Supporting research data services by LOD based authority data

Timo Borst

Workshop: Metadata and Persistent Identifiers for Social and Economic Data

Berlin
7.-8.5.2012
Overview

1. Metadata handling in current information systems

2. Linking metadata to authority data (persons)

3. Embedding authority data into library and information workflows

4. Summary and outlook
Metadata handling in current information systems

- Information systems
  - String based capture with (if at all) internal control of names, institutions, subjects etc.

- Library systems
  - Linking to internal identifiers (e.g. German PPN identifier), in a collaborative, but 'closed' environment (intransparent and non-accessible to others)
  - Use of proprietary control characters and delimiters to "tag" identifiers
  - No continuous policy for introducing authority files (no general guideline, but case based introduction)
Cataloguing and metadata maintenance

- Aggregation: In a distributed & networked environment, metadata is harvested and aggregated from different sources (not only library systems, but repositories, publishers’ data, web pages etc.)
- Especially concerning research data, you cannot expect metadata to be located in one place (different registries for different disciplines and data providers)
- Lessons to be learned from harvesting publication metadata:
  - Heterogenous in spelling -> ambiguous names
  - Compliance in spelling
  - Significant amount of duplicates
  - Registries for research data (yet) do not help
Cataloguing and metadata maintenance

Approach to tackle these issues (among others for disambiguation, e.g. algorithmic and social)

- Metadata fields for persons, subjects and organizations are linked to identifiers which are persistent, accessible and linkable according to LOD principles (stable URIs, HTTP-requests, content negotiation, statements about these identifiers etc.)
- Local customizations (language dependent labels for persons, subjects etc.) still must be possible
- In the future, registries for research data may handle author identification better than traditional biographical data
Data modelling based on Semantic Web / SKOS standards

- Modelling of concepts…

<table>
<thead>
<tr>
<th>URI</th>
<th>lang_de</th>
<th>lang_en</th>
<th>lang_fr</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://zbw.eu/stw/descriptor/19664-4">http://zbw.eu/stw/descriptor/19664-4</a></td>
<td>Finanzmarktkrise</td>
<td>financial crisis</td>
<td>crise financière</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>broader Terms</th>
<th>narrower Terms</th>
<th>altLabel</th>
<th>close Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;stw/descriptor/18750-1&gt;</td>
<td>&lt;stw/descriptor/18750-1&gt;</td>
<td>“Krise der Finanzmärkte”</td>
<td>de</td>
</tr>
</tbody>
</table>

Table 1 Concept representation with URIs.

- …and persons according to SemWeb principles and SKOS standard:
Example PND-/GND-Record

```xml
<gnd:DifferentiatedPerson rdf:about="http://d-nb.info/gnd/118598546">
  <gnd:preferredNameForThePerson>Benedikt &lt;XVI., Papst&gt;</gnd:preferredNameForThePerson>
  <gnd:preferredNameEntityForThePerson rdf:parseType="Resource">
    <gnd:epithetGenericNameTitleOrTerritory>Papst</gnd:epithetGenericNameTitleOrTerritory>
    <gnd:counting>XVI.</gnd:counting>
    <gnd:personalName>Benedikt</gnd:personalName>
  </gnd:preferredNameEntityForThePerson>
  <gnd:variantNameForThePerson>Ratzinger, Joseph</gnd:variantNameForThePerson>
  <gnd:variantNameEntityForThePerson rdf:parseType="Resource">
    <gnd:forename>Joseph</gnd:forename>
    <gnd:surname>Ratzinger</gnd:surname>
  </gnd:variantNameEntityForThePerson>
  <gnd:gender rdf:resource="http://d-nb.info/standards/vocab/gnd/Gender#male"/>
  <gnd:dateOfBirth>1927</gnd:dateOfBirth>
  <gnd:dateOfBirth>16.04.1927</gnd:dateOfBirth>
  <gnd:biographicalOrHistoricalInformation>Sohn von Joseph und Maria Ratzinger, Papst seit 19.04.2005</gnd:biographicalOrHistoricalInformation>
  <gnd:professionOrOccupation rdf:resource="http://d-nb.info/gnd/4059756-8"/>
  <gnd:professionOrOccupation rdf:resource="http://d-nb.info/gnd/4044561-6"/>
  <gnd:placeOfActivity rdf:resource="http://d-nb.info/gnd/4062404-3"/>
  <gnd:gndIdentifier>118598546</gnd:gndIdentifier>
  <gnd:oldAuthorityNumber>(DE-588c)310182379</gnd:oldAuthorityNumber>
</gnd:DifferentiatedPerson>
```
Additional context information provided by authority data

- preferred labels
- alternate labels
- gender
- life data
- profession and occupations
- family relationships
- (potential) links to other (external) information like
  - other identifier systems (VIAF, ORCID, INSI, PND, …)
  - additional researcher information (from DBPedia or other LOD-datasets)
  - list of publications associated with the person
  - contact data (to be maintained by the researcher)
Additional context information provided by authority data

…all this data is supposed to be used for identifying and disambiguating persons!
Embedding authority data into library and information workflows

Example: da|ra (registration agency for social and economic data)

- **Uploading metadata**

  - Autosuggesting names with additional information for identification [http://zbw.eu/beta/econ-ws/examples/suggest2.html](http://zbw.eu/beta/econ-ws/examples/suggest2.html)

- **Updating (and curating) metadata (from external sources)**

  - Associating the name string with the identifier whose (preferred) labels (plus additional information) are exactly or closely matched.
  
  Example: dataset [http://hdl.handle.net/1902.1/13772](http://hdl.handle.net/1902.1/13772), operation: determining contributor P. Martin
Embedding authority data into library and information workflows

- **Searching data**
  - Searching for authors/contributors: Autosuggest of names associated with an authority record + research data

- **Showing data**
  - Once associated with identifiers, results could be augmented with additional linked data presented e.g. as a tooltip (to be solved: nested resources as URIs)
Sala-i-Martin, Xavier (1963 - ) Wirtschaftswissenschaftler; Tätig als Prof. für Wirtschaft an der Harvard University.
Martinez-Vazquez, Jorge (1948 - ) Wirtschaftswissenschaftler; Tätig an der Georgia State University.

Martin, Philippe J. (Wirtschaftswissenschaftler; Tätig bei Sciences-Po (Paris) und CEPR; Federal Reserve Bank <New York, NY>, Centre d'Études Prospective et d'Informations Internationales <Paris>, Université Panthéon-Sorbonne <Paris>, Center for Economic Policy Research <Stanford, Calif.>.)

Selected terms:
- Martin, Philippe J.
- Martin, Philip L.

Start entering term:
### 1. 2. Title

Globalization and Emerging Markets: With or Without Crash? [Dataset]

### 1. 3. Other Titles

Add Title

### 1. 4. Principal Investigators

<table>
<thead>
<tr>
<th>Person</th>
<th>First Name</th>
<th>Middle Name</th>
<th>Last Name</th>
<th>Martin</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Philippe</td>
<td>J.</td>
<td></td>
<td></td>
<td>Centre d'Etudes Prospetives et d'Informations Internationales</td>
</tr>
</tbody>
</table>

**Add Affiliation ID**

<table>
<thead>
<tr>
<th>Person ID</th>
<th>ID</th>
<th>Vocabulary of Person ID</th>
<th>URI Authority Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>124546781</td>
<td>gnd</td>
<td><a href="http://id-nb.info/gnd/124546781">http://id-nb.info/gnd/124546781</a></td>
</tr>
</tbody>
</table>

**Add Person ID**
Summary and outlook

- The system of scholarly communication is in general dependent on a web of trust, and in particular dependent on reliable identity management
- Information and library systems yet do not support this
- Libraries and information providers could establish a backbone for authority data which is crucial for identifying persons on the web
- Semantic Web technology could help to mesh up local information with a layer of centrally provided and updated authority data
- To become first-class objects like publications, research data resp. their records should be linked from authority data
Identifier systems

- Which one to chose?

"It will take some time to determine which agent or coalition of agents in the scholarly communication game will end up providing the most reliable and useful author identification system. [J. Pitman (2012), Author Identity and Open Bibliography, http://bulletin.imstat.org/2012/03/author-identity-and-open-bibliography/]

- Criteria
  - Coverage?
  - Integration with other identifier systems?
  - Availability?
  - Machine readable formats?
  - Support by publishers?
  - Support by researchers (authorising)?
  - Terms of use / Licensing?

- Services?
Systems for (author) identification

Candidates:

- VIAF (‘Virtual Internet Authority File’ [http://viaf.org/])
  - Players: National Libraries, OCLC
  - Goal: Interlinking of national authority records
  - Status:

- ORCID
  - Players:
  - Goal:
  - Status:

- ISNI (‘International Name Standard Identifier’ [http://www.isni.org/])
  - Players: Societies, Proquest, National Libraries, OCLC
  - Goal: To serve as a ‘bridge identifier’ for different local identifier systems
  - Status: ISO Standard, supported by OCLC
Provision of and access to Linked Authority Data

- Provision (RDF, RDFa, JSON, SPARQL-Endpoint, Turtle-files, …)
- Access: ReST (other approaches?)
- Lightweight integration into applications and augmentation of local, yet unconnected information
Authority data – other categories

According to German National Library (DNB):
Systems for (author) identification (II)

- PND
  - Players:
  - Goal:
  - Status:

- URI
  - Players:
  - Goal:
  - Status: