Langzeitarchivierung von Forschungsdaten im Sonderforschungsbereich 840 „von partikulären Nanosystemen zur Mesotechnologie“ an der Universität Bayreuth

Johannes Fricke, Dr. Andreas Weber und Dr. Andreas Grandel

Universität Bayreuth
IT-Servicezentrum
Gebäude NW II
Universitätsstraße 30
95447 Bayreuth

johannes.fricke@uni-bayreuth.de
andreas.weber@uni-bayreuth.de
andreas.grandel@uni-bayreuth.de

Abstract: Im Rahmen des Infrastruktur-Teilprojekts Z2 des Sonderforschungsbereichs 840 (SFB840) hat das IT-Servicezentrum (ITS) der Universität Bayreuth die Aufgabe, ein Forschungsdatenportal zu konzipieren und bereitzustellen. In dem Forschungsdatenportal sollen alle forschungsrelevanten Daten, sowie die für eine Nachnutzung notwendigen Metadaten gespeichert und archiviert werden. Die eindeutige, persistente Referenzierbarkeit der Daten soll über zertifizierte DOIs (Digital Object Identifier) sichergestellt werden. Besonderes Augenmerk soll dabei auf die langfristige Verfügbarkeit der Daten gelenkt werden. Dabei soll auf existierende, möglichst freie Softwarelösungen zurückgegriffen werden, die an den spezifischen Bedarf des SFB angepasst werden. Diese Lösung soll später die Grundlage für die Speicherung von Forschungsdaten auch aus anderen Bereichen bilden.

1. Ausgangslage

Die nachhaltige Sicherung und die langfristige Zugänglichkeit von Forschungsdaten gewinnen für die Wissenschaft zunehmend an Bedeutung. Die wichtigsten Gründe dafür sind die Forderung nach Transparenz der Forschungsergebnisse und das Potential, das in der Nachnutzung bereits vorhandener Ergebnisse gesehen wird. Der offene Zugang zu

den wissenschaftlichen Ergebnissen wird in vielen Erklärungen der Wissenschaft gefordert [Be10]. Die Deutsche Forschungsgemeinschaft verlangt deshalb seit 2010 bei der Beantragung von Projekten eine Darstellung des Umgangs mit den wissenschaftlichen Ergebnissen [DF11]. Im Rahmen eines Teilprojektes des Sonderforschungsbereiches 840 (SFB840) [Z2] hat das IT-Servicezentrum (ITS) der Universität Bayreuth die Aufgabe, Möglichkeiten für die langfristige Speicherung und Wiederauffindbarkeit der forschungsrelevanten Daten des SFB840 zu konzipieren und bereit zu stellen. Ein Forschungsdatenportal, in dem die relevanten Metadaten abgelegt und abrufbar sind, soll den Forschungsprozess der im SFB beteiligten Wissenschaftler unterstützen. Auf die langfristige Verfügbarkeit der Forschungsprimärdaten, vorwiegend Bild- und Messdaten, wird besonderes Augenmerk gelegt. Es soll somit gewährleistet werden, dass die Daten über einen längeren Zeitraum zur Verfügung stehen und nicht in den Speicherorten der Labors durch Umzug oder Personalveränderungen verloren gehen.
Der Zugriff auf die Forschungsdaten soll über einen weltweit eindeutigen Identifier erfolgen. Über diesen Identifier können die Daten auch mit den Publikationen verknüpft werden und umgekehrt. Die Forschungsdaten werden somit zukünftig auch bei der Aufnahme der Publikationen in den Nachweiskatalogen verzeichnet. Ebenfalls kann zu den Forschungsdaten abgelegt werden, in welchen Publikationen die Daten eingeflossen sind. Die Forschungsdaten können somit auch externen Interessenten (z. B. Referees, anderen Arbeitsgruppen) zugänglich gemacht werden.
Die Beschreibung der Forschungsergebnisse mit geeigneten Metadaten ist ein wichtiger Bestandteil des Projektes, für den die Universitätsbibliothek als Partner gewonnen werden konnte. Das dort vorhandenen Wissen zu Metadaten und die langjährige Erfahrung in der Erschließung und Beschreibung von Informationsquellen soll genutzt werden, in Zusammenarbeit mit den Wissenschaftlern gute und praktikable Metadatenschemata zu entwickeln.
Der SFB840 überspannt viele Fachbereiche der Physik und der Chemie. Deshalb muss der Ansatz hinsichtlich der Art der gespeicherten Daten und der Beschreibungen der Daten flexibel sein. Die Erweiterbarkeit der Lösung auf die Archivierung von Forschungsdaten über den SFB hinaus sollte deshalb grundsätzlich möglich sein und für alle Wissenschaftler der Universität Bayreuth nutzbar sein.

2. Stand des Projekts

2.1 Beschreibung der Aufgabe

Die grundsätzlichen Komponenten für eine langfristige Speicherung von Forschungsdaten und deren Veröffentlichung ist in der, dem Antrag entnommenen, Abbildung 1 dargestellt. Die technische Umsetzung soll in diesem Papier diskutiert und konkretisiert werden.

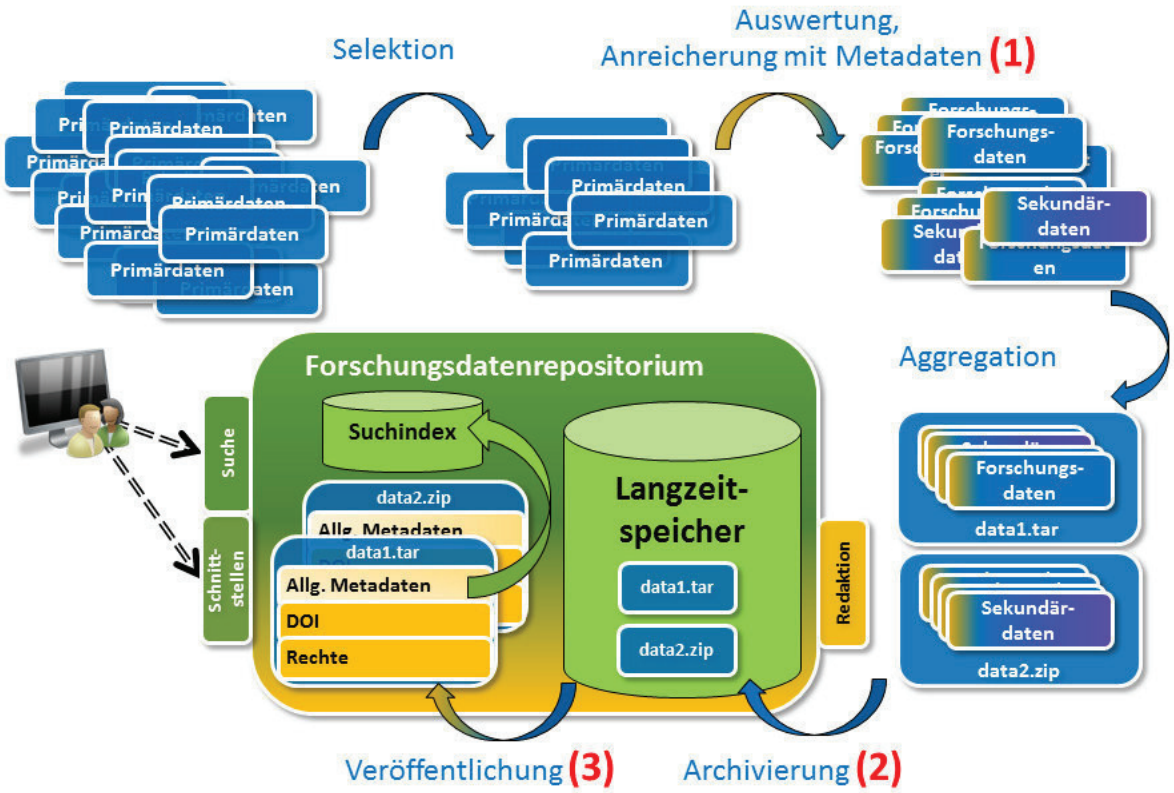

Abb. 1: Notwendige Veränderung im Umgang mit Forschungsdaten: (1) Anreicherung mit allgemeinen und fachlichen Metadaten; (2) Transfer in ein Langzeitarchiv; (3) Veröffentlichung und Referenzierbarkeit

Erläuterungen zu der Abbildung 1:

(1) Nach der Selektion der relevanten Daten werden die Primärdaten mit geeigneten Metadaten versehen, damit sie später lokalisiert und interpretiert werden können. Die allgemeinen Metadaten ermöglichen die Zuordnung der Daten zu einer Forschergruppe und einem Projekt und werden in dem Portal für das Auffinden und die Selektion der Datensätze verwendet. Die fachlichen Metadaten ermöglichen die wissenschaftliche Interpretation, Verifikation und Nachnutzung der Daten. Die Vergabe der Metadaten geht im Umfang über die Abbildung eines Laborbuches hinaus, da die Ergebnisse auch für Personen nachvollziehbar sein sollen, die den Versuchsaufbau, die verwendeten Programme und die verwendete Ausstattung nicht kennen. Die bei der Auswertung und Weiterverarbeitung der Primärdaten entstehenden „Sekundärdaten“ (Grafiken, Diagramme etc.) werden ebenfalls mit diesen Informationen versehen.

(2) Die Daten verbleiben langfristig nicht mehr auf den Speichersystemen der Forschergruppen, sondern werden in einem hochverfügbaren und zuverlässigen Langzeitarchiv abgelegt. Damit wird sichergestellt, dass die Daten über mindestens 10 Jahre verfügbar bleiben, selbst wenn sich die Struktur oder die Organisation der Forschungsgruppen ändern (z. B. Weggang von Forschern etc.). Der Zugriff auf die Daten muss für die Forscher weiterhin jederzeit möglich sein. Im einfachsten Fall werden die Daten in ein Netzlaufwerk verschoben (ähnlich

zu Dropbox), aber auch die für die Übertragung von Daten im Internet üblichen Protokolle (SFTP, SCP, CIFS etc.) können verwendet werden. Im Projekt wird die Nutzung entfernter, zentraler Langzeitspeicher konzeptionell betrachtet.

(3) Der Zugriff auf die Daten soll für externe Interessierte (z. B. Referees, andere Forschergruppen) möglich sein. Der Zugriff erfolgt über das Repository, das neben der Suchmöglichkeit in den allgemeinen Metadaten auch eine Prüfung der Zugriffsberechtigung bereitstellt. Letztere stellt sicher, dass die Daten erst nach der Veröffentlichung oder nach der Erteilung von Patenten für Externe zugänglich werden, vorab aber schon in Kooperationen (z. B. innerhalb des SFB) bereit stehen. Das Portal stellt somit auch eine Kollaborationsplattform für die Wissenschaftler dar. Die eindeutige, persistente Referenzierbarkeit der Daten wird über zertifizierte Identifikatoren erreicht. Hierzu soll der bereits seit 2005 etablierte DOI-Dienst der TIB Hannover genutzt werden, der im internationalen Kontext von DataCite errichtet wurde. Über die DOIs (Digital Object Identifier) können die Datensätze mit den Veröffentlichungen verknüpft werden und umgekehrt den Daten die Identifier der elektronischen Veröffentlichung zugeordnet werden. Über geeignete Schnittstellen werden die Metadaten für andere Systeme (z. B. übergreifenden Fachrepositorien) zugänglich gemacht.

2.2 Langzeitarchivierung

Seit Beginn des Projektes wurde insbesondere die Langfristspeicherung von Daten genauer betrachtet und eine Reihe von Veranstaltungen zu diesem Thema besucht [Bi13]. Die Evaluation ergab ein paar grundlegende Ergebnisse, die nachfolgend kurz erläutert werden.

Der OAIS-Standard (Open Archival Information System) [Spac] sollte bei der Langzeitarchivierung eingehalten werden. Das System zur Speicherung der Daten muss OAIS-konform sein, so dass verschiedene Lösungen darauf aufsetzen können. Damit wird auch eine Austauschbarkeit zwischen verschiedenen Systemen erreicht.

Wegen der Langfristigkeit der Aufgabe sollte bei der Langzeitarchivierung die Verwendung eigener Lösungen, deren Fortbestand an einzelne Personen geknüpft ist, vermieden werden. Es sollten am Markt etablierte Lösungen eingesetzt werden.

Die Beschreibung der Datensammlungen mit Metadaten sollte nach einem Standard, z.B. METS (Metadata Encoding & Transmission Standard) erfolgen [MET]. Die Metadaten müssen projektabhängig definiert werden können.

Die Sicherstellung der Integrität der Daten auf Bit-Ebene und automatische Formatkonversionen sollten durch das Speichersystem zur Verfügung gestellt werden. Die Verwaltung dieser Aufgaben sollte in dem System verankert sein.

Wegen der großen Datenmengen eignen sich einige der im Bereich der Verwaltung von digitalen Publikationen etablierte Open-Source Lösungen nicht, z.B. LOCKSS (Lots Of Copies Keep Stuff Safe) [LOC].

Die Daten sollen einfach zitiert werden können. Dazu sind geeignete Identifier zu erstellen.

Der Aufwand den die Forscher zur Aufbereitung der Daten für die Langzeitarchivierung betreiben müssen, insbesondere die Vergabe von Metadaten und der Transfer in das Archivsystem, muss gering sein. Deshalb müssen komfortable Werkzeuge für diesen Vorgang bereitgestellt werden.

Die Bereitschaft der Forscher, die Forschungsdaten als Open-Source bereit zu stellen ist sehr gering. Die Einsicht in die Daten soll aus Sicht der Forscher eher selektiv freigegeben werden. Deshalb ist ein ausgereiftes Rechtekonzept notwendig.

Die meisten Forscher kennen die Verpflichtungen zur Aufbewahrung der Daten nicht oder nicht ausreichend. Deshalb wird empfohlen, dass neben der technischen Umsetzung geeignete Richtlinien für den Umgang mit Forschungsdaten erstellt werden. Die Archivierung ist aus Sicht der Forscher nur für einen begrenzten Zeitraum notwendig, es besteht grundsätzlich nicht der Anspruch der „unendlich langen“ Bereithaltung.

Auf der Grundlage dieser Bedingungen wurde nach einem geeigneten System gesucht. Ein etabliertes Produkt im Bereich Langzeitspeicherung ist Rosetta [ExLi] von ExLibris, das in der Bayerischen Staatsbibliothek (betrieben vom Leibniz-Rechenzentrum) bereits im Bereich Digitalisierung im Einsatz ist. Rosetta bildet auch die Grundlage für die Langzeitspeicherung von Forschungsdaten an der ETH Zürich. Dort wird der Einsatz von Rosetta über das Problemfeld Forschungsdaten hinaus angestrebt, was in der Abbildung 2 aufgezeigt ist. Es wird versucht, die Bewahrung aller anfallenden digitalen Daten in einem Konzept zu vereinen und auf einer gemeinsamen Plattform umzusetzen. Die grundlegende Problematik bei der Langzeitarchivierung ist in den verschiedenen Bereichen gleich, die Vorgänge unterscheiden sich nur durch die Auswahl der Objekte und die Beschreibung mit Metadaten.

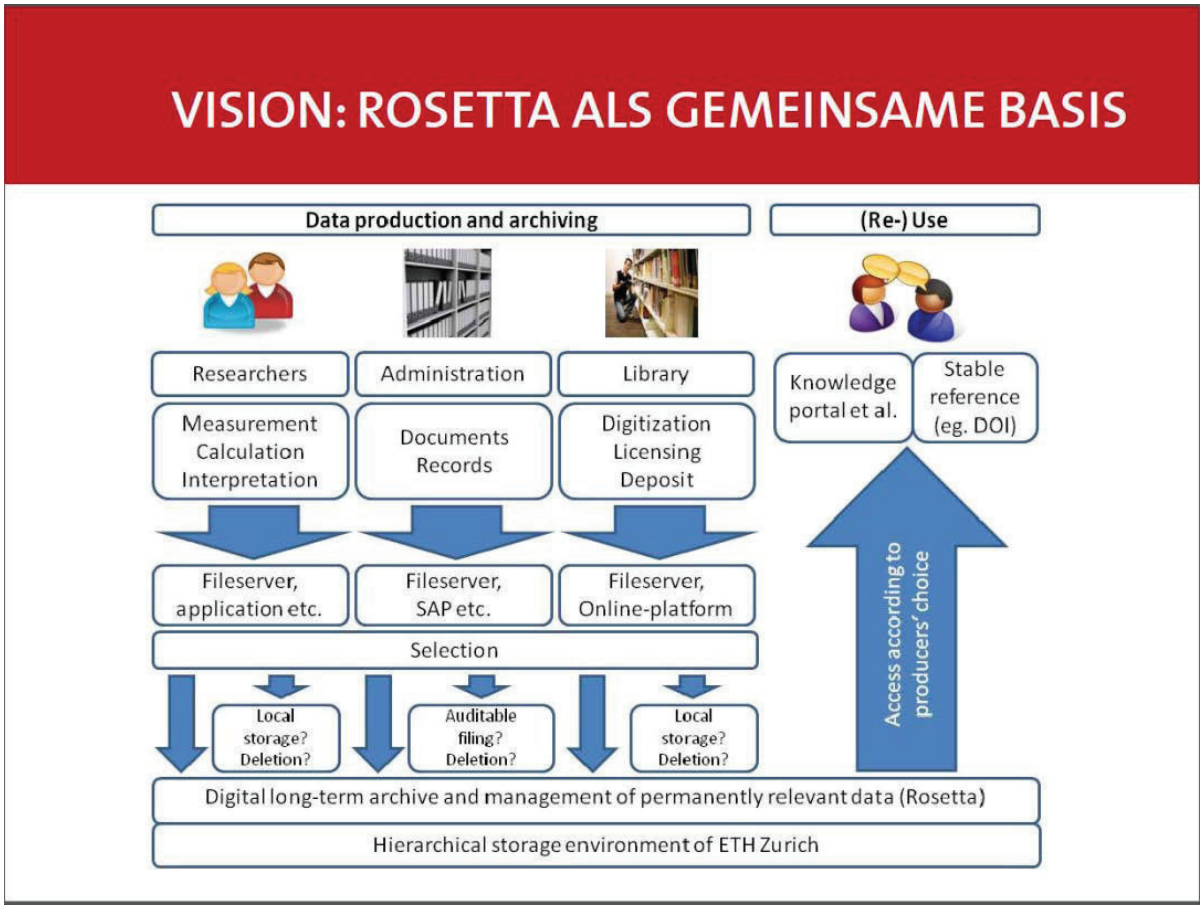

Abb. 2**: Ziel der ETH Zürich** (aus Töwe, Matthias. Forschungsdaten an der ETH Zürich, 2010 http://dx.doi.org/10.3929/ethz-a-007590685)

2.3 Transfer der Daten, Pflege der Metadaten

Zur Ablieferung der Daten durch die Forscher wird, an der ETH Zürich unter anderem auch die Software „Packer“ (ehemals docupack, vgl. Abb. 3) der Firma Docuteam [Docu] verwendet. Packer liefert standardisierte SIPs (Submission Information Package), woraus am Archivierungssystem durch das Programm Feeder (ehemals documill, vgl. Abb. 3) für Rosetta konforme AIPs (Archival Information Package) erzeugt werden. Das Zusammenspiel dieser Komponenten ist in der Abbildung 3 aufgezeigt.

Mit Packer werden die Aufgaben der Aggregation und die Anreicherung mit Metadaten, sowie der Transfer zum Langzeitarchiv abgedeckt. An dieser Stelle muss nun noch als externer Prozess die Erstellung einer DOI eingefügt werden. Die Einstellung der Daten und der Metadaten in das Archiv erfolgt über Feeder. Die Archivierung und die Bereitstellung der Daten erfolgt über Rosetta.

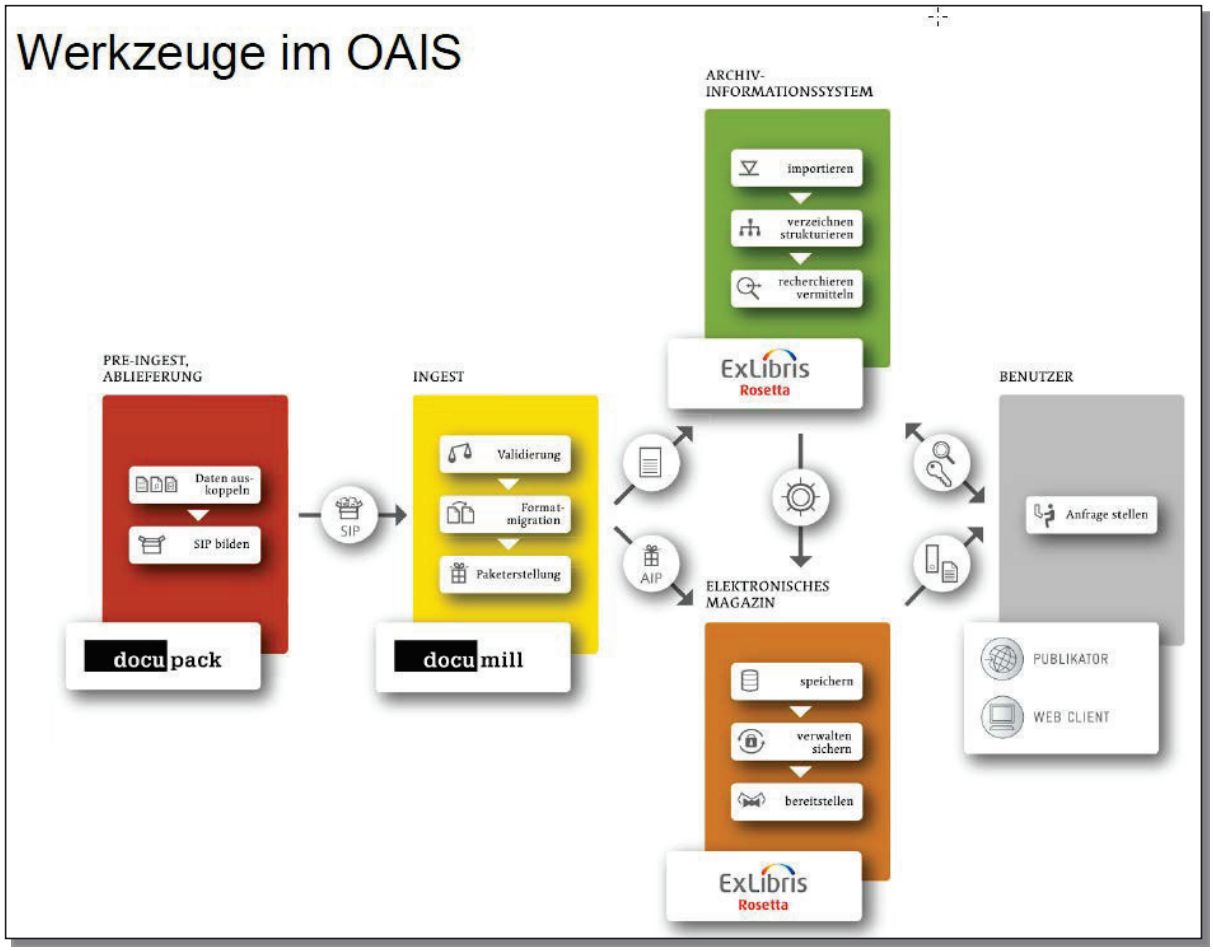

Abb 3: Eingesetzte Komponenten an der ETH Zürich (aus Strukturierung und Aufbereitung von Forschungsdaten fürs Archiv: Konzepte und Werkzeuge - Andreas Nef, Docuteam GmbH)

Diese Vorgehensweise stellt eine einfache Möglichkeit dar, Daten in das System zur Langfristspeicherung einzubringen. Eine Suche in den Metadaten oder eine nachträgliche Pflege der Metadaten wird damit aber noch nicht möglich.

Die Evaluation vorhandener Plattformen zur Erfassung und Verwaltung von Metadaten zu Forschungsdaten hat gezeigt, dass es bereits einige solcher Systeme gibt. Diese sind aber meistens für spezielle Anwendungsfälle ausgelegt. Die verwendeten Metadaten sind für alle Datensätze gleich und die Bearbeitung speziell dafür optimiert [DWB]. Eine Veränderung oder ganz flexible Gestaltung der Metadaten ist in diesen Systemen in der Regel nicht vorgesehen.

Das Problem der flexiblen Beschreibung von Objekten ist im Bereich der Bereitstellung von digitalen Sammlungen bereits bekannt. Auch hier ist die Beschreibung abhängig von den Objekten, die angeboten werden. Neben einer Reihe kommerzieller Programme gibt es die an der TU München entwickelte Software MediaTUM [Med]. Es handelt sich dabei um eine Open-Source-Software, die im Rahmen des DFG-Projekts IntegraTUM von der Universitätsbibliothek der TU München in enger Kooperation mit den Fakultäten entwickelt wurde. Die Metadaten sind hier frei definierbare und flexible Workflows für verschiedene Aufgaben erstellbar. Damit steht ein Portal zur Verfügung, in welchem

die Daten übersichtlich dargestellt, mit Vorschaubildern versehen werden und die Metadaten durchsucht und angezeigt werden können. Weiterhin ist eine Schnittstelle zur Anbindung eines Archivsystems bereits implementiert. Der hohe Grad an Modularisierung von mediaTUM erlaubt den einfachen Austausch einzelner Komponenten und damit scheint mit MediaTUM auch für die Beschreibung von Forschungsdaten geeignet. Werden auch die digitalen Sammlungen mit MediaTUM verwaltet, ergäbe sich damit zudem ein Synergieeffekt. Weiterhin ist zu erwähnen, dass die Universitätsbibliothek der Hochschule der Bundeswehr in Neubiberg, die Universitätsbibliothek der Universität Augsburg und die Katholische Universität Eichstätt bereits mediaTUM für die Verwaltung von digitalen Sammlungen einsetzen. Somit sollte die Weiterentwicklung langfristig gesichert sein und es sollten weitere Kooperationspartner zur Verfügung stehen.

3. Aussichten

Die Evaluation der Komponenten, die als Basis für die Verwaltung von Forschungsdaten in Hinblick auf die Langfristspeicherung eingesetzt werden sollen, ist weitgehend abgeschlossen.

Die Bayerische Staatsregierung hat die Software Rosetta für die Nutzung durch die Bayerischen Hochschulen lizenziert [DiLa]. Für die langfristige Speicherung der Forschungsdaten des SFB 840 wird somit der Einsatz von Rosetta angestrebt. Zunächst wird hier auf die Installation in der BSB zurückgegriffen, ein späterer lokaler Einsatz von Rosetta wäre im Falle von Engpässen in den Ressourcen, z.B. Datenübertragung ebenfalls möglich. Der Umgang mit Rosetta wird derzeit genauer betrachtet.

Die Komponenten der Firma Docuteam sollen für den Transfer nach Rosetta eingehender evaluiert werden. Gespräche über eine Zusammenarbeit hierzu haben mit der BSB bereits stattgefunden. Mit diesen Komponenten wäre ein einfacher Transfer der Daten in den Langfristspeicher realisiert und damit eine einfache Basislösung gefunden.

Für den Aufbau eines Forschungsdatenportals wird der Einsatz von MediaTUM eingehend in einer lokalen Testinstallation der Software evaluiert. Gespräche mit den Entwicklern von MediaTUM haben bereits die Grundlage für eine mögliche Kooperation geschaffen. Eine wichtige Aufgabe wird es sein, MediaTUM und Rosetta in geeigneter Weise miteinander zu verbinden. Die Einbindung der DOI-Generierung ist ebenfalls zeitnah zu betrachten.

Die Erarbeitung von Metadatenschemata wird anhand einzelner Experimente exemplarisch durchgeführt. Hierzu sind intensive Gespräche mit den Wissenschaftlern notwendig. Ebenfalls wird versucht, einfache Werkzeuge für einen automatisierbaren Transfer der Metadaten zu entwickeln. Die Speicherung der Daten von Spektrometern soll dabei die Pilotanwendung darstellen.

Literaturverzeichnis

[Be10] Berliner Erklärung über den offenen Zugang zu wissenschaftlichem Wissen, 2003; OECD Declaration on Access to Research Data from Public Funding, 2004, Allianz der deutschen Wissenschaftsorganisationen, Grundsätze zum Umgang mit Forschungsdaten, 2010

[Bi13] Bibliothekartag 2013, Session Forschungsdaten-Repositorien - Infrastrukturen zur dauerhaften Zugänglichkeit von Forschungsdaten, 11.03.2013, http://www.bid-kongress-leipzig.de/t3/index.php?id=26;

Tagung: Digitale Langzeitarchivierung an Hochschulen, http://www.ub.hu-berlin.de/de/ueber-uns/veranstaltungen/tagung-digitale-langzeitarchivierung-an-hochschulen;

Treffen der INF-Projekte, 11.4.2013, Göttingen

[DF11] DFG-Antrag: Leitfaden für die Antragstellung, 2011. http://www.dfg.de/formulare/54_01/54_01_de.pdf.

[DiLa] Pressemitteilung „Digitales Langzeitgedächtnis für die bayerischen Hochschulen: http://www.bayern.de/Pressemitteilungen-.1255.10476313/index.htm

[Docu] Hauptseite der Firma Docuteam GmbH aus Baden-Dättwil, Schweiz. Motto: „Informationen strukturieren und erhalten“: http://www.docuteam.ch/

[DWB] Hauptseite des Deutschen Klimarechenzentrums zum Thema Langzeitarchivierung: http://www.dkrz.de/daten/langzeitarchivierung?set_language=de;

Hauptseite der Diversity Workbench, einer virtuellen Forschungsumgebung für Biodiversität: http://diversityworkbench.net/Portal/Main_Page

[ExLi] Internetseite der ExLibris Gruppe über Rosetta / Langzeitarchivierung unter der Überschrift „Eine neue Methode zum Erhalt des kulturellen Erbes und kumulierten Wissens“: http://www.exlibrisgroup.com/de/default.asp?catid=%7B12040B74-9794-451D-9369-28C0F2980F1B%7D

[LOC] Hauptseite von „Lots of Copies Keep Stuff Safe“ des LOCKSS Programms, welches in Zusammenarbeit mit den Stanford Universitätsbibliotheken betrieben wird: http://www.lockss.org/

[Med] Hauptseite von mediaTUM: https://mediatum.ub.tum.de/

[MET] Hauptseite des METS-Standards der Library of Congress (LOC): http://www.loc.gov/standards/mets

[Spac] Space data and information transfer systems -- Open archival information system (OAIS) -- Reference el http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284

[Z2] Projekt Z2 im Fortsetzungsantrag zum Sonderforschungsbereich 840 „von partikulären Nanosystemen zur Mesotechnologie“ http://www.sfb840.uni-bayreuth.de/de/z-projects/index.html