gesis
Leibniz-Institut für Sozialwissenschaften

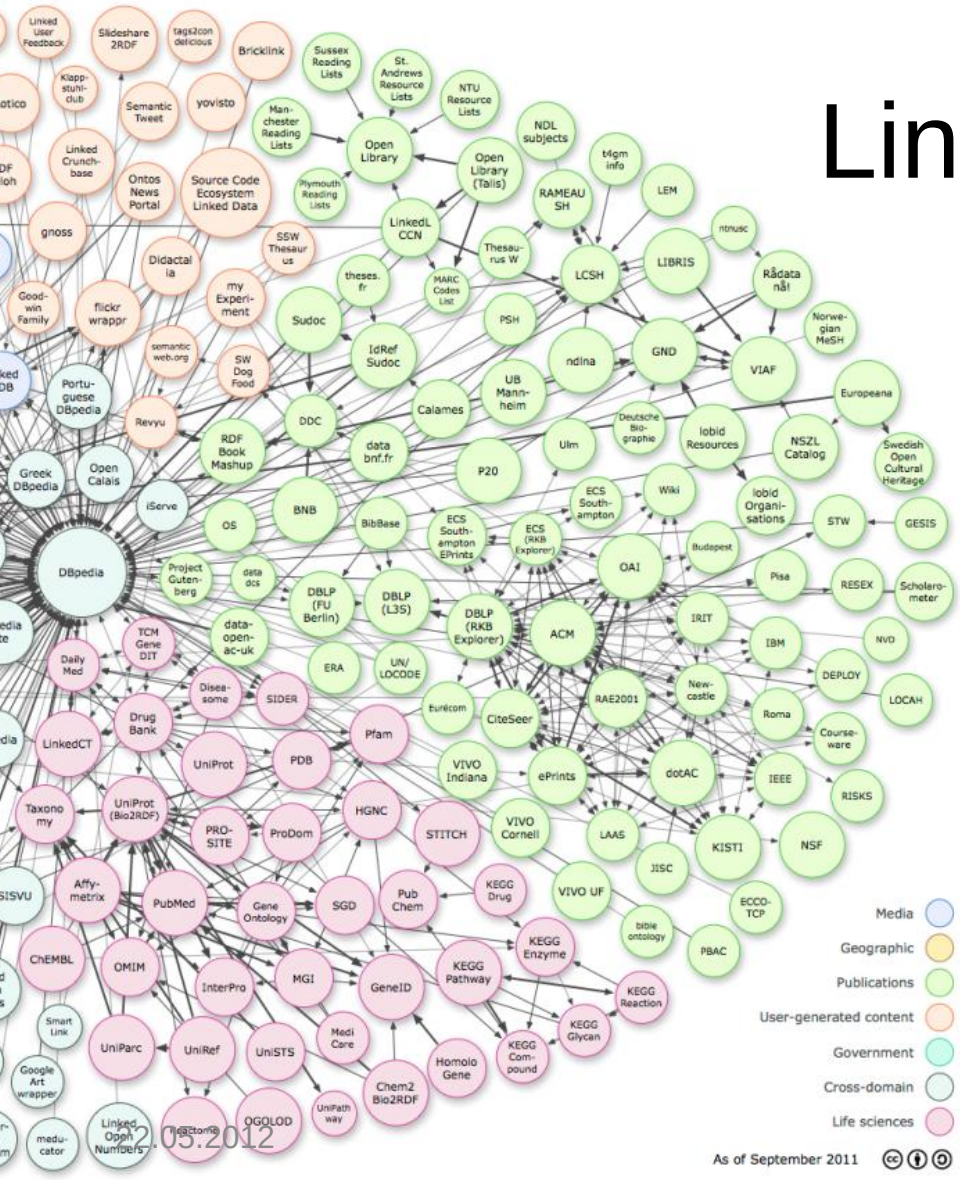
Anwendungsszenarien und Potenziale von Linked Open Data im wissenschaftlichen Forschungsprozess

Benjamin Zapilko

FIS Fachtagung 2012, 22.05.2012

Agenda

- Linked Open Data … und was nun?
- Wissenschaftler als LOD-Nutzer
- Anwendungsszenarien für LOD im Forschungsprozess
  - Forschungsdaten und Informationen recherchieren
  - Forschungsdaten analysieren
- Zusammenfassung



Linked Open Data…

- Immer mehr Daten in der LOD Cloud
- Semantische Technologien immer ausgereifter
- Erste interessante Anwendungen sind da
- „Jeder“ kann seine Daten ins Web bringen

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

… und was nun?

- Wie gelangen Daten in die LOD Cloud? Wie kann man LOD nutzen?
  - Technologie-Expertise vorhanden
  - Projekte und Initiativen, die Datenanbietern Unterstützung anbieten
- Was kann man mit LOD machen?
  - Erste Anwendungen existieren, aber meist explorierender oder visualisierender Art
- Warum LOD? Für wen? Und wie?
  - abhängig von Daten, Domäne und Nutzern
  - Argumente wie Offenheit und Transparenz als Motivation oft nicht überzeugend genug und politisch oft schwierig

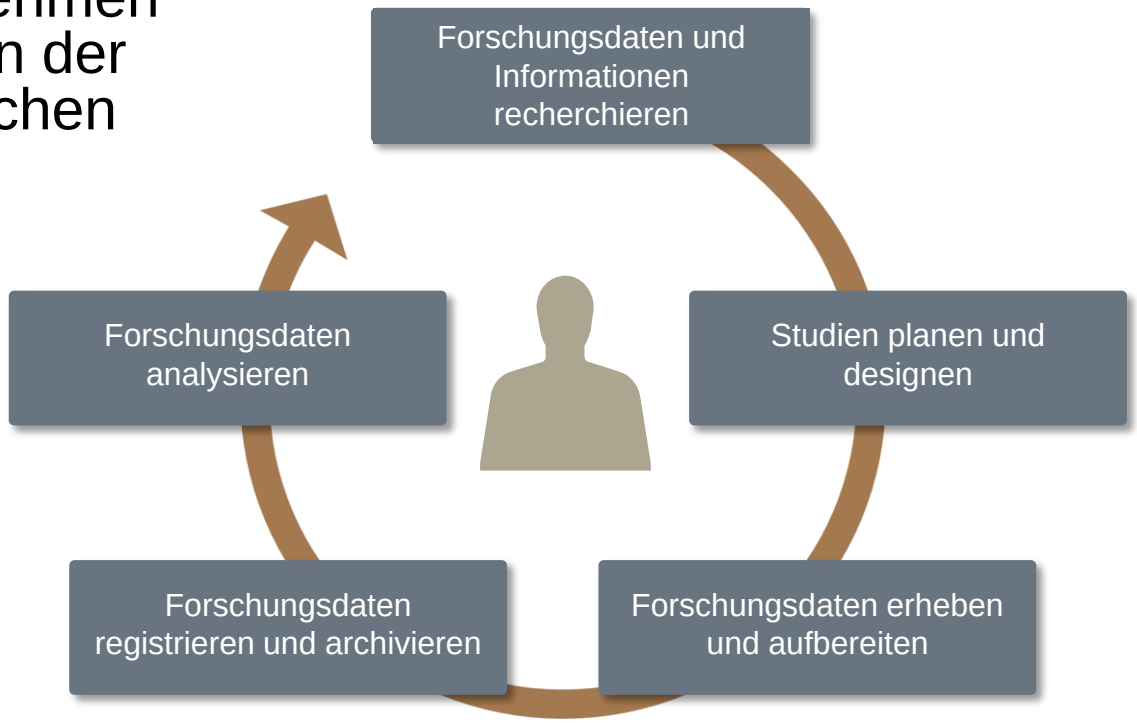
Anreize für Datenanbieter schaffen!
Mehrwerte für Nutzer schaffen!

Nutzergruppe: Wissenschaftler

- Betrachten des Forschungsprozesses eines Wissenschaftlers und darin enthaltene typische Arbeitsschritte
  - von Domäne zu Domäne unterschiedlich
  - verschiedene Schwerpunkte
  - typische Arbeitsabläufe
  - manche Aufgaben sehr zeitintensiv und wiederholen sich immer wieder
- Erkennen von Defiziten, Schwierigkeiten und Problemstellungen
  - Können Teile durch den Einsatz von LOD „aufgewertet“ oder vereinfacht werden?
- Identifizieren von möglichen Mehrwerten und Anreizen durch den Einsatz von LOD

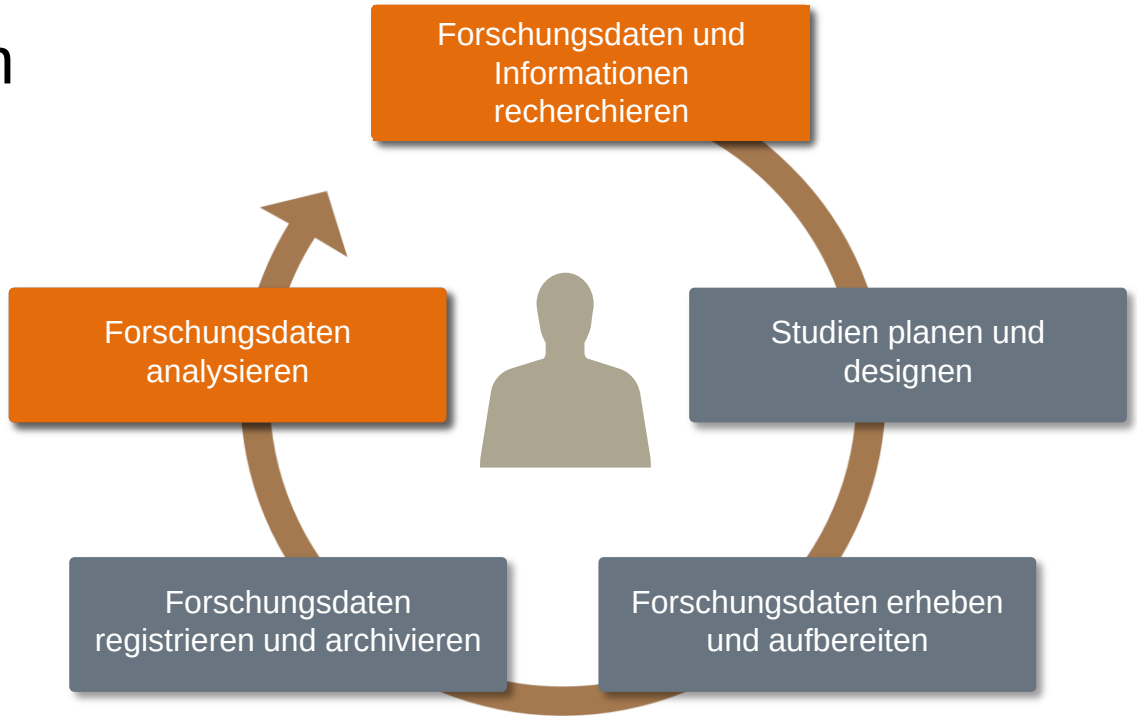
Forschung in den Sozialwissenschaften

- Forschungsdaten nehmen eine zentrale Rolle in der sozialwissenschaftlichen Forschung ein
- 5 Phasen, die innerhalb eines Forschungsprojekts durchlaufen werden



Fokus auf Sekundärforscher

- Forschungsdaten und Informationen recherchieren
- Forschungsdaten analysieren



Forschungsdaten und Informationen recherchieren

Wie ist die allgemeine Meinung zum Thema „Gewalt gegen Kinder in Familien“?

Wurden zu dieser Fragestellung bereits Umfragedaten ausgewertet?

Gibt es zu dieser Fragestellung weitere oder vergleichbare Forschungsdaten?

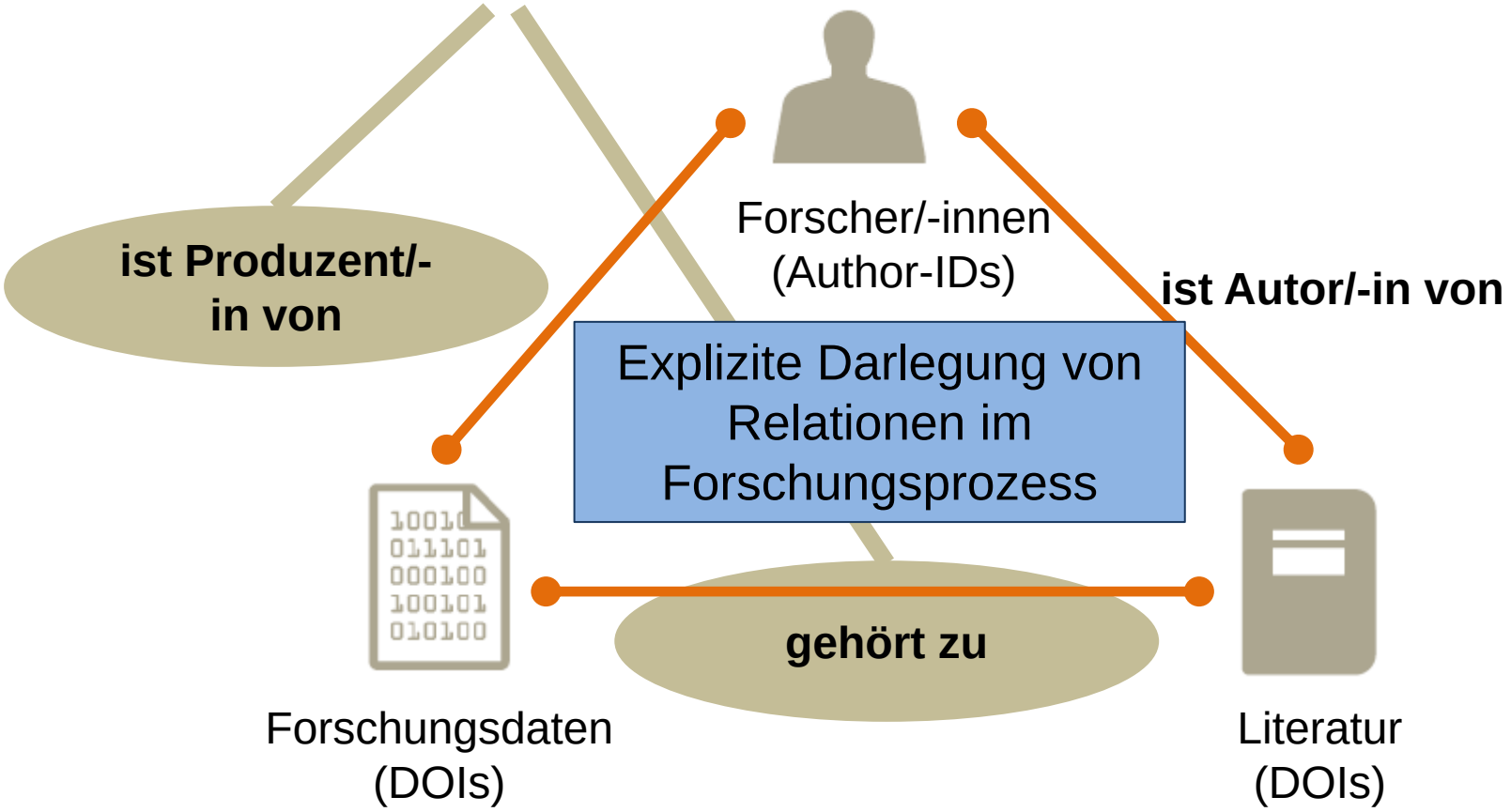
Welche Forschungsdaten und welche Publikationen helfen mir bei der Beantwortung dieser Fragestellung?

Mögliche Probleme und Hürden

- Recherche über heterogene und meist nicht miteinander verbundene Informationstypen
  - Verschiedene Metadatenstandards
  - Unterschiedlicher Erschließungsgrad
- Recherche gebunden an Institutionen, Datenbanken, Informationsportale etc. oder ganz offen (z.B. Google)
  - Forschungsdaten und Informationen werden dezentral gehalten, gepflegt und aufbereitet (u.a. historisch, rechtlich und disziplinär bedingt)
  - Im Web finden sich relevante Informationen, die von Fachdatenbanken nicht erfasst werden
  - In Fachdatenbanken finden sich relevante Informationen, die durch Websuche nicht gefunden werden

Integrierte Recherchen nur bis zu gewissem Grad möglich!
Datenintegration und Datenmodellierung äußerst aufwändig!

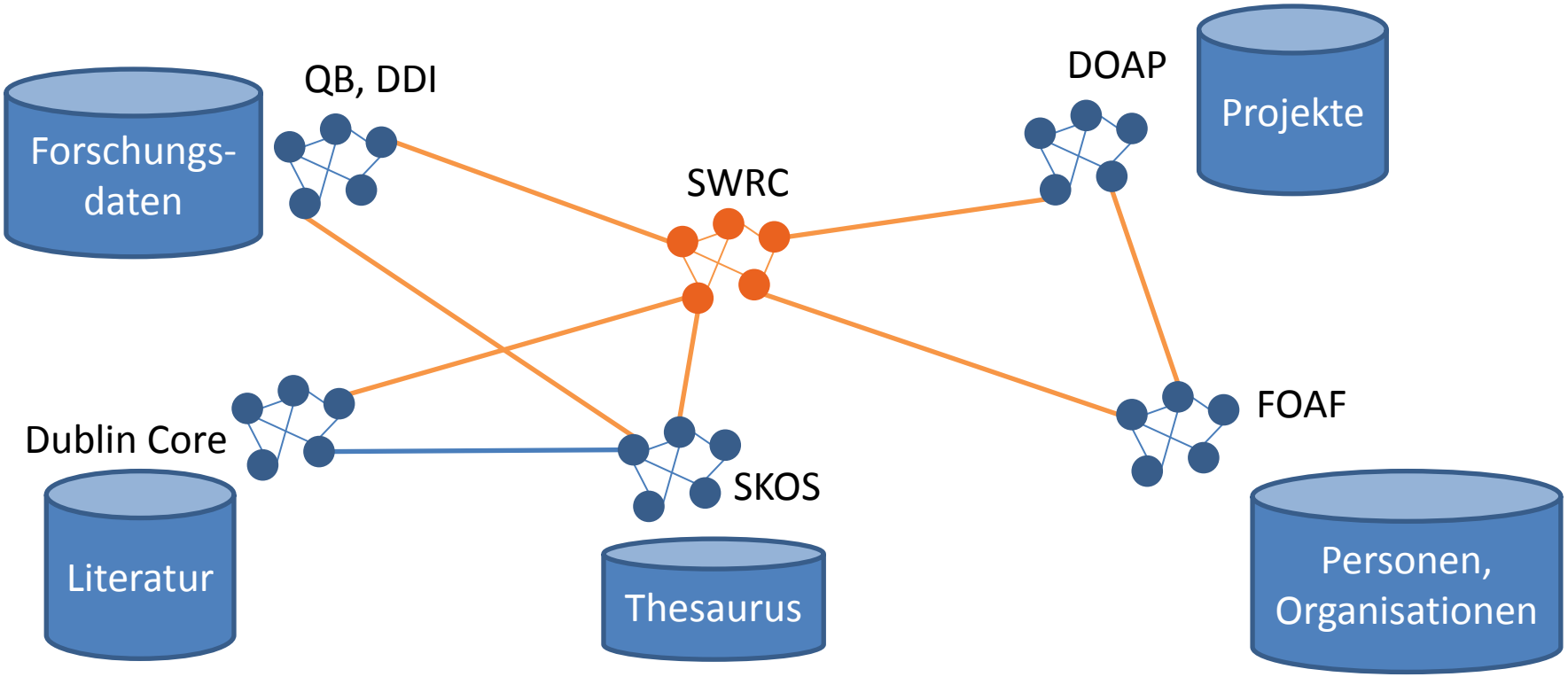
Fehlende Links herstellen



Forschungsprozess darlegen

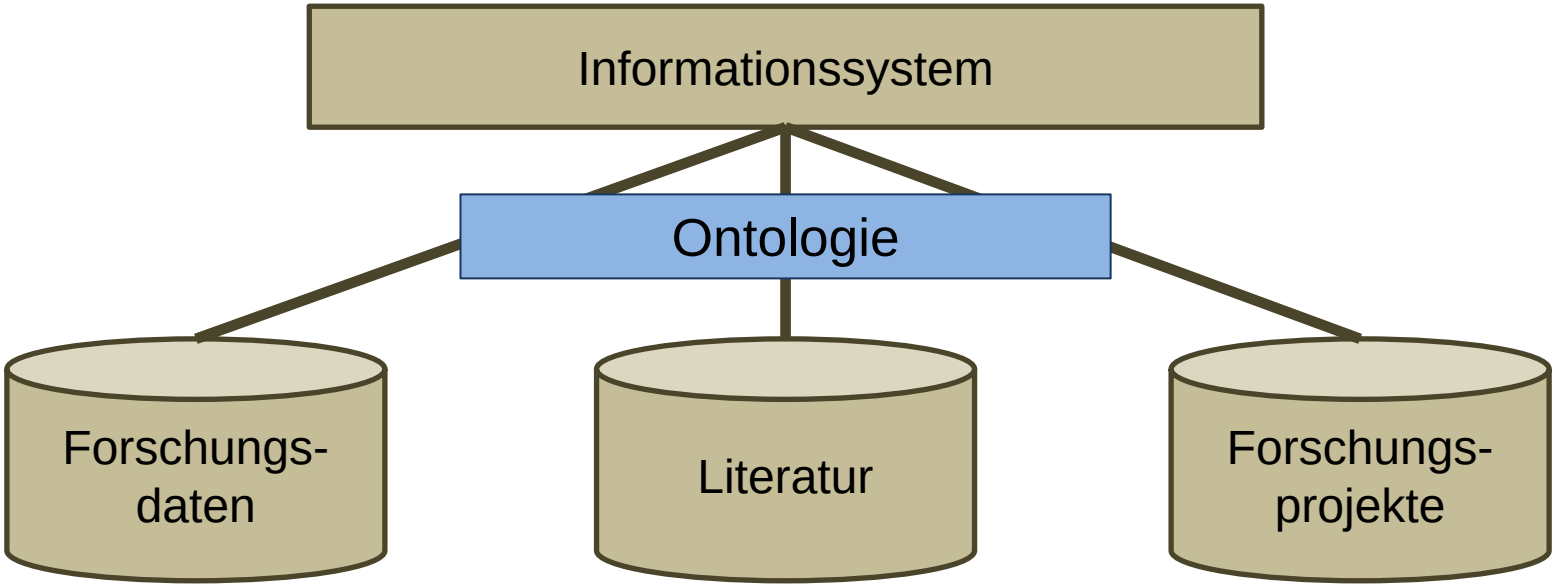
- Einheitliches Datenmodell über heterogene Informationstypen ist aufwändig
- Forschungsprozess zusammenführen durch den Einsatz von Ontologien
  - z.B. die SWRC Ontologie (Semantic Web for Research Communities)
  - über physikalische Grenzen hinweg
  - Datenbestände können unberührt bleiben
  - Gültige Metadatenstandards können beibehalten werden

Forschungsprozess darlegen



Einbettung in bestehende Infrastrukturen

- Ontologien als Modellierungsebene einfügen
  - „schnell“ implementierbar
  - bestehende Infrastrukturen bleiben bestehen



Neue Links herstellen

- DFG-Projekt „InFoLiS – Integration von Forschungsdaten und Literatur in den Sozialwissenschaften“
- gemeinsam mit UB und Uni Mannheim
- Kernziel: Automatisches Erkennen von Studienzitationen in Publikationen
- Output von Links im RDF
  - Diese Link-Information kann in Datenbestände eingebunden werden (z.B. in die Metadaten)

```
<daraDOI1> <isCitedBy> <ssoarURN1>.
<daraDOI2> <isCitedBy> <ssoarURN1>.
<daraDOI3> <isCitedBy> <ssoarURN2>.
<daraDOI4> <isCitedBy> <ssoarURN3>.
...
```

Neue Links herstellen

Veröffentlichung:
<ssoarURN1-Link>

da|ra

Zitierte Studien:
<daraDOI1-Link>
<daraDOI2-Link>

SSOAR

Neue Links herstellen

Veröffentlichung:
<ssoarURN1-Link>

da|ra

```
<bibo:AcademicArticle rdf:about="http://lod.gesis.org/ssoar/ID/0168-ssoar-123
  <dcterms:isPartOf rdf:resource="http://lod.gesis.org/ssoar/"/>
  <dcterms:issued>2009</dcterms:issued>
  <dcterms:type>Zeitschriftenartikel</dcterms:type>
  <dcterms:creator>Poustka, Fritz</dcterms:creator>
  <dcterms:language>en</dcterms:language>
  <rdfs:seeAlso>Link to full text: ../files/peer/Peer_125-2010-07-15.txt</rdf
  <dcterms:title>Pilot evaluation of the Frankfurt Social Skills Training for
  <dcterms:creator>Duketis, Eftichia</dcterms:creator>
  <owl:sameAs rdf:resource="http://lod.gesis.org/ssoar/ID/0168-ssoar-123844"/
  <bibo:issn>1435-165X</bibo:issn>
  <dcterms:license>http://www.ssoar.info/en/home/how-to-deposit-documents/leg
  <dcterms:creator>Schmötzer, Gabriele</dcterms:creator>
  <foaf:homepage rdf:resource="http://www.ssoar.info/ssoar/View/?resid=12384"
  <dcterms:identifier>http://nbn-resolving.de/urn:nbn:de:0168-ssoar-123844</d
  <dcterms:creator>Birnkammer, Sabine</dcterms:creator>
  <dcterms:extent>S. 327-335</dcterms:extent>
  <dcterms:creator>Schlitt, Sabine</dcterms:creator>
  <dcterms:source>European Child &amp;amp; Adolescent Psychiatry, Jg. 18, H. 6</d
  <bibo:producer rdf:resource="http://lod.gesis.org/ssoar/producer/GESIS"/>
  <dcterms:creator>Herbrecht, Evelyn</dcterms:creator>
  <dcterms:creator>Bölte, Sven</dcterms:creator>
  <dcterms:abstract xml:lang="en">The objective of this pilot study was to ev
</bibo:AcademicArticle>
```

Zitierte Studien:
<daraDOI1-Link>
<daraDOI2-Link>

SSOAR

Neue Links herstellen

Veröffentlichung:
<ssoarURN1-Link>

Veröffentlichung:
<Autor>
<Titel>
<Jahr>
<URN>
<Link>

da|ra

Zitierte Studien :
<Studientitel>
<Jahr>
<DOI>
<Link>

<Studientitel>
<Jahr>
<DOI>
<Link>

Zitierte Studien:
<daraDOI1-Link>
<daraDOI2-Link>

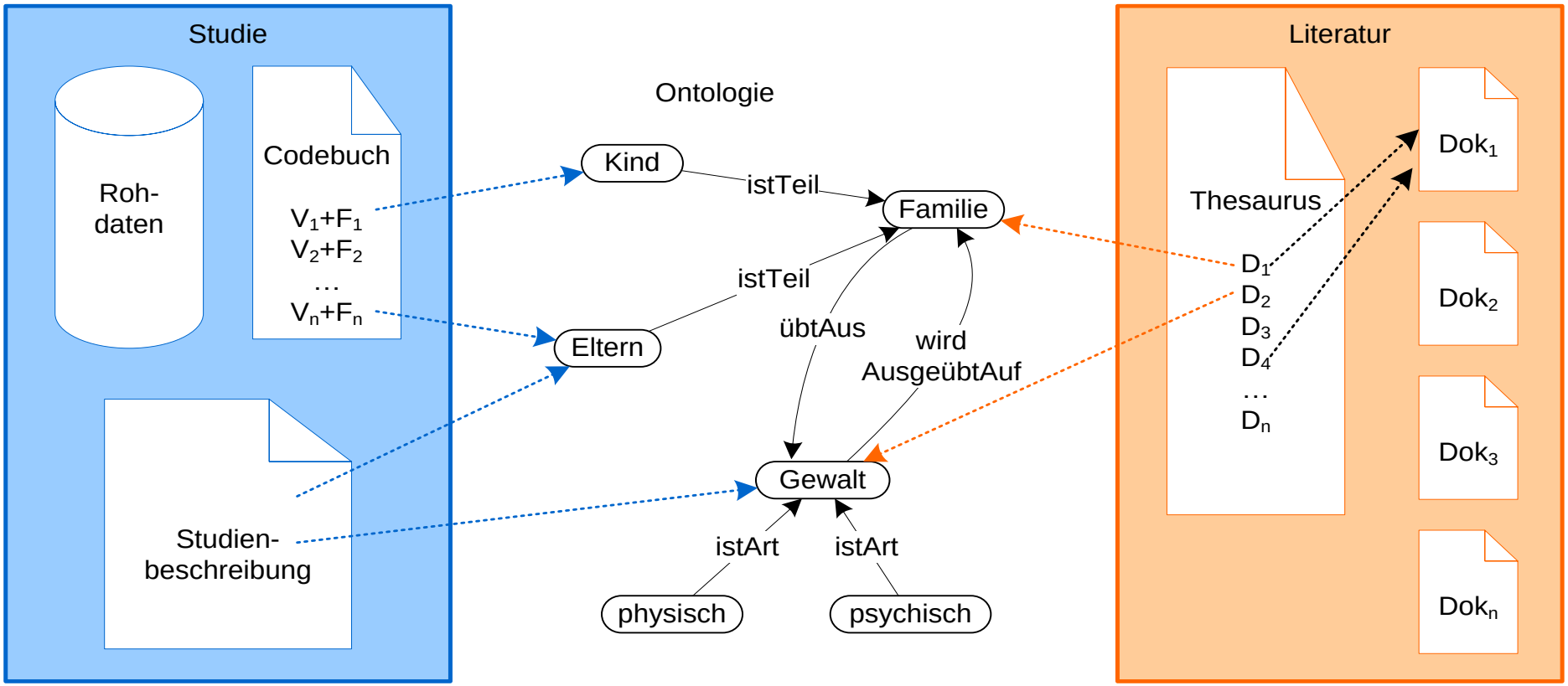
SSOAR

Neue Links herstellen

- Anreichern eigener Datenbestände durch externe Datenquellen
  - z.B. DBpedia, Gemeinsame Normdatei (GND), VIAF, Literatur wie UB Mannheim, hbz, Bibsonomy, …
  - Alternativ: „nur“ Links zu externen Datenbeständen setzen
- Vernetzung von Thesauri, Klassifikationen, etc.
  - z.B. für Search Term Recommendation und weitere IR-Dienste
  - Thesauri können auch bei Vernetzung zu externen Quellen nützlich sein, z.B. Daten mit denselben Keywords in externen Datenbeständen finden
- Vernetzung von Variablen, Indikatoren, Codelisten
  - Suche nach gleichen/ähnlichen Variablen bei unterschiedlichen Forschungsdatenzentren oder verschiedenen Studienserien

Höherer Erschließungsgrad

- Einsatz von Ontologien zur Inhaltserschließung



Mehrwerte für Wissenschaftler

- vernetzte, organisations- und datenbankübergreifende Recherche
- Präzisere Suchergebnisse durch eine semantisch ausdrucksstärkere Inhaltserschließung

Forschungsdaten analysieren

Haben Arbeitslosen- und Ausländerquote Auswirkungen auf die Angst, den Job zu verlieren?

In welchen Formaten liegen die recherchierten Daten vor?

Passen die Daten überhaupt zusammen?

Verfügen die Daten über eine statistische Qualität, dass ich sie verwenden kann?

Sind die Daten, die ich gefunden habe, die die ich tatsächlich benötige?

Mögliche Probleme und Hürden

- Daten liegen in unterschiedlichen Formaten vor (pdf, html, csv, spss, …)
  - Konvertierung notwendig
- Datenschemata oder verwendete Codelisten passen nicht zusammen
  - Schema Matching und Datenharmonisierung notwendig
- Daten liegen in verschiedenen Aggregationsebenen vor
- Statistische Qualitätsbeurteilung (Varianzen, Bias, etc.) oft nicht auf Anhieb möglich

Vorverarbeitung von Daten notwendig, um diese beurteilen, analysieren und weiter verarbeiten zu können!

Datenaustausch und -konvertierung

- Forschungsdaten als LOD veröffentlichen
  - Gemeinsame RDF/S oder OWL-Basis trotz unterschiedlicher Metadatenstandards
  - RDF als Austauschformat verwenden
  - Einheitlicher Datenaustausch, -import, -export
- Automatisierung und Vereinheitlichung von Konvertierungsprozessen
- Datenintegration

Schema Matching

Harmonised Unemployment Rate by Gender (Quelle: Eurostat)

	2010-10		2011-01		2011-04	
	Male	Female	Male	Female	Male	Female
Austria	4.4	4.0	4.4	4.6	4.3	4.1
Belgium	7.9	8.3	7.7	7.8	7.5	8.1
Germany	7.2	6.8	6.8	6.4	6.5	6.1
France	9.1	10.3	8.9	10.3	8.6	10.3
…	…	…	…	…	…	…

Schema Matching

sed Unemployment Rate by Gend

Country	2010-10		2011-01			
	Male	Female	Male	Female		
Albania						
Austria	4.4	4.0	4.4	4.6		
Belgium	7.9	8.3	7.7	7.8		
Bulgaria	7.2	6.8	6.8	6.4		
…						
France	9.1	10.3	8.9	10.3	8.6	10.3
…	…	…	…	…	…	…

Year
2007
2008
2009
2010
2011

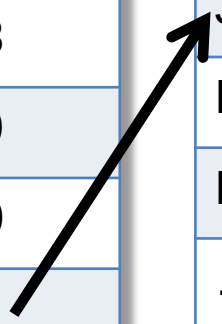
Month
January
February
March
…

Schema Matching

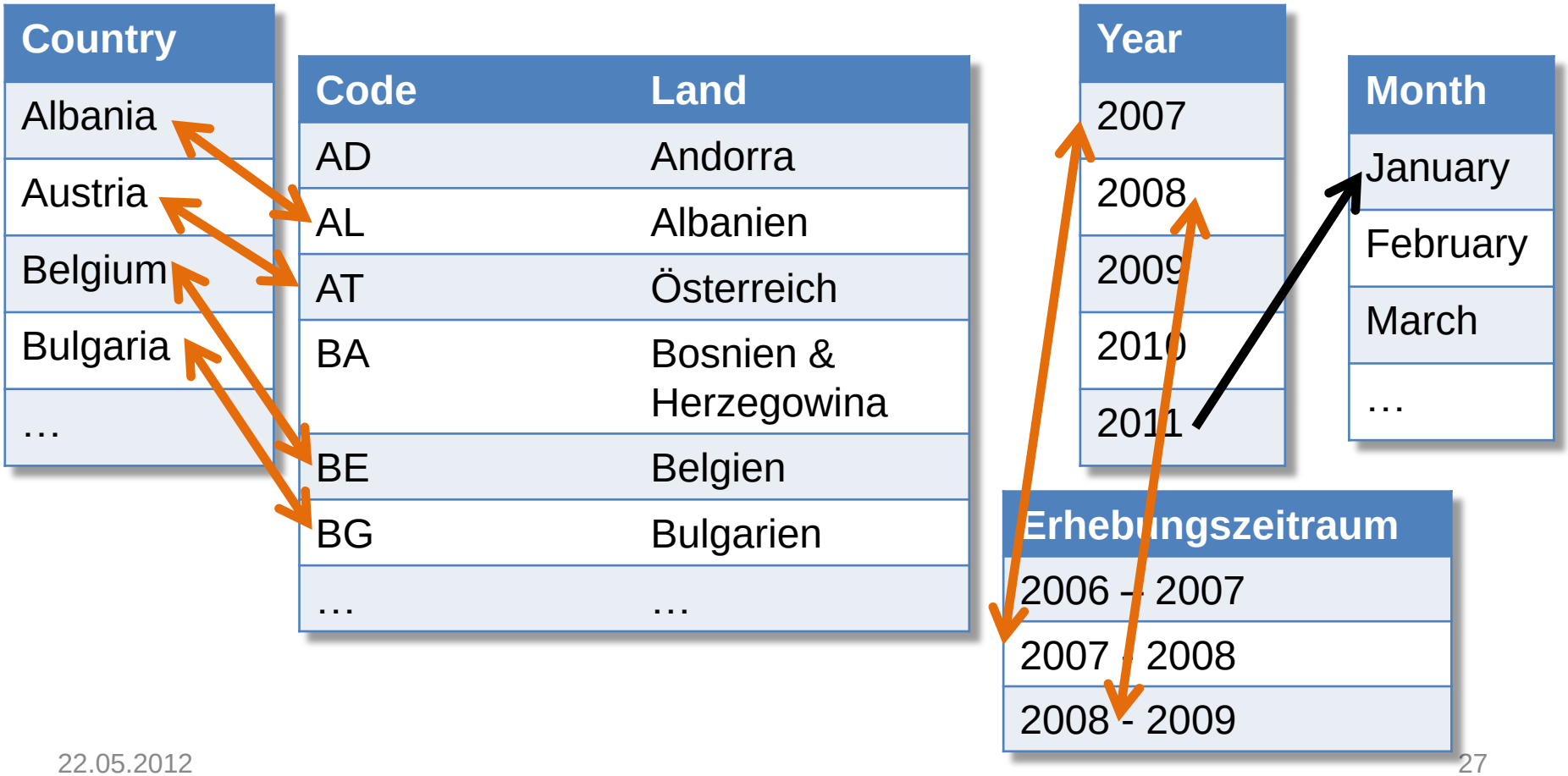
Country
Albania
Austria
Belgium
Bulgaria
…

Year
2007
2008
2009
2010
2011

Month
January
February
March
…

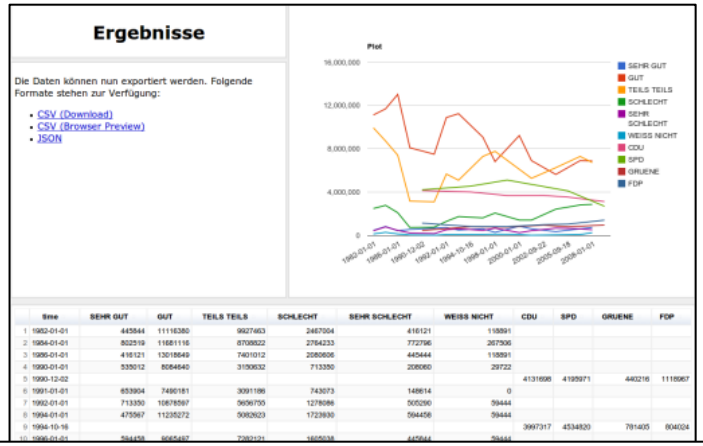


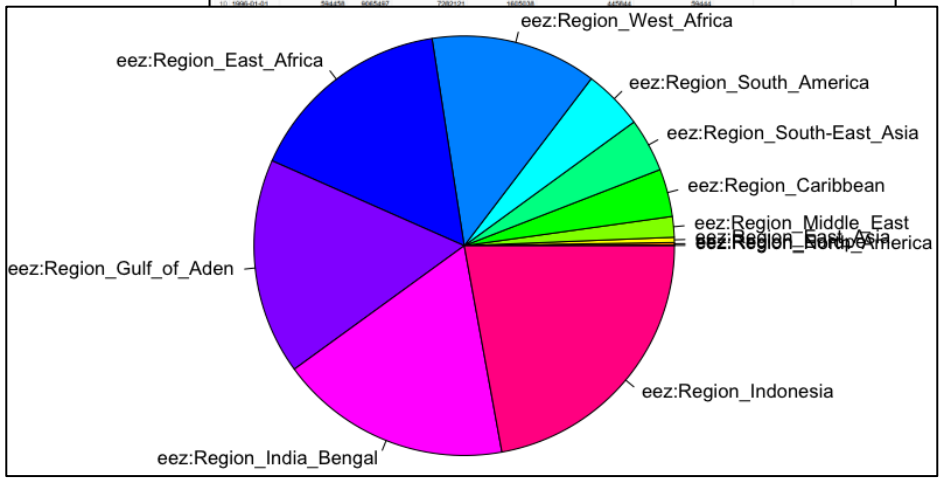
Schema Matching



Weiterverarbeitung von Forschungsdaten

- Anbindung an Analyse-, Grafik- und Statistikprogramme
  - z.B. OLAP-Systeme, Google Visualization API, R Project
  - Unterstützung bei der Beurteilung von Relevanz und statistischer Qualität
  - Ersteinschätzung von Daten (z.B. Trends)





Source: http://www.few.vu.nl/~wrvhage/#research

Mehrwerte für Wissenschaftler

- Unterstützung bei der Vorverarbeitung von Forschungsdaten für die Analyse
  - Konvertierung
  - Schema Matching
  - Datenharmonisierung
- Einfacherer Umgang / Interaktion mit heterogenen Forschungsdaten
  - Berechnungen durchführen
  - Visualisierungen erstellen
  - Weiterverarbeitung in Statistikprogrammen

LOD für Datenanbieter

- Interne und externe Vernetzung von Daten
  - Eigene Datenbestände um Daten aus dem Web anreichern
  - Steuerbar: Eigener Scope/Content kann klar abgetrennt werden
    - durch mehr Links nach außen oder
    - einbeziehen zusätzlicher Metadaten
- Präzisere Inhaltserschließung von Daten und Informationen durch semantisch ausdrucksstarke Ontologien
  - Unterstützung durch Suchmaschinen über Mapping zu schema.org
- Anbieten von Mehrwertdiensten basierend auf LOD-Daten
  - z.B. für die Analyse von Forschungsdaten
- Höhere Sichtbarkeit im Web

Zusammenfassung

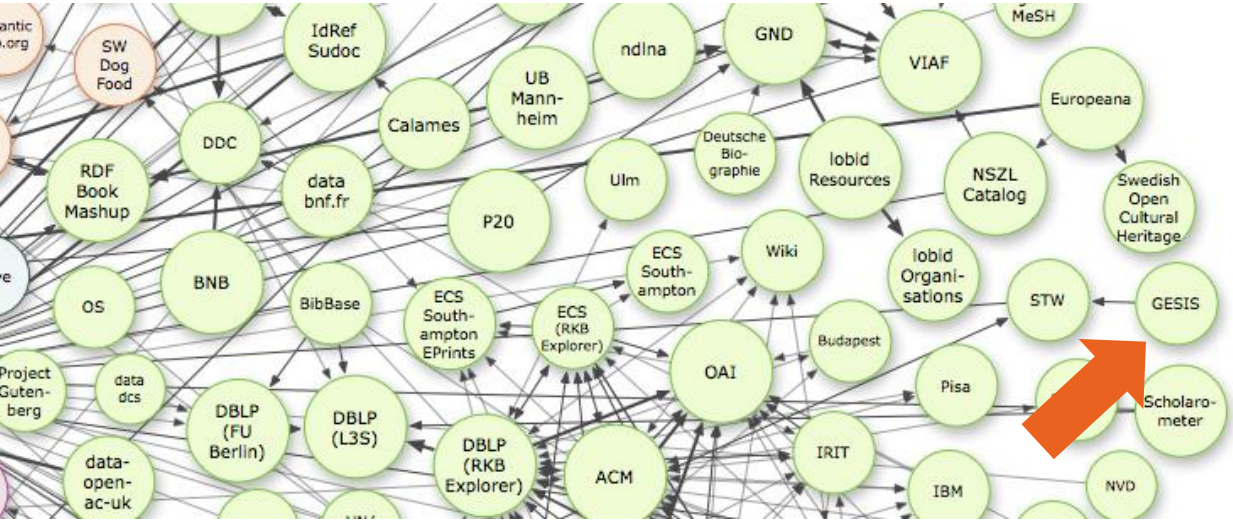
LOD im wissenschaftlichen Forschungsprozess kann

- Forschungsdaten und Informationen enger miteinander vernetzen
- Integrierte Suche über Datenbestände, Organisationsgrenzen hinweg ermöglichen
- Präzisere Suchergebnisse durch fein-granulare Inhaltserschließung liefern
- Typische Schritte der Vor- und Weiterverarbeitung von Forschungsdaten vereinfachen

Offene Punkte

- Offenheit, Datenschutz, lizenzrechtliche Problemstellungen
- Forschung transparenter und nachvollziehbar machen
  - z.B. Forschungsergebnisse veröffentlichen und nachvollziehbar machen
  - Wissenschaftler sind selbst gefragt (stark community-abhängig)
- Anwendungsszenarien von LOD für Primärforscher
  - z.B. Unterstützung bei der Planung und Durchführung von Studien

Vielen Dank!



Benjamin Zapilko
<benjamin.zapilko@gesis.org>
GESIS – Leibniz-Institut für Sozialwissenschaften
Wissenstechnologien für Sozialwissenschaften (WTS)