gesis
Leibniz-Institut für Sozialwissenschaften

Discovering Links for Metadata Enrichment on Computer Science Papers

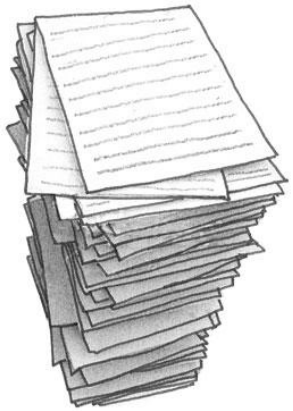
At SWIB 2012 - Cologne

Technical Report: http://bit.ly/Tiegi9

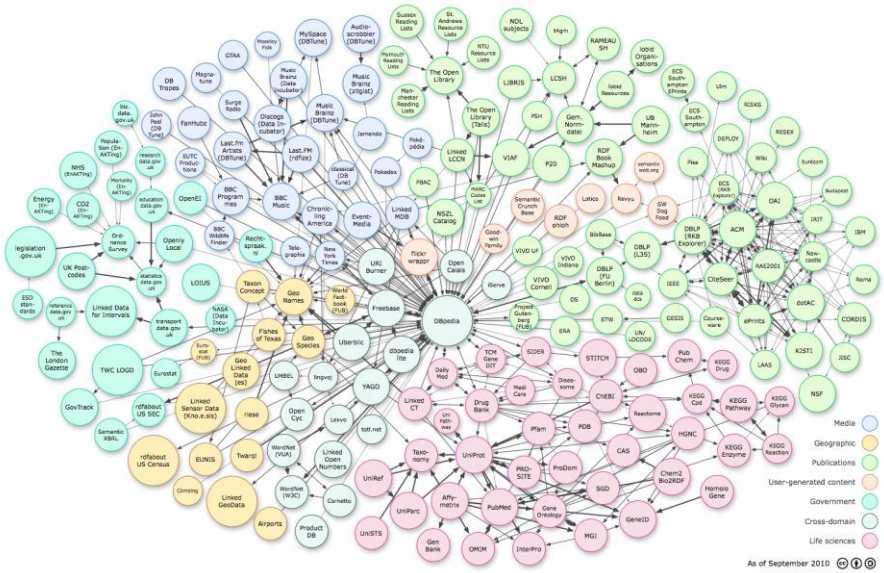
http://www.gesis.org/publikationen/gesis-technical-reports/

Johann Schaible and Philipp Mayr
GESIS - Leibniz Institute for the Social Sciences
{johann.schaible, philipp.mayr}@gesis.org

Scenario



Title, Authors, Publication Date



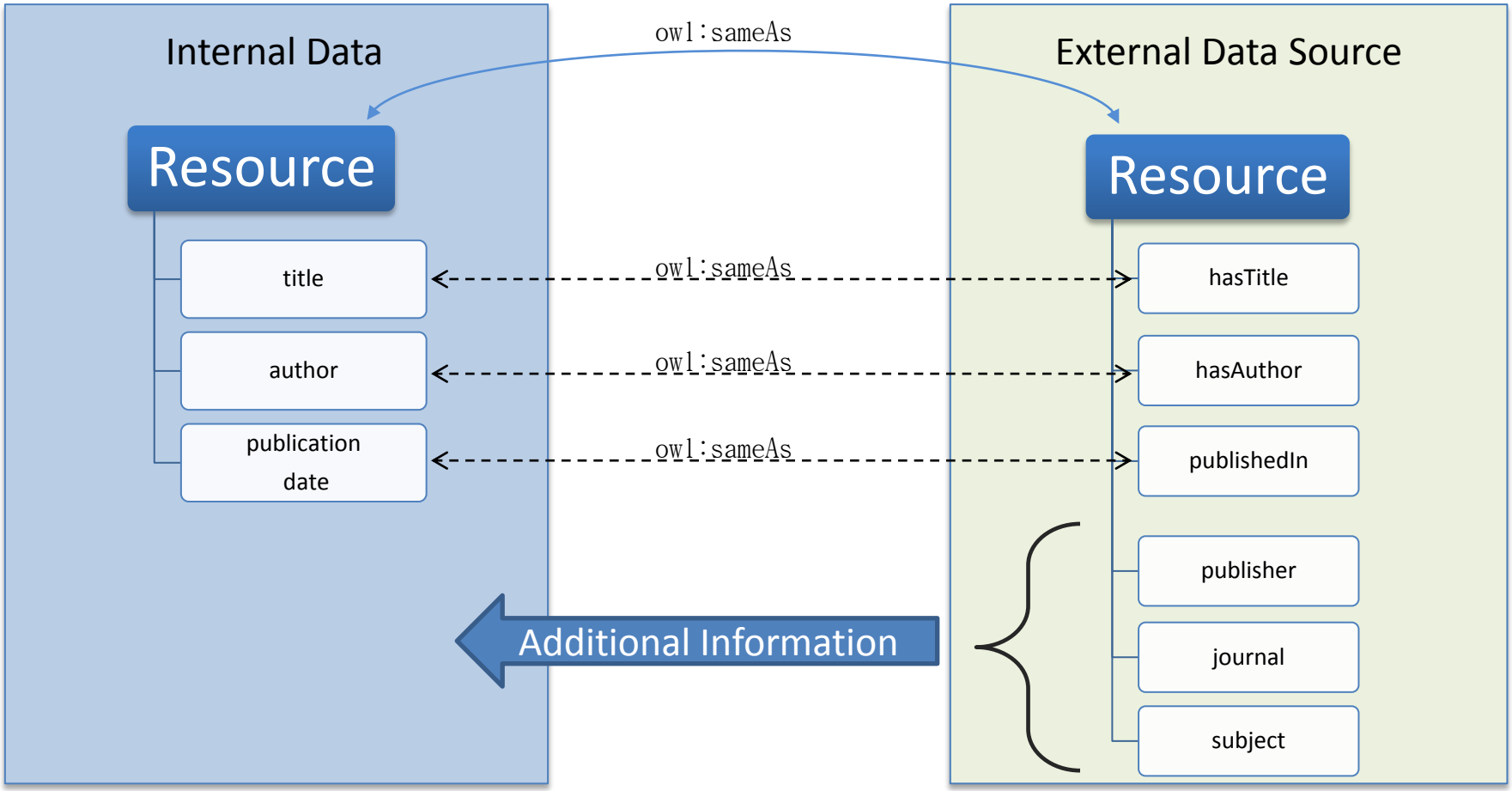
Title, Authors, Publication Date, Journal, Publisher, Conference, Abstract, Related Work, etc.

The Main Objectives

1. How to interlink internal data with the external data sources?

2. How to use an interlinking to enrich the metadata of a paper?

How to interlink Data?



gesis
Leibniz-Institut für Sozialwissenschaften

The External Data Sources

DBLP

- Data
  - About Computer Science Proceedings & Journals
  - Articles
  - Information and links about and to authors
- Access[1]
  - RKB Explorer
  - RKB SPARQL Endpoint
  - RDF/XML Dump
    - 13 GB File
    - Semantic Sitemap RKB split by year

ACM

- Data
  - Publications of the ACM
  - Details of the authors
- Access[2]
  - RKB Explorer
  - RKB SPARQL Endpoint
  - RDF/XML Dump
    - Semantic Sitemap RKB split by type

SW Conference Corpus

- Data
  - About Semantic Web Conferences & Workshops
  - Presented Papers
  - Authors, Attendants etc.
- Access[3]
  - SPARQL Endpoint
  - SNORQL Explorer
  - RDF/XML Dump Split by Conferences & Workshops

1. http://dblp.rkbexplorer.com/
2. http://acm.rkbexplorer.com/
3. http://data.semanticweb.org/documentation/user/faq

Lars' Internal Dataset

```
1.  <rdf:Description rdf:about="http://lars.org/Paper/001">
2.     <rdf:type rdf:resource="http://purl.org/linked-data/cube#DataSet"/>
3.     <dcterms:creator rdf:resource="http://lars.org/persons/johndavies"/>
4.     <dcterms:contributor rdf:resource="http://lars.org/persons/paulwarren"/>
5.     <dcterms:contributor rdf:resource="http://lars.org/persons/yorksure"/>
6.     <dcterms:title>Semantic Technology and Knowledge Management</dcterms:title>
7.     <dcterms:date>2011</dcterms:date>
8.  </rdf:Description>
9.  <foaf:Person rdf:about="http://lars.org/persons/johndavies">
10.    <foaf:name>John Davies</foaf:name>
11.    <foaf:firstName>John</foaf:firstName>
12.    <foaf:lastName>Davies</foaf:lastName>
13. </foaf:Person>
```

1. http://linkeddatabook.com/editions/1.0/
2. http://wifo5-03.informatik.uni-mannheim.de/bizer/pub/LinkedDataTutorial/
3. http://aims.fao.org/lode/bd

gesis
Leibniz-Institut für Sozialwissenschaften

A minimized DBLP & SWCC excerpt

```
<akt:Book-Section-Reference rdf:about="http://dblp.rkbexplorer.com/id/...">
        <akt:has-title>Semantic Technology and Knowledge Management.</akt:has-title>
        <akt:has-author>
            <akt:Person rdf:about="http://dblp.rkbexplorer.com/id/people-...">
                <akt:full-name>John Davies</akt:full-name>
            </akt:Person>
        </akt:has-author>
        <akt:has-date>
            <akts:Calendar-Date rdf:about="http://www.aktors.org/ontology/date#2011">
                <akts:year-of>2011</akts:year-of>
            </akts:Calendar-Date>
        </akt:has-date>
</akt:Book-Section-Reference>
```

```
<swrc:InProceedings rdf:about="http://data.semanticweb.org/.../2">
  <swc:isPartOf rdf:resource="http://data.semanticweb.org/conference/eswc/2011/proceedings"/>
  <dc:creator rdf:resource="http://data.semanticweb.org/person/dimitrios-koutsomitropoulos"/>
  <dc:title>A Structured Semantic Query Interface for Reasoning-based Search and
Retrieval</dc:title>
  <bibo:authorList rdf:resource="http://data.semanticweb.org/conference/eswc/...
/2/authorlist"/>
  <swrc:year>2011</swrc:year>
</swrc:InProceedings>

<rdf:Description rdf:about="http://data.semanticweb.org/.../2/authorlist">
  <rdf:_1 rdf:resource="http://data.semanticweb.org/person/dimitrios-koutsomitropoulos"/>
</rdf:Description>
```

Discovering Links with Silk[1]

- Input
  - Specify data sources as SPARQL endpoint or RDF/XML dump
  - Specify output file, where the links are to be saved
  - Specify linking tasks, e.g. owl:sameAs
- Output
  - SPARQL Update with discovered links
  - Discovered links are added to the specified output file

```
<http://lars.org/Paper/001>
        <http://www.w3.org/2002/07/owl#sameAs>
                <http://dblp.rkbexplorer.com/id/conf/birthday/DaviesWS11>
```

1) https://www.assembla.com/spaces/silk/wiki/dg7jfup58r4jZseJe5cbLA

How to use links for enrichment?

1. Add the discovered links to the internal dataset, thus making a hyper reference to the external data sources

2. Utilize the links to perform a query on the external data sources, thus adding their metadata to the internal dataset

Adding the links

- Advantage
  - Following links leads to all further information provided by other data publishers
  - Minimum of effort needed to include the discovered links
  - Automatic up-to-date, if external data provider change their data
- Disadvantage
  - Reliance on the external data provider. (→ If URIs are changed)
  - dereferencing of the link (→ Web representation, RKB Explorer, XML representation)

```
1.  <rdf:Description rdf:about="http://lars.org/Paper/001">
2.    <rdf:type rdf:resource="http://purl.org/linked-data/cube#DataSet"/>
3.    <dcterms:creator rdf:resource="http://lars.org/persons/johndavies"/>
4.    <dcterms:contributor rdf:resource="http://lars.org/persons/paulwarren"/>
5.    <dcterms:contributor rdf:resource="http://lars.org/persons/yorksure"/>
6.    <dcterms:title>Semantic Technology and Knowledge Management</dcterms:title>
7.    <dcterms:date>2011</dcterms:date>
8.
9.    <owl:sameAs>http://dblp.rkbexplorer.com/id/conf/birthday/DaviesWS11</owl:sameAs>
10.
11. </rdf:Description>
```

Performing a query to retrieve data

- Advantage
  - All information is stored internally
  - No reliance on the external data provider
- Disadvantage
  - More effort needed for designing a query
  - Not up-to-date if external data provider change their data

```
1.  <rdf:Description rdf:about="http://lars.org/Paper/001">
2.    <rdf:type rdf:resource="http://purl.org/linked-data/cube#DataSet"/>
3.    <dcterms:creator rdf:resource="http://lars.org/persons/johndavies"/>
4.    <dcterms:title>Semantic Technology and Knowledge Management</dcterms:title>
5.    <dcterms:date>2011</dcterms:date>
6.
7.    <akt:article-of-journal>
8.     Foundations for the Web of Information and Services
9.    </akt:article-of-journal>
10.   <akt:has-web-address>
11.     http://dx.doi.org/10.1007/978-3-642-19797-0_5
12.   </akt:has-web-address>
13. </rdf:Description>
```

Silk – lessons learned

- Silk Usability
  - Silk Workbench is very well structured and intuitively to use
  - The drag-and-drop functionality is very user friendly and connecting two properties with a comparator is straightforward
  - Silk has its own syntax for defining linkage rules
  - Loading big RDF dumps takes long. No progress bar is shown
  - If no links are found, Silk just displays an empty screen, without any messages
- Silk Results
  - Each dataset was compared with itself. Silk found all matches easily
  - Two datasets with a different schema but with the same resources. Silk found all matches, but defining linkage rules was not straightforward
  - Comparing more that 2 properties often resulted in an error message stating, that Silk was not able to execute queries in parallel.
  - Silk's linkage learning function did not work

Conclusion

- Datasets from all involved data source have to be known (→ on schema and instance level)
- Knowhow in RDF, Linked Data, link discovery tools, and SPARQL are needed for a good and effective enrichment
- “Computer Science Papers” is a good demonstration use case, but how is it with data from other domains?

Questions and Discussion

Thank You