Tagungsberichte

Helge Steenweg, Kassel

Langzeitarchivierung als Herausforderung

Ein Bericht über die Tagung „Digitale Langzeitarchivierung an Hochschulen“ vom 29. bis 30. April 2013 an der Humboldt-Universität zu Berlin

Vom 29. bis 30. April fand im Auditorium des Grimm-Zentrums eine gemeinsame Tagung der Firma Ex Libris und der UB der Humboldt-Universität zum Thema „Digitale Langzeitarchivierung an Hochschulen“ statt.

Die stetig voranschreitende Digitalisierung und die wachsende Archivierung von Forschungsdaten[1] stellt insbesondere in Forschung und Lehre eine große Herausforderung dar. In allen Bereichen, seien es Archiv-, Bibliotheks- oder Museumswesen, entstehen rapide anwachsende Mengen elektronischer Medien, die sich platzmäßig effektiv verwalten lassen und den ungeheuren Vorteil eines schnellen Zugriffs und einer umfassenden Volltext-Recherche haben. Neben den textuellen Materialien kommen zukünftig vermehrt multimediale Medien (Fotos, Filme, Musik, Forschungsdaten) zum Tragen. Zusätzlich zu der Menge der auftretenden Daten stellen die unterschiedlichen Formate dieser multimedialen Daten aber eine große Problematik für den Zugriff und die kommende Überlieferung dar, die gelöst werden muss. Wenn es zukünftig nicht gelingt, diese sich durch Soft- und Hardware-Entwicklung stetig verändernden Dokumente für kommende Generationen lesbar zu halten, droht eine sogenannte „digitale Lücke“ oder ein „digitales Vergessen“[2].

Insgesamt besteht theoretisch Einigkeit darüber, dass alle digitalen Medien regelmäßig auf die neueste Programm- und Speichertechnik übertragen werden müssen. Neben bundesweiten Förderungen[3] haben in letzter Zeit im Wissenschaftsbereich vor allem landesweite Förderungen Praxisformen der Langzeitarchivierung (LZA) ermöglicht (bwFLA[4]: Baden Württemberg Functional Long-Term Access; DA-NRW[5]: Digitales Archiv Nordrhein-Westfalen). Im kommerziellen Sektor nimmt das Produkt Rosetta der Firma Ex Libris an Bedeutung zu.

Entsprechend hoch war das Interesse der Tagungsteilnehmer – im Auditorium des Grimm-Zentrums gab es nur wenige freie Plätze. Nach einer Begrüßung durch die Veranstalter, vertreten durch Andreas Degkwitz (UB Humboldt-Universität zu Berlin) und Ulrich Jüngling (Ex Libris), erläuterte Peter Schirmbacher (Humboldt-Universität zu Berlin) die Probleme und Fragestellungen, die sich bei der Langzeitarchivierung von Forschungsdaten ergeben. Er machte dies an einer Auswertung einer Umfrage zum Umgang mit Forschungsdaten an der Humboldt-Universität (HU) deutlich, die im ersten Quartal 2013 durchgeführt wurde und einen Rücklauf von 500 Fragebögen quer über alle Fachbereiche der HU hatte. Forschungsdaten werden demnach vor allem als Nachweisinstrument für Forschungen angesehen, zu Reputationszwecken und zur Nachnutzung ins Netz gestellt. Die Verantwortung für die Speicherung, Sicherung und Archivierung der eigenen Forschungsdaten wird vorrangig in der Eigenverantwortung, gesehen. Der benötigte (geschätzte) Speicherbedarf des eigenen Datenarchivs, das nach Meinung der befragten Wissenschaftler am ehesten im eigenen Institut angesiedelt sein sollte, bewegt sich größtenteils unterhalb von 100 GB (60 Prozent). An Serviceleistungen der Universität in diesem Zusammenhang werden hauptsächlich die Bereiche „Speicherplatz“ sowie „Hilfestellungen zu technischen und rechtlichen Fragen“ genannt.

Wolfram Neubauer (ETH-Bibliothek, Zürich) berichtete von dem Projekt „Data Curation“ der ETH-Bibliothek, das zwischen Datenmanagement und Langzeitarchivierung angesiedelt ist. Auch in Zürich wurde eine Umfrage zum Thema Langzeitarchivierung und zu Wünschen der Nutzer in diesem Bereich durchgeführt, die eine sehr ho-

1 Siehe oben; Langzeitarchivierung von Forschungsdaten – Eine Bestandsaufnahme, hrsg. v. H. Neuroth, St. Strathmann, A. Oßwald, R. Scheffel, J. Klump, J. Ludwig, Glücksstadt, 2012., als Download: http://nestor.sub.uni-goettingen.de/bestandsaufnahme/nestor_lza_forschungsdaten_bestandsaufnahme.pdf.

2 Vgl. http://de.wikipedia.org/wiki/Langzeitarchivierung.

3 Vgl. Projekt NESTOR – www.langzeitarchivierung.de.

4 Vgl. http://bw-fla.uni-freiburg.de/wordpress/?page_id=4; ein Projekt des Rechenzentrums Freiburg mit Projektpartnern.

5 Vgl. http://da-nrw.hki.uni-koeln.de/projects/danrwpublic; ein Projekt der Historisch-Kulturwissenschaftlichen Informationsverarbeitung an der Universität zu Köln mit Projektpartnern.

he Rücklaufquote hatte. Darauf aufbauend wurde das Projekt „Data Curation“ bzw. „Digitaler Datenerhalt“ aufgesetzt. Ausgehend von der Erkenntnis, dass gerade bei Forschungsdaten die Mitwirkung der Forscher bei der Dokumentation der Daten sehr wichtig ist, diese aber oftmals nur wenig Zeit darauf verwenden können und mit weiteren Management- und Erhaltungsentscheidungen möglichst nicht belastet werden wollen, hat die ETH-Bibliothek in diesem Projekt ein Datenmanagement entwickelt, das deutlich vor der Abgabe der Daten einsetzt. Als wichtig wird insbesondere die Zitierbarkeit der Daten angesehen, die in Zürich mit Hilfe des Digital Object Identifiers (DOI) verlässlich gewährleistet wird. Die Langzeitarchivierung wird softwaretechnisch für strukturiert vorliegende Daten über Rosetta abgedeckt. Im Falle der Archivierung von Forschungsdaten müssen zunächst Daten ausgewählt, strukturiert und dokumentiert werden, bevor sie in Rosetta weiterverarbeitet werden können. Diese Software wurde in Zusammenarbeit mit der Firma Docuteam erstellt. Andreas Nef (Docuteam GmbH, Baden-Dättwil) beschrieb in seinem Vortrag „Strukturierung und Aufbereitung von Forschungsdaten fürs Archiv: Konzepte und Werkzeuge“ diese Software-Tools, die in dem ETH-Projekt eingesetzt werden, eingehender. Es handelt sich dabei um die bereits im Archivwesen genutzten Softwarebestandteile „Docupack“ und „Documill“, die für die Abgabe von Forschungsdaten stark erweitert wurden. In Docupack können die Forschergruppen selbständig ihre Primärdaten einstellen und mit beschreibenden Metadaten versehen. Diese Informationen werden dann mittels Documill für die weitere Nutzung in Rosetta aufbereitet.

Hans-Dieter Weckmann informierte in seinem Beitrag „Ein integrativer Ansatz für die Langzeitarchivierung einer Hochschule“ über ein Langzeitarchivierungs-Projekt an der Heinrich-Heine-Universität Düsseldorf, das seit 2012 als Partner die Universitätsbibliothek, das Universitätsklinikum und das Zentrum für Informations- und Medientechnologie in Düsseldorf zusammenführt. Das Vorprojekt hatte zum Ziel, einen Entwurf, eine Spezifizierung und eine Evaluierung einer geeigneten technischen Infrastruktur zu erstellen und den zu erwartenden Kostenrahmen festzustellen. Als Datenlieferanten im Vorprojekt fungierten die Universitätsbibliothek, einige Forscher und die Universitätsklinik; als Software wurde Rosetta eingesetzt. Aus dem Ergebnis des Vorprojekts wird derzeit ein Vorschlag an die Universitätsleistung erarbeitet mit Vorschlägen zum Aufbau eines Kompetenz-Zentrums für langfristige digitale Datenerhaltung und zur Entwicklung einer Service-Schnittstelle. Diese Schnittstelle soll nicht nur die Abgabe von Forschungs- und Bibliotheksdaten zur Langzeitarchivierung, sondern darüber hinaus ein „Ressourcen-Sharing“ mit der Universitätsklinik und die Unterstützung des „Enterprise Content Managements“ der Hochschule (Verwaltungsdaten) ermöglichen. Der benötigte Disk-Storage wird auf mindestens 1 PByte veranschlagt, bei einem geschätzten jährlichen Wachstum von mindesten 100 TByte; das hardwaretechnische Backup soll über Tape-Libraries (1,5 PByte, 2013 zusätzlich 1,75 PByte) an zwei getrennten Standorten erfolgen.

Josh Weisman (Ex Libris) wies in seinem Vortrag: „Rosetta – Digital Preservation solution for Universities“ auf die Notwendigkeit und die Bedingungen von Langzeitarchivierung hin. Die Firma Ex Libris hat die Software Rosetta entwickelt, um digitale Medien über einen langen Zeitraum mit hoher Datenintegrität archivieren zu können. Rosetta wird weltweit von Institutionen eingesetzt.

Nach einer Zusammenfassung der Beiträge des ersten Tages durch Andreas Degkwitz hielt Helmi Ben Hmdia (TIB Hannover) einen Vortrag zum Thema: „Digital Preservation of 3D Objects. DuraArk: Durable Architectural Knowledge“. Aufbauend auf den Ergebnissen des DFG-geförderten Projekts „Probado“ (2006–2011) wird von mehreren europäischen Partnern (Koordination: TIB Hannover) ein EU-gefördertes Projekt „DuraArk (Durable Architectural Knowledge)“ durchgeführt, das zum Ziel hat, Methoden und Softwaretools zur Langzeitarchivierung von 3D-Architektur-Modellen zu entwickeln. Auch hier ist vor allem daran gedacht, im Vorfeld Workflows zur Integration von 3D-Daten in die Rosetta-Umgebung zu definieren und softwaretechnisch umzusetzen.

Über den praktischen Einsatz von Rosetta innerhalb der Bayerischen Staatsbibliothek (BSB) berichtete Matthias Groß, ergänzt von Klaus Ceynowa, zu einem DFG-Antrag im Bereich Langzeitarchivierung mit dem Vortrag: „Rosetta DPS: Implementierung und Einsatz an der Bayerischen Staatsbibliothek – und ein Ausblick auf den Nationalen Leistungsverbund LZA“. Die Langzeitarchivierungsaktivitäten der BSB werden durch eine Kooperation mit dem Leibniz-Rechenzentrum in München ermöglicht, das die technische Infrastruktur zur Verfügung stellt. Der Archivspeicher der BSB belegt mit über einer Milliarde Dateien derzeit 476 TByte; im Jahr 2012 erfolgten knapp 800.000 Downloads ganzer Werke mit einem Downloadvolumen von 76 TByte. Im Durchschnitt der letzten Monate stieg das Langzeitarchiv pro Monat um ca. 5 TByte an. Eine derartige Datenmenge bedingt Anforderungen an Performanz, Datendurchsatz und Verfügbarkeit. Das Vorprojekt (2010–2012) mit Ex Libris an der BSB wird derzeit ausgeweitet auf drei weitere Universitätsbibliotheken; hinzu kommen dabei neben weiteren textuellen

Medien auch Audio-Daten (Werbefunkarchiv). Klaus Ceynowa berichtete anschließend über einen Antrag im Rahmen der DFG-Ausschreibung „Neuausrichtung überregionaler Informationsservices", Themenfeld Langzeitarchivierung. Aufbauend auf den Ergebnissen der bayerischen LZA-Projekte ist daran gedacht, einen „Nationalen Leistungsverbund LZA" aufzubauen, der bundesweit über eine Geschäftsstelle die LZA-Bemühungen bündeln und koordinieren könnte. Über einen kooperativen Ansatz sollen Bibliotheken respektive Bibliotheksverbünde mit größeren Rechenzentren zusammenarbeiten und möglichst spartenübergreifend bedarfsgerecht archivieren. Da keiner der beiden Anträge in diesem Themenfeld seitens der DFG bewilligt wurde und eine erneute Ausschreibung in diesem wichtigen Bereich angekündigt ist, soll dieser Antrag modifiziert eingereicht werden.

Über die Möglichkeiten der „Archivierung elektronischer Ressourcen mit LOCKSS am Beispiel des deutschen LOCKSS-Netzwerkes" berichtete Niels Fromm (Humboldt-Universität zu Berlin). Das Open-Source-Projekt LOCKSS wurde 1999 an der Stanford-University begonnen und hatte zunächst die Archivierung lizensierter E-Journals der Verlage unter Wahrung der rechtlichen Gegebenheiten im Focus. Seit 2004 im Produktivbetrieb, sind über das LOCKSS-Netzwerk über 1.000 Zeitschriften von 520 Verlagen[6] zugreifbar. LOCKSS als kooperatives System mit Focus auf „Bitstream-Preservation" mittels Festplatten (sogenannte LOCKSS-Boxen) möchte diese Daten sammeln, bereitstellen und bewahren. Die LOCKSS-Boxen sind aus Sicherheitsgründen im LOCKSS-Netzwerk verteilt; für ein abgesichertes Speichern sind stets mindestens sieben Kopien notwendig, die regelmäßig überprüft werden (Integrität mittels Prüfsummen). In Deutschland besteht das LOCKSS-Netzwerk aus einem Kompetenzzentrum und neun Boxen, die im Rahmen des Projektes LuKII an neun Institutionen aufgebaut wurden. Konvertierungen aus dem Originalformat erfolgen innerhalb von LOCKSS nur nach Bedarf (Anforderung eines Mediums). In der Praxis erweisen sich die hohe Bereitstellung von Speicherkapazität durch die Fokussierung auf mindestens sieben Kopien eines Mediums und die zu erbringende Rechenleistung bei der kontinuierlich zu betreibenden Prüfsummen-Überprüfung als Handicap für die Ausweitung dieser Open-Source-Software von ihrer ursprünglichen Zweckgebung der Archivierung von textuellen Medien (Zeitschriften in pdf/A) auf die Langzeitarchivierung aller Medientypen (z.B. auch AV-Medien).

Das Tagungsprogramm und die PDFs der Vorträge sind über eine entsprechende Seite der Universitätsbibliothek der Humboldt-Universität abrufbar[7].

Unabhängig davon, welche Software letztlich für die Langzeitarchivierung an Hochschulen favorisiert bzw. bereits eingesetzt wird, hat diese Tagung als Erkenntnis gebracht, dass es zum einen Praxisansätze für die dringend benötigte Langzeitarchivierung gibt, und zum anderen, dass die Durchführung von Langzeitarchivierung mit viel Arbeit und Geld verbunden ist. Durch die unterschiedlichen Beispiele und Projekte auf der Tagung kam aber ein weiterer Aspekt zum Tragen, der sehr wichtig erscheint und zukünftig auch die Rolle der Bibliotheken nachhaltig verändern dürfte: Datenmanagement an der Hochschule beginnt weit vor der eigentlichen Archivierung. Bei der Durchführung von Langzeitarchivierung ist viel Aufklärungsarbeit (s. Umfrage HU, Beitrag Schirmbacher) zu leisten, stabile Strukturen in Hard- und Software zu schaffen (s. Beitrag Weckmann), einfache und selbsterklärende Abgabesoftware zu schreiben (s. Beitrag Nef) und eine Möglichkeit für eine übergeordnete nationale Langzeitarchivierung mittels Kooperationen aufzuzeigen (s. Beitrag Groß/Ceynowa). Hier entsteht großer Bedarf seitens der Nutzer und eine Chance für Bibliotheken, mittels ihrer vorhandenen Kompetenz bei der Bearbeitung von Daten und Medien noch stärker in Forschungsprozesse eingebunden zu werden.

Helge Steenweg
Universitätsbibliothek Kassel
Diagonale 10
34127 Kassel
steenweg@bibliothek.uni-kassel.de

6 Vgl. http://www.lockss.org.

7 Vgl. http://www.ub.hu-berlin.de/ueber-uns/oeffentlichkeitsarbeit/tagung-digitale-langzeitarchivierung-an-hochschulen.