Towards a Semantic Data Library for the Social Sciences

International Workshop on Semantic Digital Archives
29.09.2011

Thomas Gottron, Christian Hachenberg, Andreas Harth, Benjamin Zapilko
Current Situation

Digital Libraries and Archives are facing challenges derived from the distributed publishing paradigm of the web.

- Different data formats
- Lack of metadata annotation and documentation

Often disconnected from each other and from the web.

-> Providing key services (e.g. collecting and classifying information) grows in complexity.

Researchers cannot use distributed data from the web as they are used to do in Digital Libraries or Archives.
Is there an impact of unemployment rate on the fear of not finding a job?

12/10/11
Semantic Digital Libraries

- Supported by Semantic Web and Social Networking Technologies

- Key Challenges
  - Information integration based on different metadata sources
  - Interoperability with other systems (not only DLs)
  - Robust, user-friendly and adaptable UIs

Focus lies often on textual-style data (e.g. literature)
...but researchers want to use numerical data for their research (statistics, survey data, etc.)
Statistical Analysis

- Web portals for retrieving and analyzing data
  - e.g. ZACAT, SOEPinfo, QuickCalcs
- Performing calculations directly on Linked Data
  - e.g. variance, linear regression
- Connection to established and powerful statistics tools
  - SPARQL client for the R Project
SDMX and SCOVO

- SDMX (Statistical Data and Metadata Exchange)
  - Very complex information model
  - Focus on collection, reporting and validation of statistics

- SCOVO (Statistical Core Vocabulary)
  - Describing statistics with datasets, items and dimensions
DDI (Data Documentation Initiative)

- Documenting the entire research data lifecycle
- Very complex documentation model
- Describing and maintaining survey data (micro data)
- Micro data is a basic source for aggregated data (statistics)
The RDF Data Cube Vocabulary

“A proposed standard for publishing statistical data and metadata according to the Linked Data principles, based on SDMX” (Publishing Statistics with the Data Cube Vocabulary, Richard Cyganiak)
The RDF Data Cube Vocabulary

- **Purpose**
  - Dissemination of statistics over the web as Linked Data

- **Requirements**
  - High-fidelity representation of statistical information
  - Linking with other information assets
<table>
<thead>
<tr>
<th>Year</th>
<th>Austria</th>
<th>Belgium</th>
<th>Germany</th>
<th>France</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010-10</td>
<td>4.4</td>
<td>7.9</td>
<td>7.2</td>
<td>9.1</td>
</tr>
<tr>
<td>Female</td>
<td>4.0</td>
<td>8.3</td>
<td>6.8</td>
<td>10.3</td>
</tr>
<tr>
<td>2011-01</td>
<td>4.4</td>
<td>7.7</td>
<td>6.8</td>
<td>8.9</td>
</tr>
<tr>
<td>Male</td>
<td>4.6</td>
<td>7.8</td>
<td>6.4</td>
<td>10.3</td>
</tr>
<tr>
<td>Female</td>
<td>4.3</td>
<td>7.5</td>
<td>6.1</td>
<td>8.6</td>
</tr>
<tr>
<td>2011-04</td>
<td>4.1</td>
<td>8.1</td>
<td>6.1</td>
<td>10.3</td>
</tr>
</tbody>
</table>

Harmonised Unemployment Rate by Gender (Source: Eurostat)
Statistics in RDF

<qb:Observation>
  <qb:dataset rdf:resource="/id/teilm020#ds"/>
  <sex rdf:resource="/dic/sex#M"/>
  <geo rdf:resource="/dic/geo#DE"/>
  <dcterms:date>2011-01</dcterms:date>
Framework for a Semantic Data Library

- Addressing key challenges of Semantic Digital Library services for survey or statistical data

- Composed of modules to address main obstacles for reusing data e.g. in the Social Sciences
Prototype Implementation

- **Use Case**
  - Analyzing correlations between election votes and subjective rating of economical situation

- **Data**
  - IT.NRW: election votes of German parties at the Bundestagswahl 1994-2009
Prototype Implementation

- Technical features
  - Generating RDF on-the-fly via wrappers
    • Connecting to a web interface and converting the provided CSV to RDF
  - Using Data Cube Vocabulary for modeling data
  - Linking temporal and geographical dimensions
Modules 1/4

- Common Identifier Format
  - Making not only data sets, but also underlying dimensions, measures, etc. referencable
  - Using URIs, because standard for publishing Linked Data

- Common Exchange Format
  - Using RDF Data Cube Vocabulary, because
    • Open, non-proprietary metadata model in RDF
    • Based on established SDMX information model
    • Provides semantic and self-descriptive annotation of data
Towards a Semantic Data Library for the Social Sciences
Towards a Semantic Data Library for the Social Sciences
Modules 2/4

- Retrieval of Data
  - Support for different complexities of search tasks
  - Retrieval of metadata and underlying data itself

- Examples:
  - Retrieval of data documentation to evaluate relevance, quality and suitability of data
Towards a Semantic Data Library for the Social Sciences

IT.NRW

GESIS

RDF store

RDF wrapper

SPARQL

Query generator

Bitte Maße und Dimensionen zur weiteren Bearbeitung auswählen

Maße für IT.NRW:
- Wähleranteil
- Wähleranteil der Parteien (absolut)
- CDU
- Die Linke
- FDP
- Grüne
- Die Republikaner
- Sonstige

Maße für ALLGEMIN:
- Empfundene Zufriedenheit mit der persönlichen wirtschaftlichen Lage
- Empfundene Zufriedenheit mit der allgemeinen wirtschaftlichen Lage Deutschlands

Dimensions:
- Zeit
- Ort (Niveau)
SELECT ?time ?value ?cat

WHERE {

  ?s dcterms:date ?time .
} UNION {

  {?s dcterms:date ?time .
   ?s gesis:valuelabel ?l .
  }
}

ORDER BY (?time)
Modules 3/4

- Data Linking and Integration
  - Identification of dimensions, measures, etc. (requires a detailed data description)
  - Examples:
    - Linking of temporal or geographical dimensions
    - Linking different levels of frequencies or areas, e.g. different observation intervals

→ Code lists can help!
Towards a Semantic Data Library for the Social Sciences
Towards a Semantic Data Library for the Social Sciences
Modules 4/4

- Preview and Analysis
  - Presenting key characteristics of data features e.g. in a graphical way
  - Simple calculations directly on the data, e.g. for a first insight

- Data Export and Referencing
  - Using data in full statistics application (or other tools)
  - Reproduction of data (and research results) at any time
  - Referencing via unique identifier
Towards a Semantic Data Library for the Social Sciences

User interface

Google Viz API

Result
Open Issues 1/2

- Dealing with privacy
  - Survey data has to be anonymized
    -> formalise, model and describe implications on kind and type of data

- Merging, aggregating and integrating data automatically needs

Open Issues 2/2

- Publication of self-created data
  - Unusual in the Social Sciences
  - Possibility to make own data citable might have an impact
  
  -> Creates citations and reputation
Conclusion

What‘ve we done:
- Prototype implementation of a real-world use case
- Definition of an associated requirements analysis
- Framework for the publication of and access to data

Questions?
benjamin.zapilko@gesis.org