



Linked Open Projects

Nachnutzung von Projektergebnissen als Linked Data

Kai Eckert (Dipl.-Inf., Dipl.-Wirt.-Inf.)
kai.eckert@bib.uni-mannheim.de

Magnus Pfeffer (Dipl.-Inf., M.A. LIS)
magnus.pfeffer@bib.uni-mannheim.de



Überblick

- Nachnutzung von Projektergebnissen
- Projekte
- Linked Data Service der UB Mannheim
- Ausblick



Was denn für Projekte?

- Projekte, die Daten erzeugen
 - In vielen (Bibliotheks-)Projekten werden Daten erzeugt
- Aber auch in anderen Bereichen:
 - Umfragen
 - Census of Marine Life
 - Studien
 - ...



Projekte

Große
Projekte

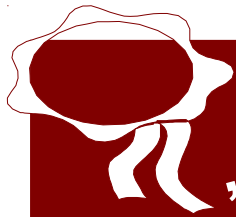
Kleine Projekte

Studenten

Wissenschaftler

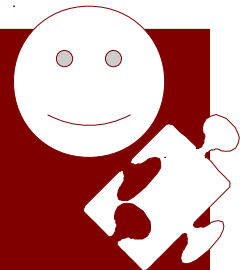
Arbeits-
gruppen

BMBF,
DFG...



„Hohe
Qualität“

„Spielereien“





Was passiert mit den Daten?

- Die Nachnutzung der Daten ist wichtig
 - ... laut den Projektanträgen
- Die Nachnutzung ist schwierig
 - „Die Daten sind ja gar nicht repräsentativ...“
 - „Da fehlen doch aber noch Daten...“
 - „In der Qualität kann man das aber nicht verwenden...“
 - „Da haben wir doch gar kein Feld mehr frei...“
 - „Das ist ja gar nicht in XYZ 4.3...“
 - „Wo kämen wir denn da hin...“
 - ...



Und Recht haben sie!

- Unterschiedliche Ansprüche bei Datenerzeugern und Datennutzern.
 - Konkrete Anwendungen können oft im Vorfeld noch gar nicht bekannt sein.
 - Integration in bestehende Datenbanken ist schwierig.
- Trotzdem sind die Daten wertvoll.
- Insbesondere in der Gesamtheit aller Projekte



Wo ist denn nun das Problem?

- Nachnutzung im Ganzen schwierig
- Nachnutzung erfordert
 - Auswahl von Daten
 - Aggregation von Daten
 - Integration von Daten

→ Hoher Aufwand für jede Datenquelle



Umsetzung **Bisher**

Bisher

Zugang

Einheitlicher Zugriff
Möglichkeit der Selektion

Archive im Web

Datenformat

Einheitliche Syntax
Maschinenlesbar

CSV, XML

Vokabular

Nachvollziehbare Semantik
Projektübergreifend

Spaltenüberschrift
Feldbeschreibung

Ressourcen

Verknüpfung mit den
Ursprungsdaten

Lokale/
uneinheitliche ID



Umsetzung **Bisher**

Bisher

Zugang

Einheitlicher Zugriff
Möglichkeit der Selektion

Archive im Web

← Unhandlich
← Grob

Datenformat

Einheitliche Syntax
Maschinenlesbar

CSV, XML

← OK

Vokabular

Nachvollziehbare Semantik
Projektübergreifend

Spaltenüberschrift
Feldbeschreibung

← Uneinheitlich

Ressourcen

Verknüpfung mit den
Ursprungsdaten

Lokale/
uneinheitliche ID

← Nicht eindeutig



Alternative: Linked Data

	Bisher	Linked Data
Zugang		
Einheitlicher Zugriff	Archive im Web	URI Dereferencing
Möglichkeit der Selektion		Sparql
Datenformat		
Einheitliche Syntax		
Maschinenlesbar	CSV, XML	RDF
Vokabular		
Nachvollziehbare Semantik	Spaltenüberschrift	externes Vokabular
Projektübergreifend	Feldbeschreibung	Verknüpfbar
Ressourcen		
Verknüpfung mit den Ursprungsdaten	Lokale/ uneinheitliche ID	URI



Projekte



Projekte mit Metadaten

- Konkordanzen
 - Verknüpfung Katalogeintrag mit Katalogeintrag
 - Verknüpfung Normdateneintrag mit Normdateneintrag
- Anreicherung
 - Verknüpfung Katalogeintrag mit sonstigen Daten
- Erschließung
 - Verknüpfung Katalogeintrag mit Normdateneintrag
- Datenservice
 - Katalogdaten für Forschung und Lehre
 - Zentrale Normdatenverwaltung



Automatische Erschließung

- Projekt
 - Automatische Verschlagwortung nach STW
 - DFG Projekt (2007-2009)
 - Basis: Aufsatzdaten aus den Nationallizenzen
 - Umfang: > 10.000 Abstracts
 - Ziel: Qualitative Bewertung der Verfahren
 - Daten im CSV-Format



Automatische Erschließung

■ Ausgangsdaten

■ Aufsatzdaten

- MAB2, Konversion nach Dublin Core einfach
- LCSH Schlagwörter vorhanden

■ STW

- Als Linked Data (SKOS) verfügbar

■ Herausforderungen

- Trennen der STW und LCSH
- Trennen von automatischer und manueller Erschließung
- Trennen der Ergebnisse von unterschiedlich parametrisierten Indexierungsläufen



Automatische Erschließung

- Umsetzung 1: Eigene Sub-Properties
 - LCSH oder STW über concept scheme erkennbar
 - manuell oder automatisch
 - Lauf 1..n

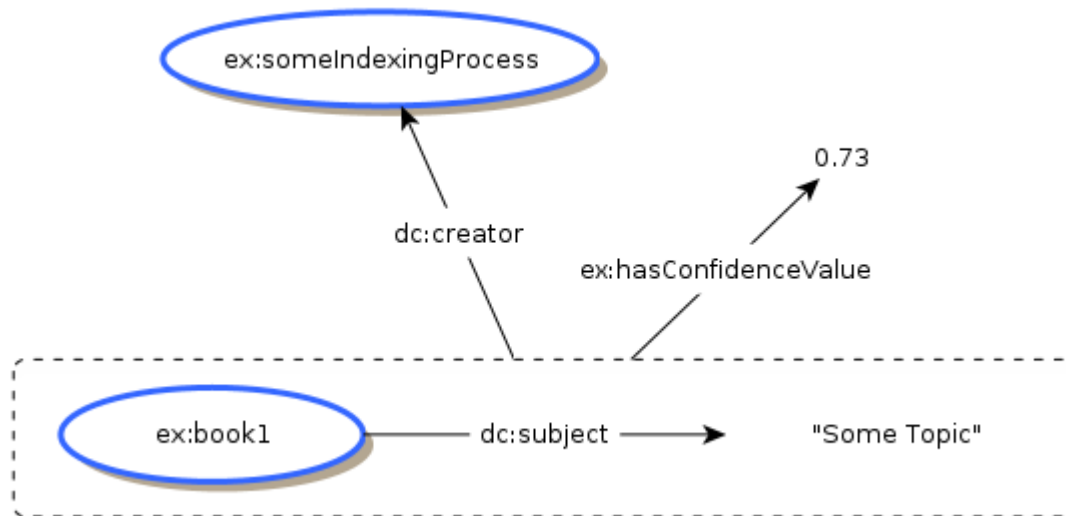
 - Trennung durch Auswahl der sub-property
 - Zusammenführung durch Mapping
sub-property → dc:subject

→ Skaliert nicht für viele Indexerläufe



Automatisches Erschließen

- Informationen über einzelne Zuweisung
- Metametadaten / Provenienz





Automatische Erschließung

- Umsetzung 2: Named Graphs
 - Ausgangsdaten im eigenen Graph
 - LCSH wahlweise als dc:subject oder sub-property
 - Ergebnisse jedes Laufs des Indexers in separaten Graph
 - Parameter, etc. als Properties des Graph-URLs
 - STW wahlweise als dc:subject oder sub-property
 - Trennung durch explizite Auswahl der Graphen in SparQL
 - Zusammenführung durch Aggregation der Graphen

- Skaliert für viele Indexierungsläufe



Anreicherung von Katalogdaten

- Projekt
 - SWBPlus
 - Kooperation im Südwestverbund
 - Zusatzinformationen für den Katalog:
 - Inhaltsverzeichnis
 - Klappentexte
 - Cover
 - Umfang: Alle Neuerscheinungen im Verbund
 - Geschlossener Nutzerkreis



Anreicherung von Katalogdaten

- Ausgangsdaten
 - Katalogdaten SWB
 - MAB2, wahlweise MARC21
 - Anreicherungsinformationen
 - XML Daten
 - PDF / Bilddateien
 - FTP Download
- Herausforderungen
 - Abrufen der Anreicherung einzelner Titel
 - Trennen der Primär- und Anreicherungsdaten
 - Anreicherung über Verbundgrenzen hinweg



Anreicherung von Katalogdaten

- Umsetzung 1: getrennte Properties
 - SWB-Katalogdaten in DC
 - Anreicherungsinformationen mit eigenem Vokabular

 - Umsetzung 2: named graphs
 - Eigene Graphs für die unterschiedlichen Daten
 - Freie Wahl des Vokabulars für Primär- und Katalogdaten
- Beide Varianten möglich
- Named graphs mit Vorteilen



Konkordanzen

- Projekt
 - Abgleich SWB/HeBIS
 - Umfang: 22 Millionen Titeldaten
 - Ziel: Übernahme von Erschließungsdaten
- Daten
 - Verbundabzüge in MAB2
 - Abgleichsdaten im CSV-Format
- Herausforderungen
 - Transitive Hülle
 - Schiere Datenmenge



Konkordanzen

- Umsetzung
 - Eigene Property für „Äquivalenz in Bezug auf Sacherschließung“
 - Eigener Graph zur Zusammenfassung der Äquivalenzinformationen aus einem Abgleichslauf
 - Definition der Transitivität in OWL



Linked Data Service (LDS) der UB Mannheim



Linked Data Service der UB Mannheim

- Prototyp und Testsystem
 - Vergleich von Softwaresystemen
 - Skalierung bei großen Datenmengen
 - Alternativen bei Darstellungen in RDF
 - Testbed für die W3C Library Linked Data Incubator Group
 - Testbed für die DCMI Metadata Provenance Task Group
- Showcase
 - HTML Ansicht und RDF/XML Download
 - Metadaten und Normdaten als Linked Data
 - Verknüpfung mit bestehenden Angeboten
 - Hosting von verschiedenen Projektdaten



Datensets

- SWB-Titeldaten
 - 13.349.777 Titel
- HeBIS-Titeldaten
 - 9.283.428 Titel
- RVK
 - 813.586 Konzepte

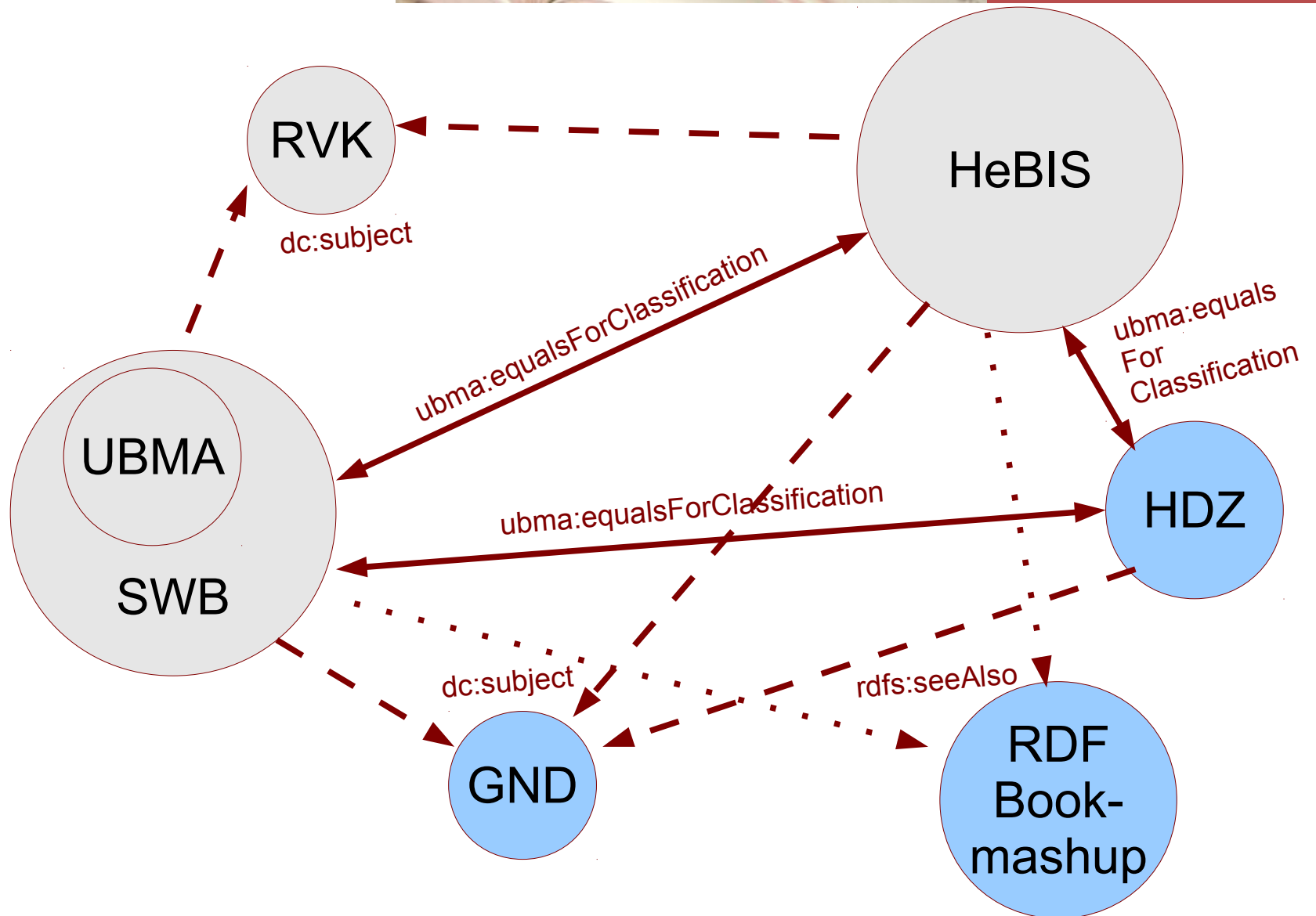
→ In Summe mehr als 178 Millionen RDF-Tripel



URI	Beschreibung
<a href="http://data.bib.uni-mannheim.de/data/swb/<PPN>">http://data.bib.uni-mannheim.de/data/swb/<PPN>	Titel aus dem Südwestdeutschen Bibliotheksverbund (SWB), <PPN> entspricht der internen Identifikationsnummer des SWB.
<a href="http://data.bib.uni-mannheim.de/data/hebis/<PPN>">http://data.bib.uni-mannheim.de/data/hebis/<PPN>	Titel aus dem Hessischen Bibliotheksinformationssystem (HeBIS), <PPN> entspricht der internen Identifikationsnummer des HeBIS.
<a href="http://data.bib.uni-mannheim.de/data/hbz/<PPN>">http://data.bib.uni-mannheim.de/data/hbz/<PPN>	Titel aus dem Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (HBZ), <PPN> entspricht der internen Identifikationsnummer des HBZ.
<a href="http://data.bib.uni-mannheim.de/data/rvk/<NOTATION>">http://data.bib.uni-mannheim.de/data/rvk/<NOTATION>	Klassifikationen nach der Regensburger Verbundklassifikation (RVK). <NOTATION> entspricht der RVK-Notation, das Leerzeichen wurde durch _ ersetzt (z.B. FB_1575).



URI	Bezeichnung	Hinweis
dc:title	Titel	Haupttitel in Vorlageform plus Zusatz
dc:creator	Urheber	Autoren in Ansetzungsform "Nachname, Vorname" oder Bezeichnung der Körperschaft.
dc:contributor	Mitwirkende/r	Weitere Beteiligte, wie Herausgeber o.ä., in Ansetzungform "Nachname, Vorname" oder Bezeichnung der Körperschaft.
dc:date	Zeitangabe	Erscheinungsjahr
dc:publisher	Verleger(in)	Verlag
isbd:P1008	has edition statement	Informationen zur Auflage
dc:identifizier	Identifikator	Die ISBN als URI: uri:ISBN:<ISBN>
dc:language	Sprache	Die Sprache, nach ISO 639-2 (B)
dc:subject	Thema	Schlagworte nach SWD mit URI der DNB, Klassifikationen nach RVK mit URI der UB Mannheim (s.o.).
rdfs:seeAlso	siehe auch	Links zu anderen Linked Data Angeboten, z.Z. zum Book Mashup Dienst der FU Berlin.
ubma>equals For Classification	Andere Ausgabe	Hierbei handelt es sich um das Ergebnis eines Projekts der UB Mannheim, das durch diesen Dienst leicht nachgenutzt werden kann.





Linked Data Service (Public Beta)

- Website / Dokumentation
 - <http://data.bib.uni-mannheim.de/>
- Sparql Endpoint
 - <http://data.bib.uni-mannheim.de/sparql/>
- URI-Dereferencing (httpRange-14)
 - 303 Verweis auf HTML oder RDF Darstellung
- Technische Basis
 - Zur Zeit Virtuoso Universal Server
 - Andere Systeme in Evaluation

An Entity of Type : [Bibliographic Resource](#), from Data Source : <http://data.bib.uni-mannheim.de/data/swb/305776088>, within Data Space : <data.bib.uni-mannheim.de>

[Forward Links](#)[Backward Links](#)

<http://data.bib.uni-mannheim.de/data/terms>equalsForClassification> <http://data.bib.uni-mannheim.de/data/hebis/214418405>
<http://data.bib.uni-mannheim.de/data/hebis/212760327>

Creator Stuckenschmidt, Heiner

Subject <http://d-nb.info/gnd/4072806-7>
<http://d-nb.info/gnd/4138762-4>
<http://d-nb.info/gnd/4049534-6>
http://data.bib.uni-mannheim.de/data/rvk/ST_205
<http://d-nb.info/gnd/4827894-4>
<http://d-nb.info/gnd/4242662-5>
<http://d-nb.info/gnd/4197730-0>
<http://d-nb.info/gnd/4480091-5>

Publisher Springer

Date 2009

Title Ontologien Konzepte, Technologien und Anwendungen

type [dcterms:BibliographicResource](#)

Raw Linked Data formats: [N3/Turtle](#) | [RDF/JSON](#) | [ODATAVATOM](#) | [ODATAJSON](#) | [RDF/XML](#)

[Linked](#) [Data](#) 

Trace ID: 7efc75f622cd



Einfache Nachnutzung

■ Standardisierte Schnittstelle:

```
select distinct ?titles where {  
  <http://data.bib.uni-mannheim.de/data/swb/305776088>  
  ubma:equalsForClassification ?titles.  
}
```

titles

```
http://data.bib.uni-mannheim.de/data/hebis/214418405  
http://data.bib.uni-mannheim.de/data/hebis/212760327
```



Studentenprojekt

- Rechercheportal Wirtschaftsinformatik
- Laufzeit: 1 Jahr (HWS 2010, FSS 2011)
- 10 Studenten
- Aufbereitung bibliographischer Daten, Mashup
- Zugriff auf Daten der UB Mannheim via LDS
- Bedingung: Jegliche Datenaufbereitung wird als RDF, bzw, via Sparql-Endpoint zur Verfügung gestellt.



Ideen und Pläne

- Ergänzung der geladenen Daten
 - Verlinkung von Personen und Körperschaften mit der PND
 - Vollständige Implementierung der RVK als Linked Data
- Integration von Projektergebnissen und neuen Daten
 - Automatische Inhaltserschließung
 - Automatisch und manuell erstellte Konkordanzen
 - Nutzungsdaten aus Katalog und Ausleihsystem
- Provenienzinformationen
- Know-How-Transfer



Viel Spaß beim Spielen!

<http://data.bib.uni-mannheim.de/>

Vielen Dank für Ihre Aufmerksamkeit!

<http://slideshare.net/kaiec>

