Einsatz von Dataspaces für die inkrementelle Informationsintegration in der Medizin

S.H.R. Wurst [1], G. Lamla [1], F. Prasser [1,2], A. Kemper[2], K.A. Kuhn[1]

[1] Lehrstuhl für medizinische Informatik
Technische Universität München
Klinikum rechts der Isar der TU München
Ismaninger Str. 22
D-81675 München
sebastian.wurst@tum.de

[2] Lehrstuhl für Datenbanksysteme
Technische Universität München
Boltzmannstr. 3
D-85748 Garching bei München

Abstract: Nach der Entschlüsselung des menschlichen Genoms ergeben sich für die medizinische Forschung weitreichende Möglichkeiten, die auf Informationsintegration und der Zusammenführung phänotypischer und genotypischer Daten beruhen. Um die komplexen Integrationsanforderungen der medizinischen Domäne handhaben zu können, wird ein inkrementeller Ansatz vorgeschlagen, der auf dem Dataspace Paradigma beruht. Anwendungsdomäne ist die Integration klinischer Datenbanken mit Forschungsdatenbanken und Biobanken.

1 Einleitung und Fragestellung

Die Bedeutung der Informationsintegration hat in der postgenomischen Ära massiv zugenommen, da phänotypische und genotypische Daten zusammengeführt werden können und die translationale Forschung unterstützt werden muss [Al08, Ku08]. Translation „from bench to bedside to community and back“ beschreibt in diesem Kontext den Kreislauf von genomischer, molekularer und klinischer Datensammlung hin zu individualisierter Behandlung mit Verlaufsbeobachtung und Evaluation bis zur erneuten Hypothesengenerierung. Translationale Medizin eröffnet neue Einblicke in Krankheitsmechanismen und unterstützt so die Ermittlung persönlicher Risiken und die Festlegung personalisierter Therapien [Al08, Ku08]. Informationsintegration spielt vor diesem Hintergrund eine bedeutende Rolle in Forschungsprogrammen weltweit [Na06, Hi07].

In diesem Artikel fokussieren wir auf die Integration von klinischen Datenbanken mit Forschungsdatenbanken und Biobanken. Zu den Zielen gehören die Zusammenführung unterschiedlicher Forschungsdatenbanken, die Übernahme klinischer Daten in Studiendatenbanken, die möglichst gemeinsame und einheitliche Erfassung von

Datenfür Forschungs- und Versorgungszwecke, die Integration von Biobanken und die Verwendung klinischer Datenbanken zur Rekrutierungsunterstützung für klinische Studien. [Hi07, Na06]

Verteilung, Fragmentierung und **semantische Heterogenität** sind für die medizinische Informationsverarbeitung zentrale Herausforderungen. Die Fragmentierung unseres Gesundheitswesens im niedergelassenen Bereich, ambulante und stationäre Behandlung im Krankenhaus, Fach- und Abteilungsstrukturen, sowie die Trennung zwischen Grundlagenforschung, klinischer Forschung, Versorgungsforschung und klinischer Versorgung führt zu verteilter, semantisch heterogener Datenhaltung. Häufig fehlt eine Identifikationsmöglichkeit von Patienten zwischen Einrichtungen, oft auch innerhalb der Einrichtungen selbst. Die verteilten Daten sind unterschiedlich strukturiert und eine terminologische Kontrolle fehlt häufig. Soweit in der Krankenversorgung strukturierte Formulare zum Einsatz kommen, sind diese oft zwischen verschiedenen Abteilungen nicht abgeglichen, und es können gleiche Attribute in verschiedenen Dokumenten und Tabellen vorliegen. Eine forschungsbezogene Dokumentation erfordert einen höheren Strukturierungsgrad und standardisierte Terminologien, aber auch die im Kontext von Studien strukturiert erhobenen Daten können derzeit i.a. nicht semantisch zusammengeführt werden. Auf der Ebene der Informationssysteme selbst besteht ebenfalls eine erhebliche Heterogenität: Die Tatsache, dass eine Systemeinführung sehr kosten- und zeitaufwendig sein kann, führt einerseits zur beharrlichen Weiterverwendung von Legacy-Systemen, andererseits zu ad-hoc Lösungen ohne ausreichende Integrationskonzepte.

Medizinische Informationsverarbeitung findet zudem in einem hoch komplexen System statt, in dem **dynamische Interaktionen** zwischen Technologie, Menschen in sehr verschiedenen Rollen und komplexen Organisationsstrukturen ablaufen. Die Technologieeinführung ist ein dynamischer Prozess, der die Arbeitsabläufe und damit auch die Anforderungen an die Informationsverarbeitung permanent verändert [WB05]. Hinzu kommen ständige Veränderungen durch neue diagnostische und therapeutische Verfahren sowie stark im Fluss befindliche Rahmenbedingungen der Finanzierung im Gesundheitswesen. Die Softwareentwicklung in der Medizin sollte aufgrund dieser Dynamik die Entwicklungszyklen verkürzen und den Anwendungsentwickler möglichst nahe mit dem Endanwender zusammenbringen. Sequentielle Vorgehensmodelle für das Softwareengineering sind hierfür schlecht geeignet. Insbesondere in einer Umgebung mit sich verändernden Anforderungen und sehr großen und sehr komplexen Informationssystemen sollte die Softwareentwicklung iterativ und hochpartizipatorisch sein, um eine maximale Anpassungsfähigkeit an Versorgungs- und Forschungsprozesse zu gewährleisten. [LK04]

Von zentraler Bedeutung sind in der Medizin der **Datenschutz** und die Verwaltung von **Zugriffsrechten**. Der Zugriff auf Patientendaten ist im Versorgungskontext nur gestattet, wenn ein Behandlungszusammenhang besteht. Für die Forschung ist die Basis die Einverständniserklärung des Patienten (Informed Consent). Ein Integrationsansatz muss die teilweise sehr komplexe Rechte- und Rollensituation abbilden, Konzepte der De- und ggf. Re-Identifikation umfassen sowie die Intellectual Property Rechte der Forscher beachten.

2 Methode

Um die genannten Anforderungen zu bewältigen, wurden unterschiedliche Informationsintegrationsansätze betrachtet. Es gibt etablierte Architekturkonzepte wie Globales Schema, Multidatenbanksysteme, föderierte Datenbanksysteme, mediator-basierte Datenbanksysteme oder Peer-Datenmanagement-Systeme, die auch zur Informationsintegration in der biomedizinischen Forschung eingesetzt werden. Abseits von relationalen Datenbankschemata existieren semi-strukturierte Ansätze, insbesondere für den Austausch von genetischen Daten. Für die Modellierung von semantischen Zusammenhängen werden Ontologien eingesetzt [Lo07].

Bei der praktischen Informationsintegration können etablierte Architekturen jedoch nicht immer alle Anforderungen erfüllen, da Daten häufig über nur lose verknüpfte Datenquellen verteilt sind und sich auch außerhalb von DBMS befinden [FHM05]. Für die Speicherung und Verarbeitung von strukturierten bzw. unstrukturierten Daten können unterschiedliche Paradigmen identifiziert werden. Wesentlich ist die Unterscheidung zwischen dem Schema First (SFA) und dem No Schema Approach (NSA). Bei einem SFA erfolgt vorab ein Integrationsschritt für die Schemata der Komponentensysteme, was eine Erstellung komplexer Mappings und konsolidierter Schemata erfordert. Abfragen auf einer SFA Integration besitzen eine klar definierte Semantik und eindeutige Ergebnisse, die Erstellung einer SFA ist jedoch aufwendig und teuer, und Teile der integrierten Daten werden unter Umständen selten oder nie verwendet. Bei einer NSA Integration werden alle Datenquellen direkt eingebunden, indem bspw. eine Schlüsselwortsuche darauf zur Verfügung gestellt wird. Eine semantische Integration der Datenquellen ist nicht notwendig, Abfragen besitzen jedoch auch keine präzise Semantik. Die o. g. Architekturen folgen einem SFA, während ein NSA bspw. im Bereich von Internetsuchmaschinen angewandt wird [Va07].

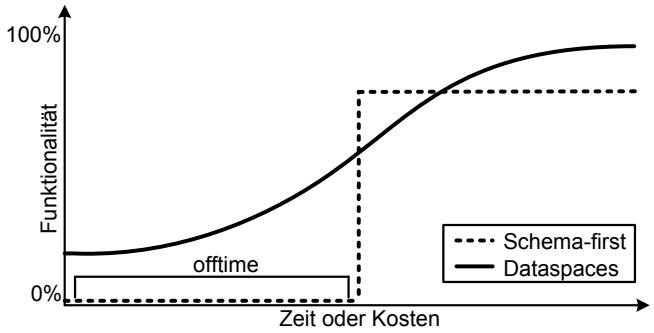

Abbildung 1: Pay-as-you-go vs. Schema-First [Di07]

Die Idee des Dataspace Ansatzes besteht darin, die Vorzüge eines NSA mit einem SFA zu kombinieren, ohne die Vorzüge des SFA aufzugeben. Dabei soll von Anfang an der Zugriff auf alle Daten unterstützt, jedoch keine volle Kontrolle über die Daten ausgeübt werden. Im Gegensatz zu herkömmlichen Datenintegrationsansätzen stellt eine DSSP den arbeitsintensiven Aspekt der Datenintegration zurück, bis er absolut notwendig ist.

Zusätzlicher Aufwand für eine engere Integration erfolgt inkrementell und bedarfsorientiert („pay-as-you-go“). Dadurch sind keine langen Vorlaufzeiten notwendig, erste Services können sofort angeboten, mit der Zeit ausgebaut und durch werkzeugunterstütztes Herstellen semantischer Beziehungen verbessert werden [FHM05].

Der Dataspaces Ansatz eignet sich somit zur Erreichung einer mit dem SFA vergleichbaren Integrationstiefe, ohne jedoch die für die SFA typische lange Vorlaufzeit zu erfordern. Durch die bedarfsorientierte Integration werden nicht nur unnötige Integrationsschritte eingespart, es können auch im Nachhinein Anpassungen vorgenommen werden, wenn neue Anforderungen auftreten.

3 Ergebnis

Durch wechselnde Anforderungen und sich verändernde Rahmenbedingungen verliert eine statische Integration mit der Zeit an Wert und muss regelmäßig überarbeitet und verändert werden. Dieselben Gründe, die für eine agile Softwareentwicklung in der Medizin sprechen, sprechen auch für einen evolutionären Ansatz bei der Informationsintegration. Der Dataspaces Ansatz erlaubt es, schneller und zielgerichteter auf diese Entwicklungen einzugehen als ein SFA.

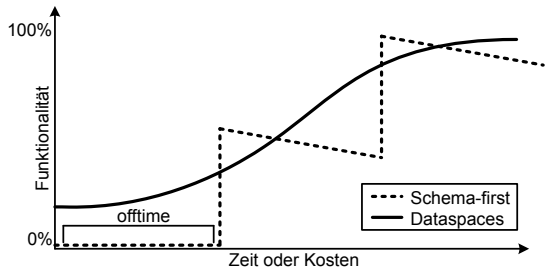

Abbildung 2: Evolutionäre Informationsintegration in der Medizin

Zur Realisierung einer DSSP in der Medizin benötigt man ein Kernschema zur globalen Identifikation von Patienten und ein generisches Datenmodell, das die Integration von strukturierten, semi-strukturierten und unstrukturierten Daten ermöglicht. Dabei ist die Verfügbarkeit der Daten aus den verteilten Datenquellen zunächst zu gewährleisten. Data Mapping erfolgt durch Zusammenführen von Patienten aus unterschiedlichen Datenquellen. Schema Mapping erfolgt Schritt für Schritt durch Zusammenführen von Schemata bzw. Attributen durch Abbildung, Konvertierung von beispielsweise Einheit und Sprache, und durch Einbindung von Terminologien und Versionierung. Datenintegration erfolgt nach dem „pay-as-you-go“ Prinzip, aufbauend auf der Koexistenz von lose integrierten heterogenen Datenquellen und stark integrierten Kerndaten.

Für die Umsetzung bietet sich wegen der Anforderung, mit möglichst geringem Aufwand Änderungen durchführen zu können, eine service-orientierte Architektur an [Wu07]. Die Kerneigenschaften wie Technologieunabhängigkeit, einfache Wiederverwendung von Komponenten und entkoppelte Entwicklung ermöglichen es, auch auf Ebene der Softwareentwicklung, mit den sich ändernden Rahmenbedingungen und Anforderungen umzugehen, und damit ein evolutionäres Modell zur Softwareentwicklung umzusetzen. Generell werden service-orientierte Architekturen als gute Entwicklungsbasis für die medizinische Anwendungsdomäne angesehen [Ku07].

Die DSSP wurde nach den Maßgaben als leicht-gewichtige Lösung zur Einbindung der verfügbaren Datenquellen entwickelt. Die Softwarearchitektur ist in drei Schichten untergliedert. Auf der Serviceebene findet die Kommunikation mit Datendiensten der Datenquellen statt. Diese sind als Wrapper realisiert, die in der Datenquelle enthaltene Daten in ein generisches Schema transformieren und der DSSP zur Verfügung stellen. In internen Diensten der DSSP werden darüber hinaus die erarbeiteten Informationen zur semantischen Integration der Daten persistent gehalten. Die darüber liegende Prozessschicht kontrolliert die Interaktionen zwischen den Services und bietet Schnittstellen an, die von dafür entwickelten Anwendungen aufgerufen werden können. Diese Anwendungen teilen sich in zwei Gruppen: Zum Einen Anwendungen für Verwaltung und inkrementelle Durchführung der semantischen Integration, und zum Anderen Anwendungen, die auf die DSSP zugreifen, um die integrierten Daten zu verwenden. Diesen werden auch Funktionen zur Verfügung gestellt, um eine kontext-erhaltende Oberflächenintegration nach einem Single Sign On Prinzip zu realisieren. Dafür wird ein Dienst im Sinne einer Trusted Third Party zur Verfügung gestellt, in dem der Anwender Authentifizierungsinformationen für die entsprechenden Komponentensysteme hinterlegen kann. Durch die dezentrale Auswertung der Berechtigungen bleiben Autonomie und insbesondere Berechtigungskonzept der Datenquellen unangetastet [Wi05].

Zum Nachweis der Machbarkeit des Ansatzes wurde eine prototypische Implementierung der beschriebenen Konzepte und Komponenten am Klinikum rechts der Isar in München durchgeführt. Dabei wurden das am Klinikum im Einsatz befindliche Patientendatenmanagementsystem SAP IS-H, das integrierte Klinische Arbeitsplatzsystem Siemens i.s.h.med, das Studiendatenmanagementsystem/CDMS Infermed Macro, sowie die Biobank des Klinikums eingebunden. Die DSSP wurde mit open-source Technologien realisiert (MySQL, Hibernate, Java, JSF).

4 Diskussion

Die konzeptionelle Idee einer DSSP in der klinischen Forschung besteht darin, einen Dataspace über alle für die Forschung relevanten Daten zu spannen und damit die Integration schrittweise durchzuführen, um eine kontinuierliche Verbesserung der Integration in den vorrangig benötigten Bereichen zu erreichen. Während in der Informationsintegration die Aspekte Verteilung, Autonomie und Heterogenität Beachtung erhalten, fokussiert der Dataspace Ansatz primär auf die Heterogenität. Das Ziel der beschriebenen Entwicklung besteht daher auch darin, Lösungen für die anderen

beiden Aspekte bereit zu stellen. Große Einschränkungen entstehen einerseits durch gesetzliche und regulatorische Beschränkungen in der klinischen Forschung, anderseits durch Legacysysteme, deren Autonomie nicht überwunden werden kann. Neben dem primären Ziel der Informationsintegration ist zur Nachvollziehbarkeit der Information und zur Datenherkunft und -entstehung auch eine Oberflächenintegration in die Komponentensysteme notwendig, um die Daten im Originalzusammenhang betrachten zu können. Mit der Entwicklung des beschriebenen Systems wurden erste Schritte unternommen, die genannten Anforderungen umzusetzen.

Bisherige Entwicklungen im Bereich Dataspaces haben einen anderen Fokus und auch andere Anforderungen. iMemex [Di06] realisiert eine Personal Information Management (PIM) Lösung in Form einer DSSP. PayGo [Ma07] und DBLife [De07] setzen die Integration von Ressourcen aus dem Web um. Informationsintegrationsansätze in der Medizin haben zwar einen ähnlichen Fokus aber verwenden andere Ansätze. caBIG bzw. caGrid [ca07] fokussiert auf die interinstitutionelle Zusammenarbeit in der Krebsforschung. Das Hauptziel ist die Standardisierung und die Entwicklung von Werkzeugen und Datenstandards. Die Heterogenität von Datenmodellen wird dabei nicht adressiert und Abfragen gegen diese werden nur unterstützt wenn diese gemäß Richtlinien für Datenstandards erfolgen. MIMM [Hi07] realisiert ein föderiertes Datenbanksystem und der Ontology-based Mapping & Unification [Ma06] Ansatz eine Art mediator-basiertes Datenbanksystem.

5 Ausblick

Das System wird derzeit angepasst und erweitert, um in verschiedenen Kliniken am Klinikum rechts der Isar und unter Wahrung der spezifischen Klinikanforderungen produktiv eingesetzt zu werden. Entsprechende Arbeiten befinden sich teilweise kurz vor dem Abschluss.

Bisher unberücksichtigte Fragestellungen umfassen u.a. Modelle für die Repräsentation und Qualität der Abfrageergebnisse (Ranking, Verweise auf die Originaldaten) und Fragen des Interaktionsdesigns. Zur Unterstützung komplexerer Abfragen auf strukturierten, unstrukturierten und semistrukturierten Daten werden unter anderem Ansätze zur verteilten Anfragebearbeitung mittels HyperQueries untersucht [Ke01]. Da die Daten in Dataspaces mit Unsicherheit behaftet und häufig inkonsistent sind, verdienen Methoden und Werkzeuge zur Darstellung der Datenherkunft und -entstehung weitere Beachtung. Konzepte für domänenspezifische Anwendungen, beispielsweise zur Weiterverwendung von Daten unter den strengen Anforderungen von Studienregularien können entwickelt werden. Ein weiterer nicht in dieser Arbeit adressierter Aspekt ist die Ablaufintegration.

Literaturverzeichnis

[Al08] Altman, R.B. et al.: Commentaries on “Informatics and Medicine: From Molecules to Populations”. Methods Inf Med. 2008;47(4): S. 296-316.

[ca07] The caBIG Strategic Planning Workspace: The Cancer Biomedical Informatics Grid (caBIGTM): Infrastructure and Applications for a Worldwide Research Community. MEDINFO 2007, K. Kuhn et al. (Eds), IOS Press, 2007.

[De07] DeRose, P.; Shen, W.; Chen, F.; Lee, Y.; Burdick, D.; Doan, A.; Ramakrishnan, R.: DBLife: A Community Information Management Platform for the Database Research Community (Demo). CIDR 2007: S. 169-172.

[Di06] Dittrich, J.: iMeMex: A Platform for Personal Dataspace Management. Proceedings of the 2nd NSF sponsored workshop on PIM, In conjunction with ACM SIGIR 2006, Seattle,Washington.

[Di07] Dittrich, J.: Aktuelle Trends: Eine Reise von Hauptspeicherdatenbanken zu Dataspace Management Systemen. ULDB 2007.

[FHM05] Franklin, M. J.; Halevy, A. Y.; Maier, D.: From databases to dataspaces: a new abstraction for information management. SIGMOD Record 34(4): S. 27-33 (2005).

[Hi07] Hibbert, M.; Gibbs, P.; O'Brien, T.; Colman, P.; Merriel, R.; Rafael, N.; Georgeff, M.: The Molecular Medicine Informatics Model (MIMM). MEDINFO 2007, K. Kuhn et al. (Eds), IOS Press, 2007.

[Ke01] Kemper, A.; Wiesner, C.: HyperQueries: Dynamic Distributed Query Processing on the Internet. Proceedings of the 27th VLDB Conference (2001): S. 551-560

[Ku07] Kuhn, K. A.; Giuse, D. A.; Lapão, L.; Wurst, S. H. R.: Expanding the Scope of Health Information Systems - From Hospitals to Regional Networks, to National Infrastructures, and Beyond. Methods Inf Med 2007; 46: S. 500–502.

[Ku08] Kuhn, K. A. et al.: Informatics and medicine. From molecules to populations. Methods Inf Med. 2008;47(4): S. 283-295.

[LK04] Lenz, R.; Kuhn, K. A.: Towards a continuous evolution and adaptation of information systems in healthcare. International Journal of Medical Informatics (2004) 73, S. 75-89.

[Lo07] Louie, B.; Mork, P.; Martín-Sánchez, F.; Halevy, A. Y.; Tarczy-Hornoch, P.: Data integration and genomic medicine. Journal of Biomedical Informatics 40(1): S. 5-16 (2007).

[Ma06] Maojo, V.; García-Remesal M.; Billhardt, H.; Alonso-Calvo, R.; Pérez-Rey D.; Martín-Sánchez, F.: Designing New Methodologies for Integrating Biomedical Information in Clinical Trials. Methods Inf Med 2006; 45: S. 180–185.

[Ma07] Madhavan, J.; Jeffery, S. R.; Cohen, S.; Dong, X.; Ko, D.; Yu, C.; Halevy, A. Y.: Web-scale Data Integration: You can only afford to Pay As You Go. CIDR 2007: S. 342-350.

[Na06] National Center for Research Resources: Clinical and Translational Science Awards. http://www.ncrr.nih.gov/clinical_research_resources/clinical_and_translational_science_awards/, letzter Zugriff: 15.05.2009.

[Ol05] Oliveira, I. C.; Oliveira, J. L.; Sanchez, J. P.; López-Alonso, V.; Martin-Sanchez, F.; Maojo, V.; Sousa Pereira, A.: Grid Requirements for the Integration of Biomedical Information Resources for Health Applications. Methods Inf Med 2005; 44: S. 161–167.

[Va07] Vaz Salles, M. A.; Dittrich, J.; Karakashian, S. K.; Girard, O.R.; Blunschi, L.: iTrails: Pay-as-you-go Information Integration in Dataspaces. VLDB 2007: S. 663-674.

[WB05] Wears, R. L.; Berg, M.: Computer technology and clinical work: still waiting for Godot. JAMA. 2005 Mar 9;293(10): S. 1261-1263.

[Wi05] Wimmer, M.; Ehrnlechner, P.; Fischer, A.; Kemper, A.: Flexible Autorisierung in Datenbank-basierten Web Service-Föderationen. Informatik Forsch Entw 2005; 20: S. 167-181.

[Wu07] Wurst, S. H. R.; Lamla, G.; Schlundt, J.; Karlsen, R.; Kuhn K. A.: A Service-oriented Architectural Framework for the Integration of Information Systems in Clinical Research. IEEE CBMS 2008.