Towards Real-Time Calculation of Financial Indexes using XBRL and Linked Data

Andreas Harth, Alexander Büscher
With input from Benedikt Kämpgen, Forschungszentrum Informatik (FZI)
Outline

- Motivation
- Linked Data Principles
- Data Modelling with Vocabularies
- Real-time Data Integration
- Conclusion
Motivation

- Regulators use stress-test scenarios (e.g. stock markets crashes, sudden increase in unemployment rates) to predict the effect of adverse events on the financial system.
- Lenders use similar analyses to assess credit risk in their client portfolio.
- Running such “what-if” scenarios often requires integrated access to a variety of data from multiple providers.
- Ideally, systems supporting analysts should be able to run scenarios regularly and on recent data, and support interactive ad-hoc query capabilities.

Key factors in building such systems are:
- Availability of data (the more data is available, the more accurate the results of the analyses)
- The ability to include new data sources easily (having data available in standardised formats helps) and
- The ability to scale to large amounts of data.
Key Factors

- Availability of data
  - Access openly available data sources, such as Eurostat (SDMX over HTTP), Worldbank (JSON API over HTTP), SEC EDGAR (XBRL over FTP)
  - Versus commercial data providers, restricted data access
- Include new data sources easily
  - Build on a uniform interface to access data (URIs and HTTP) and a uniform data model (RDF)
  - Model and map data with vocabularies and ontologies (in RDFS and OWL)
  - Versus manual integration in Excel
- Scale to large amounts of data
  - Access, integrate and query data in near-realtime with an optimised toolchain
  - Versus running analyses once a year
Outline

- Motivation
- **Linked Data Principles**
- Data Modelling with Vocabularies
- Real-time Data Integration
- Conclusion
G8 Open Data Charter
Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

http://www.w3.org/DesignIssues/LinkedData.html

“The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.”
Linked Data Example: Deutsche Bank AG

- http://edgarwrap.ontologycentral.com/cik/1159508#id
- http://km.aifb.kit.edu/services/crunchbase/api/organization/deutsche-bank#id
- http://dbpedia.org/resource/Deutsche_Bank
- https://www.wikidata.org/entity/Q66048
- https://viaf.org/viaf/130112035/
- ...

Towards Real-Time Calculation of Financial Indexes using XBRL and Linked Data
JSON-LD Syntax

```json
{
  "@context": {
    "@vocab": "http://km.aifb.kit.edu/",
    "url": {
      "@type": "@id"
    }
  },
  "@id": "person/jane",
  "@type": "Person",
  "name": "Jane Doe",
  "jobTitle": "Professor",
  "telephone": "(425) 123-4567",
  "url": "http://www.janedoe.com/"
}
```
Wrappers

- Many sources use non-RDF data formats (e.g., CSV, SDMX, XBRL) and non-HTTP (REST) data access (e.g., FTP, SOAP)
- Thin data access layer ("wrappers") allow for uniform data access and representation
- We provide access to various sources using a Linked Data interface
  - Eurostat: http://estatwrap.ontologycentral.com/
  - SEC EDGAR: http://edgarwrap.ontologycentral.com/
- Wrappers deployed on Google App Engine; perform data conversion on demand
Outline

- Motivation
- Linked Data Principles
- **Data Modelling with Vocabularies**
- Real-time Data Integration
- Conclusion
Modelling Taxonomies

- SKOS – Simple Knowledge Organization System
- Main class is skos:Concept
- Concepts can be organised in a hierarchy via skos:narrower/skos:broader

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix : <#> .
:table-g4 a skos:Concept ;
  rdfs:label "Europe (continent)."@fr .
:table-g49 a skos:Concept ;
  skos:broader :table-g4 ;
  rdfs:label "Petits pays de l’Europe (groupement de pays)."@fr .
:table-g493 a skos:Concept ;
  skos:broader :table-g49 ;
  rdfs:label "Belgique (pays)."@fr .
```
Multi-Dimensional Data
RDF Data Cube Vocabulary

The diagram illustrates the RDF Data Cube Vocabulary, showing how different components and properties are related. It includes concepts such as `qb:DataSet`, `qb:SliceKey`, `qb:ComponentProperty`, `skos:Concept`, `sdmx:Concept`, and `qb:DimensionProperty`. The relationships and properties are indicated by arrows, showing the hierarchical and interconnected nature of the vocabulary.

The document likely discusses the use of this vocabulary in the context of real-time calculation of financial indexes using XBRL and linked data.
Correspondence between things (e.g., the GDP dataset at http://estatwrap.ontologycentral.com/id/nama_gdp_c#ds) and the source (e.g., the RDF file at http://estatwrap.ontologycentral.com/data/nama_gdp_c)
Outline

- Motivation
- Linked Data Principles
- Data Modelling with Vocabularies
- **Real-time Data Integration**
- Conclusion
Data Integration Pipeline

- Data access using Linked Data
- Data integration using Vocabularies and Mappings
- Querying using SPARQL, the query language for RDF
### KPIs in Prototype

<table>
<thead>
<tr>
<th>No.</th>
<th>KPI</th>
<th>Data origin(s)</th>
<th>Corresponding URIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>totalTurnover / numberOfLocations</td>
<td>SEC, DBpedia</td>
<td>us-gaap2009:totalTurnover</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>dbpedia-owl:numberOfEmployees</td>
</tr>
<tr>
<td>2</td>
<td>Non-interestIncome / PersonnelCosts</td>
<td>SEC, SEC</td>
<td>us-gaap2009:noninterestIncome</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>us-gaap2009:LaborAndRelatedExpense</td>
</tr>
<tr>
<td>3</td>
<td>totalTurnover / numberOfLocations</td>
<td>SEC, DBpedia</td>
<td>us-gaap2009:totalTurnover</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>dbpedia-owl:numberOfLocations</td>
</tr>
<tr>
<td>4</td>
<td>( \sum_{t=0}^{n} NetIncomeLoss(t) )</td>
<td>SEC</td>
<td>us-gaap2009:NetIncomeLoss</td>
</tr>
<tr>
<td>5</td>
<td>Price of shares</td>
<td>SEC, Yahoo</td>
<td>yahoo:Close</td>
</tr>
</tbody>
</table>
Performance of Prototype

- Calculation of KPIs for commercial U.S. banks
  - 294 banks with SIC '6022 'STATE COMMERCIAL BANKS'
  - We selected seven banks: FIFTH THIRD BANK, WILMINGTON TRUST, M&T BANK CORP, COMMERCE BANCSHARES, ASSOCIATED BANC-CORP, BANCORP SOUTH, BANK OF NEW YORK MELLON

<table>
<thead>
<tr>
<th>No.</th>
<th>Data origin(s)</th>
<th>Incl. http:GET</th>
<th>Excl. http:GET</th>
<th># of triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SEC + DBpedia</td>
<td>15m32s</td>
<td>3.16s</td>
<td>55k</td>
</tr>
<tr>
<td>2</td>
<td>SEC</td>
<td>13m58s</td>
<td>2.79s</td>
<td>51k</td>
</tr>
<tr>
<td>3</td>
<td>SEC + DBpedia</td>
<td>15m42s</td>
<td>3.55s</td>
<td>55k</td>
</tr>
<tr>
<td>4</td>
<td>SEC</td>
<td>13m48s</td>
<td>2.88s</td>
<td>51k</td>
</tr>
<tr>
<td>5</td>
<td>SEC + Yahoo</td>
<td>17m02s</td>
<td>11.30s</td>
<td>1.203k</td>
</tr>
<tr>
<td>All</td>
<td>SEC + DBpedia + Yahoo</td>
<td>18m14s</td>
<td>25.24s</td>
<td>1.211k</td>
</tr>
</tbody>
</table>

- Data access and transfer makes up a large part of the overall time
Outline

- Motivation
- Linked Data Principles
- Data Modelling with Vocabularies
- Real-time Data Integration
- Conclusion
Conclusion

- Differences between computed values and directly access values
- Many data sources already available as Linked Data
  - Several EU data sources as native Linked Data, also some national statistics organisations (e.g., Istat)
  - Eurostat, Worldbank, SEC EDGAR, Crunchbase via wrappers
  - Open availability of data could still be better
  - We would like to use European data sources for XBRL
- Toolchain works in prototypes – from live data sources from multiple providers to query results in minutes, including provenance
- Future work
  - Improve modelling of “computation arcs”
  - Investigate SHACL/ShEx for validation (rather than OWL)
  - Develop further demonstrators with real-world XBRL data and use cases