Von der Datenablage zum Serviceangebot Langzeitarchivierung von Forschungsdaten an der ETH-Bibliothek

Vortrag am 101. Deutschen Bibliothekartag
Hamburg, 24. Mai 2012

ETH-Bibliothek
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

INHALT

1 Abstract

Digitale Forschungsdaten stellen ein wesentliches Ergebnis des wissenschaftlichen Forschungsprozesses dar und können als Grundlage für neue Arbeiten dienen. Damit die Nachvollziehbarkeit des Forschungsprozesses im Sinne der guten wissenschaftlichen Praxis gesichert ist und eine Wiederverwendung von Daten möglich bleibt, müssen verschiedene Voraussetzungen geschaffen werden. Aufbauend auf der sicheren Speicherung der Originaldaten muss mit aktiven Massnahmen sichergestellt werden, dass die Daten über lange Zeiträume hinweg zumindest wiedergegeben werden können. Gleichzeitig müssen die Daten aus fachlicher Sicht so umfassend dokumentiert sein, dass eine wissenschaftlich seriöse Weiterverwendung erfolgen kann.

Während die fachliche Dokumentation nur durch die Datenproduzenten gewährleistet werden kann, sollten diese von den weiteren Management- und Erhaltungsprozessen möglichst entlastet werden. Als technisch-naturwissenschaftliche Hochschule mit hoher Forschungsaktivität sieht sich die ETH Zürich mit besonderen Herausforderungen konfrontiert. Die Schulleitung hat daher die ETH-Bibliothek beauftragt, die digitale Langzeitarchivierung von Forschungsdaten voranzutreiben.

Der Vortrag zeigt, wie sich dieses Mandat in das Umfeld an der ETH Zürich einfügt, insbesondere gegenüber den Forschenden sowie gegenüber den Informatikdiensten als Partner für die Speicherung und andere Dienste. Dabei soll verdeutlicht werden, wie sich die Rolle der Bibliothek aus strategischer Sicht verändert und welche Anforderungen dies an die interne Verankerung eines solchen Projekts in der Bibliothek stellt.

Die ETH-Bibliothek setzt die Software Rosetta (Ex Libris) ein. Anhand der Erfahrungen aus aktuellen Pilotprojekten werden konkrete Bedürfnisse der Forschenden erläutert und die bis jetzt erarbeiteten Lösungen vorgestellt. Dabei werden die Randbedingungen in den Forschungsgruppen und die daraus resultierenden Möglichkeiten und Grenzen für die praktische Umsetzung angesprochen.

2 Ausgangslage

Seit inzwischen mindestens zehn Jahren beschäftigt sich die ETH-Bibliothek mit Fragen der digitalen Langzeitarchivierung. Anfänglich konzentrierten sich die Überlegungen vor allem auf typische Bibliotheksdokumente wie lizenzierte Verlagszeitschriften, nativ digitale Hochschulpublikationen und Digitalisate[1]. Forschungsdaten wurden dabei zunächst explizit oder implizit ausgeklammert.
In der Zwischenzeit ist das Bewusstsein für die Risiken bei der Aufbewahrung von digitalen Daten gewachsen und innerhalb der Hochschule ist die digitale Langzeitarchivierung von Forschungsdaten zu einem eigentlichen Thema geworden. Unter anderem spielt sie eine Rolle im

[1] Siehe z.B. Töwe, Matthias and Piguet, Arlette. Konzeptstudie E-Archiving. Version 1.2 Konsortium der Schweizer Hochschulbibliotheken; Konsortium der Schweizer Hochschulbibliotheken (2005). http://dx.doi.org/10.3929/ethz-a-004990905

Risikomanagement der ETH Zürich bei der Vermeidung von Risiken für die Reputation der Hochschule aufgrund von nicht belegbaren Ergebnissen. Folgerichtig nehmen Forschungsdaten im laufenden Projekt *Digitaler Datenerhalt* eine zentrale Position ein, während daneben wie bisher eigene Daten der Bibliothek sowie des Archivs der ETH Zürich betrachtet werden. Das verbindende Merkmal ist nun, dass diese Daten in der Regel an der ETH Zürich erzeugt wurden und nur hier vorhanden sind.
Lizenzierte Verlagsinhalte, bei denen der dauerhafte Archivzugriff über das jeweilige Lizenzende hinaus im Mittelpunkt steht (Post Cancellation Access), werden derzeit nicht berücksichtigt. Wegen der universellen Verbreitung dieser Inhalte werden zur Sicherung ihrer dauerhaften Verfügbarkeit internationale Initiativen wie LOCKSS[2] und Portico[3] geprüft, die mit ihren unterschiedlichen Ansätzen helfen können, die Belastung auf viele Schultern zu verteilen.

Die nachfolgenden Ausführungen legen das Schwergewicht auf das Teilprojekt *Forschungsdaten*.

3 Zielsetzung des Projekts

Eine eigentliche Langzeitarchivierung zur Erhaltung der Nutzbarkeit von Daten über lange Zeiträume ist naturgemäss vor allem für Forschungsdaten nötig, die sich entweder aus prinzipiellen Gründen nicht wiederbeschaffen lassen (z.B. Beobachtungsdaten) oder deren Reproduktion mit unverhältnismässig hohem Aufwand verbunden wäre. Während dieser Aspekt nur für einen Teil der an der ETH Zürich anfallenden Forschungsdaten zutrifft, gelten an der ETH Zürich zusätzlich allgemein gültige Richtlinien für Integrität in der Forschung und gute wissenschaftliche Praxis[4], die unter anderem gewisse Vorgaben zur Aufbewahrung von Rohdaten enthalten, um die Nachvollziehbarkeit insbesondere von publizierten Forschungsergebnissen zu gewährleisten. Abgesehen von reinem Speicher gibt es bisher keine geeignete Infrastruktur, die allen Forschenden zur Verfügung stehen würde, um diese Vorgabe zu erfüllen und bei Bedarf eine Langzeitarchivierung zu unterstützen.

Dementsprechend ist die gegenwärtige Praxis in den verschiedenen Forschungsgruppen recht unterschiedlich. In vielen Fällen werden Daten z.B. bei Abschluss einer Dissertation lediglich auf CD-ROM oder anderen Offline-Medien abgeliefert und von der betreuenden Professorin oder dem Professor aufbewahrt. Er oder sie muss dann die relevanten Daten heraussuchen, wenn eine Anfrage eingeht. Eine eigentliche Langzeitarchivierung findet nicht statt und meist ist auch keine Migration der Datenträger nach bestimmten Zeiträumen vorgesehen.
Auf der anderen Seite gibt es sehr wohl Gruppen mit hohem Problembewusstsein, die klare Regeln für die redundante Ablage in offenen Formaten und ein regelmässiges Umkopieren der Speichermedien festgelegt haben und diese auch umsetzen. Die Kehrseite dieser Umsicht ist der beträchtliche Zeitaufwand, der mit diesen Massnahmen verbunden ist und der nicht zwingend zu den Kernaufgaben von Wissenschaftlerinnen und Wissenschaftlern gehören müsste. In den wenigsten Fällen gibt es Personal in den Forschungsgruppen, das längerfristig und pri-

[2] Lots of Copies Keep Stuff Safe: www.lockss.org
[3] Portico: www.portico.org
[4] Richtlinien für Integrität in der Forschung und gute wissenschaftliche Praxis an der ETH Zürich vom 14. November 2007 (Stand 25. Oktober 2011);
http://www.rechtssammlung.ethz.ch/pdf/414_Integrit%C3%A4t_Forschung.pdf

mär für diese unterstützenden Aufgaben angestellt ist.
Es drängt sich daher die Frage auf, ob nicht Infrastruktureinrichtungen wie die ETH-Bibliothek geeignete Dienstleistungen in diesem Bereich anbieten könnten. Das Ziel liegt nicht darin, funktionierenden Angeboten Konkurrenz zu machen. Wenn in einem Fach Datenzentren existieren, die breit akzeptiert sind und genutzt werden, sind sie in aller Regel der richtige Ort für Daten in diesem Fach. Es gibt jedoch solche Datenzentren nicht für die ganze Breite der Fächer einer Hochschule und nicht alle Daten sind dazu geeignet, in einem zentralen und in der Regel öffentlichen Repositorium deponiert zu werden.

Das Projekt *Digitaler Datenerhalt* adressiert somit gleichermassen die Aufbewahrung von Foschungsdaten für befristete Zeiträume von mindestens zehn Jahren als auch die Langzeitarchivierung von dauerhaft relevanten Daten. Die Projektbezeichnung soll dies zum Ausdruck bringen sowie darauf hinweisen, dass für eine sinnvolle spätere Nutzung von Daten nicht nur technische Massnahmen nötig sind, sondern auch eine umfassende wissenschaftliche Dokumentation erfolgen muss. Dies schliesst eine vorbereitende Strukturierung und Beschreibung von Daten unter der Kontrolle der Forschenden ein. Explizit ausgeschlossen wird dagegen zunächst die Unterstützung eines darüber hinaus gehenden Managements von Forschungsdaten während der laufenden Bearbeitung im gesamten Forschungsprozess. Ein solches umfassendes Datenmanagement müsste dagegen zweifellos ein wesentlicher Bestandteil einer Virtuellen Forschungsumgebung sein.

4 Anforderungen der Forschenden

In verschiedenen Vorgängerprojekten wurde ein grundlegendes Verständnis der Langzeitarchivierung digitaler Daten und ihrer Herausforderungen erarbeitet. Über die besonderen Anforderungen beim Umgang mit Forschungsdaten und den aktuellen Stand an der ETH Zürich war demgegenüber wenig bekannt. Daher wurde 2010/2011 eine flächendeckende Online-Umfrage bei allen Forschungsgruppen der ETH Zürich durchgeführt und durch eine persönliche Kontaktaufnahme begleitet. Von 80% der Forschungsgruppen der ETH Zürich wurde die Umfrage beantwortet[5].

Einige Vermutungen wurden durch die Umfrage bestätigt, wobei die Antworten für die einzelnen Fächer sehr unterschiedlich ausfallen. Innerhalb einzelner Departemente besteht ebenfalls noch eine grosse Heterogenität, die zum Teil im genauen Forschungsgegenstand und in den verwendeten Methoden begründet ist.

Da die Teilnehmerinnen und Teilnehmer der Umfrage nach ihrer Bereitschaft zur Mitwirkung bei der Erarbeitung geeigneter Abläufe gefragt worden waren, konnten im Anschluss an die Umfrage zunächst vier Forschungsgruppen als Pilotpartner gewonnen werden. Bereits die Arbeit mit nur diesen vier Gruppen sowie Aussagen weiterer Interessenten führten zu den folgenden Beobachtungen.

Die Forschenden formulieren z.T. dezidiert ihre Anforderungen, die sehr individuell ausfallen können. Wo bereits eigene Lösungen bestehen, wird erwartet, dass ein Alternativangebot min-

[5] Einige Ergebnisse finden sich hier:
Töwe, Matthias and Scheid, Susanne. User expectations in archived research data. Eidgenössische Technische Hochschule Zürich, ETH-Bibliothek (2011). http://dx.doi.org/10.3929/ethz-a-006691697

destens das Gleiche leistet wie das bestehende System. Diese Vorstellungen sind mit einer Standardlösung von vorneherein nicht abzudecken, die für einen rationellen Betrieb notwendig ist. Es muss daher in Kauf genommen werden, dass nicht alle Bedürfnisse bedient werden können. Dies führt zwangsläufig zu einer gewissen Ernüchterung bei allen Beteiligten, die notwendiger Bestandteil des Klärungsprozesses für die konkrete Umsetzung im Projekt ist. Daher sollten mit der Entgegennahme von differenzierten Anforderungen keine überzogenen Erwartungen an ihre vollständige Realisierung geweckt werden.

Das Thema Langzeitarchivierung wird nicht oft explizit erwähnt, aber in vielen Fällen implizit vorausgesetzt, weil das Konzept der Langzeitarchivierung, das die Gedächtnisorganisationen in den letzten Jahren entwickelt haben, nicht bekannt ist. Häufig beschränkt sich das Verständnis auf die dauerhafte Speicherung und es gibt nur ein gewisses Unbehagen, weil diese Lösung viele Fragen offen lässt. Wie mit allen Partnern ist es daher sehr wichtig und nicht ganz einfach, zu einem gemeinsamen Verständnis und einer gemeinsamen Sprache zu kommen.

Während die Langzeitarchivierung für potentiell unbeschränkte Zeiträume nicht für alle Gruppen von grosser Bedeutung ist, wurden von sehr vielen Gruppen Bedürfnisse im Forschungsdatenmanagement angemeldet, die den Rahmen des jetzigen Projekts sprengen. Als Beispiel sei der Wunsch nach einer flächendeckenden Unterstützung von elektronischen Laborjournalen genannt, die bisher auf Ebene der Forschungsgruppen gehandhabt werden.

Konsequent weiter gedacht umfassen die Wünsche weitere Elemente zu einer Virtuellen Forschungsumgebung im weiteren Sinne, d.h. zu einer Arbeitsumgebung, in der von der optimalen Versorgung mit digitalen Informationen über die Datenproduktion- und organisation sowie die gesamte Analyse und Bearbeitung der Daten und die anschliessend Publikation bis hin zur Langzeitarchivierung alle Schritte des Lebenszyklus‘ abgedeckt werden können.
Vielversprechende Ansätze für das Datenmanagement und bereits produktive Plattformen mit einem hohen Anteil an Eigenentwicklung gibt es vor allem in den Lebenswissenschaften[6] - wobei die Langzeitarchivierung allerdings bisher kaum berücksichtigt wird.
Die ETH-Bibliothek bietet naturgemäss bisher vor allem Bausteine in der Informationsversorgung[7] und verwandten Bereichen an (z.B. bibliographische Dienstleistungen[8]), die nun um die Langzeitarchivierung ergänzt werden sollen.

Die Integration aller Elemente zu einer echten Forschungsumgebung bleibt jedoch ein Fernziel, zumal Schätzungen der Kommission Zukunft der Informationsinfrastruktur in Deutschland zeigen, welche immensen Herausforderungen allein durch die Vielzahl an Fächern entstehen, für die – wenn auch im Idealfall auf einer vielfach nutzbaren technischen Basis – je nach Zuschnitt mehrere hundert eigene Umgebungen nötig werden könnten.[9]

---

[6] Beispiele sind **B-Fabric** (http://fgcz-bfabric.uzh.ch/) des Functional Genomics Center Zurich (http://www.fgcz.ch/) sowie **openBIS** (http://www.cisd.ethz.ch/software/openBIS) des Center für Information Sciences and Databases (CISD, http://www.cisd.ethz.ch/) am Departement of Biosystems Science and Engineering (D-BSSE http://www.bsse.ethz.ch/).
[7] Zusammengefasst im Wissensportal der ETH Zürich: www.library.ethz.ch
[8] ETH e-citations (ETH Institutional Bibliography): http://e-citations.ethbib.ethz.ch/
[9] Seite B82 in: Gesamtkonzept für die Informationsinfrastruktur in Deutschland; Empfehlungen der Kommission Zukunft der Informationsinfrastruktur der Leibniz-Gemeinschaft im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder, Berlin, 2011; http://www.leibniz-gemeinschaft.de/?nid=infrastr&nidap=&print=0

5 Aufgabenteilung

Die ETH-Bibliothek betreibt diejenigen Applikationen mit eigenem Personal, die unmittelbar bibliothekarischen Bezug haben, z.B. das Wissensportal (Ex Libris Primo) und den Verbundkatalog NEBIS (Aleph) sowie Plattformen für Digitalisierungsprojekte wie e-rara.ch[10] und retro.seals.ch[11]. Dabei tritt sie vielfach auch als Dienstleister für Dritte auf, da es sich um landesweite Angebote mit zahlreichen Partnern handelt.

Hardware und Speicher werden von den Informatikdiensten der ETH Zürich bereitgestellt. Die Informatikdienste sind der zentrale IT-Dienstleister für die ETH Zürich, während die ETH-Bibliothek für die Hochschule die Federführung im Bibliotheks-, Informations- und Sammlungsmanagement hat. Diese Aufgabenteilung setzt sich so auch im Hinblick auf die Langzeitarchivierung von Forschungsdaten fort. Die ETH-Bibliothek betreibt die Anwendung Rosetta und steht in direktem Kontakt mit den Forschenden. Sie nutzt dafür die Server-, Netzwerk- und Speicherinfrastruktur der Informatikdienste sowie weitere Basisdienste wie z.B. die zentrale Benutzerverwaltung. Die Speicherumgebung wird derzeit vollständig erneuert und soll im Vollausbau weitere Möglichkeiten eröffnen, Speicher bedarfsgerecht für Objekte in der Langzeitarchivierung zuzuweisen.

6 Eine bibliothekarische Aufgabe?

Die oben beschriebenen Aufgaben im Zusammenhang mit der Aufbewahrung und Langzeitarchivierung von Forschungsdaten könnten dazu verleiten, diese sehr eng mit klassischen bibliothekarischen Aufgaben gleichzusetzen. Nicht zuletzt wird Bibliotheken von den Forschenden zugetraut, eine Rolle bei der Erhaltung digitaler Daten zu übernehmen. Hervorgehoben werden ihre institutionelle Stabilität, ihre auf Langfristigkeit angelegte Arbeitsweise sowie ihre Erfahrung im Umgang mit Metadaten.

Es gibt jedoch wichtige Gründe, die dafür sprechen, eine breitere Sicht einzunehmen, die teilweise das eigene Verständnis der Rolle der Bibliotheken beeinflussen kann und sollte: Es geht bei der Langzeitarchivierung von Forschungsdaten nicht um den bedarfsgerechten Aufbau eines kohärenten eigenen Bestandes, wie ihn Bibliotheken normalerweise anstreben, sondern darum, für „fremde Bestände“ von Kunden sinnvolle und am Bedarf orientierte Dienstleistungen zu erbringen.

Es ist offensichtlich, dass den Forschenden als Datenproduzenten eine entscheidende Rolle zukommt. Sie müssen beurteilen, welche Daten in welcher Form für welche Zeiträume verfügbar bleiben müssen und sie können Annahmen darüber treffen, wer diese Daten später für welche Zwecke nutzen wird. Es ist ebenso ihre Aufgabe, die Daten so zu dokumentieren, dass eine seriöse Wiederverwendung zumindest für Fachkolleginnen und Fachkollegen möglich bleibt.
Diese Beschreibung lässt bereits erkennen, dass die digitale Langzeitarchivierung von Forschungsdaten in der Tat typische Züge der Archivierung trägt: Die Bewertung ist von zentraler Bedeutung, ebenso das Treffen von Annahmen über zukünftige Nutzerinnen und Nutzer sowie über mögliche Fragestellungen, deren Beantwortung das archivierte Material zu einem späte-

[10] E-rara.ch (Alte Drucke aus der Schweiz): http://www.e-rara.ch/
[11] Retro.seals.ch (Schweizer Zeitschriften): http://retro.seals.ch/

ren Zeitpunkt erlauben soll. Es ist daher ausgesprochen hilfreich, dass das Archiv der ETH Zürich Teil der ETH-Bibliothek ist und Mitarbeitende des Archivs aktiv in das Projekt eingebunden sind.

Archivarinnen und Archivare versuchen seit einiger Zeit, mehr Einfluss auf den vorarchivischen Bereich zu nehmen, um die spätere Ablieferung ans Archiv möglichst unproblematisch zu gestalten. Dieser Trend verstärkt sich mit dem Übergang zu hybriden oder digitalen Verwaltungsarchiven und er wird auch bei Forschungsdaten eine Rolle spielen. Hier muss mit einer Vielzahl von möglichen Formaten gerechnet werden und jede Reduktion der Formatvielfalt im Vorwege der Langzeitarchivierung begünstigt den Erfolg der späteren Erhaltungsmassnahmen. In diesem Sinne ist es das Ziel, Forschende bereits zu einem frühen Zeitpunkt im Lebenszyklus digitaler Forschungsdaten zu beraten, welche Konsequenzen Entscheidungen im Entstehungsprozess der Daten auf deren spätere Nutzbarkeit haben.

Diese Beobachtung hat Konsequenzen für die verwendeten Werkzeuge, die verstärkt Know-how aus dem Archivbereich aufgreifen sollten und weist auch darauf hin, dass das Profil zukünftiger Datenkuratorinnen und –kuratoren nicht nur im Bereich von Forschungs- und vergleichbaren Daten stärker archivarisch geprägt sein dürfte.

7 Individuelle Wünsche vs. handhabbare Dienstleistung

Der Begriff der Forschungsdaten wird aus Projektsicht weit gefasst, da letzten Endes die Forschenden selber bestimmen, welche Daten aufbewahrt werden müssen. Daher soll es hinsichtlich der Daten*typen* zunächst keine Einschränkung geben. Es wird jedoch weiterhin angestrebt, die Vielfalt der Datei*formate* und Zeichencodierungen in Zusammenarbeit mit den Forschenden zu beschränken. Nur dann können Methoden der digitalen Langzeitarchivierung überhaupt sinnvoll eingesetzt werden.
Zudem wird erst die konkrete Zusammenarbeit mit den Forschenden zeigen, inwieweit das Verständnis der Langzeitarchivierung, das die Gedächtnisorganisationen bisher für ihre eigenen Daten gewonnen haben, auf Forschungsdaten in Naturwissenschaften und Technik übertragen werden kann und sollte. Wie mehrfach erwähnt wurde, spielen andere Aspekte offenbar eine grössere Rolle. So wurde beispielsweise von einzelnen Forschenden angemerkt, dass ihre Messgeräte schneller obsolet werden als Dateiformate. In solchen Fällen verlieren vorhandene Daten u.U. rasch an Wert und Bedeutung, und wo dies möglich ist, wird neuen Messungen z.B. mit höherer Auflösung der Vorzug vor der Verwendung älterer Daten gegeben.

Von den Forschenden wurden unterschiedliche Anwendungsfälle beschrieben, die jeweils andere Anforderungen beinhalten. Breit vorhanden ist das Bedürfnis, Daten befristet aufzubewahren, um den Ansprüchen an die Nachvollziehbarkeit der eigenen Arbeit zu genügen. In vielen Gruppen besteht der weitergehende Bedarf, alle im Zusammenhang mit einer Manuskripteinreichung und mit der späteren Publikation eines wissenschaftlichen Artikels verbundenen Daten dauerhaft verfügbar zu halten. Ein wichtiger Gesichtspunkt dabei ist die Zitierbarkeit mit Hilfe eines Digital Object Identifiers (DOI), der vom DOI-Desk[12] der ETH Zürich im Rahmen von DataCite[13] für verlässlich gespeicherte Daten registriert wird. In anderen Fällen stehen weitgehend unbearbeitete Rohdaten im Mittelpunkt, auf denen spätere Auswertungen aufgebaut werden können.

[12] http://www.doi.ethz.ch
[13] http://www.datacite.org

In erstaunlich wenigen Fällen besteht bisher das Interesse, Daten öffentlich zugänglich zu machen, die unabhängig von einer Publikation archiviert werden. Dies führt zu teilweise komplexen Anforderungen im Berechtigungsmanagement, die derzeit nicht in jedem Fall befriedigend umgesetzt werden können. Die ETH-Bibliothek begrüsst es, wenn öffentlich finanzierte Forschungsergebnisse frei zugänglich gemacht werden und wird Forschende, die Daten freigeben möchten, dabei unterstützen. Die Bibliothek als Dienstleisterin wird aber ihr Angebot zur Langzeitarchivierung nicht mit einer Verpflichtung zur Veröffentlichung der Daten verbinden.
Es wird jedoch erwartet, dass es zukünftig vermehrt Vorgaben zur Publikation von Daten von den Institutionen der Forschungsförderung geben wird und mit der Zeit greifbarere Anreize hierfür geschaffen werden.

Die Handhabung der Forschungsdaten soll durch geeignete Software unterstützt werden. Die Vielfalt der Anforderungen kann naturgemäss nicht mit einer Standard-Software abgedeckt werden, wobei die grössten Herausforderungen im Datenmanagement bzw. aus Sicht des OAIS-Modells im Pre-Ingest bestehen. Doch auch unabhängig vom Funktionsumfang einer Software muss die ETH-Bibliothek in der Zukunft Wege finden, ihre Dienstleistungen für Forschungsdaten einerseits flexibel auf die Bedürfnisse eines möglichst grossen Teils der Kunden anzupassen und gleichzeitig mit beschränkten Ressourcen einen schlanken Betrieb mit einheitlichen Arbeitsabläufen sicherzustellen.

Wie immer beim Aufbau einer Dienstleistung für die gesamte Hochschule ist damit zu rechnen, dass das Angebot für zahlreiche Forschungsgruppen eine sinnvolle Unterstützung darstellt, aber längst nicht alle Bedürfnisse abdecken wird. Es ist sogar anzunehmen, dass bestimmte Forschungsgruppen das Angebot aus verschiedensten Gründen nicht annehmen werden. Unter anderem ist es einzelnen Forschenden durch vertragliche Vereinbarungen mit Industriepartnern untersagt, Daten ausserhalb ihrer eigenen Kontrolle abzulegen, so dass sie grundsätzlich keine zentrale Infrastruktur zur Ablage ihrer Daten nutzen können. Eine in solchen Fällen allenfalls denkbare Ablieferung verschlüsselter Daten mit dem Ziel der Langzeitarchivierung ist nicht sinnvoll, weil dadurch die notwendige Analyse und spätere Erhaltungsmassnahmen weitgehend verhindert werden.

8 Verwendete Software

Es wurde bereits beschrieben, dass im Projekt neben der eigentlichen Langzeitarchivierung auch ein Datenmanagement unterstützt werden soll, das es den Datenproduzenten erlaubt, ihre Daten zur Übergabe an das Langzeitarchiv vorzubereiten.

Die Funktionen der Langzeitarchivierung gemäss dem OAIS[14]-Referenzmodell werden im Konzept der ETH-Bibliothek von Rosetta (Ex Libris) abgedeckt, und zwar gleichermassen für Forschungsdaten, Geschäftsunterlagen des Archivs der ETH Zürich und Bibliotheksinhalte. Während Bibliotheksdaten in der Regel aus anderen Applikationen bzw. zumindest aus strukturierten Ablagen übernommen werden können, ist dies bei den anderen Datentypen nicht gewährleistet. Geschäftsunterlagen stammen zwar aus strukturierten Ablagen, müssen aber in vielen Fällen nach Gesichtspunkten der Archivierung neu geordnet und erschlossen werden. Zudem

[14] Open Archival Information System gemäss ISO 14721; frühere Entwurfs-Version als Pink Book (2009) des CCSDS (Consultative Committee for Space Data Systems) zugänglich:
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

nimmt das Archivpersonal eine Bewertung der Unterlagen vor und wählt nur die seinem Auftrag entsprechenden und für die Erfüllung seiner Aufgaben relevanten Dokumente für die Archivierung aus, während der Rest vernichtet wird.

Ein ähnlicher Schritt ist bei Forschungsdaten nötig, wenn diese aus der alltäglichen Arbeitsumgebung der Forschenden heraus für die Archivierung ausgewählt, strukturiert und dokumentiert werden. Diese Art von Bearbeitung ist im OAIS-Modell nicht vorgesehen und wird auch von Rosetta nicht abgedeckt. Statt die Komplexität der Langzeitarchivierung durch die notwendigen Funktionen weiter zu erhöhen, sollen diese Aufgaben losgelöst vom Langzeitarchiv und unter der Kontrolle der Datenproduzenten bzw. des Archivs der ETH Zürich durchgeführt werden.
Wegen des engen Bezugs zur archivischen Arbeitsweise richtete sich das Augenmerk bald auf eine Software für den Archivsektor. Die Software Docupack ist ein so genannter Package Handler, der genutzt wird, um Submission Information Packages (SIP) bzw. Archival Information Packages zu prüfen und bei Bedarf anzupassen.
Für die Zwecke des Projekts werden derzeit Erweiterungen der Open Source Software entwickelt, die es erlauben, den Editor für Dateistruktur und Metadaten lokal mit abgestimmten Voreinstellungen verfügbar zu machen. Strukturierung und Erschliessung erfolgen dann an einem Speicherort der jeweiligen Forschungsgruppe und damit unabhängig von später gewünschten Zugriffsberechtigungen. Für den Ingest ins Langzeitarchiv erzeugt Docupack eine METS-Datei des SIP, die alle relevanten Informationen enthält.

9 Rolle der Bibliothek

Einige Auswirkungen des Engagements für die digitale Langzeitarchivierung innerhalb der Hochschule auf die Rolle der ETH-Bibliothek wurden bereits erwähnt. Bei dem stark durch technische Fragestellungen geprägten Thema der digitalen Langzeitarchivierung besteht zudem immer die Gefahr, dass entsprechende Vorhaben vor allem als Softwareprojekte wahrgenommen werden. Diese Wahrnehmung birgt mehrere Risiken. Von aussen kann der Eindruck entstehen, sofern eine brauchbare Software eingesetzt werde, sei die digitale Langzeitarchivierung „gelöst“. Wesentliche Herausforderungen werden so nicht gesehen und auch nicht beantwortet und es dürfte nicht zuletzt schwierig werden, die nötigen Ressourcen einzuwerben.

Dies gilt umso mehr, da die digitale Langzeitarchivierung in vielen Fällen Züge einer Versicherung für den Bedarfsfall trägt, während der kurzfristige Nutzen je nach Ausgangslage nicht offensichtlich ist. Hier hat es sich als nützlich erwiesen, bereits frühzeitig die Diskussion sowohl mit Forschenden als auch mit den Informatik-Support-Leitern der Institute und Departemente zu führen. Letztere sehen in vielen Fällen einen unmittelbaren Nutzen in der Straffung der Datenspeicherung in ihrem Zuständigkeitsbereich und in der Schaffung stabilerer Strukturen. Für die Forschenden ist die bereits erwähnte Möglichkeit zur Vergabe und Registrierung von DOI ein zentrales Argument.

Auch für die Verankerung des Projekts *Digitaler Datenerhalt* innerhalb der ETH-Bibliothek spielen die oben genannten Überlegungen eine Rolle: Wie kann vermittelt werden, warum Ressourcen in ein Projekt fliessen, das sich einerseits mit recht „unbibliothekarischen“ Forschungsdaten beschäftigt und bezogen auf die Inhalte der Bibliothek womöglich noch zu einer Duplizierung von Daten führt, die in Online-Anwendungen vorhanden sind? Schliesslich besteht ja bereits heute der selbstverständliche Anspruch, dass die Bibliothek keine Daten verliert, son-

dern sie auf Dauer pflegt. Diese Fragen sind ernst zu nehmen. Ihre Beantwortung berührt einerseits zentrale strategische Überlegungen und erfordert andererseits eine regelmässige konkrete Überprüfung, ob und wie die zugrundeliegenden Konzepte in jedem Anwendungsfall umgesetzt werden sollen.

10 Fazit

Die digitale Langzeitarchivierung von Forschungsdaten ist eng verbunden mit dem Datenmanagement im allgemeinen und stellt aus strategischer Sicht einen weiteren Schritt auf dem Weg einer noch engeren Integration der Bibliothek und ihrer Dienstleistungen in den Forschungsprozess insgesamt dar. Das Engagement der ETH-Bibliothek wird von der Mehrheit der Forschenden ausdrücklich begrüsst.

Die digitale Langzeitarchivierung von Forschungsdaten beinhaltet grosse Herausforderungen aus funktionaler, organisatorischer und technischer Sicht, die sowohl nach aussen gegenüber den Mittelgebern als auch nach innen kommuniziert werden müssen. Zudem ist eine intensive Zusammenarbeit sowohl mit den Forschenden als auch mit Partnern wie den Informatikdiensten beim Aufbau geeigneter Angebote unerlässlich. Sie kann zudem neue Möglichkeiten für attraktive Dienstleistungen eröffnen, wie verschiedene Rückmeldungen zeigen.

Kontakt

ETH Zürich, ETH-Bibliothek
Dr. Matthias Töwe
Leitung Digitaler Datenerhalt
Rämistrasse 101
8092 Zürich
Tel. +41 44 632 60 32
matthias.toewe@library.ethz.ch
www.library.ethz.ch

ETH-Bibliothek Zürich, 17. Mai 2012