Transforming Statistical Linked Data for Use in OLAP Systems

Benedikt Kämpgen, Andreas Harth
7th International Conference on Semantic Systems, Graz
7 September 2011
Outline

1. Scenarios of Statistical Linked Data (SLD)
2. Approach – OLAP Systems
3. Mapping Multidimensional Model and SLD
4. Experiments
5. Lessons Learned
6. Related Work
7. Conclusion
Statistical Linked Data – Scenarios

**Scenario 1:** Influence of gross domestic product growth on unemployment fear?

- GDP growth statistic
- Unemployment fear survey

**Scenario 2:** Comparison of Eurostat EU 2020 indicators?

- Greenhouse gas emission
- Employment rate
- Energy intensity
- ...

Semantic Web technologies promise interoperability
SLD - Definition

Statistics

“collection, analysis, interpretation, and presentation of masses of numerical data” Merriam Webster Dictionary

Linked Data principles

- URIs as names for all relevant things (e.g., dataset)
- HTTP URIs to look up those names (e.g., http://estatwrap.ontologycentral.com/id/tsieb020#ds)
- At lookup, useful information using the standards RDF, SPARQL (e.g., location of actual data)
- Reuse of URIs from other sources (e.g., two statistics talking about the same country)

http://www.w3.org/DesignIssues/LinkedData.html
SLD - Challenges

- Distributed data, e.g., sources distributed over servers
- Heterogeneous schemas, e.g., statistics as n-ary properties with time, location, ...
- Web Scale, e.g., Eurostat with 5000 datasets; UK treasury data COINS with 3-5 Mio rows

How to prepare statistical Linked Data for analysis?
Approach - Overview
Online Analytical Processing (OLAP) systems commonly used in business for analysing statistical data

- Multidimensional view of data, i.e. the numerical data depends on its many dimensions, e.g., time, location

- Interactive, navigational operations
  - Selection (dimensions, e.g., location, time)
  - Projection (metric, e.g., average GDP)
  - Drill-down/Roll-up (granularity, e.g., federal states)
  - Slice/Dice (filter, e.g., Germany, years after 2000)

- OLAP clients on OLAP servers (Data Warehouses)
Approach – Overview (2)

ETL

OLAP server

Runtime

OLAP-Client

OLAP-Client

OLAP-Client

?
Multidimensional Model and SLD reusing ontology

Requirements: Automatically and scalable prepare SLD to answer complex questions using OLAP systems

Approach: Mapping Multidimensional Model and SLD reusing ontology

Multidimensional Model (MDM) RDF Data Cube vocabulary (QB)
Institute of Applied Informatics and Formal Description Methods (AIFB)

Approach – Overview (3)

- OLAP-Client
- ETL
- OLAP server
- Linked Data
- RDF
- QB
- MDM
- OLAP Client
- OLAP Client
- OLAP Client

Runtime

ETL

OLAP server
Multidimensional Model (MDM)

Advantages
- No standard but common MDM
- Expressive enough

- Hypercube (Cube), e.g., survey data
- Fact, e.g., questionnaire
- Dimension, e.g., time
- Hierarchy, e.g., all-year-month-day
- Measure, e.g., average of GDP
- ...

Illustration of common MDM
RDF Data Cube vocabulary (QB)

- **Advantages**
  - Based on Statistical Data and Metadata Exchange (SDMX)
  - Self-descriptive data
  - Available datasets

- `qb:DataSet` – collection of statistics
- `qb:DataStructureDefinition` – defines structure of statistics
- `qb:ComponentProperty` – property used for dimensions, attributes, metrics
- `qb:Observation` – statistic
- ...

Cyganiak et al. - The RDF Data Cube vocabulary
Mapping – Multidimensional Model and RDF Data Cube vocabulary

<table>
<thead>
<tr>
<th>MDM</th>
<th>RDF (QB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Hypercube (Cube)</td>
<td>qb:DataSet + qb:DataStructureDefinition</td>
</tr>
<tr>
<td>Fact</td>
<td>qb:Observation + qb:DataSet</td>
</tr>
<tr>
<td>Dimension</td>
<td>qb:ComponentProperty</td>
</tr>
<tr>
<td>Dimension Member</td>
<td>Given by qb:codeList of skos:Concept OR instances of rdfs:range of qb:ComponentProperty</td>
</tr>
<tr>
<td>Hierarchy</td>
<td>Depends on Dimension Members, e.g., for Members of xsd:date, Hierarchy of all-year-month-day</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Measure</td>
<td>qb:MeasureProperty + possibly appropriate Aggregation Function (e.g., sum, avg, min, max, count)</td>
</tr>
<tr>
<td>Multicube</td>
<td>Cubes sharing Dimensions and Members (linked by owl:sameAs)</td>
</tr>
</tbody>
</table>

Prefixes, see http://prefix.cc/

Mapping terms of common MDM to SPARQL queries on RDF using QB
Approach – Overview (4)

- ETL
- OLAP server
- SPARQL results on QB RDF

Mapping

Runtime

OLAP-Client
...
Experiments – Implementation

- OLAP client: xmla4js
- OLAP server: Mondrian/XMLA
- ETL pipeline: PHP web service
- SPARQL engine: qcrumb.com
Experiments – GDP growth influencing unemployment fear?

- **Datasets**
  - Survey data about German employees' fear of unemployment in the last few years (http://lod.gesis.org/lodpilot/ALLBUS/ZA4570v590.rdf#ds)
  - GDP growth of European countries per year as provided by Eurostat (http://estatwrap.ontologycentral.com/id/tsieb020#ds)

- **Wanted result**

<table>
<thead>
<tr>
<th>Year</th>
<th>GDP growth</th>
<th>Percentage of people without fear of unemployment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>2010</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>...</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
MDM of Unemployment fear and GDP growth

- Cubes
  - Survey, GDP growth
- Dimensions
  - location, geo, time, answer
- Measures
  - Avg GDP growth
  - Sum of No-answers/Sum of all answers * 100
- Shared
  - location ≡ geo
  - Germany ≡ DE
  - time (xsd:date)
Experiments – Trends in Eurostat EU 2020 indicators?

- Datasets
  - Employment rate
    (http://estatwrap.ontologycentral.com/id/t2020_10#ds)
  - Greenhouse gas emissions
    (http://estatwrap.ontologycentral.com/id/t2020_30#ds)
  - ...

- Wanted result

<table>
<thead>
<tr>
<th>Country\Metric</th>
<th>Employment rate</th>
<th>Greenhouse gas emissions</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Belgium</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>...</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
## Experiments – Evaluation

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Datasets</th>
<th>Triples</th>
<th>ETL pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Unemployment fear and GDP growth</td>
<td>2</td>
<td>20 268</td>
<td>4m 23s</td>
</tr>
<tr>
<td>2a) EU 2020 indicators</td>
<td>4</td>
<td>24 636</td>
<td>10m 54s</td>
</tr>
<tr>
<td>2b) EU 2020 indicators</td>
<td>8</td>
<td>35 482</td>
<td>27m 18s</td>
</tr>
</tbody>
</table>

- Successful integration
- Bottleneck: SPARQL queries on LD
- More experiments and work needed to fully evaluate requirements

![Number of triples (x-axis) and execution time in seconds (y-axis)](chart.png)
Lessons Learned

**Automatically** and **scalable** prepare SLD to answer **complex** questions using OLAP systems?

- Automatic not always possible
- Some datasets not properly modelled, e.g., no qb:DataStructureDefinition
- Room for performance improvements
- For each query, the whole set of datasources is queried
- Links between shared Dimensions and Members are found and resolved procedurally with canonical table
- More complex questions
  - Publishers do not use Hierarchies, e.g., SKOS
  - QB does not include aggregation functions
Related Work

- **OLAP-like operations on Web sources without Semantic Web technologies**
  - E.g., Google Public Data Explorer, Tableau, Needlebase, Google Squared
  - **Problem:** Without semantic technologies, still much manual work in integrating heterogeneous datasets

- **OLAP-like operations on Web sources with Semantic Web technologies**
  - E.g., Marko Niinimäki and Tapio Niemi – An ETL Process for OLAP Using RDF/OWL Ontologies, 2009
  - **Problem:** No focus on the problem of semantic heterogeneity in datasources to be integrated in an MDM. No Linked Data
Conclusion

- Common OLAP systems can be used for analysing statistical Linked Data
- More and more data reusing RDF Data Cube vocabulary [1]
- Further research needed [2]
- Current work: Open-Source Driver for various OLAP clients

Backup: Class Diagram

Class diagram of MDM
Backup: Architecture