Ein generisches Metadatensystem für den edoc-Server

Uwe Müller
u.mueller@cms.hu-berlin.de

Bibliografische Daten bilden den Zugangsschlüssel für den Bestand einer jeden Bibliothek. Das Gleiche gilt analog für Metadaten in Bezug auf elektronische Dokumente. Für die Erfassung und Verwaltung dieser Metadaten auf dem edoc-Server sowie für die Recherche und Präsentation wurde ein Metadatensystem entwickelt, das in diesem Artikel vorgestellt wird.

Metadaten – was ist das?

Als Metadaten – »Daten über Daten« – bezeichnet man strukturierte und typisierte Informationen, die nicht für sich selbst stehen, sondern andere Daten – typischerweise elektronische Ressourcen – beschreiben. Auch wenn der Begriff relativ neu ist und zumindest seine heutige Verbreitung erst mit der Entwicklung des World Wide Web einherging, basieren die dahinter liegenden Konzepte im Wesentlichen auf jahrhundertealter bibliothekarischer Praxis: Vor allem aus Gründen der Auffindbarkeit, aber auch um die wichtigsten Informationen über Bücher und Zeitschriften auf einen Blick zur Verfügung haben zu können, wurden und werden alle in einer Bibliothek vorhandenen Werke nach inhaltlichen und formalen Kriterien erschlossen und katalogisiert [1]. Dabei werden nicht nur Informationen aus der jeweiligen Publikation in ein – heutzutage elektronisches – Bibliothekssystem übertragen, die in dieser schon explizit enthalten sind (z. B. der Titel, der Name der Autoren oder der Verlag). Eine vollständige Sacherschließung beinhaltet auch die Einordnung der Werke in eine Klassifikationssystematik (z. B. RVK[1] oder DDC[2]) und eine verbale Beschreibung, die frei oder auf der Grundlage kontrollierter Vokabularien (z. B. Schlagwortnormdatei[3] oder Thesauri) erfolgen kann.

Modell und Format

Metadaten können in unterschiedlichen Metadatenmodellen erfasst werden. Beispielsweise hat sich für die Beschreibung von Webressourcen *Dublin Core*[4] als Standard etabliert – ein Modell, das flach strukturiert ist und mit wenigen Elementen auskommt [2]. Damit ist es einfach implementier- und interpretierbar und ermöglicht so eine hohe Interoperabilität. Allerdings stößt man auch relativ schnell an die Grenzen dessen, was mit diesem einfachen Modell ausgedrückt werden kann. So ist es nicht verwunderlich, dass sich im Laufe der Zeit eine Vielzahl spezialisierter Metadatenmodelle mit höherer Ausdrucksmächtigkeit für bestimmte Anwendungsbereiche bzw. Fachgebiete entwickelt hat: *Qualified Dublin Core*, LOM[5] für Lehr- und Lernmaterial, EAD[6] für den Archiv- und Museumsbereich, ONIX[7] für den Austausch bibliografischer Produktdaten sowie *MetaDiss*[8] für Dissertationen sind nur einige ausgewählte Beispiele.

Während Metadatenmodelle eine abstrakte Beschreibung der Datenstruktur bilden, erfolgt die konkrete Darstellung in so genannten Metadatenformaten.

1 Regensburger Verbundklassifikation, wird z. B. an der UB der Humboldt-Universität verwendet, siehe http://www.bibliothek.uni-regensburg.de/Systematik/systemat.html
2 Dewey Decimal Classification, bildet u. a. die Grundlage für die Sachgruppen der Deutschen Nationalbibliografie, siehe z. B. http://www.ddc-deutsch.de/
3 http://www.ddb.de/professionell/swd.htm
4 www.dublincore.org/
5 Learning Object Metadata, siehe http://ltsc.ieee.org/wg12/
6 Encoded Archival Description, siehe http://www.loc.gov/ead/
7 Online Information eXchange, siehe http://www.editeur.org/
8 entwickelt durch *Die Deutsche Bibliothek* und das Projekt *Dissertationen Online*, siehe http://deposit.ddb.de/metadiss.htm

Auch wenn die Begriffe oft synonym verwendet werden und viele Standards Modell und Format als Einheit definieren, handelt es sich um zwei unterschiedliche Ebenen: Beispielsweise lässt sich das Dublin-Core-Modell in Form eines XML-Schemas repräsentieren (etwa für das OAI-Protokoll, siehe Seite 55 [3]). Ebenso können aber auch eine RDF-Syntax, die in HTML üblichen Metatags [2] oder eine Datenbank für die Darstellung von DC-Metadaten verwendet werden.

Allerdings können die anachronistischen Zettelkataloge nicht als Metadatenformat im engeren Sinne betrachtet werden. Denn Tim Berners-Lee, der Erfinder des WWW, definiert: »Metadaten sind maschinenlesbare Informationen über elektronische Ressourcen oder andere Dinge.«

Metadaten für edoc

Zur Beschreibung der auf dem Dokumenten- und Publikationsserver verfügbaren Veröffentlichungen ist eine Vielzahl von Metadaten erforderlich. Dabei ist natürlich in erster Linie an die Attribute der klassischen Sach- und Formalerschließung zu denken: Titel, Autoren, Erscheinungsdatum, Sprache, Schlagwörter, Einordnung in genormte Sachgruppen, Abstracts und die URL, unter der der Volltext abgerufen werden kann. Diese so genannten beschreibenden Metadaten dienen der Auffindbarkeit der Publikationen vermittels Recherchewerkzeugen wie Suchmasken oder dynamischer Browsingseiten (siehe Abb. 4) sowie der Generierung von Startseiten, die neben dem Link zum Volltext die wichtigsten Informationen der Veröffentlichung enthalten (siehe Abb. 5). Daneben sind aber auch administrative Angaben wie das letzte Änderungsdatum des Datensatzes, die Größe und Formatversion von Dateien und rechtliche Informationen Bestandteil der Metadaten, die auf dem edoc-Server erfasst werden.

Das für den edoc-Server verwendete Datenmodell orientiert sich an Dublin Core, enthält aber im Vergleich zu diesem einfachen Standard zahlreiche Erweiterungen. Neben der Vielfalt der Metadatenattribute bietet das entwickelte

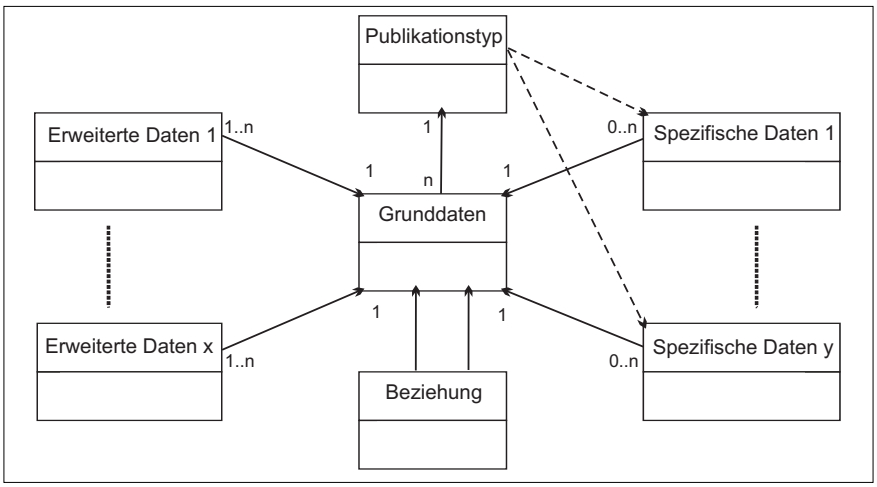

Abb. 1: Prinzipschema des Metadatenmodells für den edoc-Server.

Datenmodell und die darauf basierende Datenbank die Möglichkeit, dass

- bestimmte Attribute pro Datensatz mehrfach auftreten können (z. B. Autor, Schlagwort, URL),
- die jeweils verwendeten Metadatenattribute von der Art der Publikation abhängen können und das Datenmodell für die Integration neuer Publikationstypen einfach erweitert werden kann,
- Beziehungen zwischen einzelnen – in Dublin Core unabhängigen – Elementen dargestellt werden können (z. B. zwischen Format und URL),
- Beziehungen zwischen mehreren Publikationen wie beispielsweise hierarchische Zusammenhänge modelliert werden können.

Die Abbildung 1 zeigt die grundsätzliche Struktur des Datenbankmodells, mit dem die Metadaten des edoc-Servers verwaltet werden. Die Kerndaten jeder Publikation werden in der zentralen Tabelle *Grunddaten* gespeichert. Darunter befinden sich neben dem eindeutigen Identifikator vor allem administrative Daten, die sich auf den Status des Dokuments und der Metadaten selbst beziehen. Auch der Typ der Publikation (z. B. Dissertation, Zeitschriftenartikel) ist hier hinterlegt. Deren Eigenschaften werden in der Tabelle *Publikationstyp* gespeichert. Dazu zählen insbesondere die jeweils einzubindenden Spezialisierungstabellen, die auf der rechten Seite der Abbildung angedeutet sind. So treten beispielsweise Gutachter nur in den Metadaten von Qualifizierungsarbeiten (Dissertationen, Habilitationsschriften usw.) auf, während Heftnummern nur für periodisch erscheinende Publikationen (Zeitschriften, Schriftenreihen usw.) erfasst werden. Auf der linken Seite sind die übrigen Metadatenelemente angedeutet, die üblicherweise für sämtliche Arten elektronischer Publikationen auftauchen – z. B. Autoren, Titel, Schlagwörter oder die Links zu den Volltextversionen. In der Tabelle *Beziehung* werden Zusammenhänge zwischen mehreren Publikationen abgebildet – u. a. zu welchem Konferenzband ein Paper gehört oder auf welches Werk sich eine Rezension bezieht.

Architektur des Metadatensystems

Neben den Dissertationen werden auf dem edoc-Server inzwischen zahlreiche weitere elektronische Publikationen veröffentlicht. Vor allem deren unterschiedliche und vergleichsweise komplexe Metadatenmodelle machten die Entwicklung eines neuen Metadatensystems erforderlich. Der prinzipielle Ansatz für ein modulares und erweiterbares Datenmodell ist bereits im vorhergehenden Abschnitt beschrieben worden. Natürlich ist das Problem mit dem Entwurf und der Realisierung einer solchen Datenbank nur zum Teil gelöst. Denn ein vielgliedriges und modulares Datenschema hält auch für die Entwicklung von Werkzeugen zur Datenerfassung und -verwaltung sowie zur Recherche und Präsentation einige Herausforderungen bereit:

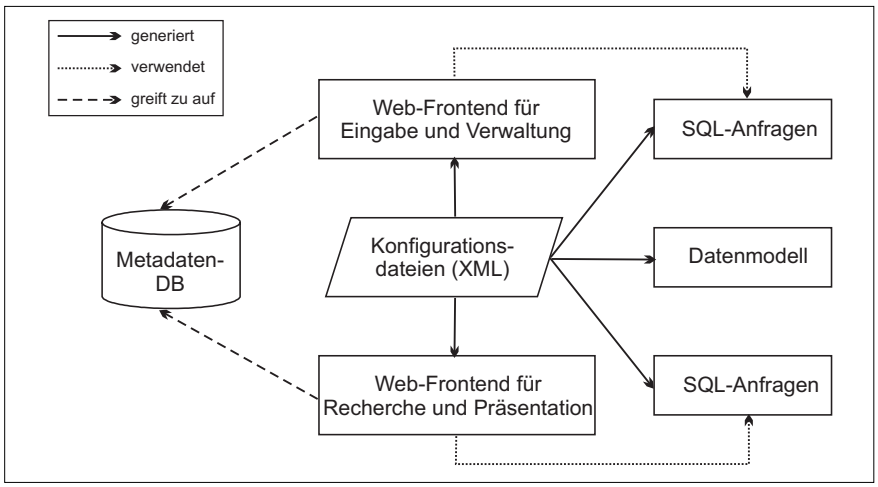

Abb. 2: Arbeitsweise des konfigurierbaren Metadatensystems.

Im Grunde genommen müssen die Erfassungsmasken und die Präsentationsansichten für jeden einzelnen Publikationstyp entwickelt werden. Denn diese unterscheiden sich nicht nur in Bezug auf das Metadatenmodell, sondern auch in der jeweiligen Darstellung im Web.

Daraus ergab sich die Motivation, nicht jeden Anwendungsfall des Metadatensystems separat zu implementieren, sondern ein generisches System zu entwickeln und die spezifischen Eigenschaften der einzelnen Publikationstypen in Konfigurationsdateien auszulagern. Die Funktionsweise des Metadatensystems ist in Abb. 2 dargestellt. Aus den in XML kodierten Konfigurationsdateien werden das jeweilige Datenmodell und die dazugehörigen SQL-Anfragen erzeugt sowie die Weboberflächen generiert. Dazu zählen neben den Erfassungs- und Verwaltungsmasken für die bibliothekarischen Mitarbeiter vor allem die Recherche- und Präsentationsansichten bis hin zu den Einstiegsseiten für die einzelnen Publikationen.

MetaIn: Erfassung und Verwaltung

Um die Eingabe und Bearbeitung der in der Datenbank gespeicherten Metadaten über eine Webschnittstelle zu ermöglichen, wurde mit MetaIn ein generisches Werkzeug zum Erstellen von Eingabemasken für komplexe Datenstrukturen entwickelt. Dabei wird dem Mitarbeiter durch eine entsprechende Konfiguration eine für die Datenverwaltung optimierte Sicht auf das Datenmodell präsentiert (siehe Abb. 2 oben). *MetaIn* kann allein durch Modifikation der XML-basierten Konfigurationsdateien an sämtliche Publikationstypen angepasst werden und somit auch für zukünftige, zur Entwicklungszeit noch nicht spezifizierte Publikationsvorhaben eingesetzt werden. Neben der Konfigurierbarkeit besteht die wesentliche Eigenschaft von MetaIn darin, dass der Erfassungsprozess linearisiert wird und dem Benutzer somit eine Bearbeitungsreihenfolge vorgeschlagen wird. Die Konfigurationssprache erlaubt es mithin auch, genau zu spezifizieren, unter welchen Bedingungen welche Eingabemaske erscheinen soll. Allerdings besteht auch die Möglichkeit, diesen Ablauf zu verlassen und frei zu navigieren.

Abb. 3 zeigt eine Eingabemaske für die Erfassung der Metadaten von Dissertationen. Für diesen Publikationstyp können, wie hier zu sehen, beispielsweise auch Gutachter und Betreuer in den Metadatensatz aufgenommen werden. Auf der linken Seite ist die aktuelle Position im Eingabeprozess dargestellt. Dort lässt sich die lineare Abfolge auch jederzeit unterbrechen und an anderer Stelle fortsetzen. Die eingegebenen Metadaten werden erst nach nochmaliger Kontrolle durch den Bearbeiter unter dem Menüpunkt *Speichern* in die Datenbank übernommen.

MetaOut: Suche und Browsing

Die für die Internetrecherche gängigste Methode zum Auffinden von Ressourcen ist die Verwendung von Suchmasken – z. B. bei Suchmaschinen, aber auch auf einzelnen Webservern – die in Form einer einfachen oder einer so genannten erweiterten Suche daherkommen. Diese Möglichkeit der Recherche in den Metadaten bietet auch der edoc-Server. Neben einer strukturierten Suchmaske, mit der durch die explizite Eingabe von Suchbegriffen gezielt nach Veröffentlichungen gesucht werden kann, sind aber auch alle Publikationen über einen so genannten

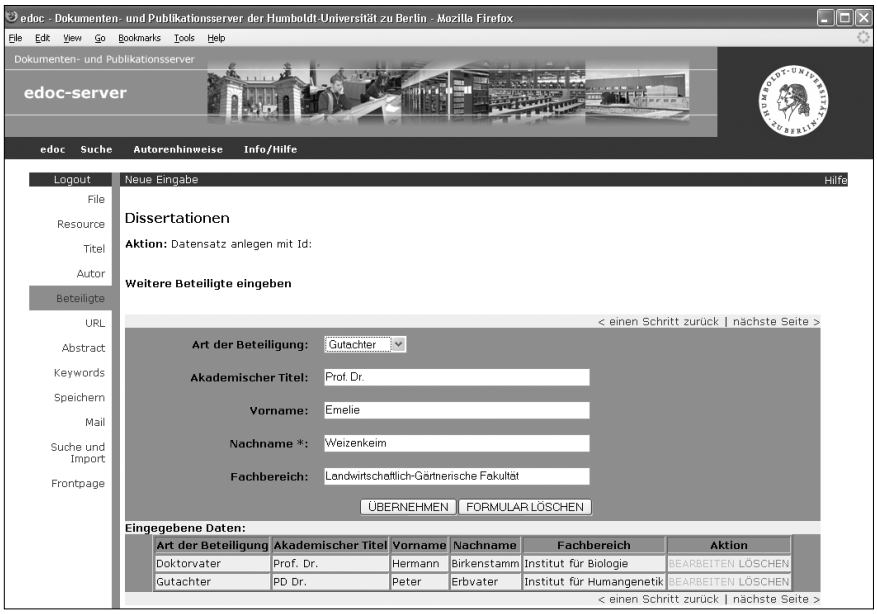

Abb. 3: MetaIn – Eingabemaske für Dissertationen.

Abb. 4: Beispiel einer Browsingansicht mit Dissertationen aus dem Bereich Innere Medizin.

Browsing-Zugang erreichbar. Damit sind Auflistungen von Publikationen nach unterschiedlichen Kriterien gemeint, die sich durch eine intuitive Navigationsstruktur verändern und konkretisieren lassen und so ohne die Verwendung von Suchformularen direkt zu den einzelnen Dokumenten führen. Je nach Art der Publikation und den jeweils erfassten Metadaten lässt sich dieser Webzugang individuell gestalten und anpassen.

In Abb. 4 ist ein Screenshot zu sehen, der einen fachlich orientierten Zugang zu den Dissertationen auf dem edoc-Server zeigt – in diesem Falle nach der Regensburger Verbundklassifikation (RVK). Oben ist der bisherige Suchpfad zu sehen, an dessen einzelne Positionen man durch einen einfachen Klick auch jederzeit zurückkehren kann. Auf der linken Seite ist der aktuelle Navigationsbaum dargestellt, über den die Recherchekriterien verfeinert werden können. Im rechten Hauptbereich werden die Publikationen angezeigt, die sich innerhalb des Suchpfades befinden. Zusätzlich dazu wird hier eine Einschränkung der Auswahl über den Anfangsbuchstaben des Autors angeboten.

Als weitere systematische Einstiege werden für Hochschulschriften die nur auf einer Hierarchieebene definierten Sachgruppen der Deutschen Nationalbibliografie gemäß der Dewey Decimal Classification (DDC) sowie die Universitätsstruktur, also die Fakultäten und Institute der Autoren, angeboten. Für periodisch erscheinende Publikationen und Sammelbände können darüber hinaus nach Erscheinungsdatum geordnete Listen und Inhaltsverzeichnisse für die einzelnen Ausgaben generiert werden. Aussehen und Struktur sämtlicher Ansichten sowie die Beziehung zum Datenmodell und der Datenbank werden in XML-Dateien und Xipe-Templates[9] konfiguriert. Dasselbe gilt auch für die Einstiegsseiten zu den einzelnen Publikationen (siehe Abb. 5) und die Webformulare zur Verwaltung der Metadaten.

Implementation und Schnittstellen

Sowohl MetaIn als auch MetaOut sind vollständig in PHP[10] implementiert worden. Dabei wurde weitgehend auf die Möglichkeiten der objektorientierten Programmierung zurückgegriffen. Als Template-System kam Xipe zum Einsatz. Für die Konfiguration der Systeme wurden XML-basierte Konfigurationssprachen entwickelt, die ohne tiefe Programmierkenntnisse angewendet werden können. Als Datenbank für das laufende System auf dem edoc-Server wird das am Computer- und Medienservice betriebene Sybase-DBMS verwendet. Der Betrieb des Metadatensystems mit mysql ist jedoch ohne weiteres möglich.

Da sich das Metadatensystem in den Gesamtzusammenhang des Publikationsprozesses eingliedern muss, verfügt es neben den beschriebenen Hauptbestandteilen MetaIn und MetaOut über zahlreiche weitere Schnittstellen.

Einige der zu erfassenden Metadaten bestimmter Publikationstypen (z. B. der Dissertationen) werden bereits zuvor in das Katalogsystem der Universitätsbibliothek eingegeben. Um Fehler und unnötige Arbeitsschritte zu vermeiden, kann ein bestehender Datensatz über das MAB-Format[11] in MetaIn importiert werden. In Abb. 3 verbirgt sich diese Funktion unter dem Menüpunkt *File*.

Eine weitere in die Erfassung integrierte Schnittstelle ist das Hochladen von Dateien. Dieses wird im Zusammenhang mit der Eingabe der URL (siehe Abb. 3) realisiert: Die automatisch vorgeschlagene und manuell anpassbare Adresse wird beim Upload dazu verwendet, die hochgeladenen Dateien an die korrekte Stelle auf dem Server zu kopieren. Anschließend wird automatisch überprüft, ob der Vorgang erfolgreich war und die eingegebene URL tatsächlich auf die hochgeladene Datei verweist.

Auch die gesetzlich geregelte Weitergabe der Metadaten an *Die Deutsche Bibliothek* (DDB) ist innerhalb von MetaIn realisiert und kann entsprechend konfiguriert werden. Zukünftig soll dieses Meldeverfahren, das bisher über E-Mail abgewickelt wird und damit einem *push*-Ansatz folgt, über OAI realisiert werden (siehe Seite 55 [3]). Die Deutsche Bibliothek holt sich dann die Informationen über eine OAI-Schnittstelle von den einzelnen Hochschulservern ab (*poll*).

Diese Methode ist auf einem anderen Gebiet bereits Realität: Zum Datenexport an den ProPrint-Dienst sowie an andere so genannte Service Provider besitzt das Metadatensystem eine OAI-Schnittstelle,

9 Xipe ist ein PHP-basiertes Template-System für die standardisierte HTML-Ausgabe und Teil von PEAR, der bekanntestes Sammlung von PHP-Programmbibliotheken, siehe http://pear.php.net/package/HTML_Template_Xipe

10 verwendet wurde die Version 4.3

11 Austauschformat für bibliografische Daten, siehe http://www.ddb.de/professionell/mab.htm

die das interne Metadatenmodell auf die jeweils angeforderten Modelle abbildet (siehe dazu die Artikel auf den Seiten 55 [3] und 65 [4]).

Auch die auf dem edoc-Server erzeugten URNs[12] und die entsprechenden URL-Zuordnungen werden durch Die Deutsche Bibliothek bereits über eine OAI-Schnittstelle abgefragt. Die DDB betreibt mithilfe dieser Daten unter anderem einen Resolver[13], der auch als PlugIn für den Webbrowser von Mozilla zur Verfügung steht[14]. Die Erzeugung der URNs innerhalb des edoc-Namensraums[15] und die Zuordnung zu den einzelnen URLs erfolgt vollautomatisch durch Datenbank-Trigger und stellt ebenfalls eine Schnittstelle zum Metadatensystem dar. Die URNs werden auf der Einstiegsseite neben den Links zum Volltext angezeigt (siehe Abb. 5).

Abb. 5: Einstiegsseite zu einer Publikation mit bibliografischen Daten und Verweis zum Volltext.

Fazit und Ausblick

Das Metadatensystem stellt eine weithin konfigurierbare und somit generische Datenbankanwendung dar, mit der die Publikationen auf dem edoc-Server und deren Metadaten flexibel verwaltet und präsentiert werden können. Es stellt komfortable Eingabemasken und eine übersichtliche und intuitive Recherche-schnittstelle bereit und kann für neue Publikationstypen angepasst werden, ohne dass ein Eingriff in den Programmcode notwendig wird.

Wie andere Softwareentwicklungen auch, ist das Metadatensystem in erster Linie für den Eigenbedarf, also für die unmittelbare Verwendung auf dem edoc-Server entwickelt worden. Allerdings ist der Anpassungsaufwand für die Nachnutzung des Systems wegen des generischen Ansatzes als vergleichsweise gering einzuschätzen. Dieser Aufwand hängt vor allem von der Art des zu verwendenden Datenmodells ab und umfasst im Wesentlichen die Erstellung eigener Konfigurations- und Templatedateien. Eine vollständige Nachnutzung des Systems wird derzeit an der HTWK Leipzig[16] vorbereitet.

Institutionen, die beabsichtigen, einen Dokumentenserver neu aufzusetzen, finden in dem beschriebenen Metadatensystem eine gute technologische Basis.

Auch wenn es auf den ersten Blick vielleicht verwundern mag: Das für den edoc-Server entwickelte Metadatensystem stellt Funktionen bereit, die in klassischen Bibliothekssystemen nicht implementiert sind, und ist daher als eigenständiges System neben dem Online-Katalog der Universitätsbibliothek notwendig. Das betrifft nicht in erster Linie das Datenmodell selbst. Aber die generischen Erfassungsmasken und Browsingseiten, das konfigurierbare Erscheinungsbild – kurz: das Gesicht des Dokumenten- und Publikationsservers – sowie die beschriebenen Schnittstellen lassen sich mit dem in der Universität verwendeten System[17] nicht oder nur mit sehr hohem Aufwand realisieren. Auf der anderen Seite schlagen die Nachteile der teilweise doppelten Datenhaltung zu Buche. Entsprechende Schnittstellen für den Abgleich von Datensätzen müssen noch realisiert werden.

Ebenso stehen die Integration der Erstellung und Verwaltung digitaler Signaturen in das Metadatensystem sowie die Unterstützung von Open-Access-Funktionen auf der Agenda. Letzteres erfordert unter anderem, dass auch externe Benutzer – Autoren und Herausgeber – schreibend auf das Metadatensystem zugreifen können. Hierfür sind entsprechende Erweiterungen erforderlich.

Literatur

[1] Oehlschläger, S.: Metadaten – Alter Wein in neuen Schläuchen? Abschlussworkshop des Projektes META-LIB und 1. Metadaten-Workshop der Arbeitsstelle für Standardisierung Der Deutschen Bibliothek. *Zeitschrift für Bibliothekswesen und Bibliographie*. 2003 (4).

[2] Müller, U.: Dublin Core: Metadaten für WWW-Inhalte. *RZ-Mitteilungen* Nr. 20, Juni 2000, S. 41–42. URL: http://edoc.hu-berlin.de/docviews/abstract.php?id=20281

[3] Müller, U.: OAI-Protokoll: Metadaten für alle. *cms-journal* 27, August 2005, S. 55–58.

[4] Schulz, M.: ProPrint – Der Print-on-Demand-Service für Dokumenten- und Publikationsserver. *cms-journal* 27, August 2005, S. 65–68.

12 Uniform Resource Name: URI mit dem Schema urn, realisiert dauerhafte, ortsunabhängige Bezeichner für Internetressourcen.
13 http://www.persistent-identifier.de/?link=610
14 http://www.persistent-identifier.de/?link=550
15 urn:nbn:de:kobv:11-100
16 Hochschule für Technik, Wirtschaft und Kultur, Fachbereich Buch und Museum, http://www.htwk-leipzig.de/bum/
17 verwendet wird das Bibliothekssystem ALEPH 500™ von der Firma Ex Libris