Cornelia Diebel

Sammlung und Langzeitarchivierung von E-Journals an der Deutschen Nationalbibliothek

Der Beitrag beschreibt die Strategie und den Stand der Umsetzung der Deutschen Nationalbibliothek zur Sammlung von Internetdokumenten, sogenannten Netzpublikationen. Die technischen Voraussetzungen bezüglich der Nutzung von Metadatenformaten, der Festlegung von zu sammelnden Dateiformaten und zu nutzenden Schnittstellen werden beschrieben, genauso wie die besonderen Herausforderungen bei der Sammlung von E-Journals, die in der Zuordnung von verschiedenen Teilen zum Ganzen liegen. Empfehlungen zweier deutschen Initiativen in Bezug auf die Sicherstellung von Perpetual Access im Bereich des digitalen Contents insbesondere von E-Journals werden aufgegriffen und am konkreten Beispiel der Nutzung von Metadaten im NLM-DTD/Schema und des Schemas von CrossRef darstellt.

1. Einleitung

Mit Inkrafttreten des Gesetzes über die Deutsche Nationalbibliothek (DNBG)[1] vom 22. Juni 2006 (BGBl. I S. 1338) hat die Deutsche Nationalbibliothek den Auftrag der Sammlung, Erschließung, Verzeichnung und Archivierung von sogenannten unkörperlichen Medienwerken erhalten. Der Begriff „Medienwerke in unkörperlicher Form" steht synonym für Netzpublikationen oder Online-Publikationen. Darunter werden elektronische Veröffentlichungen verstanden, die über ein öffentliches Netz, heute über das Internet, verfügbar sind. Die Sammelpflicht umfasst sowohl Internetpublikationen mit Entsprechung zum Print-Bereich wie E-Books und E-Journals als auch web-spezifische Medienwerke wie Webseiten. Weitere Beispiele für Netzpublikationen sind Hochschulprüfungsarbeiten, Musikdateien und Digitalisate.

Um diesen umfassenden Auftrag erfüllen zu können, geht die Deutsche Nationalbibliothek beim Aufbau der Sammlung von Netzpublikationen stufenweise vor. Zunächst wird die einzelobjektbezogene Sammlung von Netzpublikationen mit der oben genannten Entsprechung zum Printbereich vorangetrieben. In einem weiteren Schritt werden automatisierte Verfahren zur Sammlung ganzer Gruppen von Objekten, wie etwa vollständiger Websites, entwickelt.

2. Technische Konzepte

Grundlegendes Konzept der Deutschen Nationalbibliothek bei der Sammlung von Netzpublikationen ist die Entwicklung und Nutzung automatisierter Verfahren. Nur durch eine weitgehend automatische Bearbeitung des

[1] http://www.gesetze-im-internet.de/dnbg/index.html (Stand 02.01.2012)

täglichen Zugangs kann die Menge der Netzpublikationen, die in den Sammelauftrag der Deutschen Nationalbibliothek fallen, bewältigt werden.

Für die Umsetzung von automatisierten Verfahren im Zusammenhang mit der Übernahme von Netzpublikationen in das Archiv- und Katalogsystem (Ingest) der Deutschen Nationalbibliothek müssen drei grundlegende Voraussetzungen erfüllt sein. Das ist zuerst die Verwendung von standardisierten Metadatenformaten für die Beschreibung und den Nachweis der Netzpublikationen im Katalog- bzw. Recherchesystem. Genauso wichtig sind die Festlegung der aus Sicht der digitalen Langzeitarchivierung (dLZA) sammelfähigen Dateiformate und die Definition von Schnittstellen zur Übertragung der Objekte auf Ablieferserseite und auf Seiten der Deutschen Nationalbibliothek.

2.1 Standardisierte Metadatenformate

Bibliotheken haben seit Jahrhunderten Metadaten ihrer Sammlungsobjekte vorliegen. Inzwischen haben deskriptive Metadaten auch außerhalb der Bibliotheken eine hohe Bedeutung erlangt und werden bereits bei der Produktion von Objekten, auch bei der Produktion von Netzpublikationen erstellt. Diese bereits beim Produzenten erstellen Metadaten nachzunutzen und nicht erneut per Autopsie zu erstellen, ist ein wichtiges Ziel der Deutschen Nationalbibliothek. Aus diesem Grund sammelt die Deutsche Nationalbibliothek in der jetzigen Ausbaustufe des Geschäftsprozesses für Netzpublikationen lediglich Objekte, die bereits mit Metadaten geliefert werden können. Unterstützte Metadatenformate sind augenblicklich ONIX 2.1[2], das im Bereich der Verlage einen verbreiteten Standard darstellt, MARC21[3] als nationales und internationales bibliothekarisches Austauschformat, das auch von einigen Verlagen bedient werden kann und zusätzlich XMetaDissPlus[4] als etabliertes Metadatenformat im Bereich der universitären Repositorien.

Für all diese Metadatenformate wurden Mindestanforderungen an die bibliografische Beschreibung definiert und in sogenannten Metadaten-Kernsets[5] veröffentlicht. Die Mindestanforderungen orientieren sich an wichtigen Elementen der bibliografischen Beschreibung, die in der Regel von Datenproduzenten bereits erfasst wurden. Beispiele hierfür sind der Titel der Publikation, Angaben zu Autorinnen und Autoren, zu Verlagen und zu Erscheinungsjahren.

[2] http://www.editeur.org/15/Previous-Releases/#About%20R%202.1 (Stand 02.02.2012)
[3] http://www.loc.gov/marc/ (Stand 02.02.2012)
[4] http://www.d-nb.de/standards/pdf/ref_xmetadissplus_v2-1.pdf (Stand 02.02.2012)
[5] http://www.d-nb.de/netzpub/ablief/kernset_metadaten.htm (Stand 02.02.2012)

2.2 Dateiformate

Um die wichtige Funktion der digitalen Langzeitarchivierung von Online-Ressourcen gewährleisten zu können, ist es von großem Vorteil, wenn die Dateiformate der zu sammelnden Objekte Standardformaten mit offenen Spezifikationen folgen. Der Vorteil offener Dateiformate besteht in der prinzipiellen Möglichkeit der Nachbildbarkeit von Anwendungsprogrammen, wenn diese bereits vom Markt verschwunden sind oder von aktuellen Betriebssystemen nicht mehr unterstützt werden. Aktuell sammelt die Deutsche Nationalbibliothek Netzpublikationen hauptsächlich in PDF (alle Ausprägungen, bevorzugt aber PDF/A), EPub und bei Bildformaten TIFF und JPEG.

2.3 Schnittstellen

Für die Übertragung der Netzpublikationen (jeweils Metadaten mit zugehörigem Objekt) bietet die Deutsche Nationalbibliothek zurzeit drei Schnittstellen zur Ablieferung: ein Webformular und zwei automatisierte Ablieferungsverfahren, hier je ein Push- und ein Pullverfahren. Das Push-Verfahren nutzt ein Ablieferungskonto (einen sogenannten Hotfolder), über das per SFTP- oder WebDAV-Schnittstelle aktiv Objekt- und Metadaten an die Deutsche Nationalbibliothek übertragen werden können. Abgeliefert werden einzelne Transferpakete, die gezippt als Container jeweils ein Objekt mit einem dazugehörigen Metadatensatz enthalten. Das Pull-Verfahren setzt auf dem OAI Protocol for Metadata Harvesting auf. Hierüber können alle für die Verarbeitung notwendigen Metadaten und über die Auswertung eines abgesicherten Bereitstellungslinks auch das Objekt vom Produzenten zur Abholung bereitgestellt werden.

3. Herausforderungen bei der Sammlung von E-Journals

Die in Abschnitt 2 beschriebenen technischen Konzepte sind für die Sammlung von E-Books in der Deutschen Nationalbibliothek bereits umgesetzt und stellen so eine breitflächige Sammlung von monografischen Netzpublikationen sicher. Der Zugang an monografischen Netzpublikationen des Jahres 2011 beträgt über 93.000 Publikationen, inkl. Hochschulveröffentlichungen. Für eine ähnlich umfassende Sammlung von E-Journals sind weitere Herausforderungen zu meistern.

Als Zeitschrift wird ein Werk definiert, das sich durch ein fortlaufendes und mehr oder weniger regelmäßiges Erscheinen von Monografien abgrenzt. Zeitschriften als fortlaufende Sammelwerke haben keinen von vornherein geplanten Abschluss und bestehen in der Regel aus mehreren Beiträgen. Der Aspekt des fortlaufenden Erscheinens bedingt bei der Sammlung von E-Journals die Notwendigkeit der Zuordnung von einzelnen Teilen (Heften oder Artikeln) eines E-Journals (also den eigentlichen zu sammelnden Objekten) zu dem

jeweiligen Titel eines E-Journals. Hierfür wird es notwendig, in den Ablieferprozessen Identifier wie die ISSN oder die Identifikationsnummer der Zeitschriftendatenbank zu definieren, die durch festgelegte Abgleichsverfahren die automatische Zuordnung von Teilen einer Zeitschrift zum Zeitschriftentitel ermöglichen.

Neben der Zuordnung und Verknüpfung von zusammengehörigen Teilen besteht im Bereich der Zeitschriften die Schwierigkeit, dass kein umfassend verbreitetes Standardmetadatenformat vorhanden ist.

3.1 Sicherstellung eines dauerhaften Zugriffs bei E-Journals

Die Aufgabe der Deutschen Nationalbibliothek besteht vor allem darin, die langfristige Verfügbarkeit von Inhalten herzustellen. Im Rahmen der Sammlung von E-Journals trägt die Deutsche Nationalbibliothek dazu bei, sowohl einen dauerhaften Zugriff als auch die digitale Langzeitarchivierung sicherzustellen.

Mit Aspekten der Sicherstellung eines dauerhaften Zugriffs auf wissenschaftliche Informationen und hier vor allem auf die besonders wichtigen Inhalte von E-Journals haben sich in den letzten Jahren zwei Initiativen in Deutschland beschäftigt.

Die Studie „Dauerhaften Zugriff sicherstellen: Auf dem Weg zu einer nationalen Strategie zu Perpetual Access und Hosting elektronischer Ressourcen in Deutschland“[6], in Auftrag gegeben von der Allianz der deutschen Wissenschaftsorganisationen, widmet sich dem Thema Erhalt der dauerhaften Verfügbarkeit (Perpetual Access) von digitalen Inhalten, besonders im Bereich von E-Journals und hier im Bereich von lizenzpflichtigem Material, das bislang häufig nicht von Bibliotheken gehostet wurde. Bibliotheken richten ihre Bestände jedoch zunehmend auf rein elektronische Varianten aus: Im Unterschied zu gedruckten Materialien ist der Zugriff auf elektronische Ausgaben des lizensierten Contents dann nicht mehr gewährleistet, wenn die Daten bei den Produzenten nicht mehr zur Verfügung stehen.

Die Studie, auch Beagrie-Studie genannt, spricht insgesamt 30 Empfehlungen in den Bereichen technische Infrastruktur, Standards, Geschäftsmodelle, Kosten, Finanzierung und Organisatorisches aus. Wichtig in diesem Zusammenhang ist vor allem die Empfehlung Nr. 10: „Technische Richtlinien und Anforderungen als Ergänzung zu Lizenzvereinbarungen erarbeiten und vereinbaren. Auf diese Weise sollten verbreitete Standards gefördert werden, z.B. Nutzung von NLM-DTD/Schema.“[7]

[6] Deutsche Forschungsgemeinschaft im Auftrag der Allianz der deutschen Wissenschaftsorganisationen, 2009

[7] vgl. Deutsche Forschungsgemeinschaft im Auftrag der Allianz der deutschen Wissenschaftsorganisationen, 2009, S. 16 und S. 101

Die von der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder (GWK) einberufene *Kommission Zukunft der Informationsinfrastruktur (KII)* betrachtet in ihrem teilweise auf die Beagrie-Studie bezugnehmenden Gesamtkonzept acht zentrale Handlungsfelder, die für die Sicherstellung einer Informationsinfrastruktur in Deutschland zentrale Wirkung haben.[8] Beteiligt waren u.a. Wissenschaftsorganisationen, Hochschulen, DFG, Bibliotheken und auch die Deutsche Nationalbibliothek. Das zweite der acht miteinander in Beziehung stehenden Handlungsfelder beschäftigte sich mit Konzepten im Bereich Hosting/Langzeitarchivierung. Auch die hier gegebenen Empfehlungen im technischen Bereich umfassen die Festlegung verbindlicher Standards und Verfahren.

Im Rahmen der Gespräche und Kontakte mit vielfältigen Ablieferern konnte die Deutsche Nationalbibliothek – wie oben erwähnt – die Erfahrung machen, dass im Bereich der E-Journals wenig gültige Standards existieren, so dass die ausgesprochenen Empfehlungen von der Deutschen Nationalbibliothek ausdrücklich unterstützt werden.

3.2 Metadatenformate NLM-DTD und CrossRef

Zu den technischen Voraussetzungen zur Ablieferung von E-Journals gehört die Etablierung von einer überschaubaren Anzahl verwendeter Metadatenformate. Dies wäre ein wesentlicher Schritt in Richtung der automatischen Übernahme von Inhalten und Nachweisen auch für E-Journals. Neben dem erwähnten Datenformat der National Library of Medicine (NLM-DTD/-Schema) bedienen einige Verlage, vor allem diejenigen, die die DOI nutzen, auch das Schema von CrossRef. Die Untersuchung beider Metadatenformate hat ergeben, dass für die Speicherung von Zugriffs- und Objektinformationen auch im Hinblick auf die Bedürfnisse der Deutschen Nationalbibliothek alle wichtigen Informationen enthalten sind.

NLM Journal Archiving and Interchange Tag Suite

An der NLM wurde die sogenannte Journal Archiving and Interchange Tag Suite[9] mit dem Ziel entwickelt ein allgemeines Format bereitzustellen, mit dem Verlage und Archive Zeitschrifteninhalte austauschen können. Die Tag Suite besteht aus verschiedenen DTDs, eine für Bücher und drei für Zeitschriften, letztere unterschieden in eine sogenannte grüne DTD für die Archivierung (Archiving and Interchange Tag Set), eine blaue für Publikationszwecke (Journal Publishing Tag Set) und eine orangefarbene Authoring-DTD, die für Autorenarbeitsplätze eingesetzt werden kann. Alle DTDs erlauben sowohl die Strukturierung von Content als auch von Metadaten. Die zitierte Studie benennt die blaue Journal-Publishing-DTD als die am meisten

[8] vgl. Kommission Zukunft der Informationsinfrastruktur, 2011, S. 48

[9] http://dtd.nlm.nih.gov/ (Stand 02.01.2012)

verwendete. Hier können für einzelne Objekte sowohl Journal-Metadaten als auch Artikel-Metadaten angelegt werden.

Wichtige Metadatenelemente für die Strukturierung der Zeitschriftenartikel in den Nachweissystemen der Deutschen Nationalbibliothek sind die Ausgabebezeichnung, die Nennung von Autorinnen oder Autoren, der Titel, das Erscheinungsdatum und die Zuordnung zur zugehörigen Zeitschrift.

Die Ausgabebezeichnung ordnet einen Zeitschriftenartikel einer Ausgabe (issue) oder einem Band (volume) zu. In der Journal-Publishing-DTD sind die Elemente „issue" und „volume" Teil der Artikel-Metadaten.

Autorinnen oder Autoren, aber auch andere beteiligte Personen können über das Element „contrib" und ein zugehöriges Attribut transportiert werden.

Über die Struktur „article-meta" und das Container-Element „title-group" können verschiedene Arten von Titeln transportiert werden. Das Erscheinungsdatum wird im Element „pub-date" festgehalten, das wiederholbar ist. Über das dazugehörige Attribut „pub-type" können Veröffentlichungszeitpunkte verschiedener Formen transportiert werden, wie z.B. „EPub" für elektronische Publikation (nicht zu verwechseln mit dem Format-Standard EPUB).

Für die Zuordnung zu einem Zeitschriftentitel ist die Struktur „journal-meta" und das darin enthaltene Containerelement „journal-title-group" vorgesehen, das den Titel der Zeitschrift im Element „journal-title" mit weiteren Elementen für Untertitel, übersetzte Titel usw. klammert.

CrossRef

Verlage, die ihre Publikationen mit Digital Object Identifiern (DOI) versehen, müssen diese bei einer Agentur registrieren. Dies kann z.B. über CrossRef geschehen. Das Ziel aus Sicht von CrossRef bei der Ablieferung von Metadaten ist die Beschreibung des Objekts, auf das sich die DOI bezieht. Das CrossRef-XML-Metadata-Deposit-Schema[10] eignet sich daher nicht für die Strukturierung von Inhalten, aber für die Übertragung bibliografischer Metadaten und ist hierfür bei einigen Verlagen in Benutzung

Alle der oben benannten relevanten Metadaten können über CrossRef-Elemente ebenfalls abgeliefert werden. Genaue Angaben zur Ausgabebezeichnung werden über die Struktur „journal_issue" genauso abgebildet wie Angaben zur Zeitschrift selbst über die Struktur „journal_metadata". Angaben zu Artikeln können mit Angaben zu Titeln und Personen über die Struktur „journal_article" abgebildet werden.

[10] http://www.crossref.org/help/Content/CrossRef%20Schema/deposit_schema.htm (Stand 02.01.2012)

4. Ausblick

Es ist zu wünschen, dass möglichst viele Beteiligte im Publikations- und Archivierungsprozess die Notwendigkeiten erkennen, die sich für die langfristige und optimale Nutzbarkeit digitaler Angebote stellen, und sich über einheitliche technische Anforderungen verständigen. Die Voraussetzungen zur dauerhaften Nutzung der E-Journals müssen auch bei Lizenzierungsmodellen ins Bewusstsein gerückt werden. Erst die Bereitstellung von weitgehend standardisierten Metadaten und die Nutzung von Schnittstellen ermöglicht die effiziente Sicherstellung des dauerhaften Zugriffs auf den relevanten digitalen Content. Hierzu müssen nicht nur gut funktionierende Ingest-Prozesse aufgebaut werden, sondern es ist auch notwendig, sich ständig darüber auszutauschen. Bibliotheken müssen sich untereinander verständigen, und Verlage müssen mit Bibliotheken das Gespräch suchen, um die Notwendigkeiten beider Seiten verstehen zu können und Standards zu vereinbaren.

Literaturverzeichnis

Deutsche Forschungsgemeinschaft im Auftrag der Allianz der deutschen Wissenschaftsorganisationen (2009): Dauerhaften Zugriff sicherstellen. Auf dem Weg zu einer nationalen Strategie zu Perpetual Access und Hosting elektronischer Ressourcen in Deutschland. Charles Beagrie Limited in Zusammenarbeit mit Globale Informationstechnik GmbH. Online verfügbar unter http://www.allianzinitiative.de/fileadmin/hosting_studie_d.pdf (Stand 02.01.2012).

Kommission Zukunft der Informationsinfrastruktur (2011): Gesamtkonzept für die Informationsinfrastruktur in Deutschland. Empfehlungen der Kommission Zukunft der Informationsinfrastruktur im Auftrag der Gemeinsamen Wissenschaftskonferenz des Bundes und der Länder. Online verfügbar unter http://www.leibniz-gemeinschaft.de/?nid=infrastr (Stand 02.01.2012).