TK III: Infrastrukturen für eResearch

DO.UBT (Digital Objects @ UBT)

Forschungsdatenmanagement im SFB 840

Vortragende:

Claudia Piesche

IT-Servicezentrum
Universität Bayreuth

0921-555855

claudia.piesche@uni-bayreuth.de

Universitätsstr. 30
95447 Bayreuth

Co-Autor:

Dr. Andreas Weber

IT-Servicezentrum
Universität Bayreuth

0921-555851

andreas.weber@ uni-bayreuth.de

Universitätsstr. 30
95447 Bayreuth

DO.UBT (Digital Objects @ UBT)

Forschungsdatenmanagement im SFB 840

Claudia Piesche, Dr. Andreas Weber

IT-Servicezentrum
Universität Bayreuth
Universitätsstr. 30
95447 Bayreuth
claudia.piesche@uni-bayreuth.de
andreas.weber@ uni-bayreuth.de

Abstract: Die Diskussion über sinnvolle Langzeitverfügbarkeit von Forschungsdaten hat in den letzten Jahren an Bedeutung gewonnen. Insbesondere die Anforderungen zur Langzeitspeicherung an Forschungsdaten aus öffentlich geförderten Drittmittelprojekten müssen in den Blick genommen werden. Während globale Forschungsdatenportale, beispielsweise für Klimadaten schon seit längerem standardisiert und öffentlich verfügbar sind, gibt es aktuell keine einheitliche Unterstützung für die große Menge von Forschungsdaten, die in kleinen, individuellen und spezialisierten Forschungsbereichen entstehen. In diesen Fällen muss die Langzeitspeicherung durch individuelle lokale Lösungen unterstützt werden. Betrachtet man weiterhin die sich verändernden Anforderungen, kann eine Verlagerung des Interesses von der bloßen Speicherung von Daten hin zur Unterstützung des eResearch konstatiert werden. Insofern sollten erwähnte lokale Lösungen im Idealfall wichtige Schritte der Erzeugung und Transformation publizierter Forschungsergebnisse abbilden. Das Teilprojekt „INF Z2“ des Sonderforschungsbereichs (SFB) 840 hat sich zum Ziel gesetzt, eine Infrastruktur zur Unterstützung der Langzeitspeicherung von Forschungsdaten zu implementieren, welche gleichzeitig ermöglicht, den Entstehungsprozess daraus abgeleiteter Forschungsergebnisse nachzuvollziehen.

1 Ausgangssituation

Im Rahmen des „Teilprojekts INF“ (Z2) des Sonderforschungsbereichs 840 der Universität Bayreuth wird eine Infrastruktur konzipiert und umgesetzt, die eine langfristige Speicherung und Auffindbarkeit der im SFB erzeugten Forschungsdaten und eine Rekonstruktion des Entstehungsprozesses publizierter Forschungsergebnisse ermöglicht. Zusätzlich sollen die Anforderungen für eine erfolgreiche Zusammenarbeit der Forscher innerhalb des SFB und mit externen Kollegen erfüllt werden.

Durch die Veröffentlichung der Publikation „OECD Principles and Guidelines for Access to Research Data from Public Funding“ [OECD07] im Jahr 2007 stieg die Bedeutung der

Langzeitspeicherung von Forschungsdaten stetig. Einer der Hauptgedanken dieser Richtlinien ist die sinnvolle Wiederverwendung und Nachnutzung einmal generierter Forschungsdaten. Im Idealfall soll der Zugriff auf vorhandene Forschungsergebnisse nachfolgende Forschungsprozesse beschleunigen und durch Synergieeffekte eine spürbare Kostenersparnis entstehen. Gleichzeitig fordern namhafte Journale den Zugriff auf die den Publikationen zugrunde liegenden Forschungsdaten und dadurch eine Nachvollziehbarkeit des Forschungsprozesses durch Reviewer und Leser.

Diese Gedanken wurden von wichtigen deutschen Förderorganisationen, wie der DFG als Voraussetzung für eine Förderfähigkeit in die Forschungsrichtlinien aufgenommen. [DFG06] Insofern stellt der Umgang mit Forschungsergebnissen und deren Langzeitspeicherung eine Voraussetzung für die öffentliche Förderung von Forschung dar und muss bei entsprechenden Anträgen immer berücksichtigt werden. Eine ähnliche Zielvorstellung für IT-Teilprojekte in SFBs findet sich auch bei [Ef10]:

"Thus, the project should serve all or most of the scientific projects. It should consider existing standards and get connected to existing repositories, data bases or the like. It is also required to maintain, manage, document, and backup the data in cooperation with a local library or computing centre in a sustainable, permanent, and stable system."

Obwohl die Anforderungen an die Langzeitspeicherung von Forschungsdaten und -ergebnissen seit geraumer Zeit bekannt sind, entwickeln sich Ansätze für unterstützende Infrastrukturen zur Langfristspeicherung nur sehr langsam. Nur in einigen Forschungsbereichen gibt es bisher globale Repositorien für die Bereitstellung von Forschungsergebnissen, beispielsweise zur Nachnutzung in nachfolgenden Projekten. Ein Grund dafür ist, dass die globale Sammlung von Daten zu einem bestimmten Forschungsgebiet eine hohe Standardisierung der Daten voraussetzt, wie es zum Beispiel im Human Genome Project (HGP) [HGP14] oder der Klimaforschung der Fall ist.

Für die Mehrheit der Forschungsgebiete ist eine globale Standardisierung nicht möglich, da es sich oft um hoch spezialisierte und teilweise sogar einzigartige Forschungsaktivitäten handelt. Eine allgemeine Beschreibung dieser sehr speziellen Experimente, Messungen, Beobachtungen und Ergebnisse ist sehr zeitaufwändig und wird von den meisten Wissenschaftlern vermieden. Daher gibt es, wenn überhaupt, nur sehr individuelle Lösungen zur Beschreibung und Langzeitspeicherung von Forschungsergebnissen, die meist aus einzelnen Projekten entstanden sind.

Das bisher übliche Vorgehen zur Sicherung von Forschungsdaten (Abbildung 1) besteht in den meisten Fällen aus einer Aufbewahrung auf lokalen Speichermedien der Wissenschaftler beziehungsweise auf gemeinsam genutzten Netzlaufwerken einer Forschungsgruppe. Dies führt allerdings zu limitierten Zugriffsmöglichkeiten durch die jeweiligen Einzelpersonen und / oder Gruppen. Dies wiederum ist der Grund für oft heterogene Zusatzinformationen, weil die Beschreibung der Daten im Verantwortungsbereich des einzelnen Mitarbeiters liegt.

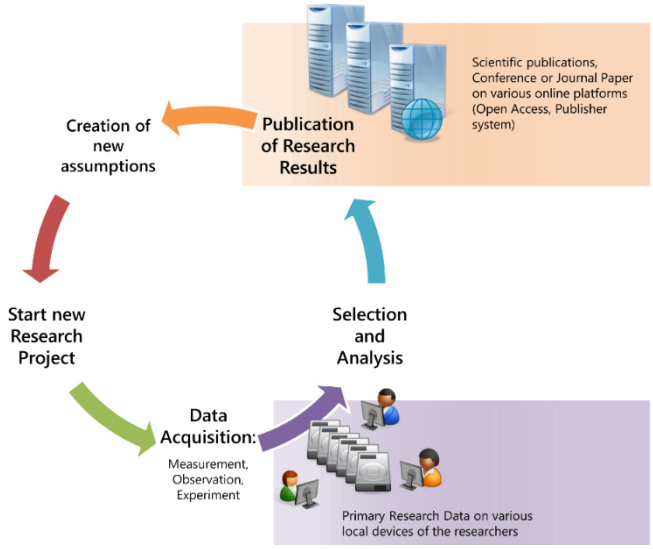

Abbildung 1: Traditioneller Lebenszyklus von Forschungsdaten

Abbildung 1 zeigt ein weiteres Problem der traditionellen Datensicherung: Der traditionelle Lebenszyklus von Forschungsdaten erlaubt einen öffentlichen Zugang nur auf Publikationen, die in dafür vorgesehenen Systemen veröffentlich wurden. Die Forschungsdaten selbst sind in den lokalen Speichersystemen verborgen und nicht zugänglich.

2 Related Work

Aufgrund der heterogenen Bedeutung des Begriffs Forschungsdatenmanagement, gibt es eine große Bandbreite verwandter Themen. Zusätzlich dazu führte die breitere Beschäftigung mit diesem Thema innerhalb der Forschergemeinschaft in den letzten Jahren zu einer großen Vielfalt von Forschungsansätzen, die sich mit der Anreicherung von Publikationen mit zusätzlichen Informationen oder der Erweiterung von Publikationen durch Hinzufügen von Forschungsdaten beschäftigen.

Zuerst soll an dieser Stelle ein wichtiger Aspekt betrachtet werden, der sich mit der Anreicherung publizierter Daten mit Informationen zu Struktur und Bedeutung beschäftigt. Aufbauend darauf wird die Idee von ‚Enhanced Publications' gezeigt, die grundsätzlich auf der Anreicherung von Publikationen basiert und noch einen konzeptionellen Schritt weitergeht. Als letzter Bereich im Forschungsdatenmanagement werden bestehende Infrastrukturen, beziehungsweise Konzepte zur technischen Unterstützung von e-Research Ansätzen gesichtet.

Das Feld zugehöriger Forschungen wird dabei teilweise durch die Zuhilfenahme von Beispielprojekte und an anderer Stelle durch die Darstellung des akademischen Standpunkts zum jeweiligen Bereich dargestellt. Den Startpunkt bildet dabei die

Betrachtung der wissenschaftlichen Beschäftigung mit Langzeitspeicherung, Datenmanagement und den daraus erwachsenden Anforderungen für ein angemessenes Konzept zur Datenverwaltung, Aufbewahrung und Publikation. Ergänzt wird diese Betrachtung durch die Darstellung der Projekte ‚RADAR' als Umsetzung eines integrierten Forschungsdatenrepositoriums, welches einen Großteil der konzeptionellen Anforderungen erfüllt. Schließlich wird der Ansatz zur Anreicherung von Publikationen mit notwendigen Strukturinformationen und Informationen zur Bedeutung und zum Inhalt untersucht.

2.1 Langzeitspeicherung und Management von Daten im Allgemeinen

Eine Vielzahl von Forschergruppen ist beteiligt an der wissenschaftlichen Auseinandersetzung mit dem Thema Datenmanagement beziehungsweise Forschungsdatenmanagement.

[LE13] beschreibt die Langzeitspeicherung von Daten als Zusammenspiel aus Erhaltung des Bitstroms (bitstream preservation) und der Nachnutzbarkeit von Forschungsdaten in inhaltlicher sowie technischer Hinsicht. Wegen der Anforderung der Nachnutzbarkeit, beziehungsweise Wiederverwendbarkeit einmal produzierter Daten, ist es notwendig zusätzlich zu den reinen Daten, Metainformationen zu Entstehungsprozess und –kontext zu sichern. Auf der Grundlage eines Lebenszyklus-Modells und dem ‚Curation Continuum' entwickeln die Autoren hierfür eine Checkliste zur Planung eines sinnvollen Forschungsdatenmanagements. In dieser Hinsicht legen sie die Grundlage für einen angemessenen Zugang zu diesem Thema, allerdings ohne konkrete Hinweise für die Umsetzung in eine geeignete Infrastruktur zu geben. Trotzdem ist es natürlich wichtig, das Ziel eines spezifischen Datenmanagements nicht aus dem Blick zu verlieren, wie es auch bei [OSN12] und [Ne12] festgestellt wird. Verschiedene Ziele bedingen zwangsläufig verschiedene Sichtweisen auf Langzeitspeicherung und Datenverwaltung.

Das 'Data Curation Continuum' in [TH08] befürwortet die These, dass Forschungsdaten in verschiedenen Domänen erstellt, bearbeitet, verändert und publiziert werden. Insofern spannen die Charakteristika der Objektdaten ein Kontinuum auf zwischen Start- und Endpunkt des Forschungdatenlebenszyklus. Betrachtet man die Charakteristika in diesen verschiedenen Phasen des Lebenszyklus (Erzeugung, Bearbeitung und Publikation) genauer, erkennt man, dass es sich in Wahrheit nicht um verschiedene Eigenschaften handelt, sondern um jeweils die gleiche Eigenschaft in verschiedenen Stati. Zusätzlich zum Kontinuum der Charakteristika von Datenobjekten gibt es die Idee, anhand dieser Charakteristiken die Forschungsdatenumgebung in drei strukturelle Bereiche einzuteilen (Forschung, Zusammenarbeit und Öffentlichkeit). Jeder dieser Bereiche wird dabei verknüpft mit einer passenden Repositoriums- beziehungsweise Datenspeicher-Infrastruktur [TH08].

Weil unser Ansatz für die Verwaltung von Forschungsdaten im SFB 840 alle Phasen des Lebenszykluses unterstützen soll, muss das Konzept des ‚Data Curation Continuums' berücksichtigt werden. Daher wird es notwendig sein, die drei Kernkonzepte innerhalb der enstehendenen Infrastruktur zu implementieren.

2.2 Ein integriertes Forschungsdatenrepositorium - Projekt ‚RADAR'

Das Hauptziel des Projekts ‚RADAR', welches durch die DFG gefördert wird, ist die Implementierung eines integrierten Forschungsdatenrepositoriums als Dienstleistung für Forschungseinrichtungen, mit dessen Hilfe Forscher in der Lage sein werden, Forschungsdaten zu sichern und zu publizieren. [TIB13].

Auf der Grundlage des 'Data Curation Continuum' stellt das Projekt einen zweistufigen Prozess zur Verfügung [PWRW12]. Einerseits wird das Repositorium einen generischen Ansatz für diverse Forschungsbereiche unterstützen, indem es in einer ersten Stufe die bloße Datenaufbewahrung anbietet. Die zweite Stufe besteht in einem weitergehenden Konzept zur Aufbewahrung und Veröffentlichung von Forschungsdaten. Da sich das Projekt gerade im Stadium der Umsetzung befindet, kann der Grad der Überscheidung und Beeinflussung mit vorliegendem Ansatz nicht abschließend eingeschätzt werden. Unser fachübergreifender Ansatz geht allerdings über den Zwei-Stufen-Ansatz von RADAR hinaus. Trotzdem wird es wichtig sein, den zukünftigen Forstschritt und die Ergebnisse des Projekts zu beobachten.

2.3 Strukturinformationen - Projekt ‚Prospect'

Jede Publikation ist gekennzeichnet durch Strukturinformationen wie Dateityp, Dateiformat, Speicherort und assoziierte Daten / Informationen / andere Publikationen [WB09]. Neben der Möglichkeit der manuellen Eingabe dieser Informationen durch den Autor wäre es wünschenswert, diese automatisch während des Publikationsprozesses von Forschungsdaten zu erfassen. In technischer Hinsicht ist es leicht, Informationen zu Größe, Anzahl, Format und hierarchischer Struktur der Datenobjekte zu extrahieren. Die automatische Erfassung assoziierter Datenobjekte und derer Beziehung ist hingegen wesentlich schwieriger umzusetzen. Eine Idee dazu ist, Publikationen mit Hilfe einer passenden Markup Sprache basierend auf XML anzureichern. Autoren müssen dann alle relevanten Stichwort, Sätze oder Bereiche einer Publikation markieren und mit einem entsprechenden Tag versehen. Damit können passende Datenobjekte später dynamisch miteinander verknüpft werden.

Derzeit gibt es diverse Markup Sprachen für spezielle Forschungsgebiete, so zum Beispiel im Bereich der Naturwissenschaften die Chemical Markup Language, die Mathematics Markup Language und die Biology Markup Language.

Das Projekt ‚Prospect' [RSC14], initiiert durch die Royal Society of Chemistry, bietet durch eine Begriffs-Ontologie die Möglichkeit, Publikationen anzureichern. Es gibt Drop-Down Menüs zur Auswahl wenn ein vorhandener Begriff benutzt wird. Dabei stammen die Begriffe der Ontologie aus der Gene Ontology, der Sequence Ontology oder der Cell Ontology. Außerdem ist es möglich die Markierung von verschiedenen Begriffe zu konfigurieren, so dass der Leser in der Lage ist, den Text leicht zu beurteilen.

Das Problem der vorgestellten Ansätze sowohl im vorgenannten Projekt als auch bei anderen Markup Sprachen ist die fehlende Abstraktion und damit Übertragbarkeit auf andere Forschungsgebiete. Jede individuelle Markup Sprache ist nur im assoziierten

Forschungsbereich nützlich einsetzbar. Daher ist das Konzept an dieser Stelle nur sinnvoll in fachspezifischen Repositorien.

2.4 Semantische Informationen - "Scientific Publication Packages (SPP)"

Die steigende Zusammenarbeit von Forschern in den letzten Jahren über Forschungsgruppen hinweg hat neue Anforderungen erwachsen lassen, Forschungsdaten gemeinsam zu nutzen und Forschungsergebnisse nachnutzbar zu machen. Neben den reinen Forschungsdaten zählen dazu auch Informationen über Bearbeitungsschritte der Rohdaten bis hin zum publizierten Ergebnis. Die dafür benötigten Informationen gehen weit über bisher betrachtete Strukturinformationen (siehe vorhergehender Abschnitt) der Datenobjekte hinaus und können als semantische Informationen bezeichnet werden, die Forschungsdaten mit Metadaten zum wissenschaftlichen Kontext und Forschungsprozess bereichern.

[Hu06] beschreibt einen Weg, um alle benötigten Informationen, inklusive Primärdaten, zugehörigen Metadaten und Informationen zur Entstehung der Daten in sogenannten ‚Scientific Publication Packages' zusammenzufügen. Diese Pakete enthalten dann alle betreffenden Struktur- und Semantikinformationen zum jeweiligen Forschungsdatum, die Beziehungsinformationen zu anderen Daten, die Beschreibung zu Erzeugung von Forschungsergebnissen aus den Rohdaten, verknüpfte Publikationen und assoziierte Kontext-, Herkunfts- und Administrationsmetadaten. Die Zusammenfassung all dieser Informationen und Daten ermöglicht es, komplexe Strukturen als einzelnes Digitales Objekt zu betrachten. Insofern erlauben SPPs die einfache Veröffentlichung, effektiven Austausch und Verbreitung komplexer Forschungsdaten.

Die konzeptionelle Grundlage von SPPs ist eine erweitertes ABC Modell, welches wiederum eine Ontologie für wissenschaftliche Modelle, deren Ursprung und Herkunft ist. Dabei wird RDF benutzt um die Beziehungen verknüpfter Komponenten zu beschreiben. Im Gegensatz zu XML ist RDF durch den Einsatz eines graphbasierten Modells sehr gut geeignet, stark verknüpfte Ressourcen darzustellen, anzufragen und darin zu navigieren [Hu06].

Der hier vorgestellte Ansatz der SPPs verdient eine weitergehende Beobachtung, selbst wenn er sich aktuell noch in der konzeptionellen Phase befindet. Die wichtigsten Herausforderungen, mit denen auch wir uns in der Beschäftigung mit dem Forschungsdatenmanagement konfrontiert sehen, werden durch die Idee der SPPs erfüllt. Das größte Defizit ist bisher die fehlende Implementierung und / oder eine realistische Umsetzung in eine technische Infrastruktur, weshalb es für den Einsatz im SFB bisher nicht geeignet ist.

2.5 Erweiterte Publikationen – Projekte ‚Driver' + ‚Driver II'

[WB09] definieren: *"An Enhanced Publication is a publication that is enhanced with research data as evidence of the research, extra materials to illustrate or to clarify or post-publication data like commentaries and ranking."* Die beiden Driver Projekte

beschäftigen sich genau mit der Möglichkeit, solche erweiterten Publikationen zu realisieren. Forschungsergebnisse sollen mit den gewünschten Informationen vor / während der Publikation versehen werden. Dafür wurden im Projekt existierende Standards, Infrastrukturen und Konzepte evaluiert und anschließend eine Infrastruktur entwickelt, die verschiedene Konzepte integriert.

Im Gegensatz zu unseren Anforderungen werden Forschungsergebnisse in den Driver Projekten in Form von Publikationen betrachtet und nicht als Forschungsdaten, die während des gesamten Forschungsprozesses entstehen.

3 Projektziel / Nutzen

Zu Beginn des Teilprojekts ‚INF Z2' war die Ausgangssituation, wie in der Ausgangslage beschrieben. Forschungsdaten wurden / werden lokal gespeichert und sind dann nur einzelnen Wissenschaftlern zugänglich. Ganz speziell, wenn der jeweilige Mitarbeiter die Forschungseinrichtung verlässt, sind die Daten ohne ausreichende Beschreibung und Dokumentation des Entstehungsprozesses dann nicht mehr nutzbar und müssen im schlechtesten Fall erneut generiert werden.

Um diese Situation zu verbessern, soll eine lokale Lösung konzipiert und implementiert werden, die neben der reinen Speicherung von Forschungsdaten eine Reihe zusätzlicher Anforderungen erfüllt. Diese sind im Detail im Abschnitt Herausforderungen beschrieben.

Das traditionelle Konzept der Langfristspeicherung betrachtet hauptsächlich die physikalische Speicherung von Rohdaten oder der kompletten Publikation, die möglicherweise noch individuell um beschreibende Informationen ergänzt werden kann. Als Konsequenz ergeben sich zwei Richtungen für die Entwicklung unterstützender Infrastrukturen und Modelle, einerseits die physikalische, bitweise Speicherung von Daten und auf der anderen Seite die Online-Veröffentlichung von Papers, Journalbeiträgen oder ganzen Büchern.

Im Gegensatz dazu schlägt unser Ansatz eine Brücke zwischen diesen beiden Sichtweisen und das Ziel ist ein integriertes System. Der komplette Forschungsprozess wird unterstützt hinsichtlich der Aufbewahrung / Speicherung von Primärdaten über die Beschreibung transformierender Ereignisse bis hin zur Publikation von Ergebnissen. Alle im Zuge dieser Prozesskette entstehenden Datenobjekte können gespeichert und mit beschreibenden Informationen versehen werden. Daher ist es am Ende möglich, die Entstehung eines Forschungsergebnisses nachzuverfolgen und damit die Qualität der Forschung zu belegen beziehungsweise zu sichern.

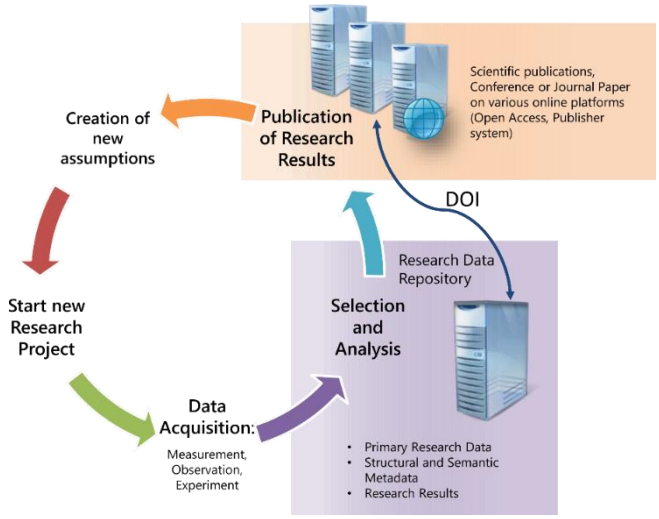

Abbildung 2: Erwünschter Lebenszyklus von Forschungsdaten im SFB 840

Wie in Abbildung 2 zusätzlich zu sehen, wird mit Hilfe der Vergabe von DOIs für relevante Datenobjekte und deren Informationen ein öffentlicher Zugang auf definierte Daten gewährt.

Eine wichtige schon im derzeitigen Status des Projekts erkennbare Erkenntnis ist, dass die Anforderungen der Forscher sehr genau identifiziert werden müssen bevor man die Umsetzung einer unterstützenden Infrastruktur sinnvoll beginnen kann.

Aus den bisherigen Gesprächen mit den Forschern des SFB 840 kann man schließen, dass die reine Langzeitspeicherung von Forschungsdaten nicht ausreichend ist. Insbesondere für Forschungsleiter und Projektmanager stellt die Reproduzierbarkeit von Forschungsergebnissen und deren Entstehung einen enormen Vorteil gegenüber bisherigen Lösungen dar. Dieser Vorteil rechtfertigt auch den zusätzlichen Aufwand für die Datenpflege und das Datenmanagement.

4 Herausforderungen

Diversität. Das Spektrum der am SFB beteiligten Forschungsfelder ist sehr breit. Entsprechend unterschiedlicher Natur sind die in den einzelnen Teilprojekten anfallenden Forschungsdaten. Da für diese hoch spezialisierten Forschungsfelder in der Regel keine überregionalen Repositorien zur Sammlung dieser Daten existieren, muss lokal die Infrastruktur dafür bereitgestellt werden. Ein hohes Abstraktionsniveau bei der Beschreibung der Forschungsdaten mit möglichst generischen Metadaten ist eine wesentliche Anforderung an die Lösung.

Reproduzierbarkeit. Neben den eigentlichen Forschungsprimärdaten sollen auch die publizierten Ergebnisse und deren kompletter Entstehungsprozess gespeichert werden.

Deshalb muss eine Architektur gewählt werden, die alle Bearbeitungsschritte bis hin zum publizierten, mit eindeutiger DOI versehenen Ergebnis abbilden kann. Ausgehend von der DOI kann der Entstehungsprozess des Ergebnisses wieder rekonstruiert werden.

Langfristspeicherung. Für die zuverlässige Verfügbarkeit und Nachnutzbarkeit der Daten muss die Speicherung den Normen der Langfristspeicherung (OAIS-Modell) genügen. Eine flexible Autorisierung muss den Zugriff auf die Daten regeln, so dass die unberechtigte Nutzung verhindert wird, Kollaborationen aber dennoch möglich sind.

5 Lösung

Kooperationen. Bei der Umsetzung wird weitgehend auf bereits bestehende Lösungen aufgebaut, weswegen intensive Kooperationen zwischen vielfältigen Partnern sinnvoll sind. Bezüglich der Strategien der Überführung der Daten in den Langfristspeicher und der Auffindbarkeit der Metadaten bestehen Kontakte zur TU München (Nutzung der Komponente MediaTUM) und der ETH Zürich (Anwendung der Docuteam-Software). Bei der Langfristspeicherung besteht eine Kooperation mit der Bayerischen Staatsbibliothek (BSB) die eine Mitnutzung der Installation der Software „Rosetta“ der Firma ExLibris ermöglicht.

Forschungsdatenportal. Im aktuellen Lösungsansatz werden alle Verarbeitungsschritte vom Rohdatum zum publizierten Ergebnis abstrakt als Knoten beschrieben, denen spezifische Daten und Metadaten zugeordnet werden können. Hierzu eignet sich das Framework MediaTUM, das eine flexible Definition von Metadatenschemata erlaubt. Der Entstehungsprozess von Forschungsergebnissen (Bilder, Grafiken, Tabellen, diverse Daten) muss ausgehend von den Primärdaten analysiert und alle Prozessschritte möglichst abstrahiert als Knoten abgebildet werden. Für die Knoten muss neben dem assoziierten Metadatenschema der Uploadprozess für die Daten definiert und implementiert werden.

Langfristige Auffindbarkeit. Für den Knoten eines publizierten Ergebnisses wird über die TIB Hannover eine DOI vergeben und somit die weltweite Sichtbarkeit des Forschungsergebnisses über DataCite gewährleistet. Alle Knoten, die zur Rekonstruktion des Ergebnisses notwendig sind, werden bei der Vergabe der DOI automatisch in Rosetta gespeichert, so dass die Langfristverfügbarkeit des Ergebnisses und dessen Genese gesichert ist.

Datenschutz. Der Schutz vor unberechtigtem Zugriff auf die Daten wird durch das Autorisierungsmodul von MediaTUM gewährleistet, das für jeden Knoten eine sehr differenzierte Zugriffssteuerung erlaubt. Neben der Anmeldung über LDAP an dem lokalen IDM können auch Benutzer im System selbst angelegt werden. Benutzer können flexibel Gruppen zugeordnet werden, wodurch Kooperationen zwischen Wissenschaftlern abgebildet werden können.

6 Status Quo

Die spezifischen Anforderungen bezüglich des Forschungsprozesses wurden durch strukturierte Interviews und einen Fragebogen evaluiert. Die jeweiligen Antworten wurden analysiert hinsichtlich Gemeinsamkeiten und Unterschieden zwischen den einzelnen Forschern und Forschungsgruppen.

Wegen der großen Bandbreite beteiligter Forschungsbereiche im SFB 840 gibt es keinen einheitlichen Forschungsprozess. Ganz im Gegenteil ist es so, dass die befragten Wissenschaftler unterschiedliche Arbeitsschritte in individueller Abfolge durchführen, um zu Ergebnissen zu kommen. Weiterhin haben sie unterschiedliche Gewohnheiten in Bezug auf Speicherung und Kommentierung von Daten, selbst innerhalb einer Forschungsgruppe.

Aus diesem Grund ist der erste Schritt des Teilprojekts, die Umsetzung der Anforderungen in einem Prototyp des Datenrepositoriums, basierend auf der MediaTUM Plattform.

Beschriebenes Vorgehen beinhaltet dabei folgende Schritte:

1. Analyse schon existierender Veröffentlichungen im Forschungsbereich
2. Definition und Implementierung notwendiger Objekttypen im Repositorium
3. Definition und Implementierung zugehöriger Metadaten-Schemata
4. Testen des Uploads für definierte Objekttypen

6.1 Analyse existierender Veröffentlichungen

Der erste Schritt ist die Analyse dreier schon existierender Veröffentlichungen aus verschiedenen Bereichen des SFB. Dabei liegt das Hauptaugenmerk auf der Erfassung publizierter Forschungsergebnisse und deren Klassifikation (Daten, Bilder, Grafiken, Tabellen, Text). Anschließend muss deren Entstehung aus Rohdaten Schritt für Schritt nachvollzogen und dokumentiert werden.

6.2 Definition und Implementierung notwendiger Objekttypen

Aus der Analyse ergeben sich alle Knoten des Forschungsprozesses, die im Datenrepositorium gesichert werden müssen. Diese werden anschließend in diskrete Typen (beispielsweise Primärdaten, Transformation, Ergebnis) unterteilt, die ähnliche oder gleiche Charakteristiken aufweisen. Auf Basis dieser Klassifikation werden neue Objekttypen im System implementiert, was durch die Nutzung von MediaTUM sehr flexibel möglich ist. Die Implementierung umfasst neben der Definition standardisierter Objektattribute, die Darstellung des Objekttyps und zugeordneter Daten und Informationen, sowie automatisierte Prozess während des Uploads von Daten dieses Objekttyps.

6.3 Definition und Implementierung zugehöriger Metadaten-Schemata

Ein Metadaten-Schema stellt einen Filter zur Darstellung ausgewählter Metadaten eines speziellen Objekttyps dar. Während die tatsächlichen Metadaten immer komplett in der Datenbank gespeichert werden, kann die Darstellung an den jeweiligen Anwendungsfall angepasst werden. Beispielsweise ist es nicht notwendig in einer Überblicksansicht alle Metadaten eines Objekts anzuzeigen, wohingegen die Bearbeitungsmaske für die Metadaten alle veränderbaren Metadatenfelder benötigt. Daher besitzt jeder Objekttyp diverse Metaschemata je nach Anwendungsfall.

7 Ausblick

Die aktuell anstehende Herausforderung im Projekt ist die Implementierung der Beziehungen zwischen verschiedenen Datenobjekten. Diese Implementierung muss konform mit allen anderen Anforderungen erfolgen. Im Besonderen muss die Nachvollziehbarkeit der Entstehung und Transformation beteiligter Forschungsdaten berücksichtigt werden. Weiters müssen zielgerechte und passende Darstellungsformen dieser Beziehungen im User Interface entwickelt werden.

Literaturverzeichnis

[Ba10] Baru, C.K. (2010): Integration or mashup? Challenges in bringing together heterogeneous scientific data - In: Curdt, C. & Bareth, G. (eds.): Proceedings of the Data Management Workshop, 29. 30.10.2010, University of Cologne, Germany. Kölner Geographische Arbeiten, 90:1-6, Köln, doi: 10.5880/TR32DB.KGA90.2.

[DFG06] Deutsche Forschungsgemeinschaft DFG (2006): Position paper: "Scientific Library Services & Information Systems – Funding priorities through 2015". http://dfg.de/download/pdf/foerderung/programme/lis/pos_papier_funding_priorities_2015_en.pdf. 2014-10-28.

[Ef10] Effertz, E. (2010): The Funder's Perspective: Data Management in coordinated programmes of the German Research Foundation (DFG) - In: Curdt, C. & Bareth, G. (eds.): Proceedings of the Data Management Workshop, 29. 30.10.2010, University of Cologne, Germany. Kölner Geographische Arbeiten, 90:35-38, Köln, doi: 10.5880/TR32DB.KGA90.7.

[EFKR10] Eichler, M., Francke, T., Kneis, D., Reusser, D.E. (2010): The GOLM-Database Standard – A framework for time series data management based on free software - In: Curdt, C. & Bareth, G. (eds.): Proceedings of the Data Management Workshop, 29. 30.10.2010, University of Cologne, Germany. Kölner Geographische Arbeiten, 90:39-44, Köln, doi: 10.5880/TR32DB.KGA90.8.

[GRBIO14] GRBIO: The Global Registry of Biorepositories. http://grbio.org. 2014-10-27.

[HGP14] HGP: The Human Genome Project. http://www.genome.gov/10001772. 2014-10-27.

[Hu06] Hunter, J. (2006): Scientific Publication Packages – A Selective Approach to the Communication and Archival of Scientific Output. - The International Journal of Digital Curation, 1 (1): 33-52, doi: doi:10.2218/ijdc.v1i1.4.

[KSS09] Karstensen Elbaek, M., Schmeltz Pedersen, G. & Sierman, B. (2009): New Technologies and Communities - In: Vernooy-Gerritsen, M. (ed.): Emerging Standards for Enhanced Publications and Repository Technology – Survey on Technology, 19-106.

[LE13] Ludwig, J. & Enke, H. (eds.) (2013): Leitfaden zum Forschungsdaten-Management – Handreichungen aus dem WissGrid-Projekt. 01|G09005A-G, Ispra: BMBF.

[Ne12] Neuroth, H. (2012): Vorgehensweise - In: Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. & Ludwig, J. (eds.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, 51-59.

[OECD07] OECD (2007): OECD Principles and Guidelines for Access to Research Data from Public Funding. http://www.oecd.org/science/sci-tech/38500813.pdf. 2014-10-27.

[OSN12] Oßwald, A., Scheffel, R. & Neuroth, H. (2012): Langzeitarchivierung von Forschungsdaten: Einführende Überlegungen- In: Neuroth, H., Strathmann, S., Oßwald, A., Scheffel, R., Klump, J. & Ludwig, J. (eds.): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, 13-21.

[PWRW12] Potthoff, J., van Wezel, J., Razum, M. & Walk, M. (2012): Anforderungen eines nachhaltigen, disziplinübergreifenden Forschungsdaten-Repositoriums. https://www.dfn.de/fileadmin/3Beratung/DFN-Forum7/konferenzband/02-Anforderungen_eines_nachhaltigen__disziplinuebergreifenden_Forschungsdaten-Repositoriums.pdf. 2014-08-05.

[RSC14] Royal Society of Chemistry (2014): Project "Prospect" - Linking compounds and concepts in articles. http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp. 2014-10-29. http://www.rsc.org/Publishing/Journals/ProjectProspect/Examples.asp. 2014-10-29.

[TIB13] TIB, FIZ Karlsruhe, LMU, IPB & KIT/SCC (2013): RADAR – Research Data Repositorium. DFG-Antrag. http://www.radar-projekt.org/display/RD/Projektantrag. 2014-10-29.

[TH08] Treloar, A. & Harboe-Ree, C. (2008): Data management and the curation continuum: how the Monash experience is informing repository relationships. http://valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf 2014-08-20.

[Ve09] Verhaar, P. (2009): Object Models and Functionalities - In: Vernooy-Gerritsen, M. (ed.): Enhanced Publications – Linking Publications and Research Data in Digital Repositories, 92-131.

[WB09] Woutersen-Windhouwer, S. & Brandsma, R. (2009): Enhanced Publications, State of the Art - In: Vernooy-Gerritsen, M. (ed.): Enhanced Publications – Linking Publications and Research Data in Digital Repositories, 19-91.