Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1101

Prozessmodellierung mittels BPMN in Forschungsinfrastrukturen der Digital Humanities

Christoph Kuras[1] Thomas Eckart[2]

Abstract: In den Geistes- und Sozialwissenschaften werden Arbeitsprozesse zunehmend unter den Gesichtspunkten Reproduzierbarkeit, Wiederverwendbarkeit, erhöhte Transparenz und Flexibilität bewertet. Entsprechend werden verstärkt sogenannte „Scientific Workflows" eingesetzt, um erstellte Verarbeitungsketten zu persistieren, verständlich und austauschbar zu machen. Bestehende Portale leiden allerdings an einer Uneinheitlichkeit der verwendeten Beschreibungslösungen und können nur in Teilen die an sie gestellten Anforderungen erfüllen. Dies ist insbesondere im Umfeld komplexer Forschungsinfrastrukturen, wie sie in den letzten Jahren verstärkt entwickelt werden, hoch relevant. Als Alternative können standardisierte Modellierungen auf Basis der BPMN (Business Process Model and Notation) die Vergleichbarkeit und Kombinierbarkeit von Teilschritten erleichtern und als Hilfsmittel dienen, um den systematischen Konflikt der Reproduzierbarkeit von Forschungsergebnissen unter Beibehaltung maximal möglicher Flexibilität aufzulösen. Entsprechende Modellierungen bieten zudem den Vorteil, die direkte Ausführbarkeit der beschriebenen Arbeitsschritte im Rahmen etablierter und quelloffener Workflow-Management-Systeme (WfMS) zu ermöglichen.

Keywords: Scientific Workflows; Forschungsinfrastrukturen; BPMN

1 Prozess- und Workflowmodellierung in wissenschaftlichen Communities

Der zunehmende Einfluss quantitativer Analysen und Verfahren auf die alltägliche Arbeitsweise der Geistes- und Sozialwissenschaften in den letzten Jahren haben zu einem zunehmenden Einfluss IT-basierter Hilfsmittel und Werkzeuge geführt. Damit vollziehen einige dieser Fachbereiche eine Entwicklung zu Arbeitsweisen wie sie bisher eher typisch für die Natur- und Lebenswissenschaften sind und wecken gleichzeitig vergleichbare Ansprüche bzgl. der Reproduzierbarkeit konkreter Experimente und Analysen (sogenannte „Reproducible Science") oder der effizienten Wiederverwendung bestehender Systeme für neue Datenbestände.

[1] Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, 04109 Leipzig, Deutschland
ckuras@informatik.uni-leipzig.de
[2] Universität Leipzig, Abteilung Automatische Sprachverarbeitung, Augustusplatz 10, 04109 Leipzig, Deutschland
teckart@informatik.uni-leipzig.de

doi:10.18420/in2017_112

Folgerichtig hat dieser Trend zur zunehmenden Nutzung von Scientific-Workflow-Management-Systemen (SWfMS) geführt, die sich in ähnlicher Form in den bereits angesprochenen Vorbildern großer Beliebtheit erfreuen[3]. Die konkret verwendeten Werkzeuge haben meist gemein, dass sie primär datengetrieben arbeiten und sich, unter Ausklammerung typischer Interoperabilitätsprobleme, durch Verkettung von Teilschritten zu komplexen Workflows kombinieren lassen.
Offensichtlich ist hier die Parallele zum, in der Betriebswirtschaft verbreiteten, Geschäftsprozessmanagement. Wie auch Geschäftsprozesse sollen Scientific Workflows beschreiben, was getan wird, jedoch nicht zwangsläufig Informationen zur konkreten Umsetzung enthalten, um die leichtere Verständlichkeit für eine breite Gruppe von Adressaten, etwa auch Personen mit geringen technischen Vorkenntnissen, zu verbessern (vgl. [Wo09]). Im deutlichen Kontrast zum Geschäftsprozessmanagement spielen in den SWfMS die betrieblichen Aspekte jedoch keine Rolle. Ziel von SWfMS ist es primär, komplexe computergestützte Experimente automatisieren zu können, wobei auch auf verteilte Daten und Dienste zugegriffen werden kann, so dass rechenintensive Workflows auch auf leistungsschwachen Rechnern lauffähig sind (vgl. [Wo09]). Bemerkenswert ist der teils synonyme Gebrauch der Begriffe *Prozess* und *Workflow* in den entsprechenden Publikationen, deren Unterscheidung zumindest aus betriebswirtschaftlicher Sicht eine hohe Bedeutung aufweist.

Als ein Ergebnis dieser Entwicklung haben sich Portale herausgebildet, die dem Austausch konkreter Workflows dienen. Beispielhaft sei hier auf das Portal *myExperiment.org* verwiesen, auf dem Nutzer Workflows für verschiedene Systeme teilen können. Im konkreten Einsatz (dokumentiert durch öffentliches Nutzerfeedback) ergeben sich allerdings diverse Probleme. Für eingetragene Workflows bestehen kaum Qualitätskontrollen: es existiert keine systematische Überprüfung der versprochenen Funktionalität, häufig sind extern eingebundene Dienste nicht oder nicht mehr verfügbar. Einblendungen weisen teils darauf hin, dass Verarbeitungsketten veraltete Dienste nutzen, dennoch ist dieser Zustand für viele Nutzer unbefriedigend (vgl. [my17a]). Offensichtlich ist auch die fehlende Einheitlichkeit verwendeter Beschreibungsmittel: verwendet werden verschiedenste Beschreibungssprachen, was teils zu einer Abhängigkeit zwischen Werkzeug und Modell führt und eine integrierte Nutzung von Modellen erschwert. Bei genauer Betrachtung der Workflows fällt zudem auf, dass viele Workflows die funktionale Beschreibung (Was?) nicht von Implementierungsdetails (Wie?) trennen, so dass teils technische Arbeitsschritte (wie spezifische Datentransformationen mit Referenzen auf konkrete Technologien) beschrieben werden und zu einer Vermischung der fachlichen und der operativen Ebene führen (vgl. [my17b]).
Gerade diese, im Geschäftsprozessmanagement übliche, Trennung sorgt allerdings für die Verständlichkeit und Nachvollziehbarkeit der dargestellten Vorgänge bei IT-fernem Personal, wie sie gerade für interdisziplinäre Arbeitsgruppen bzw. den gesamten Bereich der eHumanities typisch sind. Geprägt sind diese Fachbereiche durch die unterschiedlichsten Ausmaße technischer Fertigkeiten der beteiligten Personen, so dass die erfolgreiche Ausführung konkreter Workflows nicht in allen Fällen gelingen wird. In der Konsequenz kann ein Portal

[3] Wie etwa Kepler (https://kepler-project.org) oder Taverna (https://taverna.incubator.apache.org) in der Bioinformatik.

wie myExperiment.org auch als Sammlung von Insellösungen betrachtet werden, deren effektiver Wert und Nachvollziehbarkeit für Dritte variiert.

2 Forschungsinfrastrukturen

Als Antwort auf die angesprochenen Tendenzen wurden in den letzten Jahren verschiedene Lösungsstrategien erarbeitet. Unter diese fällt auch der Aufbau dedizierter Forschungsinfrastrukturen, die den Fachwissenschaftlern (mehr oder weniger) integrierte Anwendungen anbieten. Als ein Vorteil dieses Ansatzes wird gesehen, dass gemeinschaftlich und auch fachübergreifend erarbeitete Standards und Anwendungen helfen, den vorhandenen Bestand an Daten und Werkzeugen auch für unerfahrene Nutzer systematisch zu erschließen, Doppelarbeit und die Schaffung „isolierter" Ressourcen zu vermeiden und somit deren Wiederverwendung zu erleichtern. Konkrete Ausprägungen solcher Forschungsinfrastrukturen sind zum Beispiel CLARIN ([HK14]) oder DARIAH ([GH16]), die teils überlappende Nutzergruppen der Geistes-, Sozial- und Kulturwissenschaften adressieren.

Der grundlegenden Idee solcher Umgebungen wird allerdings teils mit Skepsis begegnet. Grundsätzliche Vorbehalte beinhalten vor allem die Frage nach der angemessenen Einbindung der Fachwissenschaftler (vgl. [BMS16]), nach fehlender Flexibilität und Spezifität entwickelter Lösungen, um im volatilen Forschungsbetrieb von Nutzen zu sein sowie die Gefahr des Aufoktroyierens ungeeigneter Lösungen durch Fachfremde. Pointiert zusammengefasst wurde dies zum Beispiel durch Joris van Zundert:

A 'one size fits all' approach would be a disastrous underestimation of the specific needs of humanities research. The essence of humanities research is in its diverse, heterogeneous and ephemeral nature. [Zu12, S.8]

bzw.

The quick technology shifts and development of thinking that research innovation requires cannot be supported through these large unified infrastructures. [Zu12, S.11]

Als alternative Lösung schlägt er den dezentralen und Community-getriebenen Einsatz von Microservices vor, die unter Wahrung der Möglichkeit optionaler Verkettung zu komplexen Verarbeitungsketten auf Basis des JSON-Formats kurzlebige Forschung erlaubten. Zentrale Ziele der Systematisierung von Forschung durch Forschungsinfrastrukturen, wie die Nachvollziehbarkeit von Ergebnissen, Interoperabilität[4], Motivation zur offenen Bereitstellung sowie öffentliches Announcement verfügbarer Ressourcen, langfristige Vorhaltung relevanter Ergebnisse usw., wird durch diesen Ansatz allerdings nicht adressiert. Dass van Zunderts Vorstellung einer sich selbst unterhaltenden „Forschungsinfrastruktur", die als Grundlage des Austausches lediglich das Internet benötigt, optimistisch gedacht ist, zeigt sich auch an den Problemen des bereits beschriebenen *myExperiment.org*, das

[4] Die sich natürlich auch durch die Verwendung einfacher JSON-Formate nicht automatisch einstellt.

seiner geschilderten Vision recht nahe kommt, aber weder Qualität, Verständlichkeit oder gar Persistenz der angebotenen Dienste garantieren kann. Wenig Beachtung erfährt auch die Rollenverteilung, die beteiligte Personen typischerweise im Entwicklungsprozess einnehmen: Fachwissenschaftler mit zusätzlicher hoher Kompetenz in den Bereichen Softwaredesign und -entwicklung, Usability, Metadatengenerierung und -verteilung usw. sind und bleiben weiterhin eine Ausnahme.
Trotzdem bleibt aus der Perspektive der verschiedenen Infrastrukturprojekte die Frage nach angemessenen Antworten auf diese berechtigten Einwände und einer geeigneten Interessenabwägung im dargestellten Kontext. Um im Wettbewerb mit konkurrierenden Ansätzen die geforderten Leistungsmerkmale sicherzustellen (wie Verfügbarkeit, Qualität, Integration, Nutzerorientierung, Offenheit), ist zwangsläufig ein planvolles Vorgehen nötig und die Bereitschaft, die eigene Rolle im wissenschaftlichen Umfeld zu überdenken. Dabei sollte im Sinne einer verbesserten Nutzerorientierung in einem extrem volatilen Umfeld die eigene Rolle stärker als Dienstleister begriffen werden, der für seine Kunden (unabhängig von unmittelbarer finanzieller Abhängigkeit) Lösungen bzw. Vorarbeiten anbietet. Entsprechend ist hier die Parallele von Forschungsinfrastrukturen und dem wirtschaftlichen Betrieb eines privaten Unternehmens zu betonen und die Chance zu ergreifen, von vorhandenen Erfahrungswerten des Geschäftsprozessmanagements zu profitieren.

3 Forschungsinfrastrukturen im Wettbewerb

Aufgrund der Parallelen zu privatwirtschaftlichen Unternehmen greifen Scientific Workflows zum Zwecke der Modellierung von Arbeitsabläufen in Forschungsinfrastrukturen zu kurz, da sie wesentliche ökonomische Konzepte völlig außer Acht lassen und sich weitestgehend auf die Verkettung von ggf. verteilten Arbeitsschritten beschränken. Diese Konzepte sind aber notwendig, damit Betreiber von Forschungsinfrastrukturen tatsächlich effektiv, effizient und kundenorientiert agieren können[5]. Dabei können interne Faktoren, die im Einflussbereich des Unternehmens liegen und externe Faktoren, außerhalb des unternehmerischen Einflussbereiches, unterschieden werden. Letztere beinhalten die Kundenzufriedenheit, aber auch Wettbewerber, während interne Faktoren die Effizienz und Wirtschaftlichkeit, welche den unternehmerischen Erfolg gewährleisten sollen, widerspiegeln (vgl. [BKR12, S.3f]).

Die Vorteile einer durch Geschäftsprozessmanagement unterstützten unternehmerischen Ausrichtung von Forschungsinfrastrukturen sind vielfältig. Besonders hervorzuheben ist hier das „Marktumfeld“, das durch eine hohe Volatilität geprägt ist. Es liegt in der Natur der Forschung, einem schnellen und kontinuierlichen Wandel unterworfen zu sein. Änderungen des Marktumfeldes betreffen dabei nicht nur die Nutzer und deren sich ändernde Anforderungen selbst, sondern auch das Agieren möglicher Wettbewerber, die etwa plötzlich neue Angebote entwickeln oder sich anderen Nutzergruppen zuwenden können (vgl. [BKR12, S.3f]; [Ch15, S.17ff]). Auch im wissenschaftlichen Bereich kann es zum Wettbewerb mit

[5] Obwohl Nutzer in akademischen Forschungsinfrastrukturen meist keine monetären Gegenleistungen erbringen, werden die Begriffe *Kunde* und *Nutzer* hier synonym verwendet.

akademischen oder kommerziellen Anbietern kommen. Da Projekte im akademischen Umfeld zumeist eine Förderung beziehen, besteht der Anreiz zu einer wettbewerbsfähigen Gestaltung der Organisationsstruktur jedoch oft nicht von Beginn an, was mit dem Ende der Förderung jedoch zu erheblichen Problemen oder sogar zur Einstellung des Vorhabens führen kann. Damit einher geht die Notwendigkeit, möglichst effizient zu wirtschaften, um Fördermittel zielgerichtet einsetzen, aber auch, um zum Ende der jeweiligen Förderphase die Sinnhaftigkeit des Weiterbetriebes rechtfertigen zu können.
Dabei spielen durchaus auch nicht-monetäre Aspekte eine Rolle. Unternehmen müssen durchgehend kundenorientiert agieren, um ihre Leistungen am Markt erfolgreich absetzen zu können. Forschungsinfrastrukturen müssen analog von Beginn an nutzerorientiert handeln, da sie andernfalls Gefahr laufen, ihr Angebot abseits des eigentlichen Bedarfs anzusiedeln und somit ihre primären Ziele nicht erfüllen zu können. Dabei müssen nicht nur die *benötigten* Leistungen angeboten werden: wenn Unternehmen ihre Leistungen in unzureichender Qualität anbieten, wechseln Kunden, unter Abwägung aller Kosten (auch Such- und Informationskosten), möglicherweise zu anderen Anbietern. Da im akademischen Bereich monetäre Kosten keine vergleichbare Rolle spielen, ist dieser Aspekt von besonderer Relevanz. Entsprechend müssen für Forschungsinfrastrukturen eine hohe Qualität ihres Angebotes sowie eine hohe Zufriedenheit ihrer Nutzer ein herausragendes Ziel sein: etliche der in [Zu12] angesprochenen Probleme sind vor allem durch mangelnde Nutzer- bzw. Kundenorientierung zu begründen.
Viele dieser Ziele werden durch den Einsatz von Geschäftsprozessmanagement (GPM) unterstützt. Dabei handelt es sich um ein Management-Konzept, das auf mehreren Ebenen ansetzt (vgl. [Ga12, S.1ff]). Auf der strategischen Ebene werden zunächst die unternehmerischen Ziele und die zentralen Geschäftsfelder und die dazugehörigen Kernprozesse festgestellt. Diese werden auf der fachlich-konzeptionellen Ebene als Geschäftsprozesse modelliert. Darüber hinaus werden Kennzahlen definiert, die mit den Zielen bzw. den kritischen Erfolgsfaktoren der strategischen Ebene in Verbindung gebracht werden. Solche Kennzahlen werden systematisch in Form eines Kennzahlensystems dokumentiert und reichen von allgemeinen Kennzahlen wie Durchlaufzeiten und Fehlerraten bis hin zu spezifischen Kennzahlen, die nur in bestimmten Teilaufgaben des Prozesses gemessen werden, etwa die Anzahl von Sätzen im Rahmen der Erzeugung von Textkorpora. Auf der operativen Ebene werden die zuvor modellierten Prozesse durch zusätzliche technische Spezifikationen ausführbar gemacht - erst ein ausführbarer Prozess wird Workflow genannt (vgl. [Ga15, S.5]). Während der Ausführung durch ein Workflow-Management-System werden die zu Beginn definierten Kennzahlen erfasst und mit vorher definierten Soll-Werten verglichen, wodurch Rückschlüsse auf Schwachstellen und Probleme gezogen werden können und letztlich die Erreichung der unternehmerischen Ziele kontrolliert werden soll.
In dieser Arbeit werden hauptsächlich die fachliche und die operative Ebene betrachtet. Für einen umfassenden Einsatz von GPM müssen jedoch unbedingt alle Ebenen bzw. Phasen Beachtung finden. Tabelle 1 enthält eine Zusammenfassung der wesentlichen Vorteile, die der Einsatz von GPM für Unternehmen bzw. für Betreiber von Forschungsinfrastrukturen mit sich bringen kann.
Ein wichtiger Bestandteil des GPM ist die Modellierung von Geschäftsprozessen und deren

spätere voll- oder teilautomatisierte Ausführung in Form von Workflows. Aufgrund der oben beschriebenen Parallelen zu privatwirtschaftlichen Unternehmen und der damit verbundenen Ziele können auch akademische Serviceanbieter von diesem Konzept profitieren.

Vorteil	**im akademischen Bereich**
Effizienzsteigerung und Kostensenkung	
- Straffung der Abläufe - Reduzierung der Durchlaufzeiten	- Effiziente Nutzung von Fördermitteln - Schnellere Forschungsergebnisse durch Integration halb-automatisierter und manuell ausgeführter Aufgaben in den Arbeitsablauf
Transparenz	
- Klare Spezifikation der Leistung - Klare Verantwortlichkeiten - Leichtere Kommunikation zwischen Fach- und IT-Bereich - Dokumentation	- Abstimmung mit Fachbereichen - Nachvollziehbarkeit auch für IT-fernes Personal - Reproduzierbarkeit (ausführbare Workflows) - Schnellere Einarbeitung für rasch wechselndes Personal
Flexibilität	
- Ergebnisorientierte Integration über Funktionsbereiche hinweg - Bessere Ausrichtung an Kundenbedürfnissen	- Schnelle Reaktion auf geänderte Anforderungen durch neue Forschungsfragen

Tab. 1: Vorteile des Einsatzes von GPM

Konkret ermöglichen Modelle von Geschäftsprozessen eine bessere Kommunikation zwischen dem Fach- und dem IT-Bereich. Während Scientific Workflows keine Trennung des Adressatenkreises vornehmen, wird im GPM strikt zwischen der fachlichen und der operativen Ebene, welche die technischen Details zur Ausführung der Prozesse (Workflows) enthält, unterschieden. Diese Art der Dokumentation führt neben der verbesserten Anforderungsermittlung auch dazu, dass alle Mitarbeiter, unabhängig ihrer Domäne, eine gemeinsame Vision der Prozesse entwickeln können. In Bezug auf akademische Organisationen, die oft eine hohe Personalfluktuation aufweisen, kann dies ein zusätzlicher Vorteil sein, da die Einarbeitung in die Prozesse dadurch erleichtert wird. Darüber hinaus wird unter allen Prozessbeteiligten ein stärkeres Bewusstsein für die zu erbringende Leistung und, welche Rolle die Beteiligten konkret im Prozess einnehmen, gefördert. Eine umfassende Dokumentation und Möglichkeiten der schnellen Anpassbarkeit können als wesentliche Voraussetzungen für den Erfolg von Forschungsinfrastrukturen gesehen werden (vgl. [Gn17, S.16]).

Besonders leistungsfähig sind ausführbare Prozesse, sogenannte Workflows. Diese enthalten alle technischen Spezifikationen oder Arbeitsanweisungen, um den Prozess automatisiert ausführen zu können (vgl. [Ga15, S.5]). Dies geschieht durch ein Workflow-Management-System (WfMS), welches bei Bedarf Echtzeitinformationen während der Laufzeit der Workflows liefert. Neben der internen Dokumentation eignen sich Prozessmodelle auch dazu, die Abläufe für den Kunden transparenter zu gestalten, etwa durch vereinfachte

Prozessmodelle bzw. Service Blueprints (vgl. [Fl09, S.215]). So kann der Nutzer auch bei einer längeren Bearbeitungsdauer nachvollziehen, in welchem Arbeitsschritt der Prozess sich gerade befindet und ist somit toleranter gegenüber längeren Wartezeiten, sofern sie der Komplexität des ausgeführten Schrittes entspricht. Gerade in Forschungsinfrastrukturen, bei denen die Nutzer selbst Daten einbringen und bezüglich der Verarbeitungsdauer keine Erfahrungswerte mitbringen, kann dies sehr nützlich sein. Zusätzlich können durch längerfristige Analysen der Laufzeitdaten Flaschenhälse identifiziert werden, die dann, etwa durch Restrukturierung des Prozesses oder den Austausch der in den Teilschritten des Workflows enthaltenen Arbeitslogik einzelner Schritte, gezielt adressiert werden können.

Ein wesentlicher Unterschied zu Scientific Workflows ist zudem, dass Geschäftsprozesse im Besonderen dafür ausgelegt sind, auch manuell durchgeführte Arbeitsschritte oder Unterprozesse zu berücksichtigen. Dieser Aspekt kommt vielen Communities entgegen, in denen die Skepsis gegenüber einer vollautomatisierten Verarbeitung in einem als „Blackbox“ empfundenen System groß sein kann. Die Steigerung der Prozesstransparenz durch den Einsatz von Modellen ist daher ein wichtiger Aspekt. Durch den Einsatz von Geschäftsprozessmanagement können manuelle Bearbeitungsschritte bestehen bleiben, erhalten aber durch die Modellierung definierte In- und Outputs sowie eine nachvollziehbare Position im Gesamtprozess. Um unnötigen Aufwand zu vermeiden, werden dabei allerdings lediglich solche Aufgaben modelliert, die der Erreichung des jeweiligen Prozessziels dienlich sind. Im folgenden Abschnitt sollen anhand eines konkreten Beispiels diese Punkte verdeutlicht werden.

4 Korpuserzeugung - Modellierung am Beispielprozess

4.1 Prozessbeschreibung

In diesem Abschnitt wird beispielhaft ein Prozess des CLARIN-Zentrums Leipzig vorgestellt. Anhand der Ist-Situation sollen wesentliche Probleme aufgezeigt werden, die mit Hilfe eines WfMS adressiert werden können. Auf Basis des Prozessmodells des Soll-Zustands sollen anschließend einige Vorteile des Einsatzes eines WfMS verdeutlicht werden. Der Prozess zur Erstellung von Textkorpora (Datenbanken) aus den Inhalten gecrawlter Webseiten ist eine über viele Jahre gewachsene Verkettung einzelner Tools, die jeweils bestimmte Aufgaben übernehmen, um aus Textdokumenten ein Textkorpus, bestehend u.a. aus einer nach ihrer Frequenz sortierten Wortliste, extrahierten Sätzen und Wort-Kookkurrenz-Informationen zu erzeugen ([GEQ12]). Die verwendeten Softwarelösungen sind eine Sammlung von Java-Programmen sowie diversen Python-, Perl- und Shell-Skripten. Die Ausführung erfolgt bereits automatisiert, wobei die einzelnen Schritte von einem Python-Skript gesteuert werden, welches außerdem eine zentrale Konfiguration enthält.
Die verwendete Lösung hat über die Zeit stark an Komplexität gewonnen, woraus sich in der Praxis verschiedene Nachteile ergeben. Insbesondere haben nicht mehr alle Beteiligten eine klare Vorstellung von Art und Reihenfolge der Verarbeitungsschritte. Dadurch, dass

die einzelnen Schritte in verschiedenen Programmiersprachen implementiert sind und keine grafische Repräsentation existiert, dauern sowohl Einarbeitung als auch Fehlersuche unnötig lange. Konkrete Ausführungen und ihre Konfiguration werden nur schwach dokumentiert, was die Fehlersuche zusätzlich erschwert. Bei Erweiterungen des Prozesses entstehen zudem leichter Ineffizienzen, da bei der Implementierung von Ergänzungen immer nur ein kleiner Ausschnitt des Prozesses betrachtet wird und mögliche Auswirkungen auf andere Arbeitsschritte nicht beachtet werden. Die zu verarbeitenden Daten sind oftmals sehr groß, sodass bei länger andauernden Ausführungen, die bis zu mehreren Monaten benötigen können, häufig Aussagen zum aktuellen Bearbeitungsstand gewünscht sind. Dies ist aufgrund des unzureichendem Monitorings nur über die Betrachtung der Ausgabedateien einzelner Teilschritte bzw. eine entsprechende Abschätzung möglich und deshalb sehr ungenau. Nach jeder Ausführung und Erstellung eines Korpus' besteht der Bedarf einer Abschätzung der Qualität des Ergebnisses, welcher derzeit vorwiegend durch manuelle Tests auf der erstellten Datenbank abgedeckt wird. Darüber hinaus wurden bereits testweise Teile des dargestellten Prozesses zur Nutzung größerer Rechenkapazitäten an ein externes Rechenzentrum ausgelagert, was zusätzliche Koordinationsmechanismen notwendig macht. Dabei wirkt sich an der bisherigen Lösung insbesondere als nachteilig aus, dass einzelne Teilschritte nur aufwändig aus dem komplexen Skript zur Steuerung des Ablaufes herausgelöst werden können und damit ihre Wiederverwendung erschwert wird.
Ziel ist es daher, den Prozess mit einem Workflow-Management-System ausführen und überwachen zu können. Dabei sollen zusätzliche Messungen durchgeführt werden, die Hinweise auf mögliche Flaschenhälse oder Qualitätsprobleme bezüglich der Verarbeitung liefern. Dabei können zur Steigerung der Ergebnisqualität auch manuelle Prüfungen durch Mitarbeiter modelliert werden.

4.2 Modellierung mit BPMN

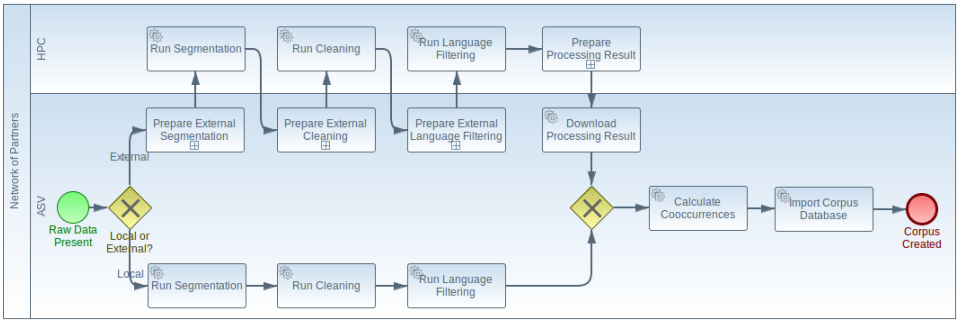

Abb. 1: Prozessmodell Korpuserzeugung (BPMN)

Die BPMN (Business Process Model and Notation) ist ein Standard der OMG (vgl. [OM11]). Ein großer Vorteil von BPMN ist neben ihrer weiten Verbreitung, dass damit modellierte Prozesse ausführbar gemacht und somit vollständig durch ein WfMS gesteuert

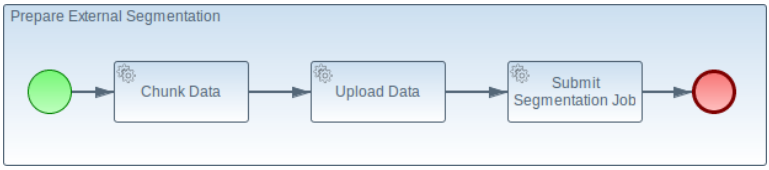

Abb. 2: Details des Subprozesses zur Vorbereitung der externen Satz-Segmentierung

und überwacht werden können. Mittlerweile existieren zahlreiche Ausführungsumgebungen, die den BPMN-Standard unterstützen, etwa jBPM, Activiti oder Camunda.

Abb. 1 zeigt den geplanten Soll-Zustand des oben beschriebenen Prozesses. Die Swimlane-Modellierung verdeutlicht die Kooperation zwischen dem CLARIN-Zentrum Leipzig und einem High-Performance-Computing-Rechenzentrum (HPC-RZ), wobei jede Lane einen Prozessbeteiligten repräsentiert. Obwohl es sich beim HPC-RZ um einen externen Prozessbeteiligten handelt, wird auf eine Modellierung in getrennten Pools verzichtet, da die Möglichkeit des unmittelbaren Zugriffs auf die Rechenkapazitäten besteht. Zudem würde eine Modellierung in separaten Pools lediglich die Kommunikation über einen Nachrichtenfluss erlauben, was durch die direkte Einbindung des Rechenzentrums in den Verarbeitungsfluss aber nicht der Realität entspräche (vgl. [GL13, S.131]). Kooperationen zwischen Organisationseinheiten sind mittels Scientific Workflows nur schwer abbildbar. Im dargestellten Prozessmodell findet die Verarbeitung entweder nur lokal oder teilweise im HPC-RZ statt, wofür Daten zum Rechenzentrum transferiert werden müssen. Die ersten drei Verarbeitungsschritte des externen Pfades, die Vorverarbeitung, enthalten jeweils einen als Subprozess modellierten Vorbereitungsschritt, dessen Details nur für die Ausführung bzw. IT-Personal von Interesse ist (vgl. Abb. 2). Bei einer ins HPC-RZ ausgelagerten Vorverarbeitung, muss das Endresultat nochmals bearbeitet werden, bevor es im CLARIN-Zentrum weiterverarbeitet wird. Die Details hierzu sind ebenfalls in einem Subprozess verborgen, da sie für ein grobes Verständnis des Ablaufes nicht benötigt werden. Ab diesem Zeitpunkt ist die Verarbeitung für beide Zweige wieder identisch. Es folgen Arbeitsschritte zur Berechnung von Kookkurrenzen[6] und der abschließende Import des Korpus in eine Datenbank. Die Flexibilität der gewählten Modellierungsform zeigt sich insbesondere anhand der einfachen Erweiterbarkeit des dargestellten Prozesses.

Abb. 3 stellt den oben beschriebenen Prozess, erweitert um einige zusätzliche Elemente, dar. Die Aufgabe für die lokale Segmentierung wurde hier mit einem Fehler-Event versehen. Tritt bei der Verarbeitung ein Fehler auf, wird ein Techniker informiert, der versucht, den Fehler zu beheben. Ist dies erfolgreich, wird die Bearbeitung erneut gestartet und andernfalls mit einem Fehler beendet.

Zusätzlich wurde nach der Kookkurrenzberechnung und der Erstellung der Wortliste eine

[6] Hierbei entstehen auch die Wortlisten. Die Modellierung in nur einem Task ist dennoch sinnvoll, da die Aufgaben hier vom gleichen Tool in einem Aufruf desselben bearbeitet werden.

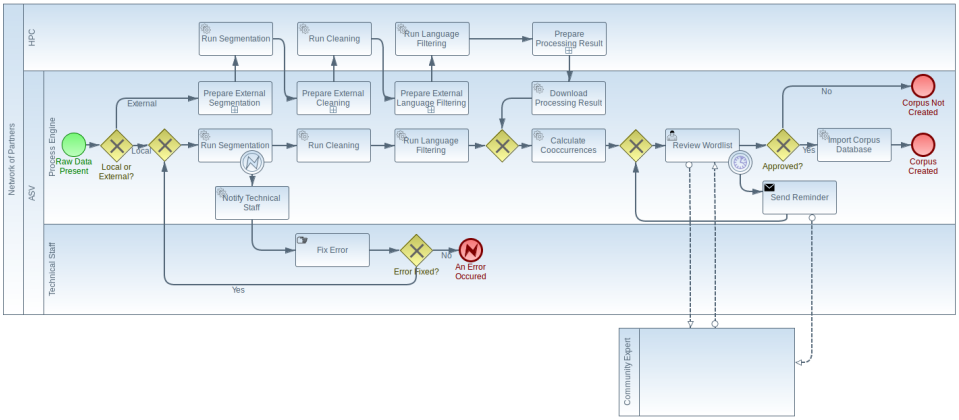

Abb. 3: Prozessmodell Korpuserzeugung, erweiterte Version (BPMN)

Benutzeraufgabe[7] eingefügt, deren Inhalt etwa die Prüfung der ersten 1000 Wortlisten-Einträge durch einen Community-Experten sein kann. Stellt der (externe) Experte, dessen Prozess hier als Blackbox modelliert wurde, Fehler fest, wird die Datenbank nicht importiert und der Vorgang beendet. Zusätzlich wurde der Review-Aufgabe ein Timer-Event hinzugefügt, welcher nach Ablauf einer definierten Zeitspanne dem Experten eine Erinnerung zur Bearbeitung der Aufgabe senden soll. Eine direkte und umfassende Integration von Fachbereich und IT ist somit gewährleistet. Hinsichtlich der konkreten Implementierung automatisch von IT-Systemen ausgeführter Teilschritte ergeben sich zudem die Vorteile, dass einerseits diese Teilschritte in anderen Prozessen wiederverwendet werden können und andererseits bei Bedarf die Implementierung einfach ausgetauscht werden kann, ohne den Gesamtzusammenhang dabei zu stören.

Die hier dargestellten Adaptionen stellen nur eine kleine Teilmenge möglicher Erweiterungen dar und können natürlich auch für deutlich größere und komplexere Arbeitsprozesse oder involvierte Nutzerrollen durchgeführt werden. Grundsätzlich erfüllt die BPMN alle Voraussetzungen, um auch wissenschaftliche Arbeitsabläufe adäquat abbilden zu können und stellt in Verbindung mit einem umfassenden Geschäftsprozessmanagement ein mächtiges Werkzeug dar. Die in einigen Communities möglicherweise als überflüssig wahrgenommenen ökonomischen Aspekte können gerade in großen Forschungsinfrastrukturen helfen, Leistungen kundenorientiert und effizient zu erstellen und dabei eine Brücke zwischen Fach- und IT-Bereich zu schlagen.

[7] Eine gute Übersicht der Aufgabentypen bietet [GL13].

5 Fazit

In komplexen Forschungsinfrastrukturen bestehen zahlreiche Herausforderungen, darunter vor allem der Bedarf, Arbeitsabläufe zu dokumentieren sowie nachvollziehbar und wiederverwendbar zu machen. Im akademischen Bereich haben sich Scientific Workflows herausgebildet, deren Prinzip dem des Geschäftsprozessmanagements entlehnt ist, aber ökonomische Aspekte außer Acht lässt. Im Rahmen großer Forschungsinfrastrukturen sind letztere jedoch unverzichtbar; Wirtschaftlichkeit kann hier zu einem wichtigen Erfolgsfaktor werden. Die vorliegende Arbeit zeigt, dass Konzepte des GPM sinnvoll auch auf den akademischen Bereich übertragen werden können. Letztlich handelt es sich bei GPM aber nicht nur um ein Werkzeug zur Strukturierung von Prozessen, sondern um einen Managementansatz, der zu einer kundenorientierten und effizienten Leistungserstellung führen soll.

Für weiterführende Arbeiten ist es daher sinnvoll, die Bedeutung von GPM als Management-Konzept für Forschungsinfrastrukturen zu analysieren. Damit einher geht die Entwicklung konkreter Kennzahlensysteme, wobei sich diese Kennzahlen immer an der Erreichung spezifischer Ziele orientieren sollten[8], welche einer Strategie entstammen, um dem Aufwand der Messung einen Nutzen gegenüber stellen zu können. Da die Ziele von Forschungsinfrastrukturen sehr unterschiedlich sein können, sollten sich künftige Arbeiten damit befassen, in welchem Umfang Kennzahlen auch in unterschiedlichen Szenarien wiederverwendet werden können. Diese ließen sich damit auch im Sinne eines Benchmarkings für Forschungsinfrastrukturen einsetzen. Einen guten Überblick über wichtige Vorarbeiten dazu finden sich in [Gn17]. Darüber hinaus kann die BPMN auch weiter an das konkrete Umfeld angepasst werden. Um einzelne Schritte für Fachdomänen im Geistes- und sozialwissenschaftlichen Kontext nachvollziehbarer zu machen, wäre es unter anderem denkbar, diese mit Termen aus TaDiRAH, einer Taxonomie von Forschungsaktivitäten aus den Geisteswissenschaften, zu verknüpfen (vgl. [Ta17]).

Literaturverzeichnis

[BKR12] Becker, J.; Kugeler, M.; Rosemann, M., Hrsg. Prozessmanagement - Ein Leitfaden zur prozessorientierten Organisationsgestaltung. Springer Gabler, 2012.

[BMS16] Busch, A.; Meister, J.; Schumacher, M.: Wo bleibt eigentlich der einzelne Fachwissenschaftler? Bibliothek Forschung und Praxis, 40:278–282, 2016. Abgerufen am 6.4.2017, doi:10.1515/bfp-2016-0028.

[Ch15] Christ, J. P.: Intelligentes Prozessmanagement - Marktanteile ausbauen, Qualität steigern, Kosten reduzieren. Springer Gabler, 2015.

[8] Etwa „Angebot aufbereiteter Textkorpora in 20 Sprachen“ oder „Verfügbarkeit konkreter Textannotationen in mindestens 80% aller Korpora“ als Beispiel für quantitative Kennzahlen. Für qualitative Metriken könnten z.B. die relativen Häufigkeiten der einzelnen Annotationen gemessen und mit einer typischerweise zu erwartenden Verteilung als Soll-Wert verglichen werden.

[Fl09] Fließ, S.: Dienstleistungsmanagement - Kundenintegration gestalten und steuern. Gabler, 2009.

[Ga12] Gadatsch, A.: Grundkurs Geschäftsprozess-Management - Methoden und Werkzeuge für die IT-Praxis: Eine Einführung für Studenten und Praktiker. Springer Vieweg, 2012.

[Ga15] Gadatsch, A.: Geschäftsprozesse analysieren und optimieren - Praxistools zur Analyse, Optimierung und Controlling von Arbeitsabläufen. Springer Vieweg, 2015.

[GEQ12] Goldhahn, Dirk; Eckart, Thomas; Quasthoff, Uwe: Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In: In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). 2012.

[GH16] Gradl, Tobias; Henrich, Andreas: Die DARIAH-DE-Föderationsarchitektur - Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen. Bibliothek. Forschung und Praxis, 40(2):222–228, 2016.

[GL13] Göpfert, J.; Lindenbach, H., Hrsg. Geschäftsprozessmodellierung mit BPMN 2.0 - Business Process Model and Notation. Oldenbourg Verlag München, 2013.

[Gn17] Gnadt, T.; J., Stiller; Schmitt, V.E.; Thoden, K.: Faktoren und Kriterien für den Impact von DH-Tools und Infrastrukturen, Technical Report. DARIAH-DE Working Papers, (21), 2017. Göttingen.

[HK14] Hinrichs, Erhard; Krauwer, Steven: The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), S. 1525–1531, Mai 2014.

[my17a] myExperiment - Workflows - Pathways and Gene annotations for QTL region. `https://www.myexperiment.org/workflows/16.html`, 2017. Abgerufen am 3.4.2017.

[my17b] myExperiment - Workflows - Success-Abandonment-Classification. `https://www.myexperiment.org/workflows/140.html`, 2017. Abgerufen am 3.4.2017.

[OM11] OMG: BPMN 2.0 Spezifikation. Website, 2011. `http://www.omg.org/spec/BPMN/2.0/PDF/`, Abgerufen am 9.4.2017.

[Ta17] TaDiRAH - Taxonomy of Digital Research Activities in the Humanities. `http://tadirah.dariah.eu`, 2017. Abgerufen am 5.4.2017.

[Wo09] Wolstencroft, K.; Fisher, P.; De Roure, D.; Goble, C: Scientific Workflows. OpenStax CNX, 2009.

[Zu12] Zundert, Joris van: If you build it, will we come? Large scale digital infrastructures as a dead end for Digital Humanities. Historical Social Research, 37(3):165–186, 2012.