Archivierung von digitalen Dokumenten in Bibliotheken[1]

Von Marianne Dörr

Einführung

Archivbibliotheken wie Die Deutsche Bibliothek – oder auf Landesebene die Bayerische Staatsbibliothek sehen sich wie die Archive zunehmend mit Aufgabe der Verwaltung und der langfristigen Vorhaltung bzw. Verfügbarmachung von digitalen Dokumenten konfrontiert. Dabei stellt sich die rechtliche Situation für die bibliothekarische Archivierungspflicht digitaler Dokumente derzeit in Deutschland (auf Bundes- und auf Länderebene) folgendermaßen dar: Publikationen auf Datenträgern unterliegen bereits der gesetzlichen Ablieferung in Analogie zu Printpublikationen. Bei den Netzpublikationen ist die Frage der Pflichtablieferung in Deutschland, wie übrigens auch in der Mehrheit der anderen europäischen Staaten, noch nicht gesetzlich geregelt. Die Deutsche Bibliothek hat ein Projekt „Elektronische Depotbibliothek“ ins Leben gerufen, in dem sie mit Verlagen wie dem Springer Wissenschaftsverlag u.a. praktische Modalitäten eines Transfers und einer Speicherung elektronischer Netzpublikationen erprobt, um damit Erfahrungen zu sammeln, die in eine Gesetzesnovellierung einfließen können.[2]

Die Spanne der elektronischen Publikationen, mit denen Bibliotheken umgehen müssen, ist sehr breit. Als Beispiele seien genannt:
(Verlags-)Publikationen auf Datenträgern[3]
- bibliographische oder Volltext-Datenbanken,
- Ratgeber, z.B. Hilfe bei der Steuererklärung,
- Demoversionen von Programmen bzw. Einführungen in die Nutzung von Word, Excel etc.,
- Spiele, Faktensammlungen ...

Netzpublikationen
- Online-Datenbanken,
- elektronische Zeitschriften,
- elektronische Dissertationen und Hochschulschriften,
- Web-Ressourcen / thematische Web-Sites,
- retrodigitalisierte Dokumente

Für die wissenschaftliche Publikation und Dokumentation stehen innerhalb der elektronischen Publikationen die elektronischen Zeitschriften im Zentrum des Interesses – und der Nutzung. Diese gelangen physisch aber schon gar nicht mehr in die Bibliotheken, die ihren Nutzern nur den Zugang über den Verlagsserver anbieten können. Partiell werden zur Absicherung bereits Klauseln in die Lizenzabkommen aufgenommen, die der Bibliothek bei Kündigung des Lizenzvertrags Sicherheitskopien für die „abonnierte Laufzeit” garantieren sollen.

---

[1] Zum Thema Digitale Archivierung ist in den letzten Jahren eine fast nicht mehr überschaubare Menge an Publikationen erschienen, die größtenteils online zugänglich sind. Hier sei nur auf zwei grundlegende Publikationen verwiesen, die ihrerseits zahlreiche Referenzen enthalten: Preservation Management of Digital Materials. A Handbook: Neil Beagrie – Maggie Jones *http://www.jisc.ac.uk/dner/preservation/workbook/* und RLG Digital Archives Attributes Working Group: Trusted Digital Repositories: Attributes and Responsibilities *http://www.rlg.org/longterm/repositories.pdf*

[2] Informationen zum Umgang der Deutschen Bibliothek mit Netzpublikationen unter: *http://deposit.ddb.de/netzpub/web_netzpublikationen_root.htm*

[3] Als Beispiel für die Menge der Publikationen seien die Zugangszahlen der Bayerischen Staatsbibliothek genannt. Der Zugang an Datenträgern ist von der Pflichtablieferung dominiert. Bei den CD-ROMs beispielsweise kamen 2000 1.600 Stück (1999: 1.466) über die Zugangsart Pflichtablieferung in die Bibliothek; nur 269 (1999: 181) wurden gekauft.

Auf der anderen Seite gibt es partiell Absichtserklärungen großer Verlage, dauerhaft auch die Archivierung zu übernehmen. Entsprechende Verlautbarungen gibt es vom Springer-Verlag, der andererseits im Projekt der Deutschen Bibliothek *Elektronische Depotbibliothek* kooperiert. Auch die American Society of Physics als Zeitschriften verlegende Fachgesellschaft hat zugesagt, die langfristige Archivierung ihrer Zeitschriften selbst zu übernehmen. Andere Verlage, wie Elsevier, verhandeln bereits mit großen Bibliotheken, die später nationale Archivserver für die Verlagspublikationen werden sollen. Die Landschaft ist auch hier uneinheitlich.

Noch erscheint die Mehrheit der elektronischen Zeitschriften parallel zu einer weiterexistierenden Printausgabe. Von den in der Elektronischen Zeitschriftenbibliothek Regensburg[4] nachgewiesenen rd. 8.500 Zeitschriften (Stand Anfang 2001) werden „nur" 900 ausschließlich elektronisch publiziert. Diese Zahl kann jedoch schnell ansteigen. Außerdem nehmen Unterschiede zwischen den gedruckten und elektronischen Versionen einer Zeitschrift zu, z.B. durch die Einbindung von Applets u.Ä. in die digitale Ausgabe.

Einen weiteren zunehmend bedeutenden Posten bei den Netzpublikationen machen die Online-Dissertationen aus. Im Zuge der an den Universitäten laufenden Umgestaltung der Promotionsordnungen werden immer mehr Dissertationen online publiziert. In einem kooperativen Projekt[5] wurden Maßgaben für eine Abfassung und XML-Konvertierung von Dissertationen entwickelt, die auch den Ansprüchen der Langzeitarchivierung entgegenkämen. Bei der Mehrheit der Universitäten und Universitätsbibliotheken kann die zusätzliche Arbeit einer Konversion von Word-, LaTeX und anderen Formaten in das „neutrale" XML jedoch nicht geleistet werden. Die Deutsche Bibliothek hat angeboten, für Online-Dissertationen als Depotbibliothek zu fungieren und entsprechende Transfer-Optionen und Metadatenschnittstellen eingerichtet. Während im Jahr 1998 der Zugang an Online-Dissertationen bei DDB noch bei 100 lag, waren es 1999 bereits 470 und im Jahre 2000 1380. Die große Mehrheit davon liegt als PDF-Files, nicht in XML-Codierung vor.

Bei den anderen Netzpublikationen, die Bibliothekare in Analogie zum Printbereich oft als „graue Publikationen" bezeichnen, herrscht eine bunte Breite vor: Qualität, Formate, Präsentation etc. sind ausgesprochen heterogen. Außerdem entzieht sich die wachsende Masse dieser Web-Sites einer systematischen Erschließung und Sammlung.

Versuche des automatischen Einsammelns („harvesten") von Web-Publikationen bestimmter Domainbereiche, die im Rahmen des Nedlib-Projekts[6] europäischer Nationalbibliotheken unternommen wurden, waren nicht sehr erfolgreich: Es gibt bereits zu viele dynamisch erzeugte Websites, die sich über Harvester nicht mehr „einfangen" lassen.

Für den explodierenden Bereich der Web-Sites und Web-Publikationen müssen neue Selektionskriterien für die Archivierung bzw. Archivwürdigkeit entwickelt werden. Veröffentlichte Ansätze hierzu aus nationalbibliothekarischer Sicht gibt es international bereits in Australien, Finnland und einigen anderen Staaten.[7]

Bibliotheken wie die Bayerische Staatsbibliothek dürfen die Frage der Sammlung und Archivierung von Web-Publikationen aber nicht nur aus der Sicht der nationalen oder regionalen Überlieferungsbildung und Archivierungspflicht sehen, sondern müssen auch den fachlichen Aspekt berücksichtigen, der aus ihrer Verantwortung für die Literatur- bzw. Informationsversorgung in bestimmten Disziplinen im Rahmen des fachlichen Verteilungsplans der Deutschen Forschungsgemeinschaft[8] resultiert. Hier hat die BSB, wie andere SSG-Bibliotheken

---

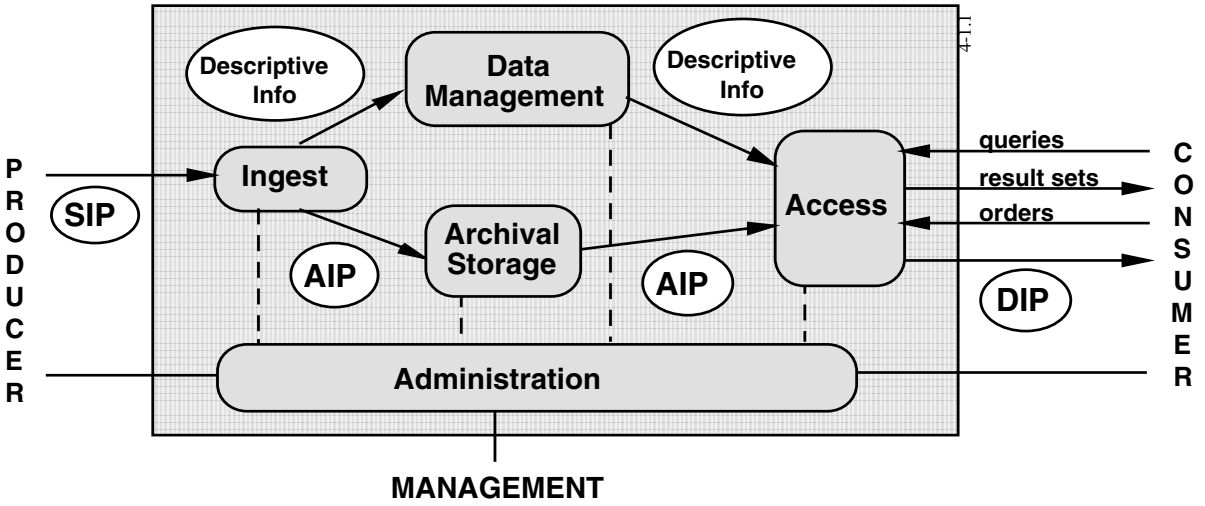
[4] *http://rzblx1.uni-regensburg.de/ezeit/fl.phtml?bibid=UBR*
[5] *http://www.dissonline.de/*
[6] *http://www.kb.nl/coop/nedlib/*
[7] Vgl. für Finnland: *http://www.lib.helsinki.fi/eva/english.html* und für Australien *http://pandora.nla.gov.au/selectionguidelines.html*
[8] Welche Bibliothek für welches Fach zuständig ist, kann leicht über das Online-System Webis ermittelt werden, das mit Unterstützung der DFG aufgebaut wurde: *http://webis.sub.uni-hamburg.de/*

auch, in einem ihrer Sondersammelgebiete, nämlich der Geschichtswissenschaft, mit der datenbankgestützten Erschließung wissenschaftlicher digitaler Ressourcen[9] begonnen. Es gibt aber derzeit noch kein Verfahren bzw. tragfähiges Konzept zu deren auch physischer Speicherung und langfristiger Zugangssicherung.

OAIS als Modell für ein Archivierungssystem

Noch an keiner Bibliothek existiert ein Depotsystem für elektronische Publikationen, das allen vielfältigen Aspekten einer Langfristarchivierung gerecht wird. Allerdings gibt es inzwischen einen relativ breiten Konsens über ein Referenzmodell für die Konzeption und Implementierung eines solchen Systems mit seinen Funktionalitäten. Dies ist das vom Consultative Committee for Space Data Systems (CCSDS) entworfene OAIS (Open Archival Information System)-Modell[10]. Das OAIS-Modell will die Funktionalitäten, die für eine langfristige Archivierung von Informationen notwendig sind, in ihrem Zusammenspiel als Referenzmodell abbilden. Damit wird auch erreicht, dass eine Verständigungsbasis zwischen Institutionen, die mit dieser Aufgabe befasst sind oder befasst sein werden, gegeben ist.

Die folgende Graphik stellt nur die oberste Ebene der Funktionalitäten dar. Für alle Prozesse existieren detaillierte Feinmodelle.



(OAIS-Funktionalitätenmodell: *http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-1.pdf*)

Bei der Auseinandersetzung mit dem Modell im Rahmen des bereits erwähnten Nedlib-Projekts europäischer Bibliotheken wurde moniert, dass keine Verbindungen bzw. Schnittstellen zu existierenden bibliothekarischen Arbeits- und Datenverwaltungssystemen vorgesehen sind. Außerdem wurden Erweiterungen vorgeschlagen, die die speziellen Vorgänge bzw. Erhaltungsmaßnahmen (z.B. Migration, Emulation) betreffen, die für eine Langzeitarchivierung von Daten notwendig sind. Die im Modell genannte Funktionalität „archival storage" umfasst nur den Komplex des Backups i.S. von Datensicherung und -speicherung. Der darüber hinausgehende Bereich der Erhaltungsmaßnahmen soll unter „preservation" gefasst werden. Die entsprechenden Modifizierungen sind in die schriftliche Fassung einer Funktionalitätenbeschreibung für ein Depotsystem elektronischer Publikationen eingegangen, die von der Koninklijke Bibliotheek in Den Haag erstellt wurde.[11]

Bisher haben zwei Bibliotheken – nach entsprechenden Ausschreibungsverfahren – Aufträge zur Entwicklung und Implementierung eines Depot- bzw. Archivierungssystems, das auf dem

[9] *http://www.bsb-muenchen.de/fachinfo/gesch/infoW.htm http://webis.sub.uni-hamburg.de/*
[10] *http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf*
[11] *http://www.kb.nl/coop/nedlib/meetings/paris/GEN-264.pdf*

OAIS-Modell beruhen soll, erteilt. Es handelt sich um die Koninklijke Bibliotheek und um die British Library. Bei beiden ist die Firma IBM zum Zuge gekommen.

Im OAIS-Modell wird die Bedeutung von Information über die abzuspeichernden Daten schon darin deutlich, dass jeder Prozess von der Erstellung von „information packages", also von Metadaten, begleitet wird. Elektronische Publikationen stellen besondere Anforderungen an die Verzeichnung. Nicht mehr nur bibliographische, sondern auch technische und administrative Metadaten müssen erfasst werden. Für die Verzeichnung in deutschen Bibliothekskatalogen sind die Regeln für alphabetische Katalogisierung in wissenschaftlichen Bibliotheken (RAK-WB) um Sonderbestimmungen für AV und – in einem zweiten Schritt – für elektronische Medien ergänzt worden (RAK-NBM „Non Book Materials"). Ebenso wurde das Austauschformat MAB (Maschinelles Austauschformat für Bibliotheken) um Kategorien für den Transport technischer Informationen (650–659, z.T. mit Unterfeldern) erweitert. Dies ist jedoch noch nicht ausreichend:

Für einen zielgerichteten Zugriff müssten Feldinhalte bzw. die Syntax der Eintragungen normiert werden. In der Regel sind diese Kategorien außerdem in den großen Katalogsystemen nicht indiziert, d.h. nicht dezidiert such- und zugriffsfähig. Die verzeichneten Informationen sind vorwiegend darauf ausgerichtet, einem Benutzer mitzuteilen, was er zu einer Nutzung der entsprechenden elektronischen Publikationen benötigt. Der Aspekt Verwaltung der Daten steht demgegenüber im Hintergrund.

Im internationalen Rahmen entstand inzwischen eine Reihe von Ansätzen zur Entwicklung ausgefeilterer Metadatenformate für elektronische Publikationen. Diese wurden in einem „White Paper" von der Arbeitsgruppe der beiden großen amerikanischen Bibliotheksverbünde OCLC und RLG zusammengestellt.[12] Ein gemeinsames anerkanntes und breiter angewendetes Format existiert indessen noch nicht.

Vorherrschende Ziele von Metadaten-Verzeichnung im Kontext der Bestandserhaltung bzw. Archivierung sind:

- Technische Informationen sammeln und dokumentieren, die Bestandserhaltungsentscheidungen bzw. -maßnahmen unterstützen
- Dokumentation von erfolgten Bestandserhaltungsaktivitäten
- Garantie der Authentizität der digitalen Objekte geben
- Rechte am digitalen Objekt dokumentieren
- Sicherung der Nutzbarkeit des digitalen Objekts

In Deutschland gibt es innerhalb des vom BMBF geförderten CARMEN-Projekts Arbeitspakete[13], die sich direkt mit dem Thema Metadaten für Rechte und Archivierung beschäftigen – konkrete, i.S. von anwendbaren Ergebnissen resultierten hieraus jedoch noch nicht.

Die Mehrheit der genannten Projekte ist allgemein auf digitale Daten ausgerichtet. Es stellt sich jedoch die Frage, ob nicht noch speziellere Metadatensets für bestimmte Datentypen bzw. Anwendungen definiert werden müssen. Dies gilt besonders für den ersten genannten Aspekt, die technischen Metadaten. Ansätze hierfür gibt es z.B. im Bereich der Stillbilder. Dies ist zwar nur ein kleiner Bereich der digitalen Daten; andererseits aber von zunehmender Bedeutung, da Bildagenturen, Kulturinstitutionen und Privatleute zunehmend digitale Bilder erzeugen und archivieren. Vielleicht gerade deswegen sind fast parallel Initiativen der Industrie und der genannten Kulturinstitutionen (angestoßen von der bereits genannten OCLC/RLG-Arbeitsgruppe) entstanden.

---

[12] *http://www.oclc.org/digitalpreservation/presmeta_wp.pdf*
[13] *http://www.sub.uni-goettingen.de/carmen/*

Als Beispiel für Standardisierungsansätze sollen diese beiden kurz vorgestellt werden.

Metadaten für Still-Bilder

NISO Draft Technical Metadata for Digital Still Images[14]

DIG 35 Metadaten-Initiative[15] der Digital Imaging Group (Adobe, Canon, Eastman Kodak, Fuji, Hewlett-Packard, IBM, Intel)
http://www.digitalimaging.org/

Das Ziel beider Initiativen liegt in der Schaffung einer dateiformat-unabhängigen Metadaten-Definition und eines Austauschformats für Informationen über Still-Bilder und für die Verwaltung von Bildern.

Folgende „Designprinzipien" werden definiert:

- Aufsetzen auf existierenden Standards
- Erweiterbarkeit und Skalierbarkeit
- Konsistenz
- Internet-Fähigkeit (deshalb XML-Implementierung angestrebt)
- Fokus auf mittel-/langfristigen Perspektiven
- Erhaltung der Information – nicht der Daten

Beide Entwürfe beinhalten Kategorien zur Beschreibung der „Basic Image Parameter, der „Image Creation" und der „History". DIG35 umfasst weitergehend auch Metadaten zur Bildinhaltsbeschreibung (Content Description) und zu den Bildrechten (Intellectual Property Rights (IPR)).

Der NISO-Ansatz geht in den berücksichtigten Bereichen teilweise über DIG35 hinaus:

Im einzelnen werden behandelt:

- Basic Image Parameters (Format, Compression, Photometric Interpretation, Segments ...)
- File (Image Identifier, File Size, Checksum, Orientation, Display Orientation, Preferred Presentation ...)
- Image Creation (Source Type, Scanning Agency, HostComputer, Scanning System, Camera Capture Settings ...)
- Imaging Performance Assessment (Spatial Metrics – Image Width, Length ..., Color Map, TargetData – TargetType, – ID, Profiles ...)
- Change History (Image Processing-Software, Actions, Previous Image Metadata)

Inzwischen haben sich einige Institutionen zur testweisen Implementierung des NISO-Dictionaries bereit erklärt. Von den dortigen Erfahrungen wird es auch abhängen, ob das Dictionary das endgültige Standardisierungsverfahren durchlaufen wird.

Die Einsicht, dass für elektronische Publikationen ein Mehr an Metadaten notwendig ist, um alle Informationen für eine langfristige Verwaltung und Pflege zugriffsfähig zu haben, ist weit verbreitet. Ein Hauptproblem der Metadatenerhebung und Verzeichnung für elektronische Publikationen wird jedoch der Aufwand sein, der für die Institutionen damit verbunden ist. Wenn nicht Verfahren zu einer weitgehend automatisierten Metadatenverzeichnung entwickelt werden, die z.B. softwaregesteuert parallel zur Erzeugung der Daten geschieht, wird die Durchsetzung von ausführlichen Metadaten-Formaten allein aufgrund fehlender Kapazitäten in den erzeugenden und verwaltenden Institutionen schwierig bis unmöglich sein.

---

[14] NISO (National Information Standards Organization) ist eine non-profit association, die Standardisierungsnotwendigkeiten im Bereich der Informations- und Kommunikationssysteme identifiziert und mit den jeweiligen involvierten Institutionen in die Wege leitet. Der Entwurf des technischen Dictionaries ist zugänglich unter: *http://www.niso.org/standards/resources/Z39_87_trial_use.pdf*

[15] *http://www.bgbm.org/TDWG/acc/Documents/DIG35-v1.0-Sept00.pdf*