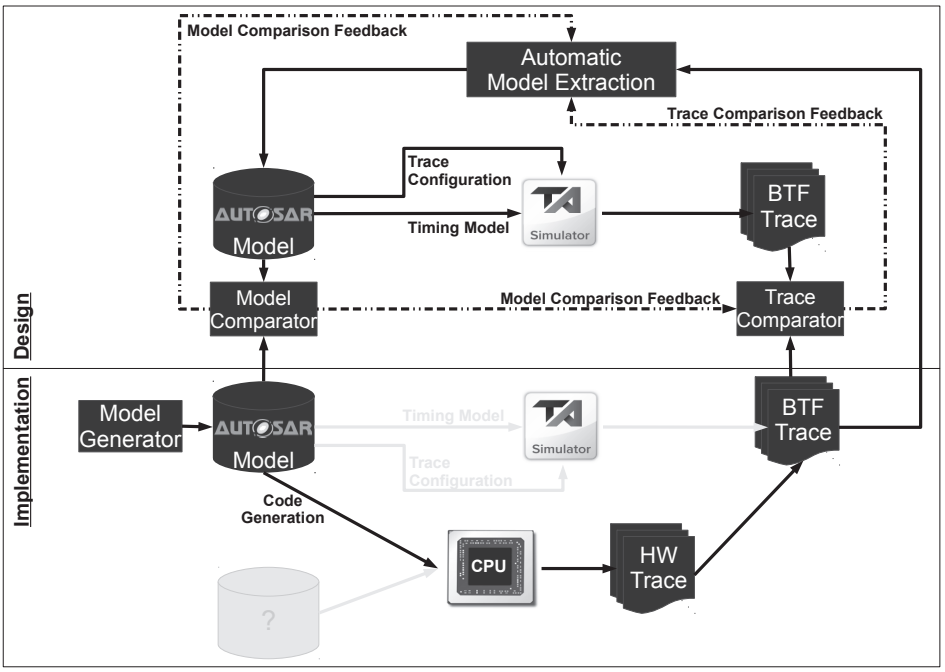


Figure 2: Operational Adoption & Validation Phase

In the second phase, the next step towards an operational use of the reverse engineering approach is taken. For that, the method for model extraction developed up to that point is validated using hardware traces instead of trace recordings generated by the discrete event simulation.
Starting point for this are once again the exact kind of models as used before with the simulation. Those however are now executed on real hardware by using code generation. The resulting hardware traces are then transformed into the trace format of the simulation, called BTF Trace Format. This process is just a simple mapping of elements from a binary trace format, which is used to compress the data within a hardware trace, to a more usable

trace format.
From here on, as depicted in Figure 2, the rest of the work flow remains the same. That way, the differences between the simulation and execution can be determined and the automatic model extraction adapted accordingly.
Since this approach is still model-based, the quality of the resulting model can be determined by comparing it with the implementation model. That way the automatic model extraction can again be improved until it continually produces models of the desired quality.

3. Operational Phase

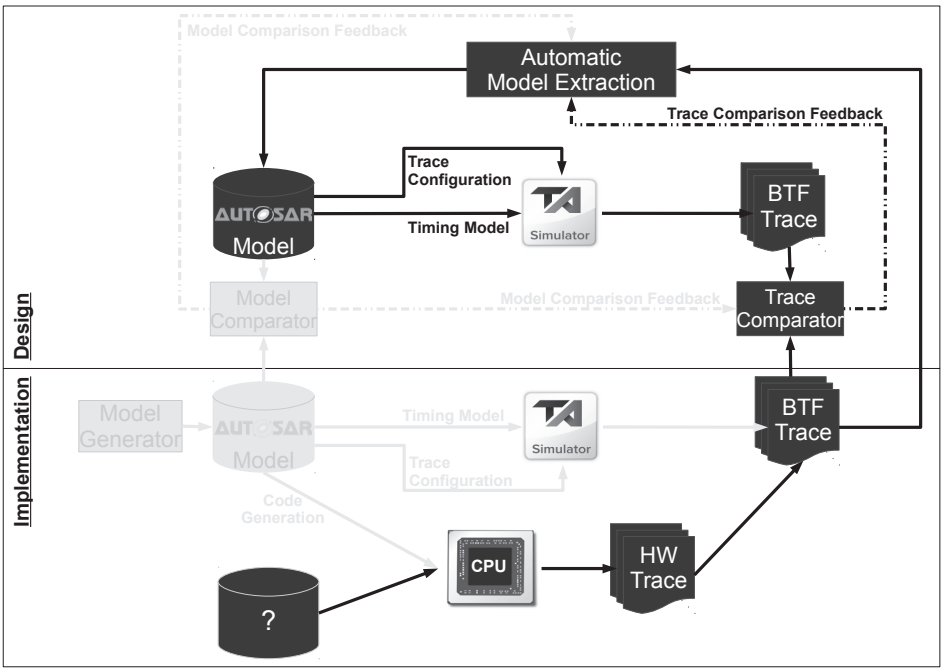


Figure 3: Operational Phase

Last but not least, the developed reverse engineering approach is ready for operation. While the starting point of the previous phases was each time an implementation model, this time it is not the case. Instead, the system has to be considered as a black box for different reasons. On the one hand, for example, the system might have not been developed model-based and thus there was never a model available. On the other hand the source code generated from a model might also have been altered manually afterwards without synchronizing those changes with the model.
As a consequence, all the information available about the system comes from hardware traces and there is no possibility to assess the resulting model with regards to an imple-

mentation model. However, as depicted in Figure 3, the work flow of the reverse engineering approach remains the same. So, since this approach already produced models of a desired quality before throughout the previous two phases, the user can be confident in the resulting model this time, too.
In addition to this fact, a statement about the quality of the resulting model is made by comparing it with other trace recordings from the original source. This is also done equivalently to the previous two phases by simulating the gained model and afterwards comparing the resulting trace with the other trace recordings from the origin source. That way the needed confidence from the user in the resulting model should be achieved.

5 Preliminary Results & Future Work

The preliminary results obtained so far demonstrate not only the feasibility of the proposed approach but also the necessity of research in this area. The following paragraphs give an overview on the current achievements and developments.
Currently, the transition between the first and second phase is in process. In a first step the work focused on extracting individual properties of a model one by one from trace recordings. For example, the used scheduling algorithm is determined from a trace by using constraint programming. At first, significant changes in the state of a task, such as task preemptions, are detected within the trace. Based on those a corresponding constraint describing the expected behaviour of a scheduling algorithm is created. In case of a fixed-priority scheduler, for each task preemption a constraint expressing that the priority of the preempting task is higher than that of the preempted one would be added to the constraint system of the represented scheduling algorithm. Finally, if the constraint system has a solution, this means that there was no fact militating against the scheduling algorithm under investigation.
Another example of determining a property of the system behaviour is the specification of the dynamic runtime of functions. For that, the individual runtimes of a function are calculated from the trace at first. Those values are then used for a goodness-of-fit test to examine, how well they fit a statistical distribution. Further impressions on the work of extracting individual properties of a model can be found in [SSD+13].
For further development on the reverse engineering approach, AUTOSAR as the targeted meta-model was replaced by AMALTHEA [ama13]. AMALTHEA is an open-source meta-model, which is not only compliant to AUTOSAR but also continuously covers all information during the development of software for embedded systems [SSDM13]. That way it was possible to gather more interesting or relevant pieces of information in the model.
Up to now a set of heuristic methods to determine the most critical parameters from trace recordings is available. The quality of the developed automatic modelling approach is ensured by a model comparison. For this, the Eclipse Modelling Framework (EMF) was used, since both, AUTOSAR and AMALTHEA, have a meta-model description available in EMF. Thus, a plain comparison of models was possible by comparing element by element.

Figure 4 provides sample results from such a comparison. For that chart, 20 random mod-

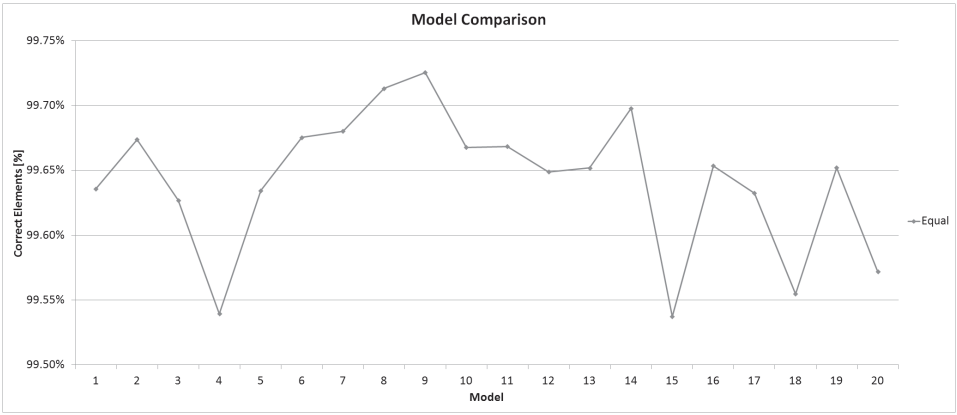


Figure 4: EMF Model Comparison

els were created. Each model represents a purely periodic task set without communication in a preemptive, fixed-priority scheduling environment. The size of the task sets varied from four to a maximum of 18 tasks. Within those, between 1374 and 2797 functions where called, where each function had a constant execution time. The length of the trace was chosen such that all tasks occurred at least once. The comparison of the reverse engineered model with the original model showed, that at least for such a simple example nearly every detail could be determined from the trace recordings. First validations of the current achievements on real hardware yielded promising results too.

In order to be also able to make a statement about the capability of the newly developed approach compared to existing ones, a benchmark is currently defined. Due to the fact that Huselius provides the most related work, the results published in [Hus07] will be considered for this.

Furthermore, one of the major tasks in the near future will be the comparison of characteristics from trace recordings to those of a model. In the course of this research different metrics will be examined and developed in order to define a comparison framework. Such metrics will be on the one hand performance measures of the system such as response times of tasks or execution times of functions, but on the other hand also abstract ones such as the pattern of task interferences. The former are analysed using once again statistical goodness-of-fit tests such as the two-sample Kolmogorov-Smirnov test. Based on the patterns, coherences between the individual metrics are examined by performing correlation, regression and association analyses. Since this is an obvious field of application for machine learning techniques such as supervised learning or pattern recognition, this activity represents a good starting point to also include those techniques in the automatic model extraction.

Before that however, an expressive evaluation of the automatic modelling approach must be defined in order to make a statement on the impact of those techniques on the model quality. In parallel to those activities also the work on validating the preliminary results on

real hardware will be intensified.

6 Acknowledgments

This work was partially funded by the ITEA2 project AMALTHEA (ITEA2-09013) through the German Ministry of Education and Research (BMBF) under the funding code 01IS11020F. I would like to thank my supervisors Jürgen Mottok and Gerald Lüttgen as well as Michael Deubzer for the many valuable discussions regarding this research.

References

[AHNW06] Johan Andersson, Joel Huselius, Christer Norstrom, and Anders Wall. Extracting simulation models from complex embedded real-time systems. In *Software Engineering Advances, International Conference on*, pages 7–7. IEEE, 2006.

[ama13] http://www.amalthea-project.org, November 2013.

[AMS+13] Martin Alfranseder, Matthias Mucha, Stefan Schmidhuber, Andreas Sailer, Michael Niemetz, and Jürgen Mottok. A modified synchronization model for dead-lock free concurrent execution of strongly interacting task sets in embedded systems. In *Applied Electronics (AE), 2013 International Conference on*, pages 1–6. IEEE, 2013.

[aut13] http://www.autosar.org, November 2013.

[Bal90] Osman Balci. Guidelines for successful simulation studies. In *Simulation Conference, 1990. Proceedings., Winter*, pages 25–32. IEEE, 1990.

[CSCS13] Federico Ciccozzi, Mehrdad Saadatmand, Antonio Cicchetti, and Mikael Sjödin. An automated round-trip support towards deployment assessment in component-based embedded systems. In *Proceedings of the 16th International ACM Sigsoft symposium on Component-based software engineering*, pages 179–188. ACM, 2013.

[Gai86] Jason Gait. A probe effect in concurrent programs. *Software: Practice and Experience*, 16(3):225–233, 1986.

[GLT03] Bas Graaf, Marco Lormans, and Hans Toetenel. Embedded software engineering: the state of the practice. *Software, IEEE*, 20(6):61–69, 2003.

[HA05] Joel Huselius and Johan Andersson. Model synthesis for real-time systems. In *Software Maintenance and Reengineering, 2005. CSMR 2005. Ninth European Conference on*, pages 52–60. IEEE, 2005.

[HAHP06] Joel Huselius, Johan Andersson, Hans Hansson, and Sasikumar Punnekkat. Automatic generation and validation of models of legacy software. In *Embedded and Real-Time Computing Systems and Applications, 2006. Proceedings. 12th IEEE International Conference on*, pages 342–349. IEEE, 2006.

[HKHP07] Joel Huselius, Johan Kraft, Hans Hansson, and Sasikumar Punnekkat. Evaluating the quality of models extracted from embedded real-time software. In *Engineering of Computer-Based Systems, 2007. ECBS'07. 14th Annual IEEE International Conference and Workshops on the*, pages 577–585. IEEE, 2007.

[HLL05] Abdelwahab Hamou-Lhadj and Timothy C Lethbridge. Measuring various properties of execution traces to help build better trace analysis tools. In *Engineering of Complex Computer Systems, 2005. ICECCS 2005. Proceedings. 10th IEEE International Conference on*, pages 559–568. IEEE, 2005.

[Hus07] Joel Huselius. *Reverse Engineering of Legacy Real-Time Systems: An Automated Approach Based on Execution-Time Recording*. PhD thesis, Mälardalen University, 2007.

[KKM11] Holger M Kienle, Johan Kraft, and Hausi A Müller. Software Reverse Engineering in the Domain of Complex Embedded Systems. *InTech*, 2011.

[Kra10] Johan Kraft. *Enabling timing analysis of complex embedded software systems*. PhD thesis, Mälardalen University, 2010.

[KWK10] Johan Kraft, Anders Wall, and Holger Kienle. Trace recording for embedded systems: Lessons learned from five industrial projects. In *Runtime Verification*, pages 315–329. Springer, 2010.

[Law09] Averill M Law. How to build valid and credible simulation models. In *Simulation Conference (WSC), Proceedings of the 2009 Winter*, pages 24–33. IEEE, 2009.

[MW10] André Marburger and Bernhard Westfechtel. Graph-based structural analysis for telecommunication systems. In *Graph transformations and model-driven engineering*, pages 363–392. Springer, 2010.

[QHE12] Sophie Quinton, Matthias Hanke, and Rolf Ernst. Formal analysis of sporadic overload in real-time systems. In *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012*, pages 515–520. IEEE, 2012.

[RS13] Salman Rafiq and Adriaan Schmidt. Systematic Modeling of Workflows in Trace-Based Software Debugging and Optimization. In *ICSEA 2013, The Eighth International Conference on Software Engineering Advances*, pages 241–248, 2013.

[Sar04] Robert G Sargent. Validation and verification of simulation models. In *Simulation Conference, 2004. Proceedings of the 2004 Winter*, volume 1. IEEE, 2004.

[SSD+13] Andreas Sailer, Stefan Schmidhuber, Michael Deubzer, Martin Alfranseder, Matthias Mucha, and Jürgen Mottok. Optimizing the task allocation step for multi-core processors within AUTOSAR. In *Applied Electronics (AE), 2013 International Conference on*, pages 1–6. IEEE, 2013.

[SSDM13] Andreas Sailer, Stefan Schmidhuber, Michael Deubzer, and Jürgen Mottok. AMALTHEA - Plattform für kontinuierliche, modellbasierte Entwicklung. In *Tagungsband Embedded Software Engineering Kongress (ESE), 2013*, pages 538–543, 2013.

[TA] Timing-Architects. TA Toolsuite Version 13.09.0. TA Academic & Research License Program. [Online]. Available: http://www.timing-architects.com.

[TVD12] J Trumper, Stefan Voigt, and J Dollner. Maintenance of embedded systems: Supporting program comprehension using dynamic analysis. In *Software Engineering for Embedded Systems (SEES), 2012 2nd International Workshop on*, pages 58–64. IEEE, 2012.

[VdAW04] Wil MP Van der Aalst and AJMM Weijters. Process mining: a research agenda. *Computers in industry*, 53(3):231–244, 2004.

[WEE+08] Reinhard Wilhelm, Jakob Engblom, Andreas Ermedahl, Niklas Holsti, Stephan Thesing, David Whalley, Guillem Bernat, Christian Ferdinand, Reinhold Heckmann, Tulika Mitra, et al. The worst-case execution-time problem - overview of methods and survey of tools. *ACM Transactions on Embedded Computing Systems (TECS)*, 7(3):36, 2008.

Querying Probabilistic Ontologies with SPARQL

Joerg Schoenfisch

Data & Web Science Research Group
University of Mannheim, Germany
http://dws.informatik.uni-mannheim.de/

joerg@informatik.uni-mannheim.de

Abstract: In recent years a lot of efforts was put into the field of Semantic Web research to specify knowledge as precisely as possible. However, optimizing for precision alone is not sufficient. The handling of uncertain or incomplete information is getting more and more important and it promises to significantly improve the quality of query answering in Semantic Web applications. My plan is to develop a framework that extends the rich semantics offered by ontologies with probabilistic information, stores this in a probabilistic database and provides query answering with the help of query rewriting. In this proposal I describe how these three aspects can be combined. Especially, I am focusing on how uncertainty is incorporated into the ABox and how it is handled by the database and the rewriter during query answering.

1 Introduction

In recent years several standards for knowledge representation, e.g. RDF[1] or OWL[2], were recommended by the W3C, and methodologies, like OntoClean [GW09], for putting these recommendations to work evolved. Those formalisms and methods focus on specifying knowledge as precisely as possible. OntoClean, for example, categorizes expressions in an ontology into rigid and non-rigid ones and discourages the usage of non-rigid knowledge to avoid false reasoning results.

However, as Paulheim and Pan already stated, optimizing the Semantic Web for precision alone is not sufficient [PP12]. It is trivial to design a system that always achieves very high precision, but subsequently it will suffer from very low recall. The impact for linked data or ontologies is so severe that IBM decided not to use structured knowledge bases in this way for their famous Jeopardy-winning Watson system, as it would only have been able to answer about two percent of the posed questions [KBP+12]. Instead, they developed a system that allows for imprecise or uncertain knowledge to improve recall and achieve a better overall performance.

The handling of uncertain or incomplete information is getting more and more important,

[1] http://www.w3.org/standards/techs/rdf
[2] http://www.w3.org/standards/techs/owl

not only in the domain of expert systems or query-answering systems like Watson, but also in other real world applications:

- *Data integration* is the mapping of two or more heterogeneous data sources, e.g. ontologies without a common upper ontology or database with a different schema, into one combined knowledge base. In many cases it is not possible to find perfect matches between all concepts and properties of the different sources. However, quite often there are concepts that almost correspond to each other or have similar meanings. Mapping those with some specific confidence value yields an increased benefit for the resulting integrated knowledge base.
- *Information extraction* is another field where uncertain information plays an important role. Information extraction tries to generate meaningful, structured knowledge from merely unstructured data in the form of natural text. Although natural language processing provides good tools to derive information from unstructured text, natural language has many ambiguities that require context or background knowledge to discern the correct meaning of a term or statement (cf. [TDD+13]).

For all these cases query answering can be improved by supporting uncertain information properly. I suggest a framework that combines the rich semantic knowledge available within ontologies and linked data with the benefit of an increased recall provided by the incorporation of imprecise knowledge when answering queries. This framework shall also scale well for large amounts of information as for example in Linked Open Data. My contribution will thus be the combination of the probabilistic results from the database with the information from a Semantic Web reasoner.

There are already various implementations for formalisms that incorporate uncertain or probabilistic knowledge, e.g. probabilistic relational databases [HAKO09], bayesian networks [WMGH08], or markov logic networks [JS12]. The former are focused on querying large relational data, however, without rich semantics, whereas the latter belong to the field of probabilistic programming and are more concerned with statistical relational learning.

Storing and querying large amounts of semantic data is already possible in various implementations, e.g. Virtuoso[3], Stardog[4], OWLIM[5] or Oracle[6], but these applications do not support any notion of uncertainty.

Probabilistic databases are around for about a decade and are in a very mature state [TDD+13]. With their focus on large databases and their foundation in relational database management systems, they are a natural choice for the combination of probabilistic knowledge bases and querying those with SPARQL.

The OWL 2 recommendation defines a special profile, OWL 2 QL[7], that is designed to enable query answering in relational databases with large amounts of instance data in mind.

[3] http://virtuoso.openlinksw.com/
[4] http://stardog.com/
[5] https://www.ontotext.com/owlim
[6] http://www.oracle.com/technetwork/database/options/spatialandgraph/overview/rdfsemantic-graph-1902016.html
[7] http://www.w3.org/TR/owl2-profiles/#OWL_2_QL

$$\exists sisterOf \sqsubseteq Woman \quad [1.00] \quad (1)$$
$$\exists motherOf \sqsubseteq Woman \quad [1.00] \quad (2)$$
$$\exists marriedTo \sqcap Woman \sqsubseteq Wife \quad [1.00] \quad (3)$$
$$marriedTo \equiv marriedTo^{-1} \quad [1.00] \quad (4)$$
$$\exists motherOf \sqsubseteq \exists marriedTo \quad [0.65] \quad (5)$$

$$marriedTo(Alice, Bob) \quad [0.30] \quad (6)$$
$$sisterOf(Alice, Carol) \quad [0.80] \quad (7)$$
$$motherOf(Alice, Dave) \quad [0.95] \quad (8)$$

Figure 1: Example used to illustrate my approach in the proposed framework. Numbers in square brackets represent probabilities.

This is achieved by limiting the expressivity of the ontology, leading to a reduced complexity of the reasoning task for the knowledge base. Due to this, answers to a query can be computed through a technique called query rewriting. The main benefit of query rewriting is the fact that it needs no information about the instance data (ABox) of a knowledge base, but only of its schema (TBox). This enables reasoning during query time and changing the data without the need for a separate reasoning step afterwards which results in answers to a query directly reflecting changes in the knowledge base.

The recommended language to query semantic data is SPARQL. It looks similar to SQL but it is designed to query graphs (the underlying structure of RDF or OWL) instead of relational data.

The example shown in Figure 1 will be used throughout Section 3 to illustrate my approach in the framework. The ontology consists of 5 axioms in the TBox and 3 assertions in the ABox stating the following:

1. A sister of someone is a woman
2. A mother of someone is a woman
3. Someone who is married and a woman is a wife
4. The married property is symmetric
5. Alice is married to Bob with a probability of 0.30
6. The probability of Alice being Carol's sister is 0.80
7. The probability of Alice being Dave's mother is 0.95

I will show in detail how a query is rewritten using the certain TBox axioms 1 – 4. How the uncertain Axiom 5 can be incorporated during rewriting is briefly discussed at the end of Section 3 and is part of future work. I describe which of the assertions 6 – 8 are relevant answers to the query and how to compute the probabilities of the final result from their given probabilities.

The rest of this paper is structured as follows: In Section 2, I am discussing related work.

Section 3 details the approach of my framework. The current status and short term goals of my work are stated in Section 4. Section 5 outlines future work.

2 Related Work

Related work can be divided roughly into two fields: probabilistic querying, and Semantic Web reasoning. Probabilistic deductive databases, e.g. as proposed by Lakshmanan et al. [LS94], are also closely related to my approach. However, they employ technologies different from query rewriting and do not focus on large scale data.

2.1 Probabilistic Querying

The two main areas of research in probabilistic querying are probabilistic programming and probabilistic databases [TDD+13]. Both have developed rather independently during the last decade but they share a common goal in making probabilistic inference scalable.

Probabilistic programming mainly focuses on learning statistical models from existing data. Its main concern is determining whether there exists a solution to a query rather than finding all solutions as probabilistic databases do. There are several formalisms, e.g. markov logic networks [RD06], bayesian networks [Jen96], or languages and tools like ProbLog [KTD07], to describe probabilistic models. These formalisms and the tools using them, for example MarkoViews [JS12] and BayesStore [WMGH08], can handle large amounts of uncertain data, however, none of them incorporates Semantic Web reasoning.

Similarly, probabilistic databases like Trio [Wid04] or MayBMS [HAKO09] are designed for querying large probabilistic knowledge bases, but they are only handling relational data without rich semantic information. Probabilistic databases usually employ a *possible world semantics*: all possible values for uncertain values are represented at once. To enable efficient inferencing, they exploit independence assumptions, most commonly tuple-independence (all database tuples or rows are independent from each other) or block-independent (groups of tuples are independent from each other). The research on probabilistic databases draws from many years of experience with relational databases. From those, they adopt relational algebra – and thus SQL – or Datalog.

Theobald et al. [TDD+13] give a more thorough overview of these two research areas, their methodologies, differences, and commonalities.

2.2 Semantic Web Reasoning in the Context of Query Answering

In the context of query answering, there are two commonly used approaches for Semantic Web reasoning with large scale instance data: *materialization* and *query rewriting*.

For materialization all knowledge that can be inferred from known facts is explicitly stored when the data is loaded. Thus, no reasoning is needed later on, as long as the data is not modified. If the data is modified the materialization has to be computed anew which can be quite expensive. For example OWLIM uses this approach.

Query rewriting, on the other side, does not change or add data but rather transforms a query over a knowledge base in such a way that it produces sound and complete answers without the need for any reasoning inside the database. In the case of OWL 2 QL and similar description logics this means that the query is expanded with information from the TBox such that implicit knowledge can be extracted without the need for any consideration of the ABox during reasoning. This is achieved by limiting the expressiveness of the description logic which is used to describe the data. Calvanese et al. [CDL05] give a thorough overview in the their paper about the $DL\text{-}Lite$ family of description logics. The OWL 2 QL profile is based on $DL\text{-}Lite_R$. PerfectRef, one of the first rewriting algorithms for semantic web reasoning, is also introduced in their paper. Applications that use query rewriting for reasoning are for example Stardog or the Maestro system [CDL$^+$11]. None of the applications for large scale semantic web reasoning incorporates uncertain information.

In my framework I want to adopt a query rewriting algorithm. There are already some implementations that combine those algorithms with traditional relational databases, e.g. Owlgres [SS08], Oracle RDF Semantic Graph, or YARR [SO13]. However they are outdated (Owlgres), or not readily available (Oracle, YARR).

3 Approach

As outlined in the introduction I suggest to use a mature probabilistic relational database system and combine it with an existing query rewriting algorithm for reasoning. I use SPARQL as query language, which is recommend by the W3C for Semantic Web applications. Figure 2 shows an illustration of my approach. The two main parts of the implementation will be the query rewriter and the probabilistic database system. The query rewriter will use the knowledge from the TBox of one or more consistent ontologies to extend the queries. The information of the ABox will be stored in the database, where the computation of the results takes place. Ontologies have to be provided in OWL format, queries will be posed in SPARQL. The communication between the rewriter and the database is done via SQL (with probabilistic extensions). The results will be presented to the user as tables, as is common for databases or triple stores.

As query rewriter I am going to use the implementation of IQAROS[8] described by Venetis et al. [VSS12]. IQAROS loads the TBox of an ontology in OWL format and rewrites queries formulated in Datalog. To be able to pose queries in SPARQL and compute answers through the probabilistic database, I will implement two additional translation steps: one from SPARQL to Datalog before the rewriting takes place, and another from Datalog to SQL with probabilistic extensions for the probabilistic database.

[8] IQAROS is one of the few openly available implementations. It can be downloaded under this link: `https://code.google.com/p/iqaros/`

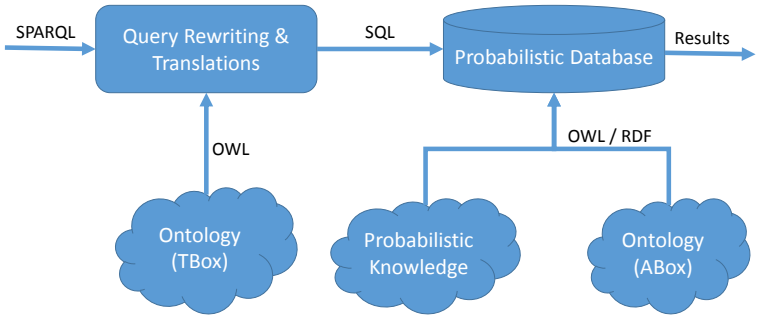


Figure 2: Illustration of the different steps of my approach and the technologies involved.

Given the example from the introduction (Fig. 1), a SPARQL query asking for all wifes looks like the following:

```
SELECT ?person WHERE { ?person a Wife . }
```

Translating SPARQL queries with simple triple patterns like that to Datalog is straightforward. Other features like optional joins or aggregates, e.g. COUNT or MAX, are more complex. These constructs are not part of Datalog, however, they are supported by SQL and probabilistic databases and thus can be used with some special handling in the translation and rewriting steps. This is the query translated to Datalog:

```
q(?person) :- Wife(?person)
```

The relevant axioms of the TBox that will be used by the rewriter to extend the query are then the following: a) a wife is married and a woman (Axiom 3), b) a person who is the mother of someone is a woman (Axiom 2), c) a person who is the sister of someone is a woman (Axiom 1), d) the relation marriedTo is symmetric (Axiom 4). The extended query looks like this:

$$q(?person) :- Wife(?person) \tag{9}$$
$$q(?person) :- marriedTo(?person, ?x), Woman(?person) \tag{10}$$
$$q(?person) :- marriedTo(?x, ?person), Woman(?person) \tag{11}$$
$$q(?person) :- marriedTo(?person, ?x), sisterOf(?person, ?y) \tag{12}$$
$$q(?person) :- marriedTo(?x, ?person), sisterOf(?person, ?y) \tag{13}$$
$$q(?person) :- marriedTo(?person, ?x), motherOf(?person, ?y) \tag{14}$$
$$q(?person) :- marriedTo(?x, ?person), motherOf(?person, ?y) \tag{15}$$

The initial query remains unchanged (9). The next two queries (10 and 11) use the definition of wife (Axiom 3) and the symmetry of marriedTo (Axiom 4). The following two

queries (14 and 15) use the same two axioms and additionally the definition that every sister is a woman (Axiom 1). The last two queries (12 and 13) are similar to the two before, but use the axiom about motherOf (Axiom 2).

Because of the restricted expressivity of OWL 2 QL, all rewriting algorithms only produce unions of conjunctive queries or non-recursive Datalog programs. The expressive power of those queries is equal to that of relational algebra, i.e. essentially SQL. Thus the rewritten Datalog query can also be fully translated to SQL. The translation of Query 9 looks as follows:

```
SELECT abox.instance AS wife, abox.conf() AS prob
FROM abox
WHERE abox.class = 'Wife'
```

The function *conf()* is provided by the probabilistic database and returns the probability for a single row of the result.

Query 10 is translated the following way:

```
SELECT abox.instance AS wife,
       (abox.conf() * married.prob) AS wife_prob
FROM abox
LEFT JOIN (SELECT relation.subject AS person,
                  conf() AS prob
           FROM relation
           WHERE relation.subject = abox.instance AND
                 property = 'marriedTo'
           GROUP BY relation.subject) married
    ON married.person = abox.instance
WHERE abox.class = 'Woman'
```

For the sake of brevity I omit the rest of the queries. They are constructed in a way similar to Query 10.

The last query also shows how the probabilities in a single query are aggregated. I assume that all atoms of a query are mutually independent and thus their probabilities for the individual answers can simply be multiplied to compute that of the final result. For the given example, this assumptions is valid. However, it is easy to formulate a query that contains atoms with dependencies, for example $q(?child) :- motherOf(?x, ?child)$, $Woman(?x)$. Clearly, $motherOf(?x, ?child)$ depends on $Woman(?x)$, as every mother has to be a woman, and thus their probabilities may not simply be multiplied to compute the result. How to handle dependencies in detail is part of my current work (cf. Section 4).

After translating all queries, I concatenate them using `UNION ALL` and send the resulting single query to the probabilistic database. For this small example it will return the following result:

```
 wife          | conf
=====================
 Alice         | 0.240
 Alice         | 0.285
```

For each row of a result, there exists an explanation (cf. [KPHS07]) why it is part thereof. A single explanation contains all axioms and assertions involved in the computation of one result row. The explanations for the results above are the following:

$$\begin{gathered}\exists marriedTo \sqcap Woman \equiv Wife \\ \exists sisterOf \sqsubseteq Woman \\ marriedTo(Alice, Bob), [0.30] \\ sisterOf(Alice, Carol), [0.80]\end{gathered} \tag{16}$$

$$\begin{gathered}Wife \equiv \exists marriedTo \sqcap Woman \\ \exists motherOf \sqsubseteq Woman \\ marriedTo(Alice, Bob), [0.30] \\ motherOf(Alice, Dave), [0.95]\end{gathered} \tag{17}$$

In the case of Alice, there are two different explanations for her being a wife, each with its own probability. When aggregating those rows, it is again important to consider if they are independent or not. For independent rows, I can simply multiply the complementary probability of each row and use the complementary probability of this product as the resulting probability of the aggregation. In this case, however, the two results are dependent because their explanations both contain the assertion $marriedTo(Alice, Bob)$, $[0.30]$. In every world $marriedTo(Alice, Bob), [0.30]$ has to be true, however there are three possible combinations for $sisterOf(Alice, Carol), [0.80]$ and $motherOf(Alice, Dave), [0.95]$ for which Alice is a wife:

$$\begin{gathered}P(marriedTo(Alice, Bob)) = 0.30 \\ P(sisterOf(Alice, Carol)) = 0.80 \\ P(motherOf(Alice, Dave)) = 0.95\end{gathered} \tag{18}$$

$$\begin{gathered}P(marriedTo(Alice, Bob)) = 0.30 \\ P(sisterOf(Alice, Carol)) = 0.80 \\ P(\neg motherOf(Alice, Dave)) = 0.05\end{gathered} \tag{19}$$

$$\begin{gathered}P(marriedTo(Alice, Bob)) = 0.30 \\ P(\neg sisterOf(Alice, Carol)) = 0.20 \\ P(motherOf(Alice, Dave)) = 0.95\end{gathered} \tag{20}$$

Calculating the probability for each of this worlds by multiplying the individual atoms and summing them up gives the following final result:

```
 wife          | conf
=====================
 Alice         | 0.297
```

I plan to use the probabilistic database system MayBMS [HAKO09]. MayBMS is a mature and stable implementation that extends PostgreSQL. It defines its own query language which is an extension of SQL with uncertainty-aware constructs. MayBMS offers a solid and scalable foundation for my approach.

To store an ABox inside MayBMS, there is either the possibility to specifically create a schema for a given ontology, i.e. tables for every property, class, etc., or to develop a generic schema that can store any ontology. If the schema is tailored specifically for an ontology, this could provide performance benefits. However, for the framework to be as flexible as possible I will create a schema that is able to store arbitrary ontologies and probabilistic information.

It is an open question how to describe uncertain information in an ontology. To my knowledge there is no standard or recommendation that describes how this should be done. RDF and OWL offer the possibility for reification (statements about another statement), but this construct is cumbersome, hard to read and understand, and badly supported in editors. There was an incubator group, URW3-XG[9], at W3C but no working group was established to follow up on their work and propose a recommendation.

Furthermore, there are several possibilities which constructs of the ontology are allowed to be probabilistic. Technically, probabilistic information can be attached to either ABox elements only, TBox elements only, or both. Supporting uncertainty in the ABox is the easiest part, as the probabilistic database already supports this and does most of its handling. The rest of the computation can be done inside the query as shown in the example. Uncertainty in the TBox, however, needs to be incorporated into the query rewriting step, which will be more complicated. An example for uncertainty in the TBox is Axiom 5. It adds another possibility for Alice to be married. Probabilistic TBox axioms need special handling during the rewriting as every query derived using that axiom is directly influenced by its probability. In the beginning of my work, I will solely focus on probabilistic information in the ABox.

4 Current Status

So far, I implemented the translation step from SPARQL to Datalog for simple queries, e.g. without aggregates, optional graph patterns, sorting, etc. This enables me to pose simple SPARQL queries to IQAROS and get a rewritten query in Datalog.

Currently, I am detailing the theoretical work on how to aggregate the probabilities after the results are returned from the database. It is only valid to multiply the probabilities, or respectively the complementary probabilities, if all the explanations for a result are mutually independent. The assumptions used during Section 3 need further considerations,

[9] http://www.w3.org/2005/Incubator/urw3/

especially when extending uncertainty to the TBox.

My next steps are the following:

1. Defining a mechanism to include uncertainty into an ontology. As a first prototype, I will use annotations which are already available and can be used on almost all constructs.
2. Creating a database schema to store the ontology. In its first version, as a proof of concept, this schema will contain class assertions and object property assertions only. However, it can be easily extended later on. I plan to keep the TBox as a separate file and not to store it in the database. This way it will be easier to make changes with an editor of choice, e.g. Protégé[10].
3. Implementing the translation from Datalog to SQL.
4. The last step towards a working prototype is the aggregation of the results returned from the database. Here, the final probabilities are computed and can then be displayed to the user.

I plan to finish these goals within the next few months, my long term goals for the p.h.d. are stated in the section future work.

5 Future Work

There are 4 directions in which I plan to continue to work after reaching my midterm goals: Extending the framework, benchmarking its performance, and using it during a case study.

- On the one side I will extend the notion of uncertainty to the TBox and allow the incorporation of imprecise information. For this, I will have to adapt the query rewriting algorithm within IQAROS to account for probabilities on axioms. On the other side, I want to implement support for additional constructs of SPARQL, e.g. unions, optional graph patterns, or aggregates.
- Together with the theoretical work on the handling of uncertainty in the reasoning step, it is also interesting to investigate the influence of the closed or open world assumption on the query results.
- Furthermore, I want to test my framework for performance and scalability. One possibility for this is to use one of the established benchmark frameworks for Semantic Web reasoning, e.g. LUBM[11], SP^{2}Bench[12], or BSBM[13]. Those benchmarks all use generators to synthetically generate knowledge bases. These generators could be adapted to also create imprecise information.

[10] http://protege.stanford.edu/
[11] http://swat.cse.lehigh.edu/projects/lubm/
[12] http://dbis.informatik.uni-freiburg.de/forschung/projekte/SP2B/
[13] http://wifo5-03.informatik.uni-mannheim.de/bizer/berlinsparqlbenchmark/

- Additionally to synthetic benchmarks, I want to apply my framework to riskmanagement as an alternative approach to markov logic networks, which we are currently using in a project concerned with this topic [vSOS14].
- I will further distinguish my work from the approaches used in probabilistic deductive databases. A deeper analysis of the work of Hernich et al. [HKLG13] about the relation of Datalog and description logics is also needed.

Acknowledgements

I want to thank my supervisor Christian Meilicke for his support and the fruitful discussions about the planned framework and my approach. Additionally, I also want to thank the reviewers for their valuable comments.

References

[CDL05] Diego Calvanese, Giuseppe De Giacomo, and Domenico Lembo. DL-Lite: Tractable description logics for ontologies. *AAAI*, 5:602–607, 2005.

[CDL+11] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, and Domenico Fabio Savo. The Mastro system for ontology-based data access. *Semantic Web*, 2:43–53, 2011.

[GW09] Nicola Guarino and Christopher A. Welty. An overview of OntoClean. *Handbook on ontologies*, pages 1–20, 2009.

[HAKO09] Jiewen Huang, Lyublena Antova, Christoph Koch, and Dan Olteanu. MayBMS: a probabilistic database management system. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 1071–1074. ACM Press, 2009.

[HKLG13] André Hernich, Clemens Kupke, Thomas Lukasiewicz, and Georg Gottlob. Well-founded semantics for extended datalog and ontological reasoning. In *Proceedings of the 32nd symposium on Principles of database systems*, pages 225–236, New York, New York, USA, 2013. ACM Press.

[Jen96] Finn V. Jensen. *An introduction to Bayesian networks*, volume 210. UCL press London, 1996.

[JS12] Abhay Jha and Dan Suciu. Probabilistic databases with MarkoViews. *Proceedings of the VLDB Endowment*, 5(11):1160–1171, 2012.

[KBP+12] Aditya Kalyanpur, Branimir K. Boguraev, Siddharth Patwardhan, J. William Murdock, Adam Lally, Christopher A. Welty, John M. Prager, Bonaventura Coppola, Achille Fokoue-Nkoutche, Lei Zhang, Yue Pan, and Zhao Ming Qui. Structured data and inference in DeepQA. *IBM Journal of Research and Development*, 56(3):351—-364, 2012.

[KPHS07] Aditya Kalyanpur, Bijan Parsia, Matthew Horridge, and Evren Sirin. Finding all justifications of OWL DL entailments. *The Semantic Web*, pages 267—-280, 2007.

[KTD07] Angelika Kimmig, Hannu Toivonen, and Luc De Raedt. ProbLog: A Probabilistic Prolog and Its Application in Link Discovery. In Rajeev Sangal, Harish Mehta, and R K Bagga, editors, *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, volume 7, pages 2462–2467, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[LS94] Laks V. S. Lakshmanan and Fereidoon Sadri. Probabilistic Deductive Databases. *SLP*, pages 254–268, June 1994.

[PP12] Heiko Paulheim and Jeff Z. Pan. Why the semantic web should become more imprecise. *What will the Semantic Web look like*, 2012.

[RD06] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine Learning*, 62(1-2):107–136, January 2006.

[SO13] Joerg Schoenfisch and Jens Ortmann. YARR!: Yet Another Rewriting Reasoner. *CEUR Workshop Proceedings*, 1015:19–25, 2013.

[SS08] Markus Stocker and Michael Smith. Owlgres: A Scalable OWL Reasoner. *OWLED*, 432, 2008.

[TDD+13] Martin Theobald, Luc De Raedt, Maximilian Dylla, Angelika Kimmig, and Iris Miliaraki. 10 years of probabilistic querying-what next? *Advances in Databases and Information Systems*, pages 1–13, 2013.

[vSOS14] Janno von Stülpnagel, Jens Ortmann, and Joerg Schoenfisch. IT Risk Management with Markov Logic Networks. *Proceeedings of the 26th Conference on Advanced Information Systems Engineering (CAiSE 2014)*, to appear, 2014.

[VSS12] Tassos Venetis, Giorgos Stoilos, and Giorgos Stamou. Incremental query rewriting for OWL 2 QL. *complexity*, 7:5, 2012.

[Wid04] Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. *Technical Report*, pages 1–22, 2004.

[WMGH08] Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakisy, and Joseph M. Hellerstein. BayesStore: Managing large, uncertain data repositories with probabilistic graphical models. In *Proceedings of the VLDB Endowment*, volume 1, pages 340–351, 2008.

Towards Enabling Design Rationale Capture by Asking the Right Questions

Mathias Schubanz

Institut für Informatik, Informations- und Medientechnik
BTU Cottbus - Senftenberg
Postfach 10 13 44
03013 Cottbus
M.Schubanz@b-tu.de

Abstract: Over the last thirty years many research has been conducted to capture the ”how” and ”why” behind design decisions. This information is known as design rationales (DR). Approaches to capture, store, preserve, and use DR have emerged from research activities. However, as of today they only found exceptional application within industrial practice. Rationales have been analysed in respect to its nature, its structure, and its quality. Additionally, some researchers performed analyses on them. They found out that DR have the potential to sustainably contribute to design, re-design, testing, maintenance, and improving product quality over the whole product-life cycle. Within this paper a research proposal is presented striving to tackle the exceptional use of DR in software documentation. We want to promote the capture of DR by paying more attention on the questions ”What information has to be captured by rationales?”, ”How detailed should the captured information be?”, and ”How should rationale capture be integrated into the development process?”. It is the goal to promote the use of DR within the whole software development process. In short, within this paper we introduce the topic of DR, related research and elaborate on the intended research on the topic of DR capture.

1 Introduction

Design rationales document the reasoning behind given actions and decisions. They provide an answer for the ”how” and ”why” which is often asked in relation to decisions made. Therefore, they play an important role in decision-making. For instance, when one needs to decide on one out of a set of available alternatives, they provide insight to the arguments, the criteria considered, and potential consequences. Rationales also document the assumptions behind decisions which often, if not explicitly asked for, stay implicit in the heads of those involved. They are substantially connected to argumentation and sometimes even are used as a synonym for argumentation [FLMM91]. Hence, capturing DR avoids tacit knowledge to remain in the stakeholders' heads.

The concept of documenting the decision-making process and its connected rationales did not originate in the software engineering domain. It has initially been applied in urban architecture and the policy domain [BCMM08]. In the late 80's several approaches to doc-

ument group meetings and argumentation within the field of human-computer interaction design were published [CMW96, CR91, CY91, MYM89]. Based on the corresponding focus of capturing rationales of software and interface designs, the term *design rationale* (*DR*) got well established. Thus, *DR* and *rationale* are treated as a synonym from here on.

Kunz and Rittel [KR70] were one of the first to approach the field of DR. They shaped rationales as *the reasons behind existing decisions*. According to MacLean et al. [MYBM91] proper rationales also include the design space. In other words, design rationales also consider how a given artefact is located in the space of existing design options. Rationales also play an important role in software architecture documentation. Correspondingly, Tang and Han [TH05] defined *Architecture Rationale* to record the reasons of requirements enhancements and its design rationales. However, considering all available definitions of DR would go beyond the scope of this paper. Following, we refer to the definition of MacLean et al. [MYBM91] when talking about DR.

DR are said to support multiple software engineering activities as, for instance, design, re-design, test, and maintenance [DMMP06]. Nevertheless, DR has been employed only exceptionally in industrial practice yet [Bur08]. There is a clear disparity between the its potential and the use of DR. Therefore, this paper presents a research proposal which is intended to result in a DR capture framework providing concrete guidance to developers when capturing rationales. This framework is expected to contain a set of reference questions grouped into templates. These questions shall be answered to capture rationales during software development. The framework is further expected to contain guidelines supporting the integration of rationale capture into existing software development processes. It shall be rounded out by a reference implementation of DR capture in a development process.

The goal of the dissertation is to resolve the uncertainty connected to DR capture (see Section 4) by providing a reference framework which guides developers when capturing DR. The work shall provide answers for the following research questions:

RQ-1: What information has to be captured by rationales?

RQ-2: How detailed should the captured information be?

RQ-3: How should rationale capture be integrated into the development process?

The remainder of the paper is structured as follows: Section 2 motivates the usage of DR. Within Section 3 related work with a focus on software engineering is presented. Section 4 discusses research gaps related to the successful application of DR. Finally, a research outline including previous work is presented in Section 5.

2 Motivation

The essential capability of DR is to reflect the intention behind a decision. One of the most important application scenarios in this respect is the usage of DR as a means of communication. As such they have initially been used in urban architecture and in the political domain [BWMK09]. In the latter, the *IBIS* approach [KR70] by Kunz and Rittel played an

important role. According to Carrol and Rosson [CR91] DR *build a pertinent understanding* of the context, the users, the tasks, the technologies, and the situations considered. This is of particular interest in the software engineering domain as there is a variety of stakeholders participating in, for instance, requirements engineering and design. This ranges from developers, who intend to understand the work of their colleagues, up to the technical sales and distribution department. They use DR to reduce the complexity contained in a software design. In other words, DR essentially *serve the perception and understanding of a given design* [BCMM08, Cas96, CY91, DMMP06, KLVV06, MYBM91, Sag12, SH94]. Conklin and Yakemovic [CY91] underpin these assumptions by a case study. At the NCR corporation they introduced a DR management system in an industrial setting for the scope of a particular project. The authors report several positive effects which can be attributed to captured rationales. At the same time they also came across several typical rationale management problems. For instance, participants did from time to time not capture DR properly. This could be attributed to a diverse perception of *what rationales are* and *when they need to be captured*. Correspondingly, Conklin and Yakemovic report in their case-study description that at one point in the project a particular feature was planned to be realised by the development team without a concrete reason attached to it. The responsible software engineer expected everybody else in the team to be familiar with the reasons for the feature request. In fact, nobody else but him knew, that this particular feature had to be introduced due to a number of customer complaints.

A positive conclusion from the case study is that DR support *knowledge preservation*. They contribute to preserve information in relation to an existing software artefact, like a component. Imagine the following: the lack of information in the previous example would not have been discovered. Even worse, the responsible software engineer would have left the company. At that point the corresponding design knowledge would have been lost. Documenting DR could prevent this. Within the same case study [CY91] the authors report a successful case of knowledge preservation. Prior to the completion of a detailed software design one of the centrally involved software engineers left the company. Nevertheless, on the basis of the DR documentation the remaining team did manage to finish the detailed design without difficulty.

Tang and Han [TH05] claim that DR also play a central role for *architecture verification*. Hence, DR support software designers when they check whether the architecture model is complete, the rationale of the design decisions is sound, and the design can satisfy the given system requirements. This way errors could be revealed. Analogously, during the mentioned case study [CY91] the authors manually converted the documentation repository including its DR from the *itIBIS* to the *gIBIS* format. This conversion revealed *11* additional design problems, whereof *7* most likely would not have been discovered until code would have been written.

DR can also be applied to support several other capabilities. For instance, DR could *support reasoning* on a given set of alternatives [MYBM91]. Hence, DR also *facilitate release planning* [Ruh11]. Alternatively, DR also can *facilitate decisions within a variant space* [TB12]. A detailed overview on capabilities of DR outlined in the literature within the last 25 years is provided by Sagoo [Sag12].

Despite of the various application scenarios in which DR could contribute, they have not

found much application in industrial practice yet. Known applications in the domain of software engineering are mainly based on research activities performing case studies or concept evaluations. They have found more application in the field of aviation industries [BWMK09]. There are several barriers to a successful DR capture as, for instance, the high effort to capture DR, maintain the information, as well as no clear understanding of how they could be handled. Within the set of problems a central role could be attributed to the uncertainty in relation to benefits of DR and a lack of certainty on what information is valuable up to which granularity. More discussion on this topic can be found in Section 4.

3 Related Work

This section will present related work in the field of DR management. The first part discusses existing approaches to represent DR. Section 3.2 provides a brief overview on DR in relation to software architectures. Finally, Section 3.3 will briefly review approaches focussing on the DR capture process.

3.1 Design Rationale Models

Probably the most important approach in the field of rationale management dates back to 1970. The IBIS approach [KR70] defined a structure to be used for documenting issues discussed in, for instance, group meetings. It has been widely used and often adopted within proceeding research activities. Several other also similar approaches have been proposed in the late 80's and early 90's. One of these is the *Questions-Options-Criteria (QOC)* approach by MacLean et al. [MYBM91]. They altered the IBIS approach by adding relevant criteria as well as the opportunity to evaluate arguments. This is why the authors claim to support a design space exploration. Lee [Lee89] proposed another important approach providing an expressive grammar to capture DR in structured way. With its seven node types and fifteen link types it contains a high complexity [BD04].

On the basis of the previously mentioned approaches several similar models and corresponding tools have been developed. For instance, the IBIS derivatives *itIBIS* [CY91], *gIBIS* [CY91], *rIBIS* [RE91], *PHI* [FLMM91], *DRed* [BWMK09], and *QAR* [HOK14] were proposed. Some derivatives of QOC are *DREAM+TEAM* [LP07], *IVMM* [TB12], *EvoPL* [SPBL12], and *RUSE* [Wol08]. A derivative of DRL is *RATspeak* [BB08].

3.2 Software Architecture Documentation

Among the numerous approaches to DR some had a dedicated focus on software architecture. For instance, Tang and Han [TH05] proposed the *Architecture Rationalization Method (ARM)* which is based on *Architecture Rationale*. The *Architecture Tradeoff Analysis Method (ATAM)* by Kazman et al. [KKC00] evaluates multiple software architectures

against a set of defined quality goals. Its output is intended to develop a set of analyses, rationales, and guidelines. Asundi et al. [AKK01] extended *ATAM* by the *Cost Benefit Analysis Method (CBAM)* which links costs and business goals to architecture decision-making. Bass et al. [BCNS06] propose to capture DR for a software architecture differently. They use to two distinct graphs, a *structural graph* and a *causal graph*.

Another aspect which relates to DR is the documentation of software architectures and the connected reasoning. Here, several models have been proposed. *Tyree's Template* by Tyree and Akerman [TA05], *Archium* by Jansen and Bosch [JB05], the *Architecture Design Decision Support System Model* by Capilla et al. [CND07], or *AREL* by Tang et al. [THV09] are just some to name in this respect.

3.3 Design Rationale Capture

For a long time *DR capture* did not receive much attention. The DR schema was the point research focussed on. Over time it became clear that the schema is not the only key to success. Hence, corresponding tools and methodologies emerged from research activities. One of the first approaches was the *graphical IBIS (gIBIS)* approach [CY91]. It tried to mitigate existing barriers to DR capture by graphical tool support. A similar approach called *MIKROPLIS* and later *PHIDIAS* was presented by McCall et al. [MBd+90]. They used the PHI approach which is based on IBIS [FLMM91]. Its essential advantage is that DR could be created by reusing already captured information from previous projects.

Shipman et al. [SIM99] went another way by proposing to capture DR without a schema. The approach captures all information on a 2D space and structures the results according to its spatial arrangement. Similarly, Reeves et al. [RS92] annotate artefact designs with textual notes within a CAD system. Thus, DR are annotated directly to the artefact design. Myers [MZG99] modified a CAD system by augmenting the symbol library with semantic information. Hence, DR is captured based on the symbol usage. The main goal of the previous approaches is to reduce the process disruptiveness by avoiding a DR schema.

On a rather abstract level Schneider [Sch06] approached the problem by proposing guidelines to DR capture which need to be turned into concrete operational practices, techniques, and/or rules. Apart from the underlying DR schema Bass et al. [BCNS06] also propose guidelines to DR capture. Their goal is to guide architects during DR capture. In contrast thereto, Casaday [Cas96] provides a set of eight templates residing on a more detailed abstraction level. In a more recent work Durdik and Reussner [DR13] capture design decisions related to the use of patterns by using a question catalogue.

4 Research Gaps

As outlined in Section 3 a considerable set of research has been conducted on the topic of DR. However, there still remains much uncertainty [Bur08] on DR usage. DR capture is accompanied by a contradiction. While there is no shortage of advocates for the value of

DR, there also is no shortage of reasons why DR are unlikely to be captured [Bur08]. There is a big uncertainty whether the payoff will justify the costs, how DR should be maintained, how it will we used, and most important, what needs to be captured. This uncertainty hindered the successful integration of DR into industrial practice. For now, there are two known cases to be mentioned when talking about successful DR capture. As outlined in Section 2 the first one is the successful application of DR during the KDS project in the NCR corporation [CY91] on the basis of the *gIBIS tool*. The second case is the successful application of the *Design Rationale Editor (DRed)* within Rolls Royce [BWMK09]. What initially started as a research project ended with a company-wide rollout of the *DRed* approach.

When comparing the lack of industrial DR usage to its occasional success the following question arises: *Why did DR not find industrial application on a broad scale yet?* Kruchten et al. [KLVV06] simply answer this by stating the following: *This is mostly because the burden to capture assumptions and decisions outweighs largely the immediate benefits that the architect may draw.* Correspondingly, one of the main reasons is the uncertainty on the usage of DR. Burge [Bur08] outlined this in three questions:

- How useful are the potential benefits of DR?
- How insurmountable are barriers to DR usage?
- Will the value of DR justify the cost?

Within future research these questions need to be answered to overcome the *barriers to DR usage*. A first contribution in this area has been proposed by Horner and Atwood [HA06]. They categorised the *barriers to DR usage* into *cognitive*, *capture*, *retrieval*, and *usage limitations*. *Cognitive limitations* consider limitations connected to human information processing. *Capture limitations* address the capture of information required to place DR in the right context, incentives to DR capture, the elicitation of tacit knowledge, political factors (Does the captured DR pose a risk to the developer or the company?), as well as the trade-off between cost and benefits of DR capture. *Retrieval limitations* are concerned about the information which needs to be captured and which is expected to be useful. Finally, the DR capture might constrained by *usage limitations* as, for instance, the uniqueness of design problems. Considering a development scenario where the design problems always differ substantially, DR re-use is an undesirable application scenario.

Another frequently discussed issue in connection to the uncertainty of DR capture is the process intrusiveness [DMMP06, Sch06]. Correspondingly, DR capture is considered to break the workflow, slow down the work progress and reduce the developers' motivation. A reduced motivation could also be caused once architects need to capture DR which reveal that they did not act as they were supposed to [DMMP06]. For instance, they decided to prefer a solution alternative which is connected to less development effort but contains a higher risk of failure. Another reason which often hinders DR capture is a lack of project resources and available time [DMMP06]. When things need to be cut due to a lack of time DR does not have the appropriate priority to be maintained. This is also due to a lack of immediate pay-off compared to the required effort [FLMM91]. Usually, the point in time when DR unfold their benefit is posterior to the time of capture [CY91].

Apart from the contribution proposed by Horner and Atwood, there is little discussion

on the problems outlined above. For now, DR research mainly focussed on DR schemes and corresponding tooling. At least as important as the schema / tool combination is an investigation on the expectations and needs of practitioners. These are a key factor to overcome the *barriers to DR usage*. Thus, this area should gain more focus in DR research. Otherwise it will remain as it has done over the years: nice ideas, but not practical [KLVV06]. First contributions have been made here by Burge [Bur08] and Tang et al. [TBGH06]. Both conducted a survey on the needs and expectations of practitioners to DR usage. They found out that practitioners consider DR as important. Additionally, practitioners criticise a lack of methodology and tool support. However, there barely is empirical work in DR research [Bur08, TBGH06]. More effort needs to be invested here.

5 Project Outline

Considering the elaborations in the previous chapters *Design Rationale (DR)* do seem to be an important means of software documentation. They capture the knowledge behind design decisions and, therefore, capture a key asset of a company's value [Lie02]. However, as of today to the best of our knowledge DR are barely used in industry in general and in software engineering in particular. Potential reasons to that have been outlined in Section 4. To overcome exactly these research gaps an essential starting point is to provide practitioners with concrete guidance on what information is relevant to be captured. Accordingly, the research questions *RQ1* to *RQ3* have been set up (cf. Section 1). To answer these questions it is intended to develop a DR capture reference framework. It is expected to contain (1) a set of concrete questions to be answered by DR, (2) guidelines to support the integration of DR capture into existing software development processes, and (3) a reference implementation of a process for a DR capture system. The questions and guidelines are intended to be grouped together into a set of reference templates. Based on a development domain, the granularity level of documentation, and the development conditions (development team, development process, development resources) a developer will need to tailor the resulting DR capture reference templates to his/her concrete development scenario. Additional guidance will be provided by the reference implementation.

5.1 Previous Work

In our previous work we augmented an approach to model *software product line evolution (SPLE)*, called *EvoPL* [PBD+11], with the basic concepts from decision-making and rationale modelling. Based on the *QOC* model ([MYBM91] , see Section 3) we support the capture of goals, requirements, decision alternatives, relevant criteria, as well as explicit DR (see [SPBL12]). Further, we provided a first IDE-integrated tool support on prototype level. In [SPP+13] we evaluated our model from [SPBL12]. To do so, we analysed the *SWT* project, a sub-project of the *Eclipse Project* [DRW04]. We retrospectively analysed the documentation of their coarse grained plans called *Themes and Priorities* tracing actions down to fine-grained commit comments. The results showed, that the introduced

modelling concepts (cf. [SPBL12]) haven reflected by analysed data. However, we were not able to systematically reconstruct explicit DR information.

Based on experiences from this research and the research gaps discussed in Section 4 the research focus has been slightly adjusted. One of the central problems which we intend to address now are the *barriers of DR usage*. As a first step to approach this problem in [SPJB14] we conducted a literature survey looking for questions to be answered by DR. Additionally, we started to work on a combination of rationale capture guidelines and the questions from [SPJB14] for the requirements engineering domain. It is intended to propose a first approach on questions-based rationale capture guidance.

5.2 Research Outline

The results from the literature survey in [SPJB14] and current research will serve as a first step to set up an initial set of DR questions and guidelines. However, the work needs to be extended. A research outline continuing this work is listed in the following steps:

1. The literature survey on questions to be answered by DR (see Section 4) needs to be enlarged to a *Systematic Literature Review* [KC07]. The goal is to gain a better understanding on the expectations outlined in rationale research.
2. The information found in the literature needs to be related to information on the expectations and needs to DR usage from the practitioner's point of view.
3. Based on the results from *step one* and *two* an initial set of question templates has to be set up. These shall serve as the basis for first guidelines to capture rationales.
4. The results from *step three* need to be tailored to a subset of domains and developments processes. Moreover, a reference implementation has to be developed.
5. The results from *step four* need to be refined in cooperation with practitioners. Finally, the usability of the defined of the DR capture framework needs to be empirically evaluated in an industrial project.

In the *first step* we want to focus on one of the central capabilities awarded to DR. They are expected to answer a given set of questions [BCNS06, BCMM08, DMMP06]. This concept can also be found in the QOC [MYBM91] and the QAR [HOK14] approach. "Why was this decision made?", "How is requirement X satisfied?", or "What are the implications of making modification Y?" are just some examples. Some of the central questions, as those just mentioned, are outlined in the literature again and again. Others, only appear seldom. Hence, an essential starting point to focus the research on is to determine the information which is required to be captured by DR. To do so, it is intended to perform a *Structured Literature Review* [KC07] on the availability of questions to be answered by DR in the corresponding literature. As outlined in Section 5.1 a first literature review has been performed already. However, this has several limitations. It needs to be enlarged to a *Structured Literature Review* considering more background information on the questions found. Thus, it is important to consider the context the questions were used in. They might originate from a theoretical background used for, e.g., motivation purposes in research papers. Alternatively, they could originate from a practical background as, e.g.,

they emerged during a case study. The results of the literature review need to be analysed in detail. Potential research questions for the literature review could be the following:

RQ-A What questions to be answered by DR are outlined in the literature?

RQ-B Which background do the questions found originate from?

RQ-C What differences can be found based on the origin of the questions?

RQ-D How have the questions found been evaluated?

RQ-E What kind of question-based guidance for DR capture has been proposed yet?

Within the *second step* we intend to prepare questionnaires and perform interviews with practitioners having different experiences and working in different domains. In detail, we want to communicate with practitioners (1) still being in education (e.g. students), (2) working in open source development projects, and (3) having industrial background. We intend to interview them using online questionnaires and/or live interviews. Therefore, we intend to find an answer to the following questions:

- What do practitioners expect of using DR? (in terms of effort, benefits, and the trade-off amongst them)
- What are the barriers to DR usage for practitioners?
- Which are the most important ones to overcome?
- What information do practitioners require when analysing, maintaining respectively evolving existing (legacy) software?
- What is the most relevant information for practitioners to be captured by DR?

In addition to the results from *step two*, we intend to analyse existing open source development projects for its documentation structure and DR usage. Based on the unstructured, decentralised development processes as well as the different information backgrounds of the developers, we expect that DR plays an important role to improve the communication within open source development projects. We want to exhibit if any DR related information is captured. If so, what information does the DR documentation reveal about DR capture, DR usage and the project itself? Furthermore, it is of interest to us, what information is often required by developers analysing existing software, performing maintenance operations, or evolving existing software.

In *step three* we want to produce initial set of question templates based on the results from *step one* and *two*. The resulting templates shall serve as the basis for a first version of guidelines to capture rationales. However, as we expect the that the questions found will reside on a rather general level we need to refine these.

The *fourth step* is intended to review the results from *step one* and *two* in respect to the domains the practitioners are working in. We think that there is no *one fits it all approach*. Based on the research proposed above, we will not be able to provide guidance for all development scenarios. Hence, the intended reference framework needs to be tailored to a subset of (1) domains, (2) team structures, and (3) development processes employed. This shall be done in collaboration with industry partners. If this is not done carefully, DR questions will remain on an abstract level. Hence, they would merely serve as a

guideline for DR capture (similar to [BCNS06]) instead of concrete guidance to DR capture. To ensure the usability and the functional relevance of the intended outcome we intend to closely collaborate with practitioners from an industrial background. Software engineering in regulated domains, including certification-oriented processes seems to be a promising candidate. The need for careful documentation is already well established and first successful industry cases exist (e.g. [BWMK09]). We also got into contact with first practitioners from the automotive domain. Step *four* will be concluded by the creation of a reference implementation of a process for a DR capture system.

Within *step five* we refine the results from *steps one* to *four*. Again, we strive to have this activity to be supported by an industry collaboration from the automotive industry. An empirical evaluation within the automotive domain (see *step four*) on the usability of the outcomes shall be performed. The setting and scope of the evaluation is not clear yet.

Finally, within the thesis it is intended to provide tool-support for the ideas from *steps one* to *five*. The DR capture framework should be accompanied by an automated configuration wizard to increase its usability. Moreover, it is intended to develop an IDE-integrated tool support, similar to the one contained in SEURAT [BB08]. DR should be captured within the development environment depending on the concrete development process and usage scenario.

References

[AKK01] Jayatirtha Asundi, Rick Kazman, and Mark Klein. Using economic considerations to choose among architecture design alternatives. Technical report, DTIC Document, 2001.

[BB08] Janet E. Burge and David C. Brown. Software Engineering Using RATionale. *Journal of Systems and Software*, 81(3):395–413, 2008.

[BCMM08] Janet E Burge, John M Carroll, Raymond McCall, and Ivan Mistrk. *Rationale-Based Software Engineering*. Springer Publishing Company, Incorporated, 2008.

[BCNS06] Len Bass, Paul Clements, Robert L Nord, and Judith A Stafford. Capturing and Using Rationale for a Software Architecture. In *Rationale Management in Software Engineering*, pages 255–272. Springer, 2006.

[BD04] Bernd Bruegge and Allen H Dutoit. *Object-Oriented Software Engineering Using UML, Patterns and Java-(Required)*. Prentice Hall, 2004.

[Bur08] Janet E Burge. Design Rationale: Researching Under Uncertainty. *AI EDAM*, 22(4):311, 2008.

[BWMK09] Rob Bracewell, Ken Wallace, Michael Moss, and David Knott. Capturing design rationale. *Computer-Aided Design*, 41(3):173 – 186, 2009.

[Cas96] George Casaday. Rationale in Practice: Templates for Capturing and Applying Design Experience. In *Design Rationale: Concepts, Techniques, and Use*, pages 351–372. Lawrence Erlbaum Associates, 1996.

[CMW96] Tom Carey, Diane McKerlie, and James Wilson. HCI Design Rationale as a Learning Resource. In *Design Rationale: Concepts, Techniques, and Use*, pages 373–392. Lawrence Erlbaum Associates, 1996.

[CND07] Rafael Capilla, Francisco Nava, and Juan C Duenas. Modeling and documenting the evolution of architectural design decisions. In *Proceedings of the Second Workshop on*

SHAring and Reusing architectural Knowledge Architecture, Rationale, and Design Intent, page 9. IEEE Computer Society, 2007.

[CR91] John M Carroll and Mary Beth Rosson. Deliberated evolution: Stalking the View Matcher in design space. *Human–Computer Interaction*, 6(3-4):281–318, 1991.

[CY91] E. Jeffrey Conklin and K. C. Burgess Yakemovic. A Process-Oriented Approach to Design Rationale. *Human–Computer Interaction*, 6:357–391, Sept. 1991.

[DMMP06] Allen H. Dutoit, Raymond McCall, Ivan Mistrik, and Barbara Paech. *Rationale Management in Software Engineering*. Springer-Verlag, New York, 2006.

[DR13] Zoya Durdik and Ralf Reussner. On the Appropriate Rationale for Using Design Patterns and Pattern Documentation. In *Proceedings of the 9th International ACM Sigsoft Conference on Quality of Software Architectures*, QoSA '13, pages 107–116, New York, NY, USA, 2013. ACM.

[DRW04] J. Des Rivières and J. Wiegand. Eclipse: a platform for integrating development tools. *IBM Syst. J.*, 43:371–383, April 2004.

[FLMM91] Gerhard Fischer, Andreas C. Lemke, Raymond McCall, and Anders I. Morch. Making Argumentation Serve Design. *Hum.-Comput. Interact.*, 6:393–419, September 1991.

[HA06] John Horner and Michael E Atwood. Effective Design Rationale: Understanding the Barriers. In *Rationale Management in Software Engineering*, pages 73–90. Springer, 2006.

[HOK14] Alireza Haghighatkhah and Harri Oinas-Kukkonen. An Argumentation-Based Design Rationale Application For Reflective Practice. In *Proceedings of the 22nd European Conference on Information Systems (ECIS)*, 2014.

[JB05] Anton Jansen and Jan Bosch. Software architecture as a set of architectural design decisions. In *Software Architecture, 2005. WICSA 2005. 5th Working IEEE/IFIP Conference on*, pages 109–120. IEEE, 2005.

[KC07] Barbara A. Kitchenham and S. Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE-2007-01, Keele University, 2007.

[KKC00] Rick Kazman, Mark Klein, and Paul Clements. ATAM: Method for Architecture Evaluation. Technical report, CMU / Software Engineering Institute, 2000.

[KLVV06] Philippe Kruchten, Patricia Lago, and Hans Van Vliet. Building up and reasoning about architectural knowledge. In *Quality of Software Architectures*, pages 43–58. Springer, 2006.

[KR70] Werner Kunz and Horst W. J. Rittel. Issues as elements of information systems. Technical report, Systemforschung, Heidelberg, Germany Science Design, University of California, Berkeley, 1970.

[Lee89] Jintae Lee. Decision Representation Language (DRL) and Its Support Environment. MIT Artificial Intelligence Laboratory Working Papers WP-325, MIT Artificial Intelligence Laboratory, August 1989.

[Lie02] Jay Liebowitz. A look at NASA Goddard Space Flight Center's knowledge management initiatives. *IEEE software*, 19(3):40–42, 2002.

[LP07] Xavier Lacaze and Philippe Palanque. DREAM & TEAM: A Tool and a Notation Supporting Exploration of Options and Traceability of Choices for Safety Critical Interactive Systems. In *Human-Computer Interaction – INTERACT 2007*, volume 4663, pages 525–540. Springer Berlin Heidelberg, 2007.

[MBd+90] Raymond McCall, Patrick R Bennett, Peter S d'Oronzio, Jonathan L Ostwald, Frank M Shipman III, and Nathan Wallace. PHIDIAS: Integrating CAD Graphics into Dynamic Hypertext. In *ECHT*, volume 90, pages 152–165, 1990.

[MYBM91] Allan MacLean, Richard M. Young, Victoria M. E. Bellotti, and Thomas P. Moran. Questions, Options, and Criteria: Elements of Design Space Analysis. *Hum.-Comput. Interact.*, 6:201–250, September 1991.

[MYM89] Allan MacLean, Richard M Young, and Thomas P Moran. Design Rationale: the Argument behind the Artifact. In *ACM SIGCHI Bulletin*, volume 20, pages 247–252. ACM, 1989.

[MZG99] Karen L Myers, Nina B Zumel, and Pablo Garcia. Automated capture of rationale for the detailed design process. In *AAAI/IAAI*, pages 876–883, 1999.

[PBD+11] Andreas Pleuss, Goetz Botterweck, Deepak Dhungana, Andreas Polzer, and Stefan Kowalewski. Model-driven Support for Product Line Evolution on Feature Level (accepted for publication). *Journal of Systems and Software (JSS) - Special Issue on "Automated Software Evolution"*, 2011. http://soft.vub.ac.be/iwpse-evol/specialissue.

[RE91] Gail L Rein and Clarence A Ellis. rIBIS: a real-time group hypertext system. *International Journal of Man-Machine Studies*, 34(3):349–367, 1991.

[RS92] Brent Reeves and Frank Shipman. Supporting communication between designers with artifact-centered evolving information spaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 394–401. ACM, 1992.

[Ruh11] Günther Ruhe. *Product Release Planning: Methods, Tools and Applications*. Auerbach Publications, 2011.

[Sag12] Jeevan Sagoo. *Design rationale for the regulatory approval of medical devices*. PhD thesis, Cranfield University, 2012.

[Sch06] Kurt Schneider. Rationale as a By-Product. In *Rationale Management in Software Engineering*, pages 91–109. Springer, 2006.

[SH94] Simon Buckingham Shum and Nick Hammond. Argumentation-Based Design Rationale: What Use at What Cost. *International Journal of Human-Computer Studies*, 40(4):603–652, 1994.

[SIM99] Frank M Shipman III and Catherine C Marshall. Spatial hypertext: an alternative to navigational and semantic links. *ACM Computing Surveys*, 31:14, 1999.

[SPBL12] Mathias Schubanz, Andreas Pleuss, Goetz Botterweck, and Claus Lewerentz. Modeling Rationale over Time to support Product Line Evolution Planning. In *Proceedings of VaMoS'12*, pages 193–199, Leipzig, Germany, 2012. ACM.

[SPJB14] Mathias Schubanz, Andreas Pleuss, Howell Jordan, and Goetz Botterweck. Guidance for Design Rationale Capture to Support Software Evolution. In *Workshop on Software-Reengineering & Evolution*, Bad Honnef, 2014.

[SPP+13] Mathias Schubanz, Andreas Pleuss, Ligaj Pradhan, Goetz Botterweck, and Anil Kumar Thurimella. Model-driven planning and monitoring of long-term software product line evolution. In *Proceedings of VaMoS'13*, pages 103–107, Pisa, Italy, 2013. ACM.

[TA05] Jeff Tyree and Art Akerman. Architecture decisions: Demystifying architecture. *IEEE software*, 22(2):19–27, 2005.

[TB12] Anil Kumar Thurimella and Bernd Bruegge. Issue-based variability management. *Information and Software Technology*, 54(9):933–950, 2012.

[TBGH06] Antony Tang, Muhammad Ali Babar, Ian Gorton, and Jun Han. A Survey of Architecture Design Rationale. *Journal of Systems and Software*, 79(12):1792–1804, 2006.

[TH05] Antony Tang and Jun Han. Architecture Rationalization: A Methodology for Architecture Verifiability, Traceability and Completeness. In *ECBS*, pages 135–144, 2005.

[THV09] A. Tang, J. Han, and R. Vasa. Software Architecture Design Reasoning: A Case for Improved Methodology Support. *IEEE Software*, 2009(March/April):43–49, 2009.

[Wol08] Timo Wolf. *Rationale-based Unified Software Engineering Model*. VDM Verlag, Saarbrücken, Germany, 2008.

Towards Usability Mining

Tom Thaler

Institute for Information Systems (IWi) at the
German Research Center for Artificial Intelligence (DFKI) and Saarland University
Campus, 66123 Saarbrücken
tom.thaler@iwi.dfki.de

Abstract: The usability of business information systems (BIS) plays an important role in the selection of corresponding software products and is therefore a crucial characteristic of differentiation. Nevertheless, there are no automatic techniques considering the process aspects of these systems in an adequate manner. Hence, the paper at hand aims at the automatic derivation of software usage reference models, which serve as a basis for measuring the usability of business information systems and its supported process. Apart from conventional event logs, further data, like element positions or sensor data, should also be taken into account. Analyzing these data in the context of business process usability and software usability bases on established techniques from the fields of business process analysis and usability engineering. The resulting method "Usability Mining" aims at the target-oriented design of BIS and an adequate technical support for business processes.

1 Introduction

The usability of business information systems (e.g. ERP, workflow, systems or even particular business software as well as production control systems) plays an important role in the selection of corresponding software products. It is defined as the extent to which a software product can be used for the effective, efficient and satisfactory achievement of business objectives by a user [IS98]. Thereby, business information systems have special characteristics in terms of process orientation since they intend the technical support of concrete business processes. Hence, usability should, in that context, be understood as the extent to which a business information system can be used for the effective, efficient and satisfactory execution of business processes.

Indeed, the supported business processes deliver a basic structure for the software and user interface design, but, at the same time, the practical application can lead to inefficiency, ineffectiveness or disaffection, e.g. in terms of stress. The reasons for that can be quite divergent. Missing or bad positioned elements as well as long loading times make a process more time consuming than necessary and lead to uncertainty and impatience at the user side. A high process complexity impedes the process execution and negatively affects the process efficiency.

Against that background, the paper at hand aims at developing a method analyzing different dimensions of usability, whereby both, the system design and the process

design, are explicitly addressed. Established process mining techniques are used for the automatical derivation of usage models which can then be enriched by manifold data like GUI information (e.g. element positions), sensor data (e.g. process data from cyber-physical systems or health data like pulse or body temperature) or additional user data (e.g. user experience). This renders it possible to analyze the real user behavior in detail and allows the target-oriented improvement of process and software design. The whole method is called "Usability Mining" and allows the consideration of different application scenarios, such as- the following:

Identification of Customer Needs. Sometimes users apply a software for processes or activities, which are not intended by the producer. Mining usage models of a software helps to identify these processes or activities and allows the further development of the software with respect to the actual customer needs (e.g. [TPL13]).

Measurement of Software and Process Usability. Software and process usability are closely connected to each other as the design of a process directly affects the implementation and vice versa. Usage models and their analysis, e.g. in terms of specific metrics, make it possible to assess the usability of processes and their implementation and can lead to target-oriented optimizations.

Controlling the Software Evolution. Further development of a software is in the nature of a product lifecycle. Nevertheless, it is challenging to evaluate whether a further development leads to the desired effect and whether it is used as it is intended. This affects both new supported processes and adapted existing processes. Since usability mining analyzes the real user behavior, it is possible to follow the software evolution from the user side. This can also be seen in [TPL13].

Inductive Usage Reference Model Development. Information about the process performance, the resource consumption and other collected data allow the inductive development of reference models with best practice (as known) character. Based on the process instances, a process model could be derived concerning different objectives like minimization of cost or resources, reduction of stress or optimization of quality.

After this introduction, section 2 describes the research methodology. Section 3 gives an overview of the related work in the relevant research fields, while section 4 presents the usability mining lifecycle with a description of the different phases and an attendant running example. Based on that lifecycle, section 5 concludes the need for further development and summarizes some particular areas of work.

2 Methodology

As the method combines different research fields from two different research disciplines, information systems (especially process mining) and software engineering (especially human computer interaction and usability engineering), it is necessary to identify relevant literature from all of the involved fields. After that, a phase model containing the different steps of usability mining (depending on the application scenario and based

on the procedure of process mining) is elaborated. The identified methods and techniques will be analyzed with respect to their applicability in the context of usability mining which results in a collection of partial solutions for specific problems and a collection of gaps. In order to fill these gaps, a design science research approach is applied [HMP04]. Thus, it is planned to apply an integrated evaluation in the following real application scenarios:

- ARIS: The ARIS Designer is a modelling platform from Software AG which is available both as a native and a web application. It is planned to analyze the process of process modelling in the context of an experiment.
- MIRROR: The research prototype MIRROR (web application) covers organizational learning processes and is used in different hospitals. It is planned to analyze these organizational learning processes in the whole and over all involved hospitals.
- DFKI: The German Research Center for Artificial Intelligence hosts various organizational web-based software e.g. for time recording, acquisition or leave requests. It is planned to analyze the travel management processes in operation.

3 Related Work

As mentioned above, in the context at hand, especially the research field of usability engineering as well as information system research in general and process mining in particular are of major importance.

Traditional methods measuring the usability of software are affected by manual and empirical approaches, whereby, in most cases, questionnaires, user interviews and observations or expert interviews are applied [Ni93]. In fact, methods like eye-tracking or click-path-analyses are partially automatable and literature concerning the application and design of such controlled experiments in terms of using adequate test methods with a meaningful configuration (e.g. sample size) is available [KLS09]. However, the environment settings (e.g. laboratory) as well as the data analyses are very expensive in most cases.

Above all, there are some works using log data as a basis for the automated analysis of software usability (e.g. [IH01, SDK06, TK05]), which also contain approaches using event patterns in order to measure specific aspects of usability (e.g. [Ho06, Sh08]). Isolated works [HR99, SE91] also derive process models (petri-nets) and address some possibilities of usability analysis. However, the used mining methods are rather rudimentary as, nowadays, aspects of process mining, like dealing with noise or a harmonization of log data of different systems, are not taken into account. In fact, the used mining techniques cover the beginning of the process mining era, e.g. [AW05, CW98], and a further consideration of current methods and techniques from the information systems research is missing.

Nevertheless, process-orientation is a core characteristic of business information systems as products like ERP or workflow systems support concrete business tasks and processes

in a technical manner. Therefore, the application scenario of usability is of high importance for the information systems research and for the design of business information systems. Current approaches, e.g. genetic algorithms [Ka06] or cluster techniques handling noisy data or avoiding spaghetti-like models [Aa11] could be helpful in that context. Not till then, it is possible to derive meaningful models or meta information enabling researchers or practitioners to draw concrete conclusions in reference to usability aspects. Especially a combination of different process or model metrics from different research disciplines – like those from [Me12] in the context of business process analysis and from [BGT05] or [IH01] in the context of usability engineering and usage analysis – might improve the evaluation of business processes and their implementation. In contrast to existing methods of usability engineering, there are also further application scenarios like the automatic derivation and further development of software reference models in general and usage models in particular [Ka07].

Indeed, both research fields address similar approaches towards an automatic derivation of process models, but the current states of research strongly diverge. A transfer, adaption and further development of both states of research will imply an enrichment of the respectively other discipline.

4 Usability Mining Lifecycle

In order to realize a concept for usability mining, the author developed a six phase lifecycle consisting of (1) user monitoring, (2) trace clustering, (3) usage model derivation, (4) usage model analysis, (5) recommendation derivation and (6) implementation. All phases are described in the following subsections with a description of input, content (possible techniques) and output. In order to exemplarily demonstrate the application of the lifecycle, a running example is being executed through the lifecycle. The example is based on a real business scenario of a translation service in the web, a detailed process description can be found in [TPL13].

4.1 User Monitoring

Process execution data (instance data) are the basis for usability mining. Depending on the analysis objectives, there are different requirements for the log data. If, for example, the intention is to develop a usage reference model in an inductive manner, it is satisfactory to fulfill the traditional log data requirements of process mining (case, task, originator, timestamp) [Aa08], while in case of an identification of usability weak points, it might be helpful to gather further information like element positions or case specific data. Collecting further information may imply the use of further data sources like an enterprise database, external services or sensors measuring vital parameters. Since software-as-a-service plays an increasing role in the business context, (web-)server logs or error logs are supposable as well, which are traditionally not considered in the context of process mining. Against that background, one needs to design a logging strategy based on the analysis objectives or the application scenario and implement it in the

addressed software. Furthermore, a consolidation of different data sources is of high importance.

Input: User interaction

Content: Logging execution data [Aa11], enhance log data with further information (e.g. from enterprise database) [JJ13] if necessary, (web-)server protocols

Output: Log data (event logs)

Running example: As the log data from the running example were already collected, there is no possibility of gathering further information afterwards. The structure corresponds to the traditional process mining requirements. Thus, there is information about the case, the task, the originator (anonymized) and the timestamp. In addition, there is also information about release changes. Thus, the log data are enriched by the system version (Table 1). Apart from the available data, in the context at hand, it would be meaningful to log some further data attributes like the language combination, the size of the customer's company, etc. In this specific scenario, these data would be of major interest in terms of the clustering of the log data.

Table 1: Sample log file of running example

case_id	task	originator	timestamp	system_ver
5812	create translation case	Customer	16.06.2011 09:01	2
5812	request calculation	Customer	16.06.2011 09:04	2
5812	offer calculation	Controller	16.06.2011 10:11	2
5812	cancel case	Controller	02.08.2011 09:16	3
5813	create translation case	Customer	16.06.2011 09:05	3
5813	order translation	Customer	16.06.2011 09:05	3
5813	offer translation	Controller	16.06.2011 09:13	3
5813	accept offer	Translator	16.06.2011 12:41	3
5813	do translation	Controller	17.06.2011 15:31	3
5813	deliver translation	Controller	17.06.2011 15:31	3

4.2 Trace Clustering

Trace clustering describes the task of clustering traces within log data concerning a specific cluster criterion and is a traditional challenge in the context of process mining. As, in general, business information systems cover a multitude of different business processes, a corresponding log file covers all these processes, too. Discovering a process model based on a non-clustered log file leads to a highly complex and not human readable model in most cases (so called spaghetti-like models). This makes it necessary to identify different processes or instance classes in order to generate several process models with less complexity or similar characteristics (e.g. [EDG13, JA10]). Depending on the objectives of the application scenario, there are also further (additional) aspects, which may serve as a criterion for trace clustering:

- Users: e.g. experience, age, groups
- Software: e.g. version, device
- Processes: e.g. variants, patterns, occurrence of loops or tasks
- Resources/Performance: e.g. time, budget, hardware, load values
- Cases: e.g. value of a shopping cart

Input: Log data (e.g. log file)

Content: Particular trace clustering techniques depending on the objective, e.g. [JA09, MGG08, SA08]

Output: Clustered log data

Running example: Since the log data of the running example contains several system version which differ in process, first of all, the log is clustered by the system versions.

4.3 Usage Model Derivation

Process mining distinguishes three different fields: (1) process discovery, (2) conformance checking and (3) enhancement [Aa11]. Process discovery aims at deriving a new process model solely based on log data, while conformance checking addresses the comparison of the as-is process to the to-be process. Enhancement focuses the derivation of new information from log data and annotating them to an existing process model.

Against that background, all of the mentioned fields play an important role for usability mining in general and in the phase of usage model derivation in particular. In that phase, one needs to derive a process model based on log data and to enrich it with further information like performance data, execution probabilities, correlation matrices and further (scenario specific) data and metrics. Today, a lot of different process mining techniques with different characteristics of the output models do already exists. For example, they differ in the context of detailedness (to what extent do the resulting models cover the trace contained in the log file), abstraction level, model type (petri-nets, EPC, FSM, etc.) [Th13] or the calculational approach. Thus, a concrete algorithm should, again, be selected depending on the objectives. In contrast to discovery and enhancement, conformance checking should be seen in the phase of usage model analysis (phase 4).

Input: Clustered log data

Content: Particular process mining algorithms depending on the objective (e.g. [AW05, Ka06, WAM06, WR11])

Output: Process model

Running example: Using the Heuristics Miner [WAM06] with default parameters leads to the process model in Figure 1.

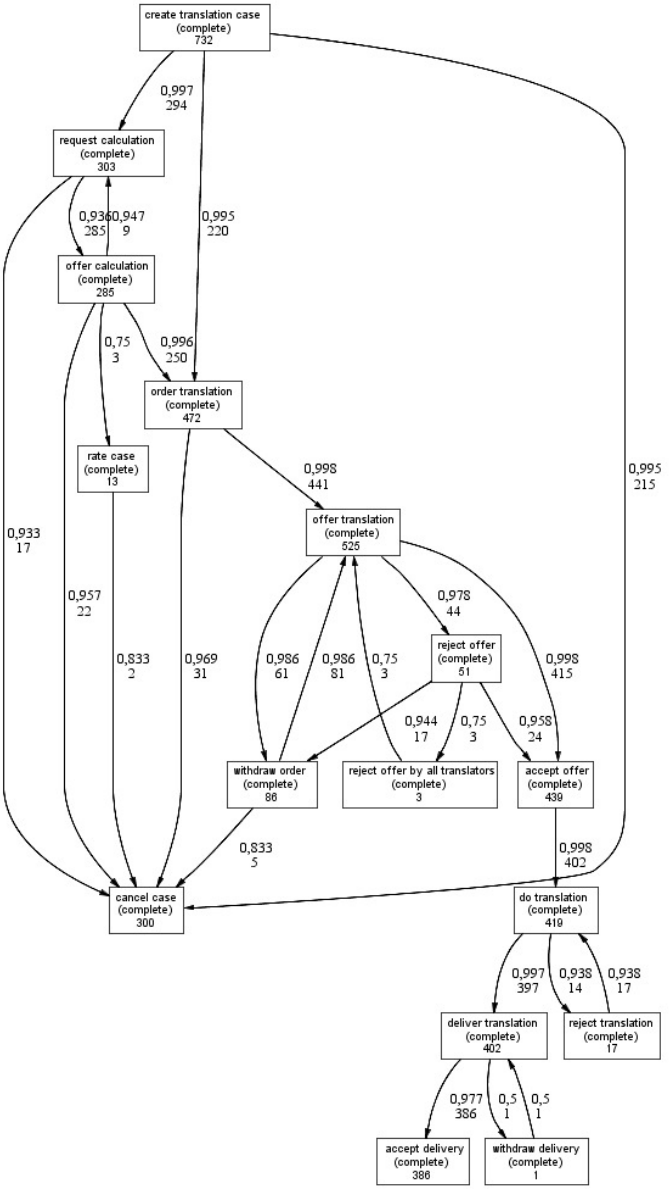


Figure 1: Usage process model of system version 3, mined with Heuristics Mining (ProM 5.1)

4.4 Usage Model Analysis

There are several possibilities of analyzing the usage model. First of all, many metrics from different research fields exist and are able to characterize the model(s) and give first indications to particular weak points:

- Model metrics: e.g. complexity, extent, cross-connectivity [Me12]
- Process metrics: e.g. execution count, execution time, error rates, cancellation rates
- Usability metrics: e.g. irrelevant actions, undo actions, using help function

These categories can also be broken down into further subcategories, e.g. size and complexity in terms of model metrics or placement and time aspects in terms of usability metrics. Disregarding these metrics, there are several further aspects, e.g.:

- Achievement of objectives / conformance checking: In the context of business processes, oftentimes, there are objectives which should be achieved at process executions. These could be the overall execution time of a process, the consumption of resources, etc. Also business rules obligatory at the process execution, e.g. coming from legal aspects, are possible to occur.
- Causal dependencies: Process models may contain causal dependencies between activities or process fragments, which are not evident in the process model. A correlation matrix may uncover those dependencies.
- Core and exception fragments: Oftentimes, process models contain activities or fragments which are executed in a high amount of cases (core actions) as well as those which are executed very seldom (exception actions). Knowledge about that frequency helps focusing on the most important system points during development.
- Non-supported processes: Sometimes users use a system for processes, for which it is not intended by the system producer. Identifying these processes helps improving a system against customer needs or may help identifying further business areas.
- System avoidance: Apart from the use of a system for non-supported processes, users also avoid systems at executing particular process steps. Avoiding a functionality although it is available may be an indication of a non-working or badly implemented functionality.

In a nutshell, simple statistical indicators can sometimes lead to a first hint concerning process or software usability issues, but are not able to analyze these issues in detail. In most cases, further information to the process and its execution logs is needed, which bases on the input from human experts.

Input: Process model

Content: Metrics, conformance checking techniques, correlation analysis, path analysis, behavioral analysis (based on instance tracking), etc.

Output: Metric values, conformance values, correlations, path frequencies, etc.

Running example: In order to analyze the usage / usability, the author implements a research prototype in the RefMod-Miner framework of the IWi at the DFKI. The currently implemented functionality calculates different metrics like element probabilities, duration of element and path executions as well as their frequencies with further statistical data (variance, standard derivation, median, etc.). In addition, the most or less frequent paths as well as the fastest or slowest process instances are being identified. That allows a more detailed analysis of relevant model fragments, also based on the analyst's needs.

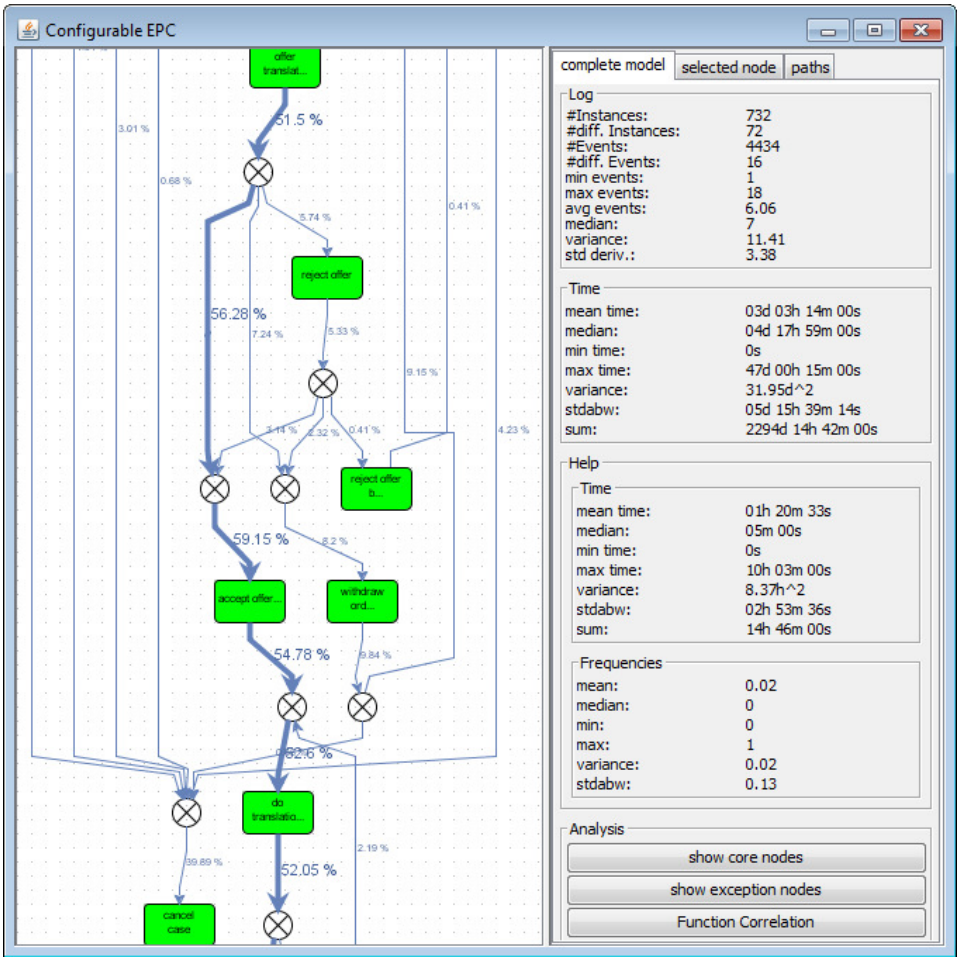


Figure 2: Process analysis with prototypical implementation in the RefMod-Miner

4.5 Recommendation Derivation

The recommendation derivation phase aims at interpreting the analysis results and delivering concrete hints concerning the process and usability improvement and the further development of the system according to the real customer needs. The recommendation should help the system producer answer e.g. the following questions:

- Are there weak points in the system concerning the usability (e.g. avoided functionalities, needless undo-actions within process execution, misuse of functionalities, missing functionalities, unclear labels, very seldom or not used buttons/functions at prominent positions, long loading times)?
- What are the core application scenarios at the user side? / Which implemented processes or functionalities are not used?

- Are there observable user profiles apart from user role or experience? E.g. some users first look for orders while others start with customer management. Are there significant differences in using a system?
- Are there observable case profiles for a process influencing its execution? E.g. minor vs. major order, standard delivery vs. transport of hazardous goods.
- Are there further functional requirements at the user side? E.g. working offline at permanent inventory since there is no network access available at some positions of a stockroom, direct access to the contact list instead of using external software.
- Are there possibilities to improve the process (e.g. user or case sensitive processing, adding new functionalities, data preloading, reorganization of forms)?

Input: Analysis results (e.g. metric values, correlations, etc.)

Content: Different approaches of automatic and manual data mining / interpretation techniques.

Output: Textual hints for improving the process or system usability, weak points, etc.

Running example: The above mentioned simple statistical information already provides an indication of weak points in a system. If e.g. the duration of an activity has a high variance, it could be the case that it cannot be handled intuitively. Users may have problems executing these tasks, thus, new employees will need much more time than troupers. Another reason might be a hidden function that is not supported by the BIS.

Considering the most probable paths, one observes that nearly 30% of the initiated translation cases, are being canceled immediately after displaying the automatically calculated translation costs. The next two paths cover successful process executions, whereby, within 17% of the cases, the automatically calculated translation costs are accepted. But in more than 21% of the cases, customers request an individual calculation. As it is not clear whether the reason for which the price could not be calculated automatically is the cost or the case, one should do both, analyze this system point and enrich the further logging strategy in order to log that aspect.

4.6 Implementation

The implementation phase covers the selection, design, planning and implementation of possible solutions for the afore generated hints. Against that background, this phase marks the beginning of a new lifecycle iteration.

Input: Textual hints (recommendations)

Content: Selection, design, planning and implementation of solutions for the identified improvement capabilities

Output: New system release

Running Example: This phase is not part of the running example, as the intention was to demonstrate the procedure identifying improvement capabilities.

5 Conclusion and Outlook

The paper at hand gives an overview of the concept of usability mining and constitutes different aspects which need to be investigated in order to be able to gather hints on the further development of a software according to the real customer needs. While the phases of user monitoring, trace clustering and usage model derivation already have an established theoretical and technical foundation which can be adapted concerning usability aspects, a detailed analysis of the resulting data seems to be challenging. In fact, there are several ideas quantifying the usability of a software system and characterizing process models. However, these ideas need to be further developed, conceptualized, implemented and evaluated. In addition, both the existing metrics in the context of business process management and usability engineering and the named ideas should be analyzed concerning their information value for the mentioned context.

Hence, the usability mining approach creates the missing link between the software engineering view and the process-oriented view on BIS, which leads to promising potentials on the design and further development of business information systems. Hence, this dissertation is in the core of information systems research.

Furthermore, the intended method has several advantages over existing approaches. It can be applied in production use and in real environments and, thus, involves the real user behavior. Thus, a deformation of measurement results is being obviated which is traditionally the case in direct observations. Moreover, the measurement and analysis of usability aspects can, in many cases, be arranged automatically or with only little input which leads to significantly lower costs and, thus, also enables small and medium enterprises to apply the method. Particularly SME often collapse on the costs of existing alternatives [PR02].

References

[Aa08] van der Aalst, W.M.P.: Decision Support Based on Process Mining. In: Burstein, F., Holsapple, C.W. (eds.) Handbook on Decision Support Systems 1. Basic Themes, Springer, Berlin, Heidelberg, 2008, pp. 637-657.

[Aa11] van der Aalst, W.: Process Mining: Discovery, Conformance and Enchancement of Business Processes. Springer, Berlin, Heidelberg, 2011.

[AW05] van der Aalst, W.M.P., Weijters, A.J.M.M., Maruster, L.: Workflow Mining: Discovering process models from event logs, 2005.

[BGT05] Balbo, S., Goschnick, S., Tong, D., Paris, C.: Leading Usability Evaluations to WAUTER. In: 11th Australian World Wide Web Conference, 2005.

[CW98] Cook, J.E., Wolf, A.L.: Discovering Models of Software Processes from Event-Based Data. ACM Transcations on Software Engineering and Methodology 7, 1998.

[EDG13] Ekanayake, C.C., Dumas, M., García-Bañuelos, L., La Rosa, M.: Slice, Mine and Dice: Complexity-Aware Automated Discovery of Business Process Models. In: Daniel, F., Wang, J., Weber, B. (eds.) Business Process Management, LNCS 8094, Springer, Berlin, 2013, pp. 49-64.

[HMP04] Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly 28, 2004, pp. 75-105.

[Ho06] Hornbaeck, K.: Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies 64, 2006.
[HR99] Hilber, D.M., Redmiles, D.F.: Extracting Usability Information from User Interface Events. ACM Computing Surveys (CSUR) 32, 1999, pp. 384-421.
[IH01] Ivory, Y., Hearst, M.A.: The State of the Art in Automating Usability Evaluation of User Interfaces. ACM Computing Surveys 33, 2001, pp. 470-516.
[Ka06] Ana Karla, M.: Genetic Process Mining. CIP-Data Library Technische Universität Eindhoven, Eindhoven, 2006.
[Ka07] Kassem, G.: Application Usage Mining: Grundlagen und Verfahren. Shaker Verlag, Aachen, 2007.
[IS98] ISO: 9241: Ergonomic requirements for office work with visual display terminals. Part 11 : Guidance on usability, 1998.
[JA09] Jagadeesh Chandra Bose, R.P., van der Aalst, W.M.P.: Context Aware Trace Clustering: Towards Improving Process Mining Results. Proceedings of the SIAM International Conference on Data Mining, SDM 2009, Sparks, Nevada, USA, 2009, pp. 401-412.
[JA10] Jagadeesh, R.P., van der Aalst, W.M.P.: Trace alignment in process mining: opportunities for process diagnostics. Proceedings of the 8th international conference on Business process management, Springer-Verlag, Hoboken, NJ, USA, 2010, pp. 227-242.
[JJ13] Jareevongpiboon, W., Janecek, P.: Ontological approach to enhance results of business process mining and analysis. Business Process Management Journal 19, 2013.
[KLS09] Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M.: Controlled experiments on the web: survey and practical guide, In: Data Mining and Knowledge Discovery, Vol. 18, Issue 1, pp. 140-181, 2009.
[Me12] Melcher, J.: Process Measurement in Business Process Management – Theoretical Framework and Analysis of Several Aspects. KIT Scientific Publishing, Karlsruhe, 2012.
[MGG08] de Medeiros, A.K.A., Guzzo, A., Greco, G., van der Aalst, W.M.P., Weijters, A.J.M.M., van Dongen, B.F., Saccà, D.: Process Mining Based on Clustering: A Quest for Precision. In: ter Hofstede, A., Benatallah, B., Paik, H.-Y. (eds.) Business Process Management Workshops, LNCS 4928, Springer, Berlin, 2008, pp. 17-29.
[Ni93] Nielsen, J.: Usability Engineering. Academic Press, Boston, 1993.
[PR02] Peisner, M., Röse, K.: Usability Engineering in Germany: Situation, Current Practice and Networking Strategie. In: 1st European Usability Professionals Association Conf., 2002.
[SA08] Song, M., Günther, C.W., Van der Aalst, W.M.P.: Trace Clustering in Process Mining. In: Ardagna, D., Mecella, M., Yang, J. (eds.) Business Process Management Workshops, LNBIP 17, Springer, Berlin, 2008, pp. 109-120.
[SDK06] Seffah, A., Donyaee, M., Kline, R.B., Padda, H.K.: Usability measurement and metrics: A consolidated model. Software Quality Control 14, 2006, pp. 159-178.
[SE91] Siochi, A.C., Ehrich, R.W.: Computer analysis of user interfaces based on repetition in transcripts of user sessions. ACM Trans. Inf. Syst. 9, 1991, pp. 309-335.
[Sh08] Shah, I.: Event Patterns as Indicators of Usability Problems. Journal of King Saud University – Computer and Information Sciences 20, 2008, pp. 31-43.
[Th13] Thaler, T.: Entwicklung einer Methode zum Process Mining unter besonderer Berücksichtigung von Organisationswissen. Semiramis Research and Service Unit, 2013.
[TK05] Ting, I., Kimble, C., D., K.: UBB Mining: Finding Unexpected Browsing Behaviour in Clickstream Data to Improve a WebSite's Design. In: IEEE/WIC/ACM International Conference on Web Intelligence, 2005, pp. 179-185.
[TPL13] Thaler, T., Fettke, P., Loos, P.: Process Mining - Fallstudie leginda.de. HMD Praxis der Wirtschaftsinformatik 293, 2013, pp. 56-66.
[WAM06] Weijters, A.J.M.M., van der Aalst, W.M.P., de Medeiros, A.K.A.: Process Mining with the Heuristics Miner-algorithm. BETA Working Paper Series WP 166, 2006.
[WR11] Weijters, A.J.M.M., Ribeiro, J.T.S.: Flexible Heuristics Miner (FHM). In: IEEE Symposium on Computational Intelligence and Data Mining, 2011.

Cooperative-Intrinsic Planning and Model-Driven Design of Business Information Systems

Sebastian Bittmann

Information Management and Information Systems, University Osnabrueck

Abstract. The planning and design of information systems is an increasingly complex and challenging task. With respect to recent developments regarding an individual as a constituting and highly creative impact on the respective organizational structures and application systems, methods are required for a sophisticated inclusion of these actors in the design of information systems. Therewith, the roles of system analysts shift from a prescribing and governmental role to a supportive actor that targets a resource-orientated vision of business information systems. Because of that methods and techniques are required that support the elicitation of potential and the effort for a collaborative identification of a development direction. "Cooperative-intrinsic" refers to the inclusion of any individual of an information system in the act of planning, in order to support agility and flexibility of an enterprise. Model-driven design aims at the support by enterprise models for the analyses and clarification of information systems, to prevent alignment faults, errors and redundancy. Furthermore, technologies need to be adapted purposefully to support the agility and flexibility of modern industries by cooperative-intrinsic planning and model-driven design of business information systems.

1 Introduction

Modelling remains a capstone in an engineering process, whereby it enables, among other aspects, the sophisticated analysis, the inquiry and elicitation of requirements as well as ultimately the design of an actual artefact. However, with its increasing importance and capabilities, a bidirectional and exclusive relation between a modeller and a model is not suitable any more. It becomes important to consider the respective lifecycle of a model and the respective actors that need to interact differently with the model. Thereby, a model continuously evolves instead of being a set of undoubtable facts. For example, enterprise modelling, initially had the primer objective for the planning and realisation of software systems [1] and in some cases is rather perceived as a method for enterprise architecture sketching [2]. However, enterprise modelling has evolved to a paradigm of holistic planning and suppresses as well as incorporates existing paradigms [3]. However, while planning is an iterative and refining process [4], modelling is usually rather result-orientated than action-oriented. Whereby, the latter seeks for completion and correctness [5], planning requires a level of information that enables the respective needed actions. Controversially, the act of enterprise modelling does not orientate towards a certain hierarchy of an enterprise and its respective planning

processes, but a single individual rather selects and interviews respective domain experts.

Therefore, enterprise modelling requires a reflective and participating contention that rather seeks for completion of the level of argument. The process of modelling rather is unaligned with the process of organisational planning. The planning of an enterprise by means of an enterprise model does not necessarily require a complete, but a holistic representation. Thereby, completion does not refer to the holism of knowledge domains, but rather to the holism about application domains. Thereby, it is required to inter-connect the different individuals that intrinsically create the enterprise by means of their actions. Thereby, the usage of upcoming and social technology in order to enable a reflective and purposeful performance of an enterprise is mandatory.

2 State of the Art

Enterprise modelling is a diverse field with propositions of various modelling languages as well as methods. Initiated by the aim of conceptual modelling to provide a graph-orientated and semi-formal representation of a domain [6], more and more evolutions of such a paradigm have to lead to modelling orientations that provide modelling concepts with enriched semantics. Most popular among them, is the idea of business process modelling [7]–[9], which supports a more holistic lifecycle of business process management. With respect to the identification of a business relevance, even languages were proposed that support the depiction of an information system [10], respectively an enterprise as a whole, through the establishment of the term of enterprise modelling [11], [12]. Thereby efforts have been undertaken in order to support the different identified domains of an enterprise, whereby one is goal modelling [13].

As the creation of a model, especially an enterprise model, requires a methodological foundation, various proposals have been made that all come with different abilities. For example, process engineering [14] divides between an as-is analysis and a to-be construction and aims at a sophisticated introduction of business processes and their related software systems. Evolved from these initiatives, the need for modelling the actual requirements has been elucidated in various works [15], [16], due to the success in modern handled software development projects. Advanced from these insights, are agile approaches for both modelling and software development have been developed [17], which primarily aim at a more efficient action-ability and rapid development of software systems. Due to the high risk of failure in human interaction, approaches such as model driven engineering abstract from such and promote models that can be executed on the run and do not further depend on development processes [18]. Despite that recent approaches such as adaptive case management [19] have identified the need for promoting actions that were previously unconsidered on a conceptual scale and try to provide mechanisms for modelling that barely provides enough information and empowers the actual individual in its own problem-solving capabilities [20]. With respect to further upcoming ideas, such as the model driven organisation [3], these advances have to be purposefully integrated in order to enable a holistic creation of an enterprise model that includes such knowledge in a relevant-to-the-actor manner (Req.

1). This requirement is valid regardless, if the model needs to be consumed by a machine or a human. Further the enterprise model actually needs to be an evolving subject and this evolution needs to be coordinated by the same processes that drive the evolution of the enterprise [21] (Req. 2). For example, strategic issues have to be valid on the long-term and operative issues have to be valid on the short-term. However, although such an evolution of an enterprise model, the content must be accessibly and adaptable by the respective actor (Req. 3) and it needs to be integrated in order to identify the value of an enterprise action with respect to the set strategy [22] (Req. 4).

3 Research Questions and Research Goal

With respect to the previously stated requirements, the ultimate goal is to show how enterprise models can be used for planning purposes and further to examine if enterprise modelling languages already provide the necessity concepts or to properly extend them. Initially, the stated requirements demand the analysis of the structure of enterprises and their respective actors. Especially, it is required to integrate various individuals who are not necessarily able to model. Therewith the initial research question for such analysis is the following.

Who are the actors that need to be involved in the creation of an information system and what is the required level of planning that needs adequate support by enterprise modelling?

As it is assumed that not every type of employee of an enterprise is necessarily capable of enterprise modelling per se, new technologies need to be identified that are capable of including the identified individuals in the process of modelling. For example, it might be necessary in a company with a high fluctuation rate to gain knowledge of experienced employees, which is usually not available in a structured sense and not part of an enterprise model. However, individuals with the same tasks will dramatically depend on such knowledge.

How can the various actors be included in order to let them contribute to the actual enterprise model in cooperative manner?

Accordingly, the sophisticated use of such technologies requires to integrate them in a certain model-driven planning process. This is even more important if such individuals from the operative staff are included in the modelling process (as discussed in the previous given example). In that sense, it is not necessarily useful to manage the best practice for sales conservations on the level of content. Moreover, its decision for relevance should be based on the level of a business process model that includes such a task, or even more efficient; based on the strategic plan that e.g. has decided to outsource the sales-department to contractors.

What is the form of a process that supports enterprise modelling as a manner of enterprise planning?

Ultimately, with having developed such a method, it is required to show any value that can be gained by the enterprise and to identify the advantages. One of the most promising advantages might be that issues, such as strategic alignment can be shifted on a conceptual scale and that enterprises can discuss these issues tangibly, identify the source of the issues within misaligned plans and maybe solve the issues by means of the methodological support, e.g. with support of artificial intelligence.

What is the use in creating plans by means of enterprise modelling and implementing a flow of enterprise modelling as an organisational planning process?

4 Research Method

The research method of the targeted research, while underlined by the social constructivism by BERGER and LUCKMANN [23] and the critical rationalism by POPPER [24], uses a configuration of multiple research paradigms, respectively methods, as proposed by FRANK [25]. Initially, a structured literature review of enterprise modelling theories is used to identify present academic work. Especially these theories are examined according to their integration of different levels of planning and their respective methodological support with respect to the initially stated requirements. Thereby, the presented work will examine the following three propositions.

Proposition 1: Reducing the cognitive effort for understanding an enterprise model is more beneficial than any syntactical or semantic precautions.

Proposition 2: The design of enterprise modelling languages requires the consideration of application domains instead of knowledge domains.

Proposition 3: Knowledge in application domains is possessed by individuals and not by enterprise models.

These propositions do not conflict with the need for a general information-, knowledge- or even experience management, but they highlight the relevance of individuals within enterprises. Thereby, the ultimate goal of the work is the transition of organizational behaviour in enterprise modelling methods for sakes of the management of enterprise models and -modelling. During the research, experiments will be chosen for strengthening the stated propositions.

References

[1] A.-W. Scheer, ARIS-business process modeling, 2nd ed. New York: Springer, 2000.

[2] M. et al. Lankhorst, Enterprise Architecture at Work: Modelling, Communication and Analysis, 3. ed. Heidelberg, New York, Dordrecht, London: Springer, 2012.

[3] T. Clark, U. Frank, V. Kulkarni, B. Barn, and D. Turk, “Domain specific languages for the model driven organization,” in Proceedings of the First Workshop on the Globalization of Domain Specific Languages - *GlobalDSL ’13*, 2013, pp. 22–27.

[4] R. N. Anthony, Planning and control systems: A framework for analysis. Division of Research, Graduate School of Business Administration, Harvard University, 1965.

[5] M. Gruninger and M. Fox, “The logic of enterprise modelling,” in Reengineering the Enterprise, J. Brown and D. O’Sullivan, Eds. Chapman & Hall, 1995, pp. 89–98.

[6] P. P.-S. Chen, “The entity-relationship model---toward a unified view of data,” ACM Trans. Database Syst., vol. 1, no. 1, pp. 9–36, Mar. 1976.

[7] A.-W. W. Scheer, “Aris--Business Process Modeling,” Nov. 1998.

[8] B. Curtis, M. I. Kellner, and J. Over, “Process modeling,” Commun. ACM, vol. 35, no. 9, pp. 75–90, Sep. 1992.

[9] P. Wohed, W. M. P. Aalst, M. Dumas, A. H. M. Hofstede, and N. Russell, “On the Suitability of BPMN for Business Process Modelling,” in Business Process Management, vol. 4102, S. Dustdar, J. Fiadeiro, and A. Sheth, Eds. Springer Berlin Heidelberg, 2006, pp. 161–176.

[10] O. Thomas, Management von Referenzmodellen: Entwurf und Realisierung eines Informationssystems zur Entwicklung und Anwendung von Referenzmodellen. Berlin: Logos, 2006.

[11] U. Frank, “Multi-perspective enterprise modeling: foundational concepts, prospects and future research challenges,” Int. J. Softw. Syst. Model., Aug. 2012.

[12] F. Vernadat, “UEML: Towards a unified enterprise modelling language,” Int. J. Prod. Res., vol. 40, no. 17, pp. 4309–4321, Jan. 2002.

[13] G. Koliadis and A. Ghose, “Relating Business Process Models to Goal-Oriented Requirements Models in KAOS,” in Advances in Knowledge Acquisition and Management, vol. 4303, A. Hoffmann, B. Kang, D. Richards, and S. Tsumoto, Eds. Springer Berlin Heidelberg, 2006, pp. 25–39.

[14] A.-W. Scheer and F. Habermann, “Enterprise resource planning: making ERP a success,” Commun. ACM, vol. 43, no. 4, pp. 57–61, Apr. 2000.

[15] W. N. Robinson and S. Volkov, “A meta-model for restructuring stakeholder requirements,” Proc. 19th Int. Conf. Softw. Eng. - *ICSE ’97*, pp. 140–149, 1997.

[16] L. Chung and J. C. S. P. Leite, “On Non-Functional Requirements in Software Engineering,” in Conceptual Modeling: Foundations and Applications, vol. 5600, A. Borgida, V. Chaudhri, P. Giorgini, and E. Yu, Eds. Springer Berlin Heidelberg, 2009, pp. 363–379.

[17] J. Erickson, K. Lyytinen, and K. Siau, “Agile Modeling, Agile Software Development, and Extreme Programming,” J. Database Manag., vol. 16, no. 4, pp. 88–100, 2005.

[18] S. Kent, “Model Driven Engineering,” in Integrated Formal Methods SE - 16, vol. 2335, M. Butler, L. Petre, and K. Sere, Eds. Springer Berlin Heidelberg, 2002, pp. 286–298.

[19] K. D. Swenson, Mastering the Unpredictable: How Adaptive Case Management Will Revolutionize the Way That Knowledge Workers Get Things Done. Meghan-Kiffer Press, 2010, p. 354.

[20] A. Fayoumi, P. Loucopoulos, and A. Fayyoumi, “Hybrid Enterprise Modelling: Integrating Modelling Mechanisms for Socio-Technical Systems Analysis and Design,” J. Softw. Eng. Appl., vol. 7, pp. 6–13, 2014.

[21] P. Loucopoulus and E. Kavakli, “Enterprise Modelling and the Teleogical Approach to Requirements Engineering,” Int. J. Coop. Inf. Syst., vol. 04, no. 01, pp. 45–79, Mar. 1995.

[22] W. Engelsman, D. Quartel, H. Jonkers, and M. van Sinderen, “Extending enterprise architecture modelling with business goals and requirements,” Enterp. Inf. Syst., vol. 5, no. 1, pp. 9–36, Feb. 2011.

[23] T. Luckmann and P. L. Berger, The Social Construction of Reality: A Treatise in the Sociology of Knowledge (Penguin Social Sciences). Penguin, 1991, p. 256.

[24] K. Popper, The Logic of Scientific Discovery. London, New York: Routledge, 2002.

[25] U. Frank, “Towards a pluralistic conception of research methods in information systems research,” University Duisburg-Essen, Institute for Computer Science and Business Information Systems (ICB), 2006.

Channel classification with hidden Markov models in mobile networks

Rafiaa Boujbel

Department of Electrical Engineering and Information Technology
Munich University of Applied Sciences
Lothstr. 64
80335 Munich
rafiaboujbel@hotmail.fr

Abstract: In telecommunication networks, Key Performance Indicator (KPI) is monitored to ensure the connection quality. Many indicators are provided by the network equipment in order to ensure higher Quality of Service (QoS) in communication networks. With the significant increase of data traffic on the mobile network, a detailed analysis of the transmission quality is becoming increasingly important. Existing classification approaches are widely considered for network traffic classification and not the channel in itself.
The aim of this work is to develop a channel estimation tool based on hidden Markov models that is able to determine transmission channel characteristics in mobile radio receivers.

1 Introduction

In the field of telecommunications engineering, the term channel refers to the medium between a transmitter and a receiver. The characteristics of wireless signal changes as it travels from the transmitter to the receiver. These characteristics depend upon the distance between the two parts, the signal path, and the environment around the path. A channel model, a model of the medium between the two parts, is used to obtain the profile of the received signal from the transmitted signal. Due to the enrichment of services and data traffic on mobile networks (e.g., downloads, videos in high resolution), channel classification is becoming crucial to ensure a favorable network quality for the user.

On the one hand, channel classification is widely associated to GSM and wireless. On the other hand, existing classification approaches are based on the network traffic and not the channel in itself.

The aim of this work is to combine channel classification and new generation mobile networks. This classification is based on hidden Markov models (HMMs) [GDM11] which are widely used and accepted in the field of digital communication. From the receiver side, a channel classification is useful for giving an overview of the channel

conditions. The general idea is to provide users an estimation tool of the channel with an easy access. The role of this tool is to give the operator an estimation of the quality and the characteristics of the transmission channel in order to assure a good use of the channel in mobile networks.

2 Approach

Consider the case of Single-Input Single-Output (SISO), packets are created and sent in a central network unit (e.g. Operation and Maintenance Center (OMC)). In the mobile radio channel, the signals are disrupted and delayed, depending on the characteristics of the channel. Due to transmission error, data packets are retransmitted and lead to delays in the corresponding IP packets. These delays depend on the packet length, the type of packet error and the used mechanism of correction namely Automatic Repeat Request (ARQ) or Hybrid Automatic Repeat Request (HARQ) [Ge11]. The evaluation of the measured jitter of the IP packets, the packet loss, and possibly higher protocol layers such as RTP leads to estimate the occurred packet error in lower layer. As it is shown in figure 1, sequence errors are generated from jitter. Received packets are converted to "1" in the case of error and "0" in the case of error-free.

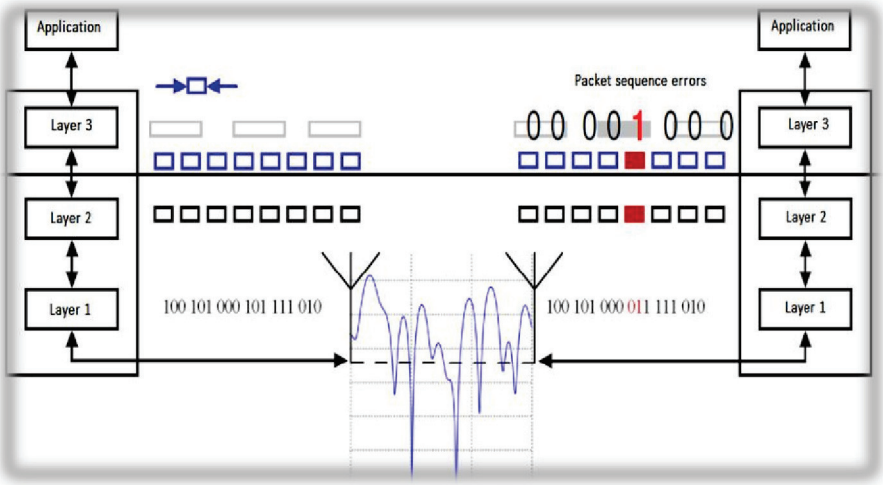


Figure 1: Schematic diagram of generating sequence errors

The functional blocks shown in figure 2 are the subject of this work, namely the channel models database and the channel estimator.

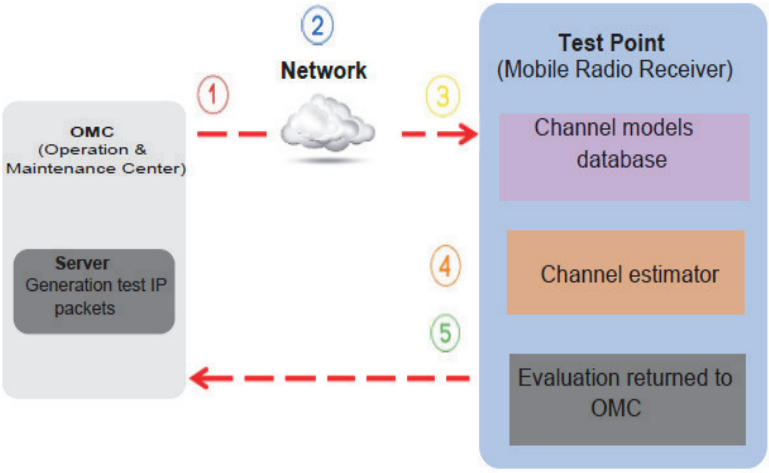


Figure 2: Functional model for channel classification in mobile networks

2.1 Channel models database

Channel models database is a database of models and precisely of HMMs. These models have been provided from a radio channel simulator on the basis of known channels. To simulate the external environment, a wireless channel emulator is used to emulate wideband radio channel characteristics such as fading profiles which are defined in the 3GPP standard.

These HMMs are built specially from sequence errors. Sequence errors are generated from Acknowledgement (ACK) and Negative- Acknowledgement (NACK) protocol message. ACK and NACK are collected via a reliable tool and sequence errors are generated as fellow: in case of erroneous packet, a NACK is received and converted to error "1". In case of error-free packet, an ACK is received and converted to "0".

2.2 Channel estimator

The inputs of the channel estimator are two models. The first model is a HMM already stored in the channel models database. The second model is a HMM generated from the measured signal. In this case, sequence errors are generated from inter-packet delay. When the inter-packet delay is higher than a certain threshold then an error is considered and a "1" is generated. When the inter-packet delay is lower than a certain threshold then an error-free is considered and a "0" is generated.

This channel estimator consists of the metric Kullback-Leibler Distance (KLD) [SY11] [ZL12]. The KLD is used to compare the similarity between these two models described above. Smaller is the KLD; bigger is the similarity between the models.

The outputs of the KLD algorithm are the channel model used and the condition of the channel. These informations are sent back as a result to the OMC.

3 Results

Considering the case of Long Term Evolution (LTE) network, a couple of measurements are done in order to guess the channel conditions used within the network. These channel conditions are namely different fading models. As it is defined in the 3GPP standard, LTE fading models are divided in low fading namely Extended Pedestrian A model (EPA5), medium fading namely Extended Vehicular A model (EVA5/EVA70) and high fading namely Extended Typical Urban model (ETU70/ETU300). The idea is to take two measurements with two different channel conditions: One measurement with good channel conditions (EPA5) and the other one with bad channel conditions (ETU70).

Assuming that we do not know the channel fading conditions used and we try to estimate them via the channel estimator.
From the received signals of our measurements, sequence errors are generated as it is described above. The channel estimator, precisely the KLD algorithm, is used to measure the distance between the sequence errors generated and each HMM stored in the channel models database. As it is shown in the tables below, the smallest absolute distance is considered and the model is classified in the channel that corresponds to the smallest distance.

Actual	**Predicted**				
	EPA5	**EVA5**	**EVA70**	**ETU70**	**ETU300**
EPA5	**231.66**	250.36	321.23	394.16	395.17

Table 1: KLD between EPA5 sequence error and HMMs in database

Actual	**Predicted**				
	EPA5	**EVA5**	**EVA70**	**ETU70**	**ETU300**
ETU70	791.35	515.92	457.29	**156.91**	226.70

Table 2: KLD between ETU70 sequence error and HMMs in database

4 Conclusions and outlook

Network traffic classification is the basis of numerous network applications. The classification method used in this work is based on HMM. Experimental results indicate that the combination of channel characteristics and statistical models leads to precise model and achieve higher classification accuracy.
As a continuation of this work, an investigation of Multiple Input Multiple Output (MIMO) channel classification is becoming crucial. The MIMO technology is promising to be one of the key techniques for wireless communications beyond 3G, for its high spectrum efficiency and reliability.

Therefore, extracting jitter values from the received signal without expensive measurement devices and adapting results found to the MIMO technology are the key point for future work.

Acknowledgments

This project is promoted by the Bavarian Ministry of Economics, Infrastructure, Transport and Technology and it is performed in cooperation with Keynote-SIGOS GmbH.

References

[GDM11] Güzelarslan B, Dippold M, Paul M.: Efficient and automatized error pattern modelling with hidden Markov models in digital communication. AEUE, 2011; pp. 417-424.

[Ge11] Gessner, C.: A concise introduction to LTE and its measurement requirements. Rohde & Schwarz, Germany, 2011; pp. 83.

[SY11] Sahraein. M, Yoon. B.: A Novel Low-Complexity HMM Similarity Measure. IEEE, 2011; pp. 87-90.

[ZL12] Zhen-Hua L, Li-Rong D.: Minimum Kullback-Leibler Divergence Parameter Generation for HMM-Based Speech Synthesis. IEEE, 2012; pp. 1495-1497.

SurgLink: Cross-media Linking in the Operating Room

Yihan Deng

Innovation Center Computer Assisted Surgery (ICCAS)
University Leipzig
Semmelweissstr. 14
04103 Leipzig
yihan.deng@medizin.uni-leipzig.de

Abstract: The clinical daily work is accompanied by different types of patient records. A successful diagnostic and therapeutic process relies on efficient access and comprehension of a patient's status. With the proceeding of treatment, the physician can easily be overwhelmed by the rapid growth in data volume of patient records. In particular, the information seeking and navigation between different media consume tremendous time and labor. Due to the lack of a cross-media connection, the retrieval of similar cases and a causality analysis among the heterogeneous data sources can merely be conducted by hand. This paper presents the preliminary stage of the PhD study and tries to raise and explain the main research questions for finding an approach to cross-media linking in the operation room. First of all, the status quo of the multimedia data storage in clinical context is introduced. Second, the special requirement of cross-media linking in the operating room is presented. Finally, the connection possibilities between different media and the feasibility of the cross-media linking are discussed.

1 Introduction

A clinical information system (CIS) serves normally as an organizational platform for the clinical information process, which is defined as socio-technical system [Rop99]. It comprises technical systems and human activities. The patient records in the CIS platform are thus distributed over several systems, which makes it difficult to obtain an integrated view. Since the first reception of a patient in the hospital, all necessary information for the diagnosis and treatment covering laboratory test, radiological and pathological investigation are stored in different sub-systems. As can be seen in Figure 1, physicians have to face numerous heterogeneous data sources. Generally, the most commonly used types of item are: 1) Text: Finding report, nurse letter, admission note, drug prescription, discharge summary, etc. 2) Image: Radiological and pathological image, endoscope image, scanned document (legacy files, important paper-based note, referral documents). 3) Audio: Dictated report for prompt recording of diagnosis and findings. 4) Video: Operation records for the purpose of teaching and quality management.
Although methods for text retrieval have been developed and optimized for some time, little work has been done to realize a clinical cross-media linking and retrieval. Along with the rapid growth in the data volume of patient records, the information seeking activity in

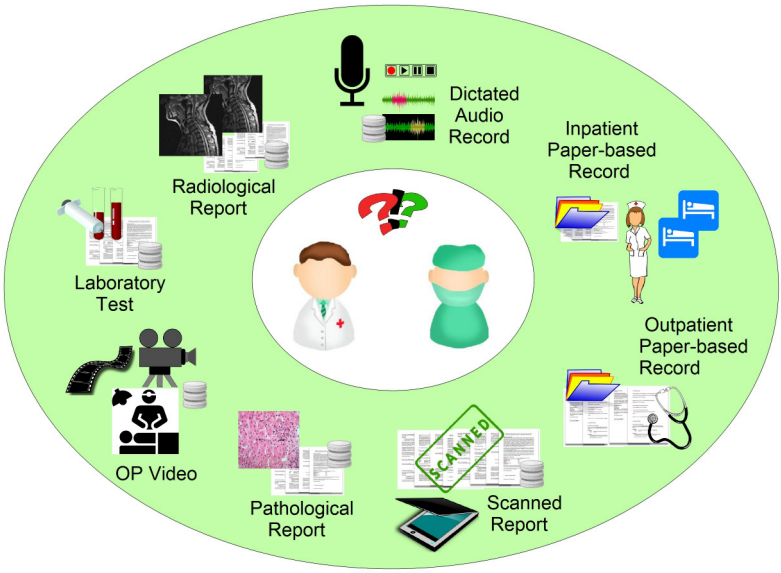


Figure 1: Status Quo of the Clinical Information System

the CIS is always a time consuming and labor-intensive task, especially for retrievals involving several media types and multiple document hierarchy. Normally, physicians need to find out the connection and diagnostic clue among different media manually. Often, they have to spend several hours on seeking a scanned document or one slide of radiological image so that the effectiveness of the treatment can be confirmed. Sometimes, it is even impossible to find out the desired document by switching between different modalities. Moreover, the searching of dictated audio snippets can only be conducted by hand, it costs a lot of time to navigate to the target excerpt in entire audio tracks, whereas the retrieval of video can be solely achieved with manually “fast forward” and “backwards”. The inconvenience by cross-media retrieval leads directly to time wasting and suboptimal planning of treatment. Even treatment error can occur due to a lacking information.

Considering such information retrieval task in critical situations as those in an operating room, it becomes clear that there is a need to support physicians in retrieving and accessing relevant information. During a surgery, a physician must not only concentrate on the operation, but also make sure that each step of the procedure is completed in full accordance with the operation plan and clinical evidences. However, the information entities are scattered among different modalities. In order to obtain detailed examination results, the surgeon has to suspend the operation process and ask the nurse or the resident to search and open relevant documents stored in the CIS. The cost for information seeking is quite high.

According to the situation in the operating room, some new information organization and linking mechanism should be implemented to help the physician in overcoming the multimedia information retrieval problems. The record of the patient should be reconstructed and presented to the physician so that a comprehensive and fore-thoughtful decision can

be made.
The remainder of this paper is arranged as follows: Section 2 raises the three research questions, which will be centered in this work. Section 3 presents related work of cross-media linking in medical domain. As next, the linking possibilities, discussion over research questions and future works are described in Section 4, 5, 6 respectively.

2 Research Questions

Due to aforementioned facts in CIS and the complex environment in the operating room, the following three research questions will be addressed in this PhD thesis:

1. **From a technical point of view, what kind of approach should be developed to realize a cross-media linking in a clinical context?**
2. **From a clinical point of view, which kind of requirements should be fulfilled for cross-media retrieval in the operating room?**
3. **Can cross-media linking techniques provide required accuracy for the usage in the operating room?**

3 Related Work

The cross-media linking is not a novel technology, as it has been already studied since the 90s of the last century. Most of methods tended to create appearance models to match the media entities with the same semantic meaning automatically. However, the application of cross-media in clinical context has still not been well explored. The reasons are twofold: First, clinical data is stored in a distributed manner, which limits its accessibility. Second reason is that the sophisticated clinical work flow involves more media in the entire process, which makes the clinical data linking and cross-media retrieval more complex than in other context. The medical context has also advantages in comparison to a general domain: Various biomedical ontologies such as UMLS[1], RadLex[2], FMA[3], etc. have been developed to summarize the terminologies and their dependencies, which provides resources for semantic annotation and inference. Moreover, the concept of digital patient modeling [Den13] has extended the semantic description between patient records from different media based on probabilities model. These knowledge bases provide the semantic foundation of data linking between clinical records from terminology level to document (record) level.
In a CIS, the image data takes the largest part of the data volumes. The current research of clinical cross-media data management has therefore focused on the linking between text

[1] http://www.nlm.nih.gov/research/umls/, accessed 11.06.2014
[2] https://www.rsna.org/RadLex.aspx, accessed 11.06.2014
[3] http://sig.biostr.washington.edu/projects/fm/AboutFM.html, accessed 11.06.2014

and image. In order to process and retrieve the image in a comparable way as textual data retrieval, the corresponding model should be developed to interpret the image or store the interpretation of image. Zrimec et al. [ZLKS09] have shown a system named *Medical Image Assistant (MIA)*. It uses domain knowledge about human anatomy and pathology as well as the medical image protocols to organize the data from the PACS system [Hua04]. The image data is transformed into a set of numerical descriptors, color, texture and structure components, while the lung anatomy and disease patterns are extracted based on high resolution CT images automatically. Attributes such as color histogram, average color intensity and texture first order are calculated for every raw image by the MIA system. During the retrieval, the features from medical protocol and CT images are stored in the global description index. The MIA has employed a relatively less complex semantic representation using knowledge based image analysis to link the multimedia data and provided a swift data access. Kozuka et al. [KTK+13] have presented a new similarity comparison method for the image retrieval. It combines image features and description extracted from finding report to facilitate the cross-media retrieval. First, the diagnostic knowledge is extracted from previous cases. Then, the correlation between diagnostic knowledge and image features is determined and the most effective features for the retrieval are derived. For the retrieval of similar cases, a physician needs to choose firstly a region of interest and input short description from finding reports referring to the chosen region. The system extracts image features from the specified region of interest. Meanwhile, the diagnostic knowledge is used to weight the image features related to keywords. In this way, the extracted knowledge from past cases can adapt the weights of image features dynamically. According to the result presented in [KTK+13], the concordance rate of texture classification is improved from 57.7% to 75.9% and the precision of subjective evaluation is improved from 61.2% to 71.3%.

4 Simulated Scenario and Requirement in the Operating Room

With the aim of understanding the real information needs of a surgeon, a series of clinical work shadowing have been conducted. The typical information needs from surgeons are, e.g., "I want to see all information related to that particular operation or the medical procedures made during the staying of a patient" or some follow-up question like "The system says there were complications occurring during the treatment of patient X. Are there some video recordings of the operation or some nurse feedbacks or differential diagnoses?".
In order to answer these questions, the system should provide an overview considering detailed dependencies among different types of media. As can be seen in Figure 2, the clinical data stream from different media for one patient can be implicitly or explicitly connected based on dependencies. If a surgeon needs to make some prompt retrieval of patient data and examination results in an operating room, the system should be able to provide not only single relevant document but also a series of relevant documents regarding dependencies rapidly. The fine-grained relations between entities within one document or between documents should be also presented to the surgeon. Moreover, a user-friendly access interface should be offered so that the surgeon can raise the query conveniently.

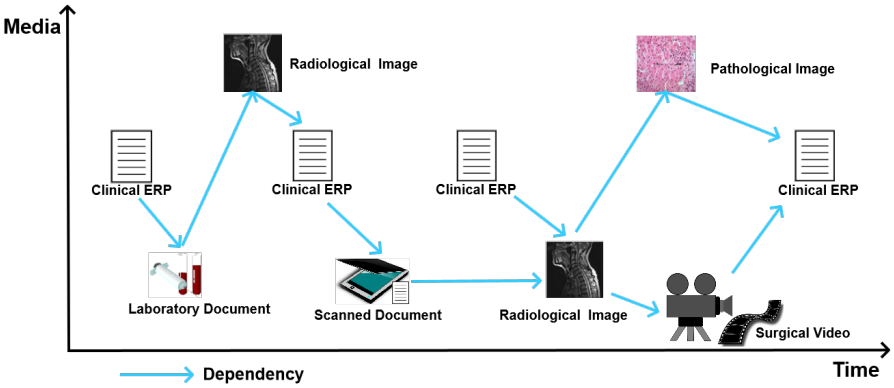


Figure 2: Dependency Chains from Different Media for One Patient

5 Linking Possibilities

Various connections among media are of interest in an operating room. The linking possibilities between text and other media types are crucial, since the text has constituted not only the representation within knowledge bases and other textual resources, but also the connection between natural language utterance. If other media can be annotated with text, the semantic roles and dependencies can be inherited from well-defined knowledge bases. Thus, the semantic description in knowledge bases can be expanded to multimedia resources. Due to the scope of the PhD study, it is assumed that each multimedia document or snippet is annotated with the corresponding text tags. Hence, the multimedia linking will mainly focus on the linking of multimedia resource through knowledge mapping. The special snippet in image, scanned document, video and audio labeled with the same semantic meaning will be clustered together. The links types between different media documents or snippets will be derived from a knowledge base such as medical ontology or patient modeling for certain diseases. Reversely, the machine learning techniques can also be applied to discover the newly emergent terminologies and dependencies so that the potential update entry for knowledge bases can be offered.

6 Conclusions and Future Work

As a preliminary stage of the PhD program, the status quo of CIS and requirement for cross-media linking in the operating room have been presented based on the clinical work shadowing and the state of the art analysis. The three research questions are answered as follows:

- **RQ1:** For cross-media linking, a unified data structure should be defined to map the documents from different media so that they can be grouped together according to

their content, while their relations can be determined based on medical ontologies and patient modeling for special disease through concept mapping. The user could get the possibility to "see both the forest and the trees". Especially for large data volume, the implicit clues hidden in the multimedia documents can be explored through the linking process, which offers additional serendipity for users.

- **RQ2:** From a clinical point view, a successful retrieval in the operating room depends on the full accordance of surgeons' need for information seeking. A swift delivery of configurable overview of patient status and an easy interaction interface will play the crucial role. The audio query interface should also be integrated to let the surgeon generate the query by their voices so that the surgeon can conduct non-stop operation with both hands.

- **RQ3:** Under the assumption that each multimedia document or snippet is annotated with the corresponding text tags, an acceptable accuracy and efficiency can be expected by using various of medical ontologies, patient modelings and machine learning methods. For optimization, the compromise between precision and recall should be decided by concrete scenarios, e.g., a result with high precision and moderate recall can help the surgeon focusing on several most relevant information instead of a result set with several pages.

At the present stage, the real data obtained from surgical data stream will be further analyzed. The features from different media will be compared. The experiment for the linking between finding report and radiological image will be conducted as a first step. In the short term, the prototype for cross-media linking will be implemented to audio and video resources. In the long term, the validated prototypical system will be integrated in the operating room to evaluate its performance.

References

[Den13] Kerstin Denecke. Model-based Decision Support: Requirements and Future for its Application in Surgery. *Biomedical Engineering*, 58(1), 2013.

[Hua04] H.K. Huang. *PACS and imaging informatics: basic principles and applications*. Wiley-Liss, 2004.

[KTK+13] K. Kozuka, K. Takata, K. Kondo, M. Kiyono, M. Tanaka, T. Sakai, and H. Kimura. Development of image-retrieval technology using radiological knowledge extracted from a clinical database of lung CT images. In *International Journal of Computer Assisted Radiology and Surgery*, pages 71–74, 2013.

[Rop99] Günter Ropohl. Philosophy Of Socio-Technical Systems. *IN SOCIETY FOR PHILOSOPHY AND TECHNOLOGY 4(3)*, 4:59–71, 1999.

[ZLKS09] Tatjana Zrimec, Iztok Lajovic, Kres, and Kreativni Sistemi. A Medical Image Assistant for Efficient Access to Medical Images and Patient Data. *Journal on Information Technology in Healthcare*, 7:261-271:pp.261–271, 2009.

Distributed Denial of Privacy

Hannes Grunert
Database Research Group
University of Rostock
Albert-Einstein-Straße 22
18051 Rostock, Germany
hg(at)informatik.uni-rostock.de

Abstract: For many years, privacy has not been considered as a critical issue in the field of software development. Recent events have led to discussions how the privacy of the user can be protected in information systems. While most approaches do only focus on increasing the security of the systems to protect private information against uncontrolled access, the aspect of enhancing privacy itself has been ignored. Our framework, called PArADISE (Privacy AwaRe Assistive Distributed Information System Environment), maximizes privacy by minimizing the amount of data stored and queried by information systems. In this paper, we present a new method to protect the anonymity of the user, even if multiple queries are combined.

1 Introduction

Assistance systems in smart environments are collecting a lot of data via different sensors (e.g. the SenseFloor project [SL08] and passive RFID tags [WT13]) in order to compute the intentions and forecast future activities of the user. In most cases the data contain more information than needed for achieving the purpose of the system.

The aim of our work is the privacy-aware propagation of the sensor- and context-information towards the analysis-tools of the assistance system. The privacy needs of the users are transformed into integrity constraints of the database system storing the collected information. Comparing the information needs and the privacy claims of the users, the privacy component of the assistance system selects, reduces, compresses or aggregates the data at the sensor as closely as possible.

Our approach is implemented in the PArADISE-Framework (Privacy-AwaRe Assistive Distributed Information System Environment) and is going to be evaluated in the Smart Appliances Lab ([EK05]) at the DFG-Graduate school MuSAMA (Multimodal Smart Appliance Ensembles for Mobile Applications).

Previous work has been limited to protect the privacy of the user in a single result set. By combining the results of multiple queries, an honest-but-curious adversary can gain access to secret information. In this paper, we propose a new methodology to observe queries using a sliding window. During the lifetime of this window, a user can not be identified.

The rest of this paper is organized as follows. In the second section the related work is

presented. The third section gives a brief overview of the superior topic. Issues regarding privacy concerns by connecting information are discussed in the next section. Some conclusions are drawn in the final section.

2 Related Work

For the compliance with privacy-protection different access-control-mechanisms and -algorithms which ensure the anonymity or pseudo-anonymity of personal records exist. Most algorithms base on k-anonymity ([Swe02]). It proposes a method that splits data into quasi identifier (QIs) and sensitive information (SIs). The parameter k shows, that k tuples are indistinguishable by using only the QIs. Thus, re-identification attacks are blocked.

Other techniques, such as l-diversity ([MKGV07]), t-closeness ([LLV07]), Data Slicing ([LLZM12]) and Bucketization ([KG06]) improve k-anonymity by preventing similarity attacks.

Unfortunately, one of the major drawbacks to these algorithms is that information can be combined to unreveal private data, even if it was highly anonymized before. In this paper, we present a new method to get rid out of this limitation.

3 PArADISE

In order to enhance privacy in assistance systems, we are currently developing a framework called PArADISE (Privacy AwaRe Assistive Distributed Information System Environment). In this prototype, the user can specify detailed privacy settings in a policy. The users privacy concerns are checked against the information need of the assistance system.

The aim is that the user offers the system enough data to operate as desired, but can not infer more information about him or her. For that task, we are currently implementing a privacy engine, which consists of the privacy policy and a privacy-aware query processor.

Our privacy model is based on the W3C draft for Privacy Preferences Platform (P3P, [W3C07]) and the Bell-LaPadula Security model ([BL73]). As an extension, we adopt continuative concepts for data selection and aggregation.

The access to information can be designed in a more fine-grained way by introducing additional gradations between locking and releasing data. Thus, both the information needs of the assistance system and the privacy rights of users can be satisfied. Possible intermediates can be obtained by selection, restriction, compression and aggregation, as well as by the insertion of the fuzzy values. For reasons of space, our privacy model is not addressed in this paper. For more information, please refer to [Gru14].

The privacy-aware query processor is composed of a pre-processor for incoming queries and a post-processor for the result set. Whereas the pre-processor modifies the query

to reduce the amount of data fetched from the data storage, the post-processor applies privacy-algorithm before transmitting the data back to the requester.

During the pre-processing stage, the preliminary query is analyzed and checked against the privacy policy of the concerned user. Every affected personal information queried by the system is monitored, whether it is uncovered by the user at all (projection) and if it can be used under user-defined constraints. This constraints can be used to decide if the revealed information are preselected or aggregated. Furthermore, it is also checked if the processing node has enough capacity (load balancing). Also, it is tested if the information system could gain enough information to produce satisfactory results. Finally, the constraints are used for a modification of the query which fulfills all needs. The reformulated query is executed on the data storage holding the personal information. The gained results are passed towards the postprocessor.

Taken the preliminary result from the intermediate stage, the post-processor checks information needs and privacy settings. This time, the result is modified with privacy-preserving algorithm like k-anonymity ([Swe02]) or data-slicing ([LLZM12]), if and only if the processing unit has enough power. Thus, the modified result is send back to the requester. A serious limitation with this approaches, however, is that they operate on a single result set. In the next chapter, we present a new method which uses a sliding window to monitor a data pool holding personal information.

4 Maintaining anonymity through multiple queries

The algorithms for achieving privacy presented in section 2 only work on single queries. The anonymization is insufficient, if a single user executes multiple queries on the database. While individual result sets for themselves can guarantee the anonymity of a user, the connection of the results can disturb privacy.

The information in the database form an universe U. The adversary do not know anything about U before he executes his first query q_1 on it. The outcome of query q_1 (after applying privacy-algorithm) is the set A. Executing a second query q_2 results in a set B. Now, the adversary has the information of both queries. Furthermore, he can connect set A and B ($A \cap B$) to get all tuples which belongs to both queries. Also, he knows which tuple appertain to set A but not to set B ($A \backslash B$), and which tuple appertain to set B but not to set A ($B \backslash A$). So, all three sets ($A \cap B, A \backslash B, B \backslash A$) have to checked against privacy-algorithm to guarantee anonymity.

Example A database db stores information about employees (name) and their income (salary). Query A sums the salary up (*SELECT SUM(salary) FROM db*), query B does the same, but without an employee named P (*SELECT SUM(salary) FROM db WHERE name != "P"*). Both queries provide only aggregated, statistical information. Nevertheless, it is obvious to see, that subtracting the result B from A reveals the salary of person P, which equals the subset $B \backslash A$.

When a database is executed over a long time, the number of queries ascends continuously. Thus, the number of subsets increases exponentially ($2^n - 1$ subsets for n queries). Guaranteeing anonymity on every subset confronts the system with two problems. First, the privacy-algorithm have to be applied on every non-empty subset. Having a huge amount of subsets leads to a long execute time, which effects the respond time of the system.

To solve these problems we propose the method shown in algorithm 4. As input, it catches a continuous stream of queries *Q*. The algorithm is executed for every new query of the system. Whilst the maximum length *n* of sliding window *W* is not exceeded, every query *q* is added to *W*.

Query *q* is analyzed and combined with the other queries of *B*. *q* is connected with a logical AND with the known queries of *B*. The same is done with the complement NOT(q). Furthermore, *q* is connected with every complement of *B*.

Thus, all subsets are stored in *B*. For every subset, an anonymization-algorithm (e.g. [Swe02]) is applied to monitor privacy claims.

If the maximum length is exceed, the oldest query *q_old* is removed from *W* and all occurrences of *q_old* are removed from *B* to realize the sliding window. Afterwards, the algorithm is repeated for the new query list.

In most cases, checking anonymity for queries from a single peer is insufficient. Experience from the field of Web Security shows that attacks can also been distributed on multiple peers. Traditional kinds of Distributed Denial of Services (DDoS, see [MR04]) attacks are flooding a specific server with rubbished queries until the system is unable to handle the traffic.

Countermeasures for DDoS includes revocation lists, rate limiting and filtering of ports and service-requests. These techniques can be assigned to privacy aspects as well. Some aspects are already covered by our privacy policy introduced in the previous chapter.

In case of privacy aspects, too many queries can lead to a huge amount of subsets, i.e. $2^n - 1$ subsets for n queries. Even if the server is able to check privacy concerns, the data can be partitioned into small parts such that every subset would leak private information. As a consequence of this, no data is uncovered, although the request is honest.

The policy allows the user to specify purposes and conditions under which a specific application can access private information. This also includes the maximum allowed frequency to query the data set. Thus, designated applications can independently query the data set. The queries for every application are stored in its own sliding window.

5 Conclusion and Further work

Our main objective is to create an environment for privacy preservation in smart environments. In this paper, we introduced an algorithm for monitoring privacy over a period of queries by using a sliding window. Additionally, we considered the setting in which the queries are committed by multiple peers.

Algorithm 1 k-anonymity on multiple queries

```
/**
* Q := a list of queries
* W := sliding window
* n := maximum length of W
* B := a set of combinded queries **/
B = ∅
//Triggered when a new query is deployed to Q
for all new q in Q do
  if |W| < n then
    W.add(q)
  else
    q_old = W.take();
    B.remove(q_old);
  end if
  if |B|! = 0 then
    for all b in B do
      //Combine every known query with the new query or its complement
      B.add(b AND q);
      B.add(b AND NOT(q));
    end for
  end if
  //Combine the new query with the complement of every known query
  B.add(q AND(NOT(b in B)))
  for all b in B do
    if count(b) < k then
      RETURN "NO_PRIVACY";
    end if
  end for
  RETURN anonymized_result(q);
end for
```

Further work needs to be performed to establish our framework. The prototype is currently limited for relational data sets. In Big Data scenarios, the collected data and the associated file formats is often heterogeneous. Research into solving this problem is already under progress. Future work will concentrate on testing the proposed algorithm under real conditions in our Smart Lab.

Acknowledgments.

Hannes Grunert is funded by the German Research Foundation (DFG), Graduate School 1424 (Multimodal Smart Appliance Ensembles for Mobile Applications - MuSAMA). I gratefully acknowledge the constructive comments of the anonymous referees.

References

[BL73] D Elliott Bell and Leonard J LaPadula. Secure computer systems: Mathematical foundations. Technical report, DTIC Document, 1973.

[EK05] José L Encarnaçao and Thomas Kirste. Ambient intelligence: Towards smart appliance ensembles. In *From Integrated Publication and Information Systems to Information and Knowledge Environments*, pages 261–270. Springer, 2005.

[Gru14] Hannes Grunert. Privacy Policy for Smart Environments. `http://www.ls-dbis.de/pp4se`, 2014.

[KG06] Daniel Kifer and Johannes Gehrke. Injecting Utility into Anonymized Data Sets. *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD)*, pages 217–228, 2006.

[LLV07] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering*, pages 106–115. IEEE, 2007.

[LLZM12] Tiancheng Li Li, Ninghui Li, Jian Zhang, and Ian Molloy. Slicing: A New Approach for Privacy Preserving Data Publishing. *Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD)*, 24(3):561–574, March 2012.

[MKGV07] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[MR04] Jelena Mirkovic and Peter Reiher. A Taxonomy of DDoS Attack and DDoS Defense Mechanisms. *SIGCOMM Comput. Commun. Rev.*, 34(2):39–53, April 2004.

[SL08] Axel Steinhage and Christl Lauterbach. Sensfloor (r): Ein aal sensorsystem für sicherheit, homecare und komfort. *Ambient Assisted Living-AAL*, 2008.

[Swe02] Latanya Sweeney. k-anonymity. A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, October 2002. doi: 10.1142/S0218488502001648.

[W3C07] W3C. Platform for Privacy Preferences (P3P) Project. `http://www.w3.org/P3P/`, 2007.

[WT13] Benjamin Wagner and Dirk Timmermann. Adaptive clustering for device free user positioning utilizing passive RFID. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*, pages 499–508. ACM, 2013.

Measuring and Monitoring the Rehabilitation of Patients on Monitoring Stations via the Analyses of poly-structured Data

Mirco Josefiok

Data Management and Data Analysis
OFFIS Institute for Information Technology
Escherweg 2
26121 Oldenburg
mirco.josefiok@offis.de

Supervisor: Prof. Dr. Dr. h.c. H.-Jürgen Appelrath

Abstract: In this paper an approach for measuring an monitoring the rehabilitation of patients on monitoring stations via the combined analyses of structured and unstructured data is presented. Here, the focus is on stroke patients in different rehabilitation phases. For clarification interviews with IT managers from hospitals and with parts of the medical staff were carried. Based on the information from the interviews a first theoretical and a prototypical approach is presented. An outlook towards the possible evaluation and further work is given.

1 Introduction

In this paper a method with an accompanying software application for measuring and monitoring a patients rehabilitation on monitoring stations will be presented. On monitoring stations lots of documentation is handwritten. For a retrospective assessment, an analyses in combination with structured data is necassary. Whilst the analyses of structured data is handled well and also analyzing unstructured data is possible a combination is not yet possible to the extent necessary [BFG+14]. Other approaches do not combine structured and unstructured data in a unified analyses process, but extend the structured database with information gathered from unstructured data. Although approaches for handling huge amounts of unstructured and structured exist in other domains none is is readily applicable to the health care sector [CCS12].

The stroke is one of the most common diseases in Germany. About six billion Euros are spent each year for the treatment of the most frequent types[1]. Most acute stroke patients are initially treated in so called stroke units which are specialized wards for taking care of stroke patients. Patients who suffer from a severe stroke and fall into a coma are treated in

[1] see diagnosis code I60, I61, I63, I65 at `www.gbe-bund.de` - last visited 2014-04-14

different phases, labeled from A to F [USG10]. Phase A describes the acute phase, phase B the early stages of rehabilitation, phase C the early mobilization, phase D the rehabilitation, phase E the outpatient follow-up and phase F the integration and participation. It is possible for each phase to be carried out by a different institution. Just in one phase a huge amount of unstructured date is created and collected for individual patients. For example, a patient can stay in phase B for a very long time depending on the process of rehabilitation. Whilst in this phase, very fine grained information about the patients condition is documented by different members of the medical staff, e.g. doctors, therapists and the nursing staff, which is normally done handwriting. No standards exist yet for this documentation. Therefore each institution handles this process using its own standards [HBB+06]. Documentation is either persisted as a more or less structured text document or the handwritten form is scanned. Structured data, which are generally present are routine records, accounting data and sensor data. Routine records and accounting data are usually used as a foundation for analysis and reports. But the less structured documentation offers valuable insight about the rehabilitation of patients [STM+11].

This scenario raises the question of how to enable a continuous, combined analysis of poly-structured data for measuring and monitoring the progress of patients rehabilitation.

2 Determine the available Information and subsequent Questions

Expert interviews were conducted in order to determine the available information, the available structured and unstructured data and to discuss possible approaches. In the first round, IT managers from six hospitals were interviewed. Interviews with head and senior physicians were conducted in a second round. The interviews lasted between one and two hours and were carried out by one interviewer. Following the interviews, one hospital offered to provide sample data for one patient from a monitoring station. The sample data consists of structured and unstructured data. The structured data is extracted from the clinical information system whilst the unstructured data is available partly as MS Word documents and partly as scanned images.

IT managers were chosen as interview partners for a first round of interviews because they have excellent knowledge about the domain and a technical background. They can hopefully provide immediate feedback about the possibility and feasibility of the presented approach here. Topics discussed during the interviews with the IT managers included:

- General information about the organization (e.g. technical equipment, number of employees)
- Systems and data (e.g. software used for reports and analysis, collected and specially prepared data, collection for unstructured data)
- Questions (e.g. typical questions for analysis, boundaries of present solutions)
- Requirements (e.g. requirements for the processing of unstructured data, analysis which should be possible)

One of the results of the interviews is that the IT landscapes of most hospitals are comparable. There is always a settlement system and a clinical information system. All other

systems are usually connected to those two, either directly or via a communication server. This is a result of the continuous efforts to integrate clinical information system more tightly [HW08]. All hospitals provide a software solution to conduct analysis and reports. These differ in terms of their approach of gathering the necessary data. One approach is to gather all relevant data directly from all systems. An alternative approach is to gather all data from the settlement and clinical information systems. Typical questions regarding method and software solutions for a combined analyses of structured and unstructured data included:

- How does one improve the process and persistance of unstructured data?
- Is it possible to offer a simple software solution for searching unstructured data?
- For which key user is it necessary to provide the possibility to work with unstructured data?

More detailed interviews with the medical staff are currently being prepared and conducted. In a first round questions regarding the actual documentation process, shortcomings of present mechanisms, scientific research questions and requirements that a method should meet were discussed, with a focus monitoring stations. As it turned out, there exists a multitude of unstructured data in monitoring stations. Questions which rose during the interviews included:

- How would it be possible to directly monitor and analyze the success of a patients treatment?
- How can the factors for a success of a patient's treatment be identified?
- How can patients be clustered by different and changing characteristics?

An continuous, combined analyses of structured and unstructured data for improving the care situation of individual patients was considered highly desirable by all interview partners. For a first assessment of the discussed possibilities, one hospital provided sample data. The data provided is typical for patients from monitoring stations. It consists of a patients core data like name, address, birthday etc. and all case based information. Different cases as well as the patient are identified by a unique number. The structured data is gathered from the settlement and the clinical information system and is prepared for being persisted in a MS SQL database with MS Analysis Services. The data can be accessed with most Data Warehouse solutions. The MS Word and image files are structured on the file system. File names contain the patient and case numbers. The individual documents are organized in sub folders for different aspects of the treatment (e.g. medical history, diagnosis, documentation, misc etc.). A corresponding XML file exists which contains meta data (e.g. core data, case data, patient information etc.). Whilst most Word documents have a repeating structure depending on their use case, the images contain scanned, hand written documents.

3 Developing an Approach for Measuring and Monitoring the Rehabilitation of Patients on Monitoring Stations

For a recurring analysis of great amounts of data with an increasing complexity of the analysis Data Warehouses in combination with Online Analytical Processing (OLAP) have been established. Analytical Information Systems (AIS) are, in differentiation to operative transaction-oriented systems, systems which support analytical and planning activities. It is not possible to integrate unstructured data into a DWH without a great deal of effort and a considerable chance of failure. In addition, it is not possible with the given technical, organizational, economic and legal circumstances [CG10].

Most institutions run their own systems for to conduct analysis and to create reports. But those systems are limited in their flexibility and extendibility. In some cases they even offer just predefined solutions for creating reports. The work is structured by the design science research process. According to Pfaff et al. the design science research process consists of six stages, which are problem identification & motivation, objectives of solution, design & development, demonstration, evaluation and communication [PTG06]. A problem identification and motivation was given in section 1, objectives of a solution were determined through interviews with experts and potential key users, although not yet concluded, gives reasonable grounds for the desired objectives. At this stage a conceptual and a prototypical approach were developed, which will both be presented in the upcoming sections.

3.1 Conceptual Approach

Based on the findings of the interviews, various technologies were examined which could support the method and the questions obtained. With MUSTANG (Multidimensional Statistical Data Analysis Engine) a software exists which is in use in various institutions in the health care sector. It is used for conducting exploratory analysis. Whilst MUSTANG offers flexible analysis it belongs to the class of AIS and only works with an underlying DWH. Therefore a direct analysis of unstructured data is not possible [KAS13]. Hence, another tool which is capable of doing so is needed. With Apache Solr exists an extendable and adaptive solution for realizing salable search applications. Apache Solr is based on Apache Lucence for the search component and Apache Tika for handling unstructured data. The downside is, that each document has to be added to Solr's index. Following that preliminary consideration an approach consisting of four elements was developed.

Harmonization of given data sources: For being able to work with heterogeneous data in a first it is necessary to harmonize the given sources as far as possible. Thus, create a unified scheme for the structured data. Together with potential stakeholders it has to be decided which data is absolutely necessary for answering the various questions raised.

Implementation of the analysis of unstructured data: It is not possible to harmonize unstructured data in the way as structured data for integrating them into an AIS.

Therefore the application of data mining technologies is necessary to extract the required information. Moreover it is important that the unstructured data can be searched.

Development of a suitable process: Depending on the systems and sources from which data should be brought together, a process must be developed. In this process must be specified which an how the data is brought together and how it should be prepared for making it available to the analysis software.

Implementation of the necessary software applications: For conducting the analysis the development and implementation of a portfolio of software applications is necessary. It is important to ensure that new data is automatically added to the data basis. Furthermore potential end users must be able to conduct analysis on their own.

3.2 Prototypical Approach

A first and very simple prototype was developed to evaluate the technical feasibility of the discussed ideas. It was applied to the sample data and the results were discussed with potential stakeholders. The prototype consists of a MUSTANG and a Apache Solr installation. The analysis of the structured data was handled by MUSTANG. Keywords from the results were then manually added to Solr's search query. It was now possible to search and find suitable documents by their meta data. Moreover it was possible to do a full text search on all MS Word documents. A correlation between the result sets has to be made manually. At this stage, the ETL process for the data warehouse is done by the hospital and the copying process for the unstructured data has to be carried out manually. In the further development process it is planned to adjust the ETL process and implement an automated procedure for deploying the unstructured data.

As a result of this first test with the provided sample data it can be concluded that the combination of an AIS with a platform for search and retrieval of information in unstructured data is not only possible but very promising.

4 Planned Evaluation and Outlook

The involvement of potential key users and stakeholders of a solution for a combined analysis of structured and unstructured data ensures that the conceptual work and the development of the prototype are valuable and appropriate. Moreover important question formulations can only be carved out in co-operation with key users. Although the work is in a very early stage, feedback on the conceptual and a prototypical approach was very positive. For a next step more and more detailed interviews with key users are planned. It is necessary to clarify further the requirements and objectives for the method and the accompanying software solution. As for the prototype further tests are needed. In addition to this date it is uncertain how to proper integrate the AIS with the search and retrieval platform. Approaches which will be explored include integration on service level and

integration on GUI level.

For demonstrating method and prototype it is planned to use them in cooperation with an hospital to measure and monitor rehabilitation of patients from a monitoring station. For this purpose, method and prototype must be kept in development. It is planned to involve the medical staff in the whole process. Two approaches are intended for evaluation of the presented approach. On the one hand, extensive expert interviews will he held. On the other hand, the prototype will be used for retrospective analysis. The result will be compared with existing ones.

References

[BFG+14] Henning Baars, Carsten Felden, Peter Gluchowski, Andreas Hilbert, Hans-Georg Kemper, and Sebastian Olbrich. Shaping the Next Incarnation of Business Intelligence. *Business & Information Systems Engineering*, 6(1):11–16, January 2014.

[CCS12] Hsinchun Chen, Roger H. L. Chiang, and Veda C Storey. BUSINESS INTELLIGENCE AND ANALYTICS: FROM BIG DATA TO BIG IMPACT. *MIS Quarterly*, 36:1–24, 2012.

[CG10] P. Chamoni and P. Gluchowski. Analytische Informationssysteme - Einordnung und Überblick. In *Analytische Informationssysteme : Business Intelligence-Technologien und -Anwendungen*, pages 3–16. 2010.

[HBB+06] Peter U Heuschmann, Marcel K Biegler, Otto Busse, Susanne Elsner, Armin Grau, Uwe Hasenbein, Peter Hermanek, Rudolf W C Janzen, Peter L Kolominsky-Rabas, Klaus Kraywinkel, and Others. Development and Implementation of Evidence-Based Indicators for Measuring Quality of Acute Stroke Care The Quality Indicator Board of the German Stroke Registers Study Group (ADSR). *Stroke*, 37(10):2551–2573, 2006.

[HW08] Rada Hussein and Alfred Winter. Towards More Integrated Implementation of Healthcare Information Systems: Using the 3LGM2 for Modeling the IHE-Scheduled Workflow Integration Profile. *2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pages 650–652, June 2008.

[KAS13] David Korfkamp, Hans-Jürgen Appelrath, and Eunice Sirri. CARESS als integrierte Auswertungssoftware zur Berechnung von {Ü}berlebenszeitanalysen. In *GMDS 2013: 58. Jahrestagung der Deutschen Gesellschaft f{ü}r Medizinische Informatik, Biometrie und Epidemiologie e. V. (GMDS)*. Deutsche Gesellschaft f{ü}r Medizinische Informatik, Biometrie und Epidemiologie, German Medical Science GMS Publishing House; D{ü}sseldorf, 2013.

[PTG06] K Peffers, T Tuunanen, and CE Gengler. The design science research process: a model for producing and presenting information systems research. In *DESRIST*, pages 83–106, Claremont, CA, 2006.

[STM+11] E. Swart, D. Thomas, S. March, T. Salomon, and O. Knesebeck. Die Eignung von Sekundärdaten zur Evaluation eines Interventionsprojekts. *Prävention und Gesundheitsförderung*, 6(4):305–311, September 2011.

[USG10] Lothar Ullrich, Gabriele Stolecki, and Matthias Grünewald. *Intensivpflege und Anästhesie*. Thieme, 2., neu berarbeitete und erweiterte auflage edition, 2010.

Improving the Detection of Comparison Arguments in Product Reviews

Wiltrud Kessler

Institute for Natural Language Processing
Universität Stuttgart
Pfaffenwaldring 5b
70569 Stuttgart
wiltrud.kessler@ims.uni-stuttgart.de

Abstract: A common way to express sentiment about some product in a review is by comparing it to a different product. In order to get meaningful information about the comparison, we need to identify the separate parts: the predicate that expresses the comparison, the two compared entities, and the aspect that they are compared in. In this work, we assume the sentence and the predicate to be given and work on argument identification and classification. We show that syntax is more helpful than context for the task. We also improve performance on argument identification by including information about the type of the comparison. We were not able to prove our hypothesis that including information on the sentiment of the predicate improves performance on argument classification.

1 Introduction

Sentiment analysis is an area in Natural Language Processing that deals with the task of determining the polarity (positive, negative) of an opinionated document or a sentence. Sentiment analysis has attracted much attention in recent year due to the growing importance of social media where users often express their opinions about products or services. In product reviews, sentiment is usually determined with regard to some target product, e.g., the sentence "X has a good lens" expresses positive sentiment towards X. A common way to express sentiment about some product is by comparing it to a different product. Such comparisons cannot be treated the same way as non-comparative sentiment expressions, because they involve more than one target entity and may involve assignment of more than one polarity, e.g., the statement "X has a better lens than Y" expresses positive sentiment towards X and less positive or maybe even negative sentiment towards Y.

For our purposes we define a comparison to be any statement about the similarity or difference of two entities. Besides the linguistic category of comparative sentences, this includes a wide variety of expressions found in user generated texts, such as "X blows away all others", "X and Y have the same sensor", or "X wins over Y". We call the word or phrase that is used to express the comparison the *comparative predicate*. A comparative predicate has three arguments: The two *entities* that are compared and the *aspect* they are compared

in. In our data, most of the entities are products (cameras, cars, phones). The term aspect denotes any attribute or part of the entity that is being compared, e.g., the lens, size or resolution of a camera. In the sentence "X has a better lens than Y" the word "better" is the comparative predicate, "X" and "Y" are the two entities and "lens" is the aspect.

In this short paper, we start with a given comparison sentence and a given comparative predicate. The task we want to solve is to find the corresponding arguments. We build on existing work [KK13] and attempt to improve argument identification and classification based on the following hypotheses: Syntax information is more helpful than context information and including information on the sentiment of the predicate and the type of the comparison improves performance.

2 Related Work

With the growing importance of user generated content in the Web 2.0, sentiment analysis (or opinion mining) which often deals with the opinions expressed in such texts has become an active research area. For an overview of methods and challenges see [Liu12]. The focus has over time shifted from document-level approaches that determine the overall polarity of a whole document to more fine-grained analyses that determine sentiment in smaller units and with respect to some target entity. One of the challenges on sentence level are comparison sentences. Jindal and Liu [JL06a] are the first to identify comparison sentences by using class sequential rules based on keywords as features for a Naive Bayes classifier. In this work we assume that we are given a set of such sentences.

Several approaches have been presented for the detection of comparison predicates and arguments. In follow-up work on their sentence identification, Jindal and Liu [JL06b] detect comparison arguments with label sequential rules and in a second step identify the preferred entity in a ranked comparison [GL08]. Approaches inspired by semantic role labeling that detect predicates and subsequently their arguments have been used for Chinese [HL08] and English [KK13]. Xu et al. [XLLS11] use conditional random fields to extract relations between two entities, an attribute and a predicate phrase.

We build upon the work of Kessler and Kuhn [KK13] who use a three step approach: predicate identification, argument identification and argument classification. In this work, we assume predicates to be given and design features to improve the last two steps.

3 Approach

The input to our system is a sentence where at least one comparative predicate has been identified. The result of our processing are three arguments for each predicate: The two entities that are being compared, and the aspect they are compared in. Any of the arguments may appear more than once or be empty.

We use the MATE Semantic Role Labeling system [BHN09] with default settings and

without the reranker. This is equivalent to the setup in [KK13]. We start with the predicates already identified. Our system performs two steps. In a first binary classification step, all arguments are identified (*argument identification*). In a second step, for each identified argument it is determined whether it is entity 1, entity 2 or aspect (*argument classification*). The system uses regularized linear logistic regression for classification.

We vary feature sets for the experiments, but always use the same feature sets for both steps. For every word in question we extract its form and the part of speech (e.g., noun, verb) as word features. The given predicate and the current argument candidate are always accessible for extracting word features. The easiest way to add more information is to include the context words before and after the candidate argument. Alternatively, the sentence can be parsed and information from the syntactic structure can be used. We use dependency parses that model the structure as a set of relations between each word and its parent. The parser is trained on news texts but applied to texts from blogs and reviews. These texts contain many errors in spelling, grammar and punctuation, so the parses may not always be correct which might lead to confusing information for the classifier. We also introduce two features to model additional sentiment and comparison information.

In our experiments we compare the following feature sets:

Base Word features extracted from current argument and predicate, position of the argument relative to the predicate (before, after or same word).

Context Base features + Word features for the three context words to the left and right of the current argument.

Syn Base features + Features extracted from the output of the MATE dependency parser [Boh10], including word features for the argument children, argument siblings, predicate parent, predicate children and syntactic path between predicate and argument. This corresponds to the setting used in [KK13]. For more detailed descriptions of the features see [BHN09].

Sent Feature to model the polarity of the predicate as defined by the MPQA list of subjectivity clues[1] [WWH05] containing 2304 positive and 4152 negative words. Possible values are positive, negative and neutral.

Type Feature to model the type of the comparison. There are four possible types (not all of them are annotated in all datasets, see data description):

- ranked comparisons (also called non-equal gradable) in which one entity is ranked as better or worse as the other, e.g., "X has higher resolution than Y",
- equative comparisons in which both entities are ranked as being equally good or bad, e.g., "X and Y have the same sensor",
- superlative comparisons in which one entity is ranked as better or worse than all other entities, e.g., "X is the best", and
- differences in which a difference between the two entities is stated, but no ranking is introduced, e.g., "picture quality is different in X and Y".

[1] `http://www.cs.pitt.edu/mpqa/subj_lexicon.html`

4 Experiments

Data. We use four datasets in our experiments: the cameras and cars datasets from the JDPA corpus[2] [KECN10], the J&L dataset[3] [JL06b], and the IMS camera review dataset[4] [KK14]. We process the data in the same way as described in [KK13] except that we add information about the type of the comparison. For J&L and IMS we use the annotations provided in the data (equative, ranked, superlative in J&L; equative, ranked, superlative, difference in IMS). For the JDPA corpus, we use the type equative if the value in the "same" slot is set to "true", otherwise we assume the type ranked. Further, we map the "scale" annotation to "aspect" in the IMS data to make the task identical to the other datasets. In IMS and J&L, the two entities are distinguished by order of appearance in the sentence. In the JDPA datasets, entities are identified as the one evaluated as better (mapped to entity 1), and the one evaluated as worse (mapped to entity 2). We extract all sentences where we find at least one comparative predicate as our data. We use annotated predicates (gold predicates) as a starting point for the experiments, as our features are tailored to arguments and do not improve predicate identification.

Evaluation Setup. We evaluate on each dataset separately using 5-fold cross-validation. We report precision (P), recall (R), and F1-measure (F1). Bold numbers denote the best result in each column and task. We mark a F1-measure result with * if the difference to *Syn* is statistically significant at $p < .05$ using the approximate randomization test [Nor89] with 10000 iterations.

Results. Table 1 shows the results for argument identification, Table 2 those for argument classification. We can see that using information from argument and predicate alone (*Base*) performs very poorly. Using syntactic information (*Syn*) outperforms using only context information (*Context*) by a considerable margin in both tasks in all datasets. The difference is significant (except for JDPA cam, entity 1 where $p = 0.0833$). We therefor use the syntactic features in the following to judge the impact of the two other features. Adding sentiment information (*Syn+Sent*) does not improve performance, it often even hurts. We were hoping for an improvement in distinguishing entity 1 from entity 2 in the JDPA data, where entity 1 (the preferred entity) occurs before the predicate for positive predicates ("A is better than B") and after the predicate for negative predicates ("B is worse than A"). The result may be due to the fact that we only consider the sentiment of the predicate itself without taking into account any context like negation that might reverse the polarity. Adding information about the type of the comparison (*Syn+Type*) improves results in nearly every setting. The difference is significant for argument identification, but only in some cases for argument classification.

[2] http://verbs.colorado.edu/jdpacorpus/

[3] http://www.cs.uic.edu/~liub/FBS/data.tar.gz

[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/reviewcomparisons/

	JDPA cam			JDPA car			J&L			IMS cam		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Base	**69.7**	12.5	21.2*	55.2	9.5	16.2*	57.8	13.1	21.4*	56.3	21.2	30.8*
Context	63.4	26.4	37.3*	60.3	27.0	37.3*	64.0	30.7	41.5*	66.0	43.3	52.3*
Syn	69.0	36.6	47.8	70.2	41.6	52.2	67.4	44.5	53.6	78.1	57.2	66.0
Syn+Sent	69.0	35.9	47.3*	**70.4**	41.3	52.1	**68.5**	44.4	53.9	**78.7**	57.0	66.1
Syn+Type	65.6	**38.8**	**48.7***	68.3	**43.0**	**52.8***	65.0	**48.3**	**55.4***	72.8	**63.0**	**67.6***

Table 1: Results for argument identification on gold predicates (Precision, Recall, F1-measure)

		JDPA cam			JDPA car			J&L			IMS cam		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Entity 1	Base	30.0	1.7	3.3*	19.5	2.4	4.2*	52.6	15.1	23.5*	47.6	20.2	28.3*
	Context	40.0	14.3	21.0	38.1	17.5	24.0*	55.0	34.6	42.5*	60.4	40.8	48.7*
	Syn	**41.5**	17.3	24.5	39.9	22.4	28.7	56.4	46.6	51.0	65.9	49.8	56.7
	Syn+Sent	41.0	16.6	23.6	**40.1**	22.2	28.5	**57.8**	46.6	51.6*	**66.5**	49.4	56.7
	Syn+Type	37.8	**18.3**	**24.7**	39.2	**24.0**	**29.7***	55.5	**49.9**	**52.5***	60.3	**56.0**	**58.1***
Entity 2	Base	29.4	2.9	5.3*	29.2	3.3	5.9*	46.6	8.1	13.8*	48.4	23.6	31.7*
	Context	36.8	13.7	20.0*	47.5	19.6	27.8*	**62.0**	22.5	33.0*	56.5	36.9	44.7*
	Syn	45.7	22.7	30.3	49.7	**30.7**	**37.9**	57.1	32.6	41.5	64.8	49.6	56.2
	Syn+Sent	**46.0**	22.7	30.4	**50.2**	30.5	**37.9**	59.4	34.1	43.3*	**65.6**	49.5	56.4
	Syn+Type	42.9	**24.9**	**31.5**	47.5	30.3	37.0*	56.7	**36.8**	**44.6***	60.4	**54.3**	**57.2***
Aspect	Base	**78.7**	27.3	40.5*	**62.9**	17.6	27.5*	46.9	8.7	14.7*	47.1	11.3	18.2*
	Context	70.7	35.6	47.4*	53.8	25.1	34.2*	52.6	19.3	28.3*	54.5	34.8	42.5*
	Syn	72.2	47.0	56.9	60.0	36.0	45.0	60.7	30.7	40.8	72.5	49.6	58.9
	Syn+Sent	72.4	46.4	56.6	59.7	35.7	44.7	**61.8**	30.7	41.0	**72.8**	49.5	58.9
	Syn+Type	70.0	**48.3**	**57.2**	59.4	**38.1**	**46.4***	55.3	**33.3**	**41.6**	68.0	**52.8**	**59.4**

Table 2: Results for argument classification on gold predicates (Precision, Recall, F1-measure)

5 Conclusions

In this short paper we present experiments on improving the identification and classification of arguments (entity 1, entity 2, aspect) of comparisons in product reviews. We show that syntax information is more helpful than context information, and that including the type of the comparison improves performance. We were not able to prove our hypothesis that including information on the sentiment of the predicate improves performance. This may be due to the fact that our feature does not take into account any context like negation that might act as a polarity reverser. For future work, we want to further investigate the interaction of sentiment and comparison arguments. We also plan to address the inherent diversity of expressions typical for user generated content by using generalization techniques, e.g., to detect product names.

Acknowledgments

The work reported in this paper was supported by a Nuance Foundation grant.

References

[BHN09] Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual Semantic Role Labeling. In *Proceedings of CoNLL '09 Shared Task*, pages 43–48, 2009.

[Boh10] Bernd Bohnet. Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of COLING '10*, pages 89–97, 2010.

[GL08] Murthy Ganapathibhotla and Bing Liu. Mining opinions in comparative sentences. In *Proceedings of COLING '08*, pages 241–248, 2008.

[HL08] Feng Hou and Guo-hui Li. Mining Chinese comparative sentences by semantic role labeling. In *Proceedings of ICMLC '08*, pages 2563–2568, 2008.

[JL06a] Nitin Jindal and Bing Liu. Identifying comparative sentences in text documents. In *Proceedings of SIGIR '06*, pages 244–251, 2006.

[JL06b] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *Proceedings of AAAI '06*, pages 1331–1336, 2006.

[KECN10] Jason S. Kessler, Miriam Eckert, Lyndsay Clark, and Nicolas Nicolov. The 2010 ICWSM JDPA Sentiment Corpus for the Automotive Domain. In *Proceedings of ICWSM-DWC '10*, 2010.

[KK13] Wiltrud Kessler and Jonas Kuhn. Detection of Product Comparisons - How Far Does an Out-of-the-box Semantic Role Labeling System Take You? In *Proceedings of EMNLP '13*, pages 1892–1897, 2013.

[KK14] Wiltrud Kessler and Jonas Kuhn. A Corpus of Comparisons in Product Reviews. In *Proceedings of LREC '14*, pages 2242–2248, 2014.

[Liu12] Bing Liu. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Nor89] Eric W. Noreen. *Computer-intensive methods for testing hypotheses – an introduction*. Wiley & Sons, 1989.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT '05*, pages 347–354, 2005.

[XLLS11] Kaiquan Xu, Stephen Shaoyi Liao, Jiexun Li, and Yuxia Song. Mining comparative opinions from customer reviews for Competitive Intelligence. *Decis. Support Syst.*, 50(4):743–754, 2011.

Wertanalyseverfahren für Kundenanforderungen

David Kuhlen
NORDAKADEMIE
david.kuhlen@datenlotsen.de

Abstract: In dieser Arbeit wurde ein Verfahren zur Anforderungsbewertung entwickelt, dass für die Probleme heterogener Bewertungsmethoden ein kombinierten Lösungsansatz darstellt. Dabei entstand ein Vorgehensmodell, das auf drei Säulen ruht: Einem Phasenmodell, das Akteuren je Dimension Methoden zuordnet; einem Prozessmodell, das den Prozess mit dem Phasenmodell verbindet und Szenarien, die einem Hersteller die Anwendung der Methoden ermöglichen.

1 Einleitung

Softwaresysteme werden profitabel, wenn ihre Qualität maximiert, ihre Kosten minimiert und ihre Entwicklungszeit gekürzt werden. Die Praxis zeigt jedoch, dass diese Ansprüche sich oft widersprechen und zwischen konkurrierenden Zielen abzuwägen ist (Vgl. [Ho08]). In vielen Fällen reicht die Entwicklungszeit nicht aus, um alle sinnvollen Features umzusetzen (Vgl. [Bo84]). Ohne eine Wertanalyse besteht die Gefahr, dass gewinnbringende Anforderungen nicht als solche erkannt und nicht umgesetzt werden (Vgl. [SR13]). Dazu ist die Forschungsfrage zu beantworten: „Wie lässt sich der Wert einer Kundenanforderung durch einen Softwarehersteller ermitteln?“. Es ist ein Verfahren zu entwickeln, das den Wert einer Kundenanforderung aus der Sicht eines Softwareherstellers zeigt. Dieses Verfahren soll die Umsetzung von Softwareanforderung mit einer betriebswirtschaftlichen Planung absichern. Es wurde für den Einsatz in Scrum-Projekten optimiert, weil bei Scrum die Untersuchung des Wertes aufgrund knapper Spezifikationen vor besonderen Herausforderungen steht.

2 Investitionen in Software und die Bewertung von Anforderungen

Investition ist definiert, als Bindung von Geldmitteln in Gütern des Anlagevermögens, wobei es sich bei Investitionen um Zahlungsreihen handelt, die mit einer Auszahlung beginnen (Vgl. [DG07]). Die Umsetzung einer Anforderung bindet Arbeitsstunden, zur Entwicklung von Funktionalität. Die entstehende Software kann, als immaterielles Vermögen, in der Bilanz aktiviert werden. Die Entwicklung erfordert zunächst Auszahlungen (Lohn der Entwickler) die später durch Einzahlungen beglichen wird. Eine Methode zur Investitionsrechnung muss das Verhältnis von Aus- und Einzahlungen bewerten. Bei dynamischen Methoden zur Investitionsrechnung wird die Zeitpräferenz, der Zahlungen berücksichtigt. Weitere Methoden erlauben es zusätzlich, Unsicherheiten bei der Investitionsrechnung zu berücksichtigen. Für die Bewertung von IT-Investitionen ist die Berücksichtigung qualitativer Effekte von besonderer Bedeutung. Insgesamt

müssen zur Bewertung von Anforderungen darum die Dimensionen Einzahlungen, Auszahlungen, Unsicherheiten, qualitative Effekte und die Zeitstruktur bewertet werden.

> Eine Anforderung beschreibt ein Merkmal, das die Lösung erfüllen muss. Sie bilden die Brücke zwischen der Ist- und der Soll-Situation (Vgl. [Br05]).

Indem eine Anforderung ein Lösungsmerkmal beschreibt, hat sie zwei Charakteristika. In erster Linie hilft sie bei der technischen Orientierung. So legt die Anforderungsdefinition fest, was das Softwaresystem leisten soll und bündelt gleichzeitig ein Zeit- und Ressourcenbudget im Projektplan (Vgl. [GM05]). Um die Soll-Situation zu definieren müssen Anforderungen ausreichend genau spezifiziert werden. So werden in einer Anforderung die späteren Schritte des Projektvorgehens detailliert spezifiziert (Vgl. [FLT92]). Die Spezifikation ist der Ausgangspunkt für die Bewertung.

3 Probleme mit der Wertanalyse von Anforderungen

Softwarehersteller entscheiden sich für die Umsetzung von Anforderungen wegen ihrer Nutzeneffekte, die sich in direkten und indirekten Einzahlungen auswirken. Gegenüber direkten Einzahlungen (etwa aus dem Verkauf einer Software), ist bei indirekten Einzahlungen teilweise unklar, wo und in welcher Höhe Erträge entstehen (Vgl. [DG07]). Erträge aus einer IT-Investition entstehen verteilt über die Organisation (Vgl. [II01]). Unternehmen wissen, dass es durch IT-Investitionen zu unerwartbaren Erträgen kommen kann (Vgl. [BS98]). Verglichen mit dem Problem der Identifikation von Ertragschancen gilt, die Quantifikation des Ertrags, als das größere Problem (Vgl. [BS98]). Der Ertrag einer IT-Investition kann schwer gemessen- und in finanziellen Größen quantifiziert werden (Vgl. [Bo84], Vgl. [FLT92], Vgl. [BH96], Vgl. [Ok06], Vgl. [SR13]). Nach der Umsetzung einer Anforderung können Erträge ausbleiben oder anders als geplant anfallen (Vgl. [Bo84], Vgl. [Br93], Vgl. [Br05]). So sind Nutzeneffekte unsicher weil verschiedener Faktoren die Anforderung beeinflussen können (Vgl. [SR13]). Die Kosten einer Software richtig einzuschätzen ist ein bedeutendes Problem der Wirtschaftsinformatik und somit bei Investitionsentscheidungen relevant (Vgl. [Bo84]). In den meisten Fällen unterschätzen Unternehmen die Kosten eines IT-Projektes (Vgl. [Ho90]). Für Softwarehersteller hängen die Kosten einer Anforderung maßgeblich vom Zeitaufwand ab, den die Entwickler zur Umsetzung benötigen. IT-Investitionen haben keine deterministischen Wirkungen und bergen Risiken, die den Wert beeinflussen (Vgl. [Hi05]). Typischerweise besteht während Entwicklungsprojekten das Risiko, dass Kunden ihre Anforderungen verändern. Fehler bei der Anforderungsdefinition wirken auf die Projektplanung. Wenn die Verteilung von Ressourcen zur Anforderungsumsetzung falsch ist, wird die langfristige Flexibilität des Entwicklungsprojektes gefährdet (Vgl. [II01]). Weil die Fehlerfreiheit einer Software kaum wirtschaftlich nachweisbar ist, bleibt weiter das Risiko von Softwarefehlern bestehen (Vgl. [KP96]). Die Erfüllung eines übergeordneten Kundenwunsches, dem mehrere Anforderungen zugeordnet sind, kann zu Abhängigkeiten zwischen den Anforderungen führen. Die Rendite einer Anforderung kann von den Risiken einer anderen Anforderung abhängen (Vgl. [WHS05]). Investitionen in Software haben den Ruf, besonders riskant zu sein,

weswegen höhere Renditen erwartet werden (Vgl. [BH96]). Aus diesem Grund verteuern sich Anforderungen künstlich und die Wirtschaftlichkeit der Anforderung leidet. IT-Investitionen haben vielfach qualitative Wirkungen, die per Definition schwer in monetäre Größen überführt werden können. Der Nutzen von IT-Investitionen kann z. B. in den Größen Effizienz, Qualität oder Effektivität ausdrückt werden (Vgl. [Po93]). Klassische Methoden zur Investitionsrechnung lassen sich nur für quantifizierbare Nutzenwirkungen (Zahlungen) einsetzen (Vgl. [BS98]). Weil qualitative Effekte schwer zu berücksichtigen sind, müssen ihre Zahlungswirkungen geschätzt werden. Investitionsentscheidungen basieren in diesen Fällen auf Annahmen, wobei subjektive Einflüsse oder gezielte Manipulationen das Ergebnis verändern können (Vgl. [II01]). Ferner fehlt bei der Investitionsrechnung vielfach der Überblick über alle Effekte um diese in der Entscheidung zu berücksichtigen (Vgl. [SR13]). Dieses führt dazu, dass gewisse Wirkungen nicht in die Wirtschaftlichkeitsanalyse einbezogen werden. Kostenabschätzungen fallen im Zeitablauf des Projektes, von der Planung bis zur Umsetzung, immer leichter (Vgl. [Bo84]). Oft ist die Veränderung der Kundenprozesse abzuwarten, bevor der Nutzen aus einer umgesetzten Anforderung vollständig sichtbar wird (Vgl. [BH96]). Oft fehlen Lebenszyklusmodelle und wichtige Informationen, um den Zahlungsverlauf bei der Wertanalyse zu berücksichtigen (Vgl. [ZSB04]).

4 Entwicklung eines Vorgehensmodells zur Wertanalyse

4.1 Konzeption eines Vorgehensmodells

Es ist ein Wertanalyseverfahren zu entwickeln, um Softwareherstellern zu ermöglichen die Anforderungen auszuwählen, die den Wert des Betriebsergebnisses maximieren. Keine eigenständige Methode kann zur Analyse des Wertes von Kundenanforderungen geeignet sein. Sinnvoll ist daher der Einsatz verschiedener Bewertungsmethoden, die sich jeweils für Kalkulation einer Dimension eignen. Verschiedene Methoden können in einem Vorgehensmodell in Phasen und Prozesse eingebunden werden (Vgl. [Ha12]). Dazu beschreibt ein Vorgehensmodell in welcher Reihenfolge Aktivitäten durchgeführt werden (Vgl. [Wo13]), um die Bewertungsmethoden zu orchestrieren. Auch die Verantwortlichkeiten für den Einsatz der Bewertungsmethoden und die Ergebnisse der Bewertung werden im Vorgehensmodell koordiniert (Vgl. [Ba96]) zitiert bei (Vgl. [Fä09]). Dabei sind die anzuwendenden Bewertungsmethoden austauschbar. Die Auswahl von Bewertungsmethoden sollte vom Geschäftsmodell und der Branche des Endanwenders abhängen, weil diese Methoden Dimensionsgrößen wie z. B. die Zeitstruktur maßgeblich beeinflussen. Das Vorgehensmodell sieht vor, dass die Analyse des Wertes in jeder Phase von Scrum zyklisch wiederholt wird. Scrum legt eine Unterscheidung der Phasen „Backlog“, „Sprint Planning“, „Sprint“ und „Review“ nahe.

4.2 Wertanalyse entlang des Softwareentwicklungsprozesses

In der **Backlog Phase** wird die Wertanalyse zur Ermittlung der relativen Vorteilhaftigkeit (und später zur Priorisierung) einer Anforderung eingesetzt. Grundlage

dazu ist die Anforderungsdefinition. Die relative Vorteilhaftigkeit wird maßgeblich durch den Product Owner ermittelt. Die Nutzenanalyse beginnt mit der Anforderungsaufnahme. Danach werden die einzelnen Dimensionen analysiert. Für die Anforderungsaufnahme werden Anforderungsschablonen-, und für die Analyse der einzelnen Dimensionen standardisierte Methoden zur Investitionsbewertung eingesetzt. Wesentlich ist die Ermittlung der Einzahlungen. Wenn für die Software kein Preis feststeht, können die erwartbaren Einzahlungen durch die Wirkungskettenanalyse ermittelt werden. Unabhängig von der Art der IT-Investition sind stets wiederkehrende Nutzeffekte zu beobachten (Vgl. [SK09]). Die Wirkungskettenanalyse kann mit einer Prozessanalyse kombiniert werden. Um Einzahlungen zu ermitteln, kann eine Prozessanalyse die Marktleistung des Kunden im Ist- und im Sollzustand erfassen (Vgl. [Br05a]). Die Wertanalyse in der **Planning Phase** baut auf den vorherigen Ergebnissen auf. Eine genauere Aufwands- und Risikoschätzung wird mit den bereits gewonnenen Erkenntnissen kombiniert. Für die genaue Aufwandsschätzung ist das Entwickler Team verantwortlich. Entwickler können den Aufwand am besten einschätzen, da sie das Produkt erstellen. Das Entwickler Team kann Anforderungen zunächst mit der Function Point Method bewerten. Dazu werden die Funktionen analysiert, die für die Umsetzung der Anforderungen neu entwickelt oder angepasst werden müssen. So wird gleichsam die Entwicklung vorbereitet. In der **Sprint Phase** werden die ausgewählten Anforderungen umgesetzt. Insbesondere die kritischen Werte und die Werttreiber geben bei Entwicklung eine Orientierungshilfe. Trotz guter Planung kann es bei der Ausführung zu Abweichungen kommen, etwa durch Fehler oder unzureichende Mitarbeiterleistungen (Vgl. [WD08]). Auf Basis der geplanten Ergebnisse der vorherigen Phasen können die realisierten Lösungen in die Wertanalyse einbezogen werden. In täglichen Meetings (den sog. Daily Scrums) schätzt das Entwicklungsteam seinen Fortschritt in Richtung des Sprintziels (Vgl. [SS11]). Im Daily Scrum informiert sich das Entwicklungsteam über den Fertigstellungsgrad von Anforderungen und über Hindernisse. In diesem Rahmen kann die Wertanalyse wiederholt werden. Die Ermittlung des realisierten Wertes einer Anforderung kann in der **Review Phase** erfolgen. Dazu wird die fertiggestellte Softwarelösung einer (Expost) Wertanalyse unterzogen. Dieses hat den Vorteil, dass für bestimmte Dimensionsgrößen bereits Werte feststehen (wie etwa die tatsächliche Entwicklungszeit). Das Vorgehensmodell empfiehlt, dass der Product Owner anhand der Software die erwartbaren Einzahlungen erneut untersucht. Dazu können erneut die Wirkungskettenanalysen vorgenommen- und mit Prozessanalysen kombiniert werden. Durch einen vorher/nachher Vergleich von Prozessen werden die Kosteneinsparungen sichtbar (Vgl. [Hi05]). Schließlich führt die Optimierung der Geschäftsprozesskosten zum Ergebnisbeitrag der IT-Investition (Vgl. [BES05]).

5 Tragfähigkeit des Verfahrens und Ausblick

Für die Evaluation der Tragfähigkeit wurde das Modell mit den Anforderungen der Scrum Akteure verglichen. Der **Product Owner erwartet**, dass mit dem Verfahren die richtigen Anforderungen gewählt werden, um den Produktwert zu maximieren. Das Vorgehensmodell ermöglicht die Anforderungsuntersuchung, indem für die Phasen der Entwicklung einzelne Methoden vorgeschlagen werden. Die Kenntnis über die

Werttreiber erleichtert es dem Product Owner die inhaltliche Ausgestaltung einer Anforderung festzulegen. Seine Kontrollmöglichkeiten werden dadurch gesteigert. Die **Erwartungen des Scrum Masters** beziehen sich auf die Unterstützung des Projektvorgehens durch die Wertanalyseverfahren. Ihm gibt das Wertanalyseverfahren die Möglichkeit auf die Erreichung der Top-Werthebel zu achten. Abweichungen werden schnell angezeigt und der Scrum Master kann Gegensteuerungsmaßnahmen einleiten. Die Erwartung des Scrum Masters, Verschwendung zu vermeiden, wird durch den modularen Aufbau des Verfahrens erfüllt. Nicht in jeder Phase müssen alle Dimensionen erneut analysiert werden. Die Prozesse der Wirtschaftlichkeitsanalyse stellen sicher, dass die entwickelte Software wirtschaftlich ist und zeigen die erreichte Wertschöpfung durch die **Entwickler**. Außerdem soll das Vorgehensmodell Entwicklern die Entscheidung über eine sinnvolle Implementierung erleichtern. Indem das Vorgehensmodell die Werttreiber einer Anforderung zeigt, gibt es den Entwicklern eine Orientierungshilfe bei ihren Entscheidungen.

Das vorliegende Verfahren hilft Softwareherstellern das Sprintziel festzulegen. In einer weiteren Forschungsarbeit ist eine genauere Untersuchung der Aufwands-Dimension angebracht. Vor der Prämisse eines steigenden Kostendrucks in der Softwarebranche müssen Softwarehersteller das Controlling ihrer Prozesskosten verbessern.

Acknowledgement

Mein besonderer Dank gilt Herrn Prof. Dr. Andreas Speck (Universität zu Kiel) und Herrn Prof. Dr. Hinrich Schröder (NORDAKADEMIE) für ihre Betreuung.

Literaturverzeichnis

[Ba96]: Balzert, H.: Lehrbuch der Software-Technik, Spektrum Akademischer Verlag Heidelberg Berlin, Oxford, 1996, zitiert bei: Fähnrich (2009).

[BES05]: Buchta, D.; Eul, M.; Schulte-Croonenberg, H.: Strategisches IT-Management: Wert steigern, Leistung steuern, Kosten senken, Gabler / GWV Fachverlag GmbH, 2. Auflage, Wiesbaden 2005.

[BH96]: Brynjolfsson, E.; Hitt, L.: Paradox Lost? Firm-level Evidence on the Returns to Information Systems Spending, in: Management Science, 42. Jg. (1996) 4, S. 541 – 558.

[Bo84]: Boehm, B.: Software Engineering Economics, in: IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. SE-10, NO. 1 (1984), JANUARY 1984.

[Br05]: Brugger, R.: Der IT Business Case, Springer-Verlag Berlin Heidelberg, Berlin 2005.

[Br05a]: Brugger, R.: IT-Projekte strukturiert realisieren, Friedrich Vieweg & Sohn Verlag / GWF Fachverlage GmbH, 2. Auflage, Wiesbaden 2005.

[Br93]: Brynjolfsson, E.: The Productivity Paradox of Information Technology, in: Communications of the ACM, 36. Jg. (1993) 12, S. 67 – 77.

[BS98]: Ballantine, J.; Stray, S.: Financial appraisal and the IS/IT investment decision making process, in: The journal of Information Technology (1998) 13, S. 3 – 14.

[DG07]: Däumler, K. -D.; Grabe, J.: Grundlagen der Investitions- und Wirtschaftlichkeitsrechnung, NWB Studium Betriebswirtschaft, Verlag Neue Wirtschaft-Briefe GmbH & Co. KG, 12. Auflage, Ettenheim 2007.

[Fä09]: Fähnrich, K., P.: Vorlesung Softwaretechnik - Vorgehensmodelle, V-Modell XT -, Universität Leipzig, Institut für Informatik Betriebliche Informati-onssysteme, 2009, URL: http://bis.informatik.uni-leipzig.de/ de/Lehre/0910/WS/LV/SWT/files?get=2009w_swt_v_03.pdf, [Stand: 17.01.2014].

[FLT92]: Farbey, B.; Land, F.; Targett, D.: Evaluating investments in IT, in: Journal of Information Technology (1992) 7, S. 109 – 122.

[GM05]: Gadatsch, A.; Mayer, E.: Masterkurs IT-Controlling, Friedr. Vieweg & Sohn Verlag / GWV Fachverlage GmbH, 2. Auflage, Wiesbaden 2005.

[Ha12]: Hafenrichter, B.: Vorgehensmodelle zur Softwareentwicklung, Fachhochschule Flensburg, 2012, URL: http://fbim.fh-regensburg.de/~hab39652/SE/ Kapitel_2a_Vorgehensmodelle_Kurz.pdf, [Stand: 17.01.2013].

[Hi05]: Hirschmeier, M.: Wirtschaftlichkeitsanalysen für IT-Investitionen, WiKu-Verlag - Verlag für Wissenschaft und Kultur Dr. Stein & Brokamp KG, Berlin 2005.

[Ho08]: Hoffmann, D., W.: Software-Qualität, Springer 2008

[Ho90]: Hochstrasser, B.: Evaluating IT investment matching techniques to projects, in: Journal of Information Technology, (1990) 5, S. 215 – 221.

[Il01]: Irani, Z.; Love, R.: Developing a frame of reference for ex-ante IT/IS investment evaluation, in: European Journal of Information Systems, (2001) S. 1 – 9.

[KP96]: Kitchenham, B.; Pfleeger, S., L.: Software Quality: The exclusive Target, in: IEEE Software 0740-7459 (1996), S. 12 – 21.

[Ok06]: Okujava, S.: Wirtschaftlichkeitsanalysen für IT-Investitionen, WiKu-Verlag: Wissenschaftsverlag und Kulturedition Dr. Stein, Duisburg 2006.

[Po93]: Post, B., Q.: A Business Case Framework for Group Support Technology, in: Journal of Management Information Systems, Vol. 9, No. 3, (1992), S. 7 – 26.

[SK09]: Schröder, H.; Kesten, R.: Wirtschaftlichkeitsprognose und -kontrolle von IT-Investitionen: Abschlussbericht und Fallbeispiel, Arbeitspapiere der Nordakademie, Nr. 2009-03, Elmshorn 2009.

[SR13]: Surka, M.; Rogers, C: Maximizing the Value of Information Technology, in: CFO PUBLISHING, LLC, MARCH (2013), S. 9 – 20.

[SS11]: Schwaber, K.; Sutherland, J.: Scrum Guide Der gültige Leitfaden für Scrum: Die Spielregeln, o. O. Oktober 2011, URL: https://www.scrum.org/Portals/0/Documents/Scrum%20Guides/Scrum%20Guide%20-%20DE.pdf, [Stand: 12.12.2013].

[WD08]: Wöhe, G.; Döring, U.: Einführung in die Allgemeine Betriebswirtschaftslehre, Verlag Franz Vahlen, 23. Auflage, München 2008.

[WHS05]: Wehrmann, A.; Heinrich, B.; Seifert, F.: Quantitatives IT-Portfoliomanagement: Risiken von IT-Investitionen wertorientiert steuern, in: Wirtschaftsinformatik, 48, 4 (2006), S. 234 – 245.

[Wo13]: Wolf, M., R.: Software Engineering Phasenmodelle, Universität Wuppertal, BWL / Wirtschaftsinformatik, o. O., o. J., URL: http://winfor.uni-wuppertal.de/fileadmin/wolff/Downloads/Grundstudium/EWI/Phasenmodelle.pdf, [Stand: 16.01.2013].

[ZSB04]: Zarnekow, R.; Scheeg, J.; Brenner, W.: Untersuchung der Lebenszykluskosten von IT-Anwendungen, in: WI – Schwerpunktaufsatz, WIRTSCHAFTSINFORMATIK 46 (2004) 3, S. 181–187.

Towards a holistic Approach for Service Management in the Technical Customer Service Domain

Deniz Özcan

Osnabrück University
Information Management and Information Systems
Katharinenstraße 3
49074 Osnabrück
deniz.oezcan@uni-osnabrueck.de

Abstract: The Technical Customer Service (TCS) represents an essential part of the so-called Product-Service System (PSS) and requires intelligent support in its product-related operations with technical goods and knowledge intensive activities. This makes information to a fundamental contribution for a qualitative service delivery. Therefor, IT-based mobile service support systems are seen as a necessity for faster and better information attainment in service operations. Until now, studies exist investigating the usage and development of mobile service support systems but not specifically targeting the field of the Technical Customer Service. Due to this lack, this paper purposes to fill the gap by providing a holistic approach for an IT-based mobile service management in the TCS domain.

1 Introduction

Customers are increasingly demanding integrated solutions to meet their individual needs. For this purpose, manufacturing industry involves services, products and systems to respond to customer demands through product extension services [HLBT14]. Also known as Product-Service Systems (PSS) the combined provision of goods and services offers advantages and potentials for providing companies, like monetary benefits [BeBK09], a clear distinction from competitors and the enhancement of customer satisfaction [KuWD09]. The machinery and plant engineering and it's accompanying Technical Customer Service (TCS) represent such a PSS provider that is characterized by the development, production and provision of products and services to customers. The TCS operates at customers' site with the aim to handle product-specific requirements [Vdma08]. The quality of services is mainly influenced by the TCS what makes the support of the service technician essential for service delivery [ÖNFM13]. Hence, mobile assistance systems are currently used by the TCS to enhance and improve service processes and to increase the autonomy and flexibility of the service technician. In order to allow the efficient operation of the TCS, it is important to identify requirements of a mobile technical assistance system [MFÖK13]. Studies exist considering the usage of mobile assistance systems in business but to date, a systematic approach for the development of mobile service support systems for the TCS is still missing. The underlying research question in this paper is: How can a mobile service support system for the TCS be developed? The research question is adressed by

providing fundamental concepts for IT-supported TCS processes as a signpost for the specification of mobile TCS support systems.

2 Related Work and Research Approach

Most studies considering mobile assistance systems topical focus the healthcare and education sector. Only few studies examine the use of mobile devices and the generated advantages in operational procedures. Major trends have been found in the field service to gain improvements in productivity and response time by mobile devices [Dutt12] but not specifically in the TCS domain. This paper purposes to provide a holistic approach for an integrated service management in the domain of the IT-supported TCS. Thereby, prior research has been done in the underlying dissertation proposal. For the development of a mobile service support system for the TCS it is essential to identify existing requirements from relevant actors. To this end, Requirements Engineering (RE) was conducted using a multi-method approach comprising shadowing of real TCS-processes and semi-structured interviews within a company of the machinery and plant engineering industry [MFÖK13]. The results of the RE process are used as foundation for the use case derivation. Applying the use case concept according to Cockburn [Cock99] a use case-driven approach to the design of service support systems was evolved [ÖNFM13]. All together 16 use cases have been identified within a group of researchers and practitioners that have been also transferred to a DIN specification of the German Institute for Standardization (DIN) [NDMP14]. In the following, substantial results of the preliminary work are presented.

3 Technical Customer Service Processes

A service process involves activities undertaken to realize and deliver the service [BoYa05] accomplished by the service technicians in the field or back office (Figure 1).

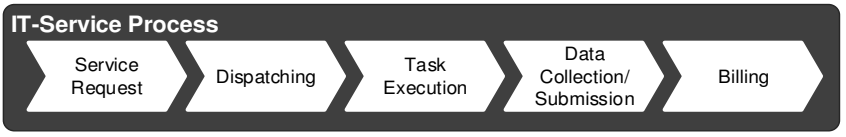


Figure 1: Technical Customer Service Process (referring to [BrTW08])

In a typical IT-service process the customer initiates the service order by contacting the TCS. The TCS is called either periodically (e.g. continous maintenance) or on-demand if an unexpected event occurs (e.g. machine breakdown). After the service request is received the service process passes the stage of dispatching where the service technicians are coordinated for the service delivery. Within the task execution the actual performance of the service, such as maintenance, repair or installation, takes place [BrTW08]. The data collection/submission comprises the acquisition of generated data and its feedback in a timely manner. The service process ends with the billing of the service order. The IT-service process can pass through different departments that operate

on the service process. Each department may require a slightly different view of the service processes where different requirements for the service management solution emerge. Stakeholders of those departments can be for example employees of the helpdesk, customer service employees, service field worker or accountants. Additionally, within the IT-service process a number of functions and sub-processes arise, e.g. installation administration or complaint management, which can vary depending on the service order [Vdma08]. Partially at the operator of the machine or at the machinery itself the service technicians dispose of current information about the state of the system, documentation and order data for their work [Walt09]. The generic TCS process can be seen as point of origin to identify determinants of mobile solutions that can be used to support TCS processes.

4 Requirements Engineering

4.1 Requirements Engineering Conception

RE is widely recognized as crucial part of Software Engineering (SE) and has established itself as a distinct research area [GrYu01]. Modelling of organisational goals constitutes a central activity of the RE process. Three main tasks are identified as integral part of requirements investigation: (1) elicitation, (2) specification and (3) validation. While requirements elicitation focusses on the comprehension of the organisational situation that the considered system aims to improve, requirements specification maps real-world needs onto a requirements model. Finally, the validation task intends to ensure that the derived specification corresponds to the original stakeholder needs [Kava02]. Considering RE in the field of PSS it is apparent that approaches for an integrated and cross-domain RE for PSS are almost completely missing but within the software and product development, RE is already prevalent.

4.2 Use Cases

Use case modelling is a common technique within SE describing functional requirements of systems [Cock99]. Therefor, the considered system is decomposed into different interaction scenarios that combined resemble a systems functionality [Lübk06]. Through use cases a complete course of interaction between an actor and the appropriate system can be constituted [CoLo01]. Depending on what special kind of request occurs and under which conditions this happens, different scenarios arise. For a structured description of use cases a template can be helpful [Cock99]. For the representation of the use cases the template by Cockburn is extended in a domain-specific way in order to capture the relevant aspects of TCS use cases (cf. Table 1). Besides the basic elements (number, name and description of the use case), especially the identification of actors and stakeholders is relevant. The primary actor is the actor, whose aims should be reached by the use case. The stakeholder has a personal interest in the use case, even if he never directly interacts with the system. Examples for a stakeholder can be companies, regulation authorities or the administrative board [Cock99].

Table 1: Use Case Template [ÖFMD13]

Title	Description
Number	Running number of the use case
Name	Title of the use case
Short Description	Brief explanation of the use case
Primary Actor	Name or description of the primary actor
Stakeholder	Name or description of the stakeholder
Precondition	Conditions that have to be satisfied in order to start the execution of the use case
Trigger	Reason to start the use case execution
Postcondition	Conditions that have to be satisfied after proper execution of the use case
Subfunction	Functions that are part of the use case
Input Data	Data relevant or needed for execution
Output Data	Data produced by the use case execution
Data Source	Origin of the data used by the use case

Further important information are the pre- and post-conditions as well as the respective triggers of the use case. Compared with the template suggested by Cockburn, this template is expanded by the following elements in order to meet the specific requirements for the illustration of IT-based processes in the TCS field: subfunctions describing functions that are part of the use case, input data which are necessary for the execution as well as the resulting output data. Moreover, for the illustration of the information flow the respective data source is described.

4.3 Specification of Use Cases for mobile Service Support Systems

In order to reduce the complexity of the system design task, 16 use cases have been specified (Figure 2).

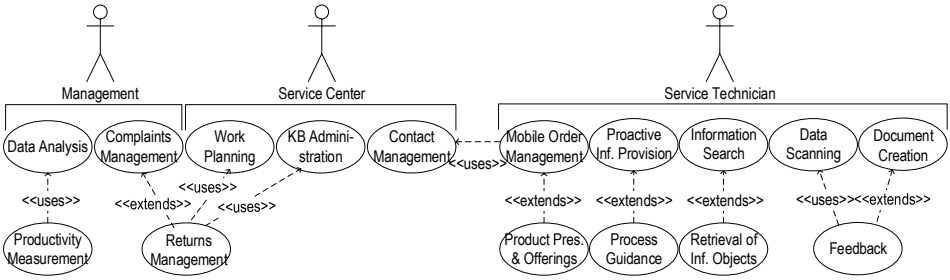


Figure 2: Use Cases [FÖMD13]

The use cases have been identified (a) by observing the daily work of service technicians and their usage of mobile TCS support systems during 77 complete service process executions and semi-structured interviews as well as (b) by examining the results of a structured literature review on requirements for and applications of mobile systems in the TCS domain. The use cases have been documented using the template provided by Cockburn (c.f. section 4.2). The system has to serve three distinct TCS stakeholders: the

management, service center and service technician. Some of the use cases make use of other use cases indicated by an <<uses>>-relation, whereas others are specialized variants indicated by <<extends>> [FÖMD13]. The management is the primary actor in the use cases data analysis and complaints management where the analysis of the operative data and the processing of warranties take place. Data and functionalities from these use cases are used further from the use cases productivity measurement (which provides analytical methods for the study of profitability and performance) and returns management (comprising the handling of complaints and returns processes). The service center represents the back office that undertakes the use cases work planning, knowledge database administration and contact management, which triggers the service order. Here again the returns management uses data resulting from activities within these use cases. The service technician at the customers' site operates through the use cases mobile order management, proactive information provision, information search, data scanning and document creation that are used or extended by product presentation & offerings, process guidance, retrieval of information objects and feedback [ÖFMD13]. These specified use cases describe IT-based TCS processes from a technical and operational term considering the user interaction. According to the template, the elementary content for each use case was derived which serves as nodal point for system specification.

5 Conclusion and Outlook

The TCS represents an important branch of the industry and is characterized by the need of product-related and customer information at the point of service. For ensuring the competitiveness and quality of TCS processes it is important to support the service technician by mobile assistance systems. With this contribution a starting point for the specification of mobile TCS support systems is provided by merging different research fields and their established methodologies. Drawing on RE concepts a consolidated list of requirements and the identification of information needs was worked out in former contributions (c.f. [MFÖK13][MDFÖ14]). Part of my future research is a prototypical implementation of the specified use cases into a mobile service support system and the empirical evaluation of the usability and productivity to prove the effectiveness of the approach. After all a holistic approach for the development of an IT-based mobile support system in the TCS domain is going to be employed beginning with requirements analysis and ending with the implementation and evaluation of a mobile service support system for the TCS.

References

[BeBK09] Becker, J.; Beverungen, D.F.; Knackstedt, R.: The challenge of conceptual modeling for product-service systems: status-quo and perspectives for reference models and modeling languages. In: Information Systems and e-Business Management 8(1), pp. 33–66.

[BoYa05] Boughnim, N.; Yannou, B.: Using Blueprinting Method for Developing Product-Service Systems. In: International Conference on Engineering Design, 2005.

[BrTW08] Brocke, J.; Thurnher, B.; Winkler, D.: Improving IT-Service Business Processes by Integrating Mobility: The Role of User Participation – Results from Five Case Studies. In: Proceedings of the International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE'08), 2008.

[Cock99] Cockburn, A.: Writing Effective Use Cases. Addison-Wesley, Boston 1999.

[CoLo01] Constantine, L.L.; Lockwood, L.A.D.: Structure and Style in Use Cases for User Interface Design. In (van Harmelan, M. Ed.): Object Modeling and User Interface Design. Addison-Wesley, Boston 2001, pp. 245–280.

[Dutt12] Dutta, S.: Field Service: Themes in Automation and Empowerment, April. 2012.

[FÖMD13] Fellmann, M.; Özcan, D.; Matijacic, M.; Däuble, G.; Schlicker, M.; Thomas, O.; Nüttgens, M.: Towards a Mobile Technical Customer Service Support Platform. In: MobiWIS, 2013, pp. 296–299.

[GrYu01] Gross, D.; Yu, E.: From Non-Functional Requirements to Design through Patterns. In: Requirements Engineering 6 (1), pp. 18–36.

[HLBT14] Hsu, M.W.; Lo, T.Y.; Bien, S.; Tsai, J.; Yang, W.H.: On the Product Development Model of Service Innovation in Manufacturing Industry. In: Anthropologist 17 (3), pp. 713–724.

[Kava02] Kavakli, E.: Goal-Oriented Requirements Engineering: A Unifying Framework. In: Requirements Engineering 6 (4), pp. 237–251.

[KuWD09] Kuo, Y.F.; Wu, C.M.; Deng, W.J: The relationship among service quality, perceived value, customer satisfaction, and post-purchase intention in mobile value-added services. In: Computers in Human Behaviour 25 (4), pp. 887–896.

[Lübk06] Lübke, D.: Transformation of Use Cases to EPC Models. In: 5. Workshop der Gesellschaft für Informatik e.V. (GI) und Treffen ihres Arbeitskreises "Geschäftsprozessmanagement mit Ereignisgesteuerten Prozessketten (WI-EPK)," 2006, pp. 137–156.

[MDFÖ14] Matijacic, M.; Däuble, G.; Fellmann, M.; Özcan, D.; Nüttgens, M.; Thomas, O.: Informationsbedarfe und -bereitstellung in technischen Serviceprozessen: Eine Bestandsaufnahme unterstützender IT-Systeme am Point of Service. In: Proceedings Multikonferenz der Wirtschaftsinformatik, 2014, pp. 2034–2046.

[MFÖK13] Matijacic, M.; Fellmann, M.; Özcan, D.; Kammler, F.; Nüttgens, M.; Thomas, O.: Elicitation and Consolidation of Requirements for Mobile Technical Customer Services Support Systems – A Multi-Method Approach. In: Proceedings of the International Conference on Information Systems ICIS 2013.

[NDMP14] Nüttgens, M.; Däuble, G.; Matijacic, M.; Peris, M.; Thomas, O.; Fellmann, M.; Kammler, F.; Özcan, D.; Balzert, S.; Koch, M.; Ahrens, D.; Friedrich, A.; Rosenkranz, N.; Böse, L.; Schlicker, M.: Use cases for mobile assistance systems in the field of technical customer service. In (DIN Ed.): DIN SPEC 91294. Beuth Verlag, Berlin 2014.

[ÖFMD13] Özcan, D.; Fellmann, M.; Matijacic, M.; Däuble, G.; Nüttgens, M.; Thomas, O.: Use Case based description of IT-supported customer service processes. In: Proceedings of XXII. International RESER Conference, 2013.

[ÖNFM13] Özcan, D.; Niemöller, C.; Fellmann, M.; Matijacic, M.; Däuble, G.; Schlicker, M.; Thomas, O.; Nüttgens, M.: A Use Case-driven Approach to the Design of Service Support Systems: Making Use of Semantic Technologies. In: Proceedings of the 5th International Symposium on Services Science (ISSS), 2013, pp. 105–116.

[Vdma08] VDMA: Entscheidungshilfe zur Einführung von Servicemanagementsystemen. VDMA Verlag, Frankfurt am Main 2008.

[Walt09] Walter, P.: Modellierung technischer Kundendienstprozesse des Maschinen- und Anlagenbaus als Bestandteil hybrider Produkte. In (Thomas, O.; Nüttgens, M. Eds.): Dienstleistungsmodellierung: Methoden, Werkzeuge und Branchenlösungen. Physica, Heidelberg 2009, pp. 129–145.

Berücksichtigung von Softwareperformanz im Entwicklungsprozess

David Georg Reichelt

Betriebliche Informationssysteme, Institut für Informatik
Universität Leipzig
Augustusplatz 10
04109 Leipzig
davidgeorg_reichelt@dagere.de

Abstract: Software-Performance-Engineering (SPE) bezeichnet verschiedene Ansätze zur Performanzverbesserung von Software. Diese wurden bisher nicht vergleichend untersucht. Der vorliegende Beitrag beschreibt ein Vorgehen, um zu überprüfen, inwiefern die SPE-Ansätze die Entstehung performanter Software sicherstellen.
Trotz des Vorhandenseins der SPE-Ansätze lassen sich bspw. beim Applikationsserver Apache Tomcat temporär im Entwicklungsversionen auftretende Performanzprobleme nachweisen. Dies deutet darauf hin, dass SPE-Techniken nicht bzw. nicht erfolgreich eingesetzt werden. Die semiautomatische Generierung von Performanztests könnte dieses Problem lösen, da mit dieser Performanzprobleme mit geringem Zusatzaufwand erkennbar wären. Dieser Beitrag schildert deshalb ein Vorgehen zur semiautomatischen Performanztestgenerierung.

1 Motivation

Softwareperformanz misst, wie effizient ein Softwaresystem in Bezug auf Zeit- und Ressourcenallokation ist [CMI11]. Indikatoren der Performanz sind u.a. Antwortzeit, Arbeitsspeicherverbrauch und CPU-Auslastung. Software Performance Engineering (SPE) bezeichnet Techniken, deren Ziel die Verbesserung der Performanz ist [WFP07].[1]

Wenngleich die Techniken des SPE in Praxisprojekten evaluiert wurden, erfolgte bisher keine Gesamt-Evaluierung an einer repräsentativen Menge von Projekten. Ein Vorgehen zur vergleichenden Evaluierung der SPE-Techniken wird hier vorgestellt. Eine erste Evaluierung anhand von Tomcat [RB14] zeigt, dass Performanzprobleme bei der Entwicklung von Tomcat 7 auftraten. Diese traten bei Aufruf von Seiten durch einen Nutzer auf. Da eine gute Einzelnutzer-Performanz Bedingung für eine gute Gesamt-Performanz ist [Pod08], soll der Einzelnutzer-Performanztest im Mittelpunkt stehen. Der Aufwand für derzeitig existierende SPE-Techniken ist sehr hoch [WFP07]; um diesen Aufwand zu reduzieren,

[1] Die Ursprungsdefinition von SPE umfasst lediglich das Refactoring der Architektur basierend auf Performanzvorhersagen, die auf Basis der Architektur abgeleitet wurden. Hier wird von Woodsides umfassenderen Definition ausgegangen.

sollen die Performanztests automatisch generiert werden. Daher soll hier ein Ansatz vorgestellt werden, um Einzelnutzer-Performanztests automatisiert zu generieren.

Dieser Beitrag gibt eingangs einen Überblick über SPE-Techniken. Anschließend werden Forschungsfragen erörtert. Darauf aufbauend wird auf das Forschungsdesign eingegangen.

2 Techniken des Software Performance Engineering

Nach [WFP07] unterteilen sich die Methoden des SPE grundlegend in zwei Kategorien: modellbasierte und messbasierte Ansätze. Bei modellbasierten Ansätzen werden Performanzinformationen an Architekturmodelle bzw. Ausführungsszenarien annotiert. Anschließend wird die Gesamtperformanz der Anwendung, bspw. durch Übertragung der Performanzinformationen in Warteschlangenmodelle und Analyse dieser, vorhergesagt. Sollte die dabei vorhergesagte Gesamtperformanz den Performanzanforderungen nicht genügen, werden Refactorings an der Architektur durchgeführt, bis eine angemessene Gesamtperformanz erreicht ist [WFP07]. Daneben existieren messbasierte Ansätze. Diese behandeln zwei Bereiche: Die eigentliche Messung und das Testen der Messwerte. Zum Messen gehört Monitoring, d.h. die Überwachung der Performanz im Einsatz befindlicher Softwaresysteme. Es ist möglich, aus den gemessenen Monitoringdaten die Architektur zu rekonstruieren und die Daten in Relation zu dieser bereitzustellen [vHWH12].

In der Praxis etabliert ist das Testen der Softwareperformanz durch Lasttests, das durch verschiedene Methoden und Werkzeuge unterstützt wird [Mol09]. Problematisch ist hier der große Aufwand für das Definieren, Ausführen und Auswerten der Tests. Aktuelle Forschung versucht den Lasttestprozess zu verbessern, bspw. indem aus Logs automatisiert potentielle Performanzprobleme erkannt werden [Jia10]. Weiterhin existieren verschiedene Arbeiten zur automatisierten Generierung von Lasttests (bspw. [SGH+14]).

In [Pod08] wird das kontinuierliche Testen der Performanz für einen Nutzer ohne zusätzliche Last gefordert. Grund hierfür ist, dass die Implementierungsqualität einzelner Abschnitte die Gesamtperformanz negativ beeinflussen kann. Ein Teil der Performanzprobleme können also durch Einzelnutzer-Performanztests erkannt werden. Probleme, die nur durch Mehrbenutzerbetrieb entstehen, können damit hingegen nicht gefunden werden. Forschung zur automatisierten Einzelnutzer-Performanztest-Generierung existiert bislang nicht.

3 Forschungsfragen

Das vorliegende Promotionsvorhaben wird empirisch die Behandlung von Softwareperformanz in Entwicklungsprojekten betrachten. Darauf aufbauend wird eine Möglichkeit untersucht, die Einhaltung von Performanzgrenzwerten besser sicherzustellen. Dieses Kapitel stellt Forschungsfragen zu diesen beiden Forschungszielen vor.

Eine zentrale Fragestellung ist es, inwiefern die beschriebenen Performanzverbesserungstechniken Performanzprobleme verhindern können. Dabei ist nach Zeitpunkt des Auf-

tretens der Performanzprobleme zu differenzieren: Beispielsweise werden Implementierungsfehler kurz vor einem Release durch anfängliches Architekturrefactoring nicht verhindert werden können. Untersucht werden muss also, zu welchem Zeitpunkt in Relation zum Release welche Arten von Performanzproblemen auftreten. Diese Erkenntnis liefert den zentralen Zugang zur Fragestellung, welche Performanzprobleme durch welche Techniken erkannt bzw. gelöst werden können. Es ergibt sich damit folgende Forschungsfrage:

1. Welche Performanzverbesserungstechniken können das Auftreten welcher Performanzprobleme an welcher Stelle im Softwareentwicklungsprozess verhindern?

Als Vorarbeit wurde gezeigt, dass sich auch bei großen Projekten Softwareperformanz ggf. stark verschlechtert [RB14]. Basierend auf der Beantwortung der ersten Forschungsfrage soll deshalb eine Methode entwickelt werden, die die Überprüfung der Erfüllung von Performanzanforderungen während des gesamten Softwareentwicklungsprozesses ermöglicht. Auch einzelne Programmeinheiten sollen analog zum Vorgehen bei funktionaler Korrektheit regelmäßig überprüft werden. Die dabei entstehenden Performanztests sind Performanz-Unit-Tests. Aufgrund der bereits in Kapitel 2 genannten Argumente für Einzelnutzer-Performanztests soll der Fokus auf diesen liegen.

Da die Implementierung umfassender Einzelnutzer-Performanz-Unit-Tests einen hohen Zusatzaufwand darstellen, werden Sie in der Praxis selten umgesetzt werden. Deshalb wird sich dieses Promotionsvorhaben mit der semiautomatischen Generierung der Tests beschäftigen. Zur Umsetzung der Testgenerierung existieren zwei Möglichkeiten: Die Generierung von Tests aus Anforderungen und die Generierung von Tests aus Monitoringdaten. Es ergibt sich somit folgende Forschungsfrage:

2. Welche Methoden eignen sich, um (semi-)automatisch Performanztests zu generieren?

4 Vorgehen

Anknüpfend an die beiden Forschungsfragen wird in diesem Kapitel auf die empirische Untersuchung des Softwareentwicklungsprozesses und anschließend auf die Möglichkeiten zur Performanztestgenerierung eingegangen.

4.1 Empirischen Untersuchung der Performanz im Softwareentwicklungsprozess

Um zu überprüfen, wann Performanzprobleme im Softwareentwicklungsprozess auftreten, wird die Performanzentwicklung großer Softwareentwicklungsprojekte untersucht werden. Hierbei können frei verfügbare Projekte als Grundlage genutzt werden. Zur Auswahl der Projekte sind Kriterien für die Mindestgröße und die Mindestanzahl an Revisionen zu definieren. Die empirische Untersuchung gliedert sich in drei Schritte: Ermittlung der Performanzverläufe, Ermittlung von Performanzproblemen und Untersuchung des Auftretens

der Performanzprobleme in Relation zum Einsatz der SPE Techniken.

Im ersten Schritt wird eine Anwendung implementiert, die aus vorhandenen Unit-Tests KoPeMe-Performanztests[2] generiert. Diese wird auf jede Revision eines Projektes angewandt, um für diese passende Performanztests zu generieren. Die Tests werden anschließend in der jeweiligen Revision ausgeführt. Im zweiten Schritt werden die sich daraus ergebenden Performanztestverläufe ausgewertet. Hierfür wird ein Auswertungswerkzeug implementiert werden, das signifikante Performanzveränderungen erkennt. Kriterien für die Signifikanz, wie bspw. ein betragsmäßig besonders großer Anstieg eines Performanzmesswertes über die Revisionen, sind noch zu definieren. Aufbauend auf den Messungen sollen Performanzprobleme erkannt werden. Hierzu sind Kriterien notwendig, um Veränderungen von Problemen zu unterscheiden. Dies ist notwendig, da bspw. neue Funktionalitäten starke Performanzveränderungen auslösen können, die keine Performanzprobleme sind. Im letzten Schritt wird durch die Auswertung der erkannten Performanzprobleme ermittelt, welche SPE-Technik zu welchem Zeitpunkt die Entstehung von Performanzproblemen unterbinden kann. Hierfür ist zu prüfen, welche SPE-Techniken angewandt wurden. Denkbar ist es, Teile der Überprüfung zu automatisieren, bspw. indem nach JMeter-Lastskripten[3] oder Palladio-Komponentenmodellen[4] im Software-Repositorium des Projekts gesucht wird. Indem geprüft wird, welche zeitliche Verteilung von Performanzproblemen mit welchem Einsatz von SPE-Techniken korreliert, kann die Wirkung der jeweiligen Technik beurteilt werden.

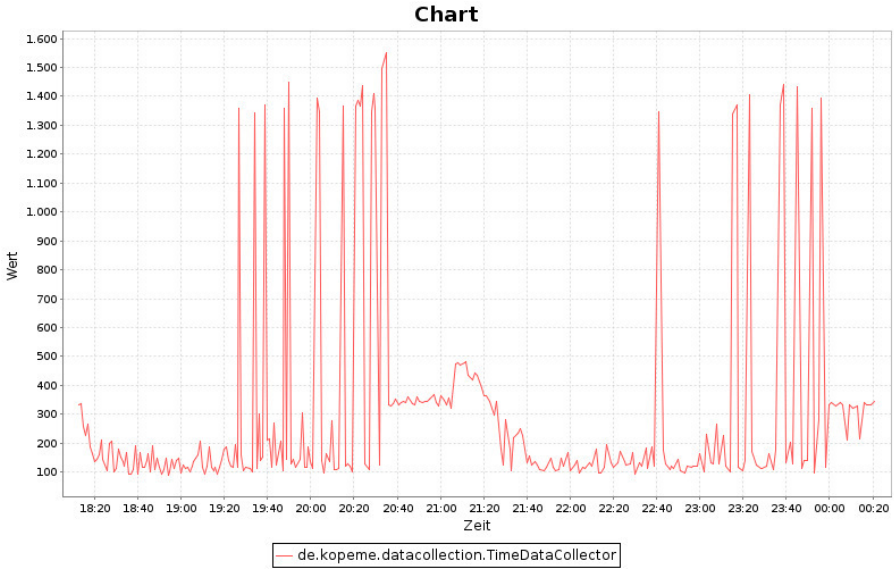


Abbildung 1: Downloadzeit einer Servletseite in verschiedenen Tomcat-Revisionen, aus: [RB14]

[2]Hierbei handelt es sich um Unit-Tests, die Performanzwerte ermitteln [RB14].
[3]JMeter ist ein Werkzeug für Lasttests, Offizielle Webseite: https://jmeter.apache.org/
[4]Palladio ist ein Werkzeug für modellbasiertes SPE, Offizielle Webseite: http://www.palladio-simulator.com

Eine Untersuchung der Performanz der Revisionen von Tomcat 6 hat im Rahmen von [RB14] stattgefunden. Abbildung 1 zeigt, wie sich die Performanz des Anwendungsfalls, eine Servlet-Seite von einem lokalen Server zu laden, verhält. Die Abszisse zeigt die Zeit der Ausführung, die jeweils einzelnen Revisionen zugeordnet werden kann, und die Ordinate zeigt die Downloadzeit der Servlet-Seite vom lokalen Server, der aus der Revision kompiliert wurde. Das starke Ansteigen der Antwortzeit legt nahe, dass Performanzprobleme mit aktuellen Techniken nicht effizient genug bearbeitet werden. Bei Tomcat wird lediglich ein Profiler zu Finden von Performanzproblemen eingesetzt.

4.2 Unterstützung des Entwicklungsprozesses

Es gibt zwei Varianten, Performanztests (semi-)automatisiert zu generieren: Unter Nutzung der Anforderungen oder der Monitoringdaten. Da Performanztests die Sicherstellung der Einhaltung von Performanzanforderungen ermöglichen sollen, können sie direkt aus den Anforderungen abgeleitet werden. Hierbei können analog zum Software Performance Engineering Performanzannotationen an UML-Modelle als Grundlage dienen. Um die dort angegebenen Grenzwerte zu überprüfen, ist es notwendig, Initialisierungsaufrufe und Parameter der zu prüfenden Anforderung manuell hinzuzufügen.

Es ist davon auszugehen, dass bei den Ausführungen von Tests die gemessenen Performanzwerte annähernd den Anforderungen entsprechend. Es ist deshalb möglich, diese als Test-Grenzwerte zu verwenden. Gegebenenfalls sind die gemessenen Performanzwerte manuell zu verändern, damit diese den Performanzzielen entsprechen.

Wird der Quelltext vollständig instrumentiert, wie es beispielsweise das Framework Kieker [vHWH12] ermöglicht, ist es denkbar, den Entwickler den Methodenaufruf aus einem Aufrufbaum auswählen zu lassen, für die ein Test erstellt werden soll. In diesem Fall sind die vorher notwendigen Initialisierungsaufrufe und die Parameter bekannt. Es ist für den Entwickler lediglich notwendig, ggf. Grenzwerte zu ändern. Eine Kombination mit der Grenzwertextraktion aus Performanzmodellen könnte es auch erlauben, Tests bis auf die Auswahl des zu testenden Aufrufs vollautomatisch durchzuführen. Ein vollautomatisches Testen aller Methoden ist nicht sinnvoll, da Performanztests sehr zeitintensiv sind.

Anschließend ist die Praxistauglichkeit des Ansatzes zu untersuchen. Dies ist möglich, indem äquivalent zum in Kapitel 4.1 dargestellten Vorgehen geprüft wird, ob bei semiautomatischer Performanztestgenerierung bei vorhandenen Projekten Performanzprobleme gefunden hätten.

Zusätztlich wird eine Evaluation an in Entwicklung befindlichen Projekten angestrebt. Hierbei wird ermittelt, wie viele Performanzprobleme diese hatten. Es wird angenommen, dass so viele Projekte analysiert werden, das Messwerte nicht durch Ausreißer verzerrt werden. Unter dieser Annahme ist der Quotient Performanzprobleme/Quelltextzeilen ein Indikator für den Einsatz guter Performanzverbesserungstechniken. Durch den Vergleich des Quotienten Performanzprobleme/Quelltextzeilen mit anderen Projekten unter Berücksichtigung der dort verwendeten Performanzverbesserungstechniken ist es möglich, die Wirkung der Performanztestgenerierung auf die Softwareperformanz zu beurteilen.

5 Zusammenfassung

Es wurde ein Promotionsvorhaben geschildert, dessen Ziel die Untersuchung der Wirkung aktueller SPE-Techniken und eines neuen Verfahrens zur semiautomatischen Performanztestgenerierung auf die Performanz der entstehenden Softwareprodukte ist.

Die Untersuchung der Wirkung aktueller Performanzverbesserungsverfahren soll durch eine Überprüfung der Performanz von frei verfügbaren Projekten erfolgen. Dabei sollen Performanzwerte einzelner Unit-Tests in Relation zum Einsatz von Performanzverbesserungsverfahren gesetzt werde. Als neues Performanzverbesserungsverfahren soll das semiautomatische Generieren von Einzelnutzer-Performanz-Unit-Tests untersucht werden. Die Generierung soll auf Basis von spezifizierten Anforderungen in UML-Diagrammen und auf Basis von Monitoringdaten eines Anwendungslaufs, bei dem Performanzanforderungen erfüllt wurden, erfolgen. Anschließend soll an frei verfügbaren Projekten geprüft werden, ob dieses Vorgehen Tests generieren kann, die Performanzprobleme erkannt hätten.

Literatur

[CMI11] Vittorio Cortellessa, Antinisca Di Marco und Paola Inverardi. *Model-Based Software Performance Analysis*. Springer, 2011.

[Jia10] Zhen Ming Jiang. Automated analysis of load testing results. In *ISSTA*, Seiten 143–146, 2010.

[Mol09] Ian Molyneaux. *The Art of Application Performance Testing: Help for Programmers and Quality Assurance*. O'Reilly Media, Inc., 1st. Auflage, 2009.

[Pod08] A. Podelko. Agile Performance Testing. In *Int. CMG Conference*, Seiten 267–278, 2008.

[RB14] David Georg Reichelt und Lars Braubach. Sicherstellung von Performanzeigenschaften durch kontinuierliche Performanztests mit dem KoPeMe Framework. In Wilhelm Hasselbring und Nils Christian Ehmke, Hrsg., *Software Engineering 2014: Fachtagung des GI-Fachbereichs Softwaretechnik*, Lecture Notes in Informatics, Seiten 119–124. Köllen Druck+Verlag GmbH, 2014.

[SGH+14] Eike Schulz, Wolfgang Goerigk, Wilhelm Hasselbring, Andre van Hoorn und Holger Knoche. Model-Driven Load and Performance Test Engineering in DynaMod. In *Proceedings of the Workshop on Model-based and Model-driven Software Modernization (MMSM '14)*, Seiten 10–11, März 2014.

[vHWH12] André van Hoorn, Jan Waller und Wilhelm Hasselbring. Kieker: A Framework for Application Performance Monitoring and Dynamic Software Analysis. In *Proceedings of the 3rd joint ACM/SPEC International Conference on Performance Engineering (ICPE 2012)*, Seiten 247–248. ACM, April 2012.

[WFP07] Murray Woodside, Greg Franks und Dorina C. Petriu. The Future of Software Performance Engineering. In *2007 Future of Software Engineering*, FOSE '07, Seiten 171–187, Washington, DC, USA, 2007. IEEE Computer Society.

Studierenden Konferenz Informatik 2014

(SKILL 2014)

Studierendenkonferenz Informatik 2014 (SKILL 2014)

Johannes Schmidt[1], Michael Becker[1], Thomas Riechert[2],
Dirk Pflüger[3], Ludger Porada[4], Sarah Oppold[5]

[1]Universität Leipzig, Abt. Betriebliche Informationssysteme, 04109 Leipzig
{jschmidt | mbecker}@informatik.uni-leipzig.de

[2]HTWK Leipzig, Fakultät Informatik, Mathematik und Naturwissenschaften
04277 Leipzig
thomas.riechert@htwk-leipzig.de

[3]Universität Stuttgart, Institut für Parallele und Verteilte Systeme
70569 Stuttgart
Dirk.Pflueger@ipvs.uni-stuttgart.de

[4]Gesellschaft für Informatik e.V. (GI)
53175 Bonn
ludger.porada@gi.de

[5]Universität Stuttgart, Fachschaft INF & SWT 70174 Stuttgart
oppoldsa@fachschaft.informatik.uni-stuttgart.de

Studierende der Informatik leisten im Rahmen von Seminar-, Bachelor- oder Masterarbeiten z.T. herausragende Beiträge, die jedoch häufig nicht sichtbar sind. Aus diesem Grund wurde im Jahr 2010 die erste *Studentenkonferenz Informatik Leipzig (SKIL)* ins Leben gerufen, die im Rahmen der GI Jahrestagung INFORMATIK 2010 stattfand [FF10, S. 1149ff]. Die Studierenden sollten auf einer Konferenz eigene Beiträge vorstellen und diskutieren. In den folgenden Jahren wurde das Konzept erweitert und als selbstständige Veranstaltung in Leipzig etabliert [SRA12, ARS11].

Mit dem Ziel, eine breitere Öffentlichkeit anzusprechen, haben wir uns entschieden, die SKIL als *Studierendenkonferenz Informatik 2014 (SKILL)* im Rahmen der INFORMATIK 2014 zu integrieren. Somit erhalten die angenommenen Beiträge die Sichtbarkeit, die sie verdienen. Das Grundkonzept hat sich jedoch nicht verändert: Ausschließlich Studierende sind aufgerufen, einen Beitrag einzureichen und auf der Konferenz vorzustellen. Der Beitragsaufruf grenzt kein Gebiet der Informatik aus, um ein möglichst breites Spektrum an Themen präsentieren zu können. Das Programmkomitee sichtet jeden Beitrag, wobei sich die Reviews dadurch auszeichnen, dass sie sehr detailliert auf Verbesserungspotentiale eingehen. Auf der Konferenz bietet sich allen Beteiligten die Möglichkeit, sich fachlich auszutauschen und näher kennenzulernen.

Aus 31 Einreichungen wurden 12 Beiträge sowie acht Poster für die Konferenz ausgewählt. Die hier veröffentlichten Beiträge zeigen einen Querschnitt durch das breite und interdisziplinäre Feld der Informatik. Sie behandeln Themen aus der Linguistik, der theoretischen und technischen Informatik bis hin zu Themen der angewandten Informatik, wie bspw. die Konstruktion von Informationssystemen oder Big Data Anwendungen.

Die breite Resonanz zur diesjährigen Studierendenkonferenz Informatik 2014 lässt uns guten Mutes in die Zukunft schauen. Wir hoffen, dass sich auch in den nächsten Jahren viele Konferenzen studentischer Arbeiten annehmen, sei es im Rahmen eines Studierendentracks oder als dedizierte Konferenz wie die Studierendenkonferenz Informatik 2014.

Wir möchten uns bei allen Unterstützern und Sponsoren bedanken, durch deren Engagement wir sowohl Preise für das Best Paper und das Best Poster ausloben als auch allen Autorinnen und Autoren eine finanzielle Unterstützung für ihre Konferenzteilnahme anbieten konnten.

Stuttgart, 25. September 2014 Johannes Schmidt, Michael Becker, Thomas Riechert, Dirk Pflüger, Ludger Porada und Sarah Oppold

Organisatoren der SKILL 2014

Programmkomitee

Besonderer Dank gilt allen Mitgliedern des Programmkomitees der Studierendenkonferenz Informatik 2014:

- Bernhard Bauer (Universität Augsburg)
- Marco Block-Berlitz (Hochschule für Technik und Wirtschaft Dresden)
- Henry Borasch (Student HTWK Leipzig)
- Andres Bruhn (Universität Stuttgart)
- Patrick Diehl (Universität Bonn)
- Steffen Dienst (Universität Leipzig)
- Thomas Efer (Universität Leipzig)
- Holger Fröning (Universität Heidelberg)
- Martin Fuchs (Universität Stuttgart)
- Stefan Funke (Universität Stuttgart)
- Olaf Herden (DHBW Stuttgart Campus Horb)
- Hans-Joachim Hof (Hochschule München)
- Walter Hower (Hochschule Albstadt-Sigmaringen)
- Axel Hummel (Universität Leipzig)
- Martin Junghanns (Universität Leipzig)
- Heiko Kern (Universität Leipzig)
- Stephan Klingner (Universität Leipzig)
- Stefan Kühne (Universität Leipzig)
- Ralf Laue (Westsächsische Hochschule Zwickau)
- Lars-Peter Mayer (Universität Leipzig)

- Kyrill Meyer (Universität Leipzig)
- Dagmar Monett Díaz (Berlin School of Economics and Law)
- Martin Pero (Unister GmbH)
- Thomas Riechert (Universität Leipzig)
- Michael Schmeißer
- Thomas Schmid (Universität Leipzig)
- Ricardo Usbeck (Universität Leipzig)
- Antonius van Hoof (DHBW Stuttgart Campus Horb)
- Karsten Weicker (HTWK Leipzig)
- Andrej Werner (Universität Leipzig)

Literatur

[ARS11] Sören Auer, Thomas Riechert und Johannes Schmidt, Hrsg. *SKIL 2011 - Studentenkonferenz Informatik Leipzig 2011: Leipzig, Deutschland, 2. Dezember 2011 ; Tagungsband*. Nummer 27 in Leipziger Beiträge zur Informatik. LIV, Leipzig, 2011.

[FF10] Klaus-Peter Fähnrich und Bogdan Franczyk, Hrsg. *INFORMATIK 2010: Service Science – Neue Perspektiven für die Informatik: 27.09. - 01.10.2010 Leipzig*. Nummer P-176 in Lecture Notes in Informatics (LNI) - Proceedings. Gesellschaft für Informatik e.V. (GI), Bonn, 2010.

[SRA12] Johannes Schmidt, Thomas Riechert und Sören Auer, Hrsg. *SKIL 2012 - Dritte Studentenkonferenz Informatik Leipzig: Leipzig, Deutschland, 25. September 2012 ; Tagungsband*. Nummer 34 in Leipziger Beiträge zur Informatik. LIV, Leipzig, 2012.

Automatische Extraktion von Fachterminologie aus kunsthistorischen Volltexten

Juliane Bredack
julianebredack@web.de
Fachhochschule Potsdam, Master Studiengang Informationswissenschaften

Abstract: Mit Hilfe eines algorithmisch arbeitenden Verfahrens können fachterminologische Mehrwortgruppen aus elektronisch vorliegenden Texten identifiziert und extrahiert werden. Inhaltlicher Schwerpunkt stellt die Einbindung von Funktionswörtern des deutschen Sprachgebrauchs in den Extraktionsalgorithmus dar. Als Datengrundlage dieser Arbeit dienten kunsthistorische Lexikonartikel des Reallexikons zur Deutschen Kunstgeschichte. Das automatische Indexierungssystem Lingo wurde in dieser Studie genutzt. Anhand selbst gebildeter Kriterien, wurden die extrahierten Mehrwortgruppen qualitativ analysiert. Es konnte festgestellt werden, dass die Verwendung von Funktionswörtern fachterminologische Mehrwortgruppen erzeugt, die als potentielle Indexterme weitere Verwendung im Information Retrieval finden können.

1 Einleitung

Inhaltsrelevante Informationen eines Fachtextes werden durch Fachwörter repräsentiert. Die entsprechende Fachterminologie, wie sie beispielsweise durch Sachbegriffe oder Eigennamen gekennzeichnet ist, liegt nicht nur in Form von Einzelwörtern vor. Konzepte bzw. Inhalte finden sich oft in Mehrwortbegriffen bzw. Fachphrasen wieder. Diese Mehrwortgruppen sind als zusammenhängende lexikalische Einheiten zu sehen, welche als Ganzes eine Bedeutung aufweisen [Le13]. Die miteinander stehenden Begriffe, wie sie z.B. in Adjektiv-Substantiv Verbindungen auftreten, erzeugen eine hohe Aussagekraft und Deutlichkeit, die ein einzelner Begriff oft nicht erreicht. Ein separat betrachteter Term kann verschiedene Inhalte ausdrücken, wohingegen der inhaltsrelevante Sinn oder die eindeutige Zuordnung zu einem Thema innerhalb einer Mehrwortgruppe sichtbar wird. Somit lassen Mehrwortgruppen eine Bedeutungsdifferenzierung auf der Ebene eines einzelnen Wortes zu [GLN12].

In Bezug auf elektronische Dokumente bzw. Dokumentkollektionen und deren inhaltlicher Erschließung, leisten automatische Indexierungsverfahren eine zuverlässige Identifizierung einzelner, inhaltsrelevanter Terme um diese in einem Index aufzunehmen. Ein zuverlässiges Suchen und Finden im Rahmen des Information Retrieval wird gewährleistet [Le13]. Neben Einzeltermen sollten jedoch auch Mehrwortgruppen im Index hinterlegt sein, um die Suchmöglichkeiten auszuweiten. Ein Dokument kann durch den Einsatz einer Mehrwortgruppe gezielt mit dieser gefunden werden, anstelle von mehreren Dokumenten, die die einzelnen Begriffe beinhalten und so unnötig viele Informationen

bereitstellen, die nicht im Sinne eines Suchenden sind. Im Folgenden wird deshalb ein Verfahren beschrieben, das eine Identifizierung von Mehrwortgruppen zulässt, um diese für ein späteres Retrieval zu nutzen.

Das für die Untersuchung eingesetzte Indexierungssystem Lingo basiert auf linguistischen Methoden und leistet eine Identifizierung von Mehrwortgruppen, wenn diese in einem Wörterbuch lexikalisiert, also vorab bekannt sind (z.B. Nutzung von kontrolliertem Fachvokabular beim Wörterbuchaufbau) [GLN12]. Die lexikalische Funktionalität wurde bereits im Digitalisierungs-, Erschließungs- und Indexierungsprojekt RDK-Web zur Erstellung einer Web-Retrievalumgebung des Reallexikon zur Deutschen Kunstgeschichte (RDK) getestet [Le06], welche sich als zweckmäßig erwiesen hat [Rea14]. Die inhaltliche Komplexität und Vielzahl der RDK Lexikonartikel lässt allerdings vermuten, dass die Anzahl unbekannter Mehrwortgruppen als weitaus höher einzuschätzen ist. Neben einer Indexierung mit bereits bekanntem Vokabular (lexikalische Funktion) ist es mit Lingo und dessen Programmmodul sequencer möglich noch unbekannte und potentiell aussagekräftige Mehrwortkonstruktionen auf algorithmischem Weg zu identifizieren.

Die automatische Identifizierung und Extraktion fachterminologischer Mehrwortgruppen erfolgte am Beispiel kunsthistorischer Fachtexte. Als Datenquelle dienten die deutschsprachigen Lexikonartikel des Reallexikon zur Deutschen Kunstgeschichte.

2 Besonderheiten kunsthistorischer Mehrwortbegriffe im RDK

Das RDK ist ein Nachschlagewerk zur Realienkunde der Kunstgeschichte, welches seit 1937 erscheint. Inhaltliche Schwerpunkte der Lexikonartikel liegen auf Architektur, Bildenden Künsten, Kunsthandwerk, Materialien, Technik und Ikonographie des deutschen Sprachgebiets. Herausgeber ist das Zentralinstitut für Kunstgeschichte in München. Die Bände des RDK erscheinen als fortsetzendes Lieferwerk [Zen14].

Im folgenden Ausschnitt des Lexikonartikels „Christus“ (Band 3, Spalte 611, RDK) sind Mehrwortbegriffe, die kunsthistorische Inhalte wiedergeben, kursiv hervorgehoben.

> Obwohl, abgesehen von der Sarkophagplastik, der erhaltene Bestand altchristlicher Monumente mit figuralem Schmuck verhältnismäßig gering ist, so ergibt sich doch bei genauer Durchsicht, daß fast alle in der Kunst des frühen MA dargestellten Szenen aus dem Leben Christi schon irgendwie in der altchristlichen Kunst vorkommen. Diese hat - neben der Verkündigung an Maria (mit apokryphen Zutaten) und der Heimsuchung Elisabeths - dargestellt: die Geburt Christi, die Anbetung der Weisen, die Darbringung im Tempel, die Flucht nach Ägypten (mit apokryphen Zutaten), den Kindermord, den zwölfjährigen Jesus im Tempel, die Taufe Jesu im Jordan, das erste Wunder Christi bei der Hochzeit zu Kana, die Brotvermehrung, die Heilung des Blinden, […]

Es wird deutlich, dass sich kunsthistorische Begriffe in Mehrwortgruppen aus einer Kombination von Sachbegriffen (Darbringung im Tempel), Sachbegriffen und Personennamen (Verkündung an Maria) oder Geografika (Flucht nach Ägypten) zusammen-

setzen. Weiterhin fällt auf, dass sich Mehrwortgruppen nicht ausschließlich aus feststehenden Kombinationen von Adjektiven und Substantiven bilden, sondern auch Funktionswörter enthalten. Unter Funktionswörtern sind in diesem Fall Artikel, Konjunktionen und Präpositionen zu verstehen.

Funktionswörter stellen eine vernachlässigte Wortart im Retrieval dar, da sie selbst keine inhaltlich relevanten Informationen vermitteln und so als Stoppwörter unbeachtet bleiben. Innerhalb einer Mehrwortgruppe erweitern sie allerdings deren syntaktische Struktur und stellen die nötige Verbindung zwischen inhaltsrelevanten Begriffen her. So verbinden Funktionswörter fachterminologische Begriffe zu einer Einheit, sodass komplexere Inhalte vermittelt werden.

Zusammenfassend lassen sich die Besonderheiten von Mehrwortgruppen und Funktionswörtern mit Bezug zur Kunstgeschichte wie folgt festhalten:

- Verknüpfung von Fachtermen durch Funktionswörter zu syntaktisch abgeschlossenen Mehrwortgruppen, z.B. „Szene aus dem Leben Jesu“. Als Verbindung zwischen inhaltsrelevanten Termen kommt neben dem Artikel dem in diesem Beispiel eine Präposition zum Einsatz. Zu beachten ist, dass ein und dieselbe Präposition unterschiedliche Beziehungen ausdrückt. So kann aus auch ein örtliches Verhältnis beschreiben, wie „Altar aus Kloster Buxheim“. Ein bestimmter Sachverhalt lässt sich allerdings auch durch verschiedene Präpositionen ausdrücken, „Darstellungen zum Leben Jesu“.

- Der Artikelausschnitt „Christus“ enthält als Indexterme in Frage kommende Begriffe wie „Leben“, „Szene“, „Tempel“ oder „Kunst“. Durch den Einsatz von Mehrwortgruppen lassen sich eindeutigere Aussagen zum Artikelinhalt treffen, z.B. „dargestellte Szene aus dem Leben Christi“, „zwölfjähriger Jesus im Tempel“, „monumentale Kunst der altchristlichen Zeit“ oder „Szene aus dem Leben Jesu“. Eine inhaltliche Spezifizierung wird deutlich.

- Bedeutungsdifferenzierung von Begriffen, z.B. „Zeit“, welcher als Indexterm zu allgemein ist, jedoch innerhalb einer Mehrwortgruppe „Kunst der altchristlichen Zeit“ seinen Zweck erfüllt.

Als vorrangig wird in dieser Studie die Integration von Funktionswörtern (Artikel, Konjunktionen, Präpositionen) in den Extraktionsalgorithmus angesehen, um unbekannte und komplex strukturiertere Mehrwortbegriffe mit kunsthistorischem Aspekt zu extrahieren. So wird im Folgenden eine Gruppe, bestehend aus mindestens drei Wörtern, als Mehrwortgruppe angesehen.[1]

[1] Potentielle Mehrwortgruppen, die sich auch aus zwei Bestandteilen zusammensetzen könne (z.B. Adjektiv-Substantiv Verbindungen) blieben in dieser Studie unberücksichtigt.

3 Automatische Identifizierung und Extraktion von Mehrwortgruppen

3.1 Das Indexierungssystem Lingo

Das Indexierungssystem Lingo[2] ist wörterbuchgestützt. Die durch eine Indexierung gewünschten Ergebnisse werden flexibel mittels der dafür vorgesehenen Konfigurationsdateien gesteuert. Einzelne Programmmodule realisieren eine Indexierung, die in ihrer Funktionsweise und Ergebnissen aufeinander aufbauen. So stützt sich das Programmmodul sequencer auf die vorab im Indexierungsprozess erfolgte grammatikalische Normierung des Textes. Die korrekte Identifizierung aller Zeichenketten setzt deren Lexikalisierung in einem Grundformwörterbuch voraus [LV06]. Lingo Wörterbücher für eine Grundformerkennung sind folgendermaßen aufgebaut:

```
automatische=automatisch #a
extraktion=extraktion #s
fachterminologischer=fachterminologisch #a
mehrwortgruppe=mehrwortgruppe #s
mit=mit #c
lingo=lingo #s
```

Links vom Gleichheitszeichen befindet sich die Wortform und rechts davon die zugeteilte Grundform. Jeder Eintrag ist zudem mit einer Wortklasse, die Angaben zur Wortart (z.B. Adjektiv #a oder Substantiv #s) enthält, gekennzeichnet. Durch die im Wörterbuch eingetragenen Terme werden die Bestandteile des folgenden Satzes nach dem Indexierungsprozess zuverlässig identifiziert.

„Automatische Extraktion fachterminologischer Mehrwortgruppen mit Lingo."

Eine Indexierung führt zu diesem Ergebnis:

```
<Automatische=[(automatisch/a)]>
<Extraktion=[(extraktion/s)]>
<fachterminologischer=[(fachterminologisch)/a]>
<Mehrwortgruppen=[(mehrwortgruppe/s)]>
<mit=[(mit/c)]>
<Lingo=[(lingo/s)]>
:./PUNC
```

Nach einer Indexierung sind die Wortklassen der einzelnen Satzbestandteile bekannt, was sich für eine algorithmische Identifizierung von Mehrwortgruppen mit dem sequencer nutzen lässt. Aus dem Wissen heraus, das jeder Term in seiner Bedeutung gekennzeichnet ist, kann durch das Bilden von Wortmustern, basierend auf den Wortklassen eines Grundformwörterbuchs, eine Identifizierung von Wortgruppen erfolgen.

[2] http://lex-lingo.blogspot.de/

Beispielsweise lässt das Muster AS eine Identifizierung von Adjektiv-Substantiv Wortfolgen zu, wie im Folgenden zu sehen ist.

Automatische [a] Extraktion [s] fachterminologischer [a] Mehrwortgruppen [s]

AS automatisch extraktion
AS fachterminologisch mehrwortgruppe

Voraussetzungen für eine zuverlässige Identifizierung von Mehrwortgruppen bilden zum einen Grundformwörterbücher mit den darin enthaltenen Termen, getaggt durch eine Wortklasse. Zum anderen müssen die auf den Wortklassen basierenden Identifizierungsmuster in der dafür vorgesehenen Lingo-Konfiguration eingebunden sein. Als Ergebnis einer erfolgten Identifikation und Extraktion, werden die Mehrwortgruppen als Listen in einer Ergebnisdatei ausgegeben. Deren Größe richtet sich nach der zu indexierenden Datei bzw. Dokumentkollektion. Die Mehrwortbegriffe werden, wie am oben gezeigten Beispiel zu sehen, in der Grundform ausgegeben [GLN12].

3.2 Wörterbuchaufbau und Wortklassen der RDK-Indexierung

Die algorithmische Identifizierung neuer, kunsthistorischer Mehrwortkonstruktionen mittels sequencer setzt voraus, dass Wörterbücher mit geeignetem Vokabular existieren. Für eine Indexierung mit Lingo können die Wörterbücher als einfache Textdateien abgespeichert und mit einem Texteditor bearbeitet werden. Für die Identifizierung fachspezifischer Inhalte wurde auf die, aus der RDK-Web Indexierung hervorgegangenen Wörterbücher und deren fachterminologischen Einträge zurückgegriffen [Le06]. Die RDK-Web Wörterbücher umfassten die für die Kunstgeschichte typischen Terme von Sachbegriffen, Personennamen und Geografika. Für die erneute Indexierung des RDK wurden alle Fachterme in einem einzigen Wörterbuch hinterlegt. In einem extra geschaffenen Wörterbuch wurden Artikel, Konjunktionen und Präpositionen lexikalisiert, um syntaktisch komplexere Mehrwortbegriffe zu erzeugen. Zusätzlich kam ein drittes Wörterbuch zum Einsatz, in dem Begriffe, die der Alltagssprache angehören, identifiziert werden.

Aufbauend auf den neu erstellten Wörterbüchern, wurden deren Einträge um neue Wortklassen erweitert. So wurden kunsthistorische Sachbegriffe, Geografika und Personenamen mit der Wortklasse E getaggt. Es erfolgte eine separate Kennzeichnung aller Funktionswörter, damit diese gezielt in den Extraktionsalgorithmus eingebaut werden konnten. Die differenzierte Kennzeichnung sollte zudem den Nutzen jener speziellen Wortarten innerhalb einer Mehrwortgruppe verdeutlichen. Artikel bekamen ein R, Konjunktionen ein U und Präpositionen die Wortklasse C zugewiesen. Weiterhin wurden für die Musterbildung die Lingo Standardwortklassen A (Adjektive) und K (Komposita) genutzt. Komposita entstehen im Rahmen des Indexierungsprozesses.[3] Ihnen wird automatisch die Wortklasse K zugeteilt (s. Tabelle 1).

[3] Die Wortbildungsmöglichkeiten im deutschen Sprachgebrauch sind vielfältig. Deswegen werden mit Lingo keine Wörterbücher lexikalisierter Komposita aufgebaut. Diese werden im Indexierungsprozess algorithmisch durch einzelne Wortbestandteile identifiziert und extrahiert [GLN 12].

Wortklasse	Bedeutung
E	Kunsthistorische Fachbegriffe
A	Adjektiv
K	Komposita
C	Präposition
R	Artikel
U	Konjunktion

Tabelle 1 zeigt die verwendeten Kennzeichnungen der Wortklassen und deren Bedeutung

3.3 Kriterien zur Bildung der Extraktionsmuster

Als Basis zur Musterbildung dienten zuvor festgelegte Kriterien. Diese folgen typischen Erscheinungsformen zusammengesetzter Terme im Deutschen. Beispielsweise stehen Adjektive als Beiwort vor einem Substantiv (mit Wortklasse E und K gekennzeichnete Terme), da diese durch das Adjektiv näher definiert werden. So werden Adjektiv-Substantiv Verbindungen als vielfach eingesetzte Kombination in den Wortmustern verwendet, z.B. ECAE – „Dreizack auf antiker Darstellung“. Funktionswörter wurden gezielt platziert, damit sie ihren syntaktischen Zweck erfüllen. Zum Beispiel wurden Artikel vorzugsweise vor Substantiven oder Adjektiv-Substantiv Verbindungen (AE, AK) platziert, wie es die Extraktionsmuster ERAE – „Symbol des eucharistischen Christus“ oder AERAE – „berühmte Komposition der italienischen Malerei“ verdeutlichen. Mehrwortgruppen beginnen und enden nicht mit einem Funktionswort. Einerseits würden Wortmuster dieser Art unvollständige Mehrwortgruppen erzeugen, da ein Funktionswort immer in Verbindung mit einer zusätzlichen Wortklasse, wie dem Substantiv, zu sehen ist. Anderseits wird in diesem Fall nur der Inhalt einer verkürzten Mehrwortgruppe repräsentiert, da Funktionswörter keine eigenständige, inhaltstragende Bedeutung besitzen. Demzufolge wurden die Wortklassen von Funktionswörtern nur innerhalb der eingesetzten Extraktionsmuster positioniert. Die Muster setzen sich aus mindestens drei, maximal aus sechs Bestandteilen zusammen, denn simple Wortmuster wie AE oder AK sind in dem Anwendungsfall nicht zielführend. Dieses Vorgehen bot die Möglichkeit eine Vielzahl von Mehrwortgruppen unterschiedlicher Länge, unter Berücksichtigung aller zur Verfügung stehender Wortklassen und deren als sinnvoll zu erachtender Kombinationen, zu bilden und zu testen.

Ein feststehendes Set von Wortmustern wurde nach umfangreichen Testindexierungen festgelegt.[4] Die für die Untersuchung genutzten Wortmuster umfassten 80 Einträge:

[4] Hierbei handelt es sich um ein Testset verwendeter Wortmuster. Es soll deshalb nicht ausgeschlossen werden, dass weitere Muster existieren, die den zuvor festgelegten Kriterien folgen.

Muster mit drei Bestandteilen:

EEE, KEE, EEK, EKK, ARE, ACE, AUE, ERE, KCE, ECK, EUE, EUK, KUE, KRE, ECE, EAE, KAE, AEE, KUK, EKE

Muster mit vier Bestandteilen:

ECEE, EUEE, EUEK, AECE, AKRK, AKUK, ECAE, ERAE, KRAE, ERAK, KCAE, AKUE, AKRE, KUEE, KCAK, EUKE, EAEE, ECKE, ERKE, EARE

Muster mit fünf Bestandteilen:

ACCEE, ACRAE, AERAK, AERAE, EERAE, AECEE, ACEUE, ECRAE, ECCAE, ECCEE, ECREE, EUERE, KUERE, EUEAE, ECEUK, ECAEE, KCREE, KCCAE, EEREE, EEUEE

Muster mit sechs Bestandteilen:

AECCEE, EECCAE, AKCRAE, KRAEAE, EUECAK, EUKRAE, KACRAE, EACRAE, EUECKE, EUERKE, AECEUE, EUECRE, AKCKUE, ECAKRE, ACEUEE, AECEEE, AKCAEE, ECRAEE, EUKRAE, ERAEAE

4 Kriterien zur Bewertung fachterminologischer Mehrwortgruppen

Zur qualitativen Beurteilung der Extraktionsmuster und den daraus resultierenden Mehrwortgruppen wurden Kriterien erstellt, an denen sich die darauffolgende Analyse orientiert.

- Abgeschlossenheit: Mehrwortbegriffe sollen nach der Extraktion als zusammenhängende Einheit (semantisch abgeschlossen) und nicht als unvollständige Satzfragmente vorliegen. Zum Beispiel stellen „abgesehen von Kloster und Wallfahrtskirche können“ oder „abendländisch von Anfang“ keine repräsentativen Mehrwortgruppen dar.

- Kombination aus Personennamen bzw. Geografika: Personennamen oder Ortschaftsbezeichnungen, die sich spezifiziert durch zusätzliche Begriffe zusammensetzen, zum Beispiel „Barocke Freskomalerei in Schlesien“ oder „Holzschnitt von Georg Lemberger“. Die Identifizierung von Personennamen und/oder Geografika erzeugt kunsthistorischen Inhalt.

- Zusammensetzungen kunsthistorischer Sachbegriffe: Mehrwortgruppen, welche sich aus zwei Fachtermen zusammensetzen. Diese repräsentieren bereits auf der Wortebene kunsthistorische Terminologie, wie durch die Mehrwortgruppen

„Wallfahrt und Pfarrkirche Maria“ oder „Sündenfall und Erlösung Christus“ deutlich wird.

- Zusammensetzungen kunsthistorischer Sachbegriffe und Alltagssprache: Mehrwortgruppen werden als positiv erachtet, wenn sich diese aus Begriffen zusammensetzen, die allgemeinerer Natur sind, jedoch durch Fachtermini präzisiert werden, wie „Verständnis der biblischen Symbolsprache“ zeigt.

5 Analyse

Die Indexierung des RDK umfasst ca. 8.000 doppelspaltige Lexikonseiten. Vier- und fünfteilige Erkennungsmuster lieferten die besten Ergebnisse, wohingegen drei- und sechsteilige Muster überwiegend inhaltlich unvollständige, nicht abgeschlossene Mehrwortgruppen identifizierten.

Die nachfolgende Analyse zeigt einen Ausschnitt positiver Ergebnisse, im Sinne der vorab genannten Kriterien, welche aus der Identifizierung von vier- und fünfteiligen Erkennungsmustern resultieren.

Wortmuster, welche mit einem substantivischen Begriff (Wortklasse E oder K) eingeleitet werden, gefolgt von einem Funktionswort (Artikel oder Präposition), welcher den ersten Term mit einer Adjektiv-Substantiv-Wortfolge in Verbindung setzt, erzeugen durchweg positive Ergebnisse.

ERAE	„theorie der schön kunst“
	„symbol der jungfäulichen geburt“
ECAE	„abwehrzauber gegen böse dämon“
	„abbildung von neun erzengel“
KRAE	„edelmetallkunst der katholisch kirche“
	„doppelfunktion der architektonisch wirkung“
ERAK	„entwicklung des abendländisch stufenportal“
	„abbildung des jüdisch tempelbau“
KCAE	„elfenbeinrelief in silbervergoldet rahmung“
	„bilderkreis auf alttestamentlich szene“
KCAK	„kunstlandschaft im romanisch kleinkirchenbau“
	„grundrißgestaltung an rheinisch bettelordensbasiliken“

ERKE „darstellung der himmelserscheinung maria“

„personifikation des wegweisend stern“

Auch anhand der getesteten Wortmuster AECE, AKRK und AKUE, welche mit einem Adjektiv beginnen, zeigen sich die positiven Ergebnisse. Das Funktionswort (Präposition, Artikel oder Konjunktion) verbindet eine Adjektiv-Substantiv-Verbindung (AE oder AK) und einen einzelnen Fachterm (E oder K). Oder im Falle fünfteiliger Muster, durch die Kombination: Adjektiv-Substantiv-Verbindung – Funktionswort – Adjektiv-Substantiv-Verbindung.

AECE „adorierend engel auf giebel“

„allegorisch auslegung auf sündenfall“

AKRK „früh holzschnittdarstellung einer schulszene“

„spiritualistisch lichtmetaphysik des neuplatonismus“

AKUE „weiß leinentuch als altarbekleidung“

„malerisch bildgestaltung und farbgebung“

AERAE „episch dichtung der höfisch zeit“

„eucharistisch bedeutung des gekreuzigt christus“

AERAK „künstlerisch möglichkeit des optisch farbaufbau“

„lehrhaft erklärung der antik götterbild“

Die Relevanz von Personenamen oder geografischen Begriffen in der kunstgeschichtlichen Terminologie muss auch bei der Extraktion entsprechender Mehrwortgruppen berücksichtigt werden, die diese Begriffe als Bestandteile enthalten. Es zeigte sich, dass vor allem Muster bestehend aus EE, Personennamen in der Kombination Vor- und Nachname, extrahieren. Da es sich bei den mit E getaggten Begriffen allerdings auch um Sachbegriffe handeln kann, sind Kombinationen von Vor-oder Nachnamen möglich, die durch einen Sachbegriff präzisiert werden.

EEREE „historia scholastica des petrus comestor“

“burgmair totenbild des Conrad celtis”

ECEE “altar von georg raphael”

„aquarell von william blake“

ECREE „romgedanke in der kunst bernini“

„christus in der protestant kunst“

Die durch ein E extrahierten Geografika werden in Verbindung mit einem Sachbegriff (E) oder Adjektiv-Substantiv-Wortfolgen (AE) spezifiziert.

KCREE	„wandgemälde in der capella greca“
AECEE	„dorisch portal am schloß aschaffenburg“
EEREE	„decretum gratiani der paris bibliothek“
ECEE	“altar im wien stephansdom”
ECREE	„altartuch aus dem zisterzienserinnenkloster zehdenick“

6 Fazit

Die positiven Ergebnisse der Analyse zeigen, dass das vorgenommene algorithmische Verfahren zur Mehrwortgruppenerkennung und Extraktion zweckmäßig ist. Das Vorgehen wurde bereits an Dokumenten mathematischen Inhalts getestet [Gö12]. Nun belegt sich die Tauglichkeit der Untersuchung auch an kunsthistorischen Datenbeständen.

Durch die eingesetzte Indexierungssoftware Lingo und dessen Programmmodul sequencer konnten fachterminologische Mehrwortgruppen mit kunsthistorischem Bezug extrahiert werden. Die Integration von Funktionswörtern in den Extraktionsalgorithmus ermöglicht zudem, syntaktisch komplex strukturiertere Mehrwortkonstruktionen zu identifizieren.

Relevant für die Erzeugung fachterminologischer Mehrwortgruppen sind die gezielt erstellten Erkennungsmuster. In der Untersuchung wurde nur ein Testset potentieller Wortmuster erstellt und analysiert. Das Bilden weiterer Muster anhand der gebildeten Kriterien hat das Potential zusätzliche, fachterminologische Mehrwortgruppen zu erzeugen. In weiteren Indexierungen können Wortmuster festgelegt werden, die fachterminologische Mehrwortgruppen in Fachtexten identifizieren, aber negative Ergebnisse von vornherein reduzieren.

Die erzeugten Mehrwortgruppen, welche durch eine Indexierung mit dem sequencer extrahiert werden, können als Empfehlung angesehen werden, um diese nach einer intellektuellen Sichtung, hinsichtlich ihrer fachspezifischen Qualität weiter zu verarbeiten. So ist es mit Lingo möglich, Wörterbücher mit Mehrwortgruppen aufzubauen, in denen diese in ihrer grammatikalisch korrekten Form (also nicht in der Grundform) lexikalisiert werden. So kann von vornherein eine Indexierung mit Mehrwortgruppen erfolgen, welche ausschließlich fachliche Relevanz besitzen.

Danksagung

Die vorliegende Arbeit entstand auf Basis meiner Bachelorarbeit am Institut für Informationswissenschaften der Fachhochschule Köln im Studiengang Bibliothekswesen. Meinem besonderen Dank gilt dem Betreuer der Arbeit Prof. Dr. Klaus Lepsky für seine Unterstützung und die Beantwortung zahlreicher Fragen, die wesentlich zum Gelingen der Arbeit beitrugen.

Literaturverzeichnis

[GLN12] Gödert, W.; Lepsky, K.; Nagelschmidt, M.: Informationserschließung und Automatisches Indexieren: Ein Lehr- und Arbeitsbuch. Springer, Berlin, 2012.

[Gö12] Gödert, W.: Detecting multiword phrases in mathematical text corpora. Köln, 2012 http://arxiv.org/ftp/arxiv/papers/1210/1210.0852.pdf

[Le06] Lepsky, K.: Automatische Indexierung des Reallexikons zur Deutschen Kunstgeschichte. In (Harms, I., Giessen, H. W., Luckhardt, H-D., Hrsg.): Information und Sprache: Beiträge zu Informationswissenschaft, Computerlinguistik, Bibliothekswesen und verwandten Fächern. De Gruyter, Berlin, 2006; S. 169-178.

[Le13] Lepsky, K.: Automatische Indexierung. In (Kuhlen, R., Semar, W., Strauch, D., Hrsg.): Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und –praxis. De Gruyter Saur, Berlin, 2013; S. 272-285.

LV06] Lepsky, K., Vorhauer, J.: Lingo: ein open source System für die Automatische Indexierung deutschsprachiger Dokumente. In ABI-Technik. Vol. 1, 2006; S. 18-28.

[REA14] Reallexikon zur Deutschen Kunstgeschichte. http://rdk.zikg.net/gsdl/cgi-bin/library.exe

[ZEN14] Zentralinstitut für Kunstgeschichte: Forschungsstelle Realienkunde. http://www.zikg.eu/forschung/forschungsstelle-realienkunde#Literatur

Erfolgreiches Lernen durch gamifiziertes E-Learning

Friedrich Pawelka, Thomas Wollmann, Jakob Stöber, Tommy Vinh Lam

Ruprecht-Karls-Universität Heidelberg – Hochschule Heilbronn
Master-Studiengang Medizinische Informatik

hpawelka@stud.hs-heilbronn.de, wollmann@stud.hs-heilbronn.de,
jstoeber@stud.hs-heilbronn.de, tlam@stud.hs-heilbronn.de

Abstract: E-Learning ist schon länger als Lernmethode für Faktenwissen etabliert. Der Erfolg hängt jedoch von der Wiederholung und der Motivation des Nutzers ab. Gamifiziertes E-Learning hat den Anspruch genau diese Lücke zwischen Motivation und E-Learning zu schließen und die Lernleistung in Bezug auf Geschwindigkeit und Dauerhaftigkeit des Wissens zu erhöhen. Dazu wird ein Konzept entworfen, welches die Vorteile einer bereits weit verbreiteten Quiz-App für Smartphones nutzt und um Aspekte zum gezielten Lernen von Faktenwissen erweitert. Dieses Konzept wird anschließend in einem Prototypen umgesetzt und getestet. Der Test zeigt erste Tendenzen, dass Motivation und Lernerfolg durch das Konzept verbessert werden. Als verbesserter Lernerfolg wird in diesem Kontext ein besseres Prüfungsergebnis gesehen. Es werden außerdem weiterführende Ansätze zur Erweiterung des Konzeptes diskutiert.

1 Einleitung

Es ist ein allgemein anerkanntes Paradigma, dass motiviertes und wiederholendes Lernen besonders erfolgreich ist. In den letzten Jahren hat das E-Learning vor allem Einzug neben den traditionellen Lernmethoden gefunden [KGÖ08]. Vorteile des E-Learnings liegen in der einfachen Teilbarkeit der Lernunterlagen und deren Orts- und Zeitunabhängigkeit. Durch Gamifizierung kann eine erhöhte Motivation im Vergleich zu traditionellen Lernmethoden erreicht werden [KGÖ08]. Wenn die E-Learning-Anwendung die Möglichkeit bietet, dass menschliche Spieler miteinander spielen und sich vergleichen können, kann dies die Motivation durch den sozialen Faktor weiter steigern [JHS08].

Das Vorbereiten auf Prüfungen ist für Studenten eine große Herausforderung. Vor allem das immer noch im universitären Umfeld notwendige Auswendiglernen von Faktenwissen ist nach [Ber14, MS12] für viele Studenten eine Tortur. Für die Lernform des Memorierens existieren bereits mehrere E-Learning-Ansätze, bei denen es sich jedoch meist nur um eine Digitalisierung des traditionellen Karteikarten-Systems handelt (z.B. CoboCards[1], BrainYoo[2]). Da diese Form des Auswendiglernens selten mit Spaß verbunden wird, kann durch die Gamifizierung solcher Lernmethoden die Motivation für das Auswendiglernen

[1] http://www.cobocards.com/de/ [Stand 23.06.2014]
[2] http://www.brainyoo.de/ [Stand 23.06.2014]

erhöht werden. Eine Gamifizierung führt jedoch nicht zwangsweise zu einem erhöhten Lernerfolg. Es kommt auf die Art und Weise der Umsetzung an [LH11, AD08].

Es wurde gezeigt, dass der Wettbewerbsfaktor maßgeblichen Einfluss auf die Motivation und somit indirekt auch auf den Lernerfolg hat. Quizbasierte Systeme bieten als eine Art der Gamifizierung genau diesen Wettbewerbsfaktor. Dort gestaltet sich ein Vergleich mit anderen Nutzern deutlich einfacher als bei einem Karteikarten-System, bei dem jeder Nutzer selbst entscheidet, ob er das gefragte Wissen korrekt wiedergegeben hat. Neben dem bekannten TV-Quiz Wer-wird-Millionär finden sich auch im E-Learning-Bereich quizbasierte Ansätze (z.B. für die Lernplattform Ilias[3]). Allen Ansätzen fehlt es jedoch am direkten Wettbewerb zwischen den Nutzern, der über einen bloßen Vergleich mit sich selbst oder anderen Nutzern hinausgeht. Die derzeit populäre Smartphone-Anwendung Quizduell bietet genau diese Funktionalität. Im Rahmen eines Pilotprojekts soll nun das erfolgreiche Spielprinzip von Quizduell auf das Memorieren im akademischen Bereich übertragen und erweitert werden.

2 Grundlagen

Im folgenden Kapitel wird kurz auf die Lernform des Memorierens eingegangen, die Grundlagen zu E- bzw. M-Learning und Gamifizierung umrissen und die Anwendung Quizduell und deren Spielprinzip erläutert.

2.1 Lernform Memorieren

Unter Memorieren wird das Auswendiglernen von Wissen verstanden (engl. Rote Learning) und wird meist als gegensätzliches Lernkonzept zu sinnerfülltem Lernen (engl. Meaningful Learning) betrachtet [Aus00]. Auswendiglernen ist eine Form des Lernens, bei dem der Lernende versucht sich den Lernstoff so einzuprägen, dass er ihn originalgetreu wiedergeben kann. Diese Form des Lernens stützt sich vor allem auf häufiges Wiederholen des Lerninhaltes. Nach Blooms Taxonomie gibt es sechs Ziele des kognitiven Lernens [BEF+56]. Bloom unterscheidet in Wissen, Verstehen, Anwenden, Analyse, Synthese und Evaluation, welche hier nach dem von ihm definierten Lern-Schwierigkeitsgrad aufsteigend aufgelistet sind. Demnach bedeutet dies, dass das Auswendiglernen von Faktenwissen weniger kognitiv anstrengend ist als das Lernen von Konzepten bzw. Anwendungswissen.

Es gibt verschiedene klassische Lernmethoden des Memorierens, wie Karteikartensysteme, Listen oder Fragebögen [Pre59, Coh79, Ebb13, Sch00]. Eine weitere Möglichkeit Faktenwissen zu lernen ist die Verwendung von Multiple-Choice-Fragen [Nov10], deren Schwierigkeit darin besteht die falschen Antwortmöglichkeiten so zu wählen, dass

[3] http://www.iliasnet.de/quizapp.html [Stand 23.06.2014]

sie nicht direkt ausgeschlossen werden können. Mit dieser Lernmethode kann neben Faktenwissen auch Anwendungswissen abgefragt werden [Lei11].

2.2 E-Learning

Als E-Learning wird die Übertragung konventioneller Lernmethoden in die digitale Form bezeichnet [Fog02]. Damit die Methoden in dieser Form weiterhin erfolgreich sind, müssen sie entsprechend angepasst werden [Fog02]. Fogg hat dazu im Foggs-Behavior-Modell drei Kernbestandteile identifiziert, die eine erfolgreiche E-Learning-Methode ausmachen: Motivationsfaktoren, Einfachheitsfaktoren und Auslöser. Motivation kann durch Erzeugung von Vergnügen bzw. Schmerz, Hoffnung bzw. Angst und soziale Akzeptanz bzw. Ausgrenzung erreicht werden. Zu Einfachheitsfaktoren gehören u.a. Zeit, monetäre Kosten und physischer Aufwand und als Auslöser werden Faktoren wie Signale oder Impulse verstanden, die zu einer Aktion auffordern [Fog09].

2.3 M-Learning

Mobile Learning oder auch M-Learning wird in der Literatur unterschiedlich definiert und benannt. Im Kontext dieser Arbeit wird darunter jedoch eine Portierung von E-Learning-Methoden auf mobile Endgeräte verstanden [KHKSB07, S.14f]. Die daraus resultierenden Vorteile dieser Lernform liegen in der Ubiquität, d.h. in der Zeit- und Ortsunabhängigkeit des Lernmaterials und der Tatsache, dass selbst alltägliche Leerzeiten, wie das Anstehen in Schlangen, nun als Lernzeit genutzt werden können [KHKSB07, S. 39f]. In diesem Kontext gilt es jedoch zwei Schwierigkeiten zu bedenken. Zum einen kann es durch die kleineren und heterogenen Bildschirme der mobilen Endgeräte und der Ubiquität zu Konzentrationsproblemen kommen [KHKSB07, S. 41f]. Zum anderen ist es notwendig, auf Grund einer meist kurzen Dauer der alltäglichen Leerzeiten den Lernstoff in kleine Sequenzen zu unterteilen [KHKSB07, S. 35].

2.4 Gamifizierung

Unter Gamifizierung (engl. Gamification) wird die Übertragung von spieltypischen Elementen und Vorgängen in spielfremde Zusammenhänge verstanden[4]. Spieltypische Elemente sind z.B. festgelegte Antwortzeiten (engl. Timed Response), Punkte-Systeme (engl. Score Keeping), Bestenlisten (engl. Highscore Lists) oder Auszeichnungen (engl. Badges) [AD08, S. 63f]. Gamifizierung wird in verschiedenen Bereichen verwendet (z.B. um Kundenbindung zu erreichen)[5] aber laut Muntean [Mun11] kann es auch erfolgreich im akademischen Lernumfeld eingesetzt werden. Dort hilft die Gamifizierung Motivation und

[4] http://wirtschaftslexikon.gabler.de/Definition/gamification.html#kurzerklaerung [Stand 23.06.2014]
[5] http://www.marketingweek.co.uk/disciplines//3018554.article [Stand 23.06.2014]

Interesse für bestimmte Arten von Lernen zu erzeugen. Viele Merkmale von M-Learning und Gamifizierung finden sich auch bei der Smartphone-App Quizduell wieder.

2.5 Quizduell

Quizduell ist die deutsche Version der schwedischen, mobilen App Quizkampen und wird von der Firma FEO Media AB entwickelt, welche neben der deutschen Version noch neun andere landesspezifische Versionen pflegen. Seit 2012 gibt es Quizduell für die mobilen Betriebssysteme iOS und Android mit inzwischen laut Google Play Store ca. 5-10 Millionen (Stand: 18.03.14) Installationen[6]. Die Deutsche Welle schrieb am 12.03.2014, dass bereits rund 23 Millionen Menschen Quizduell heruntergeladen haben[7].

Im Vorfeld eines Spiels fordert der Spieler einen anderen Mitspieler zu einem Duell heraus. Akzeptiert dieser, beginnt das Spiel mit der Auswahl einer Kategorie, aus welcher die Fragen gestellt werden sollen. Zur Auswahl stehen drei der insgesamt fünfzehn Kategorien. Bei den drei nun folgenden textuell formulierten Fragen muss der Spieler sich unter Ablauf einer Zeit für eine der vier Antwortmöglichkeiten festlegen. Nach Festlegung einer Antwort oder Ablauf der Frist leuchtet die richtige Antwort auf und, falls bereits vorhanden, die Antwort des Gegenspielers. Sind alle drei Fragen beantwortet, darf der Gegenspieler mit der Beantwortung der gleichen Fragen beginnen. Ist dies geschehen, gilt die Runde als beendet und der Gegenspieler beginnt wieder mit der Auswahl der nächsten Kategorie. Dieser Ablauf wiederholt sich sechs Runden und der Spieler mit den meisten richtig beantworteten Fragen gewinnt das Duell. Ein Sieg erhöht und eine Niederlage verringert die eigene Gesamtwertung, welche in einer Rangliste dargestellt wird.

3 Entwicklung der Lernmethode

Im Rahmen eines Pilotprojekts wird eine Methode entworfen und prototypisch umgesetzt, um den Lernerfolg im universitären Bereich zu verbessern. Die Idee dabei ist die Folgende: die Lernmethode des Memorierens soll durch eine Gamifizierung dabei helfen, in Form eines vorlesungsbegleitenden Tools den Lernerfolg im Studium zu verbessern. Die entworfene Methode und der Prototyp orientieren sich hierbei an der erfolgreichen Umsetzung der Lernmethodik von Quizduell.

3.1 Konzeption

Zunächst werden die Spiel-Mechanismen der App identifiziert und übernommen. Zum Teil sind diese Merkmale jedoch abzuändern, zu erweitern oder neu zu entwerfen, um

[6] https://play.google.com/store/apps/details?id=se.feomedia.quizkampen.de.lite [Stand 23.06.2014]
[7] http://www.dw.de/die-erfolgsapp-quizduell/av-17493265 [Stand 23.06.2014]

im neuen Kontext des universitären Lernens ebenfalls erfolgreich zu sein. Basierend auf dem Foggs-Behavior-Modell müssen durch die Merkmale vor allem Motivation und Einfachheit erzeugt werden. Dadurch wird sichergestellt, dass Studenten mehr Zeit mit der Methode verbringen, das sich wiederum positiv auf den Lernerfolg auswirken kann.

Der Prototyp wird als App für mobile Endgeräte konzipiert und fällt somit in den Bereich des M-Learnings. Die dadurch entstehenden Einschränkungen müssen bei der Konzeption und Umsetzung ebenfalls berücksichtigt werden.

Die Spiel-Mechanismen, auf denen Quizduell aufbaut, sind u.a. Wettbewerb, Rundenbasierung, Zeitbeschränkung, Punkte-Systeme und Highscore-Listen. Diese werden in dem Pilotprojekt zum Teil in geänderter Form übernommen, wie im Folgenden erläutert wird.

Das Kernelement von Quizduell ist die Möglichkeit Fragen im *Wettbewerb* gegen andere Nutzer zu beantworten. Dadurch können diese sich untereinander messen, das wiederum die Motivation zum Spielen erhöhen kann. Dieser Mechanismus bildet auch den Hauptbestandteil des Projektes, wird jedoch um zwei Punkte erweitert. Zum Einen ist es möglich eine Art Trainingsspiel ohne Gegenspieler zu starten. Diese Art zu spielen ähnelt der konventionellen Lernmethode des Memorierens am meisten und wird schon in ähnlicher Form in anderen Anwendungen unterstützt[8]. Zum Anderen kann ein Spiel mit mehr als einem Gegenspieler gestartet werden, welche die Anwendung von Anderen durch den zusätzlichen sozialen Faktor abgrenzt. Dies soll ebenfalls als motivierender Faktor dienen.

Die *Rundenbasierung* der Spiele sowie die *Zeitbeschränkung* bei der Beantwortung der Fragen bilden wichtige Elemente von Quizduell und gehören auch zu den Hauptbestandteilen des Projektes. Beide Mechanismen unterteilen den Lernprozess in kleine Abschnitte mit je einem fest definierten Ziel. Solche expliziten Ziele führen in Kombination mit einem gewissen Schwierigkeitsgrad der Fragen zu einem höheren Maß an Bemühung und wirken somit motivationssteigernd [AD08]. Darüber hinaus helfen diese Mechanismen dabei, der verminderten Konzentrationsfähigkeit zu begegnen, die durch die ubiquitäre Verwendung des Prototypen auftreten kann. Die Nutzer spielen in Pausen zwischen den Lehrveranstaltungen oder in der Bahn, befinden sich also in einer unruhigen Umgebung mit vielen Ablenkungen und können sich nur begrenzt lange konzentrieren. Die Unterteilung des Lernstoffes in kurze Abschnitte verringert nun die notwendige zusammenhängende Konzentrationszeit.

Das Konzept der zeitlichen Beschränkung wird nicht unverändert übernommen, sondern dahingehend erweitert, dass sie sich variabel einstellen lässt. Dies steigert den Wettbewerb und verhindert ein „Schummeln“ durch Nachschlagen der Lösung. Des Weiteren ist die Verwendung von Abbildungen sowohl in Fragen als auch in Antworten möglich. Die Anzahl der möglichen Antworten lässt sich ebenfalls variabel einstellen. Diese drei Erweiterungen im Hinblick auf die Darstellung der Fragen sind hilfreich, um komplexeren Lernstoff abzufragen, wie z.B. die Diagnose eines Radiologischen-Befundes oder die Lösung einer mathematischen Gleichung. Durch die Flexibilität bei der Fragestellung können diese viel konkreter formuliert werden (z.B. kann ein Radiologischer-Befund als Bild statt in verschriftlichter Form eingefügt werden).

[8] http://www.cobocards.com/de/ [Stand 23.06.2014]

Um Konkurrenzverhalten zu erzeugen muss ein direkter Vergleich mit dem Gegenspieler möglich sein. Daher ist es notwendig ein *Punkte-System* zu integrieren. Weitere Funktionen, die von Quizduell übernommen werden, sind nicht unbedingt direkt dem Spielmechanismus zugehörig, jedoch für den Aufbau der Fragenkataloge nötig. Nutzer können Fragen erstellen und einreichen, wodurch die Anzahl der Fragen schneller steigen kann und dabei auch unterschiedliche Typen von Fragestellungen einfließen können. Dies verhindert unter anderem, dass sich Fragen oft wiederholen und trägt zur höheren Motivation der Nutzer bei. Damit die Gesamtqualität der Fragenkataloge möglichst hoch ist, sollten Nutzer Fragen kommentieren können. Moderatoren müssen komfortabel auf die Kommentare zugreifen und daraufhin die Fragen anpassen können. Dieser Ansatz bietet die Möglichkeit iterativ die Qualität des Fragenkatalogs kollaborativ zu verbessern.

Neben adaptierten Merkmalen von Quizduell, muss das Projekt zusätzliche Merkmale aufweisen. Zum einen die Auswahl von konkreten Fragenkatalogen und zum anderen die Bereitstellung verschiedener Statistiken bzgl. des Lernerfolges.

Der gesamte Fragenbestand wird in Lehrveranstaltungen unterteilt. Diese entsprechen den Kategorien von Quizduell. Mehrere Lehrveranstaltungen werden zu Modulen zusammen gefasst. Sowohl Module als auch Lehrveranstaltungen können über die persönlichen Einstellungen jedes Nutzers explizit an- bzw. abgewählt werden. Während eines Spiels werden allen Teilnehmern nur Kategorien zur Auswahl gestellt, die sie alle gleichermaßen in ihren Einstellungen ausgewählt haben. Dies erhöht die Relevanz der gestellten Fragen, denn es werden nur solche Fragen gespielt, die einen Bezug zu aktuellen Kategorien des Nutzers haben, das sich wiederum motivierend auswirken kann.

Neben *Highscore-Listen* sind in der Statistik sowohl die persönliche, als auch die spielerübergreifende Lernrate über alle Fragen aufgeschlüsselt. Die Lernrate beschreibt hierbei den Zuwachs an richtigen Antworten je weiteren Fragenversuch. Der Vergleich dieser beiden Kenngrößen soll dabei als Motivationsfaktor dienen. Darüber hinaus dient die Erfassung der globalen Lernrate zwei Zielen. Zum einen wird bei einer ausreichend großen Nutzerzahl auf dieser Grundlage entschieden, ob durch den entwickelten Prototypen der Lernerfolg verbessert werden kann. Zum anderen wird die Lernrate einzelner Fragen zur Qualitätssicherung verwendet. Eine Frage deren Lernrate besonders groß bzw. besonders klein ist, sollte überarbeitet werden.

3.2 Vorgehensweise

Nach dem Ausarbeiten der initialen Idee in ein Konzept wurde die Umsetzung der Software iterativ geplant und durchgeführt. Ein Zeitplan wurde für die einzelnen Iterationen festgelegt (vgl. Abbildung 1). Im Folgenden werden kurz die bisher abgearbeiteten Meilensteine beschrieben.

In der ersten Phase wurde ein Prototyp entwickelt, der rudimentär die Funktionen der Anwendung abdeckt. Damit erste Tests möglich sind, wurden ca. 200 Fragen von den Autoren für verschiedene Themen der medizinischen Informatik eingepflegt. Dieser Prototyp wurde im Studienkurs der Autoren mit 15 Teilnehmern auf Akzeptanz getestet.

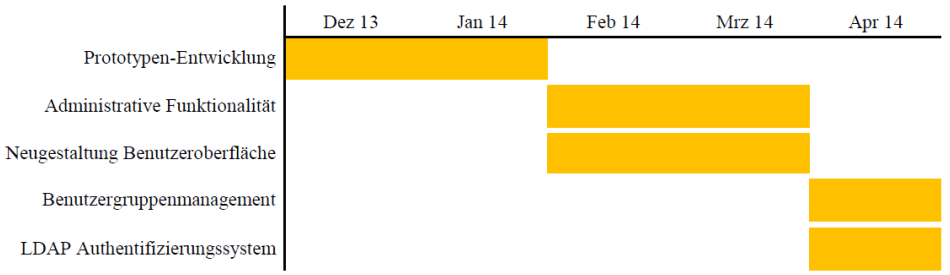


Abbildung 1: Zeitplanung

In der zweiten Phase wurden administrative Funktionalitäten und die Statistikberechnung hinzugefügt, als auch eine neue Gestaltung der Benutzeroberfläche konzipiert. Es war geplant, dass nach der ersten Phase kein weiteres Testen der Software stattfindet. Jedoch wurde aufgrund des positiven Feedbacks und der hohen Aktivität der Spieler darauf verzichtet. Stattdessen wurden weitere Nutzergruppen aus anderen Studienkursen rekrutiert. Die Teilnehmerzahl des Tests ist im April 2014 auf ca. 30 angestiegen.

Ab April 2014 ist eine fachbereichsweite Ausrollung angelaufen und das Benutzergruppenmanagement wurde deutlich erweitert. Zudem wird die Software an das lokale Authentifizierungssystem LDAP angebunden. Es wurde aus dem bisherigen Benutzerfeedback ein neues Oberflächen-Design erarbeitet, welches in dieser Phase eingeführt werden soll. Das Lernkonzept sollen mit diesen Erweiterungen in einer größeren Kohorte evaluiert werden.

3.3 Umsetzungsschwerpunkte

Bei der Umsetzung wird neben der Funktionalität auch großen Wert auf die nicht-funktionalen Anforderungen gelegt. So steht dem Nutzer die Webapplikation durch eine plattformunabhängige Implementierung auf Smartphones, Tablets und Computern zur Verfügung. Das dabei verwendete responsive Design erlaubt es die Anwendung trotz unterschiedlicher Bildschirmauflösungen und -größen adaptiv darzustellen und gegebenenfalls umzubrechen. Ein in diesem Zusammenhang besonderes Feature ist der nahtloser Übergang zwischen den Endgeräten, bei welcher der Nutzer in und zwischen jeder Spielrunde flexibel zwischen Geräten wechseln kann.

Des Weiteren ist ein ansprechendes Design und eine intuitive Bedienung wichtig, um Akzeptanz für die Anwendung zu schaffen. Die Usability soll über das Spiel hinaus die Motivation zum Lernen fördern. Ein auf diese Anforderungen entwickeltes Design wurde entworfen und mit potentiellen Benutzern evaluiert und inkrementell verbessert. Dieses wurde im Rahmen des Alpha-Tests mit getestet. Es wurde dabei Wert gelegt auf die Orientierung des Nutzers innerhalb der Anwendung und eine intuitive Spielführung.

Mit dem Konzept des konkurrierenden Lernens wurde zusätzlich ein Schwerpunkt auf die Präsentation von Statistiken gesetzt. Dabei sollen diese einerseits der Selbstüberprüfung

dienen und andererseits einen Vergleich gegenüber anderen Spielern veranschaulichen. Des Weiteren lassen sich allgemeine Präferenzen der Nutzer wie regelmäßige Nutzung der Applikation oder ausgewählte Kategorien anzeigen. Bei der Auswahl der zu präsentierenden Statistiken wurde darauf geachtet, dass möglichst anonymisierte oder nicht zur Diskriminierung fähige Daten gezeigt werden.

4 Ergebnisse

Im Folgenden werden ausgewählte Funktionen des entwickelten Prototypen vorgestellt. Für die Umsetzung wurde das PHP-Framework *Laravel*[9] verwendet.

(a) Startseite (b) Beantwortung einer Frage (c) Spielübersicht

Abbildung 2: Ausgewählte Ansichten des Protoypen

- *Spielablauf*: Ein Spiel startet damit, dass ein Gegenspieler auf der Startseite ausgewählt wird (vgl. Abbildung 2(a)). Dieser Spieler erhält eine Nachricht, dass er eine Einladung erhalten hat. Diese kann er annehmen oder ablehnen. Nimmt er an, darf er die erste von drei Kategorien aussuchen und die zugehörigen Fragen beantworten (vgl. Abbildung 2(b)). Zwischen jeder Frage wird ein Vorbereitungsbildschirm angezeigt. Dieser ermöglicht das Spiel zu pausieren, das Endgerät zu wechseln oder die vorhergehende Frage zu bewerten. Nach dem Beantworten der Fragen wird der Spieler auf die Spielübersicht geleitet. Nun hat der erste Spieler die Möglichkeit zu spielen. Dies wiederholt sich mehrere Runden. Wenn alle zu spielenden Runden vorbei sind wird ein Ende-Bildschirm angezeigt. Dieser fasst das Spiel zusammen.

[9] http://laravel.com/ [Stand 23.06.2014]

- *Kategorien auswählen*: Jeder Spieler hat die Möglichkeit über ein Auswahlmenü die im Spiel vorgeschlagenen Fragen-Kategorien einzugrenzen. Das System berücksichtigt dies und schlägt nur Gegenspieler mit gemeinsamen Kategorien vor.
- *Spielübersicht*: Die Spielübersicht (vgl. Abbildung 2(c)) zeigt die Anzahl der noch zu spielenden Runden und eine Zusammenfassung der gespielten Runden. Beantwortete Fragen können eingesehen werden und gegebenenfalls wegen mangelnder Qualität der Fragestellung gemeldet werden.
- *Fragen melden*: Wenn eine Frage gemeldet wurde, werden die Moderatoren und Administratoren darüber informiert, welche die Frage daraufhin anpassen können. Diese Funktion ist hauptsächlich für Moderatoren gedacht, da sie im Optimalfall als Lehrkräfte oder Tutoren für die Administration der Inhalte zuständig sind.
- *Fragen einreichen*: Die Anwendung bietet die Möglichkeit für ausgewählte Spieler eigene Fragen durch ein Formular einzureichen. Damit die Fragen formatiert und mit Bildern und Formeln angereichert werden können, wurde die Auszeichnungssprache Markdown[10] implementiert.
- *Fragen Review*: Die eingereichte Frage muss anschließend von einem Administrator für das Spiel freigeschaltet werden.
- *Statistiken*: Da die Statistiken Hauptbestandteil der Umsetzungsschwerpunkte sind werden hier zwei von der Anwendung generierten Diagramme erläutert.

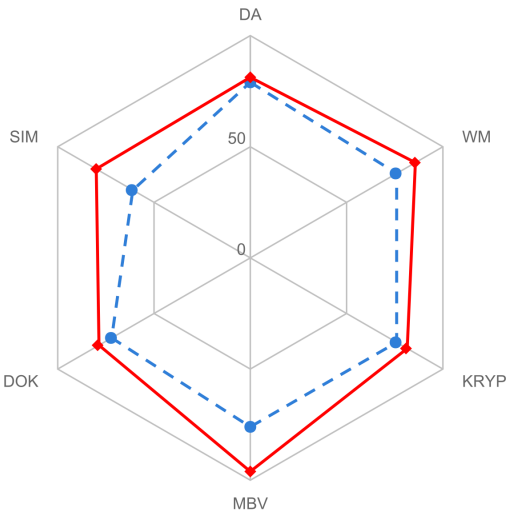


Abbildung 3: Personalisierte Wissenskarte (einzelner Spieler: durchgezogen, Durchschnitt: gestrichelt)

[10] http://daringfireball.net/projects/markdown [Stand 23.06.2014]

Die in der Anwendung bezeichnete *Knowledge Map* (vgl. Abbildung 3) zeigt die Erfolgsquote bei der Beantwortung der Fragen des Nutzers im Vergleich zur durchschnittlichen Erfolgsquote. Die Datenpunkte zeigen für jede Kategorie (Eckpunkte im Diagramm) in Prozent an, wie viele der gestellten Fragen richtig beantwortet wurden. Das Diagramm ermöglicht dem Nutzer zu vergleichen, wie gut er im Vergleich zu anderen Lernenden in den einzelnen Kategorien abschneidet und kann basierend auf diesen Informationen seinen Lernplan optimieren.

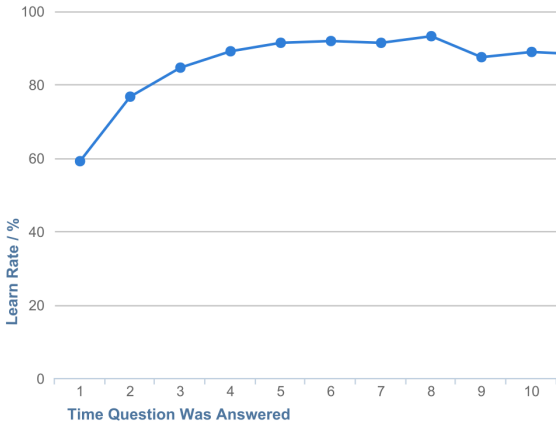


Abbildung 4: Lernrate über alle Fragen und Nutzer

Um den Erfolg der Anwendung in einer Metrik messbar zu machen, wurde eine entsprechende Darstellung (vgl. Abbildung 4) entworfen, die den Lernerfolg der Nutzer abbildet. Hierbei wird die Quote der erfolgreichen Beantwortung in Abhängigkeit zur Häufigkeit, wie oft eine Frage beantwortet wurde, dargestellt. Die Abbildung 4 stellt eine Mittelwertrechnung über alle Fragen und Nutzer dar. Die Diagramme sind auch für jede Frage einzeln abrufbar, um z.B. den Anspruch der Frage zu messen und eventuelle Unverständlichkeiten in der Frage zu erkennen. Um die Signifikanz der Messwerte zu überprüfen, kann eingeblendet werden, wie viele Antworten in die Berechnung eingeflossen sind.

5 Diskussion und Ausblick

Auf Grund der zu geringen Nutzerzahl lässt sich noch keine endgültige Aussage darüber treffen, ob das vorgestellte Konzept des gamifizierten E-Learnings das Lernen erfolgreich beeinflusst. Die Hypothese, die es zu überprüfen gilt, lautet, ob die durch den direkten Wettbewerb erzeugte Motivationssteigerung den Lernerfolg im Vergleich zu den genannten herkömmlichen E-Learning-Ansätzen signifikant erhöht. Eine Möglichkeit wäre eine Art Fall-Kontroll-Studie mit zwei Medizin-Studiengängen an unterschiedlichen Universitäten im Rahmen der Vorbereitung auf die ärtzliche Vorprüfung (ugs.: Physikum), deren

schriftlicher Teil deutschlandweit einheitlich ist. Eine Gruppe würde für die Vorbereitung die entwickelte Anwendung unterstützend benutzen, die andere nicht. Für eine Aussage über den Lernerfolg würde dann der Gruppendurchschnitt der erreichten Punkte zwischen beiden Gruppen verglichen werden. Das genaue Design, die Durchführung und die Auswertung dieser Studie wären dann Teil einer eigenständigen Arbeit.

Allerdings sind bereits in dieser Etappe erste Tendenzen zu erkennen, dass das Konzept akzeptiert wird und zu dem Erfolg des Lernens beiträgt. Die hohe Nachfrage der Testgruppe beizutreten lässt darauf schließen, dass Bedarf für eine solche Anwendung besteht. Das Feedback der Testnutzer hat ergeben, dass sich wiederholende Fragen, schlechte Übersichtlichkeit über Kategorien und eine schlechte Fragenqualität die Motivation senken. Ein Trainingsmodus mit Klausurfragen wurde gewünscht. Testnutzer, die nicht langfristig vor der Klausur anfangen zu lernen, sahen in der Anwendung keinen Vorteil. Um einer häufigen Wiederholung gleicher Fragen zu begegnen, wurde ein Auswahlalgorithmus implementiert, der versucht möglichst abwechslungsreich und für die Spieler anspruchsvolle Fragen zu stellen. Um die Qualität des Fragenkatalogs zu erhöhen wurde ein Review-Prozess eingeführt. Für bessere Übersichtlichkeit über die Kategorien sorgt eine Gruppierungsfunktion. Die Erstellung der Fragen stellt jedoch eine große Herausforderung dar. Die Fragenkataloge sind meist dozenten- und somit hochschul- und semesterspezifisch. Außerdem sind nicht alle Inhalte der Veranstaltungen abfragbar. Faktenwissen ist relativ einfach in Fragen zu fassen. Konzepte können oft mithilfe von Beispielen abgefragt werden. Somit sind auch höhere Ebenen der Bloomschen Lernzieltaxonomie abfragbar [BEF+56]. Aufzählungen abzufragen ist oft nicht möglich, da eine Antwortmöglichkeit benötigt wird, welche *nicht* in der Aufzählung enthalten ist. Somit werden nicht die Elemente der Aufzählung gelernt, sondern genau das eine Element, welches nicht dazu gehört.

Um die Abfrage von Konzepten zu erleichtern, könnten einfache Modelle in Fragen hinterlegt werden. Diese Modelle beinhalten logische Ausdrücke oder Formeln, aus denen dynamisch diverse Beispielszenarien generiert werden. Dies wirkt einer Gewöhnung an spezielle, fest formulierte Szenarien entgegen. Jedoch dürfen Fragen generell nicht zu schwer oder zu einfach formuliert werden. Dies verringert die Motivation der Nutzer und somit auch ihren Lernerfolg [Fog09]. Da die Fragen von den Nutzern selbst erstellt werden, ist die Qualität und das Niveau der Fragen sehr heterogen. Dies macht eine Prüfung und Freischaltung durch einen der Moderatoren für das Fachgebiet notwendig, damit der Fragenkatalog den Mindestanforderungen genügt. Um bereits bei der Eingabe ein Feedback über die Fragenqualität zu erhalten, bietet sich darüber hinaus ein automatischer Vergleich mit bestehenden qualitativ hochwertigen Fragen mit Hilfe von linguistischen Verfahren an. Es lässt sich sagen, dass das Konzept die Chance bietet motivierter, effizienter und mobiler zu lernen.

Literatur

[AD08] L. Ahn und L. Dabbish. Designing Games with a Purpose. *Commun. ACM*, 51(8):58–67, 2008.

[Aus00] David P. Ausubel. *The Acquisition and Retention of Knowledge: A Cognitive View*. Springer Netherlands, 2000.

[BEF+56] S. Bloom, M. D. Engelhart, Edward J. Furst, William H. Hill und David R. Krathwohl. Taxonomy of educational objectives: Handbook I: Cognitive domain. *New York: David McKay*, 1956.

[Ber14] F. Bernstorff. Der entfesselte Workload. Freiräume in modularisierten Studiengängen. *Zeitschrift für Hochschulentwicklung*, 2014.

[Coh79] M. Cohn. *Helping Your Teen-Age Student: What Parents Can Do To Improve Reading and Study Skills*. ERIC, 1979.

[Ebb13] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers college, Columbia university, 1913.

[Fog02] B. J. Fogg. Persuasive Technology: Using Computers to Change What We Think and Do. *Ubiquity*, 2002.

[Fog09] B. J. Fogg. A Behavior Model for Persuasive Design. In *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, Seiten 40:1–40:7, New York and NY and USA, 2009. ACM.

[JHS08] R. Johnson, S. Hornik und E. Salas. An empirical examination of factors contributing to the creation of successful e-learning environments. *International Journal of Human-Computer Studies*, 66(5):356–369, 2008.

[KGÖ08] B. Kleimann, M. Göcks und M. Özkilic. Studieren im Web 2.0. Studienbezogene Web- und E-Learning-Dienste. *HISBUS-Kurzinformation*, (21), 2008.

[KHKSB07] P. Krauss-Hoffmann, M. Kuszpa und M. Sieland-Bortz. *Mobile Learning – Grundlagen und Perspektiven*. Wirtschaftsverlag NW - Verlag für neue Wissenschaft GmbH, 2007.

[Lei11] S. Leitner. *So lernt man lernen: Der Weg zum Erfolg*. Nikol, Hamburg, 18. aufl. Auflage, 2011.

[LH11] Joey J. Lee und J. Hammer. Gamification in Education: What, How, Why Bother? *Academic Exchange Quarterly*, 15(2):2, 2011.

[MS12] C. Metzger und R. Schulmeister. Die tatsächliche Workload im Bachelorstudium. Eine empirische Untersuchung durch Zeitbudget - Analysen. *Bundesministerium für Bildung und Forschung*, 2012.

[Mun11] I. Muntean. Raising engagement in e-learning through gamification. In *Proc. 6th International Conference on Virtual Learning ICVL*, Seiten 323–329, 2011.

[Nov10] J. D. Novak. *Learning, Creating, and Using Knowledge: Concept Maps as Facilitative Tools in Schools and Corporations*. Routledge, 2010.

[Pre59] Ralph C. Preston. *Teaching Study Habits and Skills*. Rinehart, 1959.

[Sch00] Dale H. Schunk. *Learning theories*. Merrill Upper Saddle River, NJ, 2000.

Context-Awareness in Multiagentensimulationen auf Basis von Complex Event Processing

Daniel Steiman

Fakultät Technik und Informatik
Hochschule für Angewandte Wissenschaften Hamburg
Berliner Tor 7
20099 Hamburg, Deutschland
daniel.steiman@haw-hamburg.de

Abstract: Kontextbewusste Systeme bieten die Möglichkeit, dass Anwendungen ihr Verhalten ohne die aktive Einflussnahme von Anwendern selbstständig an vorherrschende Situationen anpassen. Dazu erfassen sie Kontextdaten, die verarbeitet und zur Steuerung der Anwendung verwendet werden. Aus diesem Grund wurden bereits Toolkits, Frameworks und Ansätze entwickelt, die eine einfache Entwicklung von kontextbewussten Anwendungen ermöglichen sollen. Die bisher auf dem Gebiet der *Context-Awareness* vorgestellten Arbeiten weisen jedoch Nachteile auf, die einen Einsatz in bestimmten Anwendungsbereichen stark einschränken oder sogar völlig ausschließen. Dazu zählen vor allem Anwendungsbereiche, die den Anspruch haben riesige Mengen von Ereignissen und damit riesige Volumen an Kontextdaten in Echtzeit verarbeiten zu können. Dieses Paper zeigt die Nachteile bestehender Ansätze auf und befasst sich mit der Entwicklung eines Konzepts, mit dem ein breites Spektrum von Anwendungen möglichst einfach kontextbewusst gemacht werden kann. Um die Einsatzfähigkeit und die Performance des entwickelten Konzepts zu untersuchen, wird es in einem prototypischen Fallbeispiel einer Multiagentensimulation verwendet.

1 Einführung

Der Begriff der *Context-Awareness*, also des Kontextbewusstseins im informationstechnischen Sinne, wurde maßgeblich von Gregory D. Abowd und Anind K. Dey eingeführt [ADB+99]. Abowd und Dey definieren außerdem den Begriff Kontext als Bezeichnung jeglicher Information die verwendet werden kann, um die Situation einer Person, eines Ortes oder eines Objekts zu beschreiben. Eine Anwendung die kontextbewusst ist, erfasst also diese Informationen aus ihrer Umgebung und wertet sie aus, um sich automatisch den aktuellen Gegebenheiten und Ereignissen anzupassen.

Jeder Vorgang in der realen Welt, als auch jede Zustandsänderung in einem System, kann laut David Luckham als ein Ereignis (*Event*) angesehen werden [Luc01]. Mögliche Ereignisse wären zum Beispiel: Ein Flugzeug landet am Frankfurter Flughafen, ein Wettersensor meldet eine Temperaturveränderung, es werden mit einer Kreditkarte 500 Euro von einem Konto abgebucht, ein Musikalbum wird online heruntergeladen. All diese Ereig-

nisse können einen Einfluss auf den Ablauf eines Prozesses nehmen und sich somit auf dessen weiteren Gesamtverlauf auswirken.

Mittlerweile besteht immer mehr der Bedarf, komplexe Abläufe in der realen Welt mit Hilfe von Multiagentensystemen zu simulieren. Dabei können je nach Simulation mehrere hunderttausend Agenten zum Einsatz kommen. Jede Veränderung des Simulationszustands bedeutet eine Kontextveränderung und somit eine riesige Menge von Ereignissen, die kontinuierlich und möglichst performant verarbeitet werden müssen.

Es besteht also nicht mehr das Problem, dass zu wenige Informationen vorhanden sind, sondern die Herausforderung besteht inzwischen darin, aus der Informationsflut die wichtigen Erkenntnisse herauszuziehen und darauf entsprechend zu reagieren. Bei bisherigen Ansätzen ist es so, dass alle Informationen, welche beispielsweise durch das Auftreten von Events entstehen, zunächst gespeichert werden. Danach können entweder manuelle Abfragen von Kontextänderungen gemacht oder automatische Benachrichtigungen über entsprechende Kontextänderungen veranlasst werden. Diese Vorgehensweise ist jedoch für eine Auswertung, der in einer Simulation kontinuierlich auftretenden Ereignisse, zu unperformant. Zudem sind die analysierbaren Daten durch den Zwischenschritt des Abspeicherns und Aufbereitens der Informationen im Normalfall bereits veraltet.

Um jedoch zeitnah auf die gegebenenfalls riesigen Mengen unterschiedlichster Ereignisse einer Anwendung reagieren zu können, wurde die recht junge Technologie des *Complex Event Processing* (CEP) entwickelt [Luc01]. In diesem Paper wird deshalb anhand eines Fallbeispiels untersucht, welche Vorteile durch den Einsatz von Complex Event Processing in einer Context-Service Komponente entstehen können. Die entwickelte Lösung soll sich dabei möglichst leicht in bestehende Umgebungen integrieren lassen und in einem breiten Spektrum von Anwendungsszenarien einsetzbar sein. Es wird außerdem gezeigt, inwieweit das in diesem Paper vorgestellte Konzept bereits bestehende Ansätze auf dem Gebiet der Context-Awareness verbessert.

Der Hauptteil dieses Papers ist in die folgenden Abschnitte unterteilt: Abschnitt 2 betrachtet verwandte Arbeiten, um den Rahmen und die Problemstellung dieses Papers genauer zu eläutern. Abschnitt 3 behandelt das Testszenario der Multiagentensimulation (3.1), den Aufbau der Testumgebung anhand seiner Komponenten (3.2) und das Design der Kontextkomponente (3.3) mit dem die Probleme verwandter Arbeiten gelöst werden. In Abschnitt 4 folgt dann die Zusammenfassung der Ergebnisse und ein Fazit.

2 Verwandte Arbeiten

In den letzten Jahren sind einige Forschungen auf dem Gebiet der Context-Awareness durchgeführt worden. Grundlegende Arbeiten, wie die Definition von kontextbewusstem Computing [ADB+99], sowie die Entwicklung einer ersten Infrastruktur für die prototypische Erstellung kontextbewusster Anwendungen [DAS01], stammen von Anind K. Dey und Gregory D. Abowd. Mittlerweile gibt es neben dem *Context Toolkit* von Dey und Abowd einige weitere Arbeiten, die sich mit Konzepten zur Entwicklung kontextbewusster Anwendungen beschäftigen. Zu den bekanntesten gehören unter anderem *Hydrogen*

[HSP+03], *SOCAM* [GPZ04], *Gaia* [RC03] und *JCAF* [Bar05]. Eine Übersicht über bestehende Arbeiten auf dem Gebiet bieten [BDR07] und [KX12]. Gaia und JCAF gehören dabei bereits zu den Ansätzen, die positive Eigenschaften und Funktionalitäten vorangegangener Arbeiten aufgreifen und diese zu einem eigenen verbesserten Ansatz weiterentwickeln. Aus diesem Grund werden primär diese beiden Arbeiten als Referenz für einen eigenes Context-Service Konzept herangezogen. Die Konzepte von Gaia als auch von JCAF verfügen jedoch trotzdem noch über einige Nachteile und ungelöste Probleme die im weiteren Verlauf betrachtet werden.

Die Kontextmodelle [BBH+10, SLP04] von Gaia und JCAF verwenden jeweils eine Tripelstruktur, um Kontextinformationen zu speichern und zu verwalten. Um verschiedene Kontextinformationen ablegen zu können, werden in JCAF APIs in Form der Interfaces *Entity*, *Relationship* und *ContextItem* vorgegeben. Entwickler sind bei der Modellierung von Kontextinformationen an die jeweiligen Tripelstrukturen gebunden und im Falle von JCAF dazu gezwungen, ihre Anwendung gegen die entsprechenden Interfaces zu implementieren. Andernfalls können Gaia und JCAF die Kontextinformationen weder verwalten noch verarbeiten. Dadurch entsteht das Problem, dass sich diese Ansätze nicht dafür eignen bereits bestehende Anwendungen context-aware zu machen, da sie sich nicht nachträglich integrieren lassen.

Die Anwendung muss also von Anfang an in den Frameworks entwickelt werden. Sowohl bei Gaia, als auch bei JCAF ist die Anwendung tief mit den Frameworks und ihrer Infrastruktur verwoben. Dies verursacht eine starke Kopplung und eine niedrige Modularität der Anwendung. Außerdem kann die Anwendung dadurch nicht mehr in einer beliebigen Sprache implementiert werden, sondern muss die Sprache des jeweiligen Frameworks verwenden.

JCAF verwendet Java RMI zur Verteilung und Kommunikation zwischen der Anwendung (Client) und dem Kontext-Service (Server). Dies hat zur Folge, dass jede Instanz die zur Laufzeit in der Anwendung verwendet wird als Java RMI Stub auf der Serverseite generiert werden muss. Dadurch wird ein entsprechend hoher Ressourcenverbrauch beim Arbeitsspeicher verursacht, welcher bei der Verwaltung von gegebenenfalls mehreren hunderttausend Agenteninstanzen nicht mehr vertretbar ist.

Context Reasoning beschreibt den Vorgang der Deduktion, also der Schlussfolgerung, von für die Anwendung relevanten Informationen aus verschiedenen Kontextinformationen. Dabei werden sogenannte *low-level* Informationen, z. B. durch Aggregation, auf eine höhere Hierarchiebene (*high-level*) abgebildet. Beispielsweise können GPS-Signale als Längen- und Breitengrad vorliegen und dann durch Reasoning auf Orte wie ZUHAUSE oder ARBEITSPLATZ abgebildet werden (siehe Abbildung 1). Ein weiteres Problem besteht darin, dass JCAF über keinerlei Möglichkeiten des Reasonings von Kontextinformationen verfügt. Gaia unterstützt Context Reasoning, welches jedoch auf aussagenlogische Ausdrücke mit den Operatoren UND, ODER und NICHT begrenzt ist. Dadurch ist das Spektrum der möglichen Operationen sehr eingeschränkt. Die Unterstützung temporallogischer Ausdrücke wird deshalb bei Gaia explizit als ein offener Punkt für zukünftige Arbeiten genannt.

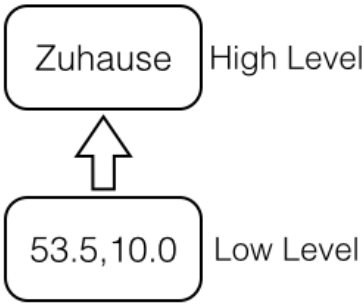


Abbildung 1: Abbildung von Koordinaten (low-level) auf Orte (high-level) durch Reasoning

Neben der Möglichkeit des Context Reasoning fehlt JCAF auch eine umfangreiche *Context Query Language* (CQL) [RWK+08] zur spezifischen Abfrage von Kontext. JCAF bietet APIs an, mit denen sich der gesamte Kontext von Objekten abfragen lässt. Jedoch lassen sich damit keine Auswahlkriterien zur Eingrenzung der Ergebnismenge, wie zum Beispiel bei einer SQL-Abfrage von Datensätzen, formulieren. Gaia dagegen unterstützt etwas spezifischere Abfragen von Kontextinformationen, welche die in [RWK+08] genannten Anforderungen an eine CQL jedoch nur zu einem Bruchteil erfüllen.

Die Performance von JCAF ist, hinsichtlich der Anforderung riesige Mengen kontinuierlich auftretender Kontextereignisse in Echtzeit verarbeiten zu können, ebenfalls nicht ausreichend. Bei einem Testlauf wurde nur ein geringer Durchsatz bei der Verarbeitung von Kontextänderungen erreicht. JCAF scheint also zumindest für möglichst echtzeitfähige Simulationen nicht in Betracht gezogen werden zu können, sobald die zu verarbeitenden Datenmengen groß werden. Da Gaia nicht öffentlich für Testzwecke zur Verfügung steht, können an dieser Stelle keine Aussagen zum Ressourcenverbrauch sowie zur Performance gemacht werden.

In den Ansätzen von Gaia und JCFA wurden also Nachteile und Probleme in den folgenden Teilbereichen identifiziert:

- Systemarchitektur
- Kontextmodellierung
- Context Reasoning
- Abfrage von Kontext
- Performance

In diesem Paper wird deshalb ein Konzept zur einfachen Realisierung von kontextbewussten Anwendungen vorgestellt, welches die zuvor identifizierten Nachteile und Probleme der bisherigen Arbeiten lösen soll.

3 Testumgebung und Beispielszenario

Das in diesem Paper vorgestellte Konzept eines Context-Service, soll bezüglich seiner Verbesserungen in den zuvor genannten Kriterien evaluiert werden. Dazu wird ein Beispielszenario für eine Multiagentensimulation [MN11] entwickelt und eine Testumgebung aufgebaut, in der die Multiagentensimulation gemeinsam mit der Context-Service Komponente als Fallbeispiel zum Einsatz kommt.

3.1 Beispielszenario

Das Beispielszenario basiert auf einem ökologischen Modell, in dem das Verhalten von Geparden und Gazellen in freier Natur simuliert wird [Eck13]. Die Geparden werden in ihrem Jagdverhalten simuliert und verfolgen dabei Gazellen. Die Gazellen fliehen wiederum vor den Geparden, solange sie genügend Energie haben. Eine Gazelle kann dem Geparden dann entweder entkommen oder sie wird von dem Geparden erlegt. Während der gesamten Simulationszeit verändern sich nun in jeder Simulationsiteration die Zustände der Gazellen und Geparden, wie zum Beispiel ihre Position, ihre Energie usw. und damit ihr derzeitiger Kontext.

Die Agenten sind grundsätzlich kontextsensitiv. Diese Kontextsensitivität ist jedoch auf auf das Bewusstsein über ihren eigenen Zustand und ihre nähere Umgebung zu einem bestimmten singulären Zeitpunkt beschränkt. Mit Hilfe des Context-Service kann diese Beschränkung überwunden werden. Er ermöglicht eine globale Erfassung von Kontexten in der Simulation über definierbare Zeitabschnitte. Dadurch können auch Ereignisse bzw. Teilkontexte, die räumlich (spatial) und ggf. auch zeitlich (temporal) weit auseinander liegen, miteinander in Korrelation gebracht werden. Die Simulation kann also nun zusätzlich auch globale Kontexte wahrnehmen und falls nötig den Simulationsablauf entsprechend anpassen. Die Kontextmuster, die während der Simulation automatisch erkannt werden sollen, können mit Hilfe von deklarativen Regeln im Context-Service definiert werden.

Um nun für den Simulationsverlauf relevante Situationen erkennen und entsprechend darauf reagieren zu können, werden die dafür benötigten Kontextdaten kontinuierlich als Ereignisse an den Context-Service gesendet. Dort kann dann überprüft werden, ob entsprechend definierte Kontextmuster vorliegen und die Anwendung darüber informieren. Die Anwendung kann nun eine entsprechende Reaktion in der Simulation einleiten und sich den aktuellen Gegebenheiten anpassen. Der Context-Service kann also so gesehen den Ablauf der Simulation beeinflussen. Die Anwendung wäre somit context-aware.

3.2 Teilkomponenten der Testumgebung

Die Testumgebung besteht im wesentlichen aus einer Multiagentensimulation, einer *Message Oriented Middleware* zur Übertragung von Nachrichten und dem Context-Service. Die drei Teilkomponenten, werden im Folgenden kurz erläutert.

3.2.1 Multiagentensimulation

Laut einer Evaluation von [RLJ06], eignet sich *Repast* für eine einfache Realisierung und vor allem für eine performante Simulation von komplexen Modellen am besten. Deshalb wird zur Implementierung und Visualisierung des Beispielszenarios, das Simulationstool *Repast Simphony*[1] verwendet. Dabei handelt es sich um ein Open Source Tool, welches neben hilfreichen APIs zur Umsetzung einer Multiagentensimulation auch eine integrierte Visualisierungsmöglichkeit bietet. Die damit realisierte Simulation des Besipielszenarios wird mit dem Context-Service verbunden.

3.2.2 Message Oriented Middleware

Zur Kommunikation zwischen dem Context-Service und der Multiagentensimulation, wird die Open Source Message Oriented Middleware *ActiveMQ*[2] mit einer *Java Message Service* (JMS) Schnittstelle verwendet. Dies dient zur Entkopplung der Kontext-Komponente vom Rest der Anwendung. Über die Middleware werden Kontextdaten als Nachrichten an den Context-Service übermittelt und dessen Antwortnachrichten von der Multiagentensimulation empfangen.

3.2.3 Context-Service

Der Context-Service empfängt kontinuierlich die Snapshots des aktuellen Simulationszustands und wertet diese in Echtzeit aus. Der Context-Service kann grob in vier wesentliche Bestandteile unterteilt werden, die im Folgenden genannt werden:

1. Context Server: Stellt die Verbindung zur Anwendung her und enthält Event Consumer. Ein Event Consumer ist in diesem Fall ein Parser, der die über die Message Queue ankommenden Ereignisse konsumiert und diese an die CEP-Engine zur Auswertung übergibt.

2. CEP-Engine: Übernimmt die kontinuierliche Auswertung der eingehenden Events und überprüft diese auf definierte Ereignismuster. Als CEP-Engine wird hier die Open Source Version von *Esper*[3] verwendet. Die Unterstützung einer SQL-basierten Context Query Language, temporaler Kontextregeln, hoher Performance sowie die freie Verfügbarkeit, sind Argumente für die Verwendung von Esper.

3. Regelbasis: Hier sind die Regeln hinterlegt, welche die Ereignismuster definieren die erkannt werden sollen und bei deren Erkennung eine entsprechende Reaktion ausgeführt werden soll.

4. Rule Listener: Lauscht kontinuierlich, ob ein in der Regelbasis hinterlegtes Ereignismuster erkannt wurde. Hier kann dann eine entsprechende Antwortnachricht generiert werden, die festlegt welche Prozedur in der Endanwendung aufgerufen werden

[1] http://repast.sourceforge.net/repast_simphony.php
[2] http://activemq.apache.org/activemq-590-release.html
[3] http://esper.codehaus.org/esper/download/download.html

soll. Optional können hier neue komplexe Ereignisse erzeugt und direkt wieder zur Auswertung an die CEP-Engine geschickt werden.

Der Aufbau des Context-Service ist in Abbildung 2 dargestellt.

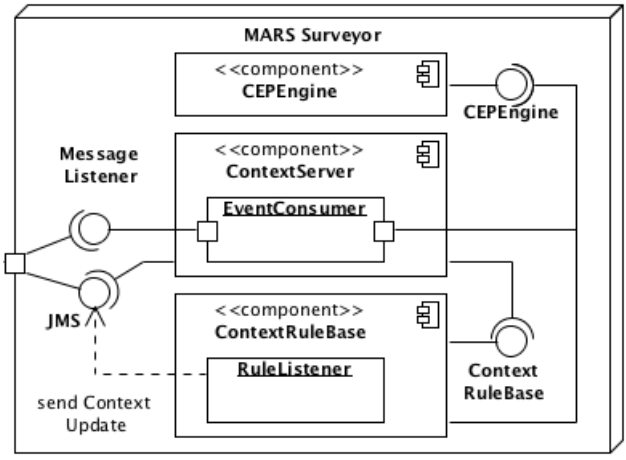


Abbildung 2: Runtime Architektur des Context-Service

3.3 Design

Das in diesem Paper vorgestellte Konzept zur Entwicklung eines Context-Service, löst mehrere der in den bereits bestehenden Arbeiten identifizierten Nachteile. Das Vorgehen zur Lösung der in Abschnitt 2 genannten Probleme, wird nun näher erläutert.

3.3.1 Systemarchitektur

Durch eine vollständige Entkopplung des Context-Service von der Anwendung mit Hilfe einer Message Oriented Middleware (MOM), werden mehrere Probleme gelöst. Da die Kommunikation der Anwendung mit dem Context-Service nun ausschliesslich über Nachrichten stattfindet, bleibt die Endanwendung modular und die technische Umsetzung des Context-Service volkommen transparent. Zusätzlich entfällt die Kommunikation per Java RMI, da auf der Seite des Context-Service primär die Verarbeitung der Kontextinformationen stattfindet und keine unnötige redundante Persistierung bzw. Verwaltung der entsprechenden Daten. Der Ressourcenverbrauch wird dadurch stark minimiert und es wird eine sehr lose Kopplung erreicht. Ein weiterer Vorteil der dadurch entsteht ist, dass die Anwendung nun in einer beliebigen Programmiersprache entwickelt werden kann. Um sich mit der Message Oriented Middleware und letztendlich mit dem Context-Service verbinden zu können, muss lediglich ein *Java Message Service* (JMS) Client in der Anwendung vorhanden sein.

3.3.2 Kontextmodellierung

Die Kontextmodellierung in diesem Ansatz basiert auf einem objektorientierten Kontextmodell [BBH+10, SLP04]. Die Kontextdaten der Anwendung werden dabei als Ereignistypen verwaltet und verarbeitet. Ein Ereignistyp ist vergleichbar mit einer Klasse. Er spezifiziert also die Eigenschaften und die Struktur aller gleichartigen Ereignisse. Ein Beispiel für einen Ereignistyp, der die Kontextinformationen von Geparden abbildet, ist in Abbildung 3 dargestellt. Die Ereignisse können in der Anwendung in einem beliebigen Format vorliegen und als Nachrichten an den Context-Service übertragen werden. Die Übersetzung der Nachrichten, in ein für die CEP-Engine verständliches Ereignisformat, kann dann im Context-Service mit Hilfe eines Adapters erfolgen. Dem Entwickler steht also völlig frei, in was für einer Struktur und in welchem Format die relevanten Kontextinformationen in der Anwendung gespeichert und verwaltet werden. Zusätzlich entfällt die Notwendigkeit vorgegebene Interfaces und die damit implizit vorgegebene Programmiersprache zu verwenden. Insgesamt entsteht der Vorteil, dass auch eine bereits bestehende Anwendung nachträglich ohne Probleme und ohne aufwändige Anpassungen context-aware gemacht werden kann.

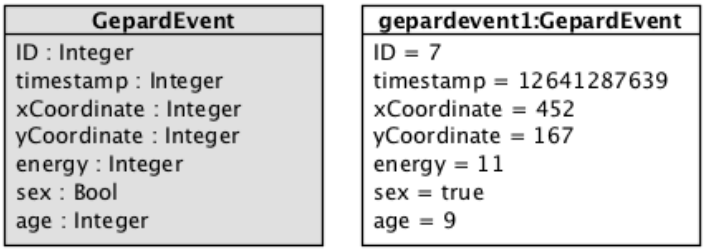


Abbildung 3: Ereignistyp und -instanz

3.3.3 Context Reasoning

In diesem Paper wird die bei JCAF fehlende und bei Gaia begrenzte Möglichkeit des Context Reasoning um wichtige Operationen und die Fähigkeit der Temporallogik erweitert. Die verwendete CEP-Engine unterstützt die Formulierung temporaler und spatiotemporaler Ausdrücke, als auch Operationen wie die Filterung, die Aggregation, das Matching, das Sortieren und das Merging von Kontextinformationen. Dadurch wird nun unter anderem die Erkennung und Auswertung von zeitlichen Zusammenhängen zwischen Ereignissen möglich. Zur Überprüfung wurden dazu in der Regelbasis des Context-Service mehrere komplexe Regeln hinzugefügt. Jeder Regel wurde ein *RuleListener* zugewiesen. Im Test wurden die Regeln erkannt und die entsprechenden Listener aktiv. Dies zeigt, dass das in diesem Paper vorgestellte Konzept die bisherigen Ansätze verbessert und den Nachteil des nicht oder nur begrenzt möglichen Context Reasonings beseitigt.

3.3.4 Abfrage von Kontext

Für die Abfrage von Kontext, wurden bereits mehrere verschiedene Arten von Context Query Languages (CQLs), wie zum Beispiel XML-basierte, RDF-basierte, Graph-basierte oder SQL-basierte CQLs entworfen. Dabei sollte die verwendete Query Language einige wichtige Kriterien und Anforderungen erfüllen [RWK+08]. Eine Übersicht und Bewertung der verschiedenen CQL Typen in [HZK06] ergibt, dass sich neben den RDF-basierten Ansätzen die SQL-basierten Ansätze besonders gut eignen. In diesem Paper wird deshalb eine SQL-basierte CQL verwendet. Dabei orientieren sich die Abfragen an der Basisstruktur gewöhnlicher SQL Querys, weshalb diese Art von CQLs als SQL-basiert bezeichnet werden. Das folgende Beispiel zeigt wie eine Abfrage aussehen könnte, mit der die IDs aller männlichen Geparden die sich in einem Gebiet zwischen 50 und 250 auf der x-Achse aufhalten, abgefragt werden können:

SELECT id **as** gepardId **FROM** GepardEvent **as** gepard **WHERE** gepard.sex = true **and** gepard.xCoordinate **between** 50 **and** 250

Mit der in diesem Paper verwendeten SQL-basierten Context Query Language, lassen sich viele der in [HZK06] genannten Aspekte von Kontext und der in [RWK+08] genannten Anforderungen an eine gute CQL bedienen. Die verwendete CQL bietet damit ein deutlich umfangreicheres Spektrum an möglichen Abfragen von Kontext und Operationen auf Kontextdaten, als die bisherigen Ansätze.

3.3.5 Performance

Die Performance, um die Ereignismengen mehrerer hunderttausend Agenten einer Multiagentensimulation in Echtzeit verarbeiten zu können, soll durch den Einsatz von Complex Event Processing erreicht werden. Der Zwischenschritt, die ggf. riesigen Volumen an zu verarbeitenden Ereignissen zunächst zu speichern, entfällt bei Complex Event Processing. Die Ereignisdaten werden unmittelbar direkt gegen die in der Regelbasis des Context-Service persistierten Ereignisregeln geschickt. Dadurch kann die Verarbeitungsgeschwindigkeit möglicherweise stark gesteigert werden, sodass mehrere hunderttausend Ereignisse pro Sekunde[4] verarbeitet werden können. Dies muss jedoch noch überprüft werden.

4 Schlussfolgerungen und Fazit

Das in diesem Paper vorgestellte Konzept zur Realisierung eines Context-Service zeigt, wie Nachteile bisheriger Ansätze behoben und fehlende Funktionalitäten umgesetzt werden können. Durch grundlegende Designentscheidungen, wie die vollständige Entkopplung des Context-Service von der Endanwendung mit Hilfe einer Message Oriented Middleware, konnten einige Verbesserungen erzielt werden. So wird vor allem die Entwicklung der Endanwendung durch die Verwendung des Context-Service nicht beeinträchtigt. Um

[4] http://.codehaus.org/tutorials/faq_esper/faq.html

den Context-Service verwenden zu können, wird lediglich ein JMS Client benötigt. Durch die lose Kopplung bleibt die Anwendung modular und kann in einer beliebigen Sprache implementiert werden. Die Architektur des Context-Service sowie die technische Umsetzung, bleibt durch die Kommunikation über Nachrichten vollkommen transparent. Dies wird durch die Verwendung eines objektorientierten Kontextmodells in Form von Ereignistypen unterstützt. Es erlaubt eine einfache Modellierung von Kontextstrukturen, die auch in bereits bestehende Anwendungen integriert werden können, ohne diese aufwändig anpassen zu müssen. Durch die Verwendung von Complex Event Processing und einer SQL-basierten Context Query Language, werden temporales Reasoning, umfangreichere Abfragemöglichkeiten und eine performantere Verarbeitung von Kontextdaten ermöglicht.

Um die Leistungsfähigkeit der Context Service-Komponente unter hoher Last zu untersuchen, müssen im weiteren Verlauf der Forschung Performancetests durchgeführt werden. Dabei wird die Anzahl der simulierten Agenten, sowie die Anzahl der in der Regelbasis des Context-Service hinterlegten Ereignisregeln variiert. Zu jeder Konfiguration, wird dann die Anzahl der verarbeiteten Nachrichten und die dafür jeweils benötigte Verarbeitungszeit dokumentiert. Dadurch soll festgestellt werden, wie leistungsstark sich der Context-Service unter den jeweiligen Konfigurationen verhält und wie das Skalierungsverhalten ist.

Bei einem guten Skalierungsverhalten könnte das Konzept letztendlich einen neuen Ansatz für andere Arbeiten bieten, in denen ein entsprechend leistungsfähiger Context-Service benötigt wird, um neue Erkenntnisse erzielen zu können.

5 Danksagung

Ein besonderer Dank gilt Prof. Dr. Thomas Thiel-Clemen, welcher mir bei der Erschließung des Themengebietes geholfen und mich bei der Umsetzung des Projekts mit Ideen und Anregungen tatkräftig unterstützt hat.

Literatur

[ADB+99] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith und Pete Steggles. Towards a Better Understanding of Context and Context-Awareness. In *Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing*, HUC '99, Seiten 304–307, London, UK, UK, 1999. Springer-Verlag.

[Bar05] Jakob E. Bardram. The Java Context Awareness Framework (JCAF); a Service Infrastructure and Programming Framework for Context-aware Applications. In *Proceedings of the Third International Conference on Pervasive Computing*, PERVASIVE'05, Seiten 98–115, Berlin, Heidelberg, 2005. Springer-Verlag.

[BBH+10] Claudio Bettini, Oliver Brdiczka, Karen Henricksen, Jadwiga Indulska, Daniela Nicklas, Anand Ranganathan und Daniele Riboni. A Survey of Context Modelling and Reasoning Techniques. *Pervasive Mob. Comput.*, 6(2):161–180, April 2010.

[BDR07] Matthias Baldauf, Schahram Dustdar und Florian Rosenberg. A Survey on Context-Aware Systems. *Int. J. Ad Hoc Ubiquitous Comput.*, 2(4):263–277, Juni 2007.

[DAS01] Anind K. Dey, Gregory D. Abowd und Daniel Salber. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-aware Applications. *Hum.-Comput. Interact.*, 16(2):97–166, Dezember 2001.

[Eck13] Malte Eckhoff. Ein Agenten-basiertes Gepardenmodell - Konzept und räumliche Wahrnehmung. Hamburg, HH, Germany, 2013.

[GPZ04] Tao Gu, H.K. Pung und Da Qing Zhang. A middleware for building context-aware mobile services. In *Vehicular Technology Conference, 2004. VTC 2004-Spring. 2004 IEEE 59th*, Jgg. 5, Seiten 2656–2660 Vol.5, 2004.

[HSP+03] Thomas Hofer, Wieland Schwinger, Mario Pichler, Gerhard Leonhartsberger, Josef Altmann und Werner Retschitzegger. Context-Awareness on Mobile Devices - the Hydrogen Approach. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 9 - Volume 9*, HICSS '03, Seiten 292.1–, Washington, DC, USA, 2003. IEEE Computer Society.

[HZK06] P.D. Haghighi, A. Zaslavsky und S. Krishnaswamy. An Evaluation of Query Languages for Context-Aware Computing. In *Database and Expert Systems Applications, 2006. DEXA '06. 17th International Workshop on*, Seiten 455–462, 2006.

[KX12] Anup Kumar und Bin Xie. *Handbook of Mobile Systems Applications and Services*. Auerbach Publications, Boston, MA, USA, 1st. Auflage, 2012.

[Luc01] David C. Luckham. *The Power of Events: An Introduction to Complex Event Processing in Distributed Enterprise Systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, USA, 2001.

[MN11] C.M. Macal und M.J. North. Introductory tutorial: Agent-based modeling and simulation. In *Simulation Conference (WSC), Proceedings of the 2011 Winter*, Seiten 1451–1464, Dec 2011.

[RC03] Anand Ranganathan und Roy H. Campbell. An Infrastructure for Context-awareness Based on First Order Logic. *Personal Ubiquitous Comput.*, 7(6):353–364, Dezember 2003.

[RLJ06] Steven F. Railsback, Steven L. Lytinen und Stephen K. Jackson. Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609–623, September 2006.

[RWK+08] R. Reichle, M. Wagner, M.U. Khan, K. Geihs, M. Valla, C. Fra, N. Paspallis und G.A. Papadopoulos. A Context Query Language for Pervasive Computing Environments. In *PerCom 2008. Sixth Annual IEEE International Conference on Pervasive Computing and Communications, 2008*, Seiten 434–440, 2008.

[SLP04] Thomas Strang und Claudia Linnhoff-Popien. A Context Modeling Survey. In *Workshop on Advanced Context Modelling, Reasoning and Management, UbiComp 2004 - The Sixth International Conference on Ubiquitous Computing, Nottingham/England*, 2004.

Entwicklung algorithmischer Skelette für CUDA am Beispiel von Affintiy Propagation

Christoph Winter

Fakultät für Informatik und Mathematik
Ostbayerische Technische Hochschule Regensburg
Prüfeninger Str. 58
93049 Regensburg
christoph.winter@st.oth-regensburg.de

Abstract: In diesem Artikel wird anhand des Clusteralgorithmus Affinity Propagation (AP) eine Bibliothek bestehend aus algorithmischen Skeletten vorgestellt, mit deren Hilfe Berechnungen auf die GPU verlagert werden können. Nach einer kurzen Beschreibung von AP wird eingehend auf die Implementierung der Bibliothek und deren Skelette eingegangen: Sowohl die abstrakten Bausteine als auch die konkrete Umsetzung und Verwendung für AP werden dargestellt. Durch das hohe Abstraktionsniveau der Bibliothek und durch Nutzung etablierter CUDA/C++ Konzepte wie Iteratoren und Funktionsobjekte entsteht eine Sammlung nützlicher Funktionen, die einerseits als Ergänzung zu bestehenden Bibliotheken, andererseits als Basis für weitere Entwicklungen dient. Die Wiederverwendbarkeit, Wartbarkeit und Übersichtlichkeit der Anwendungen werden durch deren Verwendung gesteigert. Eine kurze Analyse über das Laufzeitverhalten im Vergleich zu anderen Funktionssammlungen zeigt, dass die entwickelten Funktionen bessere Laufzeiten erzielen. Dadurch lässt sich AP unter Nutzung der Skelette im Vergleich zu einer sequenziellen Version um Faktor 40 - 50 beschleunigen.

1 Einführung

Im Rahmen eines studentischen Projektes an der Ostbayerischen Technischen Hochschule Regensburg wurde über zwei Semester der Clusteralgorithmus Affinity Propagation [FD07] mithilfe von Graphics Processing Units (GPUs) parallelisiert. Bei AP handelt es sich um ein Verfahren der Clusteranalyse, bei dem aus einer heterogenen Gesamtheit von Objekten homogene Teilmengen (*Cluster*) identifiziert werden sollen. Um die GPUs zu programmieren, wird Nvidia's Compute Unified Device Architecture (CUDA) verwendet. Diese ermöglicht es Entwicklern auf niedrigem Abstraktionsniveau die Abbildung von Berechnungen und Daten auf die Hardware genauestens zu kontrollieren. In CUDA werden Gruppen von Threads in einzelnen logischen Blöcken zusammengefasst. Während die Threads innerhalb eines Blockes synchronisiert werden können, können Blöcke untereinander nicht synchronisiert werden und müssen in jeder beliebigen Reihenfolge ausführbar sein. Blöcke sind somit ein essenzieller Bestandteil, um Parallelität zu fördern. Zur Inter-Thread Kommunikation innerhalb eines Blockes kann schneller Shared-Memory

alloziert werden, wohingegen Threads verschiedener Blöcke ausschließlich über den globalen Gerätespeicher kommunizieren können. Der Programmcode, der auf der GPU ausgeführt werden soll, wird in sog. Kernel-Funktionen beschrieben, welche auf der GPU n-mal ausgeführt werden.

Das Hauptaugenmerk bei der Implementierung von AP lag auf der Abstraktion von Algorithmus und dem CUDA Programmiermodell durch Nutzung von high-level Bibliotheken und Konzepten. Dadurch wird die Wiederverwendbarkeit und Einfachheit der implementierten Kernel und Bibliotheksaufrufe garantiert. Für viele parallele Berechnungen existieren bereits Bibliotheken wie cuBLAS[1], welche aber lediglich eine C Schnittstelle besitzen. Dadurch können Features der objektorientierten Programmierung in C++ wie Datenkapselung, Vererbung, Polymorphie und Templates nicht genutzt werden, obwohl sie von CUDA unterstützt werden. Können einzelne Teile eines Algorithmus nicht direkt mit Funktionen einer Bibliothek realisiert werden, wird häufig ein monolithischer Kernel implementiert, der weder auf ein sich veränderndes Umfeld (z. B. Einsatz auf verschiedenartigen GPUs) angepasst werden kann, noch Wiederverwendbarkeit durch Abstraktion der Berechnung vom Kernel garantiert. Anpassungen an ein anderes Umfeld müssen bei monolithischen Kerneln direkt im Algorithmus verankert werden, wobei das Grundprinzip der Trennung der Verantwortlichkeiten (»Seperation of Concerns«) verletzt wird: Kenntnis von Hardwarespezifika (z. B. Anzahl Threads pro Block, Anzahl Blöcke) wird mit Wissen um den Algorithmus, also *was* berechnet werden soll, vermischt. Hier können algorithmische Skelette nach [Col89] verwendet werden, um mittels Funktionen höherer Ordnung (HOFs) die algorithmische Struktur eines Problems von konkreten Berechnungen zu abstrahieren.

Für die Realisierung von AP ist es ausreichend, Transformationen und Reduktionen als Funktionen höherer Ordnung zu implementieren. Die Thrust-Bibliothek [BH11] stellt derartige Funktionstemplates zur Verfügung, bei welchen der Entwickler kein detailliertes Wissen über die Zielarchitektur benötigt. Er muss lediglich die Berechnungsvorschrift, die parallel abgearbeitet werden soll, spezifizieren und die Bibliothek entscheidet dann anhand der Hardware mit wie vielen Threads und Blöcken die Berechnung ausgeführt wird. Gerade bei den für AP besonders wichtigen zeilen- und spaltenweisen Reduktionen stößt Thrust jedoch an seine Grenzen. Auch hardwarespezifische Besonderheiten wie zum Beispiel CUDA-Streams können durch die Unterstützung mehrerer Backends (CUDA, OpenMP, TBB) nicht genutzt werden. Durch die verschiedenen Backends steigt auch die Komplexität des Thrust-Quellcodes, wodurch sich Erweiterungen und eigene Entwicklungen schlecht in die bestehende Codebasis einbringen lassen. Aus diesen Gründen wurde eine auf Templates basierende high-level Bibliothek für Funktionen höherer Ordnung implementiert, die auf CUDA spezialisiert, mit Thrust vollständig kompatibel und vor allem leicht erweiterbar ist.

Im Folgenden werden verwandte Arbeiten und der Clusteralgorithmus vorgestellt. Anschließend wird auf die Implementierung von AP und der dort verwendeten HOFs eingegangen. Ein Kapitel über die experimentellen Resultate zur Leistungsfähigkeit der Entwicklungen schließt die Arbeit ab.

[1] https://developer.nvidia.com/cublas [Abruf: 18.04.2014]

2 Einordnung der Arbeit

Es wurde bereits viel Forschung in Richtung des *hierarchischen Clusterings* betrieben. Anstelle fester Zuordnungen von Datenpunkten zu Clustern generiert die Methode des hierarchischen Clusterings Bäume aus sukzessive zusammengeführten Clustern oder eines sukzessive geteilten Clusters. Abhängig vom verwendeten Distanzmaß erreicht [CKO09] Speedups von bis zu 48, wohingegen eine neuere Arbeit [SD13] die sequenzielle Ausführung um Faktor 90 übertrifft.

Auch das klassische *k-Means Clustering* wurde mit CUDA in [ZG09] bereits erfolgreich auf eine GPU portiert, wobei nur die Berechnung der Distanzen auf der GPU ausgeführt wurde. Die Cluster-Centroids wurden aufgrund der GPU-Ergebnisse sequenziell von der CPU aktualisiert. [HtLlDt+09] generalisieren beide Berechnungen, sodass sowohl die Zuordnung der Datenpunkte zum Cluster als auch die iterative k-Centroids Berechnung parallel auf der GPU ausgeführt werden. Sie erzielen dabei Speedups von bis zu 40 verglichen mit der sequenziellen k-Means Variante. [ALK07] entwickelt seine vorherigen Forschungsergebnisse zum *Fuzzy c-Means Clusterings* weiter und zeigt, wie die Implementierung angepasst werden muss, um nicht nur auf Basis von euklidischen Distanzen, sondern auch auf Basis der Mahalanobis und Gustafson-Kessel Metrik zu clustern. Da der globale Speicher der GPUs beim Clustering häufig den limitierenden Faktor darstellt, ist die Arbeit von [WZH09] hervorzuheben, welche sehr große Datensätze, die in ihrer Gesamtheit nicht in den Speicher der GPU passen, mittels k-Means clustern und trotz der beschränkten Bandbreite zur GPU einen Speedup von 11 im Vergleich zur sequenziellen Variante erzielen.

In der Arbeit von [KB13] wird eine Umsetzung von Affinity Propagation auf Clustern von GPUs mit OpenCL beschrieben. Während ihr Ziel die Implementierung des Algorithmus war, legt dieses Paper Wert auf die Programmiermethodik, indem die von CUDA unterstützten Features einer Hochsprache für die Implementierung eigener HOFs in Verbindung mit der Thrust Bibliothek genutzt werden. Dadurch soll die Wiederverwendbarkeit der Kernel sichergestellt werden. Mit Muesli [CK10], SkePu [EK10] und SkelCL [SKG12] existieren bereits objektorientierte HOF-Frameworks für GPUs, welche allerdings keine Skelette für spalten- oder zeilenweise Reduktionen von Matrizen anbieten.

3 Affinity Propagation

Bei einer Eingabe von n Datenpunkten benötigt AP im Allgemeinen drei reelle, dicht besetzte $n \times n$ Matrizen, die im Folgenden mit **S**, **A** und **R** bezeichnet werden. Dabei ist die Matrix **S** die sog. Similarity-Matrix, welche die konstanten paarweisen Distanzen der Datenpunkte speichert. Der Wert $s(i,k)$ mit $i,k \in [0,n)$ gibt dabei an, wie passend Punkt i als Exemplar für Datenpunkt k wäre. Ein Exemplar stellt dabei den Mittelpunkt eines Clusters dar, dem alle zum Cluster zugehörigen Datenpunkte zugeordnet werden. Die sog. Preferences $s(k,k) \in \mathbb{R}_{\leq 0}$ geben an, wie wahrscheinlich Punkt k zum Exemplar gewählt wird: je größer $s(k,k)$ desto eher wird k zum Exemplar eines Clusters bestimmt. Aus-

gehend von $\mathbf{S}$ werden nun in einem iterativen Verfahren pro Wiederholung zwei Arten von Nachrichten zwischen allen Datenpunkten ausgetauscht. Diese werden *Responsibilities* bzw. *Availabilities* genannt und werden durch die Werte in den Matrizen $\mathbf{R}$ bzw. $\mathbf{A}$ repräsentiert.

Die Responsibility $r(i,k) \in \mathbf{R}$ bildet dabei das Verlangen von Datenpunkt i ab, dem Cluster anzugehören, dessen Exemplar Punkt k ist:

$$r(i,k) \leftarrow s(i,k) - \max_{l \neq k} \left\{ a^{t-1}(i,l) + s(i,l) \right\}, \tag{1}$$

wobei $l \in [0,n)$, $t \in \mathbb{N}$ und $\mathbf{A}$ mit 0_{nn} initialisiert wird. Der hochgestellte Index $t-1$ bezeichnet dabei den Wert der vergangenen Iteration. Die Berechnung wird realisiert, indem alle Elemente $r(i,\bullet)$ mithilfe des größten Elements der Zeile i aktualisiert werden, außer das größte Element selbst, welches mit dem zweitgrößten Element aktualisiert wird. Da sich die Rechnung jeweils nur auf eine Zeile der Matrix bezieht, können alle Zeilen parallel berechnet werden.

Nach der Berechnung der Responsibilities werden die Availabilities ermittelt. Die Availability $a(i,k) \in \mathbf{A}$ spiegelt die Bereitschaft des Punktes k wider, als Exemplar für i zu dienen. Dabei wird auch die Unterstützung all jener Knoten berücksichtigt, für die ebenfalls der Knoten k als Exemplar vorteilhaft wäre.

$$a(i,k) \leftarrow \begin{cases} \min\left\{0,\ r(k,k) + \sum_{j \neq i,k} \max\{0, r(j,k)\}\right\} & \text{falls } i \neq k \\ \sum_{j \neq k} \max\{0,\ r(j,k)\} & \text{sonst} \end{cases}. \tag{2}$$

Damit die Berechnung der Availabilities zu einem späteren Zeitpunkt effizient auf mehrere GPUs verteilt werden kann, werden zuerst die Spaltensummen ermittelt.

$$c(k) \leftarrow r(k,k) + \sum_{i \neq k} \max\{0,\ r(i,k)\}. \tag{3}$$

Da keine Datenabhängigkeiten zu anderen Spalten bestehen, können die Summen erneut vollständig parallel bestimmt und im Vektor c gespeichert werden. Im Anschluss werden die Availabilities nach Gleichung (4) aktualisiert, wobei jedes Element $a(i,k)$ anhand der Spaltensumme $c(k)$ und dem Element $r(i,k)$ transformiert wird.

$$a(i,k) \leftarrow \begin{cases} \min\{0,\ c(k) - \max\{0,\ r(i,k)\}\} & \text{falls } i \neq k \\ c(k) - r(k,k) & \textit{sonst} \end{cases}. \tag{4}$$

Die obigen Gleichungen werden wiederholt durchlaufen; dabei terminiert AP entweder nachdem die Exemplare über mehrere Iterationen hinweg stabil geblieben sind, oder nach einer a priori festgelegten Anzahl an Iterationen. Dabei ist ein Datenpunkt d_k genau dann ein Exemplar, wenn gilt $r(k,k)+a(k,k) > 0$. Um ein Oszillieren der Nachrichtenwerte zu verhindern, werden zudem alle neu berechneten Werte für $\mathbf{R}$ und $\mathbf{A}$ mit einem $\lambda \in [0.5, 1)$ gedämpft:

$$\mathbf{R^t} = \lambda \mathbf{R^{t-1}} + (\mathbf{1} - \lambda)\mathbf{R}\ , \qquad \mathbf{A^t} = \lambda \mathbf{A^{t-1}} + (\mathbf{1} - \lambda)\mathbf{A}. \tag{5}$$

4 Implementierung

Bevor auf die konkrete Umsetzung von AP eingegangen werden kann, muss zunächst die Struktur der erstellten Bibliothek erläutert werden. Anstelle eines monolithischen Kernels wird jede Berechnung an ein datenparalleles Skelett der Bibliothek weitergeleitet, welches die Berechnung auf der GPU durchführt. Abbildung 1 zeigt den grundlegenden Aufbau eines generischen Transformations- und Reduktions-Skeletts. Ähnlich dem Programmiermodell von CUDA setzt sich die Bibliothek aus drei Bausteinen zusammen: CUDA unterscheidet zwischen Kernel-Funktionen, Host-Code und Device-Code. Während Kernel-Funktionen mit der Anzahl der Threads, Anzahl der Blöcke und der benötigten Größe des Shared-Memory parametrisiert und von der GPU ausgeführt werden können, müssen Unterfunktionen eines Kernels für den Nvidia-Compiler als Device-Code markiert werden. Der Host-Code hingegen ist nur auf der CPU ausführbar und ist deshalb der Standard für alle nicht markierten Funktionen. Diese Struktur findet sich auch in der Bibliothek wieder: Der Aufruf an die Bibliothek ① läuft komplett als Host-Code auf der CPU ab. Ein einfaches Funktionsobjekt (*Closure*) in ② kapselt den Device-Code innerhalb des ( )-Operators und ein Dispatcher ③ führt das Closure aus, indem ein Kernel den ( )-Operator des Funktionsobjektes aufruft. Da der Dispatcher lediglich ein Closure ausführt, ist der Device-Code

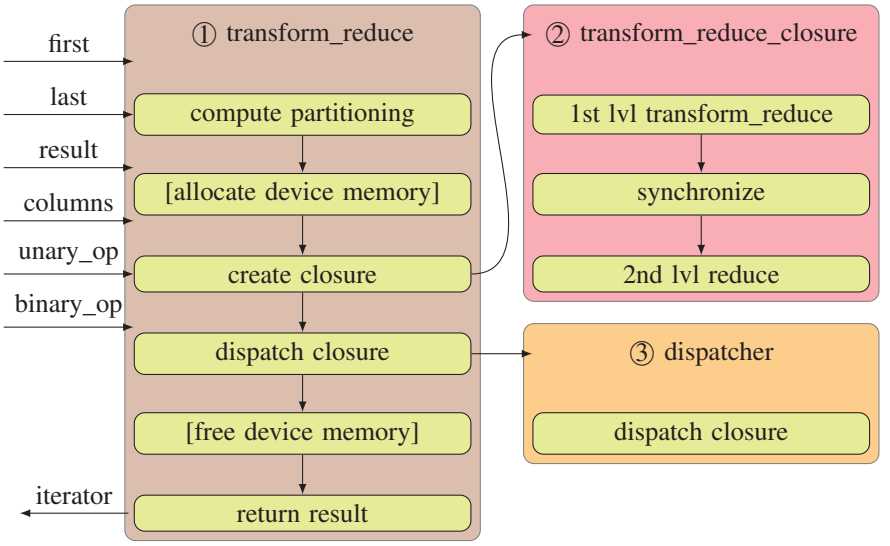


Abbildung 1: Das Blockdiagramm veranschaulicht ein datenparalleles Skelett, bei dem zuerst alle Elemente transformiert und anschließend reduziert werden. Der Aufruf gliedert sich in drei Teile, dem Host- bzw. Device-Code und dem Kernelaufruf. Im Host-Code wird sowohl die optimale Anzahl an Threads und Blöcken berechnet als auch die korrekte Menge an Shared-Memory alloziert und ein Closure, das den Device-Code enthält, erstellt. Anschließend wird das Closure vom Dispatcher in einem Kernelaufruf ausgeführt.

von der Kernel-Funktion, d. h. dem Einsprungpunkt der GPU, getrennt. Dadurch lässt sich

die Logik des Kernels komplett in einem Closure beschreiben. Der ( )-Operator des Closures enthält somit das gesamte Wissen, *wie* die Berechnung auf der GPU ausgeführt wird. Somit können verschiedene Implementierungen, z. B. unterschiedliche Zugriffsmuster auf den Speicher, getestet und iterativ verbessert werden, ohne Änderungen im Anwendungscode vornehmen zu müssen. Von den Verbesserungen profitieren alle Applikationen, die die Bibliothek nutzen, da die Bibliothek als zentrale Anlaufstelle für Berechnungen auf der GPU dient. Dadurch wird sowohl die Wartbarkeit als auch die Übersichtlichkeit aller Anwendungen gesteigert. Die eigentliche Berechnungsvorschrift, d.h. *was* berechnet werden soll, wird ebenfalls über Funktionsobjekte im Aufruf der Bibliotheksfunktion in ① übergeben. Exemplarisch wird im Beispiel sowohl eine unäre Transformation (unary_op) als auch eine binäre Operation zur Reduktion (binary_op) verwendet. In diesen Funktionsobjekten formuliert der Endanwender die Anweisungen der durchzuführenden Berechnung, also welche Daten mit welchen Operationen verknüpft werden sollen. Die Berechnung an sich ist domänenspezifisches Wissen und sollte somit von der Bibliothek strikt getrennt werden. Die Bibliothek übernimmt durch Kenntnis der Hardware lediglich die Partitionierung der Eingabedaten und die Abbildung von Threads auf Partitionen der Daten. Um die Idee zu konkretisieren, wird nun anhand der ersten Gleichung zur Berechnung der Responsibilities die Nutzung der Bibliothek demonstriert.

Nach Gleichung (1) muss zunächst die größte und zweitgrößte Summe aus a^{t-1} und s jeder Zeile bestimmt werden. Da durch die Verwendung von Templates und Iteratoren die Bibliothek von Datentypen und Zeigern entkoppelt wurde, kann die Berechnung auf das folgende Codefragment reduziert werden.

Listing 1: Berechnung des größten und zweitgrößten Elements einer Zeile

```
ap::transform_reduce_row(
    thrust::make_zip_iterator(
            make_tuple( similarities.begin(), availabilities.begin() )),
    thrust::make_zip_iterator(
            make_tuple( similarities.end(), availabilities.end() )),
    thrust::make_zip_iterator(
            make_tuple( mx1.begin(), mx2.begin() )),
    columns,
    convert_to_tuple< thrust::tuple<value_type,value_type> >(),
    reduce_tuple< thrust::tuple<value_type,value_type> >() );
```

Die Bibliothek ist vollständig kompatibel zu allen Iteratoren von Thrust. So wird hier der Zip-Iterator benutzt, um aus den einzelnen Device-Vektoren ein Structure of Arrays (SoA) zu erzeugen [BH11]. Das unäre Funktionsobjekt (convert_to_tuple) addiert die einzelnen Elemente des SoA und transformiert die Summe in ein Tupel, welches mit dem binären Funktionsobjekt (reduce_tuple) reduziert wird. Die Resultate werden pro Zeile in die Ergebnisvektoren mx1 bzw. mx2 zurückgeschrieben und alle Elemente in **R** in einem nachfolgenden, hier nicht weiter aufgeführten Aufruf, entsprechend aktualisiert.

Die Funktionsobjekte sind dabei sehr einfach aufgebaut und enthalten lediglich domänenspezifisches Wissen darüber, *was* berechnet werden soll:

Listing 2: Funktionsobjekte

```
template<class T>
struct convert_to_tuple : public std::unary_function<T,T> {

    __host__ __device__
    T operator()(const T& value) const {
        return thrust::make_tuple(
            thrust::get<0>(value)+thrust::get<1>(value),-MAX_VALUE );
    }
};

template<class T>
struct reduce_tuple : public std::binary_function< T, T, T > {
__host__ __device__
T operator()(const T& a, const T& b) const {
        typedef typename thrust::tuple_element<0,T>::type value_type;

        value_type max1 = thrust::max(
            thrust::get<0>(a), thrust::get<0>(b) );

        value_type max2 = thrust::max(
            thrust::min(
                thrust::get<0>(a), thrust::get<0>(b)),
            thrust::max(
                thrust::get<1>(a), thrust::get<1>(b)));

        return thrust::make_tuple(max1,max2);
    }
};
```

Der Aufruf der Bibliothek sorgt dafür, dass die Reduktion zeilenweise durchgeführt wird. Da alle Matrizen zeilenweise im Speicher abgelegt sind, wird jede Zeile von einem Thread-Block bearbeitet. Da nur ein Block pro Zeile verwendet wird, entfallen sowohl Synchronisationspunkte zwischen den Blöcken als auch das Allozieren globalen Device-Memories zum Speichern der Zwischenergebnisse jedes einzelnen Blocks. Stattdessen schreibt jeder Block seine Zwischenergebnisse der ersten Reduktion in den weitaus schnelleren Shared-Memory [Nvi12]. Nach der Transformation werden in einer ersten Reduktionsphase alle n Eingabedaten auf die Anzahl der Threads pro Block tpb zurückgeführt. In einer zweiten Reduktion werden die Zwischenergebnisse auf ein Endergebnis pro Zeile verdichtet. Eine einfache Heuristik übernimmt die Berechnung der Anzahl der Threads pro Block. Durch den übersichtlichen Aufbau der Bibliothek könnte die Heuristik leicht über einen optionalen Funktionsparameter durch eigene Implementierungen ausgetauscht werden.

Um die Availabilities zu berechnen, werden zuerst die Spaltensummen nach Gleichung (3) gebildet. Damit sichergestellt ist, dass zusammen ausgeführte Threads eines Blocks (*Warp*) von zusammenhängenden Speicherbereichen lesen (*memory coalescing*), bearbeitet ein Thread genau eine Spalte. Ein Block bearbeitet somit eine Gruppe zusammenhängender Spalten, wobei bei einer Eingabe von n Datenpunkten mit tpb Threads pro Block insgesamt $\lceil n/tpb \rceil$ Blöcke benötigt werden. Um die Anzahl der Blöcke und dadurch die Parallelität zu erhöhen, wird die Matrix zusätzlich zeilenweise in m gleichgroße Teile zerlegt, die jeweils parallel transformiert und reduziert werden. Die $m \times n$ vielen Zwischenergebnisse werden in einem temporär allozierten Device-Speicher abgelegt und anschließend in einer letzten Reduktionsphase zu den n Endergebnissen zusammengefasst. Ein letzter Kernelaufruf traversiert die Diagonale $diag(\mathbf{R})$ und addiert alle Elemente den

entsprechenden Spaltensummen hinzu. Nachdem die Spaltensummen nach (3) ermittelt wurden, werden alle Elemente in **A** entsprechend Gleichung (4) transformiert. Während der Transformation werden die Exemplare anhand der Ergebnisse der Transformation bestimmt, wodurch weitere Zugriffe auf den globalen Speicher der GPU entfallen.

5 Experimentelle Ergebnisse

Die Test-Plattform ist mit einer Intel Xeon E5-2667 Hexacore CPU, 6×16 GB RAM und einer Nvidia Tesla K20 mit 5 GB Speicher ausgestattet. Es wurden ausschließlich die Werkzeuge des CUDA Toolkits 6.0 verwendet, sodass der GPU-Code mit dem Nvidia Cuda Kompiler Treiber (nvcc) der Version 6.0.1, der CPU-Code hingegen mit dem GNU C Kompiler 4.7.2 übersetzt wurde. Bei der Kompilierung der Programme wurden keine Optimierungsoptionen angegeben. Die Gerätetreiber lagen in der Version 331.62 vor.

Der limitierende Faktor dicht besetzter Affinity Propagation ist der Speicherverbrauch. Dieser wird von den drei quadratischen Matrizen **S**, **R** und **A** dominiert. Hinzu kommen die Vektoren zum Speichern der Spaltensummen (c), der größten und zweitgrößten Werte pro Zeile ($mx1$, $mx2$), sowie zwei Vektoren (e^t, e^{t-1}), um AP auf Konvergenz zu überwachen. Insgesamt beläuft sich der Speicherverbrauch also auf $3n^2 + 5n$. Verzichtet man auf die Überwachung des Konvergenzkriteriums, kann der benötigte Speicher auf $3n^2 + 3n$ gesenkt werden. Bei einfacher Genauigkeit lassen sich auf einer K20 GPU somit etwa 20 k Punkte, bei doppelter Genauigkeit 14 k Datenpunkte clustern.

In Abbildung 2 werden zur Veranschaulichung die für AP entwickelten HOFs mit anderen Bibliotheken und einer sequenziellen CPU Variante verglichen. In der Abbildung werden die Messpunkte aus Gründen der Übersichtlichkeit auf einer logarithmischen Skala aufgetragen. Um das Skalierungsverhalten besser zu veranschaulichen, wurde jeweils der erste Messpunkt mit der kleinsten Problemgröße und der letzte Messpunkt mit der größten Problemgröße mit der exakten Laufzeit versehen. Ein Algorithmus skaliert dabei umso schlechter, je höher die einzelnen Graphen auf der y-Achse verzeichnet sind. Bei der sequenziellen Variante wird die Zeilensumme klassisch über zwei for-Schleifen berechnet. Die äußere Schleife iteriert dabei über die einzelnen Zeilen, die innere Schleife über die Spalten. Die sequenzielle Laufzeit ist mit bis zu 4.1 s bei der größten Problemgröße mit Abstand am langsamsten; zudem skaliert dieser einfache Algorithmus schlecht. Thrust hingegen weist jedem Wert der Eingabe einen *Key* zu, über den alle zugehörigen Werte reduziert werden. Dadurch werden sehr individuelle Reduktionen möglich, die Verallgemeinerung schlägt sich jedoch auch in einer langsamen Ausführungsgeschwindigkeit nieder. Auch wenn dieser Aufruf mit Speedups von bis zu Faktor 16 im Vergleich zum sequenziellen Aufruf weitaus besser skaliert, können die eigens für AP entwickelten HOFs und die cuBLAS Bibliothek die Laufzeit nochmals um Faktor $9 - 11$ beschleunigen. Bei dem cuBLAS Aufruf wird die Matrix mit einem Einsvektor multipliziert, was der Berechnung der Zeilensummen entspricht. Dabei müssen sowohl für Thrust als auch für cuBLAS zusätzliche Hilfsvektoren der Größe n alloziert werden: Thrust erfasst darin die *Keys* während cuBLAS den Einsvektor speichert. Des Weiteren kann bei keiner der beiden Bibliotheken vor der eigentlichen Reduktion eine Transformation ausgeführt werden.

Stattdessen müsste die Transformation in einem gesonderten Kernel der Reduktion vorangestellt werden. Für das Speichern der Ergebnisse der Transformation muss dann ggf. erneut Speicher alloziert werden. Für das Ermitteln der Spaltensummen in Gleichung (3) würde der Speicherbedarf demnach auf $4n^2+5n$ anwachsen, da eine von negativen Werten bereinigte Matrix $\mathbf{R}^+$ im Speicher alloziert werden müsste. Hinzu kommen die unnötigen Zugriffe auf den globalen Hauptspeicher der GPU, da jede Operation die Daten wieder aus dem globalen Hauptspeicher laden muss. Durch die Verschmelzung von Transformation und Reduktion innerhalb der spezialisierten algorithmischen Skelette lässt sich dieser Mehraufwand vermeiden. Für eine einfache zeilen- oder spaltenweise Summation von Ma-

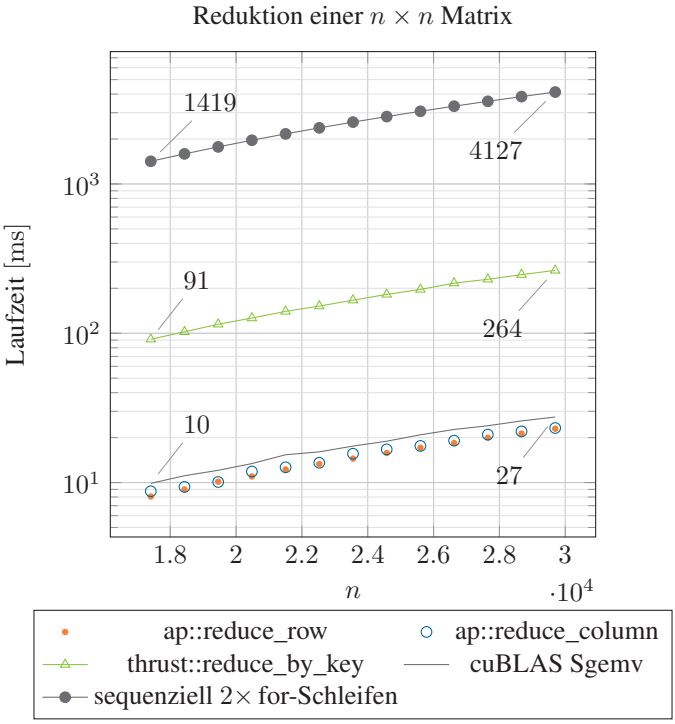


Abbildung 2: Vergleich verschiedener Bibliotheken bei der Berechnung von zeilen- und spaltenbasierten Reduktionen einer $n \times n$ Matrix bei einfacher Genauigkeit.

trixelementen lässt sich mit den neu entwickelten HOFs im Vergleich zur sequenziellen Variante ein maximaler Speedup von bis zu 185 erzielen. Der Speedup relativiert sich jedoch im Zusammenhang mit komplizierteren Berechnungsvorschriften im AP-Algorithmus: Da die Berechnungen oft Datenabhängigkeiten zu anderen Eingaben aufweisen (Gleichungen (1), (2) bzw. (4)), sind weitere Zugriffe auf den globalen Device-Speicher nötig, welche sich negativ auf die Laufzeit auswirken. Um den Speedup von AP zu messen, wurde die GPU Variante mit der originalen sequenziellen Referenzimplementierung in C von Frey & Dueck verglichen[2]. Pro Testlauf wurden 1000 Iterationen jeweils mit einfacher und

[2]http://www.psi.toronto.edu/affinitypropagation/apcluster.txt [Abruf: 18.04.2014]

Problemgröße	Sequenzielle Ausführung [s]	GPU Laufzeit [s]	Speedup
	Doppelte Genauigkeit		
1.000×1.000	21.8	0.8	27.3
2.000×2.000	88.5	2.3	38.3
4.000×4.000	342.0	8.7	39.3
6.000×6.000	759.4	19.1	39.8
8.000×8.000	1335.3	33.6	39.7
10.000×10.000	2050.0	51.6	39.7
12.000×12.000	2972.3	74.4	40
	Einfache Genauigkeit		
1.000×1.000	21	0.7	30
2.000×2.000	84.3	1.8	46.8
4.000×4.000	324.8	6.2	52.4
6.000×6.000	721	14.1	51.1
8.000×8.000	1273.8	23.8	53.5
10.000×10.000	1982	38.9	51
12.000×12.000	2857.8	55.2	51.8

Tabelle 1: Mittlere Laufzeiten von 1000 Iterationen der GPU Variante (ECC aktiviert) im Vergleich zu sequenziellem AP.

doppelter Genauigkeit durchgeführt. Jeder Durchlauf wurde fünfmal wiederholt und der Mittelwert gebildet.

Tabelle 1 zeigt, dass im Vergleich zur sequenziellen Variante Speedups von Faktor $40 - 50$ möglich sind. Rechnet man statt einfacher mit doppelter Genauigkeit, hat man mit Einbußen von bis zu $30\,\%$ zu rechnen, was in etwa dem Verhältnis der Recheneinheiten für einfache und doppelte Genauigkeit einer K20 GPU entspricht. Die Anzahl der gefundenen Cluster ist dabei sowohl bei der GPU- als auch bei der CPU-Variante gleich und bei den verwendeten synthetischen Eingabedaten unabhängig von der Genauigkeit. Auch die Zuordnung von Datenpunkten zu deren Exemplaren stimmt bei allen Varianten überein. Sind in den Ausgangsdaten exakt gleiche Punkte enthalten, kann es lediglich zu einer anderen Auswahl des Exemplars unter allen identischen Datenpunkten kommen, was jedoch keinen Einfluss auf die Güte des Ergebnisses hat. Insgesamt konnte AP durch die Nutzung der entwickelten Bibliotheksaufrufe im Vergleich zur sequenziellen Version deutlich beschleunigt werden. Durch die Zusammenführung von Transformation und Reduktion konnten sowohl Zugriffe auf das globale Device-Memory eingespart als auch der Speicherverbrauch auf das für dicht besetzte AP nötige Minimum reduziert werden.

6 Zusammenfassung und Ausblick

Im Gegensatz zu [KB13] wurde in dieser Arbeit C++/CUDA zusammen mit der Thrust Bibliothek [BH11], vor allem aber eigene high-level Erweiterungen zur Parallelisierung von Affinity Propagation verwendet. Die entwickelte Bibliothek mit ihren hochperformanten algorithmischen Skeletten ermöglicht mit geringem Programmieraufwand den Clusteralgorithmus zu implementieren, wobei die Laufzeiten der Operationen im Vergleich zu Thrust und einer sequenziellen Variante enorm verbessert werden konnten. Verglichen mit einer sequenziellen Version von AP konnte schlussendlich ein Speedup von bis zu Faktor 50 erzielt werden. Die gesamte Bibliothek basiert dabei auf dem Designprinzip der Trennung der Verantwortlichkeiten, sodass das Wissen *wie* eine Berechnung auf der Hardware optimal ausgeführt wird von der Berechnungsvorschrift, d.h. *was* gerechnet werden soll, isoliert wird. Durch diese klare Softwarearchitektur wird sowohl die Wiederverwendbarkeit der Bibliothek als auch die Wartbarkeit und Einfachheit der Endanwendung gesteigert ohne dabei auf Leistung verzichten zu müssen. Eine Masterarbeit wird die Bibliothek um ein CUDA-aware Message Passing Interface (MPI) erweitern. Dadurch wird die Bibliothek auch auf Clustern von GPUs einsetzbar sein. Affinity Propagation wird dabei weiterhin als Beispielanwendung dienen, sodass mehr Datenpunkte geclustert werden können, als der Speicher einer einzelnen GPU fassen kann. Da die Anwendung keine Kenntnis über die tatsächliche Hardware hat, muss lediglich die Bibliothek um das Wissen mehrerer GPUs erweitert werden. Die parallele Implementierung des Clusteralgorithmus für eine GPU wird bereits vom Institut für Funktionelle Genomik der Universität Regensburg[3] für ihre Forschungen verwendet. Dort clustert AP biologische Proben, um Unterschiede im Metabolismus in Abhängigkeit von biologischen Randbedingungen zu erkennen. In einer der vielen Fragestellungen wird untersucht, inwiefern sich der Stoffwechsel von Krebszellen von dem normaler Zellen unterscheidet [DVR+13].

7 Danksagung

Nvidia hat dieses Forschungsvorhaben der Hochschule mit zwei K20 Tesla GPUs im Zuge einer akademischen Partnerschaft unterstützt.

Literatur

[ALK07] Derek Anderson, RobertH. Luke und JamesM. Keller. Incorporation of Non-euclidean Distance Metrics into Fuzzy Clustering on Graphics Processing Units. In *Analysis and Design of Intelligent Systems using Soft Computing Techniques*, Advances in Soft Computing, Seiten 128–139. Springer Berlin Heidelberg, 2007.

[BH11] Nathan Bell und Jared Hoberock. Thrust: A Productivity-Oriented Library for CUDA. *GPU Computing Gems: Jade Edition*, Seiten 359–373, 2011.

[3] https://genomics.uni-regensburg.de [Abruf: 18.04.2014]

[CK10] P Ciechanowicz und H Kuchen. Enhancing Muesli's Data Parallel Skeletons for Multi-core Computer Architectures. *IEEE 12th International Conference on High Performance Computing and Communications (HPCC)*, Seiten 108–113, September 2010.

[CKO09] DJ Chang, MM Kantardzic und Ming Ouyang. Hierarchical Clustering with CUDA/GPU. *ISCA PDCCS*, Seiten 7–12, 2009.

[Col89] Murray I. Cole. *Algorithmic skeletons: structured management of parallel computation*. Pitman London, 1989.

[DVR+13] Katja Dettmer, Franziska C Vogl, Axel P Ritter, Wentao Zhu, Nadine Nürnberger, Marina Kreutz, Peter J Oefner, Wolfram Gronwald und Eva Gottfried. Distinct metabolic differences between various human cancer and primary cells. *Electrophoresis*, 34(19):2836–2847, Oktober 2013.

[EK10] Johan Enmyren und Christoph W. Kessler. SkePU: A multi-backend skeleton programming library for multi-GPU systems. *Proc. 4th Int. Workshop on High-Level Parallel Programming and Applications (HLPP-2010)*, 2010.

[FD07] Brendan J Frey und Delbert Dueck. Clustering by passing messages between data points. *Science (New York, N.Y.)*, 315(5814):972–6, Februar 2007.

[HtLlDt+09] Bai Hong-tao, He Li-li, Ouyang Dan-tong, Li Zhan-shan und Li He. K-Means on Commodity GPUs with CUDA. In *Computer Science and Information Engineering, 2009 WRI World Congress on*, Jgg. 3, Seiten 651–655, 2009.

[KB13] Marcin Kurdziel und Krzysztof Boryczko. Finding exemplars in dense data with affinity propagation on clusters of GPUs. *Concurrency and Computation: Practice and Experience*, 25(8):1137–1152, 2013.

[Nvi12] Nvidia. Cuda C Programming Guide. Oktober 2012.

[SD13] S A Arul Shalom und Manoranjan Dash. Parallel Computations for Hierarchical Agglomerative Clustering using CUDA. *Parallel & Cloud Computing*, 2(4):90–100, 2013.

[SKG12] Michel Steuwer, Philipp Kegel und Sergei Gorlatch. Towards High-Level Programming of Multi-GPU Systems Using the SkelCL Library. *2012 IEEE 26th International Parallel and Distributed Processing Symposium Workshops & PhD Forum*, Seiten 1858–1865, Mai 2012.

[WZH09] Ren Wu, Bin Zhang und Meichun Hsu. Clustering billions of data points using GPUs. *Proceedings of the combined workshops on UnConventional high performance computing workshop plus memory access workshop (UCHPC-MAW)*, Seiten 1–6, 2009.

[ZG09] M Zechner und M Granitzer. Accelerating K-Means on the Graphics Processor via CUDA. In *First International Conference on Intensive Applications and Services, INTENSIVE '09*, Seiten 7–15, April 2009.

Aspekt-basierte Sentiment Analysis

Fritjof Bornebusch, Glaucia Cancino, Alvine Nzeungang Fanseu,
Maryam Farajzadeh Jalali, Jamal Mohsen, Max Nitze, Christina Plump,
Ronald Smith Djomkam Yotedje, Hubert Fred Tchambo, Toni, Henning Ziegler

Arbeitsgruppe Rechnerarchitektur
Universität Bremen
{fritjof,cancino,alvine,maryam,mohsen,maxnitze,cplump,smith,
tfred,toni,hziegler}@informatik.uni-bremen.de

Abstract: Sentiment Analysis (engl. für Stimmungsanalyse) wird verwendet, um Meinungen, Gefühle und Emotionen aus Texten zu extrahieren. In diesem Artikel wird ein Ansatz zur Sentiment Analysis auf Basis von Aspekten beschrieben. Dieser Ansatz kann unter anderem für die automatisierte Klassifizierung von Produktkommentaren, wie sie beispielsweise auf Bewertungsplattformen vorkommen, verwendet werden. Dieser Ansatz gliedert sich in vier wesentliche Teilaufgaben: (i) Extraktion der Aspekte, (ii) Bewertung der Polarität der Aspekte, (iii) Klassifizierung der Aspekte in vorgegebene Kategorien und (iv) Bewertung der Polarität der Kategorie. Zur Umsetzung des Ansatzes werden Methoden der maschinellen Sprachverarbeitung eingesetzt.

1 Einleitung

Sentiment Analysis bezeichnet die Untersuchung von Meinungen, Gefühlen und Emotionen, die in einem Satz oder einem Text ausgedrückt werden ([Liu10]). Das grundlegende Problem bei der Sentiment Analysis ist es zu erkennen, welche Polarität ein gegebener Satz hat ([KK07]), wobei die Polarität angibt, ob der Satz sinngemäß positiv oder negativ ist. Dies wird zum Beispiel von Firmen genutzt, um einzuschätzen, wie ihre Produkte von Kunden angenommen werden. Mit den entsprechenden Rückmeldungen entwickeln sie Strategien, um die Qualität ihrer Produkte zu verbessern ([Vec10]).

Bisherige Arbeiten in diesem Bereich haben sich unabhängig von vorhandenen Entitäten *(engl.: named entity)* ([TP12]) oder Aspekten ([ME10]) immer mit der Analyse ganzer Sätze oder auch Absätze beschäftigt.

Die *SemEval-2014* ist ein Workshop der sich mit der Verarbeitung natürlicher Sprache (engl. *Natural Language Processing (NLP)*) beschäftigt. Ein *Task* der im Zuge dieses Workshops ausgeschrieben wurde ist die *Aspect Based Sentiment Analysis (ABSA)*[1]. Das Ziel in diesem Task ist es, eine Sentiment Analysis durchzuführen, die darauf basiert, dass zunächst die Polaritäten von einzelnen Aspekten eines Satzes oder Absatzes bestimmt werden. Diese werden anschließend auf vorgegebene Kategorien erweitert. Dafür wurde das

[1] http://alt.qcri.org/semeval2014/task4/

Problem in vier verschiedene Teilaufgaben unterteilt. Die erste Teilaufgabe extrahiert aus einem Satz alle vorhandenen Aspekte, die zweite bestimmt die Polaritäten der jeweiligen Aspekte, die dritte die Kategorien der einzelnen Sätze und und die letzte die Polaritäten zu den ermittelten Kategorien. Um das System evaluieren zu können, wurden Trainingsdaten während der Entwicklung zur Verfügung gestellt.

2 Verwandte Arbeiten

Die in Python implementierte Bibliothek Natural Language Toolkit (NLTK)[2] wird von Laurent Luce ([Luc12]) eingesetzt, um Twittermeldungen als positiv oder negativ zu klassifizieren. Dies wird über den NLTK eigenen Klassifizierer realisiert, da dieser durch eigene Sätze, die entweder positiv oder negativ sind, trainiert werden kann.

Die Arbeit von Pang el at. ([PLV02]) befasst sich mit der Analyse von Filmbewertungen, um zu ermittlen, ob ein Film insgesamt als negativ oder positiv bewertet wurde. Hierfür werden Bayessche Filter, die Maximum Entropie Methode und Support Vector Machines verwendet. Es wird in dieser Arbeit mit den verschiedenen Klassifizierern gearbeitet.

Eine weitere Arbeit zu diesem Thema kommt von Kim el at. ([KH06]) und befasst sich mit der automatisierten Analyse von Polaritäten aus Online Bewertungen. Es wird der Ansatz gewählt, die Gründe für eine Meinung auszuwerten und dahingehend zu analysieren, ob diese eher eine Meinung oder Tatsache beschreiben. Auch hier wird die Maximum Entropy Methode verwendet.

Die in [PLV02] und [KH06] beschriebenen Methoden verwenden Klassifizierer, die mit Daten trainiert werden und anschließend zur Analyse der Testdaten eingesetzt werden. In unserer Arbeit werden keine Klassifizierer verwendet, allerdings könnten diese zu einer weiteren Verbesserung der Ergebnisse beitragen.

Bagheri el at. ([BSdJ13]) stellen in Ihrer Arbeit eine Möglichkeit vor, wie Aspekt Terme in Bewertungen gefunden werden können. Diese Arbeit basiert auf einer eigenen Entwicklung bestehend aus dem Lernen von Aspekt-Termen und dem Anwenden von heuristischen Regeln.

Unser Ansatz basiert auf der Verwendung von Bibliotheken. Die Verwendung von selbsttrainierte Klassifizierern wäre ebenfalls eine Möglichkeit, mit der die Ergebnisse verbessert werden könnten.

3 Grundlagen

Die zu bearbeitende Aufgabe basiert hauptsächlich auf NLP-Techniken. Damit ein Computer natürliche Sprache bearbeiten bzw. verstehen kann, sind mehrere Schritte notwendig. Zunächst einmal muss die grundlegende Struktur des vorliegenden Satzes verstanden

[2] `http://www.nltk.org/`

Tabelle 1: Annotationen für den Satz „My dog likes eating sausages.“

Token	my	dog	likes	eating	sausages
Lemma	my	dog	like	eat	sausage
POS	PRP	NN	VBZ	VBG	NNS

werden (Parsing). Dies extrahiert jedoch nur das Gerüst des Satzes und nicht die tatsächliche Bedeutung desselben. Hierfür ist ein Verständnis der Semantik des Satzes (und damit natürlich zunächst einmal der einzelnen Wörter) von Nöten. Dieses Problem wird durch Bibliotheken (WordNet [Mil95], RiTa.WordNet[3]) gelöst, in denen die Bedeutungen aller Wörter und ihrer Synonyme aufgeteilt nach verschiedenen Wortgruppen enthalten sind. Mithilfe dieser Bibliotheken oder Wörterbücher können auch Beziehungen zwischen Wörtern erkannt werden. Hierauf geht der zweite Unterabschnitt dieses Kapitels ein. In einigen Teilgebieten des NLP, vor allem der Sentiment Analysis, ist weiterhin der Grundtenor des Satzes von Interesse. Hierbei steht die Frage im Vordergrund, ob der Satz eine positive oder negative Meinung darstellt.

Der folgende Abschnitt soll einige Begrifflichkeiten im Zusammenhang mit Natural Language Processing erläutern: Ein einfaches Wort, so wie es im Satz steht (das heißt, in seiner gebeugten Form) wird als *Token* bezeichnet. Hierunter fallen auch Satzzeichen und Zahlen. Wird ein Wort auf seinen Wortstamm zurückgeführt, so wird dieser als *Lemma* bezeichnet. Mit einem *Part-of-Speech Tag* wird angegeben, welche grammatikalische Eigenschaft das Wort einnimmt.

Parser geben die syntaktische bzw. grammatikalische Struktur von Sätzen unter anderem durch *Typed Dependencies* ([dMM08]) an. Diese geben Beziehungen zwischen Token des Satzes an. Der `dependent` ist dabei abhängig vom `governour` und die Relation zwischen diesen beiden Wörtern wird durch `relation` beschrieben.

Beispiel 1 *My dog likes eating sausages. Tabelle 1 zeigt die einzelnen* Tokens*,* Lemmata *und* POS-Tags *des obigen Satzes. Tabelle 2 enthält die entsprechenden* Typed Dependencies*, sowie ihre Auflösung nach* `governour`*,* `dependent` *und* `relation`*.*

Der am meisten verbreitete Parser ist der Stanford Parser ([SBMN13], [KM03]). Er wird von der Natural Language Processing Group an der Stanford University entwickelt und basiert auf den Elementen der Penn Treebank.

Im nächsten Schritt muss ein Verständnis für die semantische und lexikalische Bedeutung des Satzes entwickelt werden. Dies geschieht mithilfe der Bibliotheken WordNet [Mil95] und RiTa.WordNet. Die Bibliothek WordNet ermöglicht es, alle Bedeutungen eines Wortes zu überblicken, sortiert nach Worttyp – *noun*, *verb*, *adjective*. Weiterhin sind für jeden Wortsinn alle Synonyme und Beschreibung aufgelistet, sowie Über- und Unterbegriffe. RiTa.WordNet bietet eine Schnittstelle zu WordNet, mithilfe derer auch Informationen aus WordNet erhalten werden können, die nicht direkt abzufragen sind. Eine sehr wichtige dieser Funktionen ist der *Path based similarity algorithm* ([ZY08]). Dieser Algorithmus durchläuft Oberbegriffe der zu vergleichenden Wörter, bis ein gemeinsamer gefunden

[3] `http://www.rednoise.org/rita/`

Tabelle 2: Typed Dependencies für den Satz „My dog likes eating sausages.“

dependency	governour	dependent	relation
`poss`(dog, My)	dog	my	poss
`nsubj`(likes, dog)	likes	dog	nsubj
`root`(ROOT, likes)	ROOT	likes	ROOT
`xcomp`(likes, eating	likes	eating	xcomp
`dobj`(eating, sausages)	eating	sausages	dobj

wurde. Aus der Anzahl der nötigen Schritte, um von Wort zu Wort zu kommen, wird die Ähnlichkeit der beiden Wörter berechnet – je kürzer, desto ähnlicher.

Die Sentiment Analysis soll Wörtern, Sätzen und/oder Abschnitten eine Polarität zuordnen. Mit SentiWordNet [GGT13] gelingt dies auf Wortebene. Es wird der Thesaurus aus WordNet benutzt. Jedem Wort werden Gewichtungen für die drei Polaritäten *positive*, *negative* oder *objective* zugeordnet, wobei sich die drei Werte zu 1 addieren müssen.

Eine mögliche Implementierung, mit deren Hilfe der Grundtenor des gesamten Satz oder einzelner Teilsätze ermittelt werden kann, ist der *Sentiment Tree*. Diese Struktur wurde für die Analyse von Filmbewertungen an der Stanford Universität entwickelt und in die Stanford CoreNLP Bibliothek integriert (vgl. [SPW+13]). Der von dieser Bibliothek zur Verfügung gestellte Sentiment Tree basiert auf der *Sentiment Treebank*, die aus rund zwölftausend annotierten *Parse-Trees* besteht. Das dem Sentiment Tree zugrunde liegende *Recursive Neural Tensor Network*, welches über die Sentiment Treebank trainiert wurde, kann fünf Polaritätsklassen erkennen: *sehr negativ*, *negativ*, *neutral*, *positiv* und *sehr positiv*. Für die Erkennung der Aspekt-Term-Polaritäten fassen wir diese zu den drei wesentlichen Klassen *negativ*, *neutral* und *positiv* zusammen.

4 Implementierung

Zu Beginn werden die eingegebenen Sätze vorbereitet um anschließend auf ihnen eine Sentiment Analysis durchgeführt. Dazu werden unter anderem Annotierungen vorgenommen, die jeder weitere Algorithmus benötigt. Diese werden mit dem Satz zusammen für die weitere Verarbeitung gespeichert.

Die Implementierung gliedert sich, in Anlehnung an die vier Teilaufgaben der *SemEval-2014*, in vier Teilbereiche: Die *Aspect Term Extraction* (Abschnitt 4.1) und die *Aspect Term Polarity* (Abschnitt 4.2), die sich mit der Sentiment Analysis auf Aspekt-Ebene beschäftigen, sowie der *Aspect Category Extraction* (Abschnitt 4.3) und der *Aspect Category Polarity* (Abschnitt 4.4), die dies auf Kategorie-Ebene durchführen.

Zum Testen der Verfahren wurden die im Rahmen der *SemEval-2014* zur Verfügung gestellten Trainingsdaten verwendet. Die Daten enthalten jeweils etwa 3000 englischsprachige Sätze aus Laptop- und Restaurantbewertungen, bei denen per Hand annotiert wurden.

Das Messkriterium, um die Ergebnisse mit den Trainingsdaten zu vergleichen, ist die *F-measure*. Diese setzt sich als als harmonischen Mittel aus der *Precision*, also der Anzahl der richtigen an den gesamten Elementen, und dem *Recall*, dem Verhältnis von richtigen zu falschen Elementen, zusammen.

4.1 Aspect Term Extraction

Im ersten Teil der Aufgabe 4 der *SemEval-2014* sollen in Reviews zu einem bekannten Thema (in diesem Fall *restaurant* und *laptop*) genannte Aspekte erkannt werden. Hierfür wurden zwei Ansätze verwendet: *NounFrequency* und *TermAggregation*.

NounFrequency. Nach [HL04] ist eine gute Grundlage, um Aspekte zu erhalten, eine Liste von Nomen zu verwenden. Da dies zu einer sehr ungenauen *precision* führen würde, ist eine Untersuchung auf Häufigkeit von Nöten. Dafür werden in der Vorverarbeitung alle Nomen nach Häufigkeit sortiert in eine Liste geschrieben. Die ersten 20% werden als mögliche Aspekte betrachtet. Die 20% sind ein willkürlich gewählter Wert, der sich auf den Trainingsdaten als am besten erwiesen hat.

Dieser Ansatz führt zu zwei Problematiken: Zum einen beschreibt der Stanford Parser Wörter, die komplett in Großbuchstaben geschrieben sind, direkt als *proper nouns* (NNP(s)), womit die Häufigkeitsliste verfälscht wird (vgl. Beispiel 2). Zum anderen führen Rechtschreibfehler innerhalb der Reviews zu einer verfälschten Liste und zu falschen Häufigkeiten (vgl. Beispiel 3).

Beispiel 2 *Der Satz sei: „ DO NOT BUY IT.“ Dann werden alle vier Wörter als NNP beschrieben, dementsprechend erreichen vier Wörter, die keine Nomen sind, sowohl die Liste als auch die Liste aller Nomen eines Satzes.*

Beispiel 3 *Der Satz sei: „ Performace was great on this one.“* Performace *soll hier offensichtlich performance heißen. Letzteres hat ein relativ hohe Häufigkeit, der inkorrekte Term jedoch nicht, er wird also fälschlicherweise nicht ausgegeben.*

Ein Lösungansatz für diese Probleme ist es Wörter, die ausschließlich aus Großbuchstaben bestehen vor dem Parsieren in Kleinbuchstaben abzuändern. Worte, die dann noch nicht in der Häufigkeitsliste auftauchen könnten dann einer Rechtschreibprüfung unterzogen werden, wenn sie zum Beispiel in WordNet nicht aufzufinden sind.

Eine weitere Problematik, die durch den Ansatz der häufigsten Wörter entsteht, ist, dass Wörter, die sehr häufig vorkommen (die oberen 2%), häufig nicht Aspekte sind, sondern das Oberthema bzw. Synonyme desselben oder semantisch ähnliche Wörter. Dies gilt auch für andere Wortgruppen, welche in Tabelle 3 mit Beispielen zusammengefasst sind. Diese werden aus der Häufigkeitsliste ausgeschlossen.

Beispiel 4 *Sei das Thema* Laptop *und der Satz „ Hard drives and battery life are alright, but the touchpad doesn’t react as well as hoped and the display becomes unreadable as*

Tabelle 3: Gefilterte Wortgruppen mit Beispielwörtern

	Beispiel
person	husband, wife, mother, boyfriend
time	date, year, month, monday-sunday
location	NYC, Manhattan, street, Avenue
mistagged	everything, something, none, some, any

soon as its not raining outside.“ Aspekte wären in diesem Fall: Hard drives*,* battery life*,* touchpad *und* display.

Der beschriebene grundlegende Ansatz mit seinen kleinen bis großen Modifikationen liefert bereits eine gute Basis des Ganzen. Jedoch – wie auch in Beispiel 4 zu sehen ist – gibt es auch Aspekte, die aus mehreren Wörtern bestehen (z.B. hard drives). Hierfür werden hintereinander vorkommende Aspekte zusammengefasst. Dieser kleine Schritt bringt bereits eine große Verbesserung.

Tabelle 4 fasst die Ergebnisse für die trainingsdaten zusammen. Diese liegen mit 30% bzw. 40% F-measure weit unter den möglichen Werten (vgl. [ZY08]). Ein Verbesserungspunkt ist, mehr mit den Typed Dependencies zu arbeiten (vgl. Algorithmus *AggregatedWords*).

Tabelle 4: Ergebnisse des Noun Frequency Algorithmus

Datensatz (Domain)	Precision	Recall	F-measure
Laptop	0,25	0,38	0,30
Restaurant	0,34	0,45	0,39

Aggregated Words. Der Algorithmus erkennt zunächst Aspekte, die mehrere Wörter enthalten. Danach werden erst die Einzelaspekte identifiziert. Eine tragende Rolle spielen hierbei die Typed Dependencies, wobei nur die Relationen *nn*[4] und *amod* [5] verwendet wurden.

Beispiel 5 *Die relevanten Typed Dependencies des Satzes „It has good speed and plenty of hard drive space“ sind:* `amod`*(speed, good),* `amod`*(space, hard) und* `nn`*(space, drive). Als Aspekte erkannt würden nun:* hard drive space *und* good speed*.*

Der erste Schritt besteht daraus, Wortgruppen zum gleichen `governour` zu finden. Sind diese alle gefunden, werden alle übrigen Nomen als Aspekte identifiziert. Hier entsteht nun eine ähnliche Problematik wie im Algorithmus *Noun Frequency*: Es werden zu viele Nomen erkannt. Dies wird dadurch behoben, dass Listen von *unwichtigen* und *wichtigen* Nomen und Adjektiven angelegt werden und mithilfe dieser Listen die Ergebnisse weiter gefiltert werden. Weiterhin werden alle Wörter, die zu ausgewählten Wortkategorien aus WordNet gehören, gefiltert. Dies führt zu einer besseren *precision*.

[4] noun compound modifier: Beziehung zwischen zwei Nomen
[5] adjectival modifier: Beziehung zwischen Adjektiv und Nomen

4.2 Aspect Term Polarity

Die *Aspect Term Polarity* bestimmt für gegebene Aspekte eines Satzes die dazugehörigen Polaritäten.

Diese Ideen wurden mit Hilfe des *Sentiment Trees* verfeinert. Der Sentiment Tree ist für die Bestimmung der Polarität eines kompletten Satzes ausgelegt und muss für die Bestimmung der Aspektpolaritäten entsprechend ausgewertet werden. Erläutert wird die Funktionsweise des Algorithmus anhand des Beispiel-Satzes „The keyboard is too slik." mit dem Aspekt *keyboard*.

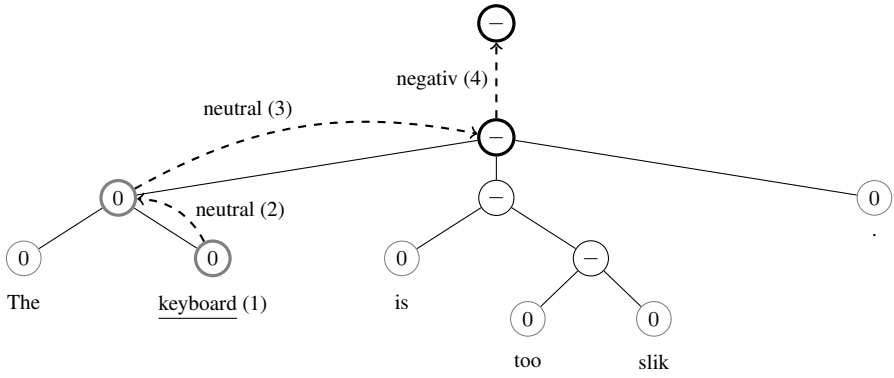


Abbildung 1: Beispiel des Sentiment-Tree-Algorithmus für den Satz „The keyboard is too slik."

Zunächst wird der Sentiment Tree für den Satz erzeugt (siehe Abbildung 1). Daraufhin wird der Knoten, der das Hauptwort des Aspekts enthält, lokalisiert. Ausgehend von diesem Knoten wird der Baum bis zum Wurzelknoten durchsucht. Wird ein Knoten mit einer nicht-neutralen Polarität (also *negative* oder *positive*) gefunden, wird diese als Polarität für den entsprechenden Aspekt angenommen. Nur wenn der Wurzelknoten erreicht wird, ohne dass eine nicht-neutrale Polarität erkannt wurde, wird der Aspekt mit *neutral* bewertet.

Im Beispiel wird ausgehend vom *Keyboard-Knoten* (1) nach einer Polarität gesucht (2). Da dessen Elternknoten keine nicht-neutrale Polarität enthalten, wird der Wurzelknoten untersucht (3). Dieser besitzt eine negative Polarität (4), wodurch der Aspekt *keyboard* insgesamt eine negative Polarität bekommt.

In Tabelle 5 sind die Ergebnisse des Algorithmus, die mit den Trainingsdaten erreicht wurden, nach Polarität getrennt aufgeschlüsselt.

Anhand der Ergebnisse ist erkennbar, dass die F-Measure für positive Polaritäten leicht besser ist als für negative Polaritäten, wobei der Unterschied bei den Restaurant-Daten größer ausfällt. Allgemein sind die Ergebnisse aber relativ gut. Besonders problematisch gestaltet sich zur Zeit die Erkennung von neutralen Aspekten, da der Sentiment Tree fast immer eine Polarität ungleich *neutral* liefert. Um trotzdem eine bessere Erkennung von neutralen Polaritäten zu gewährleisten, wurden verschiedene Modifikationen getestet. Zum Beispiel sollten nur bestimmte Knotentypen zur Bestimmung der Polaritäten heran-

Tabelle 5: Ergebnisse des Sentiment Tree Algorithmus

Datensatz (Domain)	Precision	Recall	F-measure
Laptop			
- positive	0,66	0,67	0,67
- negative	0,51	0,73	0,60
- neutral	0,33	0,08	0,12
Restaurant			
- positive	0,78	0,70	0,74
- negative	0,38	0,74	0,50
- neutral	0,40	0,11	0,17

gezogen werden. Ebenso sollte der Algorithmus nur eine festgelegte Anzahl an Elternknoten betrachten, um sicherzustellen, dass die Polarität mit dem Aspekt zusammenhängt und nicht durch einen unabhängigen Satzteil hervorgerufen wird. Dies brachte jedoch nicht die gewünschte Verbesserung.

4.3 Aspect Category Extraction

Ab Aufgabenteil 3 wird nur noch mit einem Thema gearbeitet – *restaurant*. Hierfür sind folgende Kategorien vorgegeben, in die die Aspekte sortiert werden sollen: *food*, *service*, *price*, *ambience* und die neutrale Kategorie *anecdotes/miscellaneous*. Unser Ansatz lässt sich in zwei Fälle unterteilen, je nach Vorkommnis der Aspekte.

1. Satz enthält Aspektterme:

   Mit Hilfe des Beispielsatzes „Even though its good seafood, the prices are too high." soll gezeigt werden, wie der Algorithmus funktioniert. Die Aspekte des Satzes sind *seafood* und *price*.

   - Entspricht ein Aspekt dem Namen einer Kategorie, wird diese direkt übernommen. Im Beispielsatz wäre dies *price*.
   - Als Sonderfall wird überprüft, ob der Aspekt ein Gericht bezeichnet. Dies führt zur Zuordnung der Kategorie *food*.
   - Nicht zugeordnete Aspekte werden auf ihre Ähnlichkeit bezüglich der Kategorien überprüft. Ist die *path length* (Path based similarity algorithm) geringer als 0,4 (hat sich auf den Trainingsdaten als bester Wert erwiesen), wird der Aspekt der Kategorie zugeordnet. In diesem Beispiel wäre die Kategorie *food*.

2. Satz enthält keine Aspektterme:

   Schritt 3 des ersten Falles wird für alle Nomen des Satzes (und nicht für alle Aspekte) durchgeführt. Der Schwellwert wird hier jedoch auf 0,19 gesenkt, um die Anzahl der erkannten Kategorien zu begrenzen. Falls kein ähnlichkeitswert den Schwellwert unterschreitet, wird die Kategorie *anecdotes/miscellaneous* gewählt.

Tabelle 6 stellt unsere Ergebnisse dar. Das Ergebnis ist für Basisfälle gut. Sobald ein Satz Mischungen von Kategorien enthält oder weder Aspekte noch Nomen besitzt, besteht an den Ergebnissen jedoch Verbesserungsbedarf.

Tabelle 6: Ergebnisse des Kategorieerkennungs-Algorithmus

Datensatz (Domain)	Precision	Recall	F-measure
Restaurant	0,56	0,44	0,49

4.4 Aspect Category Polarity

Dieser Abschnitt beschreibt die letzte Teilaufgabe, die die Polarität einer Aspekt-Kategorie in einem Satz bestimmt. Den vorgegebenen Kategorien (vgl. Abschnitt 4.3) werden die Wert *positive*, *negative*, *neutral* oder *conflict* zugeordnet. Die Polarität einer Kategorie wird über die Polaritäten der einzelnen Aspekte dieser Kategorie bestimmt. Da in diesem Aufgabenteil nur die Polaritäten von Kategorien bestimmt werden, müssen die Ergebnisse der zweiten und dritten Teilaufgabe vorliegen. Auf Basis der Traningsdaten wurden Regeln definiert, die dazu dienen, die Polaritäten für die Kategorien zu bestimmen. Im Folgenden werden die einzelnen Regeln kurz dargestellt.

Regel 1. Falls die Polaritäten der Aspekte innerhalb einer Kategorie gleich sind, dann ist dies auch die Polarität der Kategorie.

Beispiel 6 *„Prices are higher to dine in and their chicken tikka marsala is quite good.“ Bei diesem Beispiel wurden die folgenden Aspekte „Prices“ als* negative *und „chicken tikka marsala“ als* positive *gefunden. Die Aspekte wurden zwei verschiedenen Kategorien zugeordnet und deswegen sind die Bewertungen unabhängig voneinander, d.h. die Kategorie* food *wurde als* positive *gekennzeichnet und die Kategorie* price *wurde als* negative *markiert.*

Regel 2. Ist einer der Aspekte einer Kategorie neutral, dann hat dies keinen Einfluss auf die Polarität der Kategorie, solange mindestens eine andere Polarität vorhanden ist. Die Polaritäten der anderen Aspekte bestimmen die Polarität einer Kategorie.

Beispiel 7 *„Our agreed favorite is the orrechiete with sausage and chicken (usually the waiters are kind enough to split the dish in half so you get to sample both meats).“ Dieser Satz enthält die Aspekte „orrechiete with sausage and chicken“ und „waiters“ mit der Polarität* positive *sowie „meats“ und „dish“ mit der Polarität* neutral*. Nur der Aspekt „waiters“ gehört zur Kategorie* service*, dadurch wird die Polarität von* service positive*. Die anderen drei Aspekte haben die Polaritäten* neutral *und* positive*, dadurch wird die Kategorie* food *als* positive *deklariert.*

Regel 3. Falls die Polaritäten der jeweiligen Aspekte einer Kategorie sowohl *positive* als auch *negative* sind, dann wird die Polarität der Kategorie als *conflict* markiert.

Beispiel 8 *Zur Illustration nehmen wir den Satz: „The sweet lassi was excellent as was the lamb chettinad and the garlic naan but the rasamalai was forgettable.“ Hier gibt es vier Aspekte: „sweet lassi“ mit der Polarität* positive*, „lamb chettinad“ mit der Polarität* positive*, „garlic naan“ mit der Polarität* positive *und „rasamalai“ mit der Polarität* negative*. Dies ergibt die Kategorie* food *mit der Polarität* conflict*.*

Regel 4. Falls die Kategorie als *anecdotes/miscellaneous* identifiziert wurde und es keinen Aspekt gibt, dann wird der Stanford Sentiment Tree genutzt, um die Polarität des ganzen Satzes zu bestimmen. Dies ergibt dann die Polarität für die Kategorie *anecdotes/miscellaneous*.

Beispiel 9 *Der Satz „A guaranteed delight!“ hat keine Aspekte. Da dieser Satz keine Aspekte hat, wird der Stanford Sentiment Tree verwendet, um eine Polariät zu bestimmen. In diesem Fall ist diese* positive*.*

Wir haben unseren Ansatz auf den Trainingsdaten angewendet. Das Ergebnis ist in Tabelle 7 dargestellt. Unser Algorithmus erzielte eine F-measure von 0,85 für die Polarität *positive*, was angesichts der Neuartigkeit unseres Ansatzes ein vielversprechendes Ergebnis ist. Bei der Polarität *conflict* haben wir eine *f-Measure* von 0,16.

Tabelle 7: Ergebnisse des Aspect Category Polarity Algorithmus

Datensatz (Domain)	Precision	Recall	F-measure
Restaurant			
- positive	0,89	0,80	0,85
- negative	0,55	0,76	0,64
- neutral	0,43	0,41	0,42
- conflict	0,19	0,14	0,16

5 Evaluation

Mit der Einteilung in vier Teilbereiche durch die *SemEval-2014* wird etwas von der konventionellen Sentiment Analysis abgewichen. Für das Auffinden der Polaritäten konnten wir zum Beispiel davon ausgehen, dass die annotierten Aspekt-Terme und Aspekt-Kategorien richtig sind, um die einzelnen Algorithmen unabhängig testen zu können und nicht Fehler aus den Extraktions-Algorithmen weiter zu propagieren. Bei einer vollständigen Analyse, also einer, die Fehler weiter propagiert, wäre dies aber nicht zwingend der Fall, sodass zum Beispiel zu falschen Aspekten im folgenden Schritt eine Polarität ermittelt wird.

Deshalb haben wir auch einmal eine solche Analyse auf den Test-Daten der *SemEval-2014* durchgeführt und diese dann mit den Ergebnissen verglichen, die jeweils mit den sogenannten *Gold-Daten*[6] ermittelt wurden. Das Ergebnis für diese Analyse, dargestellt in

[6] „richtig“ annotierte Daten

Tabelle 8: Ergebnisse einer vollständigen Sentiment Analysis der Aspekte

Datensatz (Domain)	Precision	Recall	F-measure
Restaurant			
- positive	0,83	0,31	0,45
- negative	0,46	0,37	0,41
- neutral	0,31	0,03	0,05
- conflict	-	-	-

Tabelle 8 für die Aspekte und Tabelle 9 für die Kategorien, verschlechtert sich wie erwartet deutlich. Auffällig ist jedoch, dass vor allem bei den Kategorien eine sehr hohe *precision* vorliegt. Mit dem vergleichsweise hohen *recall* in den positiven Kategorien fällt hier die F-measure im Verhältnis deutlich geringer ab als erwartet.

Tabelle 9: Ergebnisse einer vollständigen Sentiment Analysis der Kategorien

Datensatz (Domain)	Precision	Recall	F-measure
Restaurant			
- positive	0,73	0,57	0,64
- negative	-	-	-
- neutral	0,08	0,01	0,02
- conflict	-	-	-

6 Zusammenfassung und zukünftige Arbeit

In dieser Arbeit wurde ein konkreter Ansatz zur Sentiment Analysis auf Basis von Aspekten entwickelt. Es wird die Polarität zu den einzelnen Aspekten und Kategorien unter Nutzung der Bibliothken Stanford CoreNLP, RiTa.WordNet, SentiWordNet und einer eigens entwickelten Datenbank, welche die Namen zu verschiedenen Gerichten enthält, bezogen auf einzelne Sätze ermittelt. Diese Bibliothken bringen bereits eigene Methoden mit, um Sätze oder Worte zu klassifizieren, oder um die entsprechenden Polaritäten zu ermitteln.

Durch die Verwendung bereits vorgegebener Methoden konnte allerdings wenig Einfluss auf das Ergebnis genommen werden. An dieser Stelle könnten Bibliotheken wie NLTK helfen, das Ergebnis nochmals zu verbessern. Sie bieten die Möglichkeit verschiedene Klassifizierer zu trainieren und damit mehr Einfluss zu nehmen.

Generell kann man in Zukunft davon ausgehen, dass die Bewertungsmöglichkeiten in Richtung verbesserte Erkennung von einzelnen Aspekttermen und von unklaren Sätzen fortschreitet.

Literatur

[BSdJ13] Ayoub Bagheri, Mohamad Saraee und Franciska de Jong. An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews. In *NLDB*, Seiten 140–151, 2013.

[dMM08] Marie-Catherine de Marneffe und Christopher D. Manning. Stanford Typed Dependency Manual, 2008. `http://nlp.stanford.edu/software/dependencies_manual.pdf`.

[GGT13] Marco Guerini, Lorenzo Gatti und Marco Turchi. Sentiment Analysis: How to Derive Prior Polarities from SentiWordNet. In *EMNLP*, Seiten 1259–1269, 2013.

[HL04] Minqing Hu und Bing Liu. Mining and summarizing customer reviews. In *KDD*, Seiten 168–177, 2004.

[KH06] Soo-Min Kim und Eduard H. Hovy. Automatic Identification of Pro and Con Reasons in Online Reviews. In *ACL*, 2006.

[KK07] Nobuhiro Kaji und Masaru Kitsuregawa. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In *EMNLP-CoNLL*, Seiten 1075–1083, 2007.

[KM03] Dan Klein und Christopher D. Manning. Accurate Unlexicalized Parsing. In *ACL*, Seiten 423–430, 2003.

[Liu10] Bing Liu. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*, 2010.

[Luc12] Laurent Luce. Twitter sentiment analysis using Python and NLTK, 2012. `http://www.laurentluce.com/posts/twitter-sentiment-analysis-using-python-and-nltk`.

[ME10] Samaneh Moghaddam und Martin Ester. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In *CIKM*, Seiten 1825–1828, 2010.

[Mil95] George A. Miller. WordNet: A Lexical Database for English. *Commun. ACM*, 38(11):39–41, 1995.

[PLV02] Bo Pang, Lillian Lee und Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. *CoRR*, cs.CL/0205070, 2002.

[SBMN13] Richard Socher, John Bauer, Christopher D. Manning und Andrew Y. Ng. Parsing with Compositional Vector Grammars. In *ACL (1)*, Seiten 455–465, 2013.

[SPW+13] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng und Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seiten 1631–1642. Association for Computational Linguistics, 2013.

[TP12] Mikalai Tsytsarau und Themis Palpanas. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, 24(3):478–514, 2012.

[Vec10] Olga Vechtomova. Facet-based opinion retrieval from blogs. *Inf. Process. Manage.*, 46(1):71–88, 2010.

[ZY08] Yu Zhang und Dit-Yan Yeung. Semi-Supervised Discriminant Analysis using robust path-based similarity. In *CVPR*, 2008.

Campus-Navigations-Applikation für Studierende mit Beeinträchtigungen

Robert Fruth, Henry Moews, David Silva Ribeaux,
Tom Tschernack, Norbert Wall, Patrick Wolfien

Universität Potsdam
Institut für Informatik und Computational Science
August-Bebel-Straße 89, 14482 Potsdam

{fruth, kraeplin, silvarib, tscherna, nowall, wolfien}@uni-potsdam.de

Abstract: Die Bewältigung des Studiums ist für Studierende mit Beeinträchtigungen noch immer mit zum Teil großen Hürden verbunden. Ein wesentlicher Faktor dabei ist die Mobilität auf dem Campus. Daher wurde in einem studentischen Projekt an der Universität Potsdam eine Applikation entwickelt, um die Orientierung auf dem Gelände der Hochschule zu erleichtern. Der vorliegende Beitrag dokumentiert die Hintergründe, das Konzept und die Umsetzung dieser Applikation.

1 Einleitung

Alle Menschen haben das Recht auf Bildung und Chancengleichheit. So sollen auch Studierende mit einer Beeinträchtigung ihrem Recht nachgehen können. Der Alltag ist für sie allerdings voller Barrieren und anspruchsvoller als für andere Studierende. Die Anforderungen, die daraus resultieren, werden unter dem Begriff Barrierefreiheit definiert. Dieser besagt im Allgemeinen, dass Barrieren entfernt werden müssen, um eine gesellschaftliche und wirtschaftliche Teilhabe bzw. Teilnahme zu ermöglichen. Die erste Hürde stellt schon der Weg zur Universität dar. Neben den sichtbaren Behinderungen, wie zum Beispiel der Seh- bzw. Gehbeeinträchtigung, gibt es auch solche, die kognitiver Natur sind. Die betroffenen Personen leiden zum Beispiel unter spezifischen Ängsten, wie der Platzangst. Diese Behinderungen werden oft außer Acht gelassen, stellen für die Betroffenen jedoch eine starke Barriere dar. Personen mit einer Sehbehinderung brauchen markante Orientierungshilfen wie die Beschaffenheit des Bodens oder die Abfolge von Türen, um den richtigen Raum zu finden. Diese Elemente müssen sich zuvor eingeprägt werden, was eine große Gedächtnisleistung fordert. Auch Studierende mit einer Gehbeeinträchtigung haben mit einer Reihe von Barrieren zu kämpfen. Unter anderem sind schwere Türen, fehlende Fahrstühle bzw. Rampen oft große Hindernisse. Für die letzteren beiden Personengruppen sind vor allem wechselnde Begebenheiten problematisch. Für Menschen mit Sehbeeinträchtigung können Umbaumaßnahmen Orientierungslosigkeit bedeuten. Studierende mit Gehbeeinträchtigung werden mit defekten Fahrstühlen und Türöffnern meist vor eine

unüberwindbare Barriere gestellt. Bei kognitiven Behinderungen hingegen brauchen die Studierenden oftmals alternative Routen über den Campus, um ihren Ängsten bestmöglich aus dem Weg gehen zu können.

An der Universität Potsdam wurde dazu eine mobile Android-Applikation entwickelt, die Studierende mit Beeinträchtigungen den Alltag erleichtern und barrierefreier machen soll. Den Anstoß zur Entwicklung dieser Applikation gab der Kurs „Barrierefreiheit in der Lehre“, welcher im Wintersemester 2013/2014 vom Lehrstuhl für „Komplexe Multimediale Anwendungsarchitekturen“ der Universität Potsdam initiiert wurde. Die Zugänglichkeit der Wege und Gebäude auf dem Campus wird durch die Applikation abgebildet. Studierende können mithilfe der Applikation, den für sie optimalen Weg bestimmen und sich anhand einer Route von der Anwendung navigieren lassen. Der optimale Weg setzt sich aus möglichst wenig Hindernissen, sowie einer akzeptablen Weglänge zusammen. Die Applikation ist in der Lage sich an die Art der Beeinträchtigung anzupassen. Zum Beispiel gibt es für Studierende mit Sehbeeinträchtigung die Möglichkeit die Applikation inklusive der Navigation per Sprachsteuerung zu benutzen. Studierenden mit einer Rot-Grün-Blindheit ist es möglich, das Aussehen der Applikation ihren Bedürfnissen anzupassen.

2 Grundlagen

Das Ziel dieses Projektes ist es, Studenten mit Behinderungen das Studium / universitäre Leben mittels Navigationstechnologien zu erleichtern. Dabei soll mithilfe eines Smartphones die Navigation auf dem Campus unter der Berücksichtigung der jeweiligen Einschränkungen ermöglicht werden.

2.1 Barrierefreie Smartphone-Bedienung

Hierzu muss als Erstes auf mögliche Einschränkungen bei der Nutzung eines Smartphones eingegangen werden. Menschen mit physischen oder motorischen Einschränkungen haben mitunter Schwierigkeiten Gesten auszuführen oder eine punktgenaue Auswahl zu tätigen. Diesen Personen kann durch eine klare Anwendungsstruktur geholfen werden. Ebenso kann durch den Einsatz einer Sprachsteuerung das Problem weitgehend umgangen werden, die von Android[1] und iOS[2] ermöglicht wird. Gehörlose und schwerhörige Menschen können durch zusätzliche Beschriftungen in Applikationen sowie durch Untertitel bei visuellen Inhalten unterstützt werden. Weiterhin kann der Einsatz des Vibrationsalarms hilfreich sein. Der Fokus dieses Prototyps liegt jedoch auf blinden und sehbehinderten Personen, wobei Menschen mit Gehbehinderungen sowie Rollstuhlfahrer nicht außer Acht gelassen werden. Sowohl Android als auch iOS bieten mit der Funktion Talkback[3] bzw.

[1] Sprachsteuerung unter Android, `https://support.google.com/websearch/answer/2940021?hl=de`, abgerufen am 03.04.2014

[2] Siri, `https://www.apple.com/de/ios/siri/`, abgerufen am 03.04.2014

[3] TalkBack, `https://play.google.com/store/apps/details?id=com.google.android.marvin.talkback`, abgerufen am 03.04.2014

VoiceOver[4] einen Mechanismus an, welcher den Bildschirminhalt akustisch wiedergeben kann. Zusätzlich zu den akustischen Ausgaben werden spezielle Gesten verwendet, die die Bedienung erleichtern. Hierzu zählen das Erkunden des Bildschirms mittels einfacher Berührung und der gleichzeitigen akustischen Ausgabe des Inhalts, das Durchblättern von Menüpunkten, Listen oder ähnlichem mittels einer Geste, sowie das Bestätigen einer Auswahl durch eine doppelte Berührung. Ein weiterer wichtiger Punkt ist die Eingabe von Inhalten. Es ist sowohl möglich über die normale Tastatur als auch über Diktierfunktionen Text einzugeben.

2.2 Campusspezifische Navigation

Eine Internetrecherche nach „NavigationsApp“ liefert für verschiedene Anwendungsfälle beispielsweise Auto-, Fahrrad- oder Fußgängernavigation umfassende Ergebnisse. Was jedoch fehlt, ist eine campusspezifische und auf persönliche Einschränkungen angepasste Navigation. Mithilfe dieser soll es möglich sein, sich auf dem Campus sowie in einzelnen Gebäuden orientieren zu können. Davon würden sowohl Blinde und Sehbehinderte als auch Menschen, die sich das erste Mal auf dem Campus befinden, profitieren. Auf drei Applikationen für sehbehinderte Personen wird näher eingegangen.

BlindSquare[5] ist eine Applikation, die Blinden und Menschen mit Sehbehinderungen helfen soll, den Alltag zu bewältigen. Dabei setzt die Applikation verstärkt auf so genannte Points of Interest (PoI). Mithilfe der GPS- und der Kompassfunktion des Smartphones können damit Informationen über die Umgebung akustisch (und visuell) ausgegeben werden: *„Das Ziel befindet sich neunzig Meter auf zwei Uhr“*.

Ariadne GPS[6] ist eine weitere Applikation zum Erkunden der näheren Umgebung. Es ist, wie auch bei BlindSquare, möglich PoI anzusteuern. Ariadne GPS bietet zusätzlich eine Funktionalität an, die kontinuierlich die aktuelle Position (z.B , die Straßen- und Hausnummer) akustisch ausgibt. Dies geschieht auf Grundlage von OpenStreetMap-Daten.

Eine weitere verfügbare Applikation ist Myway Classic[7]. Mit dieser Applikation ist es möglich Routen durch Punkte aufzuzeichnen und diese dann später mit Hilfe von Kompass und GPS abzulaufen. Je mehr Punkte in Myway Classic eingetragen werden, desto genauer kann die Navigation erfolgen.

Eine „echte“ von Start- zu Zielpunkt Navigation ist in keiner der Applikationen vorhanden[8]. Es kann zumeist nur ausgegeben werden, in welcher Richtung das Ziel liegt.

[4] VoiceOver, `https://www.apple.com/de/accessibility/ios/#vision`, abgerufen am 04.03.2014

[5] BlindSquare, `http://blindsquare.com`, abgerufen am 29.03.2014

[6] Ariadne GPS, `http://www.ariadnegps.eu`, abgerufen am 29.03.2014

[7] MyWay Classic,`https://itunes.apple.com/ch/app/myway-classic/id494282724?mt=8`, abgerufen am 29.03.2014

[8] Ran et al. nutzen zusätzliche Hardware zur Navigation [Ran04].

2.3 Kartengrundlage und Barrierefreiheitsinformationen

Um eine Route zwischen zwei Punkten berechnen zu können, wird eine Kartengrundlage benötigt. Aufgrund der freien Verwendbarkeit von OpenStreetMap(OSM)-Kartenmaterial bietet sich dessen Verwendung an. Es umfasst eine detaillierte Abbildung der Außenwelt Potsdams einschließlich der drei Universitätscampusse. Die Innenansicht von Gebäuden muss hinzugefügt werden, wobei es keinen OpenStreetMap-Standard für diese Art der Modellierung gibt. Es existiert jedoch ein Vorschlag[9], der die dreidimensionalen Gebäudedaten einschließlich der Etagen, Räume, Türen, Treppenhäuser und Fahrstühle abdeckt. Dieser Vorschlag [Goe12] wurde inzwischen von der OSM-Gemeinschaft für etwa 20 Gebäude weltweit adaptiert[10].

Neben den allgemeinen Kartendaten werden für eine benutzerspezifische Routenführung auch Barrierefreiheitsinformationen benötigt. Für die Gebäude der Universität Potsdam wurden diese am Lehrstuhl für Technologie und berufliche Orientierung mit Fokus auf Rollstuhlfahrende erstellt[11] und in eine MySQL-Datenbank überführt [Bar14]. Die Aufstellung dieser Daten orientiert sich an der DIN 18040-1 [DIN10]. Anhand der dort aufgeführten Kriterien wurden die vorhandenen Gebäude untersucht und alle Gebäudeeinheiten auf ihre Zugänglichkeit hin klassifiziert. Weiterhin wurden Konzepte der taktilen Umsetzung von OSM-Karten und Routen auf einer Stiftplatte für blinde Anwender von [Iva14] diskutiert.

3 Konzeption

Ziel des Projektes ist die Entwicklung einer Android-Applikation, welche eine Navigation auf dem Campus ermöglicht und Lagepläne der Campusse der Universität Potsdam sowie der einzelnen Gebäude bereitstellt. Die Applikation geht dabei speziell auf die Bedürfnisse von seh- und gehbehinderten Menschen bei der Routenplanung ein (siehe Abschnitt 2). Die Wahl für die prototypische Umsetzung fiel auf die Android-Plattform, da dort der App-Store leicht zugänglich ist. Weiterhin sind Testgeräte günstiger als vergleichbare Apple-Produkte. Eine Portierung auf die iOS-Plattform jedoch ist geplant. Die Umsetzung des Projekts in Form einer Client-Server-Architektur wird als notwendig angesehen, da die Routenberechnung auch auf relativ kleinen Flächen bereits sehr aufwändig ist und viele Ressourcen verbraucht. Deshalb ist es nicht möglich, die Berechnung allein auf den mobilen Endgeräten auszuführen. Auch die Kartendaten sollten über einen Zugriff vom Server geladen werden, da auf diesem Wege auch aktuelle Veränderungen auf dem Campus mit in die Berechnung und in die Anzeige einbezogen werden können (z.B. aufgrund von durch Bauarbeiten blockierte Wege).

[9] IndoorOSM, `http://wiki.openstreetmap.org/wiki/IndoorOSM`, abgerufen am 29.03.2014

[10] IndoorMapping, `http://wiki.openstreetmap.org/wiki/Indoor_Mapping#Places`, abgerufen am 29.03.2014

[11] Projekt Barrierefreiheit, `http://www.uni-potsdam.de/projekt-barrierefreiheit`, abgerufen am 29.03.2014

3.1 Karten

Die Kartendaten bestehen wie beschrieben aus drei Teilen: Karten für die Außenwelt, Karten für die Innenwelt von Universtitätsgebäuden sowie Barrierefreiheitsinformationen. Während Außenweltkarten bereits zur Verfügung standen und unverändert von OpenStreetMap verwendet werden konnten, mussten jene für die Gebäude erst erstellt werden. Bei der Speicherung ergibt sich das Problem, dass es der Arbeitsgruppe nicht gestattet ist, die Gebäudedaten öffentlich zur Verfügung zu stellen, was das Speichern auf den offiziellen OpenStreetMap-Servern verbietet. Stattdessen musste ein separater Karten-Server installiert werden. Dieser hat zwar den Nachteil, dass eigentlich vorhandene Funktionalitäten erneut implementiert werden mussten, führte aber dazu, dass genau die gewünschten Ausgaben erreicht werden konnten. Beispielsweise ist die Smartphone-Anwendung darauf angewiesen Kartenmaterial (also Bilder) verbal wiedergeben zu können. Hierzu bot sich das SVG-Grafikformat an, das mit beliebigen relevanten Informationen angereichert werden kann. Neben den Grafiken für die Smartphone-Anwendung muss das OSM-XML-Format für die Berechnung der Routen angeboten werden.

Die Barrierefreiheitsinformationen liegen in einer separaten Datenbank vor und werden gesondert gewartet. Bei der Erstellung orientiert man sich an den offiziellen Raumbezeichnungen. Falls diese fehlen oder aufgrund veränderter baulicher Eigenschaften veraltet sind, werden eigene Bezeichnungen gewählt. Um die Gebäudekarten und Barrierefreiheitsinformationen zweifelsfrei zusammenführen zu können, muss eine gemeinsame Eigenschaft in beiden Datensammlungen vorhanden sein. Es wird darauf verzichtet, die bereits vorhandenen frei gewählten Bezeichnungen manuell zu übernehmen. Stattdessen wird ein Ansatz der Ortung über Geokoordinaten, wie es in OpenStreetMap üblich ist, genutzt. Infolgedessen wurde eine Möglichkeit geschaffen, die Geokoordinaten für alle fraglichen Räume und andere Elemente wie Rampen und Fahrstühle zu berechnen und in die Datenbank einzutragen.

3.2 Benutzerschnittstelle

Um eine Navigation für die Benutzer zu ermöglichen, wird eine Route ermittelt, die durch Punkte und Teilwege, die auf dem Weg liegen, sowie deren Eigenschaften beschrieben ist. Eine solche Route sollte einerseits, entsprechend der speziellen Bedürfnisse des jeweiligen Nutzers, möglichst wenig Hindernisse enthalten und gleichzeitig möglichst kurz sein. Zur Berechnung einer Route werden ein Start- und ein Zielpunkt als Eingabe und für die Indoor-Navigation zudem gegebenenfalls die Information über die jeweiligen Etagen und den Campus benötigt. Die Berechnung der Route erfolgt auf Basis von OSM-Außen- und -Innenweltkarten und den angegebenen Ansprüchen des Nutzers über die gewünschte Route.

3.3 Applikation

Die Applikation bietet dem Nutzer Karten der drei verschiedenen Campusse der Universität Potsdam. Die Gebäude auf einem Campus werden dabei entsprechend des Grades ihrer Barrierefreiheit in einem Ampel-Farbschema angezeigt. Weiterhin sind Indoor-Karten von Gebäuden verfügbar. Diese bieten Informationen über die Barrierefreiheit einzelner Räume, beispielsweise die Türbreite, über Gänge die eventuell durch Treppen oder Schwellen unzugänglich sind sowie über die Erreichbarkeit einzelner Etagen mithilfe von Fahrstühlen oder Treppen. Außerdem werden behindertengerechte Toiletten speziell gekennzeichnet. Es wird weiterhin ein Benutzerprofil angeboten, in dem die Art der Einschränkung festgehalten wird. Weiterhin können hier Fahrstühle oder Treppen von der Routenplanung ausgeschlossen werden und es kann definiert werden, in welchem Farbschema die Karten angezeigt werden. Dies kann beispielsweise für Benutzer mit Farbsehschwäche hilfreich sein. Das Profil wird dezentral auf dem Smartphone des Nutzers im XML-Format gespeichert. Die im Profil gespeicherten Einschränkungen werden dann bei der Navigation berücksichtigt. Der Nutzer kann sich mithilfe der Applikation über den Campus navigieren lassen. Dafür muss er Start- und Endpunkt der Route festlegen. Als Standorte sind sowohl Räume innerhalb von Gebäuden als auch Gebäudeeingänge möglich. Die Applikation bietet außerdem einen Sehbehinderten-Modus. Hier wird für die Eingabe von Routen und die Navigation durch die Menüs auf Android-Talkback und Gesten gesetzt. Es kommen Einfach- sowie Doppelklicks und Wischgesten in verschiedene Richtungen zum Einsatz. Bei der Ausgabe der Route und der Menüs wird mit Audioausgaben gearbeitet.

4 Implementierung des Prototypen

4.1 Architektur

Die umgesetzte Architektur besteht aus den drei Komponenten: Applikation, Routenplanung und Kartendaten. Im Folgenden werden die einzelnen Schnittstellen näher erläutert (siehe Abbildung 1), um die Kommunikation und den Datenaustausch zwischen der Applikation, dem Kartenserver sowie der Routenplanung zu beschreiben. Die Applikation dient lediglich der Darstellung der Lagepläne und Ausgabe der Routen. Um Lagepläne zu erhalten, werden REST-Anfragen[12] an den Server gerichtet. Der Lageplan wird dabei als Vektorgrafik im SVG-Format zurückgegeben.

Eine weitere Schnittstelle bildet die Kommunikation mit der Routenplanungskomponente. Sie liefert die Routendaten beziehungsweise Wegbeschreibungen an die Applikation zurück. Hierzu werden von der Applikation Positionsdaten (z.B. GPS-Koordinaten) und Inhalte des Benutzerprofils als REST-Anfragen übertragen. Die Inhalte des Benutzerprofils werden dabei zur Anpassung der Route an die Bedürfnisse des Benutzer genutzt. Ebenfalls besteht eine Verbindung zwischen den Lageplänen und der Routenplanung. Hier werden

[12] Representational State Transfer, ein Programmierparadigma für Webanwendungen

die Karten als OSM-Dateien für die Routenplanung zur Verfügung gestellt. Zur Kommunikation werden auch hier REST-Anfragen genutzt. In den nachfolgenden Abschnitten werden nun die einzelnen Komponenten und deren Funktionalität näher beschrieben.

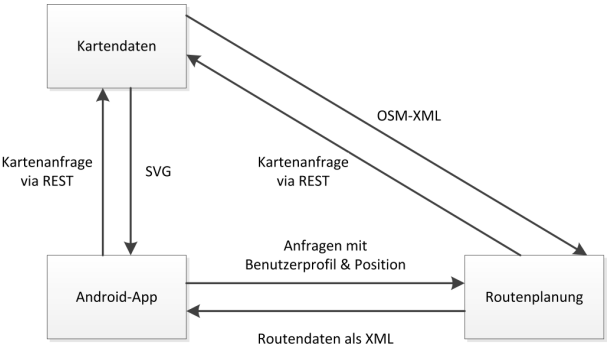


Abbildung 1: Interaktion der Komponenten

4.2 Routen

Die Routing-Anwendung ist in Java umgesetzt. Für die Implementierung der RESTful-Web-Service-Schnittstellen wurde die Bibliothek Jersey 2.0[13] verwendet. Der Webservice läuft damit auf einem Apache Tomcat 7.0[14] oder GlassFish Server 4.0[15].

Wie beschrieben, ist die Aufgabe dieser Komponente in der Gesamtarchitektur die Ermittlung einer möglichst idealen Route für den Benutzer und die Weitergabe der Routeninformation an die Applikation. Damit die Ausgabe möglichst universell auch für andere Tools außerhalb dieses Projekts nutzbar ist, werden neben den Knoten-IDs, die direkt den OSM-Daten entnommen werden, auch die Welt-Koordinaten mit ausgegeben.

Für die Routenberechnung werden die Start- und Zielkoordinaten in Längen- und Breitengrad und – für die Routenplanung in Gebäuden über mehrere Ebenen – zusätzlich die Etageninformationen angegeben. Die Anforderung an die Route werden durch eine Profil-ID bestimmt. Bei Angabe des Zugriffsorts der Kartendaten besteht auch die Möglichkeit, OSM-Kartenmaterial aus anderen Quellen zu nutzen. Da bei bereits vorhandenen Algorithmen bzw. Bibliotheken zur Routenberechnung[16] auf Basis von OSM-Daten häufig die Möglichkeit zur Festlegung von bestimmten Eigenschaften der Wunschroute fehlt, wird im Rahmen dieser Arbeit ein eigener Algorithmus auf der Basis von Dijkstra genutzt.

Die Anforderungen der Nutzer an ihre optimale Route können in einer Datei gespeichert

[13] Jersey - RESTful Web Services in Java. `https://jersey.java.net/`, abgerufen am 08.04.2014.

[14] The Apache Software Foundation: Apache Tomcat. `http://tomcat.apache.org/`, abgerufen am: 15.04.2014

[15] Oracle Corporation: GlassFish - World's first Java EE 7 Application Server. `https://glassfish.java.net/`, abgerufen am: 15.04.2014

[16] Graphhopper, `http://graphhopper.com`, abgerufen am 08.04.2014

werden. Dazu wird ein eigens für dieses Projekt entwickeltes XML-Schema genutzt. In dem Profil werden Eigenschaften möglicher Wege, so wie sie auch für die OSM-Karten definiert sind, aufgelistet und durch Zahlenwerte bewertet. Auf Basis dieser Daten werden Wege, welche solche Eigenschaften enthalten, für die Routenfindung gewichtet.

Nach der Berechnung wird die Route als Antwort der REST-Anfrage an die Applikation übergeben. Zur Repräsentation der Routendaten wird dafür ein selbst entwickeltes XML-Schema genutzt, welches sehr einfach und stark an das OSM-XML-Format angelehnt ist. Der Grund dafür ist, dass vorhandene Ausgabeformate, wie das GPS-Exchange-Format[17] (GPX), häufig überdimensioniert sind und keine direkte Einbindung der OSM-Tags oder OSM-IDs bieten. Bei dem hier genutzten Schema stehen die Knoten- und Weg-IDs als Referenz auf die OSM-Karte im Vordergrund. Über die Route sind die wichtigsten Informationen als Attribute hinterlegt, z.B. die Weglänge, die Knotenpunkte, die auf dem Weg liegen, sowie seine Eigenschaften, die direkt von den Kartendaten übernommen werden. Die Ausgabe der Route erfolgt im XML-Format. Aufgrund des hohen Datenoverheads und dem großen Aufwand beim Parsen von XML-Daten auf dem mobilen Gerät sollte in Zukunft ein alternatives Format, wie JSON genutzt werden.

4.3 Karten

Kartendaten müssen sowohl im OSM-XML- als auch im SVG-Format erstellt werden. Zunächst wurde der Ansatz verfolgt, eine Datenbank als Grundlage für beide Formate zu verwenden. Dies stellte sich jedoch als inpraktikabel heraus, da bestehende Datenbankschemata entweder für grafische Ausgaben optimiert sind (z.B. Schema von osm2pgsql[18]) und keinen OSM-XML-Export erlauben oder einen OSM-XML-Export ermöglichen und keine grafischen Ausgaben unterstützen (Schema von Osmosis[19]). In diesem Projekt wurde es jedoch vermieden zwei Datenbanken aufzusetzen, indem die Generierung der OSM-XML-Dateien dateibasiert gestaltet wurde.

Zur grafischen Generierung wurde Kartograph[20] benutzt. Dieses Programm verfügt über sehr umfangreiche Konfigurationsmöglichkeiten und erlaubt die Ausgabe von annotierten SVG-Grafiken. Über das Setzen von Bedingungen ist es möglich, Außenwelt- und Innenweltkarten getrennt zu erstellen. Alle Etagen eines Hauses werden in einer gemeinsamen Datei gespeichert und können mittels CSS-Eigenschaften ein- und ausgeblendet werden. Für die schnelle Beantwortung von Anfragen, werden die Außenweltkarten der Campusse und Innenweltkarten der einzelnen Gebäude in beiden Formaten vorgeneriert. Abbildung 2 stellt schematisch den Generierungsprozess dar. Karten mit personalisierten Kartengrenzen können unter Inkaufnahme einer erhöhten Antwortzeit angefragt und generiert werden.

Zunächst gestaltete sich die Verbindung von Außenwelt- und Innenweltkarten schwierig, da Hauseingänge auf beiden Karten nicht identisch positioniert sind und somit Sackgassen

[17] GPX: GPS Exchange Format, `http://www.topografix.com/gpx.asp`, abgerufen am 08.04.2014
[18] Osm2pgsql, `http://wiki.openstreetmap.org/wiki/Osm2pgsql`, abgerufen am 29.03.2014
[19] Osmsosis, `http://wiki.openstreetmap.org/wiki/Osmosis`, abgerufen am 29.03.2014
[20] Kartograph, `http://kartograph.org`, abgerufen am 29.03.2014

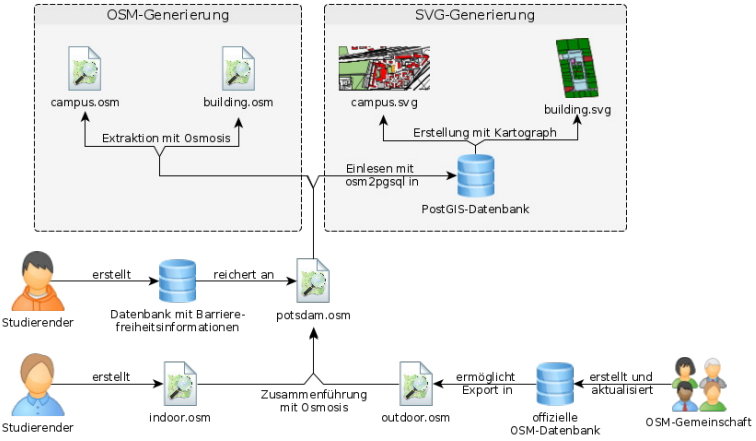


Abbildung 2: Erstellungsprozess der OSM- und SVG-Dateien

entstehen. Zur Verbindung wurden Wege zwischen zwei eigentlich identischen Punkten erstellt, die mit dem Tag highway=virtual gekennzeichnet werden. Die Routenberechnung findet somit eine Verbindung vor, die mit der Länge null betrachtet wird, weshalb diese in der Routenbeschreibung für den Benutzer nicht berücksichtigt wird.

4.4 Applikation

Die Aufgabe der Android-Applikation ist die Kommunikation mit dem Benutzer. Die Architektur ist in fünf Schichten gegliedert (siehe Abbildung 3). Ausgehend vom Core, der die Standardverwaltung der Applikation darstellen soll, gibt es hier die Kartenverwaltung. Die Kartenverwaltung ist an einen Service angebunden, welcher die Lagepläne für die einzelnen Campusse liefert. Um die Karten wie gewünscht je nach Benutzer anzuzeigen, werden Informationen aus dem Benutzerprofil benötigt. Diese sind in der Datenschicht in Form von XML-Dateien zu finden. Außerdem müssen die Kartendaten, welche als SVG zuvor heruntergeladen wurden, für die korrekte Darstellung der Route unter Nutzung eines eigens entwickelten Werkzeuges geparst werden.

Zur Darstellung der Route auf der Karte gibt es in der Logikschicht die Navigationseinheit. Sie ist wie auch die Kartenverwaltung mit einem externen Service verbunden. Neben der Hardwareschicht, die aufgrund der unterschiedlichen Struktur der einzelnen Smartphones nicht weiter beschrieben wird, existiert die Benutzerschnittstelle.

Beim Start der Applikation wird zunächst von einem sehbehinderten Nutzer ausgegangen. Dieser wird durch Audioausgaben begrüßt. Falls der Nutzer seine Sehbehinderung bestätigt, wird dies im Profil vermerkt. Handelt es sich um einen sehenden Nutzer wird mit der normalen Menüführung fortgefahren. Im Anschluss hat der Nutzer die Möglichkeit, sein Profil zu bearbeiten und die Karte einzusehen. Für Nutzer mit Sehbeeinträchtigung entfällt die Darstellung der Karten und es gibt direkt die Möglichkeit, eine Route zu

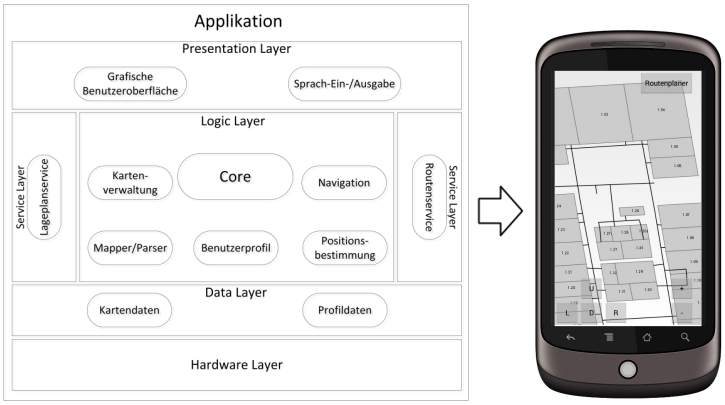


Abbildung 3: Schichtendiagramm und Screenshot einer Indoor-Kartenansicht

planen. Die Ausgabe der Route erfolgt für sehbehinderte Nutzer in Teilstücken und mit Audio-Anweisungen. Dabei wird dem Nutzer jeweils eine Anweisung vorgelesen, beispielsweise: *„Drehe dich um 90 Grad nach rechts und gehe 50 Meter“*. Weiterhin besteht die Möglichkeit, die jeweils vorherige Anweisung erneut auszugeben. Mithilfe einer Wischgeste oder dem von Android-Talkback bekannten Doppelklick auf den Bildschirm kann der Nutzer der Applikation mitteilen, dass er den Wegpunkt erreicht hat. Ihm wird dann die nächste Teilroute beschrieben, solange bis er das Ziel erreicht hat.

Das Layout der einzelnen Menüs zeichnet sich durch eine einfache und übersichtliche Struktur aus. Eine Benutzer-Eingabe ist auf verschiedenen Wegen möglich. Neben Buttons werden insbesondere für sehbehinderte Nutzer Touchgesten angeboten. So werden beispielsweise Wischgesten mit verbalem Feedback zur Auswahl von Gebäuden und Räumen während der Routenplanung eingesetzt. Die Ausgabe der Applikation erfolgt – wie für Sehende üblich – meist in Text- und Grafikform. Für Nutzer mit Sehbeeinträchtigung wird verstärkt auf Audioausgaben gesetzt.

5 Benutzertest

Im Rahmen des Benutzertests soll die qualitative Fragestellung „*Ist die Navigationsapplikation für das vorgesehene Einsatzgebiet und die unterschiedlichen, angedachten Benutzergruppen nutzbar?*“ untersucht werden. Benutzergruppen sind hierbei vorrangig die Personengruppen „uneingeschränkt Sehende“ und „sehbeeinträchtigte Personen“. Personen mit akustischen Einschränkungen werden zum jetzigen Zeitpunkt nicht betrachtet, da eine für sie nutzbare Benutzeroberfläche im engen Verbund zur Nutzeroberfläche für „uneingeschränkt Sehende“ steht; lediglich die Audioausgabe scheint an dem Punkt obsolet, da alle Informationen sowohl auditiv, als auch visuell verfügbar sind. Zur Evaluation des entwickelten Prototyps wurde ein Pilottest mit einer blinden Probandin gemeinsam mit den Entwicklern durchgeführt.

Der Pilottest kann generell in drei Phasen eingeteilt werden: Vorgespräch, Praxistest und Nachgespräch. Dabei werden der Probandin in der Anfangsphase allgemeine Informationen über Ziel, Inhalt und Aufbau des Tests zugänglich gemacht. Weiterhin werden allgemeine Informationen über die Testperson abgefragt. Während der gesamten Zeit der Testphase werden alle Handlungen der Testperson beobachtet und dokumentiert. In der Nachbereitungsphase wird eine abschließende Befragung der Probandin durchgeführt, um Schwächen der Appliktation aufzuzeigen sowie zur Erkennung möglicher weiterer Funktionen.

Vor dem Test wurde neben allgemeinen Thematiken (Vorerfahrung mit dem Smartphone) nach den am stärksten störenden Wegeigenschaften gefragt. Der Probandin zufolge sind dies für sehbeeinträchtigte Studierende vor allem große Plätze ohne markante Wegstrukturen. Als Lösung hält die Probandin einen in die Applikation eingebauten Kompass für sinnvoll, welcher ihr hilft, sich neu zu orientieren. Eine aktive Routenführung, die automatisch den korrekten Abzweig von einem großen Platz erkennt, wurde ebenfalls gewünscht. Es wurde deutlich, dass die Probandin zum Zeitpunkt des Tests aufgrund der mangelnden Verfügbarkeit bzw. Reife anderer Applikationen noch keine Vorerfahrungen mit Routing-Applikationen hatte. Während des Tests wurde festgestellt, dass die Probandin beim ersten Start der Applikation eine Hilfestellung für deren Bedienung durch die durchführenden Tester benötigte. Zurückzuführen ist dies neben zum Teil fehlenden Handlungsanweisungen in der Applikation auf die Bedienung des für die Probandin unbekannten Test-Smartphones. Hinzu kommt, dass einige Sprachausgaben mithilfe der Android Text-To-Speech-Engine falsch ausgesprochen wurden. Dies führte insbesondere bei Schlüsselworten zu Verwirrung bei der Probandin. Letztlich erfolgte auch die Bedienung der Applikation, während Talkback im Hintergrund assistierte, nicht immer, wie von der Probandin erwartet. Im Anschluss an den praktischen Test folgte ein Abschlussgespräch mit der Probandin. In diesem betonte sie einige noch offene Punkte, welche sie sich in einem vollständigen Produkt wünschen würde: Neben einer aktiven Routenführung mit Erkennung von Hindernissen und Neuausrichtung bei Verlust der Orientierung wünscht sich die Probandin Anhaltspunkte über die Beschaffenheit des Untergrunds beziehungsweise dessen Wechsel sowie Aufschluss über Bordsteinkanten, Gehwege, Treppen und Türen. Einige dieser Wünsche sollen in weiteren studentischen Arbeiten umgesetzt werden. Abschließend lässt sich sagen, dass aufgrund des prototypischen Zustands der Applikation noch keine abschließende Antwort auf die eingangs gestellte Frage gegeben werden kann.

6 Diskussion und Fazit

Ein gravierendes Problem stellt die ungenaue GPS-Ortung auf dem Campus Griebnitzsee der Universität Potsdam dar. Diese Daten sind kaum verwertbar und müssen stark normalisiert werden. Weiterhin war die Verknüpfung der Daten von Indoor- und Outdoor-Karten sowie den Barrierefreiheitsinformationen zu einzelnen Gebäuden, Räumen und Fluren problematisch. Zur Lösung dieses Problems wurden alle Informationen innerhalb einer OSM-Datei zusammengefasst. Das Bedienkonzept der Applikation mit Wischgesten und Doppelklicks wurde von der blinden Probandin positiv aufgenommen, war für sie intui-

tiv und entsprach weitestgehend der gewohnten Bedienung eines Android-Smartphones mit Talkback. Es werden sowohl bei den Karten- als auch bei den Routenanfragen REST-Services genutzt, um eine universelle Verwendbarkeit zu gewährleisten. Da die Karten als SVG-Datei zurückgegeben werden und die Routen auf GPS-Koordinaten basieren, ist ihre Nutzung in einem anderen Kontext ebenfalls möglich. Bei der Routenplanung werden die Wünsche und Einschränkungen des Nutzers berücksichtigt. Die Anfragen sind dabei einfach aufgebaut. Es müssen nur die Start- und Endkoordinaten sowie die relevanten Informationen aus dem Nutzerprofil übermittelt werden.

Die beschriebene Konzeption und Implementierung einer Campus-Navigations-Applikation für Studierende mit Beeinträchtigungen wurde positiv evaluiert, und es wurden einige Verbesserungsmöglichkeiten im Detail aufzeigt. Eine mögliche Erweiterung wäre eine umfassende Integration in die Hochschulinfrastruktur. Ein denkbares Beispiel wäre hier die Anbindung an das Vorlesungsverzeichnis, um sich direkt zu Räumen aktuell stattfindender Veranstaltungen navigieren zu lassen. Eine bereits geplante Erweiterung für blinde Nutzer ist die Ausrichtung mittels des im Smartphone vorhandenen Kompasses, um so das Finden des nächsten Checkpoints zu erleichtern. Außerdem wird an einem Ampelsystem für die Barrierefreiheitsinformationen sowohl für Indoor- als auch für Campus-Karten gearbeitet[21]. Hierbei werden dann Gänge, Räume und Gebäude je nach ihrer Barrierefreiheit eingestuft. Um eine Ortung und somit eine adaptive Navigation innerhalb von Gebäuden zu ermöglichen wäre der Einsatz eines WLAN-Triangulationsverfahrens[22] denkbar. Zudem ist die Erweiterung des Benutzerprofils um weitere Wünsche des Nutzers geplant. Ein weiteres denkbares Beispiel wäre es, bei der Routenplanung den Wunsch nach Türen mit Türöffnern anzugeben.

Literatur

[Bar14] Bartsch, Stefan. Wie barrierefrei kann man an der Universität Potsdam studieren? - Auswertung eines langjährigen Projekts der Lehreinheit WAT am Beispiel mobilitätseingeschränkter Studierender. Masterarbeit (in Arbeit), Universität Potsdam, 2014.

[DIN10] DIN 18040-1 / 2010: Barrierefreies Bauen Planungsgrundlagen Teil 1: Öffentlich zugängliche Gebäude, Beuth-Verlag Berlin, Oktober 2010.

[Goe12] Goetz, M. und Zipf, A. Extending OpenStreetMap to Indoor Environments : Bringing Volunteered Geographic Information to the Next Level. 2012.

[Iva14] Ivanchev, Mihail und Zinke, Francis und Lucke, Ulrike. Pre-journey Visualization of Travel Routes for the Blind on Refreshable Interactive Tactile Displays. In *14th International Conference on Computers Helping People with Special Needs*, 2014. (im Druck).

[Ran04] Ran, Lisa und Helal, Sumi und Moore, Steve. Drishti: An Integrated Indoor/Outdoor Blind Navigation System and Service. In *Proceedings of the Second IEEE International Conference on Pervasive Computing and Communications (PerCom'04)*, PERCOM '04, Washington, DC, USA, 2004. IEEE Computer Society.

[21] Ähnlich dem Projekt Barrierefreiheit der Universität Potsdam, siehe `http://www.uni-potsdam.de/projekt-barrierefreiheit`, abgerufen am 29.03.2014

[22] An der Universität Potsdam ist aktuell ein Projekt zu diesem Thema in Bearbeitung.

Comparison of Numerical Methods in Uncertainty Quantification

Ionut-Gabriel Farcas
M.Sc. student in Computational Science and Engineering

Institut für Informatik
Technische Universität München
Boltzmannstraße 3
85748 Garching
ionut.farcas@tum.de

Abstract: This paper presents and compares the results obtained using several methods for Stochastic Computations, used in Uncertainty Quantification. We practically present the methods using a simple ODE model. The focus is on both intrusive and non-intrusive methods, namely the Monte Carlo method, along with methods based on generalized polynomial chaos(gPC) methodology. Moreover, we asses the obtained results by comparing them with the analytical solution of the system.

1 Introduction

The purpose of this paper is to present the results and a comparison between numerical methods used in *Uncertainty Quantification* (UQ). This is motivated by the fact that Uncertainty Quantification is a new and rapidly growing domain, at the edge of applied mathematics and computational science, that receives an increased amount of attention, due to its necessity in analyzing every physical event or engineering system.

Furthermore, the reason for choosing to work with a simple model resides in the fact that our desire is to fully understand the UQ methods in a practical way, that will provide a basis for the further extension of the current work. Our interest is to have a fully functional UQ example that can be used as a one-dimensional stochastic system and that can be easily extended to several dimensions.

As most of the practical problems are of a complex kind, it requires the usage of efficient and fast numerical methods, making uncertainty quantification also very interesting from the high performance computing point of view. An increased number of dimensions introduces an overhead, caused by the large number of points used in the space discretization, for instance, via generalized tensor product, as in [Xiu08] and in this case, the dimensionality of the space has an exponential growth. There are different methods to brake the curse of dimensionality, like sparse grids [BG04] and tensor trains [DKO14]. This idea is a major research topic in UQ and its implementation is very suited for high performance architectures. Moreover, all modern UQ methods can be easily parallelized, making them

suitable for implementation on high performance architectures.

Much related work was done in comparison of UQ methods. In [Xiu08], the author compared the underlying methods using the *Burgers equation* and the *Navier Stokes equations*, in [XLS01] the used equation was again the *Navier Stokes equations* and in [WIC06], the author compared the UQ methods using the *1 d heat equation*. For us, it is also interesting to compare the results obtained with the UQ methods with the ones obtained via the analytical solution.

For the future work, it is in our great interest to extend the current version of the application code and make it usable in the case of more complex systems and at the end, to have a working UQ API.

2 Mathematical description

The used system is a nonhomogenous second order ordinary differential equation (ODE), representing the mathematical description of a second order harmonic oscillator with zero initial conditions. The reason for choosing this model resides in the fact that it has a computable analytical solution, that will be used to compare to the results obtained with the stochastic methods. Moreover, from a stochastic point of view, it can be treated both as a one and multi-dimensional system, property that makes it very suitable for understanding uncertainty quantification in both one and multi dimensions.

2.1 Deterministic model

A second order oscillator (*Figure 1*) is modeled using Newton's second law of motion and we do not insist on it, as it is beyond the purpose of this paper and there is an extensive documentation regarding this topic available in the literature.

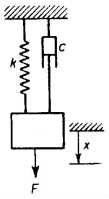


Figure 1: Second order oscillator

The model used in this paper is defined as $x : \mathbb{R}^+ \to \mathbb{R}$, by

$$\begin{cases} x'(t) = y(t) \\ y'(t) + c \cdot y(t) + k \cdot x(t) = f(t) \\ x(0) = y(0) = 0 \end{cases} \tag{1}$$

where

$$f(t) = F \cdot cos(\omega_a \cdot t) \tag{2}$$

and c represents the viscous damping coefficient $[N \cdot s/m]$, k the elasticity constant, F is the amplitude of the oscillations $[N/m^2]$ and ω_a is the angular frequency of the oscillations $[rad/s]$. Throughout this paper, the value of ω_a is set to 1.05 rad/s.

Figure 2 depicts the analytical solution for different values of k. As the focus is to analyze the results of uncertainty in one parameter of the system, we fix c = 5 $[N \cdot s/m]$ and F = 0.1 $[N/m^2]$. Thus, the only varying parameter remains k, the elasticity constant.

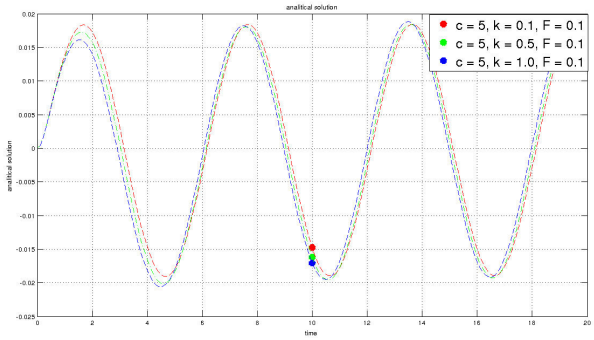


Figure 2: Analytical solution for different values of the parameters

2.2 Stochastic model

In the following, we consider a probabilistic framework and model the parameter k as a 1-variate random variable in a properly defined probability space $(\Omega, \mathfrak{A}, P)$, where Ω is the event space, $\mathfrak{A}$ is a σ-algebra and P is a probability measure, as in [Ras06].

In this paper, we model the uncertainty in (1) as a $Gaussian$ random variable $K : \Omega \to \mathbb{R}$ with the density $\rho : \mathbb{R} \to \mathbb{R}^+$, $(\overline{k},\ \sigma_k)$ the mean and the standard deviation and support $\Gamma = \mathbb{R}$. This choice is motivated by the fact that the majority of real life processes are of $Gaussian$ nature and this setup is the most favorable one to fully understand all the methods and concepts. Also, according to the authors in [Xiu08] and [Ras06], a $Gaussian$ process is fully charaterized by the first two statistical moments, i.e., the $expectation(mean)$ and the $variance$. The probability distribution function of K is:

$$\rho : \mathbb{R} \to \mathbb{R}^+, \rho(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \tag{3}$$

Thus, the stochastic model is defined as follows:

$$\begin{aligned} x : (\mathbb{R}^+ \times (\Omega, \mathfrak{A}, P)) \to (\mathbb{R}, B), x''(t,\omega) + 5 \cdot x'(t,\omega) + K(\omega) \cdot x(t,\omega) = \\ = 0.1 \cdot \cos(1.05 \cdot t), \text{with } x(0,\omega) = x'(0,\omega) = 0 \end{aligned} \tag{4}$$

2.3 Analytical solution

As it was specified in the beginning of this section, the choice of this system is motivated by the possibility to find its analytical solution and to use it for the assesment of the results obtained with the uncertainty quantification methods. With the setup presented in Section 2.1, the analytical solution will be a function that depends on time t and the elasticity constant k.

In order to compute the analytical solution, we employ the *Laplace transform* and obtain:

$$\begin{aligned} x(t,k) = c_1(k) \cdot \cos(\omega_a \cdot t) + c_2(k) \cdot \sin(\omega_a \cdot t) + \\ c_3(k) \cdot e^{\lambda_1(k) \cdot t} + c_4(k) \cdot e^{\lambda_2(k) \cdot t}, \end{aligned} \tag{5}$$

where $c_1(k)$, $c_2(k)$, $c_3(k)$, $c_4(k)$ are constants that depend on k and $\lambda_1(k)$ and $\lambda_2(k)$ are the zeros of $\lambda^2 + c \cdot \lambda + k$, depending on k as well. This second order polynomial is the associated algebraic polynomial of (1).

It is important to stress out that the form of analytical solution is determined by the nature of $\lambda_1(k)$ and $\lambda_2(k)$. With the setup presented in Section 2.1 together with the fact that $k \sim N(0,1)$ (*c.f. Section 2.2*), the zeros of $\lambda_1(k)$ and $\lambda_2(k)$ are $real$ and $distinct$, regardless of the value of k.

3 Uncertainty Quantification methods

In this section, the focus is on the presentation of the theoretical background of the stochastic methods. Moreover, for a better understaing of the methods from a computational point of view, we will present an overview of the numerical error for each method. And at the end, we will present a comparison of the methods based on convergence properties, implementation characteristics and scalability.

3.1 Sampling methods

Monte Carlo sampling, often abreviated as MCS, is one of the most commonly used methods in uncertainty quantification. According to [Xiu08] and [Iac11], the core idea of the algorithm is that one generates (independent) realizations of random inputs based on a prescribed probability distribution function and solves the problem for each generated sample. Thus, for each realization, the data is fixed and hence the problem becomes *deterministic*.

After one solves the deterministic realizations of the problem and obtaines the corresponding solutions, the statistical information can be easily extracted, e.g. $expectation(mean)$, $variance$, etc. It is important to remark that the formulas employed for obtaining the statistical information (*c.f. Algorithm 1, Step 3*) are independent of the choice of the probability density function used to generate the random inputs. The algorithm for using Monte

Carlo sampling in Uncertainty Quantification can be summarized as follows:

Algorithm 1 Monte Carlo sampling

1. Generate n samples of the random inputs, where n is user defined;
2. For each generated sample, solve the associated stochastic equation and obtain the solution $x^{(i)}(t,\omega)$;
3. Estimate the required solution statistics, e.g.
expectation(mean): $\mathbb{E}[x^{(i)}(t,\omega)] = \int_{\mathbb{R}} x^{(i)}(t,\omega)\rho(\omega)d\omega \approx \frac{1}{n}\sum_{i=1}^{n} x^{(i)}(t,\omega^{(i)})$
and variance: $Var[x^{(i)}(t,\omega)] = \mathbb{E}[(x(t,\omega) - \mathbb{E}[x(t,\omega)])^2] \approx \frac{1}{n-1}\sum_{i=1}^{n}(x^{(i)}(t,\omega^{(i)}) - \mathbb{E}[x^{(i)}(t,\omega)])^2$

3.2 Stochastic Collocation methods

As the authors in [Iac11], [JCC12] and [Xiu08] present, in collocation methods, we seek to satisfy the governing equation at a discrete set of points ("nodes") in the corresponding random space. From this, we see that, similar to MCS, we work again with a *deterministic* system. But unlike the Monte Carlo method, where the approach is based on sampling, the focus in stochastic collocations is, as presented in [Xiu08], to utilize an expansion that seeks to approximate the random process via orthogonal polynomials of random variables. The used expansion is based on *generalized Polynomial Chaos*(gPC) and for a more detailed view on this topic, see [JCC12], [Xiu08].

Let us consider an ∞-*dimensional* space $\mathbb{W}$ of orthogonal polynomials with respect to the probability measure ρ in $\mathbb{R}$, i.e.

$$\int_{\mathbb{R}} \rho(x)\cdot\phi^m(x)\cdot\phi^n(x)\cdot dx = h_{mn}^2\cdot\delta_{mn},\ \forall m,n\in\mathbb{N}_0 \tag{6}$$

where $\phi^m,\phi^n \in \mathbb{W}$, δ_{mn} is the Kronecker delta function and $h_{mn}^2 \in \mathbb{R}$ is the normalization factor. Furthermore, one can always normalize the orthogonal bases such that $h_{mn}^2 \equiv 1$ and this will be adopted throughout this paper. As the considered stochastic model is of $Gaussian$ nature with the support $\Gamma = \mathbb{R}$, the used polynomials are the $Hermite$ polynomials ($\phi^0 = 1,\ \phi^1 = 1,\ \phi^2 = x^2 - 1, \phi^3 = x^3 - 3x$ etc.), having the normalization factor $\int_{\mathbb{R}} \rho(x)\cdot(\phi^m)^2(x)\cdot dx = m!,\ \forall m\in\mathbb{N}_0$

As it is presented in [Xiu08], the Q^{th}-order gPC approximation of $x(t,\omega)$ can be obtained by projecting x onto the space $\mathbb{W}_Q$, i.e.

$$x(t,\omega) \approx x^Q(t,\omega) = \sum_{j=1}^{Q} c^j(t)\cdot\phi^j(\omega), \tag{7}$$

where $c^j(t)$ are the expansion coefficiants computed as:

$$c^j(t) = \int_{\mathbb{R}} x(t,\omega) \cdot \phi^j(\omega) \cdot \rho(\omega) \cdot d\omega = \mathbb{E}[x(t,\omega) \cdot \phi^j(\omega)], j = 1 \dots \mathrm{Q}, \tag{8}$$

where $\mathbb{E}$ is the expectation operator. From this, the statistics can be obtained easily. The expectation is calculated as (using (6) and (7)):

$$\begin{aligned}
\mathbb{E}[x(t,\omega)] &= \int_R x(t,\omega) \cdot \rho(\omega) \cdot d\omega \approx \int_R \sum_{j=1}^{Q} c^j(t) \cdot \phi^j(\omega) \cdot \rho(\omega) \cdot d\omega \\
&= \sum_{j=0}^{Q} c^j \cdot \int_R \phi^j(\omega) \cdot \rho(\omega) \cdot d\omega \\
&= c^0 \cdot \int_R \phi^0(\omega) \cdot \rho(\omega) \cdot d\omega + \sum_{j=1}^{Q} c^j \cdot \int_R \phi^j(\omega) \cdot \rho(\omega) \cdot d\omega = c^0
\end{aligned} \tag{9}$$

and the variance:

$$\begin{aligned}
Var[x(t,\omega)] &= \mathbb{E}[(x(t,\omega) - \mathbb{E}[x(t,\omega)])^2 = \int_R \Big(x(t,\omega) - \mathbb{E}[x(t,\omega)]\Big)^2 \cdot \rho(\omega) \cdot d\omega \\
&\approx \int_R \Big(\sum_{j=0}^{Q} c^j(t) \cdot \phi^j(\omega) - c^0\Big)^2 \cdot \rho(\omega) \cdot d\omega \\
&= \int_R \Big(\sum_{j=1}^{Q} c^j(t) \cdot \phi^j(\omega)\Big)^2 \cdot \rho(\omega) \cdot d\omega = \sum_{j=1}^{Q} (c^j(t))^2.
\end{aligned} \tag{10}$$

The literature presents several approaches for stochastic collocations, for instance see [JCC12], but we focus on the one based on a pseudo-spectral approach as it is presented in [Xiu08] and [XLS01]. In this approach, we compute the coefficiants $c^j(t)$ in (8) by using numerical quadrature, in the following way:

$$c^j(t) = \sum_{i=0}^{N} x(t,\omega^{(i)}) \cdot \phi(\omega^i) \cdot \alpha^i, \tag{11}$$

where $\{\omega^i, \alpha^i\}_{i=1}^N$ are a set of nodes and weights chosen such that the error between (8) and (11) is as small as possible (also called "aliasing error"). In (11), as it is presented in [Xiu08], $x(t,\omega^i)$ represents the deterministic solution with ω^i fixed.

In a 1-dimensional stochastic space where the random variable is of *Gaussian* nature, the optimal choice for nodes and weights is via *Gauss* quadrature and because the used orthogonal polynomials are the *Hermite* polynomials, we employ the *Gauss-Hermite* quadrature.

3.3 Stochastic Galerkin method

As we have seen in the previous section, the key in using a gPC approximation (7) is the evaluation of the coefficiants $c^j(t)$. Formula (8), used to compute of the coefficiants of the gPC expansion, is of little use, as it involves the unknown solution $x(t, \omega)$. An alternative method to stochastic collocations is the *Stochastic Galerkin* method. The core idea of the method is that we write (1) in a stochastic *weak form* and then we perform a projection, to obtain a (coupled) system of ODEs, as the author in [Xiu08] specifies.

To emphasize this idea, we consider the gPC approximation (7) of order Q for $x(t, \omega)$. Furthermore, let us consider a similar expansion for the random parameter k, based on the fact that k is normally distributed:

$$k(\omega) = \mu + \sigma \cdot \omega = \mu \cdot \phi^0(\omega) + \sigma \cdot \phi^1(\omega) \tag{12}$$

Thus, by plugging (7) and (12) into (1), we obtain:

$$\begin{gathered} \sum_{j=0}^{Q} (c^j)''(t) \cdot \phi^j(\omega) + c \cdot \sum_{j=0}^{Q} (c^j)'(t) \cdot \phi^j(\omega) + \\ (\mu \cdot \phi^0(\omega) + \sigma \cdot \phi^1(\omega)) \cdot \sum_{j=0}^{Q} c^j(t) \cdot \phi^j(\omega) = f(t) \end{gathered} \tag{13}$$

Next, a *Galerkin projection* is made using $\phi^k(\omega), \forall k = 0 \dots$ Q. After performing the projection, (13) becomes:

$$(c^k)''(t) + c \cdot (c^k)'(t) + \mu \cdot c^k(t) + \frac{1}{(\phi^k(\omega))^2} \cdot \sum_{j=0}^{Q} (\sigma \cdot c^j(t) \cdot e_{1jk}) = f(t) \cdot \int_R \phi^k(\omega) \cdot \rho(\omega) \cdot d\omega \tag{14}$$

where $e_{1jk} = \mathbb{E}[\phi^1(\omega) \cdot \phi^j(\omega) \cdot \phi^k(\omega)]$. Together with $\frac{1}{(\phi^k(\omega))^2}$, e_{1jk} can be evaluated analytically from the definition of the polynomials ϕ^i, as the authors in [JCC12] specify. We point out that the initial conditions in the equations (13) and (14) remain zero.

3.4 Comparison between Monte Carlo, Stochastic Collocations and Stochastic Galerkin methods

In order to efficiently implement and use the proposed methods, it is quintessential to thoroughly understand their particularities, advantages and disadvantages.

Regarding the $Monte\ Carlo$ method, its advantages are the fact that it is straightforward to use, thus easy to implement. Also, it is easily parallelizable and it is designed to work for arbitrary probability distribution functions. It features a major disadvantage, this is its accuracy depends on a large number of executions, making it computationally expensive. As the author in [Xiu08] specifies, the mean value typically converges as $\frac{1}{\sqrt{K}}$, where K is the number of realizations.

It is important to remark that the convergence of MCS is independent of the number of dimensions, thus, when the number of dimensions is high (typically ≥ 50), it is a very good choice. MCS requires the solution of the given equation, hence a numerical algorithm that automatically introduces a certain error. Moreover, as the method is based on sampling, this introduces an error as well. To summarize, the numerical error in MCS is:

$$\epsilon_{MC} = \epsilon_{sampling} + \epsilon_{num_solution} \tag{15}$$

We remark that in the literature are proposed several methods to improve the convergence of MCS by e.g. using *quasi-Monte Carlo* method. For a more detailed view, see [Fox99].

The $Stochastic\ Collocation$ method is also straightforward to use and implement, as its main focus is to approximate the coefficiants $c^j(t)$ in (8) by (11). This requires a pair of nodes and weights used for numerical quadrature and a numerical solver for approximating the exact solution of (1). All solver calls are independent from each other and when working in multi-dimensional spaces (typically ≤ 50) there are proposed several approaches, as in [BG04] or [DKO14].

Unlike MCS, it converges very fast and it requires only a small number of coefficiants to efficiently determine the statistics. Regarding the existent numerical error, it is composed by the error of the approximation via gPC expansion (7), the numerical quadrature in (11) and the numerical solver for determining an approximation of the exact solution in (11). Thus,

$$\epsilon_{SC} = \epsilon_{gPC_approx} + \epsilon_{num_quadrature} + \epsilon_{num_solution} \tag{16}$$

Stochastic Galerkin method is an intrusive method and its basic idea is that the projection is conducted in order to ensure that the error is orthogonal to the functional space spanned by the finite-dimensional basis. This is very efficient, as the error gets minimized, albeit it has an overhead when the form of (1) is complicated. Moreover, (14) is most of the time coupled, so this introduces an aditional overhead.

The main disadvantage of this method is the fact that the application code is complicated to implement and moreover, it is not reusable as in the case of the other presented methods (thus, the method is intrusive). Albeit, in the case of a "nice" form of (1), this method is very efficient, due to the fact that the convergence is exponential, as the author in [Xiu08] presents. We point out that (14) requires a numerical solver, that introduces a numerical error, along with the error caused by the gPC approximation (7) of $x(t, \omega)$.

$$\epsilon_{SC} = \epsilon_{gPC_approx} + \epsilon_{num_solution} \tag{17}$$

4 Numerical results

In this section we present the results obtained with the employed stochastic methods and to asses their quality by comparing them with the results obtained via the analytical solution. The quantities of interest are the *expectation(mean)* and the *variance* of the solutions, taken at time $t_{interest}$ = 10 s. The comparison will be done using the numerical values, the runtime, the relative errors and the used number of samples.

Table 1: Results obtained using a numerical solver

MCS					
Samples	Expectation	Variance	runtime(s)	E.R.E. [1]	V.R.E. [2]
100	0.00716	0.00603	2.5113	30.0%	71.7%
1000	0.00868	0.01075	6.1123	15.1%	49.5%
2500	0.00948	0.01505	17.6021	7.2%	29.2%
10000	0.00957	0.01801	24.2412	6.4%	15.3%
50000	0.00996	0.01988	120.6654	2.5%	6.5%
Collocations					
Points	Expectation	Variance	runtime(s)	E.R.E.	V.R.E.
20x2	0.01023	0.00802	0.4300	$\approx 8 \cdot 10^{-5}$	86.0%
20x4	0.01023	0.01404	0.5502	$\approx 8 \cdot 10^{-5}$	34.1%
20x5	0.01023	0.01812	0.7000	$\approx 8 \cdot 10^{-5}$	14.7%
20x6	0.01023	0.02019	0.8370	$\approx 8 \cdot 10^{-5}$	4.9%
20x8	0.01023	0.02115	1.1340	$\approx 8 \cdot 10^{-5}$	0.28%
Galerkin					
Points	Expectation	Variance	runtime(s)	E.R.E.	V.R.E.
20x2	-0.00343	0.00019	0.7421	133.5 %	86.0%
20x4	0.00968	0.00784	2.2512	5.3 %	63.1%
20x5	0.01017	0.01418	3.3131	0.5 %	33.3%
20x6	0.01022	0.01847	4.6418	0.03 %	13.2%
20x8	0.01023	0.02107	7.7876	$\approx 2.47 \cdot 10^{-5}$	0.96%

Moreover, having the analytical solution, we can see how much does the numerical error influence the results, by comparing the outcomes via analytical solution and the numerical solver. All simulations were performed using an Intel Core i7 2630M microprocessor and Matlab R2013b.

From Figure 2 (*c.f. Section 2.1*), we see that different values of k between 0 and 1 induce a small variance in the analytical solution, thus, we can infer that the variance in the numerical results will be small, too.

When using the analytical solution, the statistical moments are calculated using the formulas $\mathbb{E}[x(t,\omega)] = \int_R x(t,\omega) \cdot \rho(\omega) \cdot d\omega$ and $Var[x(t,\omega)] = \mathbb{E}[(x(t,\omega) - \mathbb{E}[x(t,\omega)])^2$, where the integrals are computed numerically, using *Gauss-Hermite* quadrature. We remark that the quadrature will introduce a small numerical error in the results that is $\sim 10^{-15}$. Using 15 integration nodes and weights, we obtain:

$$\mathbb{E}[x(t_{interest},\omega)] = 0.01023,\ Var[x(t_{interest},\omega)] = 0.02128 \tag{18}$$

In Table 1 [3], we present the results obtained with the stochastic methods when we use a numerical solver.

[1] Expectation relative error

[2] Variance relative error

[3] In the first column, for Collocations and Galerkin methods, the first number represents the number of integration points(N); the second one, the maximal degree of ϕ when computing the coefficiants $c^j(t)$(Q)

Table 2: Results obtained using the analytical solution

MCS					
Samples	Expectation	Variance	runtime(s)	E.R.E.	V.R.E.
100	0.00671	0.00745	0.1123	34.3%	69.9%
1000	0.01137	0.00118	0.1653	11.0%	44.3%
2500	0.00954	0.01639	0.2075	6.4%	22.9%
10000	0.00973	0.01835	0.4412	4.0%	13.7%
50000	0.01021	0.02107	2.6098	0.01%	0.29%
Collocations					
Points	Expectation	Variance	runtime(s)	E.R.E.	V.R.E.
20x2	0.01023	0.00297	0.03418	$\approx 1.77 \cdot 10^{-15}$	84.3%
20x4	0.01023	0.01401	0.01725	$\approx 1.77 \cdot 10^{-15}$	33.7%
20x5	0.01023	0.01815	0.02050	$\approx 1.77 \cdot 10^{-15}$	14.5%
20x6	0.01023	0.02023	0.02823	$\approx 1.77 \cdot 10^{-15}$	4.7%
20x8	0.01023	0.02122	0.05532	$\approx 1.77 \cdot 10^{-15}$	0.02%
Galerkin					
Points	Expectation	Variance	runtime(s)	E.R.E.	V.R.E.
20x2	0.00214	0.00568	0.3912	82.7 %	75.2%
20x4	0.00995	0.00899	0.5131	3.7 %	54.1%
20x5	0.01019	0.01628	0.6352	0.34 %	24.7%
20x6	0.01022	0.01927	0.8768	0.03 %	11.6%
20x8	0.01023	0.02127	1.2419	$\approx 4.58 \cdot 10^{-15}$	0.01%

From this, we see that all three methods converge, albeit in different manners. MCS requires a large number of samples($\sim 10^4$) to achieve convergence and this causes an overhead, due to the need for a large number of FLOPS (by an increased runtime) and thus, making the method computationally expensive. Moreover, from a computational science point of view, the inefficiency of this method is also caused by the large storage requirements, because while the number of samples increases, we need to store all computed solutions.

On the other hand, with much less used points, the Stochastic Collocations and Stochastic Galerkin methods converge very quickly and with much less computation power. Thus, we ascertain that the gPC based methods are more suitable for this application and although their overall numerical errors have three and respectively, two sources (*c.f. Section 3.4*), they produce very accurate results.

In the next simulation (*c.f. Table 2*), we use the analytical solution for the computations in MCS, Stochastic Collocations and Galerkin methods, thus eliminating the error caused by the numerical solver. In this way, we can better asses the obtained results.

As expected, the results for all methods are improved, proving that the error introduced by the numerical solver playes an important role in the final outcomes. In the case of MCS, for a large number of samples, the solution converges again, but this time, as we expect, the required number of FLOPS is drastically decreased (by approximately a factor of 60).

When using the analytical solution, the gPC based methods perform also much better, as the numerical error introduced by the numerical solver is eliminated. This leads to the conclusion that the gPC approximation (7) is very precise, even for a small number of points (i.e. a coarser truncation).

Considering a comparison between the Stochastic Collocations and Stochastic Galerkin methods, we can infer that although the Galerkin method converges slower, for a large number of points it performes better than the Collocations method. We should keep in mind that as the obtained results with these methods are similar, one is non-intrusive (Collocations) as the other is intrusive (Galerkin) - *c.f. Section 3*.

5 Conclusions

This paper underlined and compared results obtained with numerical methods used in uncertainty quantification, in a practical context. The focus was to use both intrusive and non-intrusive methods. As uncertainty is naturally present in every physical process or system, it is quintessential to consider it before any simulation is done.

The proposed example was a simple ODE model of a second order oscillator and the uncertainty was considered to be in one of the parameters. Being a simple model, it had a computable analytical solution that was used to asses the quality of the obtained results. Based on simulations, we assesed that both Stochastic Collocation method and the Stochastic Galerkin method are very suitable to treat the uncertainty in this system.

Moreover, as for the future work, we intend to transform the existing code into an UQ API and to also incorporate our sparse grids framework into the application code [4], this model can be used also as a complex system, as it is easily extendable to a higher chaos dimension.

References

[BG04] H.J. Bungartz and M. Griebel. Sparse Grids. *Acta Numerica*, pages 1–123, 2004.

[DKO14] S.V. Dolkov, B.N. Khoromskij, and I.V. Oseledets. Computation of Extreme Eigenvalues in Higher Dimensions using block tensor train format. *Computer Physics Communication*, 185:1207–1216, 2014.

[Fox99] B.P. Fox. *Strategies for Quasi-Monte Carlo*. Kluwer Academic Pub., 1999.

[Iac11] G. Iaccarino. Lectures on Uncertainty Quantification in Computational Science, Department of Mechanical Engineering Stanford University, 2011.

[JCC12] B. Jia, S. Cai, and Y. Cheng. Stochastic Collocation Method for Uncertainty Propagation. *Papers - American Institute of Aeronautics AND Astronautics*, 6:4770–4787, 2012.

[Ras06] I. Rasa. *Lectures on Probability Theory and Stochastic Processes*. U.T.PRES, 2006.

[4] see http://www5.in.tum.de/SGpp/releases/index.html

[WIC06] Q. Wang, G. Iaccarrino, and P. Constantine. Uncertainty quantification in simple linear and non-linear problems. *Center for Turbulence Reseach - Annual Research Briefs*, 2006.

[Xiu08] D. Xiu. Fast Numerical Methods for Stochastic Computations : A Review. *Communication in Computational Physics*, 5:242–272, 2008.

[XLS01] D. Xiu, D. Lucor, and C.H. Su. Stochastic Modeling of Flow-Structure Interactions using Generalized Polynomial Chaos. *Journal of Fluids Engineering*, 124:51–59, 2001.

GPU-Based Regression Analysis on Sparse Grids

Steffen Hirschmann

University of Stuttgart
Universitätsstraße 38
D-70569 Stuttgart
hirschsn@studi.informatik.uni-stuttgart.de

Abstract: Prediction and forecasting has become very important in modern society. Regression analysis enables to predict easily based on given data. This paper focuses on regression analysis on sparse grids using the existing toolbox Sparse Grid ++ (SG^{++}). The core workload of the regression analysis will be implemented on graphics cards using NVIDIA's Compute Unified Device Architecture (CUDA). Therefore, we give guidance how to get high performance when dealing with this particular problem using CUDA enabled graphics cards. We also focus on problems where the datasets are larger than the available device memory. Finally, we present test results for real-world and artificial datasets.

1 Introduction and Related Work

Machine Learning is an important field in computer science. We rely on it every day mostly in predictions and forecasts. Applications reach from traditional weather forecasts up to shopping prediction on the internet or the classification of galaxies in astrophysics. The trend of collecting larger and larger datasets and processing them is called *Big Data*.

We want to employ a traditional technique to discover relationships amongst variables: regression analysis. However, since we want to be able to deal with huge datasets, traditional regression analysis is not applicable since it does not scale linearly in the amount of data points. Therefore, we use the approach of [Pfl10]: Discretize the feature space using sparse grids. Sparse grids [BG04] are a well-known technique to avoid the so-called *curse of dimensionality*, so we are able to deal with datasets that have more dimensions than classical methods allow for. In fact, this approach scales linearly in the amount of data points.

When dealing with huge amounts of data we also need to parallelize the workload. Using graphics cards is known to give high speedups if the application is well parallelizable. However, traditional sparse grids algorithms work recursively on the grid and thus cannot be applied here. Hence, we employ iterative sparse grid algorithms as shown in [HP11]. Recently, the concerning algorithms from the toolbox SG^{++}[sgp14, Pfl10] have been ported to several platforms like Open Computing Language (OpenCL) [HP13] or the Intel®Many Integrated Core Architecture [HKP+12].

In this paper we want to show how to implement these iterative algorithms using NVIDIA's CUDA. We continue by showing how to speed up the algorithms making use of CUDA and other programming concepts. Also, we present techniques to handle datasets that are larger than the available memory of graphics cards, which is very important for dealing with *Big Data*.

After a brief introduction to the concepts of regression analysis in Section 2 we present a variational formulation and the resulting system of linear equations for doing regression analysis on sparse grids. In Section 3, we briefly introduce NVIDIA's CUDA [cud12] programming concept to the reader. Afterwards in Section 4, we show implementation details and concepts to speed up the basic algorithms. We conclude by presenting tests with real-world and artificial data in Section 5.

2 Regression Analysis

In statistics, regression analysis is a method to find relationships between variables; see for example [Syk93]. Given a dataset $S = \{(\vec{x}_i, y_i) | i = 1, \ldots, m\}$ of m d-dimensional vectors $\vec{x}_i$ and their respective function values y_i, the goal of a regression analysis is to determine a function f such that: $\forall i : y_i = f(\vec{x}_i)$. To ensure a good quality of the learned function, one usually divides the dataset S into two independent parts: The training dataset S_{train} where the function is learned from and the testing dataset S_{test} against which the learned function is validated. Typically, one chooses $\frac{|S_{train}|}{|S_{test}|} = 2$.

A simple binary classification can be achieved if the codomain of f is divided into two different classes. Depending on the value of the prediction $f(\vec{x})$ the respective class is chosen.

2.1 Regression Analysis on Sparse Grids

Sparse grids are a discretization technique for dealing with high-dimensional problems. For a general introduction we refer to [BG04]. To understand the presented algorithms, the reader needs to know that the basis functions are build from one dimensional functions via a tensor product approach. These one dimensional functions are linear hat functions, each associated to a certain level l and index i:

$$\phi_{l,i}(x) = \max\left(1 - \left|2^l \cdot x - i\right|, 0\right) \tag{1}$$

When doing a regression analysis, we want to find a function $f_N \in V_N$, where V_N is some function space. We discretize the feature space [Pfl10] using a sparse grid and use the basis functions from Equation (1) associated to the grid points as basis of V_N. A function $f_N \in V_N$ can be expressed as linear combination of the basis functions: $f_N(\vec{x}) = \sum_{i=1}^{N} \alpha_i \phi_i(\vec{x})$. To ensure well-posedness and a unique solution, we use a variational

formulation according to [Pfl10] based on a least squares fit:

$$f_N \stackrel{!}{=} \operatorname*{argmin}_{f_N \in V_N} \left(\frac{1}{m} \sum_{i=1}^{m} \underbrace{(y_i - f_N(\vec{x}_i))^2}_{\text{data term}} + \underbrace{\lambda \|\vec{\alpha}\|_2^2}_{\text{smoothness term}} \right)$$

Where the data term ensures closeness of f_N to the training data. The squared euclidean norm of the coefficient vector is a simple measure for the smoothness of a function and fast to compute. To minimize this functional, we differentiate with respect to the coefficients α_i which are to be determined and set the derivatives to zero. The resulting system of linear equations (SLE) is

$$\left(\frac{1}{m} B B^T + \lambda I \right) \vec{\alpha} = \frac{1}{m} B \vec{y}, \tag{2}$$

where I is the identity matrix which stems from the smoothness term. B stems from the data term and is defined as: $B_{i,j} = \phi_i(\vec{x}_j)$. So given $\vec{y}$ and V_N via a basis, one can solve the SLE from Equation (2) using iterative solvers as the conjugate gradient method (CG) which can handle high dimensional matrices. When applying CG to the SLE from Equation (2) we only have to provide multiplication with B and B^T. Since these two multiplications are by far the most expensive operations when solving the SLE we focus on them and parallelize them using graphics cards.

3 Compute Unified Device Architecture

The Compute Unified Device Architecture (CUDA) is a parallel computing architecture by NVIDIA, which allows to perform calculation on NVIDIA graphics cards. These calculations are done in parallel by all the cores of the graphics cards. However, if execution paths of the calculations diverge the program will be executed sequentially. This is classified as *Single Instruction Multiple Threads* (SIMT) as an extension to Flynn's taxonomy. Due to this highly parallel architecture, CUDA allows for huge speedups in several applications particularly highly parallelizable ones. Compare [cud12].

3.1 Concepts

CUDA is an extension to the C programming language. It adds three main abstractions: thread hierarchy, device memory and synchronization [cud12]. A computer hosting a graphics card is called a *host* and the card itself a *device*.

Thread Hierarchy Every CUDA thread executes a *CUDA kernel*. This is a special function which can be executed on the device and called from the host. All threads are gathered in blocks. A block can be up to three dimensional. Threads from the same block

are scheduled together, reside on the same processor core and share the so-called *shared memory*. As of compute capability 2.x, a block may contain up to 1024 threads [cud12]. Since one block may not cover the whole problem, blocks are arranged in a so-called grid. Both, block size and grid size can be up to three dimensional. The value of these sizes must be specified at kernel launch time. In order to get the best performance one has to carefully choose block and grid sizes, depending on the actual calculations and the amount of memory that needs to be shared.

Memory Hierarchy A CUDA capable device provides different kinds of memory [cud12]. The global memory forms the highest level of the memory hierarchy and has the highest access times. It is shared throughout the whole device and can be accessed from the host. Copying to global memory can be done asynchronously, i.e. parallel to kernel executions. On the lowest level of the memory hierarchy there is the shared memory. It can be accessed very fast but is only shared throughout one block of threads and very limited.

4 Implementation

In order not having to implement basic things like the abstract learning process or sparse grid construction from scratch but to focus on the GPU-based algorithms, we used a sparse grid library called Sparse Grid ++.

4.1 Sparse Grid ++

Sparse Grid ++ (SG^{++}) [sgp14, Pfl10] is a toolbox that allows to use sparse grids without great expense. It is flexible and does not need the initial overhead that has to be spent while implementing sparse grids and corresponding algorithms.

SG^{++}implements regression analysis via different classes. The class responsible for the abstract learning process uses a conjugate gradient solver to solve the SLE (2). So-called operation classes implement the underlying numerical algorithms.

4.2 CUDA Operation Classes

The basic thing we had to do is to provide an operation class that multiplies a vector with B and B^T utilizing graphics cards via CUDA as best as possible. The operation class itself has to do some administrative tasks required by SG^{++}, like measuring execution times.

To evaluate SLE (2) we need the grid's level and indices as well as the dataset. The grid (of grid size g) is passed to the operation class as matrices storing levels $(l_{k,l}) \in \mathbb{R}^{g,d}$ and indices $(i_{k,l}) \in \mathbb{R}^{g,d}$ of the basis functions. Furthermore, the dataset (of dataset size m) $(x_{i,d}) \in \mathbb{R}^{m,d}$ is known to the operation class. These matrices as well as the source

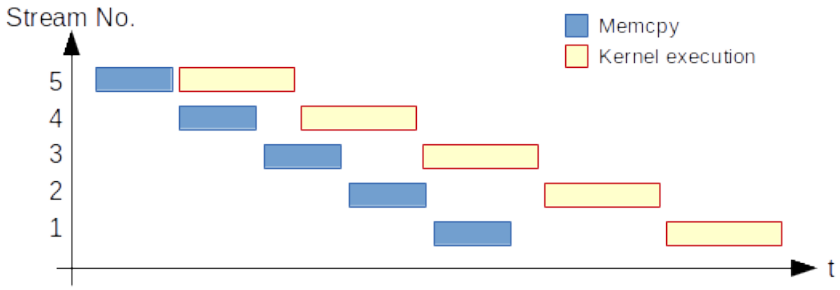


Figure 1: 5 streams each one operating on one of the 5 dataset chunks. Note, that kernel calls may not overlap. However, newer compute capabilities may allow for concurrent kernel execution.

vectors have to get uploaded into the global memory of the device. Therefore, we created a memory abstraction layer.

4.2.1 Memory Abstraction Layer

The memory abstraction layer is responsible for allocating storage on the device, uploading to the device and downloading from it. In between the memory management this layer has to call the actual CUDA kernels, which are implemented in the lowest layer of our architecture.

Streaming Since the global memory of graphics cards (up to 12 gigabytes) is very limited in contrast to host memory, datasets might be too big to fit entirely into the global memory of the device. Another motivation is that we want to reduce copying times, i.e. start the algorithms before every single data point is copied, in order to achieve good overall performance. Therefore, we have to incrementally copy small parts of the dataset onto the GPU, execute the kernels on these parts and then proceed with the next one. The applied algorithm might require for an accumulation of the partial results into one global; in some cases, synchronization might be unavoidable. This partial copy and execution is called *streaming*. There are two different approaches.

The first is to divide the dataset into N equally sized parts and immediately schedule copy and kernel execution on each one. This technique can only be used for datasets fitting onto the device since we cannot guarantee—without explicit synchronization—that only M of these N chunks reside in global memory simultaneously. This might occur if kernel execution takes a lot of time. Figure 1 shows this approach schematically.

The second approach is to divide the dataset into N parts again, however, only using M streams. These streams loop through the N parts and copy them to the card in order to afterwards execute the kernels on them. When a chunk is done, a stream proceeds with the next chunk. This technique ensures, that only M dataset chunks reside in global memory at the same time. The M streams allow to simultaneously copy a dataset chunk and execute

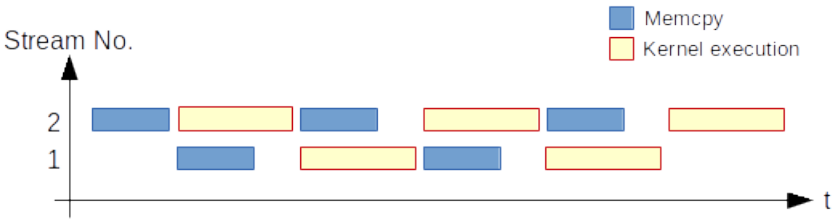


Figure 2: 2 streams iteratively processing 5 dataset chunks. Note that the kernel calls may not overlap. However, newer compute capabilities may allow for concurrent kernel executions.

a kernel on another. This approach is schematically shown in Figure 2 for $M = 2$.

4.2.2 Kernels

The kernels implement the actual multiplication with B and B^T as routines *mult* and *multTranspose*. We implemented the kernels iteratively, adapted from [HP11], rather than recursively. The high-level algorithm of the mult and multTranpose functions are shown in Algorithm 1. Note the basis function evaluation in the innermost loop according to Equation (1). The two algorithms are converted into CUDA kernels as follows: The outermost for loops are omitted and therefore m respectively g threads are launched. Each of these threads has an index corresponding to a loop indices of the former loop variable. So each thread is responsible for calculating one entry of the result vector, thereby iterating over the grid respectively dataset.

for $\gamma \leftarrow 0..g-1$ **do**
  $u_\gamma \leftarrow 0$
  for $\mu \leftarrow 0..m-1$ **do**
    $p \leftarrow v_\mu$
    for $\delta \leftarrow 0..d-1$ **do**
      $p \leftarrow p \cdot \phi_{l_{\gamma,\delta},i_{\gamma,\delta}}(x_{\mu,\delta})$
    end for
    $u_\gamma \leftarrow u_\gamma + p$
  end for
end for

for $\mu \leftarrow 0..m-1$ **do**
  $v_\mu \leftarrow 0$
  for $\gamma \leftarrow 0..g-1$ **do**
    $p \leftarrow \alpha_\mu$
    for $\delta \leftarrow 0..d-1$ **do**
      $p \leftarrow p \cdot \phi_{l_{\gamma,\delta},i_{\gamma,\delta}}(x_{\mu,\delta})$
    end for
    $v_\mu \leftarrow v_\mu + p$
  end for
end for

Algorithm 1: Pseudo code for function mult on the left. Calculates $u = Bv$. Pseudo code for function multTranspose on the right. Calculates $v = B^T\alpha$. See [HP11]. The grid has size g, the dataset m, there are d dimensions. A single basis function is described by its level l_γ and index i_γ which itself is a product of one-dimensional basis functions which have an additional index for the dimension. The variable p holds the support of a single basis function.

Using Template Metaprogramming A CUDA thread executing these kernels is most of the time busy incrementing the loop index of the innermost loop, comparing it to the number of dimensions and doing jumps from the loop end to its beginning. This has a huge impact on performance since due to the execution model graphics cards rely on, CUDA cores do not have a branch prediction like modern CPUs.

To tackle the problem of completely removing the innermost loop we employed template metaprogramming. This means, that the CUDA kernels are implemented as C++ templates with the number of dimensions as template parameter. Hence, the compiler can unroll the innermost loop because it has constant trip count. We added a wrapper routine for the kernels which unroll them for a dimensionality up to 100 by default. This bound might be increased arbitrarily. The compiler, then, creates 100 different implementations of a kernel with the inner loop unrolled. A disadvantage of this procedure is that compilation times increase tremendously—depending on the upper unroll bound. A benefit of template kernels is that a thread can keep his associated (multidimensional) grid or data point in registers. Since register count must be determined at compile time, this is only possible if the dimensionality is a compile time constant.

Utilizing Shared Memory The algorithms shown in Algorithm 1 require each thread to access the same data respectively grid point at the same instruction. To achieve better performance, we load this common data into shared memory for faster access. The threads with indices in $\{1, ..., d\}$ in a block will load the data or grid point into shared memory right before the instruction $p \leftarrow v_\mu$ respectively $p \leftarrow \alpha_\mu$. After the shared memory load the threads have to be synchronized in order not to load from undefined memory locations. Another synchronization is needed at the end of the loop to ensure every thread is done with this particular grid or dataset point.

This concept is independent of template kernels, since CUDA allows for specifying the amount of bytes of one single array, that resides in shared memory, at kernel launch time.

Parallelizing Both Outer Loops In most real world cases the grid size is much smaller than the dataset size. If the factor $\frac{m}{g}$ grows too large performance will suffer, since a thread executing a mult kernel, which is parallelized over the grid points, will iterate over all data points. To take care of this problem, the dataset is split into several parts. Every one of these parts can be calculated on different threads. There are two possibilities to achieve this. Firstly, to use a two dimensional block size (x, y) and let all threads associated with the same x component calculate the same result entry. However, we need accumulation of the result inside one single block. If we want this to take place in shared memory, we need runtime determined shared memory (which can only be a one dimensional) of size $x \cdot y$. That would require a huge amount of index calculations. Another problem is that mere a single array inside shared memory can have runtime determined size. So this approach is hardly compatible with the use of shared memory as described in the previous paragraph.

A different approach to overcome an imbalance between grid size and dataset size is to use streaming with a dataset chunk size roughly equally to the grid size. Then, every thread has to iterate over the same amount of elements (no matter which of the two kernels is

executed) and hence the overall algorithm parallelizes better. However, we would have to explicitly synchronize this as mentioned in the previous subsection. This approach usually performs much better than using a two dimensional block size and it it simpler to implement.

5 Numerical Tests

First of all, we tested the kernels separately in order to ensure functionality and measure peak performance. Afterwards we ran tests on real-world and artificial datasets to measure real application performance and compare it to the performance of a CPU.

5.1 Benchmarks

The kernels were tested on several NVIDIA graphics cards, namely GeForce 560Ti, GeForce GTX680, Tesla K10 and Tesla K20. We used the first type of streaming presented in the previous chapter. Additionally, we used shared memory. Random numbers in $[0, 1)$ created by `rand()` from the glibc served as input for the data set and the grid's levels and indices. They do not make any sense when interpreted as levels or indices of sparse grids. However, in the benchmarks we wanted to find out the theoretical computing performance of the kernels. Note, that also the vast grid sizes used in the benchmarks are not realistic. Every shown result has been averaged over ten tests. Again, d is the number of dimensions, m the dataset size and g the grid size.

The timings include all memory transfers and the accumulation of the result if using multiple cards or GPUs. Table 1 shows the achieved effective floating point operation per second for single precision. The term "operations" refers to the abstract algorithms shown in Algorithm 1 and not to the operations the GPU really executes; hence index calculations, shared memory loads, result accumulation, etc. are not included.

Table 1: Single precision peak performance. Since we made use of 10 streams, the dataset size is ten times larger than the grid size. Besides, the dataset size is doubled in dual GPU tests since parallelization between different GPUs is based on splitting the dataset. The 560Ti has too few registers, so it was not able to calculate more than 16 dimensions using 640 threads per block. Reducing the number of threads per block would have solved this problem, but we did not want to vary them.

Configuration	m	g	d	eff. GFLOPS (sp)
560Ti	1.280.000	128.000	16	300
K10 (1 GPU)	1.280.000	128.000	24	420
K10 (2 GPUs)	2.560.000	128.000	20	800
GTX680	1.280.000	128.000	24	620
2x GTX680	2.560.000	128.000	20	1.200

To compare double precision performance to single precision performance, we ran a test varying the precision and leaving all other parameters the same. This is in particular interesting for the K20, since NVIDIA presents it as an accelerator card which is ought to be very well suited for double precision computations. Other cards—mostly consumer ones like the GTX680 we used—are known to have low double precision performance. The test results are shown in Table 2. Indeed our test shows that the K20 performs vastly better than all other tested cards when it comes to double precision. It turns out, that the quotient $\frac{T_{\text{dp}}}{T_{\text{sp}}}$ is closer to the one of a CPU than to the ones of the other GPUs.

Table 2: Single vs. double precision benchmark. Measured via the fraction $\frac{T_{\text{dp}}}{T_{\text{sp}}}$ (rounded to two decimal places). Parameters: $m = g = 102400$, $d = 10$. The K20 shows a very good double precision performance in contrast to other tested graphics cards.

Configuration	$\frac{T_{\text{dp}}}{T_{\text{sp}}}$
i7-2600	1.06
K20	1.57
GTX680	5.24

5.2 Dataset Tests

To ensure good performance in real data mining settings we tested the presented implementation on real datasets and compared them to a CPU based version. The dataset tests have been conducted in single precision only. As datasets we used:

Checker Board Dataset The checker board dataset has 30.000 instances and two dimensions. The function values' heights resemble the pattern of a checker board.

Five Dimensional Astrophysical Dataset The five dimensional astrophysical dataset "Fifth Data Release" (DR5) [AMJ07] has 5 dimensions and 431908 data points. The DR5 data set is a real-world data set, that contains photometric data. Regression analysis allows to predict the spectroscopic red shift of galaxies.

Friedman1 Dataset The Friedman1 dataset [Fri91] is an artificial 10 dimensional dataset. It can feature arbitrarily many data points since it is created by a simple function:

$$y = f(x_1, \ldots, x_{10}) = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon$$

The ten variables $x_1, \ldots, x_{10}$ are all in $[0, 1]$. The latter five serve as noise only. The parameter ϵ serves as additional noise normally distributed with mean 0 and standard deviation 1. We created a training dataset with 50 million points and a testing dataset with 20

million points. The size of the training data is approximately 7.7 gigabytes and the testing data has about 3.1 gigabytes.

5.2.1 Configuration and Results

We basically used the same parameters than for the benchmarks. We did the regression analysis on a GTX680 and for comparison on an Intel®Core i7-2600. All datasets were split into two third training data and one third testing data. We used a sparse grid of level 6 for the checker board and the DR5 dataset and a sparse grid of level 4 for the Friedman1 dataset. The CG solver did 250 iterations. The parameter λ was set to 10^{-6}. We employed streaming for all of the three tests with a chunk size of approximately the grid size. The CPU implementation by Heinecke and Pflüger [HP13] is an optimized version vectorized with Intel's Advanced Vector Extensions (AVX) and parallelized using Open Multi-Processing (OpenMP). The results are shown in Table 3.

Table 3: CUDA and CPU timings of the conducted data set regressions in single precision. For a description of the data sets and the configuration we used see Section 5.2.

CUDA			**CPU**		
C. Board	DR5	Friedman1	C. Board	DR5	Friedman1
0.619667 s	77.9071 s	6 044.16 s	0.664094 s	468.269 s	48 247.5 s

The CUDA version barely pays off with the checker board dataset since it is very small. The little pay off comes from the fact that we used streaming to reduce the initial overhead of copying. Otherwise the CUDA version would perform worse. In the DR5 test the CUDA version out-speeds the CPU-based version by a factor of approximately six. This factor is enlarged once more in the Friedman1 test where the GPU performs approximaltely eight times better than the CPU.

To keep track of which optimization contributes how much to these timings we did timings of the different optimization techniques described in section 4.2.2. The results are shown in Table 4.

Table 4: Per optimization stage timings of the DR5 dataset. The abbreviations: TK stands for template kernels, ST for streaming with a dataset chunk size equaling the grid size and SM for the use of shared memory.

TK	ST	SM	Timing	eff. GFLOPS (sp)
✗	✗	✗	812.242 s	37.7886
✓	✗	✗	156.367 s	196.291
✓	✓	✗	94.2451 s	325.678
✓	✓	✓	77.9071 s	393.976

The Friedman1 test achieved a performance of about 500 effective GFLOPS (sp). This is already near the peak performance of 620 eff. GFLOPS (sp) (see section 5.1). Of course,

one could get even nearer to this bound by increasing the grid size.

6 Discussion and Future Work

First of all, we presented how to derive a SLE from the variational formulation of finding a regression function. The abstract work is done by the SG^{++}toolkit, which allowed us to focus on the main work: parallelizing the matrix vector multiplication. We presented how to map the abstract algorithms to CUDA kernels and several techniques to improve the performance. The approach scales linearly in the amount of data, which enables us to handle datasets that are much bigger than we could handle using classical techniques. In order to allow for bigger datasets than the available memory of a graphics card and to have efficient memory management, streaming has to be used. We pointed out two different variants. As far as the author knows this is the first implementation of regression analysis using sparse grids where the dataset size is allowed to exceed the memory of graphics cards. Finally, we presented benchmarks and tests emphasizing the achieved performance.

One important point is auto tuning. In the benchmarks and tests we tuned the parameters, e.g. number of streams, chunk sizes for streaming, number of threads per block and so on, by hand. There are a few general rules of thumb, but no sophisticated system which determines the best parameters. This could, for example, be done by testing different parameter combinations at compile time.

Our two approaches of using two dimensional block sizes and shared memory are hardly consistent. There might be ways to employ both techniques at once. This would be an important improvement since the grid size is almost always clearly smaller than the dataset size and using shared memory also has its advantages.

The foundations of this paper were laid in the beginning of 2013. At that time, CUDA 4.2 was the most recent release supporting compute capability 2.x. Newer CUDA releases might bring better performance and new features which may be useful in the context of this paper.

In general it can be assumed that good quality CUDA kernels run faster on NVIDIA graphics cards than, for example, OpenCL kernels (which are also able to run on graphics cards). However, other techniques—like OpenCL—have advantages of their own. For example, run-time compiled code. As a matter of fact the stake holders of a project should ponder which technique is better suited for their particular project. Since we relied on CUDA we could make use of several CUDA specific features:

CUDA streams We are able to copy and compute on different data simultaneously.

Pre-compiled kernels This can actually be a disadvantage or lead to more work—see the template kernel we had to use in order to unroll. However, if one eliminates this disadvantage there is no run-time overhead.

Explicit exploitation of the memory hierarchy We did make use of shared memory to efficiently share the same data among a block.

Explicit memory management Manually managing memory can bring better performance than using implicit memory management especially when it comes to using streams.

Atomic-operations They allow for easy merging of results of different threads or blocks. (Although they are not the best choice most of the time.)

Acknowledgments

I want to thank several people for supporting me with this paper. First of all, this paper is based on a student project the author conducted with Roberto Fontanarosa and Tobias Rupp. I also want to thank Patrick Diehl who supervised the project and enabled us to run benchmarks on a NVIDIA K20 of the University of Bonn. Therefore, my gratitude also goes to the University of Bonn. Finally, I want to thank Jun.-Prof. Dirk Pflüger who not only supervised the student project but also gave me counsel when writing this paper.

References

[AMJ07] Adelman-McCarthy, J. K. and Johnston, D. E. The Fifth Data Release of the Sloan Digital Sky Survey. *Astrophysical Journal Supplement Series*, 172:634–644, 2007.

[BG04] Bungartz, H.-J. and Griebel, M. Sparse grids. *Acta Numerica*, 13:1–123, 2004.

[cud12] NVIDIA Corporation. NVIDIA CUDA C Programming Guide, 2012. Version 4.2.

[Fri91] Friedman, J. H. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19, 1991.

[HKP+12] Heinecke, A., Klemm, M., Pflüger, D., Bode, A., and Bungartz, H.-J. Extending a Highly Parallel Data Mining Algorithm to the Intel(R) Many Integrated Core Architecture. In *Euro-Par 2011: Parallel Processing Workshops: Proceedings of the 4th Workshop on UnConventional High Performance Computing 2011 (UCHPC 2011)*, volume 7156/2012 of *Lecture Notes in Computer Science*, pages 375–384, Bordeaux, France, May 2012. Springer.

[HP11] Heinecke, A. and Pflüger, D. Multi- and Many-core Data Mining with Adaptive Sparse Grids. In *Proceedings of the 8th ACM International Conference on Computing Frontiers*, CF '11, pages 29:1–29:10, New York, NY, USA, 2011. ACM.

[HP13] Heinecke, A. and Pflüger, D. Emerging Architectures Enable to Boost Massively Parallel Data Mining Using Adaptive Sparse Grids. *International Journal of Parallel Programming*, 41(3):357–399, 2013.

[Pfl10] Pflüger, D. *Spatially Adaptive Sparse Grids for High-Dimensional Problems*. Verlag Dr. Hut, München, August 2010.

[sgp14] Technische Universität München. SG++ Documentation. website, April 2014. `http://www5.in.tum.de/SGpp/releases/main.html`.

[Syk93] Sykes, A. O. *An Introduction to Regression Analysis*. Coase lecture. Law School, University of Chicago, 1993.

Timing Attack on a Modified Dynamic S-box Implementation of the AES InvSubBytes Operation

Johannes Obermaier, Tobias Laas and Markus Roner

Institute for Security in Information Technology
Technische Universität München
Arcisstr. 21
80333 Munich, Germany
{johannes.obermaier, tobias.laas, markus.roner}@tum.de

Abstract: This paper demonstrates a novel timing attack on a software implementation of the AES decryption algorithm. The implementation was optimized to reduce its code and memory footprint by utilizing an inverse S-box operation which directly calculates the substitution values instead of fetching them from a pre-computed look-up table. This code-size optimized implementation was created as part of a laboratory for which a smart-card emulator was designed and physically tested. Later on, we noticed that the implementation shows a data-dependent execution time for which we developed a novel timing attack. It is based on a timing-model which was derived from an analysis of the implementation. The feasibility of the approach was first proved by a simulation. The subsequent application of the attack on the smart-card emulator in a real setup was successful. This paper describes the analysis done to conduct the attack and emphasizes the dangers of incautiously implemented cryptographic algorithms.

1 Introduction

The Advanced Encryption Standard (AES) is one of the most widely used standards for symmetric encryption employed in today's embedded applications. The widely used cipher implementation, AES-128, uses a key length and block size of 128 bits and is considered to be secure against all cryptanalytic attacks known to this date. The attack published by Bogdanov et al. [BKR11] reduces the computational complexity from 2^{128} to $2^{126.1}$, speeding up the process although still not making it computationally feasible. Despite its mathematical security, hardware and software implementations of AES may be vulnerable to a different class of attacks known as Side-channel Analysis (SCA) attacks. These attacks are based on exploiting the non-intended leakage of information, which occurs during physical operation of the cipher, to retrieve its cryptographic secret. The first publications, which represent a cornerstone for what has become an important field of research in both academia and industry, correspond to Kocher's papers on timing attacks published in 1996 [Koc96] and on power-analysis published in 1999 [KJJ99]. Since then a vast number of publications have dealt with novel SCA attacks and their corresponding countermeasures. A comprehensive study of power-analysis on AES implementations

is given in [MOP07] by Mangard et al.. Timing attacks are usually treated separately to those related to power consumption. In 2004, Bernstein published the first cache-timing attack to break a server using OpenSSL's AES encryption [Ber04]. In that paper, the high-precision time-stamp of the server was used for the attack. In 2005, Osvik et al. [OST05] demonstrated cache-timing attacks on a PC platform kept under control, requiring only a few hundred measurements. In that attack, the timing information was extracted from the CPU's memory cache. While there are approaches to secure speed or memory usage optimized AES implementations against timing attacks on modern 64-bit CPUs [KS09], these methods are not fully applicable to the simpler processors used in embedded applications.

Smart-cards are widely used for applications ranging from authentication and encryption to any kind of payment. Common to all those applications is the requirement for a secure and reliable operation, while preventing the extraction and replication of sensitive data from the card's embedded integrated circuit. In the case of the pay-tv smart-card (emulator) in the laboratory, a key encryption key (KEK) is stored on the smart-card and used for decryption of a session key, which is then used for the video stream decryption. The smart-card emulator uses a standard, low-cost, 8-bit ATmega microcontroller, which was utilized for testing the implementation of the AES-128 and of the timing attack. Since the embedded system on the smart-card operates widely isolated from the environment, not many indications are left to deduce relevant timing information. It is a convenient way to use power traces in order to gain this information. To exclude unwanted influence from power-analysis deviations on the timing attack however, the use of a trigger signal for timekeeping on a pin of the microcontroller is often helpful and sufficient for a proof of concept.

In Sect. 2, we begin with the explanation of an alternative implementation of the `InvSubBytes` function. We continue with its weakness and the attack's theory, together with its simulation, in Sect. 3. The practical implementation of the attack is demonstrated in Sect. 4. Finally, the results are evaluated and countermeasures are discussed in Sect. 5.

2 An alternative approach for `InvSubBytes`

In the course of an AES-128 decryption, the non-linear byte substitution `InvSubBytes` is applied to the internal state of the decryption ten times. `InvSubBytes` consists of two steps: applying an affine transformation and taking the multiplicative inverse in the Galois Field $\mathrm{GF}(2^8)$ [Nat01]. Calculating the multiplicative inverse is computationally expensive. Traditionally, this substitution and its inverse, `SubBytes`, have been implemented on microcontrollers using a substitution table (S-box), as described in the AES specification [Nat01], because that avoids calculating the multiplicative inverse at run time and thereby increases performance. As described in [DR98], "there is some ROM/performance trade-off" in every implementation, i.e. if more execution time is spent, the amount of ROM needed could be reduced. The Rijndael AES Proposal also gives an example for an optimized implementation: "For implementations where RAM is scarce, the Round Keys can be computed on-the-fly [...]" [DR98] . As an answer to those reflections, we decided that in our implementation, not only the round

keys would be computed on-the-fly, but also the entries of the S-boxes. That way more flash memory and RAM is saved because the S-boxes' values do not need to be stored. This is advantageous for devices which are very restricted in memory. The difference is clearly visible. The traditional implementation in C with static S-boxes requires 2.13 kB of flash memory on our microcontroller, but with the on-the-fly calculated S-box and inverse S-box, the demand decreases to only 1.69 kB, which is 20% less. This reduction in size comes with a performance trade-off. The execution time per decryption is increased from 20 ms to 200 ms.

There are several algorithms to find the multiplicative inverse, including the extended Euclidean algorithm and the simple brute-force search we used in our implementation, c.f. Appendix A. There is a special case when computing the multiplicative inverse in `InvSubBytes`, as the inverse of 0x00, which can occur as result of the affine transform, does not exist. In this case, 0x00 has to be returned, as specified in [Nat01]. The algorithms are usually not executed in this case, but instead 0x00 is returned immediately, which causes a shorter execution time. The question arises whether this data-dependent execution time has a negative effect on security and if it can be exploited by a timing attack.

3 Timing attack

3.1 Theory

In this subsection, the theory of a timing attack on the implementation of the AES-128 decryption is described. As already mentioned in the previous section, the brute-force search algorithm for the computation of the multiplicative inverse described in Appendix A has a data-dependent execution time, which will be exploited in this timing attack.

In the attack, random ciphertexts were decrypted on the microcontroller and the plaintext vector p and the execution time t were recorded for each decryption. We know that in its last round

$$p = k \oplus \texttt{InvSubBytes}(s) \ , \tag{1}$$

where k is the key and s is the intermediate state vector before the invocation of `InvSubBytes`. If both the output of `InvSubBytes` and p are known, the key can be computed by rearranging the equation to

$$k = p \oplus \texttt{InvSubBytes}(s) \ . \tag{2}$$

For byte s_i of the state, where $i \in \{0, \ldots, 15\}$, this means that

$$k_i = p_i \oplus \texttt{InvSubBytes}(s_i) \ , \tag{3}$$

as `InvSubBytes` and $\oplus$ are byte-wise operations. As p_i is recorded, the idea of the following attack is to derive information about the output of $\texttt{InvSubBytes}(s_i)$ given the execution time t, in order to find out k_i.

Table 1: Probabilities for the number of zeros (a) and votes for the attack.

Zeros (a)	Probability	vote for value for p_i	vote for other values
0	53.5%	$\ln(1/10000)$	$\ln(9999/(10000 \cdot 255)$
1	33.5%	$\ln(1/160)$	$\ln(159/(160 \cdot 255))$
2	10.5%	$\ln(2/160)$	$\ln(158/(160 \cdot 255))$
3	2.2%	$\ln(3/160)$	$\ln(157/(160 \cdot 255))$
4	0.3%	$\ln(4/160)$	$\ln(156/(160 \cdot 255))$
5	0.04%	$\ln(5/160)$	$\ln(155/(160 \cdot 255))$
6	$< 0.01\%$	$\ln(6/160)$	$\ln(154/(160 \cdot 255))$

For one full AES-128 decryption, `InvSubBytes` is called ten times, excluding the key expansion, which does not induce timing differences, because the key is assumed to be constant. Each of those calls applies the transformation to each of the 16 bytes of the internal state, i.e. there are 160 calls to `GF2_8InvMult`, which performs the inversions in $\mathrm{GF}(2^8)$. If the byte at the input of `GF2_8InvMult` is zero, significantly less time is needed to compute its multiplicative inverse. In this case, the output of `InvSubBytes`(s_i) is zero and $k_i = p_i$. The remaining operations of the implementation have significantly less variation in execution time. Therefore the attack focuses on the number of zeros.

First, the execution time is analyzed and the number of zeros occurring at the input of `GF2_8InvMult`, a, is estimated. More about that in Sect. 4.

For random data which is decrypted, the probability for a zeros is

$$P_A(a) = \binom{160}{a} (1/256)^a (255/256)^{160-a} \ , \tag{4}$$

under the assumption that the zeros are uniformly distributed over all invocations of `GF2_8InvMult`. The evaluated probabilities can be seen in Table 1.

It is unknown in which round and at which byte the zeros occurred, but the probability of n zeros appearing in the last round of the AES decryption, i.e. in the last invocation of `InvSubBytes`, given that there are a zeros in the whole decryption, is

$$P_{N|A}(n \mid a) = \frac{\binom{16}{n}\binom{144}{a-n}}{\binom{160}{a}} \ . \tag{5}$$

When the information acquired from the measurements is combined, the zeros arising in the last round provide so-called correct information about the key and the zeros arising in the other rounds provide so-called misleading information. They cannot be separated, since only the whole execution time is measured. Using equations (4) and (5), the probability of correct information can be computed by

$$P_{\mathrm{correct}}(a) = \sum_{n=\min(a,1)}^{a} P_{N|A}(n \mid a) P_A(a) \ , \tag{6}$$

and the probability of misleading information by

$$P_{\text{misleading}}(a) = P_A(a) - P_{\text{correct}}(a) = \begin{cases} P_{N|A}(0 \mid a) P_A(a) & a \neq 0 \\ 0 & a = 0 \end{cases} . \tag{7}$$

$a = 0$ is special because in that case, the bytes p_i in the plaintext cannot be the respective key bytes k_i, because then $\texttt{InvSubBytes}(s_i) \neq 0$ in (3), i.e. the information is correct all the time. See Fig. 1 for a plot of the probability of correct and misleading information in relation to a.

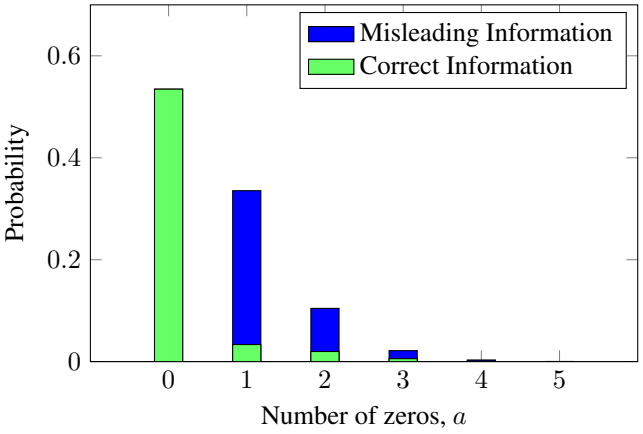


Figure 1: Probabilities of correct and misleading information in relation to the number of zeros, a. They are calculated using (6) and (7).

In order to combine the information acquired with the N individual measurements, a maximum likelihood (ML) estimator with the likelihood function L_{k_i} has been used. In order to reduce the arithmetic range during the computation, the log-likelihood function using the natural logarithm $\ln(\cdot)$ has been taken and is computed as follows

$$\ln\left(L_{k_i}(j)\right) = \ln\left(\sqrt[N]{\prod_{x=1}^{N} P_{k_i,x}(j)}\right) = 1/N \sum_{x=1}^{N} \ln\left(P_{k_i,x}(j)\right) \quad , \\ \forall i \in \{0,\ldots,15\}, \;\; j \in \{0,\ldots,255\} \quad , \tag{8}$$

where $P_{k_i,x}(j)$ is the probability that key byte $k_i = j$ in measurement x. A normalization by N has been added. The ML estimator then is

$$\hat{k}_i = \underset{j \in \{0,\ldots,255\}}{\arg\max} \; \{\ln L_{k_i}(j)\} \quad . \tag{9}$$

In the implementation, the likelihoods are stored in a likelihood matrix $\boldsymbol{V} \in \mathbb{R}^{16\times 256}$, where $v_{i+1,j+1}$ stores the likelihood for $k_i = j$. For each decryption $x \in \{1,\ldots,N\}$,

the likelihood matrix $\boldsymbol{V}$, which is initialized with zeros, is updated by adding the "vote" $\ln\left(P_{k_i,x}(j)\right)$ to the entry $v_{i+1,j+1}$. That way, after summing up the votes for all measurements, $v_{i+1,j+1} = N \ln\left(P_{k_i}(j)\right)$, c.f. (8). As a last step, each $v_{i+1,j+1}$ has to be replaced by $\exp(v_{i+1,j+1}/N)$ such that $\boldsymbol{V}$ stores the log-likelihood at the end of the estimation.

If the timing information is not considered, the probability that a single byte of the key is correct is $1/256$, because there are 256 possibilities, as usual. But if exactly one zero occurred, the probability is $1/160$ since there we assume that the zeros are distributed uniformly over the 160 inversions. Consequently, for a zeros the probability is $a/160$ and the probability for each of the other possible values is $(160 - a)/(160 \cdot 255)$, again since we assume a uniform distribution. Taking the natural logarithm leads to the voting table in Table 1 with one exception: if no zero occurred, the probability that the current bytes in the plaintext are the respective key bytes is zero. The likelihood of those bytes then is also zero, i.e. in practice a single measurement error can prevent finding the correct key, c.f. also Sect. 4. Therefore in that case, the probability for the p_i has been increased to $1/10000$ and the probability for the other values has been decreased to $(1-1/10000)/255$ instead of $1/255$.

After a certain number of measurements, there should only be one entry per key byte in the matrix which has a significantly larger vote than the other bytes, and these entries form the correct key.

3.2 Attack simulation

To test and verify the attack's theory, a simulator which is able to emulate the behavior of the AES implementation on the microcontroller was implemented in C++. If a zero is processed on the microcontroller, the algorithm skips several hundreds of cycles, which is easily measurable. The simulator therefore observes the input of `GF2_8InvMult` and counts the number of zeros, a, during decryption. Random input data is decrypted until the desired number of decryptions has been simulated and then the attack is performed. This enables the simulation of thousands of measurements in some seconds without the need for the hardware setup.

The success of the attack is highly dependent on statistical processes and so there is no fixed number of required measurements for which the attack will always be successful. The success ratio was estimated by the simulation of 4000 independent attacks, in which the number of measurements was increased in small steps from 0 to 6000. The result is depicted in Fig. 2. For less than 2000 measurements, there is almost no chance to find the correct key. The next 1000 additional measurements show the highest effect. At around 2750 measurements, the success ratio is about $50\,\%$ and at 3000, the success ratio is about $66\,\%$. It reaches $90\,\%$ at around 3550, $99\,\%$ at 4600 and finally approaches $100\,\%$. The total time needed for such an attack is only several minutes. Most of the time is required for taking the measurements. The attack itself can be computed in a few seconds on a standard computer. Eventually the attack requires neither very special equipment, nor excessive computational power and therefore can easily be performed.

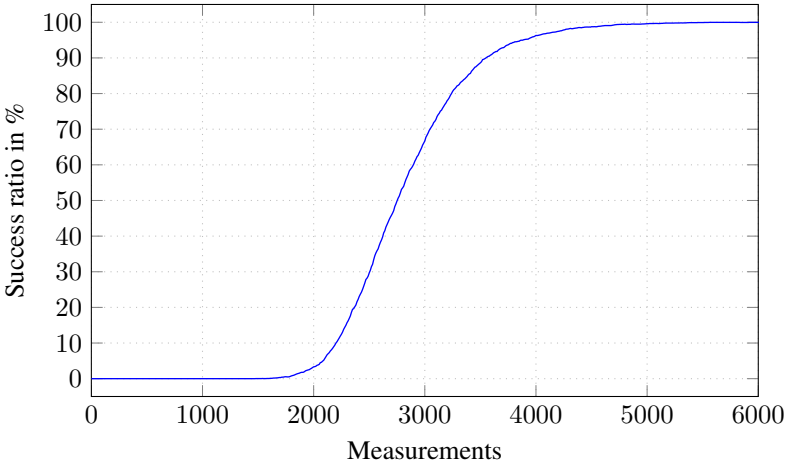


Figure 2: Success ratio in relation to the number of measurements. The curve was determined using a simulation model of the attack and executing it 4000 times with 0 to 6000 measurements each.

4 Practical implementation of the attack

The alternative implementation of the AES algorithm was loaded into the smart-card emulator, c.f. Sect. 1, so that the implementation could be tested in a real environment. The emulator was inserted into a smart-card reader, which was connected to a computer. An inexpensive USB oscilloscope was used to measure the timing, which was forwarded to the computer. The setup is depicted in Fig. 3.

The execution time is measurable in various ways, e.g. through power consumption, inter-

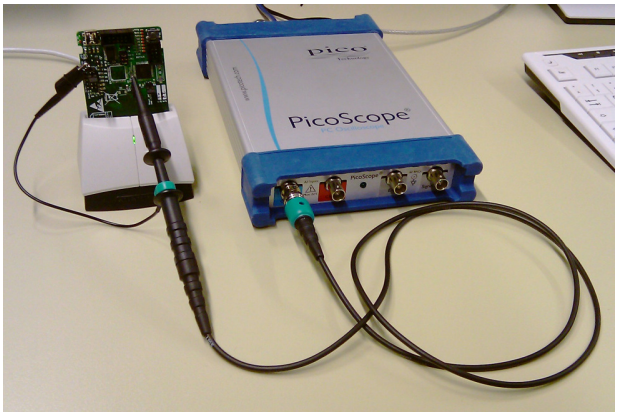


Figure 3: Measurement setup for timing attacks. The smart-card emulator is inserted into the card reader. A probe of the oscilloscope is connected to the microcontroller in the smart-card emulator. The attack is controlled by the attached PC.

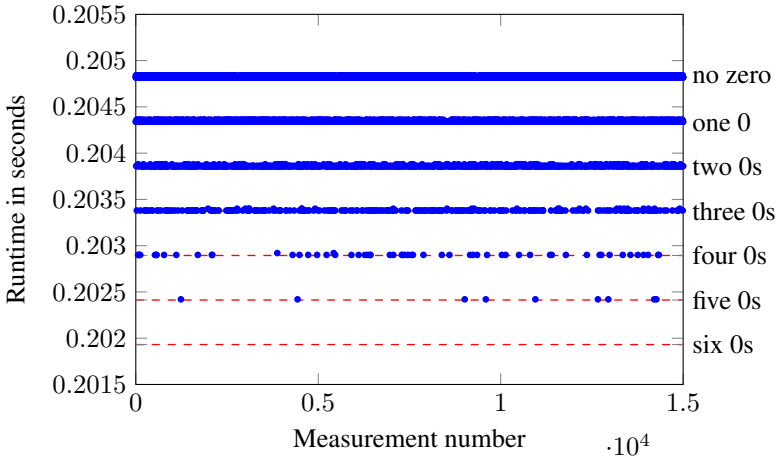


Figure 4: Input time distribution (dots) and expected run times (dashed lines). The number of points in each line complies with the predicted distribution. The dashed lines in the center of each group were automatically determined by the attack script.

face communication or dedicated pins, which indicate ongoing decryption. It is assumed to be trivial for an attacker to obtain this information through these or other methods. In our test, a microcontroller's pin was used, since it is the simplest solution and is sufficient for testing. The oscilloscope is able to determine the execution time very accurately, down to single clock cycles. The timing deviations are in the range of $0.5\,\text{ms}$ per zero at a microcontroller clock frequency of $5\,\text{MHz}$. Therefore a sample rate as low as $10\,\text{kS/s}$ is enough for these purposes and has the advantage of limiting the amount of measured data, making it easier to process. Both, the timing information and the plaintext are stored and preprocessed by a MATLAB script. Next, the number of zeros has to be estimated. The algorithm goes through the timing data and searches for the execution time, which was exceeded by more than $50\,\%$ of all measurements. The same is done for $83\,\%$. These are the cases, including some safety margin, where no zeros and one zero occurred. It gives a very good estimate for the time-saving due to each zero. The execution time for two, three and more zeros is derived from the result. There are no fixed timing values, the information is extracted by only using the expected execution time distribution. This brings

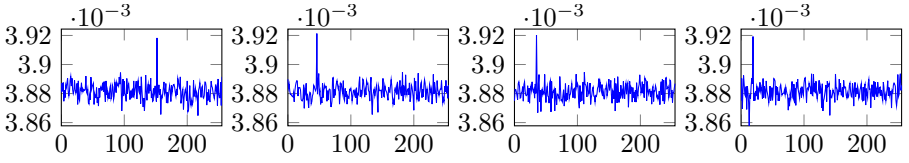


Figure 5: 4 rows of the likelihood matrix $\boldsymbol{V}$, representing the log-likelihood for each key byte. The other 12 rows look similar, only the peak position changes. The x-axes show their value, the y-axes their log-likelihood values. Each subplot shows exactly one significantly higher positive peak, which forms the correct estimated key after 15000 measurements.

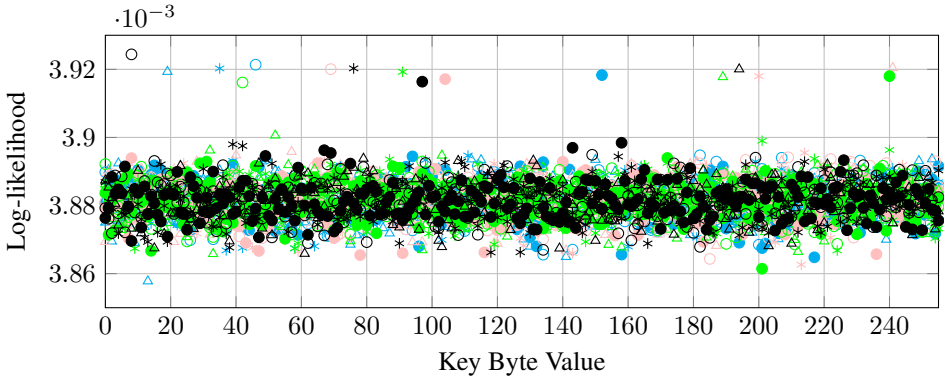


Figure 6: The full likelihood matrix $\boldsymbol{V}$. Each symbol type and shading depicts a different key byte. The incorrect estimates are concentrated in the lower plot region, the 16 correct estimates are in the upper region. Both groups separate with an increasing number of measurements. For this plot, 15000 measurements were executed.

the advantage that the script works with any microcontroller, compiler and clock speed, as long as the differences are measurable.

In order to demonstrate the performance of the estimation methods, a large measurement with 15,000 decryptions was performed. The methods work well and the result can be seen in Fig. 4. Each measurement is represented by a dot and the extracted expected execution time is shown by the dashed lines. According to the simulation in Subsect. 3.2, they contain more than enough information to extract the key. The number of zeros from the timing measurement is taken and the likelihood matrix $\boldsymbol{V}$ is computed as described in Subsect. 3.1 and plotted in two different ways in Fig. 5 and 6. One value per row in $\boldsymbol{V}$, i.e. one for each key byte, is much higher than all other 255 values in this row, as depicted in Fig. 5. It can be clearly seen that after a high number of measurements, two groups have formed in Fig. 6. The high-density area around $3.88 \cdot 10^{-3}$ contains all incorrect key byte values. In contrast to that, the 16 correct key bytes have a higher log-likelihood value of around $3.92 \cdot 10^{-3}$. At the start of the estimation, all likelihoods are the same. As the number of processed measurements increases, the correct key bytes move upwards, while the incorrect ones move downwards, on average. Finally, after a sufficient number of measurements, the correct and incorrect bytes are separated from each other and the key can be extracted.

5 Evaluation and Prospects

5.1 Countermeasures, Attack Variations and Influence on DPA Attacks

In order to defeat the timing attack, the execution time of `GF2_8InvMult` must be constant. This can be solved by modifying `GF2_8InvMult` in such a way that it does not

immediately return zero and instead uses random data whose result is discarded. The algorithm might also be replaced by a more sophisticated one, which has constant run time, but this might diminish the savings in code size.

The attack can also be used on the AES encryption, where the initial and first round are attacked. The input plaintext has to be used instead of the output. The remaining part of the attack works the same way as for the decryption.

Another variation would be a chosen plaintext attack on the encryption or decryption. If such an attack is possible, the attacker can actively guess key bytes. The number of measurements can be reduced with this method, but it requires a more sophisticated attack implementation or the manual input of data.

The implementation was also tested with a DPA attack, implemented as described in [MOP07]. It was expected that the non-constant run time leads to an increased effort for an attacker, due to trace misalignment. Nevertheless, the method was not believed to be a very secure protection against such attacks. These statements were verified experimentally and confirmed the assumptions. For normal and unprotected static S-boxes, the number of required traces was around 400. With the dynamic S-boxes and no trace alignment, even more than 1000 traces were not sufficient. After alignment of the traces, the attack was successful with around 500 traces. This shows that the effort for an DPA is increased, but the method does not provide sufficient protection.

5.2 Conclusion

We have presented a novel timing attack on our implementation of the AES decryption algorithm. As demonstrated, an attacker can find out the secret key in a negligible amount of time with an inexpensive USB oscilloscope and a common PC. Thus, the attack can be performed with very low costs. It has to be made sure that a data-dependent execution time like the one in our implementation is not even present in systems with low security requirements. But security relevant aspects such as a data-dependent execution time are missed easily. The attack shows that the way of implementing the S-boxes is critical for a secure AES implementation. But since most implementations use fixed tables, they are immune to the attack. Nevertheless, the concepts of our timing attack are useful for attacks on different devices, implementations or algorithms.

Acknowledgement

Thanks to Prof. Sigl and Oscar Guillen Hernandez, who offered and supervised the lab course, during which we developed the attack, for encouraging us to write the paper and for their valuable review and feedback.

References

[Ber04] D. J. Bernstein. Cache-timing attacks on AES, 2004. http://cr.yp.to/papers.html#cachetiming.

[BKR11] A. Bogdanov, D. Khovratovich, and C. Rechberger. Biclique Cryptanalysis of the Full AES. Cryptology ePrint Archive, Report 2011/449, 2011. http://eprint.iacr.org/.

[DR98] J. Daemen and V. Rijmen. AES Proposal: Rijndael, 1998.

[KJJ99] P. Kocher, J. Jaffe, and B. Jun. Differential Power Analysis. In M. Wiener, editor, *Advances in Cryptology — CRYPTO' 99*, volume 1666 of *Lecture Notes in Computer Science*, pages 388–397. Springer Berlin Heidelberg, 1999.

[Koc96] P. C. Kocher. Timing Attacks on Implementations of Diffie-Hellman, RSA, DSS, and Other Systems. In N. Koblitz, editor, *Advances in Cryptology — CRYPTO '96*, volume 1109 of *Lecture Notes in Computer Science*, pages 104–113. Springer Berlin Heidelberg, 1996.

[KS09] E. Käsper and P. Schwabe. Faster and Timing-Attack Resistant AES-GCM. In C. Clavier and K. Gaj, editors, *Cryptographic Hardware and Embedded Systems - CHES 2009*, volume 5747 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg, 2009.

[LN94] R. Lidl and H. Niederreiter. *Introduction to Finite Fields and Their Applications*. Cambridge University Press, 1994.

[MOP07] S. Mangard, E. Oswald, and T. Popp. *Power Analysis Attacks - Revealing the Secrets of Smart Cards*. Springer US, 2007.

[Nat01] National Institute of Standards and Technology. FIPS 197 - Advanced Encryption Standard, 2001.

[OST05] D. A. Osvik, A. Shamir, and E. Tromer. Cache attacks and Countermeasures: the Case of AES. Cryptology ePrint Archive, Report 2005/271, 2005. http://eprint.iacr.org/.

A Implementation of `GF2_8InvMult`

`GF2_8InvMult`, shown in Listing 1, computes the multiplicative inverse v' of v over $\mathrm{GF}(2^8)$, i.e. the solution to $vv' = 1$. It has been implemented using 8-bit operations only, as an 8-bit controller has been used. `uint8_t` denotes an 8-bit unsigned integer type.

If $v =$ 0x00, 0x00 is returned according to [Nat01], c.f. lines 10 and 11. The idea of the algorithm is to find a solution by expressing $vv' = 1$ as a sum, as follows:
$1 = vv' = g^{\log_g(vv')} = g^{p+p'}$, where $p = \log_g v,\ p' = \log_g v'$ and g is a base.

The rather simple and space-saving algorithm is composed of three steps. The first is to calculate $p = \log_g(v)$. There exist bases g on $\mathrm{GF}(2^8)$, for which a p can be found for every $v \neq 0$, i.e. its powers result in every element of the field, except 0. Such a g is called primitive element [LN94]. The brute-force search starts with $p = 0$, calculates g^p and increases p until $g^p = v$, c.f. the loop in lines 12 to 16. In each iteration of the loop g^p is multiplied by g. As the primitive element is of the order 255, the search will terminate after at most 255 iterations.

```
1  inline uint8_t Mult3GF2_8(uint8_t gP) /* multiply by 0x03 (x + 1) */
2  {
3       uint8_t prod = gP << 1; /* multiply by 0x02 (x) */
4       if (gP & 0x80) /* reduce modulo (x^8 + x^4 + x^3 + x + 1) */
5            prod ^= 0x1b;
6       return gP ^ prod; /* calc. result (x) + (1) */
7  }
8  uint8_t GF2_8InvMult(uint8_t v)
9  {
10      if (v == 0) /* multiplicative inverse of 0 is defined as 0 */
11           return 0;
12      uint8_t p = 0, gP = 1;
13      while (gP != v) { /* calculate 0x03^p = v, solve for p */
14           gP = Mult3GF2_8(gP);
15           p++;
16      } /* calculate the multiplicative inverse gP = 0x03^(255-p) */
17      for (gP = 1; p < 255; p++)
18           gP = Mult3GF2_8(gP);
19      return gP;
20 }
```

Listing 1: Source Code of `GF2_8InvMult` for the calculation of the multiplicative inverse

The second step is to calculate p'. g^0 has to equal $g^{p+p'}$ in $\mathrm{GF}(2^8)$. As there are only 255 different elements in the powers of g, any sum of p and p' which is a multiple of 0xFF results in the same value as 0x00. In order to find p', $0x00 = (p + p') \mod 0xFF$ should be solved by $p' = 0xFF - p$. In the code, this step has been combined with the third step, the calculation of the power $v' = g^{p'}$, c.f. lines 16 to 18. Calculating p' is omitted, instead the correct number of successive multiplications by g to compute v', $255 - p$ are executed anyway by the careful design of the loop. The result v', the inverse of v, is then returned.

Please note that the elements of $\mathrm{GF}(2^8)$ are polynomials which have an equivalent notation as an 8-bit hexadecimal number, e.g. $x + 1 = 0x03$. Choosing the primitive element $g = x + 1 = 0x03$ is beneficial, because the multiplication by it can be implemented efficiently. In `Mult3GF2_8` (lines 1 to 7), it is split into multiplying by 1, which is redundant, and by x. For this multiplication, `gP` is shifted left by one bit. If the MSB of `gP` is set, the result of the multiplication cannot be represented in $\mathrm{GF}(2^8)$ and the result still needs to be reduced modulo the polynomial $x^8 + x^4 + x^3 + x + 1$, which in this case can be simplified to subtracting this polynomial, which is equivalent to xoring the result with 0x1b [Nat01]. In the last step, `gP` and the product are added.

Regarding the security of `GF2_8InvMult`, it has to be analyzed whether the search and its intentionally simple implementation leaks any critical information. The number of iterations of the first loop depends on the input data, but together with the second loop, the total number of iterations is 255. Nevertheless, there is some minor run time difference due to the if statement, and the compiler may also implement the second loop differently than the first one. The source of the largest timing variation can easily be seen in lines 10 and 11. If the input value is zero, the function returns zero immediately and does not execute the loops, which makes the implementation vulnerable to a timing attack, as shown in Sect. 3.

Analysis of Crash Simulation Data using Spectral Embedding with Histogram Distances

Anna-Luisa Schwartz

Institut für Numerische Simulation
Universität Bonn
Wegelerstr. 6
53115 Bonn
luisa.schwartz@uni-bonn.de

Abstract: Finite Element simulation of crash tests in the car industry generates huge amounts of high-dimensional numerical data. Methods from Machine Learning, especially from Dimensionality Reduction, can assist in analyzing and evaluating this data efficiently. Here we present a method that performs a two step dimensionality reduction in a novel manner: First the simulation data is represented as (normalized) histograms, then embedded into a low dimensional space using histogram distances and the nonlinear method of Spectral Embedding/Diffusion Maps, thus enabling a much easier data analysis. In particular, this method solves the problem of comparing simulation data with small changes in the Finite Element grids due to variations of geometry or unequally fine grid structures.

1 Introduction

Numerical simulation of car crash tests has evolved since the 1980's to be an integral part of the development process for new vehicles. Due to the rapid growth in computational capacity, enabling the generation of more and much more detailed simulations, the size of the emerging simulation data increased enormously. Today, for one simulation run data is saved for more than a million finite element nodes at several hundred time steps; tests to decide on small parameter variations (plate thickness, positions for drillings, material variations) require bundles of several thousand simulation runs.

Conventionally, evaluation of this data is performed by an engineer examining several 3D-visualisations of the simulation runs simultaneously and grouping them according to observed effects. Accelerating and simplifying this difficult and time-consuming analysis through the use of Machine Learning methods is a subject of recent research [Boh13].

As the data is highly complex with dimensionality of an order of 10^8, in particular exceeding the number of data points by a factor of hundred thousand, methods of dimensionality reduction need to be applied: Since all simulation runs in a bundle concern the same car model at the same stage of development, the data exhibits many redundancies and it is to be expected that a lower-dimensional embedding will still be able to parameterize the data

according to its so-called intrinsic geometry. A 2- or 3-dimensional embedding can then present a visual overview over the relations between simulation runs in a bundle and thus be an important tool for the development engineer.

In this paper we focus on a selected method of nonlinear dimensionality reduction: Spectral Embedding, in particular Diffusion Maps, presented 2006 by Coifman and Lafon [Coi06] and already applied successfully for simulation data analysis [Erb07, Boh13]. This method is based on the use of a kernel function describing the similarities of different data points. Commonly the similarity function is derived from distance functions on the data set. So far, when applying Diffusion Maps on car data, one regards the Euclidean distance between so-called displacement vectors, containing the absolute values of the displacements at every finite element node [Boh13, Iza14]. This approach requires identical dimensions of all displacement vectors considered, meaning identical numbers of FE nodes for every simulation run. In practice, this is not always the case. Often simulations are ran to compare the effect of varying certain structural parts, thus leading to simulation runs with different geometries and consequently different finite element meshes. To compare these simulation runs, projections on auxiliary meshes can be employed, which are however computationally costly and might not be very accurate. An alternative choice of distance functions seems promising to overcome this problem.

Here, we will for the first time investigate the approach to calculate distances between simulation runs via a usage of histograms of the displacements. Representing the high-dimensional data as histograms with fewer bins than the original dimensions means a dimensionality reduction already in a preprocessing step, which will then be further refined by the application of Diffusion Maps, using an appropriate histogram distance.

The theoretical foundations of Diffusion Maps will be presented in Section 2 of this paper, as well as a short review of histogram distances. In Section 3 we introduce our proposed method and the algorithm, followed by the results obtained by applying this approach to several sets of industrial data and a discussion in Section 4.

2 Fundamental Concepts

In the following, we will refer to the data set as $X = \{x_1, x_2, \ldots, x_m\} \subset \mathbb{R}^n$, where each of the m data points represents a full simulation run. Dimensionality reduction means the search for a map (the so-called embedding) $f : X \longrightarrow \mathbb{R}^p$ with $p \ll n$, preserving intrinsic information on the data (mostly geometric information, such as certain distances). If we assume that the data set X lies on a nonlinear manifold, we also need to apply nonlinear methods to obtain a suitable embedding. Spectral Embedding methods like Diffusion Maps find such an embedding using a kernel matrix describing local similarities between data points.

2.1 Review of Diffusion Maps

Diffusion Maps [Coi06] are based on the notion of a so-called diffusion distance, motivated as follows: Imagine the data set X to be the node set of a complete weighted graph (G, ω) with $G = (X, X \times X)$. The edge weights are assigned depending on the local similarity of the adjoint nodes:

$$\omega : X \times X \to \mathbb{R}, \omega((x_i, x_j)) = k(x_i, x_j),$$

with k being a kernel function. Here one normally uses the Gaussian kernel, $k_\varepsilon(x_i, x_j) := \exp\left(\frac{d(x_i,x_j)^2}{\varepsilon}\right)$ with a scaling parameter $\varepsilon > 0$ and d the Euclidean distance; for our application however we will apply the Gaussian kernel with an application-specific distance function $d : X \times X \to \mathbb{R}$. By normalizing the edge weights with the degrees of the correspondent nodes we obtain a Random Walk on X with transition probabilities $p(x_i, x_j) = \omega((x_i, x_j))/\sum_{y \in X} \omega((x_i, y))$.
If the points are similar according to the chosen distance function, the respective transition probabilities will be high, for dissimilar points they will be low. Running the Random Walk forward in time, we obtain information about regions of high affinity within the data set, leading naturally to a definition of the diffusion distance, depending on a time parameter $t \in \mathbb{R}$:

$$D_t(x_i, x_j)^2 := \sum_{y \in X} (p_t(x_i, y) - p_t(y, x_j))^2 \frac{1}{\pi(y)}$$

with π denoting the stationary distribution of the Markov Chain and $p_t(x_i, x_j)$ the probability of transition from x_i to x_j in t steps [Coi06]. Thus, the diffusion distance sums over all possible connections between two points in t time steps, intuitively leading to a high robustness to noise.

To approximate this distance in practice, we consider the spectral decomposition of the operator p_t or, equivalently, of the t-th power of the transition matrix $\mathbf{P} = (p(x_i, x_j))_{ij}$. As $\mathbf{P}$ is not symmetric, we conjugate it by the diagonal degree matrix $\mathbf{D}$ with $\mathbf{D}_{ii} := \sum_j k(x_i, x_j)$.

The new matrix $\mathbf{A} := \mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}$ is symmetric and shares the spectrum with $\mathbf{P}$. We know from the Perron-Frobenius theorem that all eigenvalues λ_i are real, and $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$.

We denote the eigenvectors of $\mathbf{A}$ by $\{v_j\}$. Right and left eigenvectors $\{\phi_j\}$ and $\{\psi_j\}$ of $\mathbf{P}$ are obtained by conjugation:

$$\phi_j = \mathbf{v}_j \mathbf{D}^{\frac{1}{2}}, \qquad \psi_j = \mathbf{v}_j \mathbf{D}^{-\frac{1}{2}}.$$

Hence we can express the diffusion distance at time t by the spectral decomposition of $\mathbf{P}$ [Coi06]:

$$D_t(x_i, x_j)^2 = \sum_{l=2}^{n} \lambda_l^{2t} (\psi_l(x_i) - \psi_l(x_j))^2.$$[1]

[1] Remark: $\psi(x)$ denotes for an $x \in X$ the i-th entry of $\psi \in \mathbb{R}^m$, with i the index of $x = x_i$ in the indexed set X

The term for $l = 1$ is left out, as the eigenvector ψ_1 belonging to the eigenvalue $\lambda_1 = 1$ is constant.

Due to the non-increasing property of the spectrum we can approximate the diffusion distance by using only the first $p < n$ terms:

$$D_{t,p}(x,y)^2 := \sum_{l=2}^{p+1} \lambda_l^{2t}(\psi_l(x) - \psi_l(y))^2$$

Thus we can achieve a dimensionality reduction by embedding the data set into $\mathbb{R}^p$ by the diffusion maps

$$\Psi_{t,p} : \mathcal{Y} \longrightarrow \mathbb{R}^p, \qquad x \mapsto (\lambda_2^t\psi_2(x), \lambda_3^t\psi_3(x), \ldots, \lambda_{p+1}^t\psi_{p+1}(x))^T, \tag{1}$$

preserving the local diffusion distances in the originating space X up to an error depending on the ratio $\frac{\lambda_{p+1}^t}{\lambda_2^t}$.

To take out statistically unsuitable side effects, we can replace isotropic kernel functions such as k_ε by anisotropic ones [Nad08]. Here, instead of k_ε we applied the anisotropic Gaussian kernel, $\tilde{k}_\varepsilon(x,y) := k_\varepsilon / \sum_{z\in X} k_\varepsilon(x,z) \sum_{z\in X} k_\varepsilon(y,z)$, where the terms in the denominator are used as an approximation of the probability density by which the dataset X was drawn from the underlying manifold.

2.2 Review of Histogram Distances

For a consistent notation, let us formalize the notion of a (multidimensional) histogram:

Definition 1 (Histogram)**.** *For a finite set $\mathcal{I} \subset \mathbb{N}^d$ and a map $h : \mathcal{I} \longrightarrow \mathbb{R}_{\geq 0}$ we call the set $H = (h(\mathbf{i}))_{\mathbf{i}\in\mathcal{I}}$ a* histogram*. The elements of $\mathcal{I}$ represent* bins*, the assigned values $h(\mathbf{i})$ are the respective* weights*.*

To compare histograms, one can distinguish between two main classes of distances: *bin-to-bin-distances* and *cross-bin-distances*. The former only take into account the weights of matching bins, meaning they can be written as:

$$d(H_1, H_2) = \sum_{\mathbf{i}\in\mathcal{I}} f(h_1(\mathbf{i}), h_2(\mathbf{i})) \text{ for histograms } H_1 = (h_1(\mathbf{i}))_{\mathbf{i}\in\mathcal{I}} \text{ and } H_2 = (h_2(\mathbf{i}))_{\mathbf{i}\in\mathcal{I}}$$

with a suitable function $f : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}_{\geq 0}$. In particular this means that the histogram topology is completely disregarded.

Cross-bin-distances on the other hand take into account neighboring or even all bins as well. A discrete *ground distance* $d_{\mathbf{ij}}$ is utilized to define distances between the bins represented by $\mathbf{i}$ and $\mathbf{j}$. This facilitates the formulation of distances that correspond well with human perception and at the same time are more robust with respect to the choice of bin ranges. However, cross-bin-distances are obviously more computationally expensive.

Popular examples for histogram distances include the L_1- and L_2-distances of the histograms' weight vectors (bin-to-bin-distances). Note that there also exists a cross-bin histogram distance based on diffusion distances [Lin06], which undertakes the diffusion approach from a different angle than Diffusion Maps and is not further considered here. In this paper, we focus on the following distances, both widely used in image processing:

χ^2-distance The χ^2-distance is a bin-to-bin-distance, following the intuition that differences between 'large' bins are less significant than those between 'small' bins [Pel10]. It is based on the χ^2-test used in statics and is defined as:

$$\chi^2(H_1, H_2) = \frac{1}{2} \sum_{\mathbf{i} \in \mathcal{I}} \frac{(h_1(\mathbf{i}) - h_2(\mathbf{i}))^2}{h_1(\mathbf{i}) + h_2(\mathbf{i})}$$

Earth-Mover's-Distance (EMD) EMD [Rub00] is a cross-bin-distance, its name originating from the analogy to imagine one histogram as a set of earth heaps and the other as a set of holes, with a ground distance assigned to every pair of bins. The EMD then is the minimum cost of moving the earth into the holes and is obtained by solving a discrete minimum cost flow problem.
For the rather technical precise definition and set-up, we refer to [Rub00].

3 Method

We propose the following analysis procedure to compare the simulation data (cf. [Iza14]):

1. Extraction of displacement data of charasteric integral parts from the simulation data.
2. Preprocessing of data. *Here*: representation of the n_i-dimensional displacement vectors $\mathbf{v}_i$ as one-dimensional histograms $\left(h^i(j)\right)_{j=1}^{n_{\text{bins}}}$.
3. Creation of a kernel k_ε, utilizing a suitable histogram distance function as shortly discussed in Section 2.2.
4. Dimensionality reduction by Spectral Embedding. *Here*: Use of Diffusion Maps (as in Equation (1)).
5. Postprocessing and analysis of embedded data (e.g. by application of clustering algorithms). *Here*: 2D-visualisation of embedding space.

The histogram representation in step 2 allows us to apply Diffusion Maps also to compare simulation runs with varying geometries. It needs to be noted that this step represents a first layer of dimensionality reduction, where some information is lost – in our use case, this especially means that information about the positions of the displacements is discarded and only the extent of the deformation is stored.

Input: List *filelist* of m simulation data sets $\{x_1, \ldots, x_m\}$, parameter n_{bins}, time step t
Output: Histograms $\{\left(h^1(j)\right)_{j=1}^{n_{\text{bins}}}, \ldots, (h^m(j))_{j=1}^{n_{\text{bins}}}\}$, stored as a matrix $\mathbf{W} \in \mathbb{R}^{n_{\text{bins}} \times m}$
where $w_{kl} = \left(h^l(k)\right)$
Initialize displacement matrix $\mathbf{V} \in \mathbb{R}^{n \times m}$ and histogram matrix $\mathbf{W} \in \mathbb{R}^{n_{\text{bins}} \times m}$;
$i = 1$;
for *file* $\in$ *filelist* **do**
  Read and store displacement vectors in x, y, and z-direction at time step t
  $\mathbf{v}^x, \mathbf{v}^y, \mathbf{v}^z \in \mathbb{R}^{n_i}$ from *file* (with n_i number of finite element nodes in model);
  Calculate absolute displacement vector $\mathbf{v}^{\text{abs}}$: $\forall k : v_k^{\text{abs}} := \sqrt{(v_k^x)^2 + (v_k^y)^2 + (v_k^z)^2}$;
  $\mathbf{V}[:, i] := \mathbf{v}^{\text{abs}}$;
  $i{+}{+}$;
end
Determine minimum $\min_{\text{glob}}$ and maximum $\max_{\text{glob}}$ of entries in displacement matrix $\mathbf{V}$;
Define n_{bins} identically large classes for histograms (according to $\min_{\text{glob}}$ and $\max_{\text{glob}}$);
for $k \in \{1, 2, \ldots, n_{bins}\}$ **do**
  Generate a histogram $(h(j))_{j=1}^{n_{\text{bins}}}$ of displacement vector $\mathbf{V}[:, k]$ according to the defined class partitioning;
  Normalize histogram to total weight of 1;
  Store histogram weight vector in $\mathbf{W}[:, k]$;
end

Algorithm 1: Extraction and Preprocessing

Input: Matrix $\mathbf{W} \in \mathbb{R}^{n_{\text{bins}} \times m}$ of histogram weight vectors, originating from Algorithm 1;
(histogram-) distance[a] d; scaling parameter ε for Gaussian kernel, target dimension p
Output: Diffusion coordinates of data for the embedding into $\mathbb{R}^p$
Initialize distance matrix $\mathbf{D} \in \mathbb{R}^{m \times m}$;
for $i \in \{1, 2, \ldots, m\}$ **do**
  for $j \in \{1, 2, \ldots, m\}$ **do**
    $\mathbf{D}[i, j] := d_{ij} := d(\mathbf{W}[:, i], \mathbf{W}[:, j])$;
  end
end
Generate kernel matrix $\mathbf{K}^{(1)}$ with entries $k_{ij}^{(1)} = \exp\left(-\frac{d_{ij}^2}{\varepsilon}\right)$;
$\mathbf{p} := \mathbf{K}^{(1)} \cdot \mathbf{1}$, where $\mathbf{1} = (1 \ldots 1)^T$;
Generate[b] matrix $\mathbf{K}^{(2)} := \mathbf{K}^{(1)}./(\mathbf{p} \cdot \mathbf{p}^T)$ (approximation of anisotropic Gaussian kernel) ;
$\mathbf{v} :=$ sqrt$(\mathbf{K}^{(2)} \cdot \mathbf{1})$;
Generate symmetric matrix $\mathbf{K} := \mathbf{K}^{(2)}./(\mathbf{v} \cdot \mathbf{v}^T)$ conjugated to $\mathbf{K}^{(2)}$;
Perform spectral decomposition of $\mathbf{K}$ to obtain eigenvalues and eigenvectors $\{\lambda_i\}_{i=1}^{n_{\text{bins}}}, \{\psi_i\}_{i=1}^{n_{\text{bins}}}$;
Return $\mathbf{M}^T$ with $\mathbf{M} := [\lambda_i \psi_i]_{i=2}^{p+1}$;

Algorithm 2: Diffusion Maps with histograms for crash simulation data

[a] For cross-bin-distances the input d comprises in particular the respective ground distance
[b] The operation ./ denotes elementwise division

We state the combined algorithms for step 1 and 2 in Algorithm 1 and for steps 3 and 4 in Algorithm 2, adapted from [Laf04]. The matrix $\mathbf{M} \in \mathbb{R}^{m \times p}$, generated from the output of the second algorithm by $\mathbf{M}[:, i] = (\lambda_i \psi_i)^T$ for $i \in \{ 2, 3, \ldots, n_{\text{bins}} \}$, comprises in its rows the diffusion coordinates of the m data points for an embedding into $\mathbb{R}^p$. Note again that the first eigenvector to the eigenvalue $\lambda_1 = 1$ is not taken into account, as it is the constant vector and thus does not store any relevant information.

4 Results and Discussion

To analyze the discrimination ability of the algorithm, it has been implemented in Python[2] and applied to the following data sets:

Truck-Beam Structural beam of a Chevrolet C2500 Pick-Up-Truck at frontal collision, $m = 132$ simulation runs with variation of 9 parameters, each $n = 1714$ FE-nodes [Boh13].

Car-Lateral Lateral part of a medium-sized vehicle at lateral collision with varying geometry (40 to 44 integral parts, original data from German car manufacturer), $m = 143$ simulation runs with different FE-meshes, each $n_i \approx 26.100$ FE-nodes.

(a) Structural Beam (b) Bending mode 1 (c) Bending mode 2

Figure 1: *Truck-Beam* data set and different bifurcation modes

As it is difficult to find an objective measure for the algorithm's quality we chose three evaluation criteria: Simulation runs that are similar according to their visualizations should be grouped together, runs that bear no similarities to others should be recognizable as outliers and in case of bifurcations the modes should be clearly separated.
To check the consistency of the new approach we first applied the algorithm to the *Truck-Beam* data set, which can also by analyzed with the conventional method. The data set is illustrated in Figure 1. It features a clearly visible bifurcation in its bending behavior, which we wish to find in the embeddings as well. Figure 2 shows a two dimensional embedding of the *Truck-Beam* data set according to the diffusion coordinates obtained with Algorithm 2. The data points are colored corresponding to the value of the second eigenvector (resp. the first diffusion coordinate), which according to spectral clustering theory comprises the most information about the global structure of the data set [Lux07]. It can be seen, that both axes correlate with the beam's bending behavior (color-coded in the small images): the second eigenvector separates the bifurcation modes by its sign, the value of the third eigenvector is proportional to the extent of the deformation in the middle segment of the beam.

[2]We used Python 2.7.1 with an extensive use of the packages *NumPy* and *matplotlib*

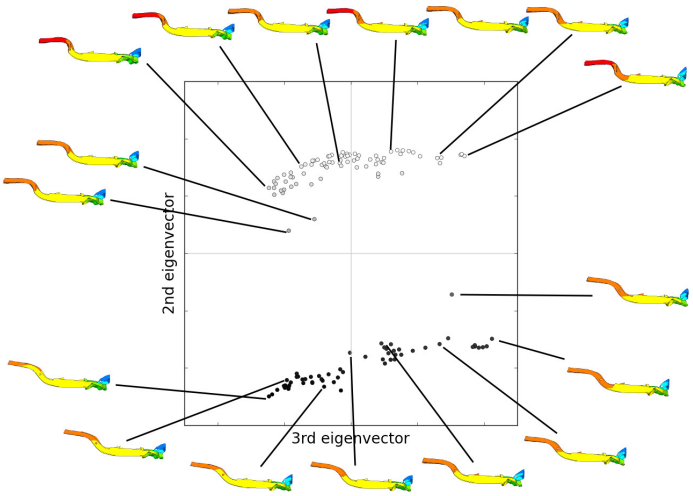
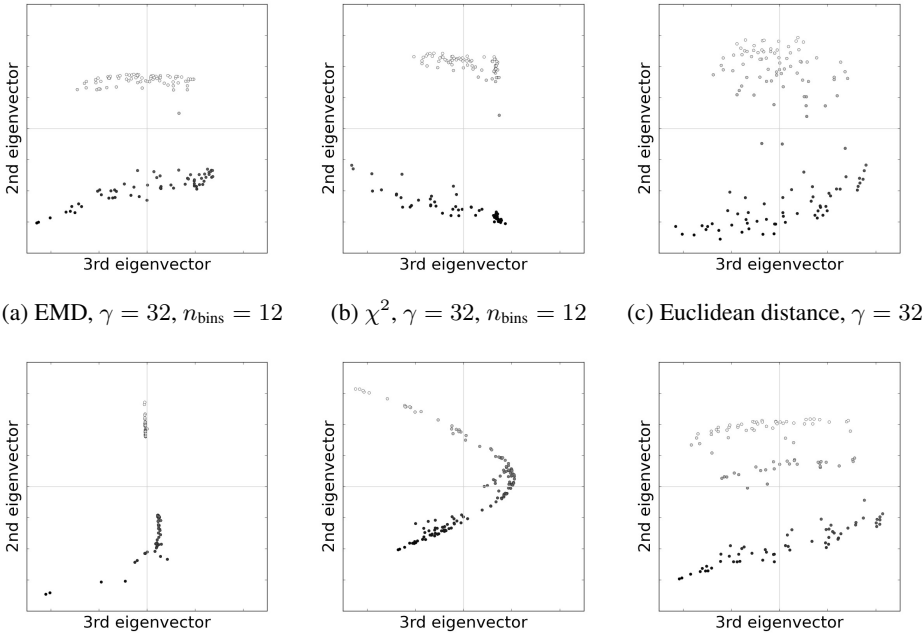


Figure 2: Embedding of the *Truck-Beam* dataset as in Figure 3a (parameters: $\gamma = 32, n_{\text{bins}} = 10$), with selected visualizations of deformations.

The first row of subfigures in Figure 3 shows the results obtained with different distance functions: 3a with histograms and the EMD histogram distance, 3b with histograms and the χ^2-distance function and 3c with the standard Euclidean distance of the displacement vectors without any histograms (which is applicable here as the FE grids for the simulation runs are identical). One can observe that the first two embeddings are similar concerning the structure of the clusters and the positioning of the data points. The Euclidean distance embedding is rather appropriate concerning the data as well, but the two groups are not as clearly separated as in the other embeddings, and the two central points are not placed to the correct bifurcation mode by the second eigenvector. These differences between histogram distances and displacement vector distances are confirmed for variations of the scaling parameter ε (details on the parameter choice see below) and are rather surprising, since by applying histograms in the preprocessing step and thus performing dimensionality reduction, we discarded some information. Considering the embeddings obtained, it must be assumed that a large part of this lost information was noise, which would make the technically less accurate new method more appropriate for this data set.

For the *Car-Lateral* data set with varying geometry the simulation runs are also grouped in an appropriate manner (see Figure 4). The two groups (upper and lower), as well as the two outliers, match the data well, especially along the second eigenvector. As the data are confidential industrial data provided by the manufacturer, visualizations cannot be shown here. It should be noted that this data set could not be analyzed with the conventional diffusion maps method, as the FE grids differ among the simulation runs. Therefore, our new method seems well applicable for the intended purpose in this first test.



(a) EMD, $\gamma = 32$, $n_{\text{bins}} = 12$ (b) χ^2, $\gamma = 32$, $n_{\text{bins}} = 12$ (c) Euclidean distance, $\gamma = 32$

(d) EMD, $\gamma = 1$, $n_{\text{bins}} = 12$ (e) EMD, $\gamma = 32$, $n_{\text{bins}} = 5$ (f) EMD, $\gamma = 32$, $n_{\text{bins}} = 20$

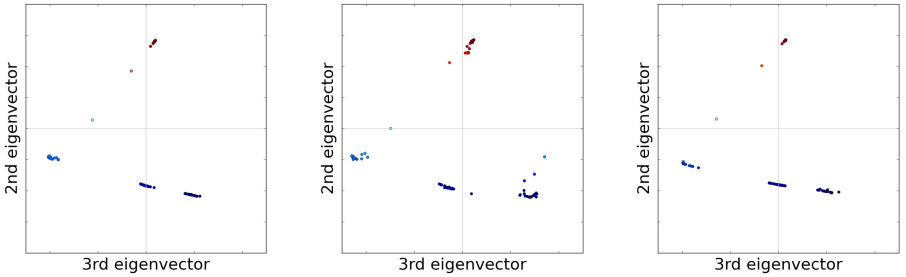
Figure 3: Different embeddings of the *Truck-Beam* dataset

4.1 Parameter choices and variations

The most important parameters to be varied in our proposed method are the ε of the Gauss kernel for Diffusion Maps and the n_{bins} for the histogram construction. The Gauss kernel's parameter determines the size of the neighborhood around a point x where the value $k_\varepsilon(x, y)$ differs from zero significantly. While there are no objective rules about suitable values, some rules of thumb have been established. As proposed for Diffusion Maps by Lafon we use the average distance between a point and its nearest neighbor [Laf04] and additionally scale it with a parameter γ: $\varepsilon := \frac{\gamma}{m} \sum_{x \in \mathcal{Y}} \min_{y \in \mathcal{Y}} \{\, d(x, y) \mid d(x, y) \neq 0 \,\}$.

For the *Truck-Beam* data set all tested choices of γ lead to reasonable embeddings, but changes apply to the axes' functions: For $\gamma = 1$ (cf. Figure 3d) the second eigenvector in addition to the information about the bifurcation modes comprises the information about the total deformation of the beam. Third and fourth eigenvector (not depicted here) take the function of indication vectors, as they each only comply information about one of the bifurcation modes and become zero for the other. For growing γ the embedding changes, to finally stay stable from $\gamma = 16$ with the point cloud shaped as in Figure $3a$.

Using histograms, the choice of n_{bins} (or the interval length $h \sim \frac{1}{n_{\text{bins}}}$ respectively) is crucial for the expressiveness of the histogram. Too small as well as too large classes will not do



(a) EMD, $\gamma = 32$, $n_{\text{bins}} = 16$ (b) χ^2, $\gamma = 32$, $n_{\text{bins}} = 16$ (c) EMD, $\gamma = 32$, $n_{\text{bins}} = 5$

Figure 4: Different embeddings of the *Car-Lateral* dataset

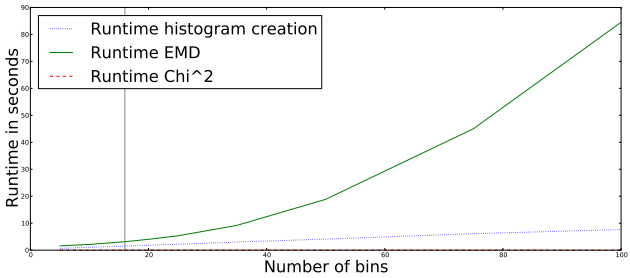


Figure 5: Practical Runtimes, evaluated for the *Car-Lateral* data set. The choice of $n_{\text{bins}} = 16$, obtained by the Sturges rule, is marked by the gray vertical line.

justice to the underlying structure: small intervals emphasize minor changes too much; large intervals obscure the structures entirely. In statistical applications, rules of thumb have been established for class intervals, depending on the number of observed data points. The displacements calculated in the car crash show structures similar to those of several normal distributions: As the nodes are all connected, a large displacement for one node influences all neighboring nodes as well. Therefore we applied one of these rules here as well: Due to its simplicity and the similar results of all rules for small data sets, we initially chose n_{bins} according to the Sturges-rule [Stu26]:

$$n_{\text{bins}} = \lceil \log_2(m) + 1 \rceil.$$

Our experiments showed, that for all data sets the Sturges rule provided very good results. Less classes obscured some of the information, as to be seen in Figure 3e with a choice of $n_{\text{bins}} = 5$ for the *Truck-Beam* data set, where the bifurcation is no longer obvious and some misplacements appear compared to the choice of $n_{\text{bins}} = 12$ proposed by the Sturges-rule. For more classes on the other hand, general structures and tendencies are to be seen for all tested bin numbers, but here as well the bifurcation becomes less obvious, cf. Figure 3f for $n_{\text{bins}} = 20$, where three agglomerations can be seen instead of two. The *Car-Lateral* data

set was even more stable for variations of n_{bins}, as to be seen in Figure 4c: even for only 5 bins the embedding looked very similar to the Sturges-rule based choice of $n_{\text{bins}} = 16$; analogously the shape of the point cloud does not change for very small bin sizes such as $n_{\text{bins}} = 100$. The negative effects of too large choices for n_{bins} are (as expected) boosted by the application of bin-to-bin-distances, such as χ^2, but still surprisingly good, especially considering the low runtime [Sch13].

The number of bins yields the initial configuration for the distance computations and thus, combined with the choice of a distance function, determines the algorithm's runtime. Here lies a large advantage of the χ^2-distance, as it runs linearly in $O(n_{\text{bins}})$, whereas EMD is usually implemented with Orlin's algorithm [Kor08], resulting in a runtime of $O({n_{\text{bins}}}^3 \log n_{\text{bins}})$. A practical evaluation is plotted in Figure 5. For all class sizes considered, calculation of the χ^2-distance took less than 0.01 seconds.

5 Conclusion

Our presented approach for the analysis of car crash simulation data proved successful for a real-world example with varying geometries. In a preprocessing step we represented the displacement data as normalized histograms and then used pairwise histogram distances to calculate a lower dimensional embedding using Diffusion Maps. The industrial data was arranged in a meaningful manner with visible correlations between the coordinate axes and the deformation behavior, which should simplify further analysis by development engineers.

Applying the method to a data set with fixed grid structures we could compare it to classical diffusion maps based on the Euclidean distances of the displacement vectors. This led to the surprising result that for our (simple) data set the distinguishing quality of the embedding was even improved by the new approach. Especially the EMD histogram distance proved stable against parameter variations. The χ^2-distance shows comparably good embedding results at much better runtimes, but with a higher sensitivity to parameter changes.

For the histogram representation in contrast to a displacement vector representation all information about positions and directions of the deformations are discarded. For the data investigated here, this did not prove problematic, but might be for more complex data sets. In general though the bundles of simulation runs generated in the car development process are all considered under the same loading conditions, so that positions of deformations should not vary. Still, to overcome these potential problems for more complex simulation runs one could further investigate the use of multidimensional histograms, e.g. 3-dimensional histograms to store displacement information on all coordinate directions instead of absolute displacements.

Acknowledgements

This paper is based on the author's bachelor's thesis [Sch13] and was written at Fraunhofer Institute for Scientific Computing and Algorithms SCAI. The author wishes to thank Prof. Dr. Jochen Garcke (Fraunhofer SCAI and Institute for Numerical Simulation, University of Bonn) and Rodrigo Iza-Teran (Fraunhofer SCAI) for their support and encouragement.

References

[Boh13] Bohn, B., Garcke, J., Iza Teran, R., Paprotny, A., Peherstorfer, B., Schepsmeier, U. and Thole, C. A. Analysis of Car Crash Simulation Data with Nonlinear Machine Learning Methods. *Procedia Computer Science*, 18:621 – 630, 2013. 2013 International Conference on Computational Science.

[Coi06] Coifman, R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[Erb07] Erban, R., Frewen, T., Wang, X., Elston, T., Coifman, R., Nadler, B. and Kevrekidis, I. Variable-free exploration of stochastic models: a gene regulatory network example. *Journal of Chemical Physics*, 126(15):155103, 2007.

[Iza14] Iza Teran, R. Enabling the Analysis of Finite Element Simulation Bundles. *International Journal for Uncertainty Quantification*, 4(2):95–110, 2014.

[Kor08] Korte, B. and Vygen, J. *Kombinatorische Optimierung - Theorie und Algorithmen*. Springer-Verlag, Berlin, 2008.

[Laf04] Lafon, S. *Diffusion Maps and Geometric Harmonics*. PhD thesis, Yale University, 2004.

[Lin06] Ling, H. and Okada, K. Diffusion Distance for Histogram Comparison. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 246–253, 2006.

[Lux07] Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[Nad08] Nadler, B., Lafon, S., Coifman, R. and Kevrekidis, I. Diffusion maps-a probabilistic interpretation for spectral embedding and clustering algorithms. In *Principal manifolds for data visualization and dimension reduction*, pages 238–260. Springer, 2008.

[Pel10] Pele, O. and Werman, M. The Quadratic-Chi Histogram Distance Family. In *Computer Vision â ECCV 2010*, volume 6312 of *Lecture Notes in Computer Science*, pages 749–762. Springer Berlin Heidelberg, 2010.

[Rub00] Rubner, Y., Tomasi, C. and Guibas, L.J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:99–121, 2000.

[Sch13] Schwartz, A.-L. Diffusion Maps und ihre Anwendung bei der Analyse von Automobildaten. Bachelor's thesis, University of Bonn, 2013.

[Stu26] Sturges, H. A. The choice of a class interval. *American Statistical Association*, 21:65–66, 1926.

Schema-Management ohne Schema? - Schema-Verwaltung in NoSQL-Datenbanksystemen

Helena Glatzel

helenaglatzel@gmx.de
Hochschule Darmstadt, Master Studiengang Informatik

Abstract: In der Anwendungsentwicklung finden sich mittlerweile immer kürzere Release-Zyklen, auf die auch im Bereich der Datenbanken reagiert werden muss. Gerade im Umfeld der Webanwendungen erfreuen sich daher NoSQL-Datenbanksysteme zunehmender Beliebtheit. Durch deren Schema-Flexibilität kann effizient auf die schnellen Änderungen von Anwendung und Datenschema eingegangen werden. Üblicherweise ist es in NoSQL-Systemen nicht erforderlich im Voraus ein Schema zu definieren. Dieses ergibt sich implizit durch die Struktur der Daten und muss nicht gesondert definiert und verwaltet werden. Dadurch sind jederzeit flexibel Änderungen am Schema möglich. Allerdings bedeutet dies auch, dass die Verwaltung des Schemas zum Großteil in der Verantwortung der Anwendung liegt. Um die Komplexität der Anwendungsentwicklung zu begrenzen, ist ein gewisses Maß an Unterstützung durch das eingesetzte Datenbanksystem wünschenswert. Dieses Paper gibt daher zunächst einen Überblick über die momentane Unterstützung des Schema-Managements durch die gängigsten NoSQL-Datenbanken und vielversprechendsten, aktuellen Entwicklungen. Im Anschluss werden die gewonnenen Erkenntnisse genutzt, um eine systemunabhängige Schema-Management-Komponente konzeptionell zu entwerfen, die sämtliche Aufgaben des Schema-Managements erfüllt.

1 Einführung

Gerade im heutigen Web 2.0-Umfeld findet man in der Anwendungsentwicklung immer kürzere Release-Zyklen. Dies wirkt sich auch auf Datenbanken und die Struktur der Daten aus. Es ist oft nicht mehr möglich im Vorfeld ein festes Schema zu definieren, das nur wenigen Änderungen unterliegt und über viele Releases stabil bleibt.

Um diesen Eigenschaften gerecht zu werden, eignen sich NoSQL-Datenbanksysteme, denn die Schema-Flexibilität bzw. Schemalosigkeit wird als eines ihrer Kern-Features betrachtet [Tiw11]. Im Gegensatz zu relationalen Datenbanksystemen muss in den meisten NoSQL-Systemen a priori kein Schema definiert werden. Dies bedeutet jedoch nicht automatisch das komplette Fehlen eines Schemas. Die Daten selbst weisen implizit ein Schema auf, selbst wenn dieses nicht explizit modelliert wird.

Weiterhin verhalten sich NoSQL-Datenbanken hinsichtlich Schema-Änderungen sehr flexibel. Es ist jederzeit möglich Objekten neue Eigenschaften zuzuweisen, ohne dass

diese im Vorfeld definiert werden müssen (in relationalen Datenbanken mit dem SQL-Statement ALTER TABLE t_name ADD column_name datatype). Da außerdem die Heterogenität von Objekten unterstützt wird, können sich Schema-Änderungen bei Bedarf auf ein einziges Objekt beschränken und müssen nicht wie in relationalen Datenbanken für die gesamte Tabelle vorgenommen werden. Dadurch können zwei Objekte der selben (logischen) Klasse durch vollkommen unterschiedliche Attribute beschrieben werden.

Die Verwaltung des Schemas obliegt in NoSQL-Datenbanksystemen im Normalfall der Anwendung. Diese Übertragung der Verantwortung für das Schema-Management bringt zusammen mit der Schemaflexibilität von NoSQL-Datenbanksystemen eine neue Ebene der Komplexität in die Anwendungsentwicklung. Da Anwendungen mit jeglichen Varianten von Objekten umgehen können sollen, ist viel Eigenentwicklung erforderlich. Dies ist sowohl kosten- und zeitintensiv als auch fehleranfällig. Daher wäre eine (teilweise) Automatisierung des Schema-Managements und somit eine tiefergehende Unterstützung seitens der NoSQL-Datenbanksysteme wünschenswert.

Dieses Paper fasst die Ergebnisse einer Masterarbeit zum Thema „Schema-Management in NoSQL-Datenbanksystemen“ [Gla14] zusammen. Nach einer einleitenden Erläuterung der wichtigen Grundlagenbegriffe wird zunächst ein Überblick der derzeitigen Unterstützung des Schema-Managements in NoSQL-Datenbanksystemen geschaffen. Zu diesem Zweck werden die Anforderungen an ein Schema-Management identifiziert und mit Hilfe dieser Kriterien die populärsten NoSQL-Systeme analysiert. Darüber hinaus werden, basierend auf dem gleichen Anforderungskatalog, aktuelle Entwicklungen in diesem Bereich untersucht.

Im Anschluss wird der Entwurf einer modularen Schema-Management-Komponente beschrieben. Diese arbeitet systemunabhängig und erfüllt sämtliche Aufgaben des Schema-Managements.

2 Was bedeutet Schema-Management?

Grundsätzlich muss von einem Schema-Management eine Möglichkeit zur Modellierung eines Datenbankschemas geboten werden. Diese strukturelle Beschreibung umfasst die Definition verschiedener Entity-Typen mit ihren Attributen. Des Weiteren können verschiedenen Constraints bezüglich der Attribute (z.B. Einzigartigkeit „unique“ oder Existenz „not null“), sowie Beziehungen zwischen gleichen und unterschiedlichen Entity-Typen definiert werden. Werden Daten in die Datenbank eingefügt oder geändert, erfolgt eine Validierung gegen das definierte Schema. [EN05]

Darüber hinaus umfasst das Schema-Management eine Unterstützung der Schema-Evolution, also der späteren Weiterentwicklung des Schemas. Außerdem sollte eine durch Schema-Updates eventuell erforderliche Migration vorhandener Daten unterstützt werden. Hier sollten die beiden Varianten *eager* und *lazy* Migration zur Verfügung stehen. Unter eager Migration ist die Anpassung aller Datensätze an das neue Schema nach einer Schema-Änderung zu verstehen. So kann sichergestellt werden, dass nach Abschluss der Migrationsroutine alle Datensätze dem neuen Schema entsprechen. Lazy Migration führt

eine Migration erst dann durch, wenn die Daten nach einer Schema-Änderung das erste mal abgefragt werden. Ältere Einträge, auf die nicht mehr oder nur noch selten zugegriffen wird, müssen nicht an das neue Schema angepasst werden und verbleiben in der ursprünglichen Schema-Version. Dadurch werden unnötige Migrationen vermieden. Allerdings müssen Anwendungen beide Versionen des Schemas unterstützen. [KSS14]

Theoretisch ist auch eine dritte Variante denkbar. Nach einem Schema-Update können sowohl alte als auch neue Schema-Version erhalten bleiben, so dass keinerlei Datenmigration erforderlich ist. [KSS14]

Soll für eine Anwendung, die zunächst ohne Schema-Management gearbeitet hat, zukünftig eine Schema-Verwaltung eingebunden werden, muss eine Möglichkeit bestehen eine Schema-Extraktion, basierend auf den bereits in der Datenbank existierenden Daten, auszuführen.

3 Status Quo Analyse

Die nachfolgend beschriebenen Anforderungen an ein Schema-Management bauen auf typischen Eigenschaften eines Schema-Managements in relationalen Datenbanksystemen und [SKS13] auf. Basierend auf diesen Kriterien wurden sechs NoSQL-Datenbanksysteme und fünf aktuelle Entwicklungen analysiert. Es wurden NoSQL-Systeme der Kategorien Column-Family- und Document-Datenbanken betrachtet. [EFH11]

Da Key/Value-Datenbanken ein Datenmodell von geringer Komplexität nutzen, in dem einfache Key/Value-Paare ohne weitergehende Struktur-Informationen gespeichert werden, ist nicht davon auszugehen, dass diese Systeme eine Unterstützung des Schema-Managements bieten können. Daher wurden keine Datenbanken dieser Kategorie untersucht. Auch auf die Betrachtung von Graphdatenbanken wurde, aufgrund ihrer Andersartigkeit zu den Kategorien der untersuchten Datenbanksystemen, verzichtet. [RW12]

Die Ergebnisse der Analysen werden im zweiten Teil dieses Kapitels vorgestellt.

3.1 Anforderungskatalog

- Es wird eine geeignete Möglichkeit zur Schema-Beschreibung und -Speicherung benötigt. Dies schließt die Beschreibung der folgenden Eigenschaften ein:
    - verschiedene Entity-Typen
    - Struktur der Entity-Typen (Attribute)
    - Einschränkung der Attribute
        - Unterstützung der benötigten Datentypen
        - Constraints wie Einzigartigkeit (unique) oder Existenz (not null)

        - weitergehende Einschränkung des Wertebereiches auf bestimmte Wertgrenzen (z.B. Value between 0 and 10)
    - Kennzeichnung des Primary Keys
    - Beziehungen zwischen gleichen und unterschiedlichen Entity-Typen
- Schema-Extraktion
- Schema-Evolution
    - Hinzufügen eines neuen Entity-Typs
    - Löschen eines Entity-Typs
    - Umbenennen eines Entity-Typs
    - Update eines Entity-Typs
        - Hinzufügen eines Attributes
        - Löschen eines Attributes
        - Umbenennen eines Attributes
        - Verschieben eines Attributes
        - Kopieren eines Attributes
- Validierung gegen das Schema
- Koexistenz verschiedener Schema-Versionen zum Erhalt der Schema-Flexibilität
- Unterstützung der Datenmigration
    - Eager Migration
    - Lazy Migration

3.2 Analyse-Ergebnisse

Wie erwähnt wurden die sechs gängigsten NoSQL-Datenbanksysteme evaluiert. Es wurden die Column-Family-Systeme HBase [Hba14], Cassandra [Cas14] und Google App Engine Datastore (im Folgenden kurz mit Datastore bezeichnet), das auf Google's Big Table basiert [GAE14], untersucht. Die Ergebnisse des Datastore basieren auf der Nut-

zung mit dem verfügbaren Python SDK. Weiterhin wurden die Document-Datenbanken MongoDB [MDB14], CouchDB [CDB14] und Couchbase [Cou14] betrachtet.

Ergänzend wurden fünf verschiedene, aktuelle Software-Entwicklungen im Bereich des Schema-Managements analysiert. Dazu gehört KijiSchema, eine Komponente des Open Source Frameworks KijiProject, das Schema-Design und -Evolution für HBase bereitstellt. [Kij14] Außerdem wurden die beiden Mapper Hibernate OGM [OGM14] und Kundera [Kun14], sowie die Library Mongoose (für MongoDB) [Mon14] und das Java-Interface Objectify-Appengine, kurz Objectify, (für den Datastore) [OAE14] betrachtet.

Die Ergebnisse der Analysen sind in Abbildung 1 dargestellt. Ein ✓ bedeutet, dass das jeweilige Kriterium (annähernd) vollständig erfüllt wurde. Wohingegen ein **(✓)** eine nur teilweise Erfüllung, ein **X** eine mangelhafte oder keine Unterstützung kennzeichnet.

Es ist zu erkennen, dass die untersuchten NoSQL-Datenbanksysteme nur eine rudimentäre Unterstützung des Schema-Managements bieten können. CouchDB und Couchbase können sogar keine der Anforderungen erfüllen.

Im Vergleich dazu können MongoDB und HBase zumindest teilweise die Aufgaben eines Schema-Managements realisieren. In diesen beiden Systemen können zwar verschiedene Entity-Typen definiert werden, jedoch kann deren Struktur nicht modelliert werden. Daher sind in diesem Fall auch die Möglichkeiten der Schema-Extraktion und -Evolution stark eingeschränkt.

Lediglich Cassandra und der Datastore, in Verbindung mit dem Python SDK, können eine relativ umfassende Unterstützung des Schema-Managements bieten. Mit beiden Systemen kann die Schema-Modellierung fast vollständig umgesetzt werden. Einzig Einschränkungen in der Verfügbarkeit von Constraints und der Definition von Beziehungen existieren. Auch Methoden zur Schema-Extraktion und -Evolution werden von Cassandra und dem Datastore teilweise bereitgestellt.

Die Analysen der zusätzlich untersuchten Software-Entwicklungen (Tabelle 1) zeigen, dass die Schema-Modellierung und auch die Validierung in diesen Fällen sehr gut unterstützt wird. Allerdings kann alleine KijiSchema einige Methoden zur Schema-Extraktion und -Evolution zur Verfügung stellen. Weiterhin bietet einzig KijiSchema eine wirkliche Unterstützung der Koexistenz mehrerer Schema-Versionen, ist aber auf HBase beschränkt. Kundera und Hibernate OGM haben den Vorteil, dass sie den Einsatz verschiedener Datenbanksysteme ermöglichen. Jedoch befindet sich Hibernate OGM derzeit noch im Beta-Stadium und kann daher noch nicht alle geplanten Funktionen anbieten.

Abschließend lässt sich sagen, dass sowohl die analysierten NoSQL-Systeme als auch aktuelle Software-Entwicklungen in diesem Bereich die Anforderungen an ein Schema-Management nur ansatzweise erfüllen. Selbst eine Kombination der beiden, mit denen sich die Möglichkeiten der Schema-Modellierung und Validierung deutlich verbessern, ist die Funktionalität nicht ausreichend. Gerade an einer Unterstützung der Schema-Extraktionen und -Evolution sowie einer Automatisierung der Datenmigration fehlt es. Aus

diesem Grund wird im nächsten Kapitel der Entwurf einer Schema-Management-Komponente beschrieben, die systemunabhängig arbeitet und sämtliche Anforderungen des Schema-Managements erfüllen kann.

Tabelle 1: Gesamtüberblick der derzeitigen Unterstützung des Schema-Managements durch NoSQL-Datenbanksysteme und aktuelle Entwicklungen

Feature	HBase	Cassandra	Datastore	MongoDB	CouchDB	Couchbase	Kiji Schema	Mongoose	Hibernate OGM	Kundera	Objectify
Schemamodellierung	(✓)	(✓)	✓	(✓)	X	X	(✓)	✓	✓	✓	✓
Entity-Typen	(✓)	✓	✓	(✓)	X	X	✓	✓	✓	✓	✓
Struktur der Entity-Typen	X	✓	✓	X	X	X	✓	✓	✓	✓	✓
Attribut-Einschränkungen	X	(✓)	(✓)	X	X	X	(✓)	(✓)	(✓)	(✓)	(✓)
Primary Key	X	✓	✓	X	X	X	✓	X	✓	✓	✓
Beziehungen	X	X	✓	X	X	X	X	✓	✓	✓	✓
Schema-Extraktion	(✓)	(✓)	✓	(✓)	X	X	(✓)	X	X	X	X
Schema-Evolution	(✓)	(✓)	(✓)	(✓)	X	X	(✓)	X	X	X	X
Add Entity-Type	(✓)	✓	(✓)	(✓)	X	X	✓	X	X	X	X
Delete Entity-Type	(✓)	✓	X	(✓)	X	X	✓	X	X	X	X
Rename Entity-Type	X	X	X	(✓)	X	X	X	X	X	X	X
Update Entity-Type	X	(✓)	X	X	X	X	(✓)	X	X	X	X
Add Property	X	✓	✓	X	X	X	✓	X	X	X	X
Delete Property	X	✓	X	X	X	X	✓	X	X	X	X
Rename Property	X	X	X	X	X	X	✓	X	X	X	X
Move Property	X	X	X	X	X	X	X	X	X	X	X
Copy Property	X	X	X	X	X	X	X	X	X	X	X
Validierung	(✓)	✓	✓	X	X	X	✓	✓	✓	✓	✓
Mehrere Schema-Versionen	X	X	(✓)	X	X	X	✓	X	X	X	(✓)
Datenmigration	X	X	X	(✓)	X	X	X	X	X	X	(✓)
Eager Migration	X	X	X	(✓)	X	X	X	X	X	X	X
Lazy Migration	X	X	X	X	X	X	X	X	X	X	(✓)

4 Entwurf der Schema-Management-Komponente

An die nachfolgend beschriebene Schema-Management-Komponente werden ebenfalls die in 3.1 festgelegten Anforderungen gestellt. Zusätzlich soll die Komponente eine Automatisierung für den Prozess der Schema-Evolution und der Datenmigration bereitstellen. Werden Operationen auf den Daten ausgeführt, die implizit das Schema verändern, ist ein automatisches Schema-Update durchzuführen. Im Gegenzug soll im Fall expliziter Schema-Änderungen eine ggf. erforderliche Datenmigration automatisch durchgeführt werden. Im Zuge der Planung einer möglichen Umsetzung der Komponente wurde ebenfalls untersucht, welche Module der bisher analysierten Systeme ggf. wiederverwendet werden können, um den Entwicklungsaufwand möglichst gering zu halten.

4.1 Architektur

Generell existieren für die Einbindung der Schema-Management-Komponente zwei verschiedene Varianten. Die Komponente kann innerhalb des Datenbanksystems oder als externe, zusätzliche Schicht realisiert werden. Um die Unabhängigkeit vom Datenbanksystem, sowie eine Austauschbarkeit des DBMS oder ggf. eine Einbindung mehrerer, verschiedener Systeme zu unterstützen, wird die entworfene Schema-Management-Komponente, wie in Abbildung 2 dargestellt, als externe, zusätzliche Schicht geplant. [vgl. KSS14]

Wie auch in [KSS14] beschrieben, ist es auf diese Weise möglich für jede Anwendung individuell zu entscheiden, ob die Schema-Management-Komponente genutzt werden, oder ein direkter Zugriff auf das NoSQL-Datenbanksystem erfolgen soll. Die Möglichkeit ohne Schema-Management zu arbeiten wird allerdings durch die Art der Datenmigration und den unterstützten Grad der Schema-Flexibilität (vgl. hierzu 4.3 Validierungsmodi) beeinflusst. Wird beispielsweise lazy Migration genutzt ist eine Umgehung der Komponente u.U. nicht möglich, falls die Anwendung nicht mit sämtlichen Schema-Versionen kompatibel ist, auch wenn die Anwendung selbst kein Schema-Management nutzt.

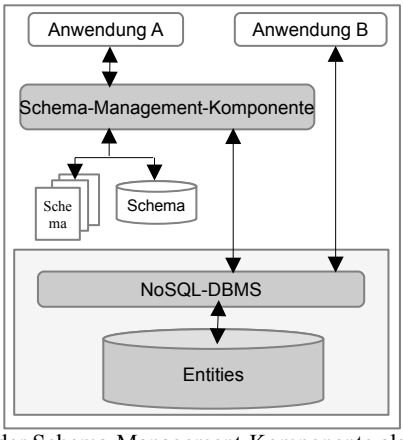


Abbildung 1: Einbindung der Schema-Management-Komponente als externe Schicht [vgl. KSS14]

Um die verschiedenen Anforderungen an ein Schema-Management zu erfüllen, ist die Schema-Management-Komponente in drei Schichten gegliedert (Abbildung 3).

Die oberste Schicht, die Daten- und Schema-Manipulations-Schicht, fungiert als Schnittstelle zum Anwender für Operationen auf den Daten und zur Schema-Evolution. In der mittleren Schema-Management-Schicht werden die unterschiedlichen Aufgaben des Schema-Managements realisiert. Diese Schicht ist in sieben Manager-Module gegliedert, damit die individuellen Anforderungen umgesetzt werden können. Die unterste Schicht, die Schema-To-NoSQL-DB-Schicht, umfasst datenbankspezifische Mapper, mit denen durchzuführende Operationen auf verschiedene Datenbanksysteme abgebildet werden.

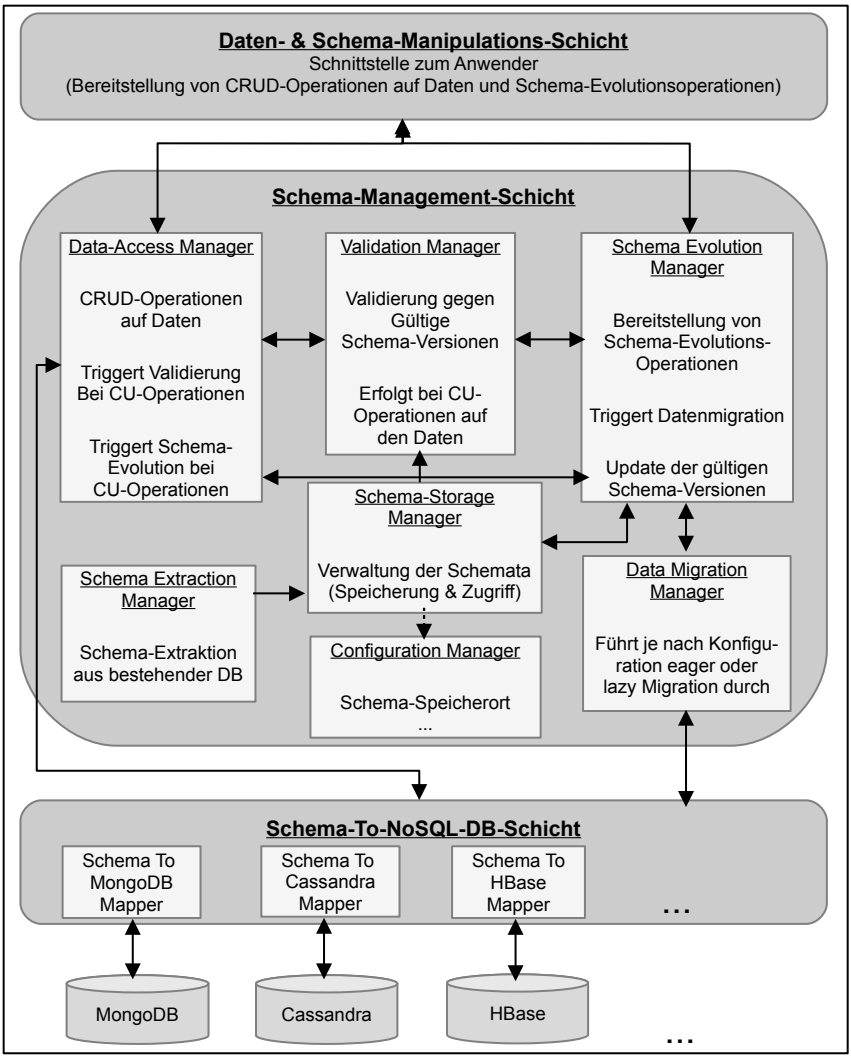


Abbildung 2: Architektur der Schema-Management-Komponente

4.2 Schema-Modellierung

Für die Umsetzung der Schema-Modellierung existieren verschiedene Möglichkeiten. So können in Cassandra die Schema-Informationen beispielsweise mit Hilfe der systemeigenen Cassandra Query Language (CQL) beschrieben werden. Auch die Nutzung von JPA-Annotationen (Java Persistence API) oder ähnlichen Abwandlungen, wie in Hibernate OGM, Kundera oder Objectify genutzt, sind denkbar.

JSON Schema [JSc14] ist eine deklarative, JSON-basierte Sprache, mit der JSON-Dokumente in ihrer Struktur beschrieben werden können. Da sich JSON Schema als Format zur Beschreibung von JSON-Dokumenten, besonders im Umfeld der Webanwendungen, etabliert hat und es viele der an die Modellierung gestellten Anforderungen erfüllen kann, wird JSON Schema in der Schema-Management-Komponente für die Schema-Beschreibung eingesetzt. Darüber hinaus verwenden viele NoSQL-Datenbanksysteme ohnehin JSON-Dokumente zur Speicherung der Daten. Aber auch die Datenstrukturen anderer Systeme lassen sich mit Hilfe dieses Formates problemlos modellieren.

Um die Schema-Verwaltung zu realisieren werden einige Attribute, wie z.B. der Entity-Typ oder die Versionsnummer, definiert, die jedes, in Verbindung mit der Schema-Management-Komponente genutzte, JSON Schema enthalten muss. Diese Eigenschaften werden in einem Metaschema definiert, gegen das alle erstellten JSON-Schema-Dokumente validiert werden.

4.3 Validierung

Werden Daten in die Datenbank eingefügt, müssen diese gegen gültige Schema-Versionen validiert werden. Hierfür existieren verschiedene Libraries für diverse Programmiersprachen (u.a Python, Java, JavaScript, PHP, ...) Diese können für die Datenvalidierung eingesetzt werden.

Für den Fall invalider Daten sind zwei verschiedene Vorgehensweisen möglich. Zum einen können invalide Daten, wie aus relationalen Datenbanksystemen bekannt, abgelehnt werden. Zum anderen ist es, dank der Schema-Flexibilität von NoSQL-Systemen, auch möglich, dass durch die Speicherung invalider Daten eine neue Schema-Version definiert wird.

Aus diesem Grund führt die Schema-Management-Komponente zwei verschiedene Validierungsmodi ein, die vom Anwender konfiguriert werden können. Wird *strictValidation* genutzt, werden invalide Daten nicht von der Komponente akzeptiert. Wird *updateValidation* konfiguriert, ruft der Data-Access-Manager für invalide Daten den Schema-Evolution-Manager auf, der die entsprechenden Schema-Updates automatisch durchführt.

4.4 Datenmanipulation

Damit der Zugriff auf die Daten über die Schema-Management-Komponente realisiert werden kann und somit eine Validierung, wie auch die automatische Schema-Evolution

und Datenmigration möglich sind, muss eine datenbanksystemunabhängige Anfragesprache zur Verfügung gestellt werden. Um die Lernkurve möglichst flach zu halten, verwendet die Komponente eine SQL-ähnliche Sprache. Voraussichtlich kann zu diesem Zweck eine Anpassung der relevanten Operationen der CQL an die Nutzung in der Komponente vorgenommen werden.

Je nach eingesetztem Datenbanksystem werden die Methoden der Anfragesprache in der Schema-To-NoSQL-DB-Schicht mit Hilfe von spezifischen Mappern auf Operationen des konkreten Datenbanksystems abgebildet. Für die Implementierung der Mapper können Funktionalitäten der datenbanksystemspezifischen Module der beiden systemübergreifenden Mapper Hibernate OGM und Kundera, nach vorheriger Anpassung, genutzt werden. Damit weitere, von diesen Mappern nicht unterstützte, Systeme von der Schema-Management-Komponente eingesetzt werden können, müssen die entsprechenden Mapper selbst entwickelt werden, falls nicht anderweitige Vorlagen existieren.

4.5 Schema-Evolution

Zur Realisierung der Schema-Evolution ist ebenfalls die Definition einer adäquaten Sprache erforderlich. Auch in diesem Fall sind verschiedene Umsetzungsvarianten denkbar. Da als Format zur Schema-Beschreibung JSON Schema gewählt wurde, könnten Evolutions-Operationen mit Hilfe von JSONiq [Jiq], einer Abfragesprache für JSON-Dokumente, durchgeführt werden. Jedoch soll die Schema-Management-Komponente nicht auf ein bestimmtes Format der Schema-Modellierung beschränkt werden, da es denkbar ist zu einem späteren Zeitpunkt ein anderes oder mehrere verschiedene Formate anzubieten. Daher wird die Nutzung einer unabhängigen Evolutionssprache bevorzugt.

Da erforderliche Anpassungen verfügbarer Lösungen, wie beispielsweise der CQL oder der von KijiSchema, relativ umfangreich sind und diese Systeme ohnehin keine vollständige Unterstützung der Schema-Evolution bieten, wird eine Eigenentwicklung vorgezogen. Zu diesem Zweck ist eine Implementierung der in [SKS13] beschriebenen Schema-Evolution-Language für die Schema-Management-Komponente vorgesehen. Diese Sprache besteht aus nur wenigen, intuitiv nutzbaren Befehlen, mit denen alle notwendigen Evolutions-Operationen ausgeführt werden können. Aufgrund des geringen Umfangs ist sie schnell zu erlernen und relativ einfach implementierbar.

Intern werden die Evolutions-Operationen, sofern entsprechende Methoden verfügbar sind, auf JSONiq abgebildet, um Änderungen an den JSON-Schema-Dokumenten durchzuführen.

4.6 Datenmigration

Die Schema-Management-Komponente unterstützt, wie oben bereits beschrieben, sowohl eager als auch lazy Migration. Die Migrationsstrategie ist Entity-Typ-basiert vom Anwender konfigurierbar.

Der Schema Evolution Manager initiiert automatisch eine erforderliche Datenmigration, die durch den Data Migration Manager ausgeführt wird. Hierfür werden die erforderlichen Migrationsschritte aus den durchgeführten Schema-Änderungen abgeleitet.

4.7 Schema-Extraktion

Eine Schema-Extraktion kann, wie in [KSS14] beschrieben, mit Hilfe eines Spanning Graph realisiert werden. Zunächst werden die Strukturinformationen der JSON-Dokumente eines bestimmten Entity-Typs eingelesen. Anschließend wird inkrementell ein Spanning Graph aufgebaut, der alle Strukturvarianten des jeweiligen Entity-Typs enthält. Aus dem Graphen kann schließlich das JSON Schema abgeleitet werden.

Dieser Prozess kann nicht nur für sämtliche Daten eines Entity-Typs durchgeführt werden, sondern auch auf einem Ausschnitt der Daten, der beispielsweise durch einen konkreten Timestamp oder ein Attribut bestimmt wird. Auf diese Weise ist die Ableitung verschiedener Schema-Varianten bzw. -Versionen eines Entity-Typs möglich. [KSS14]

5 Fazit und Ausblick

In den durchgeführten Analysen konnte festgestellt werden, dass sowohl die evaluierten NoSQL-Datenbanksysteme als auch aktuelle Entwicklungen in diesem Bereich nur eine unzureichende Unterstützung des Schema-Managements zur Verfügung stellen.

Dennoch wurden einige Teilbereiche identifiziert, in denen die Anforderungen relativ umfassend erfüllt werden. So kann, vor allem unter Einsatz zusätzlicher Aufsätze, die Schema-Modellierung weitgehend unterstützt werden. Dies ist jedoch bei Weitem nicht ausreichend, um ein annähernd zufriedenstellendes Schema-Management zu realisieren.

Aus diesem Grund wurden die gewonnenen Erkenntnisse genutzt, um eine modulare Schema-Management-Komponente zu entwerfen. Diese arbeitet systemunabhängig, erhält die Schema-Flexibilität und erfüllt alle Anforderungen, die an ein Schema-Management gestellt werden.

Als logische Fortführung dieser Arbeit bietet sich die prototypische Implementierung der entworfenen Schema-Management-Komponente an. Dies ist im Rahmen einer Projektgruppe ab dem Wintersemester 2014/15 an der Hochschule Darmstadt geplant.

Abschließend lässt sich sagen, dass die Entwicklung im Bereich des Schema-Managements von NoSQL-Datenbanksystemen noch am Beginn steht und weitere Forschungs- und Entwicklungsarbeit erforderlich ist, bevor sich erste, systemunabhängige Lösungen etablieren und den Arbeitsalltag mit NoSQL-Datenbanken merklich erleichtern.

Danksagung

Ich möchte mich bei Prof. Dr. Uta Störl bedanken, die mich während meiner Masterarbeit betreut und auch zu diesem Beitrag ein kurzes Feedback gegeben hat.

Literaturverzeichnis

[Cas14] The Apache Cassandra Project: http://cassandra.apache.org, 2014, letzter Zugriff: 21. April 2014

[CDB14] Apache CouchDB: http://couchdb.apache.org, 2014, letzter Zugriff: 20. April 2014

[Cou14] Couchbase: http://www.couchbase.com, 2014, letzter Zugriff: 21. April 2014

[EFH11] Edlich, S; Friedland, A.; Hampe, J.; Brauer, B.; Brückner, M.: NoSQL: Einstieg in die Welt nichtrelationaler WEB2.0 Datenbanken; 2. Auflage; Münschen: Carl Hanser Verlag, 2011

[EN05] Elmasri, R.; Navathe, S. B.: Grundlagen von Datenbanksystemen – Ausgabe Grundstudium; 3. Auflage. Münschen: Pearson Studium; 2005.

[HBa14] Apache HBase: https://hbase.apache.org, 2014, letzter Zugriff: 22. April 2014

[Jiq14] JSONiq: http://www.jsoniq.org, 2014, letzter Zugriff: 22. April 2014

[JSc14] JSON Schema: http://json-schema.org, 2014, letzter Zugriff: 21. April 2014

[GAE14] Google Developers, Google App Engine: https://developers.google.com/appengine/, 2014, letzter Zugriff: 21. April 2014

[Gla14] Glatzel, H.: Schema-Management in NoSQL-Datenbanksystemen. Master's thesis, Hochschule Darmstadt, 2014.

[Kij14] The KijiProject: http://www.kiji.org, 2014, letzter Zugriff: 20. April 2014

[KSS14] Klettke, M.; Scherzinger, S.; Störl, U.: Datenbanken ohne Schema? - Herausforderungen und Lösungs-Strategien mit schema-flexiblen NoSQL-Datenbanksystemen. Datenbank-Spektrum, Volume14, Nummer 2, Juli 2014.

[Kun14] Kundera: https://github.com/impetus-opensource/Kundera/wiki/Object-mapper, 2014, letzter Zugriff: 20. April 2014

[MDB14] MongoDB: http://www.mongodb.org, 2014, letzter Zugriff: 21. April 2014

[Mon14] Mongoose: http://mongoosejs.com, 2014, letzter Zugriff: 22. April 2014

[OAE14] Objectify-Appengine: https://code.google.com/p/objectify-appengine/, 2014, letzter Zugriff: 20. April 2014

[OGM14] Hibernate OGM: http://hibernate.org/ogm/, 2014, letzter Zugriff: 21. April 2014

[RW12] Redmond, E.; Wilson, J.R.: Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement; Pragmatic Programmers, 2012

[SKS13] Scherzinger, S.; Klettke, M.; Störl, U.: Managing Schema Evolution in NoSQL Data Stores. In: Proc. DBPL; 2013.

[Tiw11] Tiwari, S.: Professional NoSQL. Indianapolis, Indiana: John Wiley & Sons; 2011.

Konzeption und Entwicklung eines Informationssystems zur Unterstützung der Vertriebsplanung im Lösungsvertrieb

Markus Paatsch
Karlsruher Institut für Technologie
markus.paatsch@student.kit.edu

Zusammenfassung: Dieser Artikel beschäftigt sich mit der Konzeption, Entwicklung und Evaluation einer softwaretechnischen Lösung zur Vertriebsplanungsoptimierung in lösungsorientierten B2B-Unternehmen. Die Grundlage der Optimierungen bilden hierbei historische Vertriebsdaten, wie z.B. frühere Vertriebschancen. Es wird untersucht, wie sich eine derartige Anwendung unter Zuhilfenahme eines bestehenden Business-Intelligence-Umfeld integrieren lässt. Die konzipierte Webapplikation wird als JSF-Anwendung umgesetzt. Es wird ein Klassifizierungsverfahren implementiert, welches durch die Bibliothek weka integriert wird. Zur Optimierung werden Funktionen der Bibliothek ilog cplex benutzt. Hauptaufgabe der Anwendung ist die Maximierung des Umsatzes durch Optimierung der Zuweisungen von Vertriebsmitarbeitern zu Kunden. Alle optimierten Ergebnisse werden in übersichtlicher und anschaulicher Art visualisiert.

1 Einführung

Der einmalige Verkauf von Produkten verliert im Business-to-Business-Umfeld zunehmend an Bedeutung. Stattdessen konzentrieren sich immer mehr B2B-Unternehmen darauf, durch das Anbieten von komplexen, integrierten Lösungen langfristige Beziehungen mit Kunden zu erschaffen, um so einen größeren und kontinuierlicheren Gewinn zu erwirtschaften. Der lösungsorientierte Verkauf existiert so nur im B2B-Umfeld. Eine Lösung ist aus verschiedenen Produkten und Dienstleistungen zusammengesetzt [JKS03]. Für diese Arbeit ist nur das grundlegende Konzept von Lösungen wichtig, die spätere Lösungsdomäne ist irrelevant. Verkäufer bzw. Repräsentanten des Unternehmens spielen eine zunehmend wichtigere Rolle beim Aufbau von Kundenverbindungen. Durch die Zusammenarbeit von Verkäufern, bzw. deren gezielte Zuweisung zu bestimmten Kunden kann eine Gewinnmaximierung erzielt werden [KBB13]. Innerhalb dieser Teams existieren oft verschiedene Vertriebsrollen. Die Interaktionen zwischen Kunden und Verkäufern können durch die rasche IT-Entwicklung der letzten Jahrzehnte detailliert gemessen und in großem Umfang gespeichert werden. Diese Daten werden im Folgenden Kundeninteraktionsdaten genannt. Sie beschreiben auf verschiedenen Ebenen die Güte der Interaktion zwischen Kunde und Verkäufer [Ha12]. Mit Hilfe dieser historischen Daten ist es möglich, die Vertriebsplanung zu optimieren und somit den Ertrag eines Unternehmens entscheidend zu maximieren. Da dem Verkaufsprozess von komplexen, integrierten Lösungen bisher

relativ wenig Beachtung geschenkt wurde, existieren aktuell keine Softwaretools, die es ermöglichen, Verkaufsteams im B2B-Umfeld optimal bzw. gewinnbringend Kunden – auf Basis von Kundeninteraktionsdaten – zuzuweisen.

Das Ziel dieser Arbeit ist die softwaretechnische Planung, Umsetzung und Evaluation einer Webapplikation zur Vertriebsplanungsoptimierung. Die Webapplikation dient zur Unterstützung in der Vertriebsplanung und soll in ein Business-Intelligence-Umfeld integriert werden. Diese Arbeit beschäftigt sich dabei ausschließlich mit der softwaretechnischen Umsetzbarkeit der verschiedenen Komponenten, ihrer Interaktion untereinander und der anschließenden Integration des Systems in ein bestehendes Business-Intelligence-Umfeld.

Als Daten-Grundbausteine dienen *Gebiete*, mit darin enthaltenen *Kunden* und *Verkäufern* eines Unternehmens. Besteht eine *Verkaufsmöglichkeit* mit einem Kunden, muss diesem ein Verkäufer zugewiesen werden. Der Nutzer des Systems soll in der Lage sein, diese *Zuweisungen* zu ändern. Eine *Vertriebsplanung* für ein Gebiet ist komplett, wenn jedem Kunden dieses Gebiets ein Verkäufer zugeordnet wurde. Die Hauptaufgabe der Webapplikation besteht darin, unvollständige Vertriebsplanungen mittels Data-Mining-Techniken und linearer Programmierung optimal zu ergänzen bzw. das Optimum zu berechnen. Alle Ergebnisse werden visualisiert und können untereinander verglichen werden (dazu gehört auch der Vergleich mit dem Optimum).

Im weiteren Verlauf werden zuerst verwandte Arbeiten vorgestellt. Anschließend werden verschiedene Arbeitsschritte der Anwendung aufgezeigt. Die Anforderungsanalyse bildet die Grundlage für die nachfolgenden Abschnitte Design und Architektur sowie Implementierung. Abschließend werden die umgesetzten Anforderungen evaluiert und ein Fazit gezogen bzw. ein Ausblick gegeben.

2 Verwandte Arbeiten

In [KBF14] wurde eine umfangreiche Literaturrecherche zu Themen wie Sales Force Modelling, Sales Force Analytics, Sales Force Optimization, etc. durchgeführt. Diese Literaturrecherche liefert daher auch die ursprüngliche Grundlage für die in diesem Paper beschriebene Implementierung. Vorhergehende Arbeiten auf ähnlichem Themengebiet aus [KBF14] sind vor allem [Lo76], [SA98], [DA99] und [Ba12]. Daraus resultiert z.B. das in der Einführung angeführte Fehlen eines derartigen Tools zur Zuweisung von Verkaufsteams im B2B-Umfeld. Grundlage für dieses Paper bildet die Arbeit von Johannes Kunze von Bischhoffshausen und Jeffrey T. Becker [KBB13], welche das zugrundeliegende Optimierungsmodell der Anwendung liefert. Die verwendete Datengrundlage durch Kundeninteraktionen basiert auf der Arbeit von Francois Habryn [Ha12].

3 Methodik

Im Folgenden werden die Schritte vorgestellt, die für das System notwendig sind, um optimale Ergebnisse oder Ergänzungen mit diesem Optimierungsmodells zu berechnen. Das Vorgehen basiert grundlegend auf dem Modell von [DH07]: Der Datenaufbereitung folgt eine prädiktive Analyse (Vorhersage), um mit Hilfe dieser optimale Ergebnisse zu berechnen.

3.1 Daten

Die Webapplikation bezieht ihre Daten aus einem Data Warehouse, in dem bereits die Datenaufbereitung stattfand. Der Datenprozess beschränkt sich für das System daher nur auf die Datenbeschaffung, also die Kommunikation mit dem Data Warehouse, und die Transformation in das eigene Format. Die bezogenen Daten bieten verschiedene aggregierte Metriken für bisherige Verkaufsmöglichkeiten.

3.2 Prädiktive Analyse

Die bezogenen Daten repräsentieren Verkaufsmöglichkeiten aus der Vergangenheit. Aufgrund dieser historischen Daten kann für eine bestehende Käufer-Verkäufer-Beziehung der Verkaufserfolg vorhergesagt werden. Dies geschieht bei der prädiktiven Analyse beispielhaft mit dem Verfahren der logistischen Regression, welches ein statistisches Verfahren zur Regressionsanalyse darstellt. Dabei wird anhand einer Funktion der Zusammenhang einer abhängigen und mehrerer unabhängiger Variablen modelliert. Damit das Modell Wahrscheinlichkeiten im Bereich $p \in [0,1]$ liefert, muss die abhängige Variable binär sein. Der jeweilige Ausgang (verkauft, nicht verkauft) von vergangenen Verkaufsmöglichkeiten wird somit zur abhängigen Variable und alle gesammelten Daten zur vorherigen Interaktion werden als unabhängige Variablen modelliert.

3.3 Optimierung

Ausgehend von den vorhergesagten Verkaufswahrscheinlichkeiten sollen optimale Zuweisungen bzw. optimale Ergänzungen zu unvollständigen Vertriebsplanungen berechnet werden. Dazu wird das Optimierungsmodell aus [KBB13] benutzt:

$$max \sum_{j=1}^{n} \left(v_j \prod_{i=1}^{m} q_{ij}^{x_{ij}} \right)$$

$$s.t. \sum_{j=1}^{m} \left(x_{ij} \right) \leq W_i \quad (i = 1, \dots, m)$$

$$\sum_{i=1}^{n} \left(x_{ij} \right) = 1 \quad (j = 1, \dots, n)$$

Dabei steht *i* für Verkäufer und *j* für Kunde. x_{ij} ist eine binäre Variable: 1, wenn Verkäufer *i* Kunde *j* zugewiesen ist (0, sonst). W_i ist die maximale Anzahl an Kunden, die einem Verkäufer zugewiesen werden können. Dieser Parameter kann manuell geändert werden. v_j bezeichnet den möglichen Ertrag bei Verkauf an Kunde *j*. Die Wahrscheinlichkeit, mit welcher ein Verkäufer *i* bei Kunde *j* Erfolg hat, heißt q_{ij}.

4 Anforderungsanalyse

Ein Benutzer soll die vorgestellte Methodik anhand von verschiedenen Interaktionsmöglichkeiten steuern können. Diese werden nachfolgend durch die funktionalen Anforderungen spezifiziert. Die Rahmenbedingungen der funktionalen Anforderungen werden durch die nichtfunktionalen Anforderungen beschrieben.

4.1 Funktionale Anforderungen

- **FA1** *Anfordern einer Datenauswahl*: Der Nutzer wählt die Gebiete für die Optimierung.
- **FA2** *Verändern der Optimierungsparameter*: Verschiedene Optimierungsparameter (maximale Anzahl von Kunden pro Verkäufer, etc.) können manuell geändert werden.
- **FA3** *Ändern des Gebiets in dem Zuweisungen optimiert werden*: Der Nutzer kann in ein anderes Gebiet wechseln.
- **FA4** *Verkäufer für die Optimierung im aktuellen Gebiet an- bzw. abwählen*: Im aktuellen Gebiet können Verkäufer in die Berechnung integriert oder von dieser ausgeschlossen werden.
- **FA5** *Manuelles Anpassen der Zuweisungen zwischen Verkäufern und Kunden*: Für eine Berechnung können feste Zuweisungen gesetzt werden.
- **FA6** *Berechnung optimaler (ergänzender) Zuweisungen im aktuellen Gebiet*: Fehlende Zuweisungen zwischen Kunden und Verkäufer sollen automatisch von der Anwendung ergänzt werden.
- **FA7** *Vergleich verschiedener Berechnungen*: Verschiedene Berechnungen können miteinander oder dem Optimum verglichen werden.

4.2 Nichtfunktionale Anforderungen

- **NA1** *Zugriffsanforderungen*: Der Zugriff auf die Webapplikation erfolgt ausschließlich per Webbrowser. Dazu muss der Server bzw. die URL der Anwendung bekannt und erreichbar sein.
- **NA2** *Nutzungsanforderungen*: Alle funktionalen Anforderungen sollen in maximal 3 Sekunden bearbeitet und ausgeführt werden.

- **NA3** *Bedienungsanforderungen*: Sobald man Zugriff auf die Webapplikation besitzt, kann diese ausschließlich mit der Maus bedient werden. Durch die Integration von Tastaturbefehlen wird die Bedienung erleichtert.
- **NA4** *Fehlerbehandlung*: Jede unerlaubte Operation des Nutzers bzw. jeder durch das System erkannte Fehler wird dem Nutzer in angemessener Art und Weise angezeigt.

Durch die Definition von Anforderungen an die Anwendung kann das System anschließend näher spezifiziert werden.

5 Design und Architektur

Im folgenden Abschnitt werden verschiedene Anforderungen technisch näher spezifiziert. Die Grundlage für die Geschäftslogik bilden die funktionalen Anforderungen aus Abschnitt 4.1. Exemplarisch wird pro Schritt (Abschnitt 3) eine Anforderung betrachtet. Abschließend wird die grundlegende Architektur des Systems erläutert.

Spezifikation - Datenauswahl anfordern (Abschnitt 3.1): Alle Funktionen des Systems können nur auf einer vorhandenen Datenbasis ausgeführt werden. Aus diesem Grund wird beim Aufruf der Webapplikation zuerst eine Daten- bzw. Gebietsauswahl vom Nutzer getroffen. Die Daten sind in einem Data Warehouse gespeichert und können mit Hilfe der Vorauswahl abgefragt werden. Die bezogenen Daten müssen anschließend in ein eigenes Format überführt werden.

Spezifikation - Zuweisungen anpassen (Abschnitt 3.2): Ein Verkäufer kann (einstellbar vielen) Kunden – jedem Kunden aber nur einmal - zugeordnet werden. Wird die maximale Anzahl an Kunden pro Verkäufer überschritten, informiert die Anwendung den Nutzer per Fehlerinformation und bricht die Operation ab. Beim eigentlichen Zuweisen von Verkäufern zu Kunden wird gleichzeitig die Verkaufswahrscheinlichkeit aufgrund der vorhandenen Daten berechnet (prädiktive Analyse).

Spezifikation - Berechnung optimaler Zuweisungen (Abschnitt 3.3): Die Hauptaufgabe der Anwendung liegt darin, den erwarteten Umsatz von Vertriebsplanungskonfigurationen zu berechnen. Diese können in zwei Arten eingeteilt werden. Bei vollständig zugewiesenen Konfigurationen kann nichts optimiert werden. Der Nutzer hat bereits jedem Kunden einen Verkäufer zugeordnet. Die Anwendung hat in diesem Fall die Aufgabe, verschiedene Kenngrößen wie den erwarteten Umsatz zu berechnen. Wenn der Nutzer allerdings nicht alle Zuweisungen manuell setzt, wird die Anwendung diese Aufgabe übernehmen bzw. optimieren (unvollständige Konfiguration). Es muss sichergestellt sein, dass ausreichend berechnungsrelevante Verkäufer vorhanden sind.

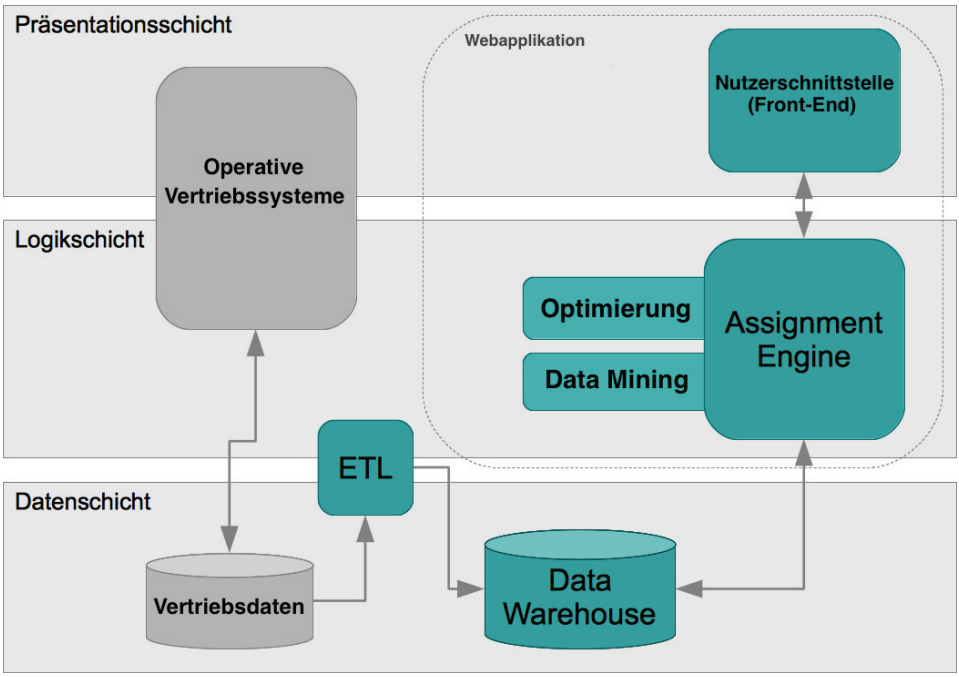


Abbildung 1: Systemarchitektur

Abbildung 1 beschreibt die Einbettung der Anwendung in die Gesamt- bzw. Systemarchitektur. Das komplette System ist in drei Schichten untergliedert (Daten-, Logik- und Präsentationsschicht), wobei die Webapplikation nur auf Logik- und Präsentationsschicht arbeitet. Daten werden aus einem bestehenden Data Warehouse bezogen, welches Schnittstellen zum Datentransfer anbietet. Die Datenschicht ist daher nicht Teil dieser Arbeit.

Die Anwendungsarchitektur in Abbildung 2 ist ein detailliert dargestellter Teilbereich der Systemarchitektur. Sie bezieht sich nur auf die Realisierung des Webapplikationsteils aus Abbildung 1. Da die Anwendung auf einem zentralen Server gehostet wird und jeder Nutzer nur per Webbrowser darauf zugreifen soll, bietet sich hier eine einfache Client-Server-Architektur an, welche wiederum als MVC-Pattern [Ga11] realisiert wird. Zusätzlich zur Grundarchitektur sind in der Abbildung auch die Datenströme der Anwendung integriert.

Mit den Anforderungsspezifikationen und dem Architekturentwurf kann das System implementiert werden.

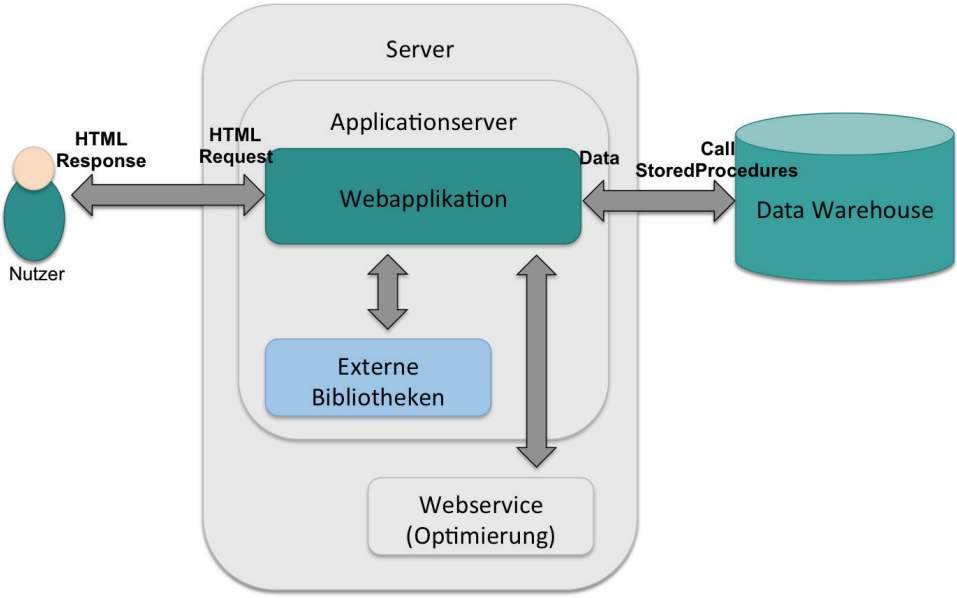


Abbildung 2: Anwendungsarchitektur

6 Implementierung

Die Webapplikation wird als JSF-Anwendung (JavaEE 7, [Ja14]) umgesetzt. Zusätzlich werden verschiedene Frameworks wie Twitter's Bootstrap [Bo14], jQuery [Jq14] sowie D3.js [D314] zur Graph-Visualisierung auf Clientseite eingesetzt. Durch das *Model-View-Controller Pattern* [Ga11] ist eine gewisse Paketstruktur vorgegeben, welche allerdings noch um Services und Parser erweitert wurde. Demnach befindet sich jeglicher HTML/Javascript-Content im View-Bereich. Das *Model* beinhaltet die interne Datenstruktur der Webapplikation. Der *Controller*, hier als Mediator umgesetzt, steuert die gesamte Kommunikation zwischen Front- und Back-End sowie mit dem Data Warehouse.

6.1 Model

Damit die Daten aus dem Data Warehouse verarbeitet werden können, müssen diese in eine entsprechende Datenstruktur überführt werden. Diese Datenstruktur ist im Model-Paket verankert. Der Mediator verwaltet ein *Map*-Objekt und hält damit alle Gebiete bzw. alle angeforderten Daten. Bei Instanziierung erhält ein Gebiet eine Standardkonfiguration ohne jegliche Zuweisungen. Alle aktuellen Änderungen von Nutzerseite werden innerhalb einer aktuellen Konfiguration abgelegt. Sobald eine Berechnung erfolgreich war, wird diese an eine Liste von allen Berechnungen angehängt. Diese Berechnungen hängen immer am aktuellen Konfigurationsobjekt,

welches alle für die Berechnung verwendeten Verkäufer und Kenngrößen, wie erwarteter Umsatz, speichert. Die innerhalb einer Konfiguration integrierten Verkäufer-Objekte verwalten ihre eigenen Zuweisungen wobei ein Zuweisungsobjekt den jeweiligen Kunden und die Verkaufswahrscheinlichkeit enthält. Außerdem verwaltet jedes Verkäuferobjekt seine eigenen vergangenen Verkaufsmöglichkeiten. Diese sind entscheidend für die logistische Regression, da sie neben dem Kunden alle Metriken und das Resultat der damaligen Verkaufsmöglichkeit speichern.

6.2 Services

Services sind als modulare Bausteine der Anwendung gedacht. Ein schneller Austausch des Optimierungsverfahrens oder des Klassifikators wird somit vereinfacht. Das Services-Paket beinhaltet folgende Funktionen:

- Kommunikation mit dem Data Warehouse – Daten (Abschnitt 3.1)
- Prädiktive Analyse (Abschnitt 3.2)
- Optimierung von Konfigurationen (Abschnitt 3.3)

Die Klasse *DataService* ist dafür zuständig, mit dem Data Warehouse zu kommunizieren. Dabei geht es nur um die Etablierung einer Verbindung und die Entgegennahme von Daten. Um die Rohdaten in die Model-Datenstruktur zu übersetzen, wird der *DataParser* aus dem Parser-Paket beansprucht. Letztendlich liefert der DataService ein komplettes Map-Objekt, gefüllt mit den angewählten Daten.

Die ursprünglich vorgestellten Techniken zum Data Mining und zur Optimierung werden nicht eigenhändig in den Klassen *ClassificationService* und *OptimizationService* implementiert. Vielmehr sind auch diese beiden Klassen Schnittstellen zu externen Bibliotheken. Eigene Daten müssen lediglich in das externe Format übersetzt werden. Im Rahmen dieser Arbeit wird für die logistische Regression das Open-Source-Framework *weka* [We14] und für die binäre ganzzahlige Optimierung das von IBM entwickelte *ilog cplex* [IC14] benutzt. Der Optimierungsansatz wurde in Abschnitt 3.3 erläutert.

6.3 Controller

Das Paket *Controller* bildet die Controller-Komponente des MVC-Patterns. Die einzige darin enthaltene Klasse *Mediator* (Vermittler-Pattern, [Ga11]) steuert die gesamte Kommunikation zwischen Front- und Back-End und zwischen der Webapplikation und dem Data Warehouse. Damit der Nutzer Funktionen des Back-Ends ansprechen kann, ist die Klasse Mediator direkt mit dem Front-End verbunden (Callback-Funktionen). Um die Nutzeranweisungen weitergeben zu können, verwaltet der Mediator ein komplettes Model-Objekt (gefüllt mit den vom Nutzer angeforderten Daten).

6.4 View

Als Nutzerschnittstelle bzw. View (MVC, [Ga11]) wird im aktuellen Kontext jeglicher HTML-/CSS-/Javascript-Content bezeichnet. Die Datenauswahl kann den Eintrittspunkt in die Anwendung darstellen. Hier sollen dem Nutzer alle vorhanden, auswählbaren Gebiete aufgelistet werden. Diese initiale Gebietsliste wird beim Seitenaufruf aus dem Data Warehouse abgefragt. Erst wenn eine nichtleere Auswahl an Gebieten getroffen wurde, ist es möglich, auf die Übersichtsseite zu navigieren. Die Übersichtsseite ist in Abbildung 3 exemplarisch mit 4 Gebieten dargestellt. Verbindungen zwischen Kunde (engl. customer) und Verkäufer bzw. Repräsentant (engl. representative) symbolisieren in dieser Ansicht (*Sales Collaboration History*) vergangene Verkaufsmöglichkeiten und deren Kardinalität. Der Nutzer hat hier die Möglichkeit, sich Daten detaillierter anzeigen zu lassen (*Additional Information*), oder den Datenbestand zu filtern. Mit einem Klick auf *What-If Analysis* kann der Nutzer Zuweisungen für ein Gebiet verändern. Im nächsten Reiter können die so veränderten Konfigurationen optimiert und verglichen werden. In allen Ebenen der Webapplikation werden (Zwischen-) Ergebnisse angemessen visualisiert.

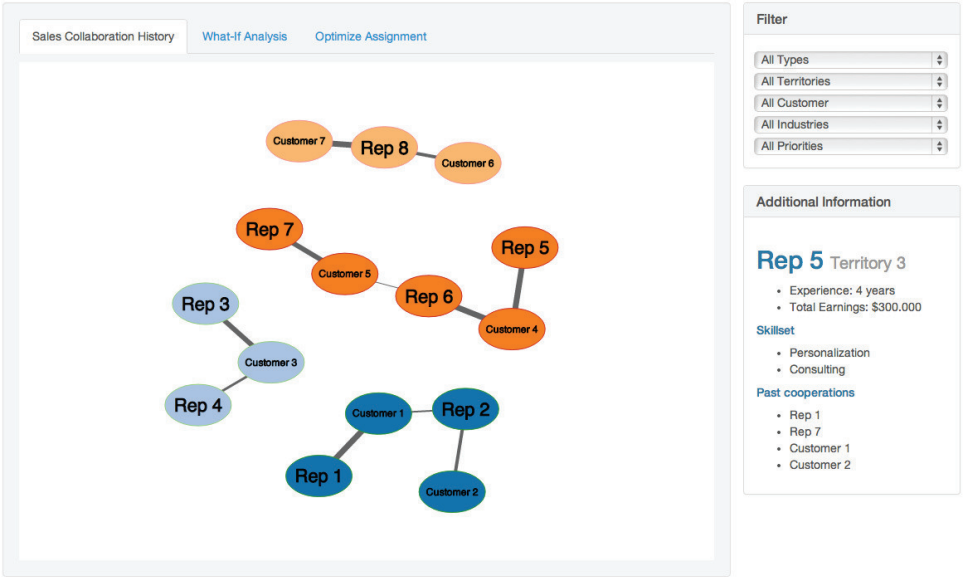


Abbildung 3: Ausschnitt der Oberfläche (Zuweisungen in Gebieten)

7 Evaluation

Nachfolgend wird untersucht, ob die entwickelte Webapplikation den spezifizierten Anforderungen aus Kapitel 4 gerecht wird.

Die einzelnen Klassen bzw. Module/Pakete des Back-Ends können mit Hilfe von Unit Tests automatisiert auf korrekte Funktionsausführung überprüft werden. Dafür wurde für jede Klasse des Back-Ends ein Unit-Test geschrieben. Unit-Tests prüfen jede implementierte Funktionalität mit vorgefertigten Ergebnissen. Bei Änderungen am Quelltext kann so leicht erkannt werden ob die Funktionalität verletzt wurde. Für das Back-End der Webapplikation wurden alle Unit-Tests mittels JUnit entwickelt. Alle Unit-Tests wurden bestanden. Nachdem jedes Modul isoliert auf seine grundlegende Funktionalität geprüft wurde, muss anschließend durch Integrationstests das richtige Zusammenspiel aller Module untersucht werden. Mit Hilfe von Test-Daten und dem Debug-Werkzeug der Entwicklungsumgebung wurde sichergestellt, dass jeder Testprozess korrekt abläuft. Nachfolgend wird exemplarisch ein solches Testszenario vorgestellt:

Berechnung nicht-vollständiger Konfiguration: Die Datenbeschaffung war erfolgreich und der Nutzer kann mit der Manipulation der Zuweisungen beginnen. Zunächst aktiviert er eine Teilmenge der Verkäufer für die Berechnung. Durch diese Funktionalität ist die Verbindung zwischen Controller und Model sichergestellt. Die What-If-Seite zeigt stets nur das an, was tatsächlich als Model im Back-End liegt. Anschließend setzt der Nutzer Zuweisungen in der Form, dass mindestens ein Kunde keinen Verkäufer zugewiesen bekommt. Unter der Annahme, dass so mindestens eine Zuweisung existiert, kann nun überprüft werden, ob die zugehörige, berechnete Verkaufswahrscheinlichkeit korrekt ist. Dies wurde mit verschiedenen Daten manuell überprüft. Das Beispiel läuft korrekt ab, also funktioniert auch die Verbindung zwischen Controller und ClassificationService. Danach kann die Berechnung gestartet werden. Auch hier wurden Beispiele manuell mit den Ergebnissen verglichen, um die Optimalität der hinzukommenden Zuweisungen zu überprüfen.

8 Fazit und Ausblick

In dieser Arbeit sollte eine Webapplikation zur Vertriebsplanungsoptimierung – basierend auf Kundenbeziehungen – softwaretechnisch geplant, umgesetzt und evaluiert werden. Durch die Entwicklung des Systems sollte grundlegend geprüft werden, ob eine solche Anwendung in ein bestehendes Business-Intelligence-Umfeld integriert werden kann. Durch die erfolgreiche Umsetzung der Webapplikation können mehrere Fragen beantwortet werden. Zum einen wurde gezeigt, dass die geforderten Funktionen (Kapitel 4) und Komponenten mit vorhandener Technologie implementiert werden können. Das bedeutet weiter, dass eine derartige Anwendung in ein bestehendes Business-Intelligence-System integriert werden kann. Zum anderen ist der Nutzen für Unternehmen mindestens genauso wichtig wie die Integration. Da das System alle Anforderungen umsetzt, kann aus bisher wenig beachteten Daten automatisiert neues

Wissen über eine optimierte Vertriebsplanung gewonnen werden. Trotz der erfolgreichen prototypischen Umsetzung kann ein derartiges System kaum mit vorhandener Software verglichen werden. Es existieren bisher keine vergleichbaren Systeme die im B2B-Umfeld auf Kundeninteraktionsdaten Vertriebsplanungsoptimierungen für Verkaufsteams durchführen.

Nachfolgend werden einige Ideen aufgezeigt, wie eine derartige Anwendung weiterentwickelt werden könnte.

- Verwerfung des Gebietsparadigma – Ein globales Gebiet
- Einführung zusätzlicher Bedingungen in der Optimierung
- (Echte) Serviceorientierte Architektur
- Realisierung für mobile Geräte
- Manuelle Erstellung einer Datenbasis zu Simulations-/Forschungszwecken
- Verbesserung des Optimierungsmodells
- Austausch des Prädiktionsalgorithmus

Für den Nutzen des noch jungen Gebiets der Customer Intimacy Analytics bzw. der Sales Force Analytics/Optimization bietet das entwickelte System einen ersten demonstrativen Grundstein und die Möglichkeit für weitere Forschungen.

Danksagung

Dieses Projekt wurde im Rahmen meiner Bachelorarbeit unter Betreuung von Johannes Kunze von Bischhoffshausen umgesetzt und seitdem weiterentwickelt. Ihm gilt ein besonderer Dank, da er mich nicht nur bei der Abschlussarbeit, sondern auch beim Verfassen dieses Papers tatkräftig unterstützt hat.

Literaturverzeichnis

[Ba12] Baier, M., Carballo, J. E., Chang, A. J., Lu, Y., Mojsilovic, A., Richard, M. J., Singh, M., Squillante, M. S., Varshney, K. R.: Sales-force performance analytics and optimization. In IBM Journal of Research and Development (56:6), S. 8:1-8:10, 2012

[Bo14] Bootstrap, 2014, http://www.getbootstrap.com, zuletzt geprüft am 19.06.2014

[D314] D3.js, 2014, http://www.d3js.org, zuletzt geprüft am 19.06.2014

[DH07] Davenport, T. H., Harris, J.G.: Competing on analytics: The new science of winning. Harvard Business School Press, Cambridge, 2007

[DH99] Drexl, A., Haase, K.: Fast approximation methods for sales force deployment. In Management Science (45:10), S. 1307-1323, 1999

[Ga11] Gamma, E.: Entwurfsmuster: Elemente wiederverwendbarer objektorientierter Software. Addison-Wesley, München, 2011

[Ha12] Habryn, F.: Customer intimacy analytics: Leveraging operational data to assess customer knowledge and relationships and to measure their business impact. Dissertation, Karlsruhe, 2012

[IC14] IBM Ilog Cplex, 2014, http://www-03.ibm.com/software/products/de/ibmilogcpleoptistud, zuletzt geprüft am 19.06.2014

[Ja14] Java EE 7, 2014, http://www.oracle.com/technetwork/java/javaee/overview/index.html, zuletzt geprüft am 19.06.2014

[JKS03] Johansson, J. E., Krishnamurthy, C., Schlissberg, H.E.: Solving the solutions problem. In McKinsey Quarterly (3), S. 116-125, 2003

[Jq14] jQuery, 2014, http://jquery.com/, zuletzt geprüft am 19.06.2014

[KBB13] Kunze von Bischhoffshausen, J., Becker, J.: Towards a Personal Relationship-Based Assignment of Client Representatives to Accounts. In IESS, S. 329-335, 2013

[KBF14] Kunze von Bischhoffshausen, J., Fromm, H.: Sales Force Analytics for the Solution Selling Firm: A Predictive Model for Assessing the Impact of Sales Team Assignments. In AMICS, 2014, i.E.

[Lo76] Lodish, L.M.: Assigning salesmen to accounts to maximize profit. In Journal of Market Research (13:4), S. 440-444, 1976

[SA98] Skiera, B., Albers, S.: COSTA: Contribution Optimizing Sales Territory Alignment. In Marketing Science (17:3), S. 196-213, 1998

[We14] Weka, 2014, http://www.cs.waikato.ac.nz/ml/weka/, zuletzt geprüft am 19.06.2014

Evaluating the Combination of two User-oriented Usability Evaluation Methods: Thinking-aloud and Questionnaires

Fabian Trautsch
fabian.trautsch@stud.uni-goettingen.de
University of Göttingen, Masters Programme Applied Computer Science

The Usability Evaluation (UE) of a product has been a popular research topic in recent years, e.g. [FIA11, HKM13]. Usability is assessed by different methods at every step of software development. Through redesign of the product and repeated assessments, the usability of a software is improved. UE also plays an important role in the development of websites. Nowadays, with web 2.0, web applications, and dynamic websites, it is even more important to test the usability of a website than it was in the past.

Usability is assessed through its major aspects. Based on the International Organization for Standardization (ISO) 9241, these are effectiveness, efficiency and satisfaction. Every UE method evaluates different aspects of usability, but there is no method that evaluates all three at once. The aim of our work was the provision of a combination of two well-known UE methods that solves this problem. Furthermore, we evaluated our method through a case study with 28 participants.

The combination of methods we proposed is a combination of two different user-oriented usability evaluation methods: Thinking-aloud and questionnaires. We have shown, that both methods give us different information about the usability of a website and therefore locate different usability problems. With our combination of methods, we are able to get data about the effectiveness, efficiency and satisfaction of the website as well as data about the context in which the website is used. These information give us the ability to determine the usability of a website and locate usability problems.

We identified and prioritized 15 usability problems in our case study. Additionally, we were able to derive concrete redesign suggestions for the tested website based on the user statements and comments during and after the usability tests.

References

[FIA11] Adrian Fernandez, Emilio Insfran, and Silvia Abrahão. Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, 53(8):789–817, 2011.

[HKM13] Nico Hamacher, Karl-Friedrich Kraiss, and Jörg Marrenbach. Einsatz formaler Methoden zur Evaluierung der Gebrauchsfähigkeit interaktiver Geräte (Utilization of Formal Methods for the Usability Evaluation of Interactive Devices). *Issues*, 55, 2013.

sGrid: Centralized management of BOINC-based research projects - Summary

Emanuel Jöbstl, Jérôme Urhausen, Elisaweta Masserova and Ainara Askar

IPD Tichy
Faculty of Computer Science
Karlsruhe Institute of Technology
Am Fasanengarten 5
76131 Karlsruhe
sGrid@ipd.kit.edu

To speed up computionally intensive research projects, work can be distributed among computers of volunteers. This process is called desktop grid computing. The Open-Source Software BOINC (Berkeley Open Infrastructure for Network Computing)[1] provides a platform for grid based volunteer computing and forms a de-facto standard.

However, BOINC was introduced in 2002, when web services like social media sites did not exist. Furthermore, BOINC is highly decentralized. There exists an API for so-known BOINC account managers, which enable users to join multiple BOINC projects using only one account. The management of scoreboards and other similar features is still handled by each project itself.

sGrid, an acronym for social grid, tries to overcome these gaps between BOINC projects, by providing a web based application which unifies and interconnects projects. To achieve this goal, sGrid offers the following features:

- Login opportunities using social network accounts.
- Central selection and comparison of projects, using only a single account.
- A cross-project reward system.
- A custom BOINC client application wrapper wich is built do be easy to use.

Using this approach, an additional appeal to share their computer for research purposes is given to internet users.

[1] BOINC, University of California, Berkeley, http://boinc.berkeley.edu

Drohnensteuerung im Programmierunterricht

Julius Höfler

Komplexe Multimediale Anwendungsarchitekturen
Institut für Informatik und Computational Science
Universität Potsdam
julius.hoefler@uni-potsdam.de

Der Umgang mit Computern und das Erlernen von Programmiersprachen wird heutzutage in vielen Schulen schon frühzeitig gelehrt. Einfache, kindgerechte Entwicklungsumgebungen (IDEs) erleichtern diese Arbeit. So erlangt die grafische Kinder-IDE „Scratch“ immer größere Beliebtheit. Sie abstrahiert typische Programmierkonzepte und stellt sie in Form von grafischen Bausteinen zur Verfügung. Diese können via „Drag & Drop“ zusammengesteckt werden. Syntax-Fehler sind somit ausgeschlossen [Ze14].

ScratchDrone ermöglicht es den Schülern, ein eigenes Programm zur Steuerung einer Flugdrohne zu schreiben und auf einer realen Drohne auszuführen. Dabei handelt es sich bei ScratchDrone um ein Framework, welches die Drohne als Laufzeitumgebung in Scratch einbindet. Die notwendigen Steuerkommandos wurden mittels der Scratch-Erweiterung „Build Your Own Blocks“ (BYOB) implementiert und werden den Schülern in Form von weiteren Bausteinen zur Verfügung gestellt. Die Drohne erhält dabei ihre Steuerbefehle durch REST-Services, welche über WiFi beispielsweise von einem Computer an sie übertragen werden. Dafür agiert die Drohne als DHCP-Server und weißt sich selbst und allen Geräten im Netzwerk eine IP-Adresse zu. Somit kann die Drohne ausschließlich von Geräten bedient werden, die sich in ihrem Netzwerk befinden. Der PC besitzt zudem eine Socket-Verbindung mit Scratch bzw. BYOB, über welche er die Steuerbefehle empfängt. Diese werden im Anschluss interpretiert und in REST-Kommandos umgewandelt.

Die Evaluierung an mehreren Schulen ergab, dass die Kinder großen Spaß an der Drohnenprogrammierung hatten [Ze14]. Das angestrebte Ziel, das Interesse am Programmierunterricht und an der Informatik im Allgemeinen zu steigern, konnte erfüllt werden. Somit soll in Zukunft der Einsatz dieser Technologie weiter ausgebaut und verfeinert werden.

Literatur

[Ze14] Raphael Zender, Julius Höfler, Patrick Wolfien, Ulricke Lucke. *ScratchDrone - Systematische Programmierung von Flugdrohnen für den Informatikunterricht*. März 2014.

Statische Informations- und Datenflussanalyse von Android-Apps

Markus Schmidt
Universität Siegen, Bachelorstudiengang Informatik
markus1.schmidt@student.uni-siegen.de

Im Zeitalter von zunehmender Überwachung durch Kriminelle oder staatlichen Institutionen gewinnt die IT-Sicherheit immer mehr an Bedeutung. Von einer Sicherheitslücke sind oft viele Endgeräte betroffen. Die Folge kann eine große Menge an ausgespähten Daten sein. Dahingehend existieren Werkzeuge zur Analyse von Applikationen wie die Datenflussanalyse, welche die technischen Aspekte, und die Informationsflussanalyse, welche die inhaltlichen Aspekte, wie zum Beispiel Sicherheitslücken im Programmcode, analysiert. In der vorliegenden Arbeit wurden einige Methoden vorgestellt. Außerdem wurde eine Analyse an einer Beispiel-App evaluiert.

Die Analyse mit der Methode Dexpler [BJK13] anhand der Beispiel-App ist eine reine Datenflussanalyse und zeigt die technischen Möglichkeiten und Grenzen der statischen Analyse auf. Für einfache Verschachtelungen im Quelltext werden zuverlässige Ergebnisse produziert, die für die weitere inhaltliche Analyse genutzt werden können. [FAR+13] Mithilfe der Analyse des Dalvik-Bytecodes der kompilierten apk-Datei eines Android-Handys wird der Datenfluss gut lesbar repräsentiert. Mit verschiedenen Repräsentationen als zum Beispiel Kontrollflussdiagramm, Quelltext und Zwischenstufen des Dekompilierens lassen sich einzelne Aspekte fokussieren und beurteilen.

Die Methoden der Informationsflussanalyse stellen als grösstes Problem die Identifikations-IDs (IMEI, IMSI, Phone Number) heraus. Daneben wird besonders der Standort abgefragt. Dies ist meist für Werbenetzwerke interessant, die bezüglich des Standortes personalisierte Werbung anzeigen und den Anwender so gezielt beeinflussen wollen. Die Ergebnisse aller Methoden unterscheiden sich nur unwesentlich. Die statische Daten- und Informationsflussanalyse kann bei der Analyse helfen, allerdings sind die Ergebnisse immer als Tendenz zu verstehen.

Literatur

[BJK13] Alexandre Bartel und Martin Monperrus Jacques Klein, Yves Le Traon. Dexpler: Converting Android Dalvik Bytecode to Jimple for Static Analysis with Soot. 2013.

[FAR+13] Christian Fritz, Steven Arzt, Siegfried Rasthofer, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves le Traon, Damien Octeau und Patrick McDaniel. Highly Precise Taint Analysis for Android Applications. 2013.

Algorithmenbeschleunigung durch FPGAs zur Massendatenverarbeitung

Christian Meltebrink
Fakultät Ingenieurwissenschaften und Informatik, Hochschule Osnabrück
Christian.Meltebrink@hs-osnabrueck.de

In dem Forschungsprojekt „Hochperformante Verarbeitung und Visualisierung von Massendaten“ (HPVis)[1] der Hochschule Osnabrück wird ein herkömmlicher Standard-PC von einem externen „Field Programmable Gate Array“ (FPGA)-Board bei der Bearbeitung von rechenintensiven Aufgaben unterstützt [WKK13].

In HPVis wird eine Programmierschnittstelle (API) entwickelt, die dem Programmierer die Kommunikation mit einem FPGA-Board erleichtert. Eine rechenintensive Applikation wird in parallele Prozesse aufgeteilt. So können bestimmte Prozesse der Gesamtapplikation auf das FPGA ausgelagert werden. Die API wird durch die Angabe der beteiligten Prozesse sowie deren Kommunikationskanäle untereinander in einer Konfigurationsdatei spezifiziert. Anhand der entsprechenden Kanalnummer überträgt die API durch bereitgestellte Funktionen Daten zwischen zwei Prozessen. Für die Kommunikation zwischen zwei Prozessen auf dem PC verwendet die API Socket-Verbindungen. Für die Kommunikation zwischen einem Prozess auf dem PC und einem Prozess auf dem FPGA wurde ein PCI-Express-Treiber entwickelt. Dieser organisiert den Datenaustausch über PCI-Express [BAS03] anhand von Scatter/Gather-Listen mit einem „Direct Memory Access“ (DMA)-Controller auf dem FPGA-Board. Die zwei unterstützten Kommunikationswege sind in Abbildung 1 dargestellt.

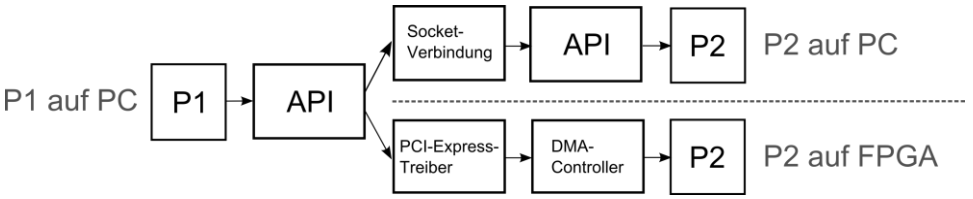


Abbildung 1: Unterstützte Kommunikationswege zwischen zwei Prozessen

Literaturverzeichnis

[BAS03] R. Budruk, D. Anderson, T. Shanley, “PCI Express System Architecture”, Mindshare PC System Architecture, Addison-Wesley Longman, Amsterdam, Sep. 2003

[WKK13] M. Weinhardt, A. Krieger, T. Kinder: “A Framework for PC Applications with Portable and Scalable FPGA Accelerators”, ReConFig 2013, Cancun, Mexiko, Dec. 2013

[1] HPVis wird durch den Europäischen Fonds für regionale Entwicklung (EFRE) und die Niedersächsische Landesregierung unterstützt (www.ecs.hs-osnabrueck.de/hpvis.html).

Intent-basierte Kommunikation im Web: Übersicht und Vergleich

Sarah Seifert

Universität Leipzig, Masterstudiengang Informatik
mam09dbm@studserv.uni-leipzig.de

Die Zahl an komplexen Webanwendungen nimmt stetig zu. Um Wiederverwendung zu gewährleisten, sollten diese Anwendungen lose gekoppelt und deren Schnittstelle einfach zu implementieren sein. Da letzteres beim häufig verwendeten SOA-Ansatz nicht gegeben ist, wurde basierend auf den im Android Betriebssystem etablierten Intents [ACS09] das Web Intents-Konzept [Kin11] entwickelt, welches diese Anforderung erfüllt. Aufgrund der sehr umfangreichen Gestaltung des Web Intents-Konzeptes wird auch an einem vereinfachten Gegenentwurf namens Web Activities [Moz13] gearbeitet.

Alle drei Intent-Ansätze weisen ein gemeinsames Grundkonzept auf. Jedes Intent ist ein Messaging-Objekt, das zwei wesentliche Eigenschaften besitzt: Das Attribut *action* repräsentiert die auszuführende Aktion in Form einer Zeichenkette (String), *data* enthält die zur Bearbeitung notwendigen (Nutz-)Daten. Ebenfalls zum Grundkonzept gehören *Services*, die eingehende Intents verarbeiten, sowie verschiedene *Handler*. Ein *Filtermechanismus* gewährleistet, dass die Intents ausschließlich an Services weitergeleitet werden, die deren Aktion verarbeiten können. Der Filter dient aber auch gleichzeitig zur Registrierung dieser Services. Je nach Ansatz können weitere optionale Attribute für das Intent-Objekt angegeben werden, die beispielsweise die Filterung verfeinern.

Sowohl Web Intents als auch Web Activities befinden sich noch in der Entwicklungsphase. Dadurch können derzeit Unstimmigkeiten auftreten oder die Anwendungsmöglichkeiten eingeschränkt sein. Das Potential dieser Ansätze ist jedoch gegeben, sodass sie sich langfristig gesehen als einfacher zu implementierende Alternativen gegenüber bestehenden Ansätzen wie SOA etablieren könnten.

Literatur

[ACS09] Frank Ableson, Charlie Collins und Robi Sen. *Unlocking Android. A Developer's Guide*. Mannig Publications Co., 2009.

[Kin11] Paul Kinlan. Web Intents. `http://webintents.org/`, 2011. Letzter Aufruf: 27. März 2014.

[Moz13] Mozilla. WebAPI/Web Activities. `https://wiki.mozilla.org/WebAPI/WebActivities`, 2013. Letzter Aufruf: 27. März 2014.

3D-Rekonstruktion mit der AR.Drone 2.0

Hendrik Geßner, Matthias Weise

Universität Potsdam - Institut für Informatik
hgessner@uni-potsdam.de
matweise@uni-potsdam.de

Ein Geologe steht in einem unbekannten Höhlensystem. Er startet vier Drohnen, die sich selbstständig ihren Weg suchen. Die zurückgelegte Strecke wird in Echtzeit kartografiert und als dreidimensionales Modell im Hauptquartier angezeigt. Die beschriebene Szene stammt aus dem Film „Prometheus - Dunkle Zeichen“ aus dem Jahr 2012 und gab die Motivation für das hier vorgestellten Framework. Es sollte überprüft werden, wie nah eine heute existierende Lösung an die im Film dargestellte 3D-Kartografierung heranreicht. Dabei wird softwareseitig nur auf kostenlose und bereits existierende Software zurückgegriffen, hardwareseitig werden eine AR.Drone 2.0 sowie ein handelsüblicher PC genutzt. In Abbildung 1 ist der komplette Ablauf dargestellt.

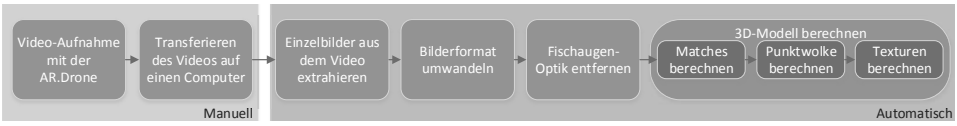


Abbildung 1: Ablauf der erforderlichen Arbeitsschritte [Quelle: eigene Darstellung]

Für die Videoaufnahmen muss die Drohne manuell gestartet und geflogen werden. Nachdem das Video auf einen PC übertragen wurde, kann die automatische 3D-Rekonstruktion gestartet werden. *FFmpeg* extrahiert Einzelbilder aus dem Video, deren Format von *ImageMagick* umgewandelt wird. Durch die Fischaugenoptik der Drohne entstehen Verzerrungen, die mit Hilfe des Lens Distortion Filter von *Gimp* entfernt werden. Für die abschließende 3D-Rekonstruktion wird *VisualSFM* eingesetzt. Die entstehenden 3D-Modelle werden im PLY-Format gespeichert.

Mit Hilfe der Drohne wurden mehrere Aufnahmen gemacht und ausgewertet, zudem wurden Vergleichsaufnahmen mit einer Smartphone-Kamera gemacht, auf denen die gleichen Verarbeitungsschritte durchgeführt wie auf den Drohnenaufnahmen. In der Gegenüberstellung fallen verschiedene Eigenschaften der 3D-Modelle aus Drohnenaufnahmen auf. So sorgt der in der Höhe anpassbare Blickwinkel für ein vollständigeres Modell als die Smartphone-Aufnahmen, allerdings finden sich Krümmungen und Artefakte, die in den 3D-Modellen aus Smartphoneaufnahmen nicht auftreten. Die Gesamtverarbeitungsdauer der Drohnenaufnahmen entspricht dem 13-fachen der Aufnahmedauer. Die hier nicht beleuchteten Aspekte Autonomie und 3D-Darstellung stellen zukünftige Arbeitsfelder dar.

Efficient Regression for Big Data Problems using Adaptive Sparse Grids

Michael Lettrich
m.lettrich@mytum.de
Fakutät für Informatik, Technische Universität München

The amount of available data increases rapidly. This trend, often related to as *Big Data* challenges modern data mining algorithms, requiring new methods that can cope with very large, multi-variate regression problems. A promising approach that can tackle non-linear, higher-dimensional problems is regression using sparse grids. Sparse grids use a multi-scale system of grids with basis functions φ with local support to circumvent the curse of dimensionality [BG04].

Sparse grid based regression [Pfl10] iteratively constructs a function $\hat{f}(x) = \sum_i \alpha_i \varphi_i(x)$ that approximates the solution of the regression problem f finding optimal weights α_i for a fixed grid using least squares optimization. For further error reduction, refinement algorithms then add new basis functions and the optimization is repeated.

Current refinement algorithms add hierarchical child functions in all d input dimensions if a neighborhood with a high local error is discovered. This results in the addition of less beneficial basis functions if not all input dimensions are equally important. Instead, we propose a heuristic method that identifies and introduces only new basis functions with the potentially highest error reduction. We found, that the average of the local error r at data points x_i weighted with the evaluation of the basis function $\varphi(x_i)$ in question is suited as a refinement indicator $\alpha \approx n^{-1} \sum_{i=1}^{n} \varphi(x_i) r(x_i)$, as we then can use the heuristics from [Pfl10] to estimate potential error reduction. The number of data points n required for a significant approximation is in $\mathcal{O}(d)$.

The algorithm has been tested with both synthetic and real world datasets from the UCI repository. We have found that our algorithm offers a 30 to 50 percent improvement in error decay over current methods while reducing the grid size between 40 to 75 percent, allowing us to tackle large, high dimensional datasets more efficiently. The accuracy remains comparable with current approaches.

References

[BG04] Hans-Joachim Bungartz and Michael Griebel. Sparse grids. *Acta Numerica*, 13:147–269, 2004.

[Pfl10] Dirk Pflüger. *Spatially Adaptive Sparse Grids for High Dimensional Problems*. PhD thesis, Technische Universität München, 02 2010.

Integrata-Stiftung für humane Nutzung der Informationstechnologie (Hrsg.)

Privatheit in der E-Society

Kongress der Integrata-Stiftung für humane Nutzung der Informationstechnologie

25. September 2014 in Stuttgart

Integrata-Stiftung für humane Nutzung der Informationstechnologie

Vorwort

Vorstand
Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
hans-peter.froeschle@integrata-stiftung .de

Der Integrata-Kongress 2014 steht unter dem Motto „Privatheit in der E-Society“.

Impulsreferate, Keynotes sowie die Präsentation und Diskussion eines Positionspapier zu Privatheit in der E-Society beleuchten die vielfältigen Aspekte des Mottos schlaglichtartig und sollen zu einer weiteren Diskussion in Wissenschaft und Gesellschaft anregen.

Das Konzept für ein Gütesiegel „Humane Nutzung der Informationstechnologie“ wird der Öffentlichkeit vorgestellt.

Abgerundet wird der Integrata-Kongress durch die Verleihung des Wolfgang-Heilmann-Preises für humane Nutzung der Informationstechnologie.

Profilierung menschlichen Verhaltens

Dr. Constanze Kurz

Forschungsgruppe INKA
Hochschule für Technik und Wirtschaft (HTW) Berlin
Wilhelminenhofstr. 75a
12459 Berlin
constanze.kurz@htw-berlin.de

Abstract: Nicht mehr nur für die sogenannte Netz-Community und technisch Interessierte ist die Regulierung und die Gestaltung der Grundzüge der Netz-, Daten- und Technologiepolitik von Bedeutung, sondern längst ein Thema für die breite Öffentlichkeit. Mit den Snowden-Enthüllungen, aber auch durch die Masse an öffentlich diskutierten IT-Sicherheitsprobleme und den Streit um die Macht der Internetkonzerne ist eine Vertrauenskrise entstanden, die es nötig macht, die Weichen für die digitale Gesellschaft neu zu stellen.

Heartbleed: Lessons from the System Failure

Christof Leng

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

Abstract: The Heartbleed bug in OpenSSL has both be described as one of the worst security vulnerabilities in the history of the internet and as a simple programming error. I argue that vulnerabilities like Heartbleed are not a coincidence or inevitable, but a result of the overall process underlying the design and implementation of current internet protocols. In order to minimize the probability of similarly catastrophic failures in the future, many aspects of the protocol engineering culture need to be revisited. Since the internet has become a fundamental infrastructure for almost every aspect of our society, it is crucial to address these issues and to (re-)establish trust in its security and reliability.

1 Introduction

On April 7th 2014, news about the discovery of the Heartbleed bug hit mainstream media worldwide. A rather trivial missing check in the Heartbeat extension of the popular OpenSSL library potentially compromised hundreds of thousands of servers worldwide and might have exposed user data, passwords, and even private encryption keys of millions of users and websites.

OpenSSL is the most popular open source implementation of the internet transport security protocols SSL, TLS, and DTLS, and is part of many Linux distributions, network appliances like home routers, and a large fraction of HTTPS-secured websites. The Secure Socket Layer (SSL) [Hic95] was an attempt to add security to TCP-based application protocols like HTTP, SMTP, or FTP. It later evolved into Transport Layer Security (TLS) [Die08], which currently is available as version 1.2, and turns the application protocols into their secured counterparts HTTPS, SMTPS, and FTPS. Datagram Transport Layer Security (DTLS) [RM12] is a version of TLS that provides security for connectionless application protocols that are based on UDP or similar transport protocols. It is used by some Virtual Private Network (VPN) software or the WebRTC protocol for browser-to-browser real-time communication, among others.

It is not the intention of this paper to blame any individual, project, or organization for the Heartbleed bug or any circumstance that may have promoted it. Anyone who dedicates time to the design, implementation, and operation of the internet deserves our deepest

respect, especially if the contributions are made in the form of open standards or free software. Instead, this discussion is an attempt to highlight potential shortcomings of the status quo and to make suggestions on how to improve the situation in the common interest.

2 Background

In order to understand the Heartbleed bug and the various mechanisms that may have promoted it, we first need to understand TLS/DTLS and its heartbeat extension.

The OpenSSL project dates back to 1998 and is based on the earlier SSLeay implementation. The SSL protocol was first developed by Netscape and publicly released as version 2.0 and 3.0. It was later standardized by the Internet Engineering Task Force (IETF) under the name TLS. The IETF publishes its standard documents under the name request for comment (RFC). The current version of TLS is 1.2. The TLS protocol is extendable and numerous extensions exist, which are also defined in RFCs. DTLS re-uses most of the TLS structure and only modifies it where necessary to preserve the unreliable datagram semantics of the underlying transport.

Both protocols use a record structure to encapsulate the application protocol messages and their own control messages. The header of the record layer consists of a content type field, a protocol version number, and a length field (see Figure 1). DTLS adds a sequence number and an epoch field. In the remainder of the paper, TLS record is used to describe both DTLS and TLS records. The type field is a protocol discriminator, which determines how the content of the record has to be processed.

2.1 Heartbeat Use-Cases

The heartbeat extension defines one of the currently five TLS record content types and was defined in RFC 6520 [TSW12]. It has multiple use cases, among them keep-alive, path MTU discovery, and mobility support.

Keep-alive is a mechanism to detect unresponsive communication partners. A host sends a keep-alive request to its communication partner, and this receiver replies with a keep-alive response by echoing the payload of the request. The sender typically sends a counter value, pointing to a local table entry to keep responses apart.

Path MTU Discovery (PMTUD) is used to determine the maximum transfer unit (MTU) on the communication path between the two hosts. The MTU is the maximum message size intermediate routers can forward to the destination. PMTUD sends a series of probe packets to pinpoint the MTU. Standard-compliant routers unable to forward such a packet will respond with a “Fragmentation Needed” ICMP message and its MTU. The sender adjusts the probe packet size accordingly and the process is repeated until the destination is reached, which will notify the sender of the success. The heartbeat extension uses a padding field with random data to generate probe packets of the required sizes. The

payload, which is echoed back upon success, is used to determine the path MTU.

The heartbeat extension was designed to accommodate future uses like mobility support. In modern internet use cases, hosts may change their public IP address at any time, often even without being able to detect the change themselves, because an intermediate middlebox (e.g., a NAT gateway) is responsible for the change. In this case, the connection breaks, because the packets can no longer be routed to the correct receiver. Heartbeats can help in two ways. First of all, a client that periodically sends heartbeat messages to the server, such that the server can detect the changed sender address even if the client sends no other data. The server then has to check the authenticity of the client. It sends a special heartbeat message as a challenge to the client. Only if the client is able to decrypt this message and echo the payload, the server sends future messages to the new address. In order to avoid resource attacks with fake client heartbeats, the server encodes all necessary information to process the reply in its heartbeat request. Mobility support in DTLS require a modified record format to identify the connection despite the address change. Therefore, it currently is only a theoretical concept.

2.2 Heartbeat Design

Heartbeat requests and replies for all different use-cases share the same message format, which is a superset of the required functionality for each scenario. The content type of this message format is 24. The message contains a one byte type field, which distinguishes between requests and replies (see Figure 1). It also contains a variable length payload field and a variable length padding field. The length of the payload field is determined by a two byte payload_length field. The length of the padding field can only be determined by subtracting the length of the other fields from the length field in the TLS record header. The formula is $padding_length = length - 1 - 2 - payload_length$. Additionally, the RFC requires the padding_length to be at least 16.

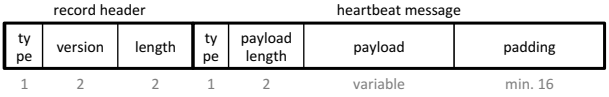


Figure 1: Heartbeat message format

In order to detect malformed heartbeat requests, the receiver must check that the message has at least the minimum length of 1 + 2 + 16, and that the content of the payload_length field is consistent with the total length of the record. Thus, $payload_length \leq length - 1 - 2 - 16$. If the latter check is missing, a malicious sender may send a payload_length field much larger than the actual record size. The receiver then not only copies the actual payload into the heartbeat response, but also the memory content adjacent to it. By sending a very small heartbeat request containing no payload or padding at all, but a payload_length of the maximum value of 65535, the attacker can receive almost 64KB of memory content

from the attacked. This is exactly how the Heartbleed attack works.

This is a very severe vulnerability, because it returns memory from more or less random locations, and repeated attacks can reveal critical content like passwords or private keys with a high probability. Additionally, the attack does not trigger any error, log message, or modify the attacked system's state and therefore leaves no trace at all.

The faulty heartbeat implementation in OpenSSL was accepted by the project maintainer on December 31st 2012 and has only been detected more than two years later. It is practically impossible to determine if it has been known to and used by attackers before its public disclosure.

3 Causes of Heartbleed

Superficially, Heartbleed is caused by the missing of a trivial security check that has been overlooked by the programmer and the reviewer. On the one hand, this may sound comforting, because it implies that there is nothing fundamentally wrong. On the other hand, because human error is inevitable, we must anticipate future bugs with similar or worse consequences. In the following, we will investigate a wide range of circumstances that have facilitated the bug and need to be addressed to reduce the probability of future vulnerabilities.

3.1 Programming Methods

The most obvious problem with the programming side is that Hearbleed is a buffer over-read bug. A buffer over-read is a type of buffer overrun, where data is read beyond the actual boundaries of the buffer. This is a problem known for more than 40 years [And72] and has well-known remedies. There is no excuse for buffer overruns in modern security-critical software, but they are still very frequent, because of the choice of **programming language**. C/C++ is probably the only still commonly used language that does not use automatic bounds checking by default. Unfortunately, C/C++ is especially dominant in systems programming and communication protocols.

Even with a language like C/C++, memory bugs can be minimized when the proper **tools** are used. Plain pointer arithmetic is more prone to buffer overruns than specialized data structures with internal boundary checks. Debugging tools like Valgrind [NS07] can help to discover memory bugs. Unfortunately, such tools would not work well with OpenSSL, because it uses (and abuses) its own custom memory management instead of using standard memory allocation [Una14]. OpenSSL also does not use comprehensive unit testing as a tool for automated quality assurance. In the aftermath of Heartbleed, an initiative was started to add more unit tests to OpenSSL [Bla14].

The challenge is aggravated by the **complexity** of the OpenSSL codebase. Currently, it consists of more than 450,000 lines of code [Ohl14]. That is more than five times the

size of Google's QUIC, an experimental protocol that combines transport functionality with transport security. OpenSSL can only be blamed in part for this complexity, because SSL/TLS itself is a very complex standard (see next Section), which requires a lot of code to implement. Nonetheless, the LibReSSL fork [Bec14] managed to remove 150,000 lines of content from the code base within a few weeks.

With an unsafe programming style and without a proper tool chain for automated tests, the manual quality assurance becomes the crucial part of the process. The OpenSSL does manual **code reviews** of submitted source code contributions. The heartbeat extension did pass that check. Code reviews are an important tool for quality assurance, but can only reduce the probability of programming errors slipping through. Eric S. Raymond's so-called Linus's Law states that "given enough eyeballs, all bugs are shallow", which may already be debatable, but in the case of OpenSSL there were definitely not enough eyeballs. Despite being used all over the internet by many companies, governments, and enthusiasts, the project has been notoriously underfunded and thus only one developer could afford to work full-time on the project. A few weeks after Heartbleed, the Linux Foundation has formed a Core Infrastructure Initiative of IT companies providing substantial funding to OpenSSL and similar projects to improve the situation [The14]. Projects like OpenSSL seem to suffer from the "tragedy of the commons": Even though the software is used by millions or billions of users, countless companies, and more than a few governments, only very few individuals take responsibility and contribute money or time and simply rely on the work of a few altruistic volunteers.

Given the circumstances, it must be rated as a combination of the effort of these individuals and pure luck that not more vulnerabilities of this magnitude have surfaced. Without a change in system programming culture and responsibility culture, the outlook is bleak.

3.2 Design Process

Heartbleed is not simply a programming problem. It roots much deeper into the design process of TLS and DTLS. First of all and most fundamentally, the whole concept of TLS vs. DTLS must put into question. As explained earlier, DTLS shares many concepts, formats, and even implementations with TLS. The goal of this approach was to minimize the overall effort and the ability to solve problems once instead of twice. Instead, a bug in the heartbeats, which only make sense for DTLS, affected the much more widespread and critical TLS. TLS's underlying transport, TCP, already provides perfectly fine PMTUD and keep-alive functionality. The mobility mechanism that could be built on top of heartbeats would not even work with TCP, because TCP connections cannot be redirected to a new IP address like this.

Even worse, DTLS only needs a heartbeat mechanism because it introduces concepts to connection-less datagram protocols which are intrinsic to connection-oriented stream protocols. A stateless UDP-based application protocol does not need a keep-alive, because there is no connection. But DTLS introduces a handshake to establish symmetric encryption and has to maintain the associated state. Thus, it needs to know, when this state can be

thrown away. UDP-based application protocols also need to avoid packet fragmentation, because fragmented packets negatively impact the loss probability. Either the application protocol provides its own PMTUD or it avoids large datagrams altogether. Yet, the "magic" DTLS layer does introduce large messages itself, e.g., the certificate exchange in the connection handshake. Trying to make two things look the same, which are fundamentally different, is dangerous **over-generalization**.

This conceptual problem becomes evident through **contradictions** in the standards document. While RFC 6520 states heartbeats are meant as "a basis for path MTU (PMTU) discovery for DTLS" [TSW12], the DTLS 1.2 specification [RM12] specifically claims "DTLS's philosophy is to leave PMTU discovery to the application". The publication of the two RFCs is just one month apart, both have been in the standardization process in parallel.

```
4.  Heartbeat Request and Response Messages

   The Heartbeat protocol messages consist of their type and an
   arbitrary payload and padding.

   struct {
      HeartbeatMessageType type;
      uint16 payload_length;
      opaque payload[HeartbeatMessage.payload_length];
      opaque padding[padding_length];
   } HeartbeatMessage;

   The total length of a HeartbeatMessage MUST NOT exceed 2^14 or
   max_fragment_length when negotiated as defined in [RFC6066].

   type:  The message type, either heartbeat_request or
      heartbeat_response.

   payload_length:  The length of the payload.

   payload:  The payload consists of arbitrary content.

   padding:  The padding is random content that MUST be ignored by the
      receiver.  The length of a HeartbeatMessage is TLSPlaintext.length
      for TLS and DTLSPlaintext.length for DTLS.  Furthermore, the
      length of the type field is 1 byte, and the length of the
      payload_length is 2.  Therefore, the padding_length is
      TLSPlaintext.length - payload_length - 3 for TLS and
      DTLSPlaintext.length - payload_length - 3 for DTLS.  The
      padding_length MUST be at least 16.

   The sender of a HeartbeatMessage MUST use a random padding of at
   least 16 bytes.  The padding of a received HeartbeatMessage message
   MUST be ignored.

   If the payload_length of a received HeartbeatMessage is too large,
   the received HeartbeatMessage MUST be discarded silently.
```

Figure 2: The section of RFC 6520 that describes the critical length check

Even without semantic contradictions, RFCs are often ambiguous and a source of misinterpretation. RFC 2616, the original HTTP/1.1 specification, has recently been replaced with six new RFCs (7230-7235), which do not add any new functionality, but simply rephrase the original standard to minimize the **misinterpretation and ambiguity** problems. This process took seven years. Most RFCs do not receive this much attention.

The information about the critical bounds check that led to Heartbleed is scattered over the message format specification in RFC 6520 (see Figure 2). The actual formula is hidden in the description of the padding field and has to be solved for payload_length for the implementation. In this particular case, the co-author of the RFC and the programmer of the OpenSSL implementation are the same person, so this is probably not a problem of text comprehension, but it proves that highlighting potential security problems was not a priority in the standardization process. The lack of **formalism** in RFCs culminates in the absence of standardized tests, which could be used to check the standard conformance of an implementation.

The biggest problem of DTLS/TLS is probably its **feature bloat**. Despite its critical nature as one of the internet's core security components and the consequential need for conciseness, the protocol suite overflows with extensions. Wikipedia currently lists 42 RFCs covering TLS (see Figure 3). This list is not comprehensive. For example, it does not include RFC 6520 (heartbeats). A complete implementation of TLS must cover all these standards, even the outdated ones, because of the need of backwards compatibility with legacy clients and servers. The enormous complexity of the standard does however hinder the widespread adoption of the current version. DTLS/TLS does not adhere to the Unix philosophy: "Write programs that do one thing and do it well", which is of special importance of security critical components, which need to minimize their attack surface.

Finally, the heartbeat extension did introduce unnecessary complexity that neither adds functionality nor security. The 16 byte minimum padding is not justified anywhere in the RFC. The PhD thesis of one of the co-authors states that "without additional random bytes this payload can be guessed easily, thus making it prone to a Known-Plaintext Attack (KPA)" [Seg12], because heartbeat implementations may use simple counters as the request payload. Modern ciphers like AES are specifically designed to be immune to KPAs, and even if a weakness that makes KPAs realistic would be found, the DTLS heartbeats would be our least concerns. Countless applications encrypt easy-to-guess plaintexts with the ciphers available in DTLS/TLS, especially all the standard internet protocols like HTTP, FTP, and SMTP, which are encapsulated in TLS. If TLS does not add random padding in general to prevent attacks, there is no point adding them in a rarely-used extension.

This issue points to a deeper problem: **cargo cult security**. Computer security and cryptography are highly complex and demanding fields, and any small mistake may jeopardize the overall security. Nonetheless, security-critical components are often designed and implemented by system programmers with little or no background in security. They mimic the methods of security experts without fully understanding the underlying principles, much like the cults of oceanic cultures who tried to mimic the "rituals" of Japanese and Allied military to attract airplanes dropping valuable cargo. We need to strengthen the ties between the systems and security communities to improve the mutual understanding.

Overall, one must consider DTLS/TLS as a failed protocol. Its complexity, contradictions, and legacy support requirements make it impossible to fix. New transport protocols like CUSP [TLLB10] or QUIC [Ros13] consequently integrate security into their own functionality instead of interfacing with TLS. Any such attempt must minimize the complexity of the protocol and provide a coherent specification. Formal methods for protocol design,

Protocol versions from SSL 2.0 to TLS 1.2:

- Hickman, Kipp E.B. "The SSL Protocol". (Internet Draft)
- RFC 6101: "The Secure Sockets Layer (SSL) Protocol Version 3.0".
- RFC 2246: "The TLS Protocol Version 1.0".
- RFC 4346: "The Transport Layer Security (TLS) Protocol Version 1.1".
- RFC 5246: "The Transport Layer Security (TLS) Protocol Version 1.2".

Extensions to TLS 1.0 include:

- RFC 2595: "Using TLS with IMAP, POP3 and ACAP".
- RFC 2712: "Addition of Kerberos Cipher Suites to Transport Layer Security (TLS)".
- RFC 2817: "Upgrading to TLS Within HTTP/1.1".
- RFC 2818: "HTTP Over TLS".
- RFC 3207: "SMTP Service Extension for Secure SMTP over Transport Layer Security".
- RFC 3268: "AES Ciphersuites for TLS".
- RFC 3546: "Transport Layer Security (TLS) Extensions".
- RFC 3749: "Transport Layer Security Protocol Compression Methods".
- RFC 3943: "Transport Layer Security (TLS) Protocol Compression Using Lempel-Ziv-Stac (LZS)".
- RFC 4132: "Addition of Camellia Cipher Suites to Transport Layer Security (TLS)".
- RFC 4162: "Addition of SEED Cipher Suites to Transport Layer Security (TLS)".
- RFC 4217: "Securing FTP with TLS".
- RFC 4279: "Pre-Shared Key Ciphersuites for Transport Layer Security (TLS)".

Extensions to TLS 1.1 include:

- RFC 4347: "Datagram Transport Layer Security".
- RFC 4366: "Transport Layer Security (TLS) Extensions".
- RFC 4492: "Elliptic Curve Cryptography (ECC) Cipher Suites for Transport Layer Security (TLS)".
- RFC 4680: "TLS Handshake Message for Supplemental Data".
- RFC 4681: "TLS User Mapping Extension".
- RFC 4785: "Pre-Shared Key (PSK) Ciphersuites with NULL Encryption for Transport Layer Security (TLS)".
- RFC 5054: "Using the Secure Remote Password (SRP) Protocol for TLS Authentication".
- RFC 5077: "Transport Layer Security (TLS) Session Resumption without Server-Side State".
- RFC 5081: "Using OpenPGP Keys for Transport Layer Security (TLS) Authentication".

Extensions to TLS 1.2 include:

- RFC 5288: "AES Galois Counter Mode (GCM) Cipher Suites for TLS".
- RFC 5289: "TLS Elliptic Curve Cipher Suites with SHA-256/384 and AES Galois Counter Mode (GCM)".
- RFC 5746: "Transport Layer Security (TLS) Renegotiation Indication Extension".
- RFC 5878: "Transport Layer Security (TLS) Authorization Extensions".
- RFC 5932: "Camellia Cipher Suites for TLS".
- RFC 6066: "Transport Layer Security (TLS) Extensions: Extension Definitions".
- RFC 6091: "Using OpenPGP Keys for Transport Layer Security (TLS) Authentication".
- RFC 6176: "Prohibiting Secure Sockets Layer (SSL) Version 2.0".
- RFC 6209: "Addition of the ARIA Cipher Suites to Transport Layer Security (TLS)".
- RFC 6347: "Datagram Transport Layer Security Version 1.2".
- RFC 6367: "Addition of the Camellia Cipher Suites to Transport Layer Security (TLS)".
- RFC 6460: "Suite B Profile for Transport Layer Security (TLS)".
- RFC 6655: "AES-CCM Cipher Suites for Transport Layer Security (TLS)".
- RFC 7027: "Elliptic Curve Cryptography (ECC) Brainpool Curves for Transport Layer Security (TLS)".

Encapsulations of TLS include:

- RFC 5216: "The EAP-TLS Authentication Protocol"

Figure 3: List of relevant RFC from the Wikipedia article on Transport Layer Security

like language security based approaches [PB11], and a more sophisticated standardization model than traditional RFCs should be considered in this context.

3.3 Computer Science Culture

The internet may be the most outstanding success of academic computer science research. Within a single generation, basic research turned into a world-spanning infrastructure that is ubiquitous to virtually every aspect of modern society. When claiming its success,

one must also take responsibility for its failures. Like many other internet protocols, the heartbeat extension is the result of an academic research project. This is no coincidence and our academic culture actually promotes the problematic practice behind it.

First of all, the **publish or perish** culture encourages to produce more instead of better results. A PhD thesis that yielded only a handful or even no paper at all is commonly considered a failure. On the other hand, the peer review process is not sufficient to ensure a minimum quality. To put it bluntly, any programming project paper can be published at some workshop as long as it comes with a few performance figures. Making matters worse, the practicality of an approach is often "proven" by patching popular open-source implementations and submitting the patch to the upstream maintainers. Responsible peer review cannot be outsourced to volunteer code reviewers from the open-source community.

While publishing positive results is too easy, publishing **negative results** is much harder[1]. Papers with a "we did not do X and here is why" approach are rare, especially at prestigious conferences. This is fatal in two ways. First, it encourages researchers to focus on building systems instead of reflecting on the underlying design principles. Second, it impedes the distribution of knowledge about fallacies, dead ends, and pitfalls.

Another problem roots in the success of the internet. It is hard to argue against the success stories of TCP or TLS. Many experts consider the transport layer a "done thing". Only minor tweaks seem practical, a paradigm shift in contrast utopian. This **conservation culture** slows down innovation in the field.

However, giving up completely on the existing infrastructure and building **castles in the air** with clean-slate internet architectures that are not practical does not help the situation either. Any realistic solution must be incrementally deployable on the current internet infrastructure. But that is actually less difficult than skeptics make it seem.

4 An Alternative Heartbeat Proposal

While I personally believe that DTLS/TLS is broken beyond repair[2] and that they should be abandoned as soon as possible, so that all energy can be focused on the design of a sane replacement, I will take the time to sketch an alternative design for DTLS heartbeats to show that it would have been simple to reduce the attack surface. That said, I believe that there is no place for keep-alive and PMTUD functionality in transport security. Those features exclusively belong to the traditional transport layer. If you need them in DTLS/TLS, it only indicates that you are doing something fundamentally wrong and should start over with your design.

First of all, the **redundant functionality** of the type field in the record header and inside the individual record type is superfluous. Using the type field in the record header exclusively as a "protocol discriminator" is completely unnecessary. No security-critical protocol should need to use more than the 256 different message types that fit in the one

[1] This paper being the exception to the rule, of course.

[2] The same goes for many of the associated technologies like ASN.1 and concepts like certificate authorities.

byte type field, because many message types imply high complexity. Actually, TLS uses just 50 message types (which already is a lot), even if all 30 different alerts are counted as individual types [IAN14].

Secondly, the design should observe the **separation of concerns**. Requests and responses, keep-alive and PMTUD messages should not share the same message type and format. Aside from simplifying the message layout and message handlers, this allows to selectively turn these features on or off as needed. A bug in an unneeded and thus disabled extension would neither impact security nor functionality of actually needed extensions.

Finally, **variable-length fields** should be avoided wherever possible. Without a length field, there is no possibility of a missing or wrong boundary check. In TLS, we are stuck with the length field in the record header, but we should not add additional length fields without need. For keep-alive and PMTUD, a simple fixed-length identifier is sufficient to map responses to outstanding requests.

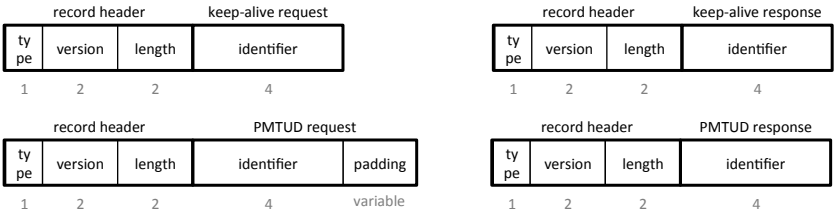


Figure 4: Alternative heartbeat message format

This leaves us with four simple message types (see Figure 4), each of them with a clear purpose and concise specification. The payloads of a keep-alive request and a keep-alive response each consist of a 4 byte identifier. A receiver only needs to check if the record content has a length of four. Even if this check was absent, an attacker could read just four bytes of random memory from the target (instead of 65,535 bytes).

The path MTU discovery needs minimal extra functionality in the form of a padding field in the request. The content of this field is irrelevant and only there to test the routers on the path. The receiver only reads and echoes the identifier. The only change in the handler is to assert a length of at least four instead of exactly four.

A specific message format for the hypothetical mobility extension seems unnecessary. As proposed in the original design, the client can use normal keep-alive messages to make address changes visible to the server. In the case of an unexpected sender address, the server can simply reply with an alert message notifying the client of the change. The server does not need to keep any state. The client can then prove its identity by initiating a TLS session resumption [SZET08]. By going through a connection reestablishment with the same master secret, the possession of the credentials can be validated by the server. If the necessary functionality is already in the standard, it is unnecessary to reinvent it.

5 Dealing with Heartbleed Attacks

Heartbleed attacks are especially dangerous because they leave no trace. An intrusion detection system (IDS) can be taught to detect Heartbleed attacks reliably [Ama14], but that helps only against future attacks. Therefore, it is practically impossible to decide if the vulnerability has been abused before its public disclosure.

The only method to know for sure is network forensics on traffic dumps of the complete timespan that the vulnerability was available. Practically no one keeps comprehensive traffic dumps for over two years (if at all). At ICSI, we keep one to two months of traffic dumps for research purposes. The Lawrence Berkeley National Lab (LBNL) keeps up to three months. With these and the updated IDS scripts, a large number of attacks can be detected after the public disclosure, but none in the time before (see Figure 5). It has to be noted that many of the detected attacks are likely to be deliberate self-tests by the system administrators.

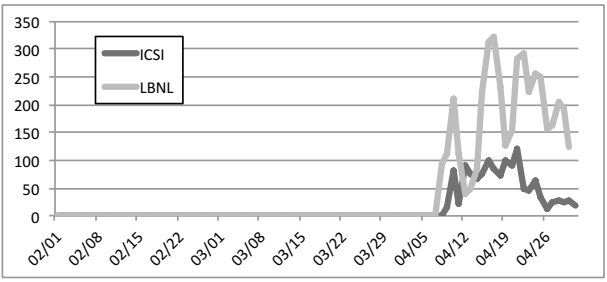


Figure 5: Detected Heartbleed attacks at ICSI and LBNL

The result only shows that the two networks were not attacked within the given timeframe. An earlier attack may have gone undetected. Also, there might have been specifically tailored attacks against other organizations. Without comprehensive traffic dumps covering the complete attack window and all possible targets, the only thing we can conclude with reasonable certainty is that the vulnerability was not used for random 'carpet bombing' attacks recently.

Network forensics is a powerful tool in this context, but it is also a dangerous one. In order to be able to detect and analyze attacks post-mortem, one must keep large amounts of potentially sensitive communication data over a long period of time. This information may become a viable target of attacks itself and thus must be secured carefully. Many users are unable to provide the necessary security and administration locally. Storing the communication data at a campus level or centrally at an internet service provider does aggravate the attack potential of the data and increases the risk of unwanted eavesdroppers.

Furthermore, the forensic analysis is only possible, because the attacks are rather simple and the confidentiality of TLS is limited (record headers are never encrypted). A better transport security or a more complex vulnerability could make forensic analysis of traffic dumps impossible. Surrendering encryption in favor of network forensics is certainly not

the right answer. There currently is no common interface that allows to capture the traffic before encryption either, and implementing one would create a dangerous new attack surface. Therefore, the ability to detect and analyze such attacks may diminish further in the future, which emphasizes the importance of a paradigm shift in protocol engineering and implementation.

6 Conclusion

Heartbleed is not just a programming error. It is a chain reaction of failures in the current way of designing and building critical internet protocols. A bug as bad as Heartbleed or even worse can happen again anytime, and it will if we do not change the culture of researching, designing, and implementing internet protocols. Because the internet has become a crucial infrastructure of the global society, it would be irresponsible not to take action.

We need to apply state-of-the-art software engineering methods and tools, need better written and more concise specifications, and should acknowledge that practical computer science is not about building everything that is possible but about understanding how to build good systems. In summary, we need to put security and reliability first and break with bad habits.

Acknowledgments

I would like to thank Johanna Amann for the measurement of Heartbleed attacks in ICSI and LNBL datasets.

References

[Ama14] Johanna Amann. Detecting the Heartbleed Bug Using Bro. `http://blog.bro.org/2014/04/detecting-heartbleed-bug-using-bro.html`, April 2014.

[And72] James P Anderson. Computer Security Technology Planning Study. Volume 2. Technical report, DTIC Document, 1972.

[Bec14] Bob Beck. LibreSSL - An OpenSSL Replacement. `http://www.libressl.org`, 2014.

[Bla14] Mike Bland. Call for OpenSSL Testing Help. `http://mike-bland.com/2014/06/02/call-for-openssl-testing-help.html`, June 2014.

[Die08] Tim Dierks. The Transport Layer Security (TLS) Protocol Version 1.2. RFC 5246, August 2008.

[Hic95] Kipp Hickman. The SSL protocol. Internet Draft, April 1995.

[IAN14] IANA. Transport Layer Security (TLS) Parameters. http://www.iana.org/assignments/tls-parameters/tls-parameters.xhtml, June 2014.

[NS07] Nicholas Nethercote and Julian Seward. Valgrind: A Framework for Heavyweight Dynamic Binary Instrumentation. In *ACM Sigplan Notices*, volume 42, pages 89–100. ACM, 2007.

[Ohl14] Ohloh Black Duck Open Hub. The OpenSSL Open Source Project. http://www.ohloh.net/p/openssl, 2014.

[PB11] Meredith L. Patterson and Sergey Bratus. The Science of Insecurity. 28c3 presentation, December 2011.

[RM12] Eric Rescorla and Nagendra Modadugu. Datagram Transport Layer Security Version 1.2. RFC 6347, January 2012.

[Ros13] Jim Roskind. QUIC: Quick UDP Internet Connections. https://docs.google.com/document/d/1RNHkx_VvKWyWg6Lr8SZ-saqsQx7rFV-ev2jRFUoVD34/, December 2013.

[Seg12] Robin Seggelmann. *SCTP: Strategies to Secure End-To-End Communication*. PhD thesis, Universität Duisburg-Essen, October 2012.

[SZET08] Joseph Salowey, Hao Zhou, Pasi Eronen, and Hannes Tschofenig. Transport Layer Security (TLS) Session Resumption without Server-Side State. RFC 4507, January 2008.

[The14] The Linux Foundation. Core Infrastructure Initiative. http://www.linuxfoundation.org/programs/core-infrastructure-initiative, April 2014.

[TLLB10] Wesley W. Terpstra, Christof Leng, Max Lehn, and Alejandro P. Buchmann. Channel-based Unidirectional Stream Protocol (CUSP). In *Proceedings of IEEE INFOCOM'10*, March 2010.

[TSW12] Michael Tuexen, Robin Seggelmann, and Michael Williams. Transport Layer Security (TLS) and Datagram Transport Layer Security (DTLS) Heartbeat Extension. RFC 6520, February 2012.

[Una14] Ted Unangst. Analysis of OpenSSL Freelist Reuse. http://www.tedunangst.com/flak/post/analysis-of-openssl-freelist-reuse, April 2014.

Bankbook – Sicherung der Privatheit in einer experimentellen Bankfiliale bei Facebook

Prof. Dr. Gottfried Vossen

Institut für Wirtschaftsinformatik
Westfälische Wilhelms-Universität Münster
Leonardo-Campus 3
48149 Münster
vossen@helios.uni-muenster.de

Abstract: Das Projekt BankBook basiert auf der Idee der Etablierung einer virtuellen Bankfiliale bei Facebook. Bisher sind viele Banken auf Facebook nicht als App, sondern lediglich mit einer Facebook-Seite vertreten, sofern sie überhaupt eine Facebook-Präsenz haben. Häufig wird diese Seite dann genutzt, um mit Kunden zu kommunizieren oder um PR-Aktionen durchzuführen oder bekannt zu machen. Das bedeutet, dass sich die Kommunikation meist an alle Kunden richtet, die angegeben haben, dass ihnen die Seite der Bank gefällt; es wird nicht individuell auf Kunden eingegangen. Dabei könnten Banken (und Kunden) Facebook besser nutzen, um die Kommunikation mit den Kunden zu unterstützen. Die BankBook-App bietet dazu neue Funktionen, die dabei helfen, einen vertrauensvollen und persönlichen Kontakt zwischen Beratern und Kunden aufzubauen, der die langfristige Verbindung zwischen beiden stärkt. Allerdings unterscheidet sich eine solche Geschäftsbeziehung deutlich von einer typischen Facebook-Freundschaft. Aus diesem und anderen Gründen setzt BankBook neuartige Konzepte um (wie z.B. ein von Facebook gelöstes Beziehungsmanagement und eine separate Datenhaltung), die im Rahmen eines Feldtests erprobt wurden und die derzeit in einer Zweitauflage des Projekts weiterentwickelt werden.

„Privatheit in der E-Society – Zwischenbilanz eines Diskussionsprozesses“

Denk- und Impulspapier Version 1.0 der Autorengruppe
Christiane Eckardt, Danilo Kardel, Christof Leng,
Jan Marco Leimeister, Michael Mörike,
Welf Schröter, Hans-Dieter Zimmermann
aus dem Fachbeirat der
Integrata-Stiftung für humane Nutzung der IT

Tübingen, Juli 2014

Übersicht über die Autorenbeiträge

Welf Schröter
Vorbemerkung

Danilo Kardel
Schutz der Privatheit

Christof Leng
Privatheit und Sicherheit

Hans-Dieter Zimmermann
Smart Cities und Privatheit

Michael Mörike
Exkurs 1: Schutz der Privatheit bei ePartizipation

Christiane Eckardt
Exkurs 2: Privatheit in der eSociety – Innerer Dialog eines Nutzers

Welf Schröter
Privatheit und virtuelle Arbeit in Industrie 4.0

Jan Marco Leimeister
Exkurs 3: Arbeitgeber und Gewerkschaften im Vergleich –
wie gut sind diese auf Crowdsourcing vorbereitet?

Vorbemerkung

Mit dem vorliegenden Denk- und Impulspapier „Privatheit in der E-Society – Zwischenbilanz eines Diskussionsprozesses“ will die Autorengruppe Christiane Eckardt, Danilo Kardel, Christof Leng, Jan Marco Leimeister, Michael Mörike, Welf Schröter, Hans-Dieter Zimmermann mit Unterstützung von Hans-Peter Fröschle und Karl Stroetmann aus dem Fachbeirat der „Integrata-Stiftung für humane Nutzung der IT“ neue Diskursanstöße setzen und in bestehende öffentliche Diskurse eingreifen. Das Papier begleitet den Integrata-Kongress 2014 unter dem Motto „Privatheit in der E-Society“.

Im Vordergrund steht dabei die Frage nach dem Menschen in den neuen virtuellen Lebensumgebungen. Welche Chancen eröffnen sich ihm? Welche Rechte und Schutzrechte sind erforderlich? Wie verändern sich Leben und Arbeiten? Was heißt Beteiligung in der „eSociety“? Die aufgeführten Gedanken eröffnen erweiternde Perspektiven. Die Textbausteine haben nicht den Anspruch alles und dieses umfassend aufzugreifen und zu bewerten. Die Autorengruppe startet mit einzelnen Punkten, will im Sinne eines „living document“ das Impulspaper in kommender Zeit fortschreiben, auf dass es alsbald nach anregender Debatte die Versionen 3.0 oder 7.0 oder 11.0 geben möge.

Die Idee der humanen Nutzung der Informationstechnik folgt ethischen und sozialen Überlegungen und möchte mehr Lebensqualität ermöglichen. Die Autorengruppe geht bei ihren Überlegungen von jenen zehn Thesen aus, die der Stifter der Integrata-Stiftung, Wolfgang Heilmann, vor einiger Zeit formulierte „Mehr Lebensqualität durch IT – 10 Thesen zur Verbesserung der Lebensbedingungen durch humane Nutzung der Informations- und Kommunikationstechnologie“:

> 1. Gesundheit und Wohlleben: Bewahrung und Wiederherstellung der körperlichen und geistigen Gesundheit.
>
> 2. Sicherheit und Freiheit: Erhaltung der inneren und äußeren Sicherheit unter Wahrung von Freiheit und Würde des Menschen.
>
> 3. Freizügigkeit und Verkehr: Herstellung und Sicherung der Freizügigkeit und menschenwürdiger Verkehrsverhältnisse zur persönlichen Begegnung.
>
> 4. Information und Kommunikation: Wiederherstellung des Vertrauens zwischen Kommunikationspartnern durch verantwortungsbewussten Umgang mit Informationen und freie Kommunikation.
>
> 5. Bildung und Ausbildung: Öffnung des Zugangs zu Bildung und Ausbildung nach Maßgabe des individuellen Leistungsvermögens, zugleich aber Sicherstellung der weltanschaulichen Neutralität von Bildungsvermittlung und Unterhaltung.

6. Arbeit und Erwerb: Schaffung von Arbeits- und Erwerbsmöglichkeiten, die auf die Menschen zugeschnitten sind, in ausreichender Zahl und Qualität, um das Gemeinwohl zu fördern und Wohlstand für alle zu erzielen.

7. Führung und Verantwortung: Weiterentwicklung der Informationstechnologie zu einem hilfreichen Instrumentarium für die Führung von Menschen durch Menschen in Wirtschaft und Gesellschaft.

8. Politik und Gesellschaft: Förderung der Partizipation der Bürger an der öffentlichen Meinungsbildung und Gestaltung des Gemeinwesens im Gleichgewicht von Freiheit, Recht und Gerechtigkeit.

9. Natur und Umwelt: Schutz der Natur und der Umwelt vor Raubbau und Zerstörung zugunsten der natürlichen Prozesse und der kommenden Generationen.

10. Kultur und Religion: Überwindung des Sinn- und Zeitnotstands moderner Menschen zugunsten einer würdigen Lebensführung mit Muße für Kultur und Religion.

Wir laden alle Interessierte ein, das Denk- und Impulspapier zu diskutieren, der Kritik zu unterziehen oder es zu erweitern.

Für die Autorengruppe
Welf Schröter
schroeter@talheimer.de

Schutz der Privatheit

Danilo Kardel

IT-Beratung und Management
dk@kardel-it.de

Auch wenn die Frage, was die Gesamtheit eines Menschen über seine Physis hinaus überhaupt ausmacht, gar nicht geklärt ist, kann seine Identität versuchsweise definiert werden als die Summe aller über ihn existierenden Informationen[1]. Das Wesen der Information ist, „die implizite (latente, an sich vorhandene) Relation zwischen dem Einzelnen und dem Allgemeinen explizit zu machen" (nach G. Werckmeisters Interpretation der Hegelschen Dialektik[2]).

Die informationelle Selbstbestimmung eines Individuums besteht darin, Produzent der Informationen über sich selbst zu sein, und verfügen zu können, wer Kenntnis von diesen Informationen erhält. Entsprechend beinhaltet die informationelle Fremdbestimmung sowohl die Produktion von Informationen über ein Individuum durch Andere („Identifizierung") als auch die vom Individuum selbst nicht kontrollierbare Verwendung und Zuweisung von Informationen. Informationelle Fremdbestimmung erfolgt z.B. innerhalb der Gesellschaft (Einordnung in gesellschaftliche Gruppen oder Zuweisung eines sozialen Status), in der Arbeitswelt, auf staatlicher Ebene, oder automatisiert durch elektronische Datenverarbeitung. Dabei werden Identifizierungsmerkmale und Identifizierungsmethoden willkürlich durch Dritte festgelegt und mit personenbezogenen „gespeicherten" Daten korreliert. Historische Beispiele für gesellschaftliche Gruppen, denen Menschen (auch) durch informationelle Fremdbestimmung zugeordnet wurden, sind „Zigeuner" oder „Juden". Aktuell findet informationelle Fremdbestimmung in besonderem Maße statt bei der Identifizierung von „Straftätern" oder „Terroristen", teilweise mit hohem Automatisierungsgrad. So kamen in den vergangenen Jahren hunderte von signature strikes der U.S. Army in ihrem war on terror zustande[3]. Die genannten Beispiele deuten bereits an, wie informationelle Fremdbestimmung für Individuen bedrohlich werden kann.

Informationelle Selbstbestimmung, d.h. die Befähigung des Individuums zur Kontrolle über Entstehung und Verbreitung von Informationen über sich selbst, ist das einzige Mittel gegen ein Übermaß an informationeller Fremdbestimmung. Wenn ein Individuum von seinem Recht auf informationelle Selbstbestimmung Gebrauch macht, entsteht Privatheit.

[1] W. Schmale, M. Tinnefeld, Privatheit im digitalen Zeitalter, Wien 2014.
[2] G. Werckmeister, DVR Band 7, 1978. http://www.hegel-system.de/de/v131-werckmeister.htm; Zugriff 20.05.2014.
[3] Nach: http://drones.procon.org/; Zugriff 20.05.2014.

Privatheit hat es immer gegeben, auch wenn unser heutiger Privatheitsbegriff sich stark von den individuellen Optionen zur Schaffung von Privatheit in der Vergangenheit unterscheidet, etwa im Mittelalter oder in der Antike. Die Idee des „Gartens als Ort der Privatheit“ (M. Tinnefeld), in dem Menschen sich unkontrolliert bewegen können, begegnet uns bereits in den ältesten überlieferten Erzählungen. Beim Garten Eden allerdings haben wir es mit einem frühen Fall von totaler Überwachung zu tun: Bei Übertretung des (nur aus einem einzigen Verbot bestehenden) Gesetzes erfolgt umgehend die Abschiebung aus dem Paradies.

Privatheit ist von zentraler Bedeutung als Fundament einer anti-totalitären demokratischen Zivilgesellschaft. Ohne Privatheit und Privatsphäre gibt es keine Demokratie, ebenso wenig wie ohne Informationsfreiheit, Meinungs-, Presse- und Medienfreiheit. Privatheitsschutz und Datenschutz sind als besondere rechtskulturelle Leistungen anzuerkennen.

Die Vereinten Nationen haben im Jahre 1948 in der Allgemeinen Erklärung der Menschenrechte das Recht auf Privatheit (right to privacy) als eines der universellen Menschenrechte erkannt. Das Bundesverfassungsgericht hat 1983 erstmals ein Grundrecht auf informationelle Selbstbestimmung anerkannt, indem es dem Einzelnen das Recht gewährt, „grundsätzlich selbst über die Preisgabe und Verwendung seiner persönlichen Daten zu bestimmen.“[4] Die EU hat mit der Konvention zum Schutz der Menschenrechte und Grundfreiheiten aus dem Jahre 1950 (auch unter der Bezeichnung Europäische Menschenrechtskonvention, EMRK) und der EU-Grundrechte-Charta (EGRC, auch Charta der Grundrechte der Europäischen Union) von 2000 fortschrittliche Standards zum Datenschutz und zum Schutz der Privatsphäre geschaffen. Artikel 8 der EMKR lautet[5]:

> (1) Jede Person hat das Recht auf Achtung ihres Privat- und Familienlebens, ihrer Wohnung und ihrer Korrespondenz.
>
> (2) Eine Behörde darf in die Ausübung dieses Rechts nur eingreifen, soweit der Eingriff gesetzlich vorgesehen und in einer demokratischen Gesellschaft notwendig ist für die nationale oder öffentliche Sicherheit, für das wirtschaftliche Wohl des Landes, zur Aufrechterhaltung der Ordnung, zur Verhütung von Straftaten, zum Schutz der Gesundheit oder der Moral oder zum Schutz der Rechte und Freiheiten anderer.

Die EU-Grundrechte-Charta enthält je einen Artikel zum Schutz der Privatsphäre und zum Recht auf informationelle Selbstbestimmung. Artikel 7 der EGRC fordert die Achtung der Privatsphäre entsprechend EMRK, Art. 8 (s.o.); Artikel 8 der EGRC erklärt den Schutzbedarf personenbezogener Daten[6]:

[4] BVerfGE 65, 1 – Volkszählung; Urteil des Bundesverfassungsgerichts vom 15. Dezember 1983; zitiert nach: http://www.servat.unibe.ch/dfr/bv065001.html#; Zugriff 25.05.2014.
[5] Zitiert nach: http://conventions.coe.int/Treaty/ger/Treaties/Html/005.htm; Zugriff 20.05.2014.
[6] Zitiert nach: http://www.europarl.europa.eu/charter/pdf/text_de.pdf; Zugriff 20.05.2014.

(1) Jede Person hat das Recht auf Schutz der sie betreffenden personenbezogenen Daten.

(2) Diese Daten dürfen nur nach Treu und Glauben für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage verarbeitet werden. Jede Person hat das Recht, Auskunft über die sie betreffenden erhobenen Daten zu erhalten und die Berichtigung der Daten zu erwirken.

(3) Die Einhaltung dieser Vorschriften wird von einer unabhängigen Stelle überwacht.

Während also das Recht auf Privatsphäre und informationelle Selbstbestimmung mittlerweile einen starken rechtlichen Schutz genießt, steht heute mit dem automatisierten Erstellen von Profilen und „Signaturen" durch Auswertung digitaler Spuren (vor allem im Internet) ein globales und zugleich umfassendes Werkzeug für informationelle Fremdbestimmung zur Verfügung. Sowohl staatliche als auch privatwirtschaftliche Einrichtungen nutzen die damit entstandenen Möglichkeiten extensiv, erstere eher unter sicherheitspolitischen Aspekten, letztere zum Zwecke individualisierter Manipulation aus ökonomischen Interessen. Die Untersuchungen von Privacy International mit einem standardisierten „Privatsphären-Index" zeigen bereits in den Jahren 2006 / 2007 bestürzende Ergebnisse über Ausmaß und Ausweitung weltweiter Überwachung[7] .

Um dem Anspruch gerecht zu werden, Privatheit und Privatsphäre als konstituierende Elemente einer anti-totalitären demokratischen Zivilgesellschaft zu schützen, ist es heute dringend erforderlich, Individualrechte zu stärken und bessere Kontrollfunktionen zu implementieren, die es dem Einzelnen ermöglichen, seine informationelle Selbstbestimmung zu behaupten bzw. wiederzuerlangen.

Bedingt durch den strukturellen Wandel der Arbeits-, Erwerbs- und Geschäftswelt und durch die Konvergenz bislang getrennter virtueller Handlungsräume erhalten virtuelle Identitäten zunehmende Bedeutung. Ein wesentlicher Baustein zur Erlangung informationeller Selbstbestimmung in der E-Society ist daher die Befähigung des Individuums zum Management virtueller Identitäten. Privatheit wird in der digitalen Welt gewährleistet durch die Autonomie und das Recht des Individuums, seine Identität innerhalb (und außerhalb) elektronischer Netze selbst zu bestimmen.

Der Schutz von Privatheit und Privatsphäre bedarf also rechtsverbindlicher Optionen auf Anonymisierung und Pseudonymisierung, ein Recht auf virtuelle Identitäten. Ein nutzergesteuertes Management virtueller Identitäten erlangt zentrale Bedeutung beim Schutz der Privatheit in einer demokratischen Informations- und Wissensgesellschaft.

Weitere wichtige Aspekte ergeben sich mit der Inkraftsetzung bzw. Wiedererlangung informationeller Selbstbestimmung gegenüber dem digitalen „Gedächtnis" des Internet. Dazu gehört das Recht auf Vergessen (werden). Das diesbezügliche Urteil des Europäischen Gerichtshofes vom 13. Mai 2014, demzufolge der Suchmaschinenbetreiber

[7] Privacy International, http://en.wikipedia.org/wiki/Privacy_International; Zugriff 20.05.2014.

Google Links zu personenbezogenen Daten unter bestimmten Bedingungen entfernen muss, weist in die richtige Richtung[8]. Die Rechte des Einzelnen gilt es, endlich mit Ernsthaftigkeit einzufordern und durchzusetzen. Das gleiche trifft zu für das in der EU-Grundrechte-Charta bestätigte Recht jeder Person, „Auskunft über die sie betreffenden erhobenen Daten zu erhalten" oder für den Grundsatz, die Verarbeitung und Weitergabe von Daten sei nur „für festgelegte Zwecke und mit Einwilligung der betroffenen Person oder auf einer sonstigen gesetzlich geregelten legitimen Grundlage" erlaubt. Schon die oberflächlichste Gegenüberstellung von EU-Recht und „unbewusst gelebter" Internet-Praxis enthüllt überdeutlich die Diskrepanz zwischen Anspruch und Wirklichkeit.

G. Werckmeister sieht durch die Entwicklung der Informationstechnologie den „Weg geöffnet für die allseitige (universelle) Entfaltung der Individuen (die Allgemeinheit des Einzelnen), für die freie Entfaltung der Persönlichkeit." Diese Perspektive zu wahren und nicht durch Kontrolle, Überwachung und Fremdbestimmung mithilfe ebendieser Technologien zerstören zu lassen, nimmt eine humane Nutzung moderner Informations-technologie sich zum Ziel. Um den Prozess eines gemeinwohlorientierten sozialen Mit-einanders zu gestalten, bedarf es aktiver Maßnahmen in Bildung und Ausbildung, Kunst und Kultur.

[8] EuGH, Rechtssache C-131/12: http://curia.europa.eu/juris/liste.jsf?language=de&num=C-131/12; Zugriff 20.05.2014.

Privatheit und Sicherheit

Christof Leng

Technische Universität Darmstadt
International Computer Science Institute Berkeley
cleng@icsi.berkeley.edu

Individuelle Privatsphäre und öffentliche Sicherheitsinteresse stehen scheinbar im ständigen Konflikt und müssen gegeneinander abgewogen werden. In der Tat benötigen Strafverfolgungsbehörden traditionell spezielle Privilegien zur Aufklärung von Straftaten, z.B. bei Wohnungsdurchsuchungen. Nicht immer kann jedoch durch weniger Privatheit mehr Sicherheit erreicht werden.

Durch den digitalen Wandel ergibt sich eine Vielzahl neuer Möglichkeiten für die Sicherheitsbehörden. Es sind nicht nur wesentlich mehr Informationen verfügbar, sondern sie sind auch in digitaler und vernetzter Form wesentlich leichter zu erlangen und automatisiert auswertbar. Ein fataler Teufelskreis führt dabei zur kontinuierlichen Aushöhlung der Bürgerrechte: Jeder Erfolg der Sicherheitsbehörden rechtfertigt die bestehenden Befugnisse; jeder Fehlschlag weckt und rechtfertigt hingegen weitergehende Befugnisse. Eine nüchterne Abwägung der Grundrechte gegeneinander ist unter diesen Umständen nicht möglich, und die einseitige Entwicklung bringt langfristig die freiheitlichen Grundfesten unserer Demokratie in Gefahr.

Bereits ein Konzept wie die Vorratsdatenspeicherung, welches von Strafverfolgern oft als unabdingbar dargestellt wird, widerspricht nach einheitlicher Rechtsprechung der Gerichte bereits in massiver Weise unserer demokratischen Grundordnung. Es zeigt sich jedoch, dass die Geheimdienste bereits seit Jahren eine sehr viel weitreichendere Totalüberwachung der Kommunikationsnetze praktizieren, weit jenseits der gesetzlichen Grundlagen und der demokratischen Kontrolle.

Diese Entwicklung entzieht sich trotz aller Enthüllungen weitgehend der öffentlichen Diskussion. Die Eingriffe in die Privatsphäre bleiben für den Einzelnen unbemerkt und zunächst ohne Folgen. Diese Unsichtbarkeit unterscheidet sich fundamental von anderen Formen staatlicher Gewalt und minimiert das Empörungs- und Betroffenheitspotential. Zum einen fühlt man sich nicht unmittelbar als Opfer, und zum anderen bleiben selbst drastische Fälle typischerweise ohne Gesicht. Wenn aber der ehemalige Ex-NSA-Chef Michael Hayden erklärt, „Wir töten auf Basis von Metadaten“, ist die Tragweite der Entwicklung kaum von der Hand zu weisen.

Zahlreiche Einzelfälle zeigen, dass die auf geheimen Überwachungsmaßnahmen basierenden Entscheidungen oft fehlerhaft aber gleichzeitig schwer anfechtbar sind. Das zentrale Problem ist jedoch, dass eine Infrastruktur geschaffen wird, deren Missbrauch in Krisenzeiten unmöglich zu verhindern ist und dann wesentlich zu einer Eskalation bei-

tragen könnte. Diese vermeintlichen Sicherheitswerkzeuge stellen somit selbst eine Gefahr für die individuelle und kollektive Sicherheit dar.

Folglich ist der Widerspruch zwischen Privatsphäre und Sicherheitsinteressen gar nicht so groß. Für die Sicherheit unserer freiheitlichen und demokratischen Gesellschaft ist die Privatheit kein Hindernis sondern eine Voraussetzung. Jeglicher Eingriff in die Privatsphäre muss daher offengelegt und äußerst sorgfältig abgewogen werden.

Daraus ergeben sich folgende Forderungen:

1. Grundrecht auf unbeobachtete Kommunikation

Die Meinungsfreiheit ist eine zentrale Stütze unserer Demokratie. Diese kann bereits vom Gefühl, überwacht zu werden, beeinträchtigt werden. Da ein stetig wachsender Anteil des innersten Lebensbereichs in den digitalen Netzen stattfindet, muss die Vertraulichkeit der ausgetauschten und gespeicherten Daten höchsten gesetzlichen Schutz genießen.

2. Keine verdachtsunabhängige Überwachung und Ausspähung

Die Unschuldsvermutung muss auch in der E-Society aufrechterhalten werden. Ohne einen konkreten und schwerwiegenden Anfangsverdacht darf das Grundrecht auf unbeobachtete Kommunikation nicht angetastet werden. Dies schließt insbesondere die flächendeckende Vorratsdatenspeicherung und Kommunikationsüberwachung, auch als Suche nach „verdächtigen“ Mustern, aus.

3. Dokumentations- und Offenlegungspflichten

Sollte aufgrund eines konkreten und schwerwiegenden Verdachts nach der Ausschöpfung anderer Ermittlungsmethoden eine richterlich angeordnete Überwachung oder Ausspähung notwendig geworden sein, muss diese nachvollziehbar dokumentiert und den Betroffenen zeitnah, z.B. nach Abschluss der Ermittlungen, angezeigt und offengelegt werden. Nur so kann die demokratiefeindliche Angst vor Überwachung minimiert werden.

4. Transparenz und Kontrolle der Behörden

Die Einhaltung der gesetzlichen Rahmenbedingungen durch die Sicherheitsbehörden ist derzeit faktisch nicht kontrollierbar. Eine personell angemessen und mit vollständigem Informationszugriff ausgestattete unabhängige Kontrollinstanz muss geschaffen werden, welche die demokratische Kontrolle der Sicherheitsbehörden wiederherstellt. Zudem müssen die vorhandenen Werkzeuge und Vorgehensweise der Sicherheitsbehörden der Öffentlichkeit transparent gemacht werden, um das Vertrauen wiederherzustellen und eine gesellschaftliche Kontrolle zu ermöglichen. Die IT-Security lehrt uns, dass transparente Methoden in der Praxis zuverlässiger sind als „security by obscurity“.

5. Unabhängige Überprüfung

Auch die Effektivität der Ermittlungswerkzeuge, welche durch die Einschränkung des Rechts auf Privatsphäre ermöglicht werden, muss regelmäßig von einer unabhängigen Stelle ermittelt und mit den möglichen Alternativen verglichen werden. Nur die Ergebnisse einer regelmäßigen objektiven Überprüfung können die Debatte wieder versachlichen.

6. Institutionelle Durchsetzung des Schutzes der Privatsphäre

Nicht zuletzt sollten die Sicherheitsbehörden den Schutz der Privatsphäre als wesentlichen Teil ihrer Aufgaben und des staatlichen Sicherheits- und Stabilitätsinteresses begreifen. Ein wesentliches Ziel muss daher die Sicherung der Privatsphäre gegenüber Kriminellen und sich außerhalb der eigenen Gerichtsbarkeit bewegenden Stellen fremder Staaten darstellen.

Smart Cities und Privatheit

Hans-Dieter Zimmermann[9]

Institut für Informations- und Prozessmanagement IPM
FHS St. Gallen Hochschule für Angewandte Wissenschaften
St. Gallen, Schweiz
hansdieter.zimmermann@fhsg.ch

Die Idee der Smart City

Weltweit stehen Städte und Gemeinden – wie auch ganze Regionen – vor immer komplexeren Herausforderungen, die es zu bewältigen gilt. Ein Ansatzpunkt sind Ideen und Ansätze der *Smart City*. Solche Herausforderungen können beispielsweise die Bewältigung der Verkehrssituation in einer Stadt oder auch Region, die Reduktion von Umweltemissionen oder der Energieverbrauch sein, oder auch die Frage wie Bürger stärker in die Gestaltung von Lösungen in einer Smart City einbezogen werden und diese aktiv mitgestalten können. Häufig werden die folgenden sechs Charakteristika oder Anwendungs- bzw. Handlungsfelder dem Konzept Smart City zugeschrieben: Smart Economy, Smart People, Smart Governance, Smart Mobility, Smart Environment, Smart Living (vgl. z.B. [2]). Der Begriff als solcher wird auch in der Wissenschaft intensiv diskutiert, vgl. z.B. [4] [10] [3]. Vor allem der Begriff *Smart* wird je nach Perspektive unterschiedlich interpretiert; in [10] werden die Perspektiven Marketing, Stadtplanung und IKT exemplarisch genannt.

Die EU definiert in einem aktuellen Bericht den Begriff wie folgt: „A Smart City is a city seeking to address public issues via ICT-based solutions on the basis of a multi-stakeholder, municipally based partnership." ([9] S. 9). Dieses Begriffsverständnis macht deutlich, dass vor allem die Informations- und Kommunikationstechnik IKT eine zentrale Rolle spielt.

Nicht zuletzt beschäftigen sich auch IT-Firmen wie z.B. IBM mit dem Thema[10]; hier dürften ökonomische Beweggründe dominierend sein.

Gemäss Nam und Pardo [10] lassen sich drei grundsätzliche Dimensionen des Konzepts Smart City unterscheiden: *Technologie*, *Mensch* und *Gemeinschaft* bzw. *institutioneller Rahmen*. Während der IKT-Fokus vieler Begriffsdefinitionen vor allem die Technologie-Dimension repräsentieren, fokussiert die Dimension Mensch auf das „Menschliche", die Kreativität des Menschen und dessen Sozialkompetenz. Die dritte Dimension adressiert die institutionellen und rechtlichen Rahmenbedingungen. Anders formuliert, die drei

[9] Mitglied des Fachbeirats der Integrata-Stiftung, Tübingen
[10] Vgl. „Welcome to TheSmarterCity"; http://www-03.ibm.com/innovation/us/thesmartercity/

Dimensionen betreffen die technischen Infrastrukturen, die dem Zusammenleben der Menschen dienen, deren Gemeinwesen entsprechend institutionell organisiert ist. Letzterer Aspekt geht einher mit der Entwicklung zu einem neuen Verständnis von Verwaltungen und Regierungen, dem Open Government (vgl. z.B. [7]). In diesem Zusammenhang übernehmen z.B. die Mitglieder eines Gemeinwesens neue Rollen, was u.a auch im Verständnis offener gesellschaftlicher Innovationen (OGI) zum Ausdruck kommt (vgl. [8]).

Bei dieser Betrachtung erhält die IKT einen anderen, eher dienenden Stellenwert als in vielen Diskussionen rund um Smart Cities, bei denen die IKT zentral im Mittelpunkt steht.

Diese Sichtweise entspricht auch weitgehend der Rolle der Informationstechnik aus Sicht der Integrata-Stiftung. Im Kontext der „10 Thesen zur Verbesserung der Lebensbedingungen durch humane Nutzung der Informations- und Kommunikationstechnologie“[11] der Integrata-Stiftung werden mit dem Begriffsverständnis der Smart City mindestens drei Thesen unmittelbar adressiert: Gesundheit und Wohlleben, Freizügigkeit und Verkehr, Natur und Umwelt; weitere Thesen zumindest mittelbar.

Den verschiedenen Interpretationen und Sichtweisen der Idee der Smart City ist gemein, dass wir die unübersehbaren Herausforderungen unserer Zeit durch integrierte und ganzheitliche Ansätze angehen müssen. Der IKT kommt dabei durchaus eine tragende Rolle zu, jedoch gibt es unterschiedliche Auffassungen darüber, inwieweit die IKT der *Treiber* der Entwicklungen ist oder als *Enabler* für kluge Lösungen angesehen wird.

Zusammenfassend kann festgehalten werden, dass die Idee der Smart City im Grundsatz die Wertvorstellungen einer Gesellschaft betrifft, d.h. wie z.B. eine Gesellschaft mit dem Verbrauch natürlicher Ressourcen, mit Umweltbelastungen oder mit ihren Mobilitätsbedürfnissen umgeht, ohne das erreichte Wohlstandsniveau reduzieren zu müssen.

Privatheit

Privatheit – oder englisch Privacy – ist ebenso ein Begriff, der kaum wirklich fassbar ist. In den aktuellen Diskussionen wird er inflationär benutzt, ohne das oft seine Bedeutung hinterfragt wird. Festzuhalten ist, dass Privatheit keinen absoluten Wert darstellt.

Das Autorenkollektiv von [1] beschäftigt sich intensiv mit der Begriffswelt rund um Privatheit und dem (scheinbaren) Gegenüber, der Öffentlichkeit. Ohne den Begriff näher eingrenzen zu wollen, sollen hier vor allem die Aspekte Datenschutz, also der Schutz persönlicher bzw. personen-bezogener Daten, sowie die informationelle Selbstbestimmung des Einzelnen als wichtige Elemente hervorgehoben werden.

Somit handelt es sich auch beim Phänomen der Privatheit um Wertvorstellungen einer Gesellschaft.

[11] Vgl. https://www.integrata-stiftung.de/startseite/stiftungsidee/mehr-lebensqualitat-durch-it/

Smart Cities und Privatheit

Betrachtet man die Begriffe Smart City und Privatheit zusammen, ergeben sich durchaus Widersprüche in Bezug auf die gesellschaftlichen Wertvorstellungen – dies kann zumindest als These formuliert werden.

Beispiel Energieversorgungsinfrastruktur: Neben den sog. Smart Grids geht es hier aus Sicht des Konsumenten vor allem um das Smart Metering (vgl. z.B. [1] S. 149ff.). Ziel ist die möglichst effiziente Energieversorgung. Die Basis hierfür sind neuartige, elektronische Stromzähler, die über eine zusätzliche Kommunikationseinheit z.B. mit dem Energieversorger oder Betreiber der Messstelle bidirektional kommunizieren können. Der Stromverbrauch kann dabei potentiell sekundengenau gemessen und weitergegeben werden. Die Messdaten werden zwar aggregiert und somit anonymisiert weiterverarbeitet, aber letztendlich liegen je Haushalt individuelle Energieverbrauchsdaten vor, die in einem Informationssystem gespeichert werden. Dies stellt gewisse Herausforderungen dar: „*Besonders deutlich wird diese Herausforderung im Bereich des Datenschutzes bzw. des Rechts auf informationelle Selbstbestimmung. Grundsätzlich stellen nämlich Energiemessdaten personenbezogene Daten dar, aus denen sich – je nach Auflösung – zumindest potentiell Aussagen zum Verhalten und zur Haushaltsausstattung eines einzelnen Elektrizitätsnutzers ableiten lassen.*“ ([1] S. 152). Ein denkbares Szenario ist auch, dass Haushalten entsprechend ihres Verbrauchs differenziert bepreist werden, um z.B. Steuerungsanreize für eine ausgeglichene Auslastung der Energieversorgungsinfrastruktur zu geben und Spitzenverbrauche zu reduzieren. Wer als Haushalt den Wert der informationellen Selbstbestimmung höher bewertet und seine individuellen Verbrauchsdaten nicht herausgeben will, wird dann möglicherweise durch eine Preisdiskriminierung bestraft. Hier wird ein Wertekonflikt deutlich. Auf diesen Konflikt mach z.B. Datoo in seinem Beitrag in The Guardian aufmerksam [5]. Im Zusammenhang mit der Übernahme des Thermostat- und Feuermelderhersteller Nest durch Google gab es ebenfalls entsprechende Diskussionen. Der Beitrag von Steiner in der NZZ vom Juni 2014 zeigt z.B. Szenarien auf, wie intelligente Thermostate im Kontext des *Internet of Things* mit beliebigen sonstigen elektronischen Geräten vom Schlaftracker bis hin zu unserem Auto vernetzt werden können [6] – mit allen Konsequenzen bzgl. Privatheit. Auch wenn Google gegenwärtig gerne als Feindbild betrachtet wird, so sind es letztendlich beliebige Firmen und Organisationen wie Energieversorger oder Stadtwerke, die – im Zusammenspiel – solche Dienste zukünftig anbieten könnten.

Ähnlich verhält es sich bei der Steuerung des privaten Individualverkehrs wie auch des öffentlichen Verkehrs. So gibt es sowohl in Deutschland als auch in der Schweiz Versicherungen, welche mittels einer Blackbox das Fahrverhalten aufzeichnen und davon abhängig die Versicherungsprämie berechnen (vgl. z.B. [11]). Dazu werden entsprechend personenbezogene Daten aufgezeichnet. Durch den ökonomischen Anreiz könnte der Fahrer motiviert werden, umsichtiger zu fahren und somit weniger Unfälle zu verursachen und weniger Emissionen zu generieren. Der Preis dafür sind die erhobenen Daten. Das viel diskutierte sog. *Raodpricing* ist ein weiteres Beispiel in diesem Kontext.

Im öffentlichen Verkehr werden Systeme erprobt z.B. durch die Schweizerische SBB – die es erlauben, dass ein Fahrgast, der mit einer entsprechenden Chipkarte ausgerüstet

ist, automatisch beim Ein- und Aussteigen erkannt wird und ihm z.B. am Monatsende der für ihn optimale Fahrpreis berechnet wird. Dieses System hat Vorteile für beide Seiten, es ist durchaus vorstellbar, dass z.B. Verkehrsflüsse optimiert werden, etc. Aber auch hier *bezahlen* wir den Mehrwert mit Bewegungsdaten.

Ein weiteres Beispiel, das dem Bereich Smart Living zugeordnet werden kann, sind Sensor-basierte Assistenzsysteme (AAL Systeme), die es älteren, z.B. auch pflegebedürftigen oder behinderten Menschen erlauben, selbständig in der eigenen Wohnung zu wohnen. Dies bedeutet sicher ein Mehr an Lebensqualität, es wird aber *erkauft* mit der Erfassung und Speicherung von Bewegungs- und Verhaltensdaten via Sensorsystemen.

Fazit

Die skizzierten Beispiele machen deutlich, dass es zwischen den Zielen einer Smart City und den Grundsätzen der Privatheit zu Zielkonflikten kommen kann. Diese basieren auf sich widersprechenden Wertvorstellungen. Deswegen ist es zentral, dass wir als Gesellschaft über eben diese Wertvorstellungen und Wertsysteme offen und als Ganzes diskutieren und den Konsens suchen. Gerne werden heute Einzelbeispiele *dafür* oder *dagegen* benutzt, um die eigene Meinung zu untermauern, ohne den Kontext zu berücksichtigen. Ebenso gerne werden bestimmte Unternehmen als Feindbilder aufgebaut, um gewisse Massnahmen, z.B. gesetzlicher Art, durchzusetzen.

Ein Ziel im Sinne der informationellen Selbstbestimmung kann es sein, dass wir selbst darüber entscheiden können müssen, ob wir mit unseren Daten einen Mehrwert *erkaufen* wollen. Die *Pflicht* entsprechende Daten zu sammeln darf nur dann entstehen, wenn es darüber einen gesellschaftlichen Konsens gibt. Um diesen zu erreichen, muss auch der *Smart Citizen* seine Verantwortung als Teil des Gemeinwesens übernehmen bzw. delegieren. Keinesfalls dürfen die Rechte des Einzelnen unter dem Vorwand eines übergeordneten Nutzens umgangen werden. Dazu gibt es in verschiedenen Ländern Europas genügend aktuelle Beispiele.

Als *Smart Citizens* muss uns auch bewusst sein, dass es heute vor allem Staaten bzw. staatliche Institutionen wie Geheimdienste sind, die praktisch ohne jegliche Konsequenz Daten auch illegal beschaffen und verwerten können.

Literaturverzeichnis

1. Buermeyer, A.E.A.·.M.B.·.U. et al.: Gleichgewicht und Spannung zwischen digitaler Privatheit und Öffentlichkeit. (2011).

2. Centre of Regional Science: Smart cities – Ranking of European medium-sized cities. Vienna (2007).

3. Chourabi, H. et al.: Understanding smart cities: An integrative framework. Proceedings of the Annual Hawaii International Conference on System Sciences. pp. 2289-2297 (2011).

4. Cocchia, A.: Smart and Digital City: A Systematic Literature Review. In: Renata Paola Dameri, C.R.-S. (ed.) Smart City: How to Create Public and Economic Value with High Technology in Urban Space. pp. 13-43 Springer International Publishing (2014).

5. Datoo, S.: Smart cities: are you willing to trade privacy for efficiency?, http://www.theguardian.com/news/2014/apr/04/if-smart-cities-dont-think-about-privacy-citizens-will-refuse-to-accept-change-says-cisco-chief, (2014).

6. Henning Steier: Was nicht nur Google die Heimvernetzung erschwert, http://www.nzz.ch/mehr/digital/nest-api-mercedes-benz-jawbone-whirlpool-google-now-1.18329044, (2014).

7. Herzberg, J.: Staatsmodernisierung durch Open Innovation – Problemlage, Theoriebildung, Handlungsempfehlungen. Deutsche Telekom Institute for Connected Cities: TICC-Schriftenreihe ; Bd. 4 (2012).

8. Lucke, J. von et al.: Offene gesellschaftliche Innovation – Die Seealemannische Definition, http://esocietybodensee2020.wordpress.com/publikationen/offene-gesellschaftliche-innovation-die-seealemannische-definition/, (2012).

9. MANVILLE, C. et al.: Mapping Smart Cities in the EU. (2014).

10. Nam, T., Pardo, T.A.: Conceptualizing smart city with dimensions of technology, people, and institutions. Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in Challenging Times - dg.o '11. pp. 282-291 ACM Press, New York, New York, USA (2011).

11. Siedenbiedel, C.: Revolution der Kfz-Versicherung, http://www.faz.net/aktuell/finanzen/meine-finanzen/versichern-und-schuetzen/nachrichten/ueberwachtes-fahrverhalten-revolution-der-kfz-versicherung-12747505.html, (2014).

Exkurs 1:
Schutz der Privatheit bei ePartizipation

Michael Mörike

Mitglied des Vorstandes der Integrata-Stiftung für humane Nutzung der IT
michael.moerike@integrata-stiftung.de

In der eSociety, in der es alle mit allen per Facebook teilen, in der die Virtualität mehr und mehr zur erlebten Realität wird, und in der auch die Welt der staatlichen Verwaltung sich mehr und mehr mit der Privatsphäre mischt, möchten wir dennoch weiterhin über uns selbst bestimmen. Wie das gehen kann, weiß heute keiner wirklich.

Im folgenden Beitrag beziehe ich Stellung und möchte einige mehr oder weniger naheliegende Möglichkeiten dazu aufzeigen. Sie sollen alle diejenigen anregen, die eParticipation gestalten.

Meine Thesen zur Verteidigung der Privatheit:

1. Hinter dem deutschen Wort Datenschutz steckt nicht genau derselbe Begriff wie hinter dem englischen privacy. Ich übersetze Privacy mit Privatheit – nicht mit Datenschutz.

2. Meine simple Definition von Privacy: Je mehr Information (über eine Person) der Öffentlichkeit verborgen bleiben, umso größer ist die Privatheit (dieser Person). Privatheit ist ein schützenswertes Gut, auf das jeder Mensch ein Grundrecht hat. Der Schutz besteht also darin, möglichst wenig an die Öffentlichkeit gelangen zu lassen, und insbesondere möglichst zu verhindern, dass Information an die Öffentlichkeit gelangt, von der die Person das nicht möchte.

3. Je nach Zusammenhang (Kontext) wird mehr oder weniger Privatheit benötigt.

   a. Solche Zusammenhänge werden heute oft als "Kreise" bezeichnet und sind - mathematisch gesehen – als Mengen von (miteinander mehr oder weniger) verbundenen Personen anzusehen.

   b. Wenn die Kreise mit den Informationen (über eine Person) sorgfältig umgehen, kann Privatheit auch in diesen Kreisen gewährleistet sein; d.h. es dringen keine vertraulichen Informationen an die breite Öffentlichkeit.

4. Der Schutz der Privatheit verlangt nach Sorgfalt und Zuverlässigkeit (der anderen). Privatheit benötigt Vertrauen (auf diese Zuverlässigkeit). Vertrauen entsteht durch langfristiges gemeinsames Handeln. Das Vertrauen besteht aus einem Versprechen (keine

Informationen weiterzugeben, Datenschutzerklärung) und dessen (saubere, unbedingte) Einhaltung.

5. Vertrauen ist gut, Kontrolle ist besser: Das gegebene Vertrauen (in eine Datenschutzerklärung) muss kontrollierbar sein. D.h. der Umgang mit den Daten muss kontrollierbar sein: D.h. jede Person, die Daten einem anderen überlässt, muss die Möglichkeit haben, den Umgang mit den Daten in jedem Einzelfall überprüfen zu können. Dazu können Logfiles dienen, die bei jeder Datenverarbeitung mitgeschrieben werden, und die der Manipulation durch den Datenverarbeiter entzogen sind. Das lässt sich mit (derzeit noch großem) technischen Aufwand lösen.

Daher meine Forderung (im Sinne der Integrata-Stiftung für humane Nutzung der IT): Jede (einzelne) Datenverarbeitung ist dem Informationseigentümer (potentiell) offenzulegen. Das erzeugt zwar sehr viele Metadaten, aber auch Vertrauen. Das ist es mir wert!

Zum Kontext der eParticipation:

Meine simple Definition der eParticipation: eParticipation ist Beteiligung am öffentlichen Leben, an der Politik mit Hilfe vom Internet, kurz Bürgerbeteiligung im Netz im Zeitalter der eSociety. Die Form kann sowohl deliberativ als auch kampagnengetrieben sein.

Wenn Bürger sich deliberativ beteiligen, z.B. indem sie Ideen oder Vorschläge einbringen, oder indem sie Zielvorstellungen und Kriterien für eine gute Sache äußern, geben sie immer auch etwas von sich selbst preis: Aus ihren Äußerungen kann man immer auch auf ihre Gedankenwelt schließen. Oft vertreten sie dabei ihre eigenen Interessen bis hin zu massiv vorgetragenen Partikularinteressen. Dabei sehen sie meist auch einen Nutzen darin, wenn sie das tun. Sie tauschen also einen Nutzen gegen ein Stück Privatsphäre und sehen überwiegend einen Vorteil darin.

Beteiligung per Netz gilt als eine mögliche Form, sich zu beteiligen. Beteiligung wird noch lange Zeit auch durch persönliche Gespräche mit anderen Bürgern, z.B. in Bürgerwerkstätten erfolgen. Derzeit gilt eine passende Mischung von Beteiligung per Internet und in Präsenzveranstaltungen als beste Form, auch Hybridbeteiligung genannt.

Schutzmechanismen der Privatsphäre bei eParticipation

1. Der Dienstanbieter von Bürgerbeteiligung per Internet sollte vertrauenswürdigen Datenschutz bieten, vor allem, wenn man sich per Klarnamen beteiligen möchte. Aber reicht das aus?

2. Man kann anonym teilnehmen, also ohne Angabe einer Identität. Das hat aus Sicht des Dienstanbieters den Nachteil, dass man dasselbe Argument mehrfach einbringen kann

und so die Gesamtdarstellung manipulieren kann. Ein Meinungsbild lässt sich so jedenfalls nicht seriös erstellen.

3. Man kann statt mit Klarnamen mit einem Pseudonym teilnehmen, also mit einer virtuellen Identität. Dies schwächt den Nachteil der Anonymität aber nur ab und hebt ihn nicht vollständig auf, denn es ist sehr aufwändig, sich mehrere virtuelle Identitäten zuzulegen, aber eben nicht unmöglich.

Anonym oder mit Klarnamen?

In vielen Fällen von Bürgerbeteiligung im Netz kann man sich nicht nur mit Klarnamen, sondern auch anonym (oder besser: mit einem Pseudonym) beteiligen. Ist das die Lösung?

Anonyme Beteiligung hat derzeit nicht das gleiche zugebilligte Gewicht, wie eine Beteiligung mit offenem Visier, also mit Klarnamen. Sie ist vielleicht besser geschützt, hat aber auch weniger Wert. Spätestens wenn ein Bürger seine Wünsche oder Vorschläge, Ziele und Kriterien auch in Präsenzveranstaltungen vorbringt, verlässt er den Schutz der Anonymität.

Schließlich gilt es nach wie vor als lobenswert, sich in einer Demokratie zu seiner Haltung zu bekennen. Ist also gar keinen Schutz erwünscht?

Mit dem neuen elektronischen Personalausweis (ePerso) ist eine Anonymisierung möglich, ohne die Identität aufgeben zu müssen. Der ePerso gestattet, nur den Wohnort abzufragen, also die Berechtigung, mitreden oder sogar abstimmen zu dürfen, ohne den Namen abzufragen oder abzugleichen. Dabei wird ein Token übergeben, der für die weitere Verfolgung der Identität genutzt werden kann. Allerdings wird das heute noch viel zu wenig genutzt. Mir ist jedenfalls kein solches Verfahren bei Bürgerbeteiligung im Netz bekannt.

Interessensvertretungen

Man kann sich in einer Gruppe von Menschen „verstecken“, die ihre Interessen gemeinsam vertritt: Ein Verein, eine Interessensvertretung oder eine andere Gruppierung von Gleichgesinnten. Dann muss der Datenschutz nur innerhalb der Gruppe gewährleistet sein, was vielleicht etwas einfacher möglich ist. Die Gruppe wird dann als Ganzes nach außen von ihrem Sprecher vertreten. Eine solche Gruppe hat meistens eine eigene Internetseite und kann intern ein internes Forum bieten, auf dem diskutiert wird.

Möglich wäre es, auch solch eine Gruppe aus den oben erwähnten Kreisen aufzubauen, wenn sich alle Mitglieder einig sind. Ein solches Verfahren ist mir derzeit aber nicht bekannt.

Meine Forderungen zur eParticipation

Sie lauten daher:

1. Einerseits ist den Menschen möglichst umfangreiche Partizipation zu ermöglichen, andererseits ist ihre Privatheit zu schützen.

2. Zu einer intensiven Partizipation kann heute die IT (Internet) wesentlich beitragen. Dies gilt insbesondere für die sie voraussetzende Bildung. Gemäß der Stiftungsidee ist sie also auch dafür zu nutzen.

3. Die IT ist so einzusetzen, dass sie dem Individuum hilft, sich eine fundierte Meinung zu bilden. Dabei müssen Argumente auf Werte zurückgeführt werden können. Da die Werte des einzelnen Menschen zu seiner Privatsphäre gehören, verdient der Schutz dieser dabei ein besonderes Augenmerk.

4. Damit die Privatheit dabei dennoch vom Bürger und von Personen, die im öffentlichen Interesse stehen, selbst gesteuert werden kann, muss Partizipation sowohl mit Klarnamen als auch anonym angeboten werden.

5. Um dennoch die Identifizierbarkeit der Person zu gewährleisten, ist der ePerso entsprechend zu nutzen.

6. Eine Diskussion von Individualinteressen sollte bei ePartizipation vermieden werden. Wenn sie sich nicht vermeiden lässt, sollte sie von den Diskutanten anonym geführt werden – eventuell mit Ausnahme der Betroffenen selbst. Dazu können Kreise dienen, die als Interessensvertreter fungieren.

7. Um Emotionen dabei zu vermeiden, die zu viel Privatheit offenlegt, ist geeignete aktive Moderation erforderlich.

8. Alle einzeln durchgeführten Informationsverarbeitungsschritte sind dem Eigentümer der Daten potentiell offenzulegen. Er sollte sie z.B. per Formular anfordern können.

9. Das Verfahren dazu ist durch einen externen Dritten zu kontrollieren und zu zertifizieren.

Exkurs 2:
Privatheit in der eSociety – Innerer Dialog eines Nutzers

Christiane Eckardt

Mitglied des Fachbeirates der Integrata-Stiftung für humane Nutzung der IT
eckardt-christiane@t-online.de

Onlinebanking ist prima. ... das kann man nachts um 23h machen, shopping im Internet ist ebenso prima – keine Verkäuferin will einem was aufschwätzen, man kann stundenlang zögern, auswählen, wieder ändern und dann braucht man es nicht nach Hause schleppen, es wird ja geliefert.

Mit Facebook kann man Belanglosigkeiten mit einem großen Bekanntenkreis austauschen (dass es „Freunde" sind – wer glaubt das denn?), das reduziert den Aufwand für soziale Kontakte enorm. Allerdings – will ich wirklich alle „Belanglosigkeiten" allen mitteilen? Da muss ich grad bei den nächsten Informationen aus dem Urlaub etwas genauer aufpassen. Und: sind es wirklich soziale Kontakte – oder sind es nur belanglose, nichts bedeutende sprachliche Absonderungen?

Im beruflichen Umfeld ist es mittlerweile ein Muss, ein aussagekräftiges XING-Profil zu haben, auch über LinkedIn im internationalen Kontext muss man sich positionieren, schließlich will man vom Kunden oder Headhunter gefunden werden.

Allerdings: die Unterwäschewerbung, die letztens auf dem Bildschirm hochpoppte, als ich dem Nachbarn meiner Mutter Urlaubsbilder am PC gezeigt habe, war mir peinlich. Ob das davon kommt, dass ich bei denen mal was bestellt habe?

Und warum bekomme ich, seitdem ich wieder Single bin, vermehrt Reisen für Singles und sonstige Singleevents angeboten? Ok, ich habe das bei Facebook gepostet, weil es mir zu dumm war, dass ich jedem erklären musste, warum Markus jetzt nicht mehr mitkommt. Was geschieht denn bei den Providern mit den Inhalten? Werten die das inhaltlich aus und verkaufen es an Werbetreibende? Ich hätte nicht geglaubt, dass mich das auch treffen kann.

Meine letzte Hotelbuchung mit der Kreditkarte hat dazu geführt, dass ich plötzlich Abbuchungen von fünf Flügen in Afrika auf meiner Rechnung fand. Die Kreditkartenfirma hat zum Glück den Schaden übernommen, aber wer liest da denn eigentlich alles mit, wenn ich buche?

Die naive Vermutung der Anonymität (keiner sieht, was ich hier mache und wie oft ich mich beim shoppen umentscheide) kollidiert irritierend mit der Erfahrung, dass alle

Informationen, die ich meinen Kontaktpersonen mitteile, auch für unbekannte Dritte zur Verfügung stehen können.

Abgesehen von den unbekannten Dritten im Netz erfordert der Schutz der Privatheit im Netz auch im direkten realen Umfeld Sorgfalt. Meine Nichte schreibt alle Passworte, die sie bei diversen Portalen benutzt, in eine handschriftliche Kladde, weil auch sie sich das nicht alles merken kann. Dummerweise hat ihr kleiner Bruder das Heft gefunden und mit seinen Freunden zusammen mal alles aufgerufen. Dumm gelaufen – und irgendwie auch „voll peinlich“ wie sie sagt, wenn die Pimpfe jetzt immer anzüglich grinsen, wenn sie ihr begegnen.

Wenn es nur nicht so mühsam wäre! Einerseits, Passworte so zu konstruieren, dass keiner so leicht drauf kommt, nirgendwo aufschreiben (auch nicht handschriftlich), jederzeit rekonstruierbar und zugreifbar, auf der anderen Seite, immer wieder ein anderes Passwort, diese auch häufig wechseln. Ja, ich gebe es zu, davon fühle ich mich überfordert. Es kostet zu viel Nerven und Zeit und es ist einfach nicht handhabbar. Im Urlaub oder beruflich unterwegs muss ich ja auch die Liste dabei haben – zumindest einen Teil davon. Auf dem Handy? Im Laptop? Auf dem Tablett? Papierausdruck?

Der Verzicht auf Onlinebanking und Onlineshopping ist ebenfalls keine Lösung. Ohne Facebook könnte zumindest ich leben, aber ohne Email, diverse Internetzugriffe pro Tag, ohne Bordkarte auf dem Smartphone und Google Maps in der fremden Stadt auf dem Smartphone wird es schwierig. Klar, früher ging es auch ohne das alles, – sagt meine Freundin und verzichtet auf den Technikeinsatz, ist dann aber heilfroh, dass ich Google Maps zur Verfügung habe, wenn wir in einer fremden Stadt sind und auch mal mit meinem Smartphone und meiner Kreditkarte schnell im Café die Theaterbuchung mache.

Also, was nun? Komplett verzichten ist unpraktisch und man ist dann auch schnell aus den realen Beziehungsnetzen seiner Peergruppe ausgeschlossen. Soll ich einen sogenannten Passwort-Safe, bei dem ein Masterpasswort die anderen Passworte abschottet, benutzen? Das eine Passwort wird man sich ja merken können, ohne es aufzuschreiben – aber das ist ja auch elektronisch verwaltet – ob das dann wirklich sicher ist? Und funktioniert das zuverlässig und sicher über das Mobiltelefon unterwegs?

Ich müsste mich mindestens aufraffen, nicht überall das gleiche simple Passwort zu verwenden. Die Idee mit dem Codesatz könnte ich mir angewöhnen – dann muss ich mir aber die Codesätze merken. Schon wieder schlecht. War das noch „ich freue mich auf den Urlaub mit meiner Freundin am 23.Juli“ obwohl ich mich schon gar nicht mehr freue, weil der Urlaub gecancelt werden musste?

Nun hat sowohl die NSA als auch der Whatsapp- Deal mit Facebook auch etwas Gutes: jeder ist etwas mehr sensibilisiert, was Privatheit anbelangt und alternative Anbieter, die höhere Sicherheitsstandards wie Ende-zu-Ende-Verbindung und Verschlüsselung anbieten, erhalten Zulauf.

Leichtsinnig – oder dumm aus Unwissenheit oder Bequemlichkeit –, sind die Jugendlichen, die private, manchmal auch intime Fotos posten oder an vermeintliche Freunde senden, die dann die Fotos in sozialen Netzwerken posten oder diese unkontrolliert wei-

terversenden. Da beschleicht einen dann die Vermutung, dass das Prinzip des Internets nicht verstanden wurde. Dieses Unverständnis trifft man aber nicht nur bei der jungen Generation, sondern auch bei deren Eltern. Da hören Lehrer schon mal den Vorschlag, doch die Handys einer Klasse einfach mal einzusammeln und die Fotos von der gemobbten Mitschülerin einfach zu löschen. Dann sei das Problem doch gelöst.

Wie mache ich das im realen Leben? Ich verschließe meine Haustüre, weiß aber sehr wohl, dass ich damit einen Einbruch nicht absolut verhindern kann. Ich lege meine Liebesbriefe im Büro nicht auf den Schreibtisch und gehe dann zur Besprechung. Die Fotos von der letzten Party, auf der es ziemlich hoch her ging, zeige ich nur Leuten, auf deren Diskretion ich mich verlassen kann. Ich erzähle nicht jedem, dass ich manchmal richtig kriminelle Träume habe oder dass mein Onkel ein Konto in der Schweiz hat. Ich achte darauf, dass mein Geldbeutel mit allen Kreditkarten beim Shopping auf der Frankfurter Zeil in der verschlossenen Innentasche der Handtasche ist und lasse die Handtasche mit Geld, Handy und persönlichen Briefen im Lokal nicht unbeobachtet weit weg von mir an der Garderobe hängen.

Ich entscheide bewusst, was für mich schützenswerte Privatheit ist und was nicht. Ich entscheide, wie und ob ich mich absichern will gegen fremde Zugriffe und Einblicke. Dabei können andere Menschen für sich zu anderen Schlussfolgerungen kommen – das gehört zur individuellen Freiheit.

Gewisse Vorsichtsmaßnahmen sind sowohl real als virtuell unerlässlich – alles kann man nicht verhindern. Für manches gibt es zusätzlich Versicherungen. Bei dem realen Diebstahl von Kreditkarten habe ich eine Telefonnummer, über die alle meine Karten gesperrt werden, und die mir den Schaden – bis hin zum Bargeldverlust in limitierter Höhe – ersetzt. Die persönliche Post in der Handtasche, die auch schützenswert ist, die ist allerdings nicht versicherbar.

Hier vermischt sich Datenschutz und Schutz meiner Privatheit, denn wer meine Kreditkartendaten missbräuchlich benutzt, hinterlässt auch ggfs. für mich nachteilige Spuren oder er kann durch diesen Zugriff mein Einkaufsverhalten nachvollziehen oder erfahren, was meine Urlaubspräferenzen sind.

Aber das wäre egal, ich hab ja nichts zu verbergen. Wirklich? Mal darüber nachdenken, ob ich im nächsten Urlaub unwissentlich den Leuten mit den kriminellen Absichten gleich wieder begegnen will.

Irritierend ist für mich der Gedanke, ob ich mich grundsätzlich anders verhalte, weil ich davon ausgehen muss, beobachtet zu werden. Und: macht es einen Unterschied für mich, ob ich glaube, dass mich (wohlmeinende?) staatliche Institutionen beobachten oder ob mich Firmen mit wirtschaftlichen Interesse beobachten oder kriminelle Vereinigungen?

Die grundsätzliche philosophische Diskussion – was ist persönliche Freiheit, was ist Privatheit im allgemeinen und im individuellen – müsste viel breiter in der Gesellschaft geführt werden. Manche Mitglieder der Generation Digital Natives haben hier ein ganz anderes Verständnis als jemand, der vor 1960 geboren ist. Und wenn man Privatheit begreift als das Recht auf individuelle Selbstentfaltung, dann muss man zugestehen, dass

es keine engen Regelungen geben kann. Wenn es für jemanden zur Selbstentfaltung dazugehört, eine – nach meiner Meinung – exhibitionistische Selbstdarstellung im Netz zu pflegen, dann soll er das tun – sofern ihm oder ihr klar ist, dass das Netz nicht wirklich vergisst.

Ja, es gibt nun Bestrebungen, dass auch Löschungen möglich sein müssen – aber mit welchem Aufwand!

Sind sich allerdings Mehrheiten einig, was geschützt werden soll, was nicht erlaubt sein soll, ist politisches Handeln erforderlich. Keiner will z.B. von der NSA pauschal ausgespäht werden. Wer weiß, was die NSA (oder eine Vereinigung, die wie auch immer Zugriff zu den NSA-Daten erhält) damit alles macht, machen könnte und wem sie die Daten weitergeben könnte. Hier ist politisches Handeln gefordert.

Für beängstigende gesellschaftliche Entwicklungen braucht es in der realen und in der virtuellen Welt Menschen, die sich zusammentun, dies in einer demokratischen Öffentlichkeit thematisieren und für Abhilfe eintreten.

Wichtig ist eine informationelle Grundausbildung – die Informationsveranstaltungen, die die Polizei dankenswerterweise in vielen Schulen anbietet, reichen nicht aus. Das bleibt nicht ausreichend hängen – dafür braucht es weitergehende pädagogische Konzepte innerhalb und außerhalb der Schule.

Panik ist nicht angebracht, Augen verschließen auch nicht.

Und im Zweifel trete ich für die Freiheit von Facebook ein. Denn mein Freiheitsbegriff ist im Sinne von Rosa Luxemburg, die sagt, dass Freiheit immer die Freiheit der anderen sein muss. Diese sozialen Anwendungen staatlicherseits verbieten zu wollen, das darf nirgendwo eine Option sein und ist hoffentlich auch nicht mehr real durchführbar.

Privatheit und virtuelle Arbeit in Industrie 4.0

Welf Schröter

Forum Soziale Technikgestaltung, Netzwerk ZIMT,
Redaktion www.bloch-blog.de

Eine der zentralen Herausforderungen der Zukunft von Arbeits- und Geschäftswelten liegt im Wandel der Identität und Subjekthaftigkeit der erwerbstätigen Person. War sie früher Teil einer kollegialen Belegschaft und eingebunden in Ort und Zeit des Betriebes, wird sie immer mehr zum Akteur im virtuellen Raum und selbst zum virtuellen Akteur. Die wachsende Virtualisierung führt dazu, dass die Zukunft der Geschäfts- und Arbeitswelten, die Zukunft der Betriebe und Beschäftigungen im virtuellen Raum entschieden wird.

Betriebe, Verwaltungen und Beschäftigte sehen sich einer Überlappung und eines Zusammenwachsens real-materieller und virtuell-immaterieller Entwicklungen gegenüber. Virtuelle und nicht-virtuelle Wirklichkeiten in den Geschäfts- und Arbeitsumgebungen greifen ineinander und erzeugen eine ganzheitlich wahrnehmbare Mischform der Arbeitswelt[12]. Orts- und zeitbezogene Arbeitswelten werden vom Trend zur Ent-Ortung und Ent-Zeitlichung der Geschäftsabläufe durchdrungen.

Das Zusammenwachsen von bisherigen und virtuellen Arbeitswelten stellt das traditionelle Verständnis von Arbeit wie auch das traditionelle „Prinzip Betrieb“ grundlegend in Frage. Der „Ort Betrieb“ (Firma, Werkstatt, Verwaltung, Büro etc.) ist nicht mehr der allein maßgebliche und dominierende Ort der Erbringung von Arbeit und Dienstleistungen. Das Konzept „Industrie 4.0“ beschleunigt zudem diese Tendenz zur Ent-Betrieblichung.

Unter dem Konzept „Industrie 4.0“, das in der Automobilbranche, im Maschinenbau und in der Chemiebranche samt Zulieferern pilotiert wird, versteht man die weitreichende Digitalisierung, Virtualisierung und Neuorganisation von Produktionsabläufen über eine vollständige Wertschöpfungskette hinweg. Dabei soll der Prozess von der Kunden- bzw. Auftragsseite her angelegt werden. Die „digitale Fabrik“ wird selbst zum Produkt.

Das Konzept „Industrie 4.0“ hat drei Gesichter: Es entfesselt einen erheblichen Rationalisierungsgrad, es bildet die strategische Grundlage für den Industriestandort im globalen Wettbewerb und es eröffnet gute Chancen für die Neugestaltung der Arbeit im Sinne der

[12] Welf Schröter: Impulse für Innovation und Beschäftigung in der Metropolregion Rhein-Neckar ASITEMA. Erfahrungsbericht und Beschreibung eines regionalen Innovationsdiskurses ASITEMA – Anstöße für ein Soziales Innovations- und TEchnologieMAnagement zur Stabilisierung und Schaffung von Beschäftigung im Wirtschaftsraum Heidelberg als Teil der Metropolregion Rhein-Neckar – Ausarbeitung und partizipative Beratung von Handlungsempfehlungen. Bericht an die Hans-Böckler-Stiftung Förderschwerpunkt „Strukturwandel – Innovation und Beschäftigung“. Projekt S-2009-283-1. Juli 2011.

Humanisierung. Die zunehmende und übergreifende Digitalisierung der Arbeitswelt birgt riesige Potenziale[13] in der Produktion und Wertschöpfung und löst gleichzeitig Ängste bei den Beschäftigten aus.[14] Dreh- und Angelpunkt einer sozialen und innovationsförderlichen Technikgestaltung muss daher die Beteiligung der Beschäftigten und deren Entwicklungs- und Qualifizierungsmöglichkeiten im Prozess sein.

In den zunehmend IT-gestützten Arbeitswelten wächst dem handelnden natürlichen Menschen eine oder mehrere virtuelle Zwillingsidentitäten („virtuelle Ichs") zu. Das technisch abstrakte „virtuelle Ich" entsteht zunächst durch die aktiven und passiven Ansammlungen von Datenmengen und Profilen. Es wird qualitativ wirkungsmächtig und beginnt das „biografische Ich" schrittweise zu beeinflussen.

Der Begriff „Identität" lässt sich auf verschiedene Weise verstehen als pädagogisch-didaktisch-erzieherisches, als psychologisches Entwicklungsziel, als beruflich-inhaltliche corporate identity, als soziokulturelles Ergebnis, als technisch verstandenes Konstrukt von Rollen und Berechtigungen. Von technischer Seiten werden Identitäten geschaffen durch E-Mail-Accounts, Websites, Mitgliedschaften in Communities, durch zugewiesene Nutzerrolle, Rechteinhaber (Plattform-Zugangsrechte), Avatare, Augmented Identity, Facebook-Präsenz (oder XING, Google, Twitter etc.), durch die Anwendung elektronischer Signaturen, durch die die Nutzungen des elektronischen Personalausweises oder der elektronischen Gesundheitskarte. Ergänzt wird diese technisch geformte Identität durch personalisierte „Butler" (Softwareagenten), Bots oder andere Automatisierungstechniken und anderes mehr.

Der technikgetriebenen Aussage „Wer bin ich im Netz und wieviele?" muss um ein nutzergesteuertes Identitätsmanagement erweitert werden, das es Individuen erlaubt, in konzeptionell geplanter Weise die eigene Identität im Netz (Virtuelles Ich") aufzubauen und zu pflegen. Dazu ist eine rechtlich abgesicherte Kultur der Anwendungen von Pseudonymen und Anonymisierungsinstrumenten unabdingbar.

Datensammelnde Entstehungsebenen virtueller Identitäten (aus Nutzersicht) können nicht-dynamisch und fremdgesteuert (z.B. elektronischer Personalausweis), dynamisch und fremdgesteuert (z.B. Suchmaschine Google, Profilbildung bei Amazon etc.), nicht-dynamisch und nutzergesteuert (z.B. hinterlegte Pseudonyme), dynamisch und nutzergesteuert (z.B. personalisierbare e-Butler/Softwareagenten in Smart Grids) sein.

Den parzellierten Deutungen von Identität gilt es, ein gesamtheitliches Verständnis von Identität gegenüber zustellen, das dem Lebenslagen-Ansatz der Sozialwissenschaften folgt.

Das neue arbeitende Individuum im „New Blended Working" bzw. in Croudsourcing-Kontexten wird von den kulturellen Prägungen und Regulationen der Sozialpartner im-

[13] Jutta Rump, Isabel Biegel: Arbeit und Freizeit. Wie wir in Zukunft leben und arbeiten werden. Mössingen 2009.

[14] Siehe hierzu: „IT und Beschäftigung – Memorandum des Arbeitskreises Informations- und Kommunikationstechnik des IG Metall Bezirks Baden-Württemberg in Zusammenarbeit mit dem Forum Soziale Technikgestaltung. Version 2.0. Verfasst vom Autorenteam Norbert Baumgarten, Hansjörg Jäckel, Johannes Katzan, Welf Schröter, Jürgen Werner, Thomas Windgassen, Ulrike Zenke und andere. Stuttgart Mai 2014.

mer weniger erreicht. Die Kollegialkultur und deren sozialisierende Kraft aus Betrieben, Belegschaften, Teams, Verbänden und Sozialpartnern verlieren ihre Interpretationshoheit in den aufkommenden virtuellen Einkommenswelten. Die strukturelle Ent-Integration des arbeitenden Individuums aus den postindustriellen Strukturen der Arbeitswelten[15] zieht eine latente Tendenz zur Ent-Sozialisation nach sich.

Angesichts des beginnenden Ent-Kohäsionsprozesses der Gesellschaft muss die Frage nach neuen Anlässen, Orten und Kulturen formuliert werden, wie und wo kompensatorische Kohäsionsorte und ebensolche Kohäsionskräfte wachsen können. Nicht die realen und virtuellen Orte des lebenslangen Broterwerbs strukturieren die neuen Einkommens-Patchwork-Biografien, sondern es sind die realen und virtuellen Orte des emotional-intelligenten Lernens als lebenslange Lernkultur.

Wir brauchen ein Denken, das nicht nur von Geschäftsprozessen her agiert, sondern die Identitätsstiftung als Ziel setzt und aufwertet[16]. Der Prozess der Identitätsstiftung umfasst zukünftig sowohl das „biografische Werden“ als auch die „virtuelle Lebenslage“. Das „biografische Ich“ verhält sich zu seinem „virtuellen Ich“ dabei ungleichzeitig.

Thesen

1. Privatheit wird gewährleistet durch die Autonomie und das Recht des Individuums, seine Identität innerhalb und außerhalb elektronischer Netze selbst zu bestimmen.

2. Privatheit ist ein zentrales Element im Fundament einer anti-totalitären demokratischen Zivilgesellschaft. Ohne Privatheit gibt es keine Demokratie.

3. Privatheit braucht die Möglichkeiten und Chancen der Herausbildung eines eigenen Identitätsvermögens und der nutzergesteuerten Identitätsstiftung in elektronischen Arbeits- und Lebensumgebungen.

4. Privatheit und virtuelle Identität(en) werden zu zentralen Kategorien der Entfaltung einer demokratischen Wissensgesellschaft.

5. Der zunehmenden Virtualisierung von Arbeit und Leben gilt es, eine bewusste Strategie der Schaffung von realen Treffpunkten zivilgesellschaftlicher Communities zur Seite zu stellen, um die soziale Kohäsion zu stabilisieren.

6. Eine humane Nutzung moderner Informationstechnologie nimmt sich zum Ziel, die Privatheit des Individuums und das nutzergesteuerte Identitätsmanagement im Netz technisch zu sichern.

---

[15] Dirk Balfanz, Welf Schröter (Hg.): Gestaltete Virtualität. Realität der neuen Medien in der Arbeitswelt – Standortbestimmung und Perspektiven. Mössingen 2010.

[16] Zu dem Kreis der Diskussionsansätze um soziale Rechte im Netz, gehört auch der gewerkschaftliche Diskurs „SozialCharta Virtuelle Arbeit“ des „ZIMT“-Projektes aus der Metropolregion Rhein-Neckar, das von den dortigen IG Metall-Verwaltungsstellen, Betriebsräten und dem Forum Soziale Technikgestaltung getragen wird.

7. Privatheit und virtuelle Identität(en) erhalten ihr zunehmendes Gewicht durch die Konvergenz bislang getrennter virtueller Handlungsräume und durch den strukturellen Wandel der Arbeits-, Erwerbs- und Geschäftswelt.

8. Dem Rückgang der Bedeutung des Ortes Betrieb steht die anwachsende Bedeutung der realen und virtuellen Urbanität in Gestalt komplexer Konzepte von Smart Cities entgegen.

9. Der Schutz von Privatheit und virtuellen Identitäten bedarf einer rechtsverbindlichen Option der Anonymisierung und Pseudonymisierung.

10. Neben der Entwicklung der biografisch-sozialen Identität („biografisches Ich“) wachsen dem Individuum in Arbeit und Leben neue virtuelle Identitäten („virtuelles Ich“) zu.

11. Die Identität einer zukünftig erwerbstätigen Person wird sich durch die Verknüpfung des „biografischen Ich“ und des „virtuellen Ich“ ergeben.

12. Um diesen Prozess im Sinne eines gemeinwohlorientierten sozialen Miteinanders zu gestalten, bedarf es aktiver Maßnahmen in Bildung und Ausbildung, Kunst und Kultur.

Exkurs 3
Arbeitgeber und Gewerkschaften im Vergleich – wie gut sind diese auf Crowdsourcing vorbereitet?

Jan Marco Leimeister

Forschungszentrum für IT-Gestaltung an der Universität Kassel
Institut für Wirtschaftsinformatik an der Universität St. Gallen (IWI HSG)
JanMarco.Leimeister@unisg.ch

Unser Leben wird immer stärker von digitalen Technologien beeinflusst. Die Digitalisierung verändert unsere Gesellschaft, unsere Wirtschaft und unser Leben mit nie gekannter Geschwindigkeit – und damit unter anderem auch die Art, wie wir arbeiten bzw. wie Unternehmen ihre Arbeits- und Leistungserstellungsprozesse organisieren. Das Internet bietet Unternehmen heutzutage die Möglichkeit, Aufgaben über das Internet weltweit zu verteilen und somit auf ein großes Reservoir an Arbeitskräften schnell und gezielt zurückzugreifen. Diese Art der Aufgabenabwicklung bzw. -organisation wird „Crowdsourcing" genannt und beschreibt im Allgemeinen die Auslagerung von Unternehmensaufgaben auf eine unabhängige Masse von Menschen über das Internet. Die Mitglieder einer Crowd agieren hierbei als Digital Worker bzw. Crowd Worker und übernehmen kollektiv Aufgaben, die typischerweise von den Mitarbeitern innerhalb eines Unternehmens vollzogen werden.

Natürlich gibt es unterschiedliche Spielarten und Ausprägungsformen von Crowdsourcing – in der Forschung unterscheidet man zumeist zwischen „internem" und „externem" Crowdsourcing. Während beim „internen" Crowdsourcing die gesamte unternehmensinterne, u.U. weltweit verteilte Belegschaft die Crowd bildet, ist die Crowd beim „externen" Crowdsourcing aus beliebigen (auch unternehmensexternen) Individuen zusammengesetzt, die oft in sog. Communities organisiert sind. Beim externen Crowdsourcing bedienen sich Unternehmen zumeist so genannten Crowdsourcing-Intermediären, die selbst eine (aktive) Community – bestehend aus Internetnutzern aus aller Welt – aufbauen und Crowdsourcing betreibenden Unternehmen die Möglichkeit bieten, ihre Aufgaben über eine aufgesetzte Crowdsourcing-Plattform auszulagern.

Der Vorteil, der für Unternehmen durch Crowdsourcing entsteht, ist die Möglichkeit, viele, zum Teil sehr spezialisierte, Arbeitskräfte flexibel, schnell und jederzeit in die Entwicklungsprozesse einzubinden. Nehmen wir zum Beispiel das Testen von Softwareapplikationen: Anstatt das Testen einer entwickelten Applikation nur von wenigen internen Mitarbeitern durchführen zu lassen, können Softwareunternehmen ihre Applikationen von vielen Crowd Workern – die meist freiwillig und mit Freude am Testen mitwirken – in allen Phasen der Entwicklung testen lassen. Dadurch können sich Unternehmen beispielsweise auf ihre Kernkompetenzen (hier bspw. die eigentliche Softwareentwicklung) konzentrieren. Einzelne Studien zeigen auch, dass Crowdsourcing Kostensen-

kungspotenziale haben kann – z.B. durch Ausnutzung von Lohnkostenunterschieden aufgrund der globalen Verteilung von Aufgaben oder durch Einbindung von niedriger qualifizierten Arbeitern, die entsprechend günstiger sind.

Nahezu jede Wertschöpfungsaktivität bzw. -aufgabe, die digital übertragbar ist, kann von der Crowd durchgeführt werden. Dazu muss diese nur konkretisiert, detailliert beschrieben und in kleine (Arbeits-) Einheiten zerlegt werden. Ähnlich den Prinzipien des Taylorismus können so komplexe Arbeitsvorgänge durch Zerlegung und Standardisierung effektiver bearbeitet werden. Die Arbeit bzw. die Aufgabenerfüllung wird dann von den Crowd Workern erbracht und die einzelnen Arbeitsergebnisse werden anschließend zu einer größeren Gesamtleistung wieder zusammengesetzt.

Crowdsourcing als Arbeitsorganisationsmodell ist relativ neu und noch weitgehend unerforscht, wird aber in zahlreichen Beiträgen und Medienberichten schon sehr kontrovers diskutiert. Hierbei wird meist aber nicht zwischen den zwei Ausprägungsformen internes und externes Crowdsourcing unterschieden – obwohl dies zwingend notwendig für das Verständnis wäre. Für Unternehmen ist externes Crowdsourcing in vielen Fällen sehr reizvoll, da hierbei der Organisations- und Betreuungsaufwand von Crowdsourcing-Intermediären erbracht wird und auf außerhalb des Unternehmens liegende Ressourcen zugegriffen werden kann.

Die Anbieter / Intermediäre, die externe Crowdsourcing-Lösungen anbieten, stehen manchmal in der Kritik aufgrund der vermuteten Arbeitsbedingungen, die sie ihren Crowd Workern bieten. So wird in einzelnen Beiträgen vor der Entstehung von „digitalen Ausbeuterbetrieben“ (engl. „digital sweatshops“) gewarnt, da die Entlohnungen für die Crowd Worker auf solchen Plattformen zum Teil sehr gering ausfallen kann und zudem nicht sicher sind. Die Entlohnung von Crowdsourcees ist zudem sehr unterschiedlich. Während bei vielen einfachen Tätigkeiten ein Reservationsgehalt von wenigen Euros pro Stunde vorliegt, gibt es auch Plattformen, auf denen Crowd Worker einen durchschnittlichen Stundenlohn von 200 Euro erhalten. Andere Entlohnungsformen und -modelle wie Sachprämien oder Bonuspunkte sind auch teilweise anzutreffen. Die Spanne der verschiedenen Entlohnungsarten und -höhen zeigt, dass etwaige Standards noch nicht etabliert worden sind. Kritisch beäugt wird auch, dass die externen Crowd Worker als Selbständige agieren – sofern Crowd Worker beispielsweise einen Großteil ihres Einkommens auf einer Plattform generieren, können auch Fragen der Scheinselbstständigkeit eine Rolle spielen. Weiterhin bestehen keinen Regelungen zu Eigentumsrechten, die beim externen Crowdsourcing aufgrund der AGBs zumeist direkt abgetreten werden.

In Bezug auf das interne Crowdsourcing müssen wir leider feststellen, dass er hierzu noch viel weniger Erkenntnisse sowohl über die Perspektive der Unternehmen als auch über die Perspektive der Crowd Worker gibt. Breit diskutiert wurde die unternehmensweite Einführung von Crowdsourcing bei IBM über das sog. ‚Liquid Programm‘, im Rahmen dessen zumeist unternehmensinterne Ausschreibungen für bestimmte Aufgaben in Form von Wettbewerben durchgeführt werden, für die sich grundsätzlich alle (weltweit verteilten) IBM Mitarbeiter bewerben können. Auch andere Unternehmen – vor allem aus der Softwarebranche – unternehmen erste Versuche mit internem Crowdsourcing, wissen aber nicht, wie diese neue Art der Arbeitsverteilung in bestehende Unter-

nehmensstrukturen- und Prozesse einzubetten ist. Außerdem ist nicht bekannt, wie man die Arbeitsprozesse im Rahmen dieser vermeintlich hierarchiefreien Arbeitsorganisationsform kontrolliert und koordiniert und wie Entlohnungsstrukturen anzupassen sind – denn bei international agierenden Unternehmen sind zwangsweise Lohnkostenunterschiede zu berücksichtigen.

Um das Potenzial von internem Crowdsourcing zu heben, sind Erkenntnisse zu Steuerungs- und Kontrollmechanismen, zur Gestaltung von effektiven Anreizmechanismen oder auch Erkenntnisse zu Erfolgsfaktoren erforderlich.

Für Crowd Worker bzw. die Belegschaft von Unternehmen, die Crowdsourcing einführen, ergeben sich durch internes Crowdsourcing weitreichende Implikationen: So ist bspw. für Crowd Worker selbstbestimmte Arbeitszeitgestaltung denkbar, da die Aufgaben per Internet verteilt werden. Ebenso ist eine Spezialisierung auf die Tätigkeiten denkbar, bei denen der Crowd Worker besondere Kompetenzen oder Präferenzen hat – und durch diese Spezialisierung ist auch die Menge an „ungeliebten" Aufgaben deutlich verringerbar. Gleiches gilt für die Arbeitnehmer im Unternehmen, die durch crowdgesourcte Aufgaben von vermeintlich ungeliebten Aufgaben (bspw. die Dokumentation von Software oder Support-Anfragen) entlastet werden können und sich auf ihre Kernaufgaben fokussieren können.

Neben diesen potentiell positiven Aspekten werden in der Literatur jedoch insbesondere potentiell negativ behaftete Auswirkungen des internen Crowdsourcings diskutiert: Die Arbeitsbedingungen können sich drastisch ändern, wenn man sich beispielsweise für jede intern ausgeschrieben Aufgabe bewerben muss und damit immer in (zunehmender) Konkurrenz zu den Arbeitskollegen steht. Damit kann unternehmensinternes Crowdsourcing auch Einfluss auf das „Miteinander" der Beschäftigten und somit auf die Unternehmenskultur haben. Ein weiterer Nachteil in diesem Zusammenhang ist die durch neue Technologien dargebotene Möglichkeit, digitale Arbeitsprozesse elektronisch zu überwachen. Für Mitarbeitervertreter gilt es entsprechend, diese Aspekte in der Zukunft genauer zu betrachten. Besonders das Thema der Mitbestimmungsrechte von Crowd Workern und deren Interessensvertretung ist in der Zukunft zu klären. Zudem gilt es auch zu untersuchen, ob und inwieweit tarifliche Standards festgelegt werden können. Durch die Internationalität von Crowdsourcing-Intermediären und die damit verbundenen internationalen Standorte, tritt hier ein weiteres Problemfeld auf.

Unter dem Strich ist zu festzustellen, dass internes und externes Crowdsourcing jeweils sowohl für Arbeitnehmer – also Crowd Worker – als auch für Crowdsourcing betreibende Unternehmen Chancen, aber auch Risiken bergen können. Insgesamt sind noch viel zu wenige Kenntnisse zu Wirkungszusammenhängen, Effekten und Gestaltungsmöglichkeiten vorhanden. Die Chancen zu nutzen – und gleichzeitig die Risiken zu minimieren – verlangt nach guten Spielregeln. Diese zu erarbeiten, erfordert wiederum eine solide Kenntnis der unterschiedlichen Spielarten, Prinzipien und Funktionsweisen. Dieser Aufgabe müssen wir uns stellen, denn Crowdsourcing wird in Zeiten der zunehmenden Digitalisierung der Arbeit immer mehr an Bedeutung gewinnen.

Vorstellung des Qualitätssiegels „Lebensqualität durch IT“ der Integrata-Stiftung

Dr. Bertram Lohmüller, Prof. Dr. Rolf Pfeiffer

Steinbeis Global Institute Tübingen(SGIT) der Steinbeis-Hochschule Berlin SHB
Waldhörnlestraße 18
72072 Tübingen
bertram.lohmueller@gimaconsult.com
rolf.pfeiffer@steinbeis-git.org

Abstract: Das Qualitätssiegel „Lebensqualität durch IT“ zeichnet Unternehmen aus, die bei dem Einsatz von Informationstechnologie und bei der Interaktion Mensch-Technik, den Menschen in den Mittelpunkt stellen. Kriterien sind u.a. Flexibilisierung der Arbeitszeit, Worklife-Balance, Integration privater mobiler Endgeräte in das Unternehmensumfeld (bring your own device), Gesundheit und Ergonomie. Das Qualitätssiegel beinhaltet ein Benchmarking und ein Qualitätsaudit. Im Rahmen des Benchmarkings werden Konzepte und Vorgehensweisen im Markt verglichen. Auf Basis der Benchmarking-Ergebnisse haben die Unternehmen dann die Möglichkeit, sich einem Qualitätsaudit zu unterziehen und das Qualitätssiegel „Lebensqualität durch IT“ der Integrata-Stiftung zu erwerben. Die Konferenz ist der Start für das Benchmarking-Projekt „Best Practices in Lebensqualität durch IT“. Fragebögen können auf der Konferenz angefordert werden. Die ersten Audits für das Qualitätssiegel sind für das Frühjahr 2015 geplant.

Der Preis für humane Nutzung der IT - heute und in Zukunft

Prof. Dr. Wolfgang Heilmann

Präsident des Kuratoriums
Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
wolfgang.heilmann@integrata-stiftung .de

Abstract: Mit dem Wolfgang-Heilmann-Preis für humane Nutzung der Informationstechnologie werden seit Jahren herausragende Leistungen gewürdigt, die eine humane Nutzung der Informationstechnologie in unterschiedlichen Zusammenhängen nachweisen. Der Wolfgang-Heilmann-Preis 2013/2014 steht unter dem Motto „Privatheit in der E-Society“.

Liebe Preisträger,

Meine Damen und Herren,

auf dem Programm steht jetzt "Preisverleihung durch den Stifter und die Laudatoren". Welcher Preis das ist und wofür er verliehen wird, das werde ich Ihnen nun kurz erläutern:

- Wir vergeben heute den **Wolfgang-Heilmann-Preis für humane Nutzung der Informationstechnologie**. Er ist 1999 aus dem Preis für integrierte Datenverarbeitung hervorgegangen, der von 1988 bis 1999 von der INTEGRATA-AG verliehen wurde. Dieser Firmenpreis hatte die Funktionalisierung und Rationalisierung der Arbeitsprozesse durch die IT im Focus. Mein Preis hat sich dagegen das Ziel der Humanisierung der IT gesetzt, wie die Stiftung selbst. Wir verfolgen mit unserem Preis also die Strategie der **Steigerung der Lebensqualität durch die IT**. Die IT ist für uns ein Mittel zum Zweck der Verbesserung der Lebensverhältnisse der Menschen, nicht vorrangig der Wirtschaft.

- Diese allgemeine Aussage haben wir dadurch präzisiert, dass wir seit 1999 Jahr für Jahr eine andere **Dimension der Lebensqualität** zum Preisthema gewählt haben. Das zeigt die Liste der bisherigen Preisverleihungen deutlich (Anlage 1). Sie reicht von "Telearbeit" über "Wissensmanagement" bis zu "Bürgernahen Anwendungen" in der Politik. So haben wir Schritt für Schritt den weiten Raum dessen, was Lebensqualität bedeutet, ausgeleuchtet und inzwischen **10 Thesen** zur Verbesserung der Lebensbedingungen durch humane Nutzung der Informations- und Kommunikations-

technologie formuliert (vgl. Anlage 2). Sie reichen von "Gesundheit und Wohlbefinden" über "Arbeit und Erwerb" bis zu "Kultur und Religion".

- Der Preis, der heute verliehen wird, gehört auf dieser Liste zur 4. Dimension der Lebensqualität: sie soll das Vertrauen zwischen den Kommunikationspartnern durch verantwortungsvollen Umgang mit Informationen und fairer Kommunikation wieder herstellen - eine Forderung, die im Zeichen der NSA-Affaire eine besondere Aktualität und Brisanz erfuhr. Wir haben bereits im Frühsommer 2013 das konkrete Thema der Preisausschreibung, nämlich:

"Privatheit in der E-Society"

festgelegt und damit die Frage in die Gesellschaft eingebracht, wie Privatheit gewahrt werden kann, auf welche Weise virtuelle Identitäten geschützt werden können - auch und vor allem in sozialen Medien. Gesucht waren Vorschläge, Konzepte und Lösungen für den Einsatz von IT, um Privatheit von physischer und virtueller Identität auf humane und gemeinwohlorientierte Weise in der E-Society zu gewährleisten.

Aus der Vielzahl der eingereichten Arbeiten wurden dann mit Hilfe einer speziell für diesen Zweck entwickelten Nutzwertanalyse die Arbeiten ausgewählt, die Ihnen im Anschluss an meine Ausführungen von den Laudatoren vorgestellt werden. Das Verfahren stützte sich dabei insbesondere auf folgende **Auswahlkriterien**:

1. Übereinstimmung mit dem Stiftungszweck der humanen Nutzung der IT,
2. Neuigkeit und Kreativität, insbesondere im Hinblick auf die Umsetzung in die Praxis,
3. Reife- und Generalisierungsgrad der vorgestellten Anwendung,
4. Bedeutung, Gewicht, Qualität des Gestaltungsvorschlags im Hinblick auf die Lebensqualität und
5. Darstellung in Bezug auf Form, Stil, Transparenz und Medienkompetenz.

Die **Preisjury** bestand neben den 4 Laudatoren, die Sie anschließend hören werden, aus weiteren 15 Damen und Herren des Kuratoriums, sowie des Vorstands und des Fachbeirats der Stiftung. Viele von uns haben Sie bereits durch ihre Beiträge am heutigen Tage kennen gelernt. Wir alle vertreten zwar unterschiedliche Disziplinen, sind also Wirtschafts- und Sozialinformatiker, Sozialwissenschaftler, Ingenieure oder Berater aus Wirtschaft und Gesellschaft, aber wir sind alle gleichermaßen überzeugt von der Sache der humanen Nutzung der IT. Die diesjährige Preisauswahl hat uns wegen der hohen Qualität der Einreichungen und der Bedeutung ihrer Thematik für das tägliche Leben der Menschen ganz besonders bewegt.

Wir hoffen nun sehr, dass Sie der bisherige Tagungsverlauf ebenfalls in Bewegung gebracht hat und Sie die Preisträger mit gebührendem Interesse begrüßen werden.

Damit übergebe ich nun an die Preisträger und Laudatoren. Zuerst wird Prof. Hieber, langjähriger Kurator der Integrata-Stiftung, das Wort ergreifen.

Laudatio Preis für humane Nutzung der IT: Friend Inspector – Bewusstsein für Privatsphäre in sozialen Netzwerken spielerisch erlernen

Prof. Dr. Günther Pernul (Universität Regensburg) und Team: Alexandra Cetto, Dr. Michael Netter, Christian Richthammer, Dr. Moritz Riesner, Christian Roth

Prof. Dr. Ludwig Hieber
Kurator der Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
Ludwig.Hieber@informatik.uni-stuttgart.de

"Bei dem Kriterium „Neuigkeit und Kreativität" wurde klar, dass erstmals ein Integrata-Preis die „Thematik Computerspiele" tangiert und damit die Gefahr von Missverständnissen einhergeht. Der in der eingereichten Arbeit dargelegte wissenschaftliche Hintergrund des Einsatzes von „Serious Games" als einem induktiven Lernansatz zur Verbesserung des Datenschutzes und der Privatsphäre in sozialen Netzwerken hat schließlich überzeugt.

Wichtig war für die Jury auch, dass mit „Friend Inspector" ein praktisch nutzbares Produkt eingereicht wurde, wenn auch zunächst beschränkt auf Facebook Anwender."

Zitat aus der Einreichung Prof. Pernul und Team „**Friend Inspector – Bewusstsein für Privatsphäre in sozialen Netzwerken spielerisch erlernen**":

„Friend Inspector basiert auf der Kombination der wissenschaftlichen Erkenntnisse zu Privatsphäre-Bewusstsein in sozialen Netzwerken und Serious Games als insbesondere bei Jugendlichen und jungen Erwachsenen effektiver Ansatz zur Erreichung von Lernzielen. Die Lernziele gehen dabei aus den wissenschaftlichen Erkenntnissen hervor und werden durch den Serious Games-Ansatz umgesetzt.

Folgende Lernziele wurden definiert:

- Verbesserung des Privatsphäre-Bewusstseins:
  Nutzer sollen die Auswirkungen ihrer Privatsphäre-Einstellungen verstehen. Dadurch soll die Lücke zwischen wahrgenommener und tatsächlicher Sichtbarkeit ihrer preisgegebenen Informationen verringert werden.

- Verbessertes Verständnis der Privatsphäre-Einstellungen:
  Nutzer sollen durch Handlungsempfehlungen die Privatsphäre-Einstellungen besser verstehen und so befähigt werden, diese gemäß ihrer Sichtbarkeitspräferenzen anzupassen.

Die Rahmenhandlung als wesentliches Element von Serious Games definiert den Kontext, in dem das Spiel stattfindet und besteht bei Friend Inspector aus zwei Teilen:

- Einerseits, Nutzer zu motivieren und herauszufordern, ihr Privatsphäre-Bewusstsein zu testen und eine möglichst hohe Punktzahl zu erreichen.
- Andererseits sollen Nutzer die Möglichkeit haben, ihre Punktzahl auf Facebook zu teilen und somit einen Wettbewerb unter Freunden um die höchste Punktzahl zu verstärken.

Dies wird in Friend Inspector durch eine herausfordernde Frage auf der Startseite sowie der Ermittlung einer Gesamtpunktzahl, welche auf Facebook geteilt werden kann, erreicht.“

Laudatio Preis für humane Nutzung der IT: Methode Empirical Living Lab (EmLiL) zur Entwicklung von Informations- und Kommunikationstechnologien

Dr. Stefan Hellfeld (FZI), Dr. Agnes Koschmider (KIT),
Prof. Dr. Andreas Oberweis (FZI)

Dr. Frank Schönthaler
Kurator der Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
frank.schoenthaler@promatis.de

„Mit Empirical Living Lab (EmLiL) hat das Karlsruher Forscherteam Hellfeld/ Koschmider/Oberweis der Preisjury eine neuartige Methode zur Entwicklung von Informations- und Kommunikationstechnologien vorgestellt. Durch die Kombination von empirischen Untersuchungen mit der Evaluation in der realitätsnahen Zukunftsatmosphäre des FZI House of Living Labs (HoLL) ermöglicht EmLiL die Entwicklung von Technologien, die die Einstellung einer individuellen situations- und anwendungsabhängigen Privatheit unterstützen. Erste Erfahrungen belegen, dass die mit EmLiL entwickelten Technologien interessante Potenziale zur Verbesserung der Lebens- und Arbeitsqualität eröffnen. Insofern steht die eingereichte Arbeit im Einklang mit dem Stiftungszweck.

Die Preisjury war sowohl von der Methode selbst als auch von der professionellen Darstellung des Konzepts und der Ergebnisse überzeugt. Die durchweg positiven Bewertungen bezüglich aller relevanten Aspekte qualifizieren die Arbeit als preiswürdig.“

Laudatio Sonderpreis „Innovative Produkte“: Trustner- Einfache und sichere Lösung zur Wahrung der Privatheit digitaler Inhalte

Dr. Harald Sondhof (Trustner GmbH)

Dr. Peter R. Walti
Kurator der Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
peter.walti@waltipartners.ch

Unter den Namen „Trustner“ wurde von Dr. Harald Sondhof ein ausserordentlich leistungsfähiges Gesamtsystem entwickelt, das die Privatheit in der elektronischen Kommunikation schützt, einfach wie Social Networks zu nutzen ist, die Kontrolle über die Inhalte und Adressaten gewährleistet und den technischen Schutz vor der Einsichtnahme durch Dritte in der End zu End Verschlüsselung sicherstellt. Mit Trustner wird die Lücke zwischen sicheren, aber komplizierten und unsicheren, aber bequemen Kommunikationsverfahren geschlossen. So ist Trustner für viele Menschen sowohl im Geschäftsleben, im Gesundheitswesen und im Finanzsektor als auch im privaten Umfeld von großem Interesse.

Der sog. Trustner Client wird – wie Skype oder Spotify – auf dem Rechner des Nutzers installiert. Dabei findet jede Kommunikation bei Trustner in sog. „Räumen“ statt, die der Nutzer mit einem Blick überschauen und jederzeit verlassen und damit löschen kann. Nur der Nutzer selbst verfügt über die Schlüssel zur Ver- und Entschlüsselung seiner Kommunikation. Trustner selbst hat als Betreiber keinen Zugriff auf die Daten der Nutzer.

Trustner bietet ein vollständig automatisches Verfahren zur Verschlüsselung elektronischer Daten. Die verschiedenen Einzelschritte, die normalerweise ein Internetnutzer bei der Verwendung eines sicheren Verschlüsselungsverfahrens durchführen muss, werden vom Trustner Client selbst übernommen.

Über einen sog. „Store“ findet der Nutzer von Trustner verschiedene Raumtypen sowie Zusatzfunktionen und Branchenlösungen. So ermöglicht der sog. *Team-Raum* die Kommunikation in der Gruppe, die von jedem Mitglied erweitert werden kann. Wer eine Einladung in einen Team Raum erhält, kann diesem beitreten, ohne dass andere Kontakte im Raum dieser Einladung zustimmen müssen. Demgegenüber steht der sog. *Club-Raum*, der die Kommunikation in einer Gruppe ermöglicht und der nur mit Zustimmung aller weiteren Mitglieder erweitert werden kann.

Als kostenpflichtige Anwendung bietet der Trustner Store u.a. das *Trustner-Sprechzimmer* an. Diese Anwendung wird von einem Trustner-Reseller (careon GmbH) für Ärzte betrieben, die mit ihren Patienten und deren Angehörigen elektronisch kommunizieren möchten. Eine weitere kostenpflichtige Anwendung ist der Trustner GlucoChat,

der in einer speziellen Form von Trustner-Sprechzimmern der aktiven Betreuung von Kindern und Jugendlichen mit Diabetes Typ 1 dient.

Überzeugt hat die Preisjury das Trustner-Sicherheitssystem durch seine hohe Benutzerfreundlichkeit, mit der die Privatheit der elektronischen Kommunikation gegenüber unerwünschter Drittzugriffe technisch geschützt und durch den Nutzer in transparenter Weise selbst kontrolliert werden kann. Die bereits angebotenen sowie das große Potential an möglichen weiteren Anwendungen lassen das Trustner System als ein Sicherheitssystem mit einer grossen Zukunft als überaus preiswürdig erscheinen.

Laudatio Sonderpreis „Innovative Produkte“: REDS – Remotely Encrypted Distributed Storage

Torben Haase und Annika Schulz (Flowy Apps Schulz & Haase)

Hans-Peter Fröschle
Vorstand der Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
hans-peter.froeschle@integrata-stiftung.de

Das Internet als weltumspannendes Informations- und Kommunikationsnetzwerk in seiner ursprünglichen Philosophie: dezentrale technische Implementierung und dezentrale Organisationsstruktur.

Das Internet in seiner aktuellen Ausprägung: Zunehmend konzentrieren große Provider den Netzverkehr über ihre Server und Knoten. Dadurch entstehen neue Machtkonstellationen, da zunehmend zentralisierte Stellen den Netzverkehr kontrollieren, konzentriert angreifbar sind und den Internetnutzer mit einem Gefühl zurücklassen, wenig bis keine Kontrolle mehr über die eigenen Daten zu haben. Verschlüsselung der Daten mit den heute gängigen Verfahren bedeutet für die Nutzer leider immer noch Komfortverlust.

Eine Situation, die das Massenmedium Internet unter Gesichtspunkten der Privatheit der Nutzer mehr und mehr in Verruf bringt.

Was läge hier näher als die ursprünglich implementierten dezentralen Strukturen des Internets wieder umzusetzen?

Genau in diese Zielrichtung geht der Vorschlag der beiden Startup-Unternehmer Torben Haase und Annika Schulz mit ihrer Einreichung zum Integrata-Preis. Eine einfach zu nutzende technische Lösung, die die Nutzerdaten und Anwendungsprogramme dadurch voneinander trennt, dass die Nutzer autonom entscheiden können, wo ihre Daten physikalisch gespeichert werden. Zudem erfolgt in dem Konzept der – erforderliche – Austausch schützenswerten Nutzerdaten mit den Providern, die auf ihren Plattformen diese Daten verarbeiten müssen, verschlüsselt. Der Provider hat die Daten zu keinem Zeitpunkt unverschlüsselt im Zugriff.

Trennung von Daten und Programm – verschlüsselte Übertragung der Daten zwischen Nutzer und Provider und das alles unter Kontrolle der Benutzer! Damit ist ein signifikanter Beitrag des eingereichten Vorschlags zur Sicherstellung der Privatheit in der eSociety gegeben. Damit verbunden ist eine deutliche Steigerung der Lebens- und Arbeitsqualität beim Umgang mit der „Alltagstechnologie“ Internet. Zusätzlich wird die vorgeschlagene Lösung als Entwickler-Framework unter einem Dual-Licensing-Modell

angeboten – zur privaten Nutzung kostenfrei und jederzeit für eine Code Inspection offen.

Ein kleines, schlagkräftiges Team mit organisatorischen, betriebswirtschaftlichen und technischen Kompetenzen, einem über den Vorschlag hinausreichenden Engagement für Transparenz und Nachhaltigkeit und dem „Feuer“ junger Unternehmensgründungen sind weitere Indikatoren für eine Preiswürdigkeit im Sinne der humanen Nutzung der Informationstechnologie.

Der Preis für humane Nutzung der IT - heute und in Zukunft

Prof. Dr. Wolfgang Heilmann

Präsident des Kuratoriums
Integrata-Stiftung für humane Nutzung der Informationstechnologie
Schleifmühleweg 70
72070 Tübingen
wolfgang.heilmann@integrata-stiftung .de

Liebe Preisträger,

ich übergebe Ihnen nun die Urkunden zur Preisverleihung und das Preisgeld.

Was die Bedeutung Ihrer Beiträge anbetrifft, so kann ich mich den Ausführungen der Laudatoren voll anschließen. Ich möchte aber noch zwei Bemerkungen in eigener Sache und als Ausblick für das nächste Jahr bringen.

Meine Damen und Herren:

- Das diesjährige Preisgeld ist jeweils zur Hälfte an die beiden ersten Preisträger gegangen. Die Preisträger des Sonderpreises "Innovative Produkte" müssen sich mit einer Urkunde und der anerkennenden Erwähnung begnügen, weil der Wolfgang-Heilmann-Preis nur mit 10 T€ ausgestattet ist. Diese finanzielle Enge wird der Weite des Themas der humanen Nutzung der IT nicht gerecht. Deshalb suchen wir **Zustifter** und **Sponsoren**, die einen finanziellen Beitrag zur Steigerung der Lebensqualität der IT erbringen können. Jeder Zustifter oder Sponsor kann jetzt eine bestimmte Dimension der Lebensqualität im hier genannten Sinne wählen, so dass der Stiftungspreis in (naher oder ferner) Zukunft um weitere **Themenpreise** erweitert würde. Zusammen mit dem Wolfgang-Heilmann-Preis für das **Gesamtthema** der humanen Nutzung der IT ergäbe das eine der Aufgabe angemessenere Differenzierung, und mehr Chancen für künftige Preisträger.

- Für das Jahr 2015 kann ich Ihnen aber eröffnen, dass unser Budget für einen weiteren Wolfgang-Heilmann-Preis ausreicht und bereits eine Preisausschreibung unterwegs ist, die thematisch direkt an den diesjährigen Preis anschließt. Wir haben von den Preisträgern gehört, dass es vor allem auf eine Stärkung des **"Privatsphären-Bewusstseins"**, d. h. auf die Einstellung einer individuellen situations- und anwendungsabhängigen Privatheit ankommt. In diesem Sinne haben wir das Thema der Preisausschreibung für 2014/15 wie folgt formuliert:

  Bildung <u>für</u> die E-Society.

  Wir versprechen uns davon viele weiterführende Vorschläge zur Steigerung der Lebensqualität in der E-Society durch bessere Ausbildung und zielgenauere Bildung.

Ziel muss der wahrhaft mündige Mensch sein, der die Möglichkeiten der modernen Informations- und Kommunikations-Technologien zu nutzen weiß, statt sich von ihnen beherrschen zu lassen.

- Das ist die **Kernfrage der E-Society**: Wird der Mensch der Zukunft Herr seiner selbst bleiben oder in einem vernetzten Alltag - nicht nur von seinen Computern und Handys, sondern bald auch von seinen Uhren, Fernsehern, Autos - kurz: allen seinen technischen und sozialen **Netzen** soweit fremdbestimmt, dass er gar keine Kraft und keine Zeit mehr hat, sich auf sich selbst und die Gesellschaft zu besinnen? Werden "Androiden" die neuen Götter sein, die uns weiterführen, oder haben wir noch immer die Möglichkeit, unser Leben, unsere Welt und unsere Gesellschaft im humanen Sinne selber (mit) zu gestalten?

  Mit dieser Frage - meine Damen und Herren - möchte ich meinen Beitrag schließen.

  Ich danke unseren Preisträgern für ihre Vorschläge zur Sicherung der Privatheit - und ich danke Ihnen allen für Ihr Kommen.

Themen Wolfgang-Heilmann-Preis 1998 – 2011/2012

Anlage 1

1998:	Telearbeit (noch INTEGRATA-Preis)
1999:	Tele-Dienste, Telekooperation, Electronic-Commerce (noch INTEGRATA-Preis)
2000:	Telelearning - Aus- und Weiterbildung in einer vernetzten Welt
2001:	Wissensmanagement als Beitrag zur humanen Nutzung der Informationstechnologie
2002:	Telemedizin - die humane Nutzung der Informationstechnologie in der Medizin
2003:	Telemanagement - Führung in virtuellen Organisationen
2004/5:	Verkehrsentlastung durch Telematik und Telekooperation
2005/6:	Humanere Nutzung der Kommunikationstechnologie
2006/7:	Sicherheit, Information und Medienkompetenz
2007/8:	Bürgernahe Anwendungen der Informations- und Kommunikationstechnologien
2009/10:	Mehr Lebensqualität durch Informationstechnologie
2011/12:	Mehr Demokratie durch Informationstechnologie

Mehr Lebensqualität durch Informationstechnologie

10 Thesen zur Verbesserung der Lebensbedingungen durch humane Nutzung der Informations- und Kommunikationstechnologie

Anlage 2

1. **Gesundheit und Wohlbefinden**
   Bewahrung und Wiederherstellung der körperlichen und geistigen **Gesundheit**.

2. **Sicherheit und Freiheit**
   Erhaltung der inneren und äußeren **Sicherheit** unter Wahrung von **Freiheit** und **Würde** des Menschen.

3. **Freizügigkeit und Begegnung**
   Herstellung und Sicherung der **Freizügigkeit** und menschenwürdiger **Verkehrs**verhältnisse zur **persönlichen Begegnung.**

4. **Information und Kommunikation**
   Wiederherstellung des Vertrauens zwischen Kommunikationspartnern durch **verantwortungsbewussten Umgang mit Informationen** und faire **Kommunikation**.

5. **Bildung und Ausbildung**
   Öffnung des Zugangs zu **Bildung** und **Ausbildung** nach Maßgabe des individuellen Leistungsvermögens, zugleich aber Sicherstellung der weltanschaulichen **Neutralität** von Bildungsvermittlung und **Unterhaltung**.

6. **Arbeit und Erwerb**
   Schaffung von **Arbeits**- und **Erwerbs**möglichkeiten, die auf die Menschen zugeschnitten sind, in ausreichender Zahl und Qualität, um das **Gemeinwohl** zu fördern und **Wohlstand für alle** zu erzielen.

7. **Führung und Verantwortung**
   Weiterentwicklung der Informationstechnologie zu einem hilfreichen Instrumentarium für die **Führung von Menschen durch Menschen** in Wirtschaft und Gesellschaft.

8. **Politik und Gesellschaft**
   Förderung der **Partizipation** der Bürger an der öffentlichen Meinungsbildung und Gestaltung des Gemeinwesens im Gleichgewicht von **Freiheit, Recht** und **Gerechtigkeit**.

9. **Natur und Umwelt**
   Schutz der **Natur** und der **Umwelt** vor Raubbau und Zerstörung zugunsten der natürlichen Prozesse und der kommenden Generationen.

10. **Kultur und Religion**
    Überwindung des Sinn- und Zeitnotstands moderner Menschen zugunsten einer würdigen Lebensführung mit **Muße** für **Kultur** und **Religion.**

Komplexität meistern in Zeiten von Big Data – wir liefern Ihnen die passenden Lösungen!

Als ganzheitlicher IT-Dienstleister decken wir mit unseren qualitativ hochwertigen Systems Management Tools, gepaart mit unserem Know-how, ein wichtiges Leistungsspektrum im Bereich komplexer Big-Data- und Cloud-Strategien ab.

automaIT – Befreien Sie Ihr IT-Potenzial!

Mit dem Provisionierungs-Werkzeug automaIT erstellen, konfigurieren, ändern und entfernen Sie IT-Software-Services auf Knopfdruck. So befreien Sie Ihre IT mit automaIT von unnötigen manuellen Abläufen.

inCharge – Management und Nutzungsanalyse von Softwarelizenzen

inCharge ist die Softwarelösung, mit der Sie Ihren Lizenzbestand zentral überwachen, managen, auswerten und unter Einhaltung der Compliance optimieren. So senken Sie Anschaffungs- und Wartungskosten und vermeiden Lizenz-Engpässe.

inspectIT – ... because performance matters!

Die Anwendungsperformance entscheidet häufig über Erfolg und Misserfolg eines Software-Services. Um Ihren Erfolg zu sichern, bietet Ihnen inspectIT alles, was Sie für die Performance-Diagnose und die Überwachung Ihres Browsers, Ihrer .NET- und Ihrer Java-Applikationen benötigen.

loadIT – Kostenlose Lastgenerierung

Das Lasttest-Werkzeug loadIT ermöglicht eine effiziente Durchführung und Auswertung von Lasttests. loadIT deckt bereits in der Entwicklungsphase Performance-Probleme und Flaschenhälse auf und simuliert realistisch den späteren Produktivbetrieb.

Weitere Infos unter:
www.novatec-gmbh.de/produkte

NOVATEC
make IT happen!

GI-Edition Lecture Notes in Informatics

P-1 Gregor Engels, Andreas Oberweis, Albert Zündorf (Hrsg.): Modellierung 2001.

P-2 Mikhail Godlevsky, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications, ISTA'2001.

P-3 Ana M. Moreno, Reind P. van de Riet (Hrsg.): Applications of Natural Lan-guage to Information Systems, NLDB'2001.

P-4 H. Wörn, J. Mühling, C. Vahl, H.-P. Meinzer (Hrsg.): Rechner- und sensorgestützte Chirurgie; Workshop des SFB 414.

P-5 Andy Schürr (Hg.): OMER – Object-Oriented Modeling of Embedded Real-Time Systems.

P-6 Hans-Jürgen Appelrath, Rolf Beyer, Uwe Marquardt, Heinrich C. Mayr, Claudia Steinberger (Hrsg.): Unternehmen Hochschule, UH'2001.

P-7 Andy Evans, Robert France, Ana Moreira, Bernhard Rumpe (Hrsg.): Practical UML-Based Rigorous Development Methods – Countering or Integrating the extremists, pUML'2001.

P-8 Reinhard Keil-Slawik, Johannes Magenheim (Hrsg.): Informatikunterricht und Medienbildung, INFOS'2001.

P-9 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Innovative Anwendungen in Kommunikationsnetzen, 15. DFN Arbeitstagung.

P-10 Mirjam Minor, Steffen Staab (Hrsg.): 1st German Workshop on Experience Management: Sharing Experiences about the Sharing Experience.

P-11 Michael Weber, Frank Kargl (Hrsg.): Mobile Ad-Hoc Netzwerke, WMAN 2002.

P-12 Martin Glinz, Günther Müller-Luschnat (Hrsg.): Modellierung 2002.

P-13 Jan von Knop, Peter Schirmbacher and Viljan Mahni_ (Hrsg.): The Changing Universities – The Role of Technology.

P-14 Robert Tolksdorf, Rainer Eckstein (Hrsg.): XML-Technologien für das Semantic Web – XSW 2002.

P-15 Hans-Bernd Bludau, Andreas Koop (Hrsg.): Mobile Computing in Medicine.

P-16 J. Felix Hampe, Gerhard Schwabe (Hrsg.): Mobile and Collaborative Business 2002.

P-17 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Zukunft der Netze –Die Verletzbarkeit meistern, 16. DFN Arbeitstagung.

P-18 Elmar J. Sinz, Markus Plaha (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2002.

P-19 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund.

P-20 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3. Okt. 2002 in Dortmund (Ergänzungsband).

P-21 Jörg Desel, Mathias Weske (Hrsg.): Promise 2002: Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen.

P-22 Sigrid Schubert, Johannes Magenheim, Peter Hubwieser, Torsten Brinda (Hrsg.): Forschungsbeiträge zur "Didaktik der Informatik" – Theorie, Praxis, Evaluation.

P-23 Thorsten Spitta, Jens Borchers, Harry M. Sneed (Hrsg.): Software Management 2002 – Fortschritt durch Beständigkeit

P-24 Rainer Eckstein, Robert Tolksdorf (Hrsg.): XMIDX 2003 – XML-Technologien für Middleware – Middleware für XML-Anwendungen

P-25 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Commerce – Anwendungen und Perspektiven – 3. Workshop Mobile Commerce, Universität Augsburg, 04.02.2003

P-26 Gerhard Weikum, Harald Schöning, Erhard Rahm (Hrsg.): BTW 2003: Datenbanksysteme für Business, Technologie und Web

P-27 Michael Kroll, Hans-Gerd Lipinski, Kay Melzer (Hrsg.): Mobiles Computing in der Medizin

P-28 Ulrich Reimer, Andreas Abecker, Steffen Staab, Gerd Stumme (Hrsg.): WM 2003: Professionelles Wissensmanagement – Er-fahrungen und Visionen

P-29 Antje Düsterhöft, Bernhard Thalheim (Eds.): NLDB'2003: Natural Language Processing and Information Systems

P-30 Mikhail Godlevsky, Stephen Liddle, Heinrich C. Mayr (Eds.): Information Systems Technology and its Applications

P-31 Arslan Brömme, Christoph Busch (Eds.): BIOSIG 2003: Biometrics and Electronic Signatures

P-32 Peter Hubwieser (Hrsg.): Informatische Fachkonzepte im Unterricht – INFOS 2003

P-33 Andreas Geyer-Schulz, Alfred Taudes (Hrsg.): Informationswirtschaft: Ein Sektor mit Zukunft

P-34 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 1)

P-35 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 2)

P-36 Rüdiger Grimm, Hubert B. Keller, Kai Rannenberg (Hrsg.): Informatik 2003 – Mit Sicherheit Informatik

P-37 Arndt Bode, Jörg Desel, Sabine Rathmayer, Martin Wessner (Hrsg.): DeLFI 2003: e-Learning Fachtagung Informatik

P-38 E.J. Sinz, M. Plaha, P. Neckel (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2003

P-39 Jens Nedon, Sandra Frings, Oliver Göbel (Hrsg.): IT-Incident Management & IT-Forensics – IMF 2003

P-40 Michael Rebstock (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2004

P-41 Uwe Brinkschulte, Jürgen Becker, Dietmar Fey, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle, Thomas Runkler (Edts.): ARCS 2004 – Organic and Pervasive Computing

P-42 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Economy – Transaktionen und Prozesse, Anwendungen und Dienste

P-43 Birgitta König-Ries, Michael Klein, Philipp Obreiter (Hrsg.): Persistance, Scalability, Transactions – Database Mechanisms for Mobile Applications

P-44 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): Security, E-Learning. E-Services

P-45 Bernhard Rumpe, Wofgang Hesse (Hrsg.): Modellierung 2004

P-46 Ulrich Flegel, Michael Meier (Hrsg.): Detection of Intrusions of Malware & Vulnerability Assessment

P-47 Alexander Prosser, Robert Krimmer (Hrsg.): Electronic Voting in Europe – Technology, Law, Politics and Society

P-48 Anatoly Doroshenko, Terry Halpin, Stephen W. Liddle, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications

P-49 G. Schiefer, P. Wagner, M. Morgenstern, U. Rickert (Hrsg.): Integration und Datensicherheit – Anforderungen, Konflikte und Perspektiven

P-50 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 1) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-51 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 2) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-52 Gregor Engels, Silke Seehusen (Hrsg.): DELFI 2004 – Tagungsband der 2. e-Learning Fachtagung Informatik

P-53 Robert Giegerich, Jens Stoye (Hrsg.): German Conference on Bioinformatics – GCB 2004

P-54 Jens Borchers, Ralf Kneuper (Hrsg.): Softwaremanagement 2004 – Outsourcing und Integration

P-55 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): E-Science und Grid Adhoc-Netze Medienintegration

P-56 Fernand Feltz, Andreas Oberweis, Benoit Otjacques (Hrsg.): EMISA 2004 – Informationssysteme im E-Business und E-Government

P-57 Klaus Turowski (Hrsg.): Architekturen, Komponenten, Anwendungen

P-58 Sami Beydeda, Volker Gruhn, Johannes Mayer, Ralf Reussner, Franz Schweiggert (Hrsg.): Testing of Component-Based Systems and Software Quality

P-59 J. Felix Hampe, Franz Lehner, Key Pousttchi, Kai Ranneberg, Klaus Turowski (Hrsg.): Mobile Business – Processes, Platforms, Payments

P-60 Steffen Friedrich (Hrsg.): Unterrichtskonzepte für inforrmatische Bildung

P-61 Paul Müller, Reinhard Gotzhein, Jens B. Schmitt (Hrsg.): Kommunikation in verteilten Systemen

P-62 Federrath, Hannes (Hrsg.): „Sicherheit 2005“ – Sicherheit – Schutz und Zuverlässigkeit

P-63 Roland Kaschek, Heinrich C. Mayr, Stephen Liddle (Hrsg.): Information Systems – Technology and ist Applications

P-64 Peter Liggesmeyer, Klaus Pohl, Michael Goedicke (Hrsg.): Software Engineering 2005

P-65 Gottfried Vossen, Frank Leymann, Peter Lockemann, Wolffried Stucky (Hrsg.): Datenbanksysteme in Business, Technologie und Web

P-66 Jörg M. Haake, Ulrike Lucke, Djamshid Tavangarian (Hrsg.): DeLFI 2005: 3. deutsche e-Learning Fachtagung Informatik

P-67 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 1)

P-68 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 2)

P-69 Robert Hirschfeld, Ryszard Kowalcyk, Andreas Polze, Matthias Weske (Hrsg.): NODe 2005, GSEM 2005

P-70 Klaus Turowski, Johannes-Maria Zaha (Hrsg.): Component-oriented Enterprise Application (COAE 2005)

P-71 Andrew Torda, Stefan Kurz, Matthias Rarey (Hrsg.): German Conference on Bioinformatics 2005

P-72 Klaus P. Jantke, Klaus-Peter Fähnrich, Wolfgang S. Wittig (Hrsg.): Marktplatz Internet: Von e-Learning bis e-Payment

P-73 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): “Heute schon das Morgen sehen“

P-74 Christopher Wolf, Stefan Lucks, Po-Wah Yau (Hrsg.): WEWoRC 2005 – Western European Workshop on Research in Cryptology

P-75 Jörg Desel, Ulrich Frank (Hrsg.): Enterprise Modelling and Information Systems Architecture

P-76 Thomas Kirste, Birgitta König-Riess, Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Informationssysteme – Potentiale, Hindernisse, Einsatz

P-77 Jana Dittmann (Hrsg.): SICHERHEIT 2006

P-78 K.-O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann (Hrsg.): Land- und Ernährungswirtschaft im Wandel

P-79 Bettina Biel, Matthias Book, Volker Gruhn (Hrsg.): Softwareengineering 2006

P-80 Mareike Schoop, Christian Huemer, Michael Rebstock, Martin Bichler (Hrsg.): Service-Oriented Electronic Commerce

P-81 Wolfgang Karl, Jürgen Becker, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle (Hrsg.): ARCS´06

P-82 Heinrich C. Mayr, Ruth Breu (Hrsg.): Modellierung 2006

P-83 Daniel Huson, Oliver Kohlbacher, Andrei Lupas, Kay Nieselt and Andreas Zell (eds.): German Conference on Bioinformatics

P-84 Dimitris Karagiannis, Heinrich C. Mayr, (Hrsg.): Information Systems Technology and its Applications

P-85 Witold Abramowicz, Heinrich C. Mayr, (Hrsg.): Business Information Systems

P-86 Robert Krimmer (Ed.): Electronic Voting 2006

P-87 Max Mühlhäuser, Guido Rößling, Ralf Steinmetz (Hrsg.): DELFI 2006: 4. e-Learning Fachtagung Informatik

P-88 Robert Hirschfeld, Andreas Polze, Ryszard Kowalczyk (Hrsg.): NODe 2006, GSEM 2006

P-90 Joachim Schelp, Robert Winter, Ulrich Frank, Bodo Rieger, Klaus Turowski (Hrsg.): Integration, Informationslogistik und Architektur

P-91 Henrik Stormer, Andreas Meier, Michael Schumacher (Eds.): European Conference on eHealth 2006

P-92 Fernand Feltz, Benoît Otjacques, Andreas Oberweis, Nicolas Poussing (Eds.): AIM 2006

P-93 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 1

P-94 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 2

P-95 Matthias Weske, Markus Nüttgens (Eds.): EMISA 2005: Methoden, Konzepte und Technologien für die Entwicklung von dienstbasierten Informationssystemen

P-96 Saartje Brockmans, Jürgen Jung, York Sure (Eds.): Meta-Modelling and Ontologies

P-97 Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther, Jens Nedon (Eds.): IT-Incident Mangament & IT-Forensics – IMF 2006

P-98 Hans Brandt-Pook, Werner Simonsmeier und Thorsten Spitta (Hrsg.): Beratung in der Softwareentwicklung – Modelle, Methoden, Best Practices

P-99 Andreas Schwill, Carsten Schulte, Marco Thomas (Hrsg.): Didaktik der Informatik

P-100 Peter Forbrig, Günter Siegel, Markus Schneider (Hrsg.): HDI 2006: Hochschuldidaktik der Informatik

P-101 Stefan Böttinger, Ludwig Theuvsen, Susanne Rank, Marlies Morgenstern (Hrsg.): Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten

P-102 Otto Spaniol (Eds.): Mobile Services and Personalized Environments

P-103 Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, Christoph Brochhaus (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007)

P-104 Birgitta König-Ries, Franz Lehner, Rainer Malaka, Can Türker (Hrsg.) MMS 2007: Mobilität und mobile Informationssysteme

P-105 Wolf-Gideon Bleek, Jörg Raasch, Heinz Züllighoven (Hrsg.) Software Engineering 2007

P-106 Wolf-Gideon Bleek, Henning Schwentner, Heinz Züllighoven (Hrsg.) Software Engineering 2007 – Beiträge zu den Workshops

P-107 Heinrich C. Mayr, Dimitris Karagiannis (eds.) Information Systems Technology and its Applications

P-108 Arslan Brömme, Christoph Busch, Detlef Hühnlein (eds.) BIOSIG 2007: Biometrics and Electronic Signatures

P-109 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 1

P-110 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.) INFORMATIK 2007 Informatik trifft Logistik Band 2

P-111 Christian Eibl, Johannes Magenheim, Sigrid Schubert, Martin Wessner (Hrsg.) DeLFI 2007: 5. e-Learning Fachtagung Informatik

P-112 Sigrid Schubert (Hrsg.) Didaktik der Informatik in Theorie und Praxis

P-113 Sören Auer, Christian Bizer, Claudia Müller, Anna V. Zhdanova (Eds.) The Social Semantic Web 2007 Proceedings of the 1st Conference on Social Semantic Web (CSSW)

P-114 Sandra Frings, Oliver Göbel, Detlef Günther, Hardo G. Hase, Jens Nedon, Dirk Schadt, Arslan Brömme (Eds.) IMF2007 IT-incident management & IT-forensics Proceedings of the 3rd International Conference on IT-Incident Management & IT-Forensics

P-115 Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther (Eds.) German conference on bioinformatics GCB 2007

P-116 Witold Abramowicz, Leszek Maciszek (Eds.) Business Process and Services Computing 1st International Working Conference on Business Process and Services Computing BPSC 2007

P-117 Ryszard Kowalczyk (Ed.) Grid service engineering and manegement The 4th International Conference on Grid Service Engineering and Management GSEM 2007

P-118 Andreas Hein, Wilfried Thoben, Hans-Jürgen Appelrath, Peter Jensch (Eds.) European Conference on ehealth 2007

P-119 Manfred Reichert, Stefan Strecker, Klaus Turowski (Eds.) Enterprise Modelling and Information Systems Architectures Concepts and Applications

P-120 Adam Pawlak, Kurt Sandkuhl, Wojciech Cholewa, Leandro Soares Indrusiak (Eds.) Coordination of Collaborative Engineering - State of the Art and Future Challenges

P-121 Korbinian Herrmann, Bernd Bruegge (Hrsg.) Software Engineering 2008 Fachtagung des GI-Fachbereichs Softwaretechnik

P-122 Walid Maalej, Bernd Bruegge (Hrsg.) Software Engineering 2008 - Workshopband Fachtagung des GI-Fachbereichs Softwaretechnik

P-123 Michael H. Breitner, Martin Breunig, Elgar Fleisch, Ley Pousttchi, Klaus Turowski (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Technologien, Prozesse, Marktfähigkeit
Proceedings zur 3. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2008)

P-124 Wolfgang E. Nagel, Rolf Hoffmann, Andreas Koch (Eds.)
9th Workshop on Parallel Systems and Algorithms (PASA)
Workshop of the GI/ITG Speciel Interest Groups PARS and PARVA

P-125 Rolf A.E. Müller, Hans-H. Sundermeier, Ludwig Theuvsen, Stephanie Schütze, Marlies Morgenstern (Hrsg.)
Unternehmens-IT:
Führungsinstrument oder Verwaltungsbürde
Referate der 28. GIL Jahrestagung

P-126 Rainer Gimnich, Uwe Kaiser, Jochen Quante, Andreas Winter (Hrsg.)
10th Workshop Software Reengineering (WSR 2008)

P-127 Thomas Kühne, Wolfgang Reisig, Friedrich Steimann (Hrsg.)
Modellierung 2008

P-128 Ammar Alkassar, Jörg Siekmann (Hrsg.)
Sicherheit 2008
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
2.-4. April 2008
Saarbrücken, Germany

P-129 Wolfgang Hesse, Andreas Oberweis (Eds.)
Sigsand-Europe 2008
Proceedings of the Third AIS SIGSAND European Symposium on Analysis, Design, Use and Societal Impact of Information Systems

P-130 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
1. DFN-Forum Kommunikations-technologien Beiträge der Fachtagung

P-131 Robert Krimmer, Rüdiger Grimm (Eds.)
3rd International Conference on Electronic Voting 2008
Co-organized by Council of Europe, Gesellschaft für Informatik and E-Voting. CC

P-132 Silke Seehusen, Ulrike Lucke, Stefan Fischer (Hrsg.)
DeLFI 2008:
Die 6. e-Learning Fachtagung Informatik

P-133 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 1

P-134 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 2

P-135 Torsten Brinda, Michael Fothe, Peter Hubwieser, Kirsten Schlüter (Hrsg.)
Didaktik der Informatik –
Aktuelle Forschungsergebnisse

P-136 Andreas Beyer, Michael Schroeder (Eds.)
German Conference on Bioinformatics GCB 2008

P-137 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2008: Biometrics and Electronic Signatures

P-138 Barbara Dinter, Robert Winter, Peter Chamoni, Norbert Gronau, Klaus Turowski (Hrsg.)
Synergien durch Integration und Informationslogistik
Proceedings zur DW2008

P-139 Georg Herzwurm, Martin Mikusz (Hrsg.)
Industrialisierung des Software-Managements
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik

P-140 Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, Dirk Schadt (Eds.)
IMF 2008 - IT Incident Management & IT Forensics

P-141 Peter Loos, Markus Nüttgens, Klaus Turowski, Dirk Werth (Hrsg.)
Modellierung betrieblicher Informations-systeme (MobIS 2008)
Modellierung zwischen SOA und Compliance Management

P-142 R. Bill, P. Korduan, L. Theuvsen, M. Morgenstern (Hrsg.)
Anforderungen an die Agrarinformatik durch Globalisierung und Klimaveränderung

P-143 Peter Liggesmeyer, Gregor Engels, Jürgen Münch, Jörg Dörr, Norman Riegel (Hrsg.)
Software Engineering 2009
Fachtagung des GI-Fachbereichs Softwaretechnik

P-144 Johann-Christoph Freytag, Thomas Ruf, Wolfgang Lehner, Gottfried Vossen (Hrsg.)
Datenbanksysteme in Business, Technologie und Web (BTW)

P-145 Knut Hinkelmann, Holger Wache (Eds.)
WM2009: 5th Conference on Professional Knowledge Management

P-146 Markus Bick, Martin Breunig, Hagen Höpfner (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Entwicklung, Implementierung und Anwendung
4. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2009)

P-147 Witold Abramowicz, Leszek Maciaszek, Ryszard Kowalczyk, Andreas Speck (Eds.)
Business Process, Services Computing and Intelligent Service Management
BPSC 2009 · ISM 2009 · YRW-MBP 2009

P-148 Christian Erfurth, Gerald Eichler, Volkmar Schau (Eds.)
9th International Conference on Innovative Internet Community Systems
I^{2}CS 2009

P-149 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
2. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-150 Jürgen Münch, Peter Liggesmeyer (Hrsg.)
Software Engineering
2009 - Workshopband

P-151 Armin Heinzl, Peter Dadam, Stefan Kirn, Peter Lockemann (Eds.)
PRIMIUM
Process Innovation for Enterprise Software

P-152 Jan Mendling, Stefanie Rinderle-Ma, Werner Esswein (Eds.)
Enterprise Modelling and Information Systems Architectures
Proceedings of the 3rd Int'l Workshop EMISA 2009

P-153 Andreas Schwill, Nicolas Apostolopoulos (Hrsg.)
Lernen im Digitalen Zeitalter
DeLFI 2009 – Die 7. E-Learning Fachtagung Informatik

P-154 Stefan Fischer, Erik Maehle Rüdiger Reischuk (Hrsg.)
INFORMATIK 2009
Im Focus das Leben

P-155 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2009:
Biometrics and Electronic Signatures
Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-156 Bernhard Koerber (Hrsg.)
Zukunft braucht Herkunft
25 Jahre »INFOS – Informatik und Schule«

P-157 Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, Peter Stadler (Eds.)
German Conference on Bioinformatics 2009

P-158 W. Claupein, L. Theuvsen, A. Kämpf, M. Morgenstern (Hrsg.)
Precision Agriculture Reloaded – Informationsgestützte Landwirtschaft

P-159 Gregor Engels, Markus Luckey, Wilhelm Schäfer (Hrsg.)
Software Engineering 2010

P-160 Gregor Engels, Markus Luckey, Alexander Pretschner, Ralf Reussner (Hrsg.)
Software Engineering 2010 – Workshopband
(inkl. Doktorandensymposium)

P-161 Gregor Engels, Dimitris Karagiannis Heinrich C. Mayr (Hrsg.)
Modellierung 2010

P-162 Maria A. Wimmer, Uwe Brinkhoff, Siegfried Kaiser, Dagmar Lück-Schneider, Erich Schweighofer, Andreas Wiebe (Hrsg.)
Vernetzte IT für einen effektiven Staat
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2010

P-163 Markus Bick, Stefan Eulgem, Elgar Fleisch, J. Felix Hampe, Birgitta König-Ries, Franz Lehner, Key Pousttchi, Kai Rannenberg (Hrsg.)
Mobile und Ubiquitäre Informationssysteme
Technologien, Anwendungen und Dienste zur Unterstützung von mobiler Kollaboration

P-164 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2010: Biometrics and Electronic Signatures Proceedings of the Special Interest Group on Biometrics and Electronic Signatures

P-165 Gerald Eichler, Peter Kropf, Ulrike Lechner, Phayung Meesad, Herwig Unger (Eds.)
10th International Conference on Innovative Internet Community Systems (I^2CS) – Jubilee Edition 2010 –

P-166 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
3. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-167 Robert Krimmer, Rüdiger Grimm (Eds.)
4th International Conference on Electronic Voting 2010
co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-168 Ira Diethelm, Christina Dörge, Claudia Hildebrandt, Carsten Schulte (Hrsg.)
Didaktik der Informatik
Möglichkeiten empirischer Forschungsmethoden und Perspektiven der Fachdidaktik

P-169 Michael Kerres, Nadine Ojstersek Ulrik Schroeder, Ulrich Hoppe (Hrsg.)
DeLFI 2010 - 8. Tagung der Fachgruppe E-Learning der Gesellschaft für Informatik e.V.

P-170 Felix C. Freiling (Hrsg.)
Sicherheit 2010
Sicherheit, Schutz und Zuverlässigkeit

P-171 Werner Esswein, Klaus Turowski, Martin Juhrisch (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2010)
Modellgestütztes Management

P-172 Stefan Klink, Agnes Koschmider Marco Mevius, Andreas Oberweis (Hrsg.)
EMISA 2010
Einflussfaktoren auf die Entwicklung flexibler, integrierter Informationssysteme
Beiträge des Workshops der GI-Fachgruppe EMISA (Entwicklungsmethoden für Informationssysteme und deren Anwendung)

P-173 Dietmar Schomburg, Andreas Grote (Eds.)
German Conference on Bioinformatics 2010

P-174 Arslan Brömme, Torsten Eymann, Detlef Hühnlein, Heiko Roßnagel, Paul Schmücker (Hrsg.)
perspeGKtive 2010
Workshop „Innovative und sichere Informationstechnologie für das Gesundheitswesen von morgen“

P-175 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 1

P-176 Klaus-Peter Fähnrich, Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für die Informatik
Band 2

P-177 Witold Abramowicz, Rainer Alt, Klaus-Peter Fähnrich, Bogdan Franczyk, Leszek A. Maciaszek (Eds.)
INFORMATIK 2010
Business Process and Service Science – Proceedings of ISSS and BPSC

P-178 Wolfram Pietsch, Benedikt Krams (Hrsg.)
Vom Projekt zum Produkt
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschafts-informatik (WI-MAW), Aachen, 2010

P-179 Stefan Gruner, Bernhard Rumpe (Eds.)
FM+AM`2010
Second International Workshop on Formal Methods and Agile Methods

P-180 Theo Härder, Wolfgang Lehner, Bernhard Mitschang, Harald Schöning, Holger Schwarz (Hrsg.)
Datenbanksysteme für Business, Technologie und Web (BTW)
14. Fachtagung des GI-Fachbereichs „Datenbanken und Informationssysteme“ (DBIS)

P-181 Michael Clasen, Otto Schätzel, Brigitte Theuvsen (Hrsg.)
Qualität und Effizienz durch informationsgestützte Landwirtschaft, Fokus: Moderne Weinwirtschaft

P-182 Ronald Maier (Hrsg.)
6th Conference on Professional Knowledge Management
From Knowledge to Action

P-183 Ralf Reussner, Matthias Grund, Andreas Oberweis, Walter Tichy (Hrsg.)
Software Engineering 2011
Fachtagung des GI-Fachbereichs Softwaretechnik

P-184 Ralf Reussner, Alexander Pretschner, Stefan Jähnichen (Hrsg.)
Software Engineering 2011
Workshopband
(inkl. Doktorandensymposium)

P-185 Hagen Höpfner, Günther Specht, Thomas Ritz, Christian Bunse (Hrsg.)
MMS 2011: Mobile und ubiquitäre Informationssysteme Proceedings zur 6. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2011)

P-186 Gerald Eichler, Axel Küpper, Volkmar Schau, Hacène Fouchal, Herwig Unger (Eds.)
11th International Conference on Innovative Internet Community Systems (I^2CS)

P-187 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
4. DFN-Forum Kommunikations-technologien, Beiträge der Fachtagung 20. Juni bis 21. Juni 2011 Bonn

P-188 Holger Rohland, Andrea Kienle, Steffen Friedrich (Hrsg.)
DeLFI 2011 – Die 9. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. 5.–8. September 2011, Dresden

P-189 Thomas, Marco (Hrsg.)
Informatik in Bildung und Beruf
INFOS 2011
14. GI-Fachtagung Informatik und Schule

P-190 Markus Nüttgens, Oliver Thomas, Barbara Weber (Eds.)
Enterprise Modelling and Information Systems Architectures (EMISA 2011)

P-191 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2011
International Conference of the Biometrics Special Interest Group

P-192 Hans-Ulrich Heiß, Peter Pepper, Holger Schlingloff, Jörg Schneider (Hrsg.)
INFORMATIK 2011
Informatik schafft Communities

P-193 Wolfgang Lehner, Gunther Piller (Hrsg.)
IMDM 2011

P-194 M. Clasen, G. Fröhlich, H. Bernhardt, K. Hildebrand, B. Theuvsen (Hrsg.)
Informationstechnologie für eine nachhaltige Landbewirtschaftung
Fokus Forstwirtschaft

P-195 Neeraj Suri, Michael Waidner (Hrsg.)
Sicherheit 2012
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 6. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)

P-196 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2012
Proceedings of the 11th International Conference of the Biometrics Special Interest Group

P-197 Jörn von Lucke, Christian P. Geiger, Siegfried Kaiser, Erich Schweighofer, Maria A. Wimmer (Hrsg.)
Auf dem Weg zu einer offenen, smarten und vernetzten Verwaltungskultur
Gemeinsame Fachtagung Verwaltungsinformatik (FTVI) und Fachtagung Rechtsinformatik (FTRI) 2012

P-198 Stefan Jähnichen, Axel Küpper, Sahin Albayrak (Hrsg.)
Software Engineering 2012
Fachtagung des GI-Fachbereichs Softwaretechnik

P-199 Stefan Jähnichen, Bernhard Rumpe, Holger Schlingloff (Hrsg.)
Software Engineering 2012
Workshopband

P-200 Gero Mühl, Jan Richling, Andreas Herkersdorf (Hrsg.)
ARCS 2012 Workshops

P-201 Elmar J. Sinz Andy Schürr (Hrsg.)
Modellierung 2012

P-202 Andrea Back, Markus Bick, Martin Breunig, Key Pousttchi, Frédéric Thiesse (Hrsg.)
MMS 2012:Mobile und Ubiquitäre Informationssysteme

P-203 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.)
5. DFN-Forum Kommunikations-technologien
Beiträge der Fachtagung

P-204 Gerald Eichler, Leendert W. M. Wienhofen, Anders Kofod-Petersen, Herwig Unger (Eds.)
12th International Conference on Innovative Internet Community Systems (I2CS 2012)

P-205 Manuel J. Kripp, Melanie Volkamer, Rüdiger Grimm (Eds.)
5th International Conference on Electronic Voting 2012 (EVOTE2012)
Co-organized by the Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-206 Stefanie Rinderle-Ma, Mathias Weske (Hrsg.)
EMISA 2012
Der Mensch im Zentrum der Modellierung

P-207 Jörg Desel, Jörg M. Haake, Christian Spannagel (Hrsg.)
DeLFI 2012: Die 10. e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V.
24.–26. September 2012

P-208 Ursula Goltz, Marcus Magnor, Hans-Jürgen Appelrath, Herbert Matthies, Wolf-Tilo Balke, Lars Wolf (Hrsg.) INFORMATIK 2012

P-209 Hans Brandt-Pook, André Fleer, Thorsten Spitta, Malte Wattenberg (Hrsg.) Nachhaltiges Software Management

P-210 Erhard Plödereder, Peter Dencker, Herbert Klenk, Hubert B. Keller, Silke Spitzer (Hrsg.) Automotive – Safety & Security 2012 Sicherheit und Zuverlässigkeit für automobile Informationstechnik

P-211 M. Clasen, K. C. Kersebaum, A. Meyer-Aurich, B. Theuvsen (Hrsg.) Massendatenmanagement in der Agrar- und Ernährungswirtschaft Erhebung - Verarbeitung - Nutzung Referate der 33. GIL-Jahrestagung 20. – 21. Februar 2013, Potsdam

P-212 Arslan Brömme, Christoph Busch (Eds.) BIOSIG 2013 Proceedings of the 12th International Conference of the Biometrics Special Interest Group 04.–06. September 2013 Darmstadt, Germany

P-213 Stefan Kowalewski, Bernhard Rumpe (Hrsg.) Software Engineering 2013 Fachtagung des GI-Fachbereichs Softwaretechnik

P-214 Volker Markl, Gunter Saake, Kai-Uwe Sattler, Gregor Hackenbroich, Bernhard Mit schang, Theo Härder, Veit Köppen (Hrsg.) Datenbanksysteme für Business, Technologie und Web (BTW) 2013 13. – 15. März 2013, Magdeburg

P-215 Stefan Wagner, Horst Lichter (Hrsg.) Software Engineering 2013 Workshopband (inkl. Doktorandensymposium) 26. Februar – 1. März 2013, Aachen

P-216 Gunter Saake, Andreas Henrich, Wolfgang Lehner, Thomas Neumann, Veit Köppen (Hrsg.) Datenbanksysteme für Business, Technologie und Web (BTW) 2013 – Workshopband 11. – 12. März 2013, Magdeburg

P-217 Paul Müller, Bernhard Neumair, Helmut Reiser, Gabi Dreo Rodosek (Hrsg.) 6. DFN-Forum Kommunikations- technologien Beiträge der Fachtagung 03.–04. Juni 2013, Erlangen

P-218 Andreas Breiter, Christoph Rensing (Hrsg.) DeLFI 2013: Die 11 e-Learning Fachtagung Informatik der Gesellschaft für Informatik e.V. (GI) 8. – 11. September 2013, Bremen

P-219 Norbert Breier, Peer Stechert, Thomas Wilke (Hrsg.) Informatik erweitert Horizonte INFOS 2013 15. GI-Fachtagung Informatik und Schule 26. – 28. September 2013

P-220 Matthias Horbach (Hrsg.) INFORMATIK 2013 Informatik angepasst an Mensch, Organisation und Umwelt 16. – 20. September 2013, Koblenz

P-221 Maria A. Wimmer, Marijn Janssen, Ann Macintosh, Hans Jochen Scholl, Efthimios Tambouris (Eds.) Electronic Government and Electronic Participation Joint Proceedings of Ongoing Research of IFIP EGOV and IFIP ePart 2013 16. – 19. September 2013, Koblenz

P-222 Reinhard Jung, Manfred Reichert (Eds.) Enterprise Modelling and Information Systems Architectures (EMISA 2013) St. Gallen, Switzerland September 5. – 6. 2013

P-223 Detlef Hühnlein, Heiko Roßnagel (Hrsg.) Open Identity Summit 2013 10. – 11. September 2013 Kloster Banz, Germany

P-224 Eckhart Hanser, Martin Mikusz, Masud Fazal-Baqaie (Hrsg.) Vorgehensmodelle 2013 Vorgehensmodelle – Anspruch und Wirklichkeit 20. Tagung der Fachgruppe Vorgehensmodelle im Fachgebiet Wirtschaftsinformatik (WI-VM) der Gesellschaft für Informatik e.V. Lörrach, 2013

P-225 Hans-Georg Fill, Dimitris Karagiannis, Ulrich Reimer (Hrsg.) Modellierung 2014 19. – 21. März 2014, Wien

P-226 M. Clasen, M. Hamer, S. Lehnert, B. Petersen, B. Theuvsen (Hrsg.) IT-Standards in der Agrar- und Ernährungswirtschaft Fokus: Risiko- und Krisenmanagement Referate der 34. GIL-Jahrestagung 24. – 25. Februar 2014, Bonn

P-227 Wilhelm Hasselbring,
Nils Christian Ehmke (Hrsg.)
Software Engineering 2014
Fachtagung des GI-Fachbereichs
Softwaretechnik
25. – 28. Februar 2014
Kiel, Deutschland

P-228 Stefan Katzenbeisser, Volkmar Lotz,
Edgar Weippl (Hrsg.)
Sicherheit 2014
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 7. Jahrestagung des
Fachbereichs Sicherheit der
Gesellschaft für Informatik e.V. (GI)
19.–21. März 2014, Wien

P-231 Paul Müller, Bernhard Neumair,
Helmut Reiser, Gabi Dreo Rodosek
(Hrsg.)
7. DFN-Forum
Kommunikationstechnologien
16.–17. Juni 2014
Fulda

P-232 E. Plödereder, L. Grunske, E. Schneider,
D. Ull (Hrsg.)
INFORMATIK 2014
Big Data – Komplexität meistern
22. – 26. September 2014
Stuttgart

The titles can be purchased at:

Köllen Druck + Verlag GmbH
Ernst-Robert-Curtius-Str. 14 · D-53117 Bonn
Fax: +49 (0)228/9898222
E-Mail: druckverlag@koellen.de

Gesellschaft für Informatik e.V. (GI)

publishes this series in order to make available to a broad public recent findings in informatics (i.e. computer science and information systems), to document conferences that are organized in cooperation with GI and to publish the annual GI Award dissertation.

Broken down into
• seminars
• proceedings
• dissertations
• thematics
current topics are dealt with from the vantage point of research and development, teaching and further training in theory and practice. The Editorial Committee uses an intensive review process in order to ensure high quality contributions.

The volumes are published in German or English.

Information: http://www.gi.de/service/publikationen/lni/

ISSN 1617-5468
ISBN 978-3-88579-626-8

Dieser Band enthält Beiträge der 44. Jahrestagung der Gesellschaft für Informatik (GI), der INFORMATIK 2014. Die Konferenz steht unter dem Motto „Big Data – Komplexität meistern". Zwei Thementage, rund 40 Workshops und Tutorials sowie vier assoziierte Konferenzen bilden das Programm dieser Jahrestagung. Die Beiträge erörtern eine breite Palette praxis- und forschungsorientierter Fragestellungen mit Schwerpunkten rund um das Thema „Big Data" sowie um komplexe Anwendungen der IT-Branche..