

# Dynamische Metadatenerfassung in Repositorien

Ein Vortrag in der Session

„Mehrwert durch Repositorien generieren“

TK 6: Bibliotheken für die digitale Wissenschaft: Fokus eScience und elektronisches Publizieren

auf dem

104. Bibliothekartag in Nürnberg

Kathleen Neumann

# Agenda

- Reposis – der Repository-Dienst der VZG
- Redundante Datenerfassung vermeiden
- Lösungsansätze und die Vision
- Stand der Dinge
- Welche Probleme zeichnen sich ab?
- Zusammenfassung und Ausblick

# Reposis – Der Repository-Dienst der VZG

- Bereitstellung, Anpassung, Wartung & Pflege eines eigenen Repository-Servers
- Dienst umfasst:
  - Installation der Webanwendung inkl. CSS-Anpassungen
  - Anpassung der Rechte, Rollen und des Workflows
  - Softwareaktualisierung und Wartung im Produktivbetrieb
  - Backup der Daten
  - Bereitstellung einer telefonischen Hotline
- mehr dazu auf <https://www.gbv.de/>



# Motivation ist, zu motivieren

- „Mehrwert durch Repositorien generieren“
  1. Datenerfassung vereinfachen
  2. Redundante Datenerfassung vermeiden
  3. Dienste und Informationen Dritter nutzen

# Zwei Lösungsansätze

## 1. Datenextraktion aus Volltexten

### Metadata Extraction

– **DRY, don't repeat yourself!**

Zenodo will automatically discover basic information like title, description, authors and publication date about your upload so you just have to validate our findings instead of copy/pasting or re-typing everything.

[zenodo.org, Zugriff am 22.05.2015]

The screenshot shows the 'EXTRACT TEXT' interface of the PDF Extract Tool. At the top, there is a 'Select PDF File' section with a 'Durchsuchen...' button and a message 'Keine Datei ausgewählt.'. Below this, the selected file 'Neumann\_2015\_05\_26.pdf' is shown, identified as a 'PowerPoint-Präsentation' with '17 pages, 559kB'. The main interface has four tabs: 'text', 'images', 'bookmarks', and 'annotations', with 'text' selected. Under 'Extraction Options', 'Use Page Rotate Attribute' is checked, and 'Extract from' is set to 'page 1'. Under 'Presentation Options', 'Raw PDF Fragments' is selected. The output window shows a table of extracted text with X and Y coordinates:

X Pos	Y Pos	Text
98.23	372.98	Dynamische Metadatenerfassung in
267.34	329.78	Repositorien
272.38	284.74	Ein Vortrag in der Session
211.06	264.34	„Mehrwert
291.94	264.34	durch Repositorien
427.08	264.34	generieren“
508.94	264.34	
34.97	243.94	TK 6: Bibliotheken für die digitale Wissenschaft: Fokus eScience und

At the bottom, there is a feedback prompt: 'Do you find this application useful? Your feedback is appreciated: [click here](#)'.

[pdf-online.com, Zugriff am 25.05.2015]

## 2. Datenextraktion aus externen Quellen unter Verwendung von Linked Open Data

# Mögliche Anwendungsszenarien

- Universitäts-Bibliographie
  - Autor meldet sich an und gibt PersonenID (z.B. GND) an
  - System findet Publikationen von diesem Autor in verschiedenen Quellen und schlägt vor, diese in die Bibliographie zu übernehmen
- Repository Datenerfassung
  - Angabe eines „Kennzeichens“ (z.B. ISSN) der zu erfassenden Publikation
  - System erkennt Art des „Kennzeichens“ und
  - nutzt verschiedene Quellen, um mögliche, bereits erfasste Informationen zu laden

# Mögliche Datenquellen

Kennzeichen	mögliche Quelle	Format
ZDB-ID	Zeitschriftendatenbank	XML (RDF)
ISSN	EZB, ZDB	XML
ISSN	DNB, Verbundkataloge (VZG, HBZ), GoogleBooks	XML (RDF, MODS) / JSON
DOI	CrossRef	JSON
...	...	...

- MIT stellt eine [Übersicht verschiedener Schnittstellen](#) zu wissenschaftlichen Ressourcen zur Verfügung

# Umsetzung – Stand der Dinge

- Autovervollständigung von Personennamen und Körperschaften sowie Integration der GND bzw. VIAF z.B. unter Nutzung von lobid.org und den Suggest-Services der VZG
- Übernahme von Daten anhand der PPN oder ZDB-ID aus GVK bzw. ZDB
- Autovervollständigung mit Hilfe bereits erfasster Metadaten im Repository

# Autovervollständigung von Personennamen

<intR><sup>2</sup>Dok[S] administrator ▾

Blättern ▾ Publizieren 📌<sup>0</sup> Suchbegriff eingeben 🔍

<b>Publikationsstatus</b>	<input type="text" value="wird bearbeitet"/>	<input type="button" value="i"/>	
<b>Publikationstyp:</b>	<input type="text" value="Buchkapitel"/>	<input type="button" value="i"/>	<input type="button" value="+"/> <input type="button" value="-"/>
<b>führende Artikel:</b>	<input type="text" value="The"/> <input type="text" value="Englisch"/> <input type="button" value="▾"/> <input type="text" value="Haupttitel"/> <input type="button" value="▾"/>	<input type="button" value="i"/>	<input type="button" value="+"/> <input type="button" value="-"/>
<b>* Titel:</b>	<input type="text" value="Next Big Thing: Web Services Collaboration"/>	<input type="button" value="i"/>	
<b>Untertitel:</b>	<input type="text"/>	<input type="button" value="i"/>	
<b>Autor(in)</b> ▾	<input type="text" value="Jeckle"/> <input type="button" value="Suchen..."/>	<input type="button" value="i"/>	<input type="button" value="+"/> <input type="button" value="-"/>
<b>Konferenz:</b>	<input type="text" value="Jeckle, Mario (1974-08-25-2004-06-11)"/> <input type="text" value="Jeckle, Mario"/>	<input type="button" value="i"/>	
<b>* Veröffentlicht am:</b>	<input type="text" value="JJJJ-MM-TT"/>	<input type="button" value="i"/>	

[intr2dok.vifa-recht.de, Zugriff am 25.05.2015]

# Zeitschriften und Serien erfassen mit ZDB-ID

## 1. Angabe der ZDB-ID

• Zeitschrift/Serie

ZDB-ID

Importieren

## 2. Datenübername aus ZDB

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF ...>
  <rdf:Description rdf:about="http://ld.zdb-services.de/resource/5674-1">
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Periodical" />
    <dc:identifrier>(DE-101)01007953X</dc:identifrier>
    <bibo:issn>0005-9366</bibo:issn>
    <bibo:issn>0005-9366</bibo:issn>
    <owl:sameAs rdf:resource="http://lobid.org/resource/ZDB5674-1" />
    <dc:identifrier>(DE-600)5674-1</dc:identifrier>
    <dc:identifrier>(OCoLc)224454535</dc:identifrier>
    <dcterms:language rdf:resource="http://id.loc.gov/vocabulary/iso639-2/ger" />
    <dc:title>Berliner und Münchener tierärztliche Wochenschrift</dc:title>
    <bibo:shortTitle>Berl. Muench. Tieraerztl. Wochenschr.</bibo:shortTitle>
    <rdau:P60333>Hannover : Schlütersche, 1938</rdau:P60333>
    ...
  </rdf:Description>
</rdf:RDF>
```

# Zeitschriften und Serien erfassen mit ZDB-ID

### Zeitschrift/Serie

Genre\*  Zeitschrift  Serie

Titel\*

Titel (html)

Titel übersetzt  Englisch

Titelabkürzung

### Einordnung

Verlag

Erscheinungsort

Identifikator  ISSN     
 ZDB-ID

Kenndaten 

Jahr	Impact Factor	Referiert
<input type="text"/>	<input type="text"/>	<span>(bitte wählen)</span> <input type="button" value="+"/> <input type="button" value="-"/>

[openagrar.bmel-forschung.de, Zugriff am 25.05.2015]

# Viele Identifizier – Eine Person

- GND, ORCID, VIAF, PubMedID ...
- Crosskonkordanz für Personen



Deutsche Nationalbibliothek

DNB-128415606

[viaf.org, Zugriff 22.05.2015]

**Gotthold Ephraim Lessing**  
Kritiker, Philosoph, Bibliothekar, Schriftsteller  
Geboren: 22. Januar 1729, [Kamen](#)  
Gestorben: 15. Februar 1781, [Braunschweig](#)

**Objekte:**

Gotthold Ephraim Lessing, Gotthold Ephraim Lessing, Porträt Gotthold Ephraim Lessing, Lessing-Haus, Lessinghaus

[Alle Bilder \(29\)](#) | [Alle Audios \(2\)](#)

**Beteiligt an:**

- **Sämtliche Schriften. 32**  
Lessing, Gotthold Ephraim. - Berlin : Voss, (1828)
- **Sämtliche Schriften. 27**  
Lessing, Gotthold Ephraim. - Berlin : Voss, (1794)
- **Gesammelte Werke. 6**  
Lessing, Gotthold Ephraim. - Leipzig : Göschen, 1855
- **Sämtliche Schriften. 20**  
Lessing, Gotthold Ephraim. - Berlin : Voss, (1794)

[Alle Objekte \(421\)](#)

**Thema in:**

- **Lessings Bedeutung für das deutsche Drama**  
Motz, Heinrich. - Basel : Schwabe, 1872
- **Gotthold Ephraim Lessing als Theologe : e. Beitr. zur Geschichte d. Theologie im 18. Jh.**  
Schwarz, Carl. - Halle : Pfeffer, 1854
- **Ironie und Autorschaft : zu Friedrich Schlegels Charakteristiken**

Gotthold Ephraim Lessing  
Quelle: [Wikimedia Commons](#)  
Informationen zum Lizenzstatus eingebundener Mediendateien (etwa Bilder oder Videos) können im Regelfall durch Anklicken dieser abgerufen werden.

**Suche:**

„Gotthold Ephraim Lessing“ in der DDB

**Externe Links:**

- Wikipedia (Deutsch)
- Internet Movie Database (IMDb)
- Bayerisches Musiker-Lexikon Online
- Bibliothèque nationale de France | National Library of France
- Wikipedia (Englisch)
- Library of Congress / NACO
- Biographie in der Ostdeutschen Biographie (Kulturportal West-Ost)

[deutsche-digitale-bibliothek.de, Zugriff 22.05.2015]

# Crosskonkordanzen mit MODS abbilden (1)

- MODS 3.6<sup>1</sup> ermöglicht nun auch mehrere Identifier für einen `<mods:name />`

**MODS News & Announcements**

- [MODS 3.6 Schema Released](#) **NEW!**
  - [MODS 3.6 Schema](#)
  - [MODS 3.6 Changes](#)
- [MODS/RDF Draft Ontology](#)

[loc.gov/standards/mods/, Zugriff am 22.05.2015]

<http://d-nb.info/gnd/128415606/about/lds>

```
<owl:sameAs rdf:resource="http://dbpedia.org/resource/Mario_Jeckle" />  
<owl:sameAs rdf:resource="http://viaf.org/viaf/85472775" />  
<gndo:gndIdentifier>128415606</gndo:gndIdentifier>
```



MODS < 3.6

```
<mods:name type="personal" authorityURI="http://d-nb.info/"  
  valueURI="http://d-nb.info/gnd/128415606">  
  <mods:displayForm>Jeckle, Mario</mods:displayForm>  
  <mods:role>  
    <mods:roleTerm authority="marcrelator" type="code">aut</mods:roleTerm>  
  </mods:role>  
</mods:name>
```

12

1 - MODS 3.6 erschienen am 05.05.2015

# Crosskonkordanzen mit MODS abbilden (2)

- neues Element `<mods:nameIdentifier />`

```
http://d-nb.info/gnd/128415606/about/lds
```

```
<owl:sameAs rdf:resource="http://dbpedia.org/resource/Mario_Jeckle" />  
<owl:sameAs rdf:resource="http://viaf.org/viaf/85472775" />  
<gndo:gndIdentifier>128415606</gndo:gndIdentifier>
```



```
MODS 3.6
```

```
<mods:name type="personal">  
  <mods:nameIdentifier type="gnd">128415606</mods:nameIdentifier>  
  <mods:nameIdentifier type="viaf">85472775</mods:nameIdentifier>  
  <mods:nameIdentifier type="dbpedia">Mario_Jeckle</mods:nameIdentifier>  
  <mods:displayForm>Jeckle, Mario</mods:displayForm>  
  <mods:role>  
    <mods:roleTerm authority="marcrelator" type="code">aut</mods:roleTerm>  
  </mods:role>  
</mods:name>
```

# Intuitive Nutzung bereits erfasster Daten

erschieden in Serie:

Titel:

ISSN:

Bandnr. der Serie:  ⓘ

Erscheinungsjahr:  ⓘ

↓

erschieden in Serie:

Titel:  ⓘ

ISSN:  ⓘ

Bandnr. der Serie:  ⓘ

Erscheinungsjahr:  ⓘ

[Prototyp deposit\_hagen, Zugriff 22.05.2015]

# Probleme

- Verfügbarkeit externer Quellen ist nicht sichergestellt
- nur selten werden normierte Austauschformate angeboten
- Aktualisierung der eigenen Daten kann zum *Denial of Service*-Angriff werden

# Zusammenfassung und Ausblick

- Aktueller Stand:
  - Statische Datenübernahme aus definierten Quellen (ZDB und GVK)
  - Integration externer Daten anhand GND und VIAF
  - Autovervollständigung mittels lobid.org, Suggest-Services der VZG und interner Daten
- das Ziel:
  - Anwender soll nicht explizit aus Quellen wählen müssen
  - bei der Erfassung der Metadaten werden aus verfügbaren Quellen passende Verknüpfungen angeboten

# Kontakt

## Herzlichen Dank für Ihre Aufmerksamkeit

Kathleen Neumann  
Verbundzentrale des GBV (VZG)  
Platz der Göttinger Sieben 1  
37073 Göttingen, Deutschland  
neumann@gbv.de