SLUB
Wir führen Wissen.

Qucosa Project

Aufbau eines Cross Media Repository

$ whoami

- Ralf Claussnitzer
  - Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB)
  - IT Abteilung seit 2008
  - Software Entwicklung und -Architektur
  - WWW
    -  @claussni
    -  github.org/claussni
    - 

Ausgangspunkt

- Institutionelles Repositorium Qucosa
  - Mandantenfähig
    - Universitäten, Hochschulen, Forschungsinstitute, Bibliotheken
  - Basierend auf OPUS 4 und TYPO3
    - Two-Tier: Web Service Backend + CMS Frontend
      - Einfache IR Software
      - Komplexes Content Management System
  - Status Quo
    - End-Of-Project = End-Of-Resources
    - Veraltetet PHP Codebasis
    - Nachfrage nach neuen/zeitgemäßen Features
    - Resource Discovery System hat Recherchemaske nicht abgelöst

Gesamtlösung für alle Digitalen Objekte - 1

- Anforderungen Hochschulschriftenserver
  - Entwicklung von Pre-Print, über Post-Print-Volltext zu Gold Open Access
  - Fokussierung auf Dissertationen im PDF Format
  - Sekundär auch Bilder und Tondokumente, DVD
  - Oft Einbettung in Webseiten der Universität
  - Integration mit anderen System „drangestrickt“
- Anforderungen der DNB
  - OAI-PMH
  - Spezielles Protokoll mit inhaltlichen Aspekten: xMetaDissPlus
- Pflichtexemplarrecht und Administrative Workflows

Gesamtlösung für alle Digitalen Objekte - 2

- Digitalisate
  - Behandlung bisher getrennt von Repositorien in „Digitalen Sammlungen“ und Bilddatenbanken
  - Born Digitals und Retrodigitalisiate teilen sich viele Eigenschaften (URIs, Dateien, Präsentation im Web)
  - Datenqualität deutlich homogener (Bilder, Beschreibungsformate wie METS/MODS)
  - Metadatenerfassung meist ausreichend standardisiert

Gesamtlösung für alle Digitalen Objekte - 3

- Sammlung vs. Langzeitarchivierung
  - Möglichst umfassende Sammlung von Netzpublikationen erfordert Verringerung der (technischen) Qualitätsanforderungen
  - (echte) Langzeitarchivierung erfordert aber hohe Qualität
  - Umwandlung nicht immer möglich (oder erlaubt)
- Definition von Service Levels
  - Nutzer schon bei der Abgabe Informieren

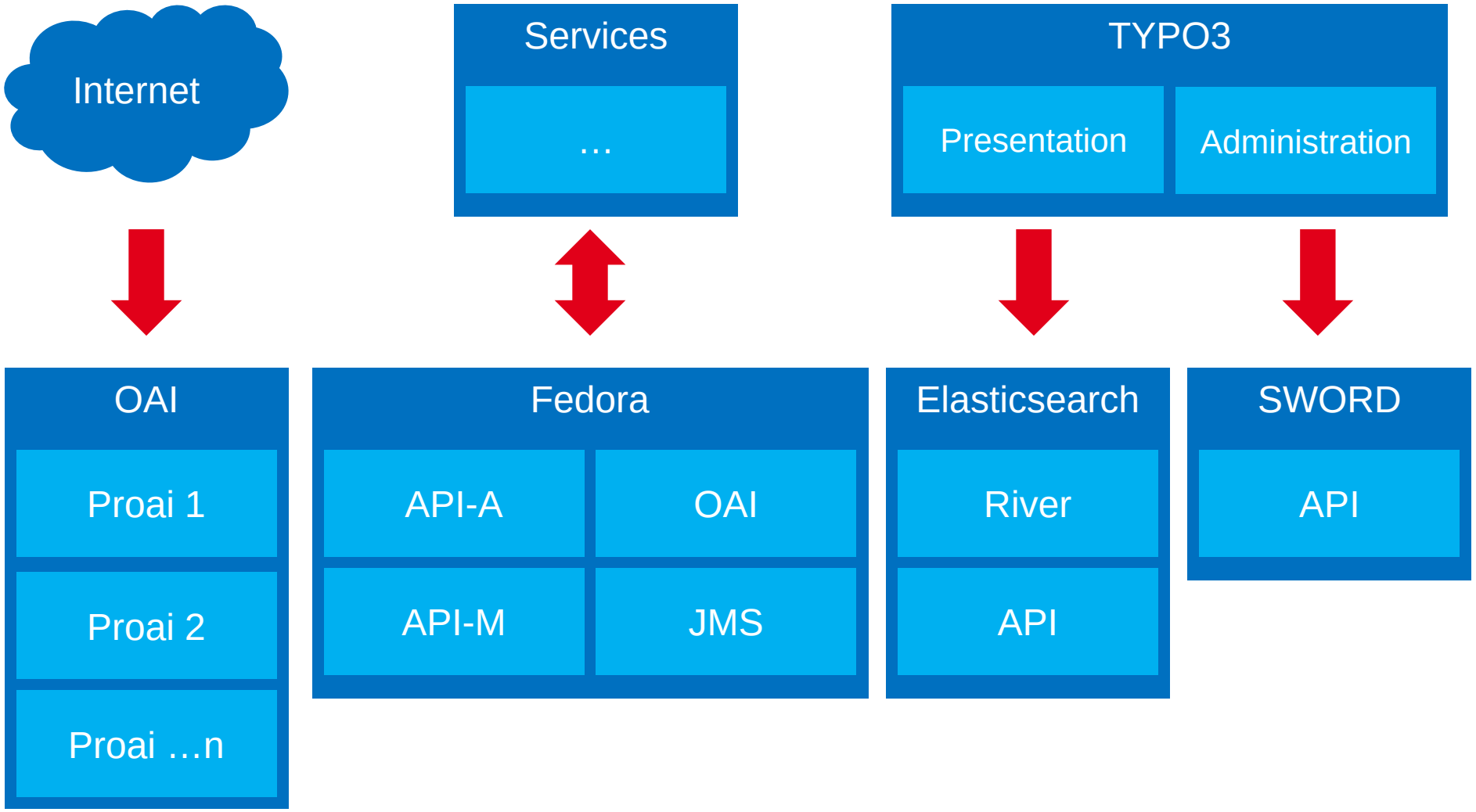
Perspektiven

- Born-Digitals und Retrodigitalisate gleichartig behandeln
  - Digitale Objekte mit Metadaten und Anhängen
- Metadatenstandards voll ausnutzen
  - Weg von Bibliothekarisch (/Katalog-) orientierten Datenbankschemata
- Organisatorisch-Technische Vorteile
  - Verringerung der Systemvielfalt
  - Weniger Technologien; Mehr Transparenz

Warum Fedora?

- Fedora fokusiert auf Infrastruktur
  - Out-of-the-box Software fokussiert IR Anwendungsfälle und GUI
    - Hohes Abstraktionsniveua
  - Loose gekoppelte Dienste als Systemmodell
  - Speicherung unabhängig vom Metadatenschema
- Unsere Anforderungen
  - Weiterentwicklung des Metadatenschemas
  - Vielzahl an Medientypen
  - Veraltetet Speicherinfrastruktur
    - Vollständige Migration ist keine Option
    - Möglichkeit von irreperablen Inkonsistenzen

Architekturansatz - 1



Architekturansatz - 2

- Weitere Verbesserungen
  - Konsequente Verwendung von Message Bus Konzepten
    - SOA mit Microservices
  - Ablösung des Elasticsearch River Index
    - ES Rivers werden nicht mehr unterstützt
  - Apache Camel / SLUB d:swarm Projekt
    - Datentransformation
    - Nachrichten Routing
  - Fedora 4.x
    - Workspaces (Entwürfe)
    - Hierarchische Datenbank
    - Federated Storage

Stand der Entwicklungen

- Fedora 3.8.0-RC1 als Repository
  - Generisch, Schemaunabhängig
- ElasticSearch 1.2.x
  - Cluster-fähiger Suchindex (Lucene Core, REST API)
  - Indexierung der Daten mit Fedora River (Eigenentwicklung)
- Fedora 4
  - nächster großer Schritt, aber: Fundamental anderes Konzept
  - Migration von Fedora 3.x erst mit Fedora 4.1 geplant
- Generisches Messaging und Application Integration
  - SLUB Datenmanagement Plattform d:swarm für Suchindex

Nachnutzung der Ergebnisse

- Open Source
  - Projektergebnisse auf GitHub
  - https://github.com/slub
- Lizensierung
  - GPLv3
    - TYPO3 Frontend
  - Apache License 2.0
    - Fedora Komponenten; Java Services
    - Elasticsearch River
  - Andere Open Source Lizenzen
    - Für Erweiterungen bestehender Fedora Commons Dienste

Thanks!

Questions?

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

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