SLUB
Wir führen Wissen.

Medienübergreifende Repositorien

Mehr als „nur“ Dokumentenserver

05.06.2014
Ralf Claußnitzer

Gesamtlösung für alle digitalen Objekte
Anforderung: Hochschulschriftenserver

- Entwicklung von Pre-Print, über Post-Print-Volltext zu Gold Open Access
- Fokussierung auf Dissertationen im PDF Format
  - Beispiel Ablieferungsform für Harvesting: Genau eine Transfer-URL, sonst ZIP-Datei
  - Ausdruckbar, Analysierbar - im Fall von PDF-A – Archivierbar
  - Anforderungen der Katalogisierung (Verbund und DNB)
  - Rechteübertragung verhältnismäßig unkompliziert
- Sekundär auch Bilder und Tondokumente, DVD
  - Angabe von Hochschul-spezifischen Metadaten über Betreuer und Hochschullehrer, etc.
- Oft Einbettung in Webseiten der UNI
  - Anpassungen am Quelltext (daher primär Webtechnologien mit geringer Einstieghürde, wie PHP MYSQL…)

Gesamtlösung für alle digitalen Objekte

Anforderung: Open Access Publishing Plattform

- Weniger qualifizierte erschlossene Beiträge bedeuten mehr Aufwand für die Bearbeiter, mehr Nachfragen, mehr Änderungen
- Unter Umständen müssen Publikationsanfragen abgelehnt werden (keine Thematische Einschränkung, aber inhaltliche)
- Unklare Lizenzsituation bzw. Patentsituation kann zum Veranlassen einer Sperrung führen
- Häufig müssen Embargofristen beachtet werden (Konflikt mit Abgabe und Ablieferungspflicht DNB)
- Neue Veröffentlichung -> neue bibliothekarische Metadaten (schwierig bei einfachen Datenbankkonstruktionen)
- Dubletten (Erkennung? Repository muss potentielle Kandidaten finden und den Administratoren vorschlagen)

Gesamtlösung für alle digitalen Objekte
Anforderung: Elektronische Pflichtexemplare (1)

- Gesetzliche geforderte Abgabe aller elektronisch Publikationen an die Bibliothek
- Basis: Pflichtexemplarrecht

„Zweck des Pflichtexemplarrechtes ist heute vorrangig die möglichst vollständige Archivierung aller Veröffentlichungen eines Landes als Zeugnis des kulturellen Schaffens, ihre bibliografische Dokumentation und die Zugänglichmachung für die Allgemeinheit. *Die Bibliotheken sind deshalb gesetzlich dazu verpflichtet, Pflichtexemplare auf unbegrenzte Zeit aufzubewahren und eine Nationalbibliographie zu erstellen.*“

(*) http://de.wikipedia.org/wiki/Pflichtexemplar

Gesamtlösung für alle digitalen Objekte
Anforderung: Elektronische Pflichtexemplare (2)

§ 7 Beschaffenheit von Netzpublikationen und Umfang der Ablieferungspflicht

(1) *Unkörperliche Medienwerke (Netzpublikationen) sind in marktüblicher Ausführung und in mit marktüblichen Hilfsmitteln benutzbarem Zustand abzuliefern.* Eine Pflicht zur Ablieferung besteht nicht, wenn die Ablieferungspflichtigen im Rahmen des § 16 Satz 2 des Gesetzes über die Deutsche Nationalbibliothek mit der Bibliothek vereinbaren, die Netzpublikationen zur elektronischen Abholung bereitzustellen. Für die Ablieferung von Netzpublikationen gilt § 2 Abs. 3 entsprechend; für die Bereitstellung zur elektronischen Abholung gilt § 2 Abs. 3 Satz 1 entsprechend.

(2) *Die Ablieferungspflicht umfasst auch alle Elemente, Software und Werkzeuge, die in physischer oder in elektronischer Form erkennbar zu den ablieferungspflichtigen Netzpublikationen gehören, auch wenn sie für sich allein nicht der Ablieferungspflicht unterliegen.* Dies gilt insbesondere für nicht marktübliche Hilfsmittel, die eine Bereitstellung und Benutzung der Netzpublikationen erst ermöglichen und bei den Ablieferungspflichtigen erschienen sind. Sie sind zusammen mit den Netzpublikationen abzuliefern oder zur elektronischen Abholung bereitzustellen.

http://www.gesetze-im-internet.de/pflav/__7.html

Gesamtlösung für alle digitalen Objekte
Anforderung: Elektronische Pflichtexemplare (3)

- Nicht nur Open Access Publikationen
- Portale mit unterschiedlich lizensiertem Inhalt
- Formatqualität? Archivierbarkeit?
- Rechtliche Bedingungen? Keine freie Verfügbarkeit bei gleichzeitiger Ablieferungspflicht über öffentliche Schnittstellen?
- Wie dem Pflichtexemplar-Gesetz genügen?
  - *Bibliothek kann Veröffentlichung im archivierbaren Format fordern, aber diese Forderung durchzusetzen ist oft nicht praktikabel*
  - Schulung und Hilfe anbieten

Gesamtlösung für alle digitalen Objekte
Anforderung: Digitalisate

- Digitalisate im Repository? Ja, weil: Digitales Objekt (Born Digital oder Digital Reformatting aka Retrodigitalisate)
- Behandlung bisher getrennt von Repositorien in „Digitalen Sammlungen“ und Bilddatenbanken
- Aber: Born Digitals und Retrodigitalisiate teilen sich viele Eigenschaften (URIs, Dateien, Präsentation im Web)
- Und: Technologisch im Grunde kein Unterschied
- Datenqualität deutlich homogener (Bilder, Beschreibungsformate wie METS/MODS)
- Metadatenerfassung meist ausreichend standardisiert
- DFG Richtlinien haben zu eingehender Standardisierung in diesem Bereich geführt
- Zu Beachten: Behandlung von Digitalisaten auf Grund ihrer Größe deutlich teurer, Synergien in der Administration helfen Aufwand insgesamt zu verringern

Sammlung vs. Langzeitarchivierung
Qualitätsabwägung

- Möglichst umfassende Sammlung von Netzpublikationen erfordert Verringerung der (technischen) Qualitätsanforderungen
- (echte) Langzeitarchivierung erfordert aber hohe Qualität
- Umwandlung nicht immer möglich (oder erlaubt)

Sammlung vs. Langzeitarchivierung
Service Levels

- Definition von Service Levels:

- LZA Ampelsystem
- Den Nutzer schon bei der Abgabe über den bereitstellbaren Service Level informieren
- Schlussfolgerung für den Export ins Archivierungssystem
- Prüfung durch Programme wie JHove (Dienste, bereitgestellt vom LZA System)
  http://www.dcc.ac.uk/resources/external/jhove2

Vorhandene Workflows vernetzen

- Workflows zur Bearbeitung von elektronischen Publikationen meist sehr individuell und Erfassungs-fokussiert
- Werkzeuge sind entsprechend spezialisiert
- Workflows für Digitalisate auf Durchsatz optimiert -> ohne Automatisierung ist dem Aufkommen eines Digitalisierungszentrums nicht beizukommen
- Produktion der digitalen Objekte unterschiedlich. Verwaltung und Präsentation hingegen sehr ähnlich!
- Technische Workflows:
  - Archivierung
  - Backup
  - Migration
  - Datenhaltung!
  - Bereitstellung

Verschiedene Präsentationsformate

- Liegen digitale Objekte in bekannten Formaten vor, können automatische (on-demand) Umwandlungen erfolgen, z.B. PDF als ePUB ausliefern
- Digitalisate werden bereits in herunterladbaren PDFs angeboten
- Inhaltsverzeichnisse, wenn Strukturmetadaten das hergeben
- Weiter denkbar:
  - Video Anzeige
  - Bildvorschau (Thumbnails)
  - Audio Streaming
  - OCR Text Einbettung

Anforderungen und Möglichkeiten

- Born-Digitals und Digitalisierungen zusammenlegen und gleichartige Objekte auch gleichartig behandeln
- Erweiterbarkeit: Metadatenstandards voll ausnutzen
- Behandlung sehr großer Dateien > 300Mb > 1Tb (Übertragung aus technischen Gründen oft beschränkt, Alternative aktiver Download)
- DRM Check, Einschätzung der Archivierbarkeit, Transparente Erklärung zu den Gründen, Möglicherweise Anpassung bei der Dokumenterstellung
- Nutzungseinschränkung: Allg.: Durchsetzen der Lizenzbestimmung. (In einigen Fällen nicht mit vertretbarem Aufwand zu realisieren)
- Nicht mehrere Backendsysteme einsetzen (z.B.: mehrere unterschiedliche Implementationen des OAI Protokolls,)
- Organisatorisch-Technische Vorteile ausnutzen

Notwendigkeit zur Erneuerung der Infrastruktur
Qucosa

- SLUB Repository Qucosa
  - Seit 2010
  - OPUS4 basiert (PHP, MySQL, Non-Standard Schema)
  - Mandantenunterstützung
  - TYPO3 Frontend
- Erfordert generischere Metadaten (keine Anwendungsfallbasierten)
- Konsequenter Einsatz von Standards
- Langzeitarchivierung mit bedenken
- Schnittstellen: Sehr unterschiedliche Anforderungen, Verbindung mit Durchsetzung von Lizenzen
- MediaTypes (application/mods+xml)!
- MediaTypes sind nicht nur andere Form der Dateiendung. Sie beschreiben auch das Protokoll

Stand der Entwicklung Qucosa

- Fedora 3.7.1 als Repository
  Generisch, Schemaunabhängig
- ElasticSearch 1.2.x
  Cluster-fähiger Suchindex, Lucene Core, REST API, Indexierung der Daten mit Fedora River (Eigenentwicklung)
- Legacy API
  Migration von OPUS4 Kern nach Fedora abgeschirmt durch Legacy API (setzt Qucosa Protokoll auf Fedora API Aufrufe um)
- Fedora 4 nächster großer Schritt, aber: Fundamental anderes Konzept.
  Migration von Fedora 3.x erst mit Fedora 4.1 geplant
- Intensivierter Einsatz Apache Camel und Apache ActiveMQ
  Generisches Messaging und Application Integration (SLUB Datenmanagement Plattform d:swarm für Suchindex)
- TYPO3 Frontend für Fedora 3 (mit Option 4) geplant

Nachnutzung der Gesamtlösung

- Lizenzierung
  - Qucosa Code GPLv3 (Legacy API nicht sinnvoll nachnutzbar)
  - Fedora Content Model als Anschauungsobjekt
  - ElasticSearch basierter Code Apache Licence 2.0

  - Quelltexte auf GitHub
    - https://github.com/slub/qucosa-webapi
    - https://github.com/slub/elasticsearch-river-fedora
    - https://github.com/slub/qucosa-fcrepo-contentmodel
    - TYPO3 Frontend soll auch dort veröffentlicht werden

Vielen Dank für die Aufmerksamkeit!

Ralf Claußnitzer

ralf.claussnitzer@slub-dresden.de