gesis
Leibniz Institute for the Social Sciences

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**

gesis Leibniz-Institut für Sozialwissenschaften

KIT Karlsruher Institut für Technologie

WeST People and Knowledge Networks

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

Research Question

Searching for relevant Data

Data Conversion

Data Analysis

Publication

Citation

Semantic Digital Libraries

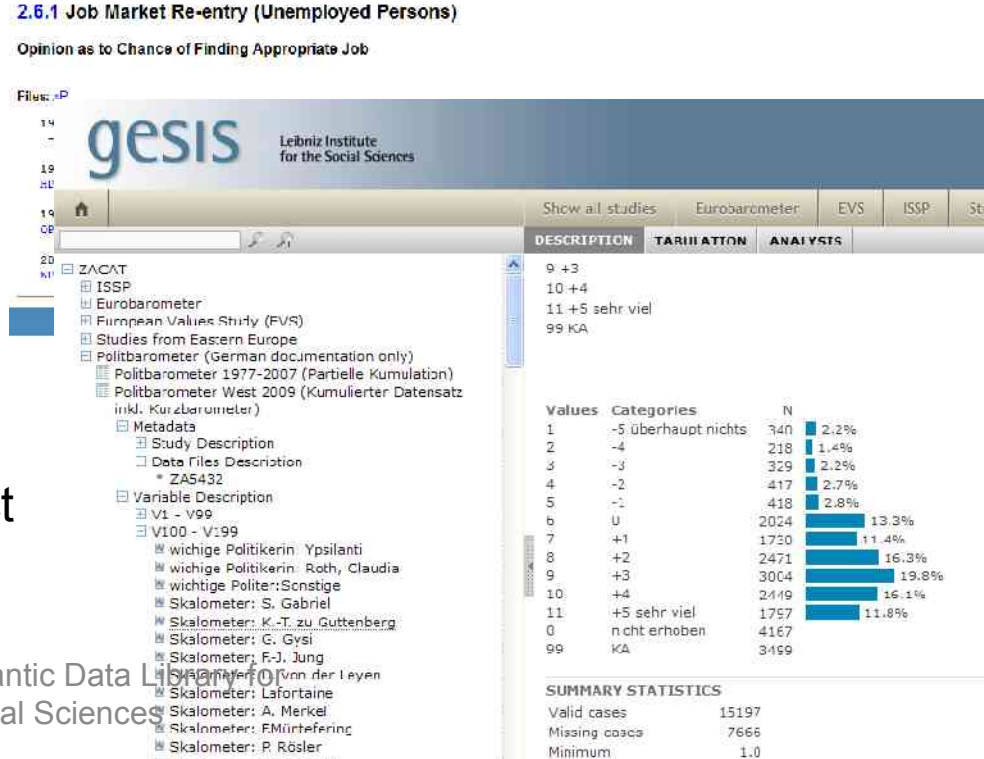
- Supported by Semantic Web and Social Networking Technologies
- Key Challenges
  - Information integration based on different metadata sources
  - Interoperability with other systems (not only DLs)
  - Robust, user-friendly and adaptable UIs

- 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
  - Focus on data, no attributes, no complex structure

DDI (Data Documentation Initiative)

- Documenting the entire research data life cycle
- Very complex documentation model
- Describing and maintaining survey data (micro data)
- Micro data is a basic source for aggregated data (statistics)
- No RDF representation yet, but effort has been started

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
  - Re-use of artifacts
- Scope
  - Data Structure Definition
    - Specification of dimensions, attributes, measures, code lists, etc
  - Dataset
    - Includes observation data, dimensions, attributes, measures, etc.

Harmonised Unemployment Rate by Gender (Source: 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
…	…	…	…	…	…	…

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>
   <sdmx-measure:obsValue>6.8</sdmx-
    measure:obsValue>
</qb:Observation>
```

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
  - GESIS: Excerpt of Cumulated ALLBUS (German General Social Survey) 1980-2008 – only NRW

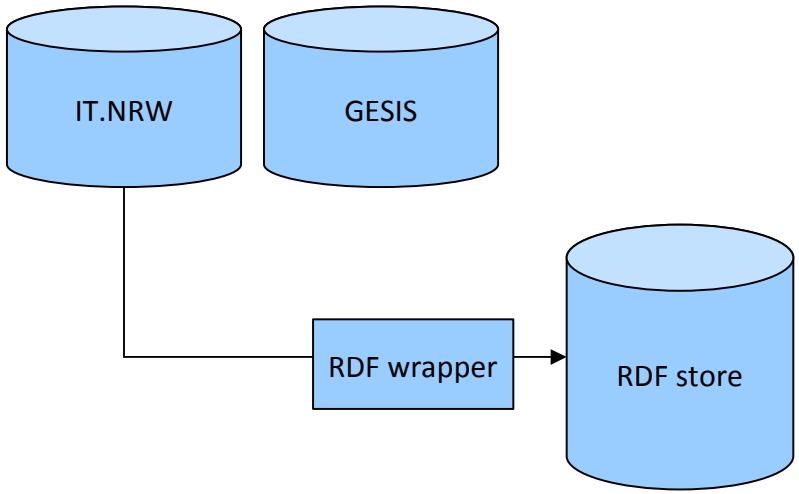
Prototype Implementation

- Technical features
  - Generating RDF on-the-fly via wrappers
    - Connecting to an web interface ant converting the provided CSV to RDF
  - Using Data Cube Vocabulary for modeling data
  - Linking temporal and geographical dimensions
  - Export data to CSV and JSON

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

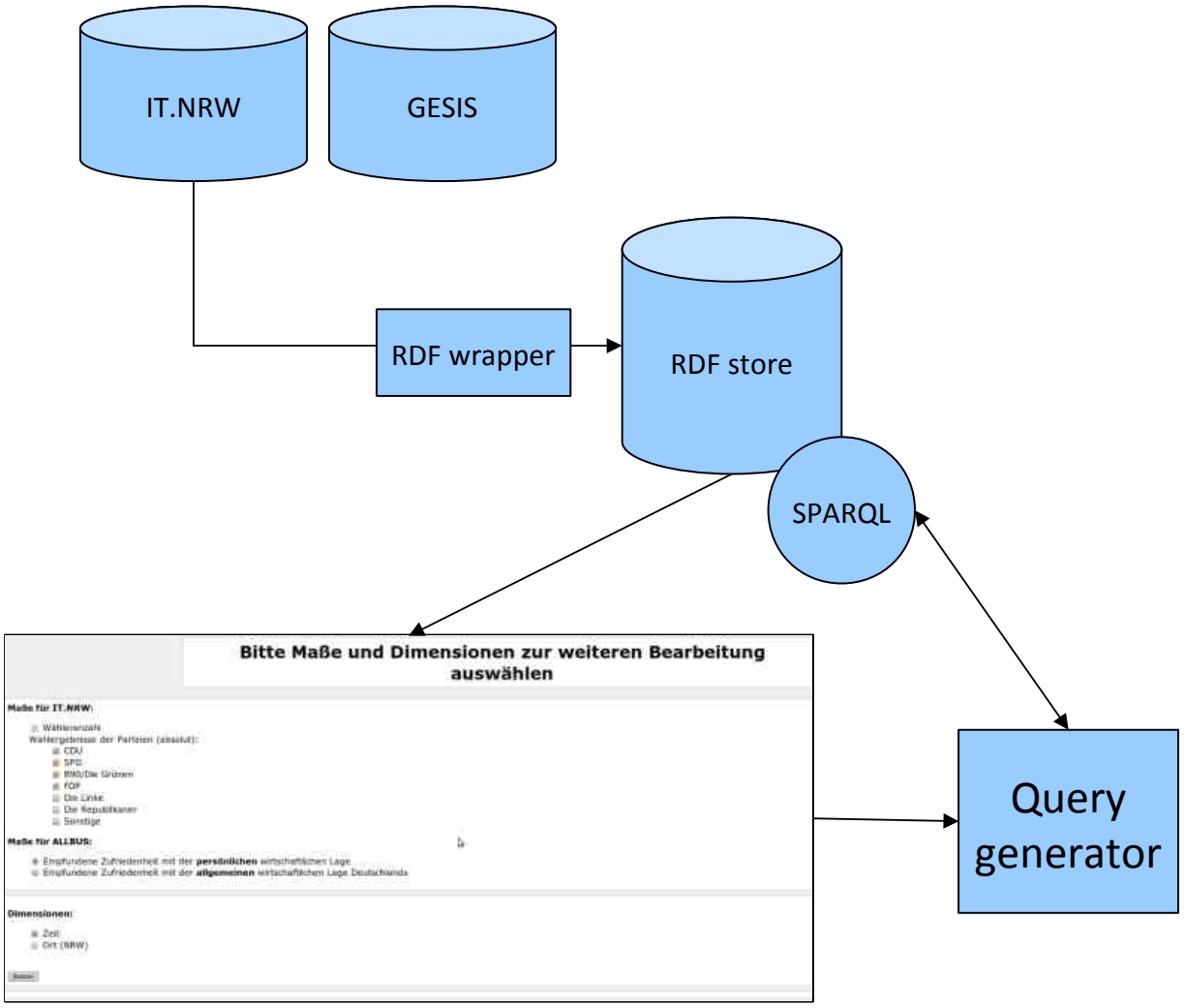

IT.NRW
GESIS
RDF wrapper
RDF store

```
<qb:Observation>
   <qb:dataset rdf:resource="./data?
     code=14111#ds"/>
   <dcterms:date>2009-09-27</dcterms:date>
   <geo rdf:resource="./geo.rdf#05"/>
   <partei
     rdf:resource="./parteien.rdf#SPD"/>
   <sdmx-measure:obsValue>2678956</sdmx-
     measure:obsValue>
</qb:Observation>
```

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
    - Retrieval of GDP of European Countries in Euros

      ⇨ a second source might provide currency conversion rates


IT.NRW
GESIS
RDF wrapper
RDF store
SPARQL
Bitte Maße und Dimensionen zur weiteren Bearbeitung auswählen
Query generator

```
SELECT  ?time ?value ?cat

WHERE {
  {
    ?s qb:dataset <http://lod.gesis.org/gesis-lod-pilot/data?code=14111#ds> .
    ?s dcterms:date ?time .
    ?s itnrw:partei ?partei .
    ?partei rdfs:label ?cat .
    ?s sdmx-measure:obsValue ?value .
  } UNION {
    ?s dcterms:date ?time .
    ?s gesis:variable <http://lod.gesis.org/lodpilot/ALLBUS/ZA4570agg.rdf#var11> .
    ?s gesis:valuelabel ?l .
    ?l rdfs:label ?cat .
    ?s sdmx-measure:obsValue ?value .
  }
}
ORDER BY (?time)
```

Modules 3/4

- Data Linking and Integration
  - Identification of dimensionen, 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!
  - Scaling and aggregation of data

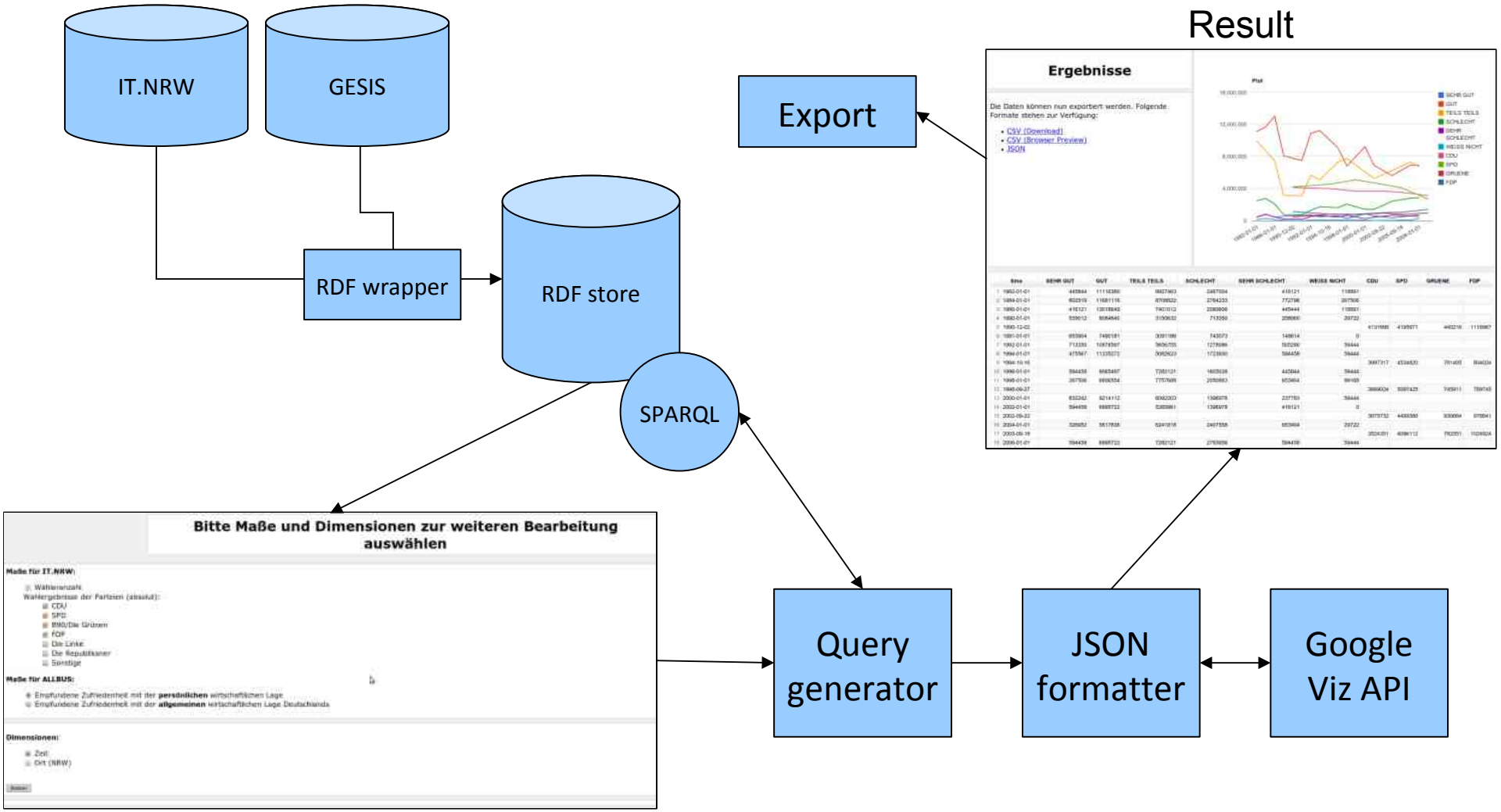
```
<skos:hasTopConcept>
    <skos:Concept rdf:about="#CDU">
    <rdfs:label>CDU</rdfs:label>
    </skos:Concept>
</skos:hasTopConcept>
…
<skos:hasTopConcept>
    <skos:Concept rdf:about="#SPD">
    <rdfs:label>SPD</rdfs:label>
    </skos:Concept>
</skos:hasTopConcept>
```

```
<skos:hasTopConcept>
     <skos:Concept rdf:about="#05554068">
     <rdfs:label>Vreden, Stadt</rdfs:label>
     </skos:Concept>
</skos:hasTopConcept>
...
<skos:hasTopConcept>
     <skos:Concept rdf:about="#05">
     <rdfs:label>Nordrhein-Westfalen</rdfs:label>
     <owl:sameAs rdf:resource="http://lod.gesis.org/lodpilot/ALLBUS/geo.rdf#50"/>
     <owl:sameAs
        rdf:resource="http://estatwrap.ontologycentral.com/dic/geo#DEA"/>
     </skos:Concept>
</skos:hasTopConcept>
```

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



User interface

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
  - Standardized information about bias and weights
  - Transformation rules?

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