Linked Data und Repositorien

Pascal-Nicolas Becker | Technische Universität Berlin | DINI Jahrestagung | Frankfurt, 27.10.2015

Digitale Repositorien

Ein Repositorium ist ein System zur sicheren Speicherung digitaler Objekte und der die Objekte beschreibenden strukturierten Metadaten.

Beispiele:

- Digitale Archive
- Institutionelle Repositorien (Preprints, Postprints, OA-Publikationen, …)
- Digital Bilddatenbanken
- Forschungsdaten-Repositorien
- …

Mehr als 2500 OA-Repositorien weltweit

Source: The Directory of Open Access Repositories, http://www.opendoar.org, abgerufen am 06.06.2014.

xxx.lanl.org / ArXiv.org

“Although the WorldWideWeb still represents only a small fraction of the overall usage, this access mode is expected to become dominant in the near future.”

Paul Ginsparg 1994

Quelle: Paul Ginsparg, *First Steps Towards Electronic Research Communication*. In: *Computer in Physics*, Vol. 8, No. 4, 1994, pp. 390-396.
Foto: Paul Ginsparg (privat)

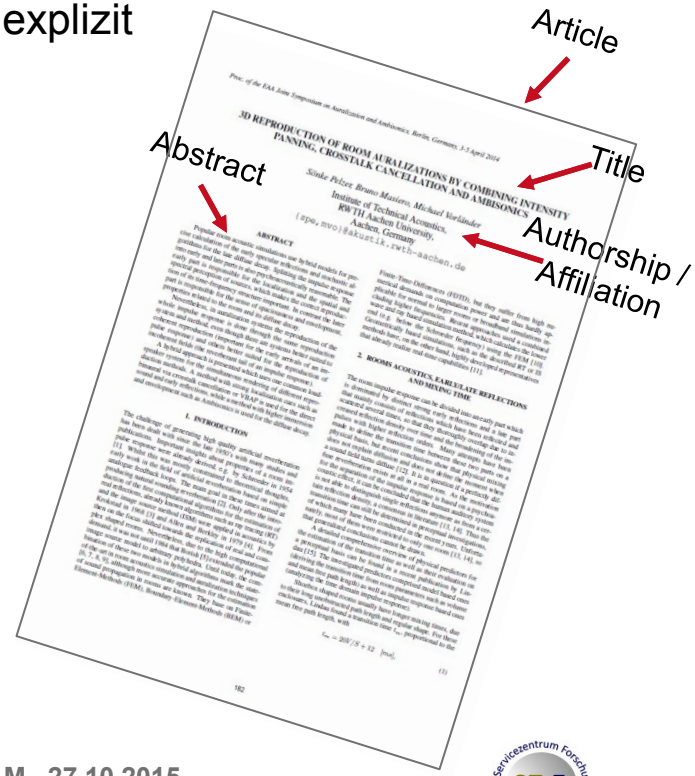
Bisheriger Datenaustausch mit Repositorien

- OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting): de-facto-Standard im Repositorienumfeld
- Aber auf das Repositorienumfeld begrenzt!
- Google unterstützt OAI-PMH seit 2008 nicht mehr (zuvor als Alternative zum Sitemap Protocol genutzt)
- „Nur“ eine Schnittstelle, kein Format

- ➢ Linked Data ist ein generischer, nativer Weg Daten auszutauschen
- ➢ Linked Data ist nicht auf das Umfeld von Repositorien beschränkt
- ➢ Linked Data ist selbst beschreibend
- ➢ Linked Data vereinfacht den Datenaustausch mit Repositorien

Repositorieninhalte und das Semantic Web

- Informationen im Internet sind auf den menschlichen Konsum ausgerichtet
- Viele Informationen werden nur implizit über den Kontext zur Verfügung gestellt
- Grundidee des Semantic Web: Implizite Informationen explizit machen, um sie automatisiert verarbeiten zu können



- Ein Repositorium „kennt“ den Titel eines Artikels
- Titel werden Fett und in größerer Schrift ausgegeben, Autorennamen kursiv
- Wieso stellen wir diese Daten nicht auch maschinell nutzbar bereit?

Repositorieninhalte sind besonders gut geeignet

Inhalte von Repositorien sind zur Bereitstellung im Semantik Web besonders gut geeignet:

- Metadaten existieren bereits in einer strukturierten Form
- Sie müssen nicht erst generiert oder aufwändig manuell zur Publikation als Linked Data erzeugt werden
- Es geht “lediglich” um die Konvertierung vorhandener Daten in RDF, um das Hinzufügen von Links und das Publizieren unter Berücksichtigung der Linked Data Principles

Was speichern Repositorien?

"Ein Repositorium ist ein System zur sicheren Speicherung *und Weitergabe* digitaler Objekte und der die Objekte beschreibenden Metadaten."

- Digitale Objekte
  - Eine oder mehrere Dateien: Dokumente (PDF, Text, …), Bilder (PNG, Tiff, …), Tabellen (CSV, …), Audio (Wave, OGG/VORBIS, …), Video, …
- Beschreibende Metadaten
  - Strukturierte Metadaten in Key-Value-Form: dc.title, dc.contributor.author, dc.description, dc.date.available, dc.subject.lcsh, dc.subject.ddc, …

Full metadata record

DC Field	Value	Language
dc.contributor.author	Lindau, Alexander	-
dc.date.accessioned	2014-02-27T14:19:30Z	-
dc.date.available	2014-02-27T14:19:30Z	-
dc.date.issued	2014-02-27	-
dc.identifier.uri	http://depositonce.tu-berlin.de/handle/11303/157	-
dc.identifier.uri	http://dx.doi.org/10.14279/depositonce-1	-
dc.description	The 'SAQI. Test Manual. v1.0' documents the complete German and English version of the SAQI. It serves as an user-oriented introduction giving valuable hints for practical application, e.g., by referring to the whisPER Matlab toolbox v1.8.0 which features a full implementation of a SAQI test. Additional resources are provided in a zip-container. It includes relevant project-related publications, illustrative audio examples, empirical test data sets, and Matlab functions for convenient later statistical analysis and plotting of SAQI test results. Folder structure in 'SAQI. Test Manual. v1.0. Additional files.zip': /1 references /2 audio files /3 mfiles Further related resources: http://www.ak.tu-berlin.de/saqi http://www.ak.tu-berlin.de/whisper	en_US

- Unmöglich die Dateien zu konvertieren (technische Probleme, manuell nicht leistbar)
- Aber wir können die Metadaten konvertieren und die Dateien verlinken!

Linked Data und Repositorien

- Bei der Entwicklung von Fedora 4 stand Linked Data im Mittelpunkt
  - Erschienen im Dezember 2014
- DSpace 5 wurde um die Bereitstellung der Metadaten als Linked Data erweitert
  - Erschienen im Januar 2015
- Linked Data Plattform als W3C Recommendatio im Februar 2015 verabschiedet
  - Spezifikation zur Nutzung von HTTP für Zugriff und CRUD-Operationen auf Servern, die ihre Inhalte als Linked Data anbieten.
- COAR Resource Type Vocabulary Draft im Juli 2015 vorgestellt
  - Linked Data Frontend bereits für den ersten Draft

Beispiel: DSpace 5

- DSpace ist die weltweit am häufigsten eingesetzte Software für OA-Repositorien
- DSpace 5.0 wurde im Januar 2015 veröffentlicht (aktuell: 5.3)
- Die TU Berlin hat Linked Data Support in DSpace 5 eingebracht:
  - Konvertierung der Metadaten in Linked Data
  - Ausgabe als RDF/XML, Turtle, N-Triples
  - SPARQL-Endpoint und Content Negotation
  - Stark konfigurierbar, anpassbar und erweiterbar
  - Standardkonfiguration mit ausgeliefert
- Test it yourself:
  http://demo.dspace.org/data/handle/10673/5/ttl
  http://demo.dspace.org/data/handle/10673/5/ttl?text
  wget -O - --header='Accept: text/turtle' http://demo.dspace.org/jspui/handle/10673/5

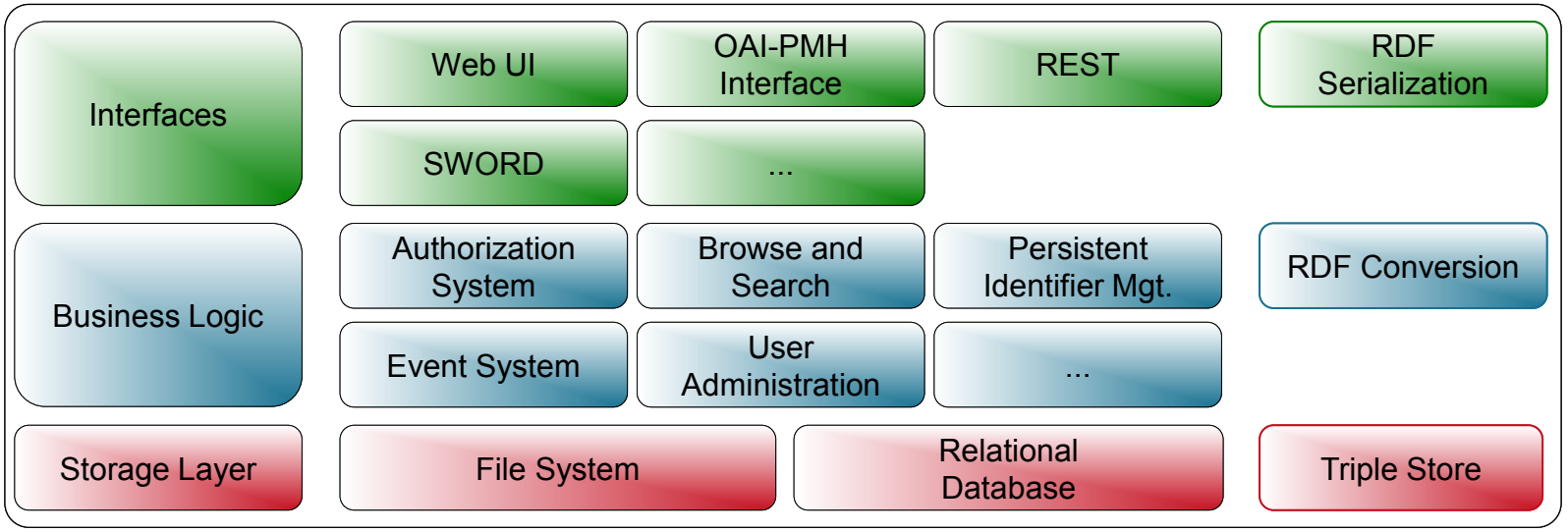
Wenn Sie DSpace 5.0 einsetzen, bitte schalten Sie den Linked Data Support ein!

Charakteristika von Repositorien

- Verschiedene Repositorien nutzen unterschiedliche Metadatenschemata
  - ➢ Die Konvertierung muss stark konfigurierbar und erweiterbar sein
- Für Metadatenschemata können bereits Vokabulare existieren (z.B. DC, LCSH, …)
  - ➢ Werte von Metadaten müssen in URIs/Links konvertiert werden können
- Repositorieninhalte ändern sich so gut wie nie (Zitierfähigkeit, Zuverlässigkeit), die Konvertierung kann aber zeitaufwändig sein
  - ➢ Daten konvertieren und in einem Cache speichern
- Repositorien erzeugen URIs zur Adressierung ihrer Inhalte
  - ➢ Bestehende URIs nachnutzen und um Content Negotiation erweitern
- Persistent Identifier (handle, DOI, …) verstoßen gegen die Linked Data Principles
  - ➢ Persistent Identifier sind in der Form als HTTP(S) URIs zu nutzen (z.B. http://hdl.handle.net/..., http://dx.doi.org/...)

Repositoriensoftware erweitern

- Hinzufügen eines Triple Stores
  - Nutzung als Cache für die konvertierten Daten
  - Zur Bereitstellung eines SPARQL-Endpoints
- Methoden entwickeln zur Konvertierung in RDF und zur Ergänzung von Links
- Ein Modul zur Bereitstellung der konvertierten Daten als RDF-Serialisierung erstellen
- Content Negotiation unterstützen



Konvertierung vorhandener Metadaten in RDF

- Repositoriensoftware um Metadatenschemata/Metadatenfelder erweiterbar
- Dublin Core wird oft genutzt, aber bei weitem nicht das einzige Schema

- ➢ Die Konvertierung muss stark konfigurierbar sein!
- ➢ Eine Konfiguration in RDF stellt die volle Funktionalität von RDF bereit
- ➢ „Reification“ zur Beschreibung der gewünschten Ergebnisse
- ➢ Platzhalter z.B. für URIs des Repositoriums
- ➢ Reguläre Ausdrücke um Literale und URIs aus Metadatenwerte zu generieren
- ➢ Es braucht ein Vokabular um solche Konfigurationen zu erstellen:
  DSpace Metadata RDF Mapping Vocabulary:

http://digital-repositories.org/ontologies/dspace-metadata-mapping/

Beispiel: DSpace Metadata RDF Mapping

```
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix dm: <http://digital-repositories.org/ontologies/dspace-metadata-mapping/0.2.0#> .
@prefix : <#> .

:title
    dm:metadataName "dc.title" ;
    dm:creates [
        dm:subject dm:DSpaceObjectIRI ;
        dm:predicate dcterms:title ;
        dm:object dm:DSpaceValue ;
    ] ;
.
```

```
:doi  dm:metadataName  „dc.identifier.doi" ;
        dm:condition  „^doi:“ ;
        dm:creates [
            dm:subject  dm:DSpaceObjectIRI ;
            dm:predicate  dc:identifier;
            dm:object [
                a dm:ResourceGenerator ;
                dm:modifier [
                    dm:matcher  „^doi:(.*)$“ ;
                    dm:replacement  „http://dx.doi.org/$1“ ;
                ] ;
                dm:pattern  „$DSpaceValue“ ;
            ] ;
        ] ;
.
```

Vermeidbare Fehler und Hindernisse

- URIs nachnutzen, wo immer möglich! Keine eigenen URIs erzeugen, wenn es bereits adäquate gibt
  - Beispiele: Für Klassifikationen, wie die Library of Congress Subject Headings (LCSH) oder die Dewey Decimal Classification (DDC), existieren bereits URIs
- Erzeugung von URIs nur für „eigene“ Entitäten, Erzeugung von URIs nur wenn ausreichend Informationen vorhanden sind
- Beispiel: Autoren
  - Können verschiedene Autoren gleichen Namens differenziert werden?
  - Sollten Autoren nicht eigene URIs erzeugen und hinterlegen können, anstatt von „fremden“ Institutionen URIs zugewiesen zubekommen?
  - Lösungen im Rahmen des Problems der Autorenidentifikation (vielleicht mittels ORCID?)

Erzeugung von Links so oft wie möglich und eindeutig, Erzeugung von URIs nur für Objekte des Repositoriums (da autoritativ)

Repositorien beschreiben

- Neben der Konvertierung der Metadaten sollten die Repositorien selbst beschrieben werden
- Wer betreibt das Repositorium? Gibt es eine SWORD-Schnittstelle? Wo ist der SPARQL-Endpoint zu finden? Wie ist das Repositorium strukturiert? …
- Auch braucht es ein Vokabular, um die Digitalen Objekte (Dateien) zu verlinken

- Für DSpace habe ich die DSpace Repository Ontology erstellt:
  http://digital-repositories.org/ontologies/dspace/
- Es wäre wichtig eine Digital Repositories Ontology zu erstellen, unabhängig von der genutzten Repositoriensoftware
- Ein Mapping zwischen einer solchen Ontologie und der DSpace Repository Ontology, der EPrints Ontology, … wäre hilfreich!
- Vielleicht im Rahmen von DINI und/oder COAR?

Technische Universität Berlin
Universitätsbibliothek
Pascal-Nicolas Becker
p.becker@tu-berlin.de

Servicezentrum Forschungsdaten und –publikationen
http://www.szf.tu-berlin.de

Repositorium DepositOnce
http://depositonce.tu-berlin.de

Diplomarbeit „Repositorien und das Semantic Web“
http://www.pnjb.de/uni/diplomarbeit/