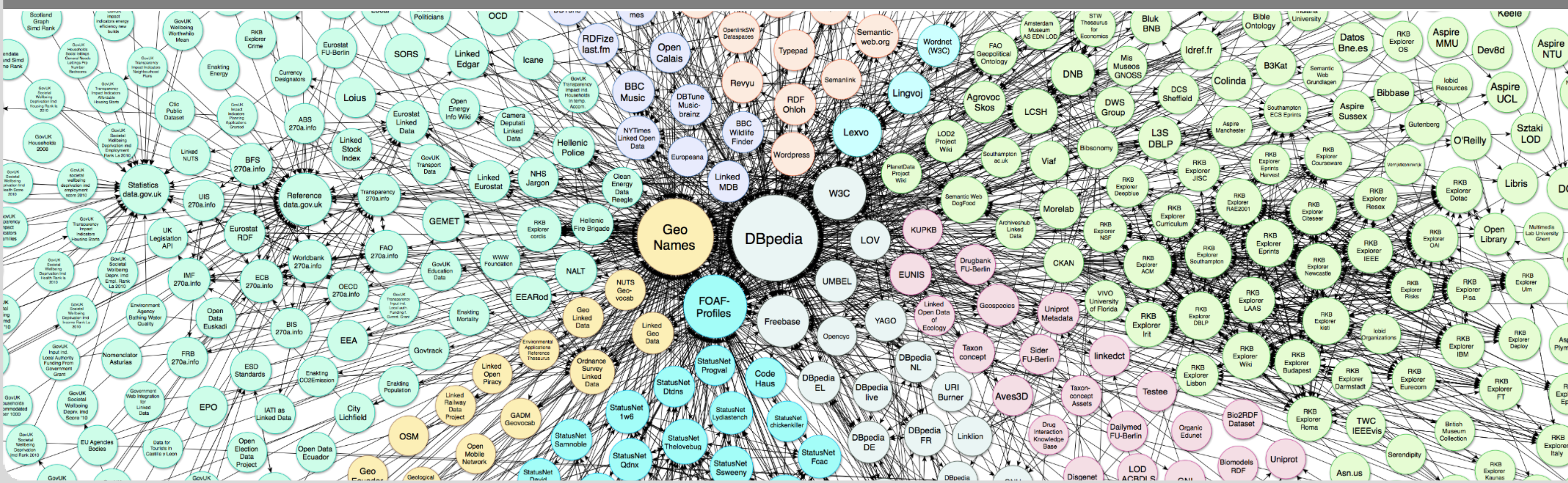
Karlsruhe Institute of Technology

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

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

WEB SCIENCE AND KNOWLEDGE MANAGEMENT



# 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/
- http://d-nb.info/gnd/16307113-5
- …

# JSON-LD Syntax

```
{
  "@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 acess 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

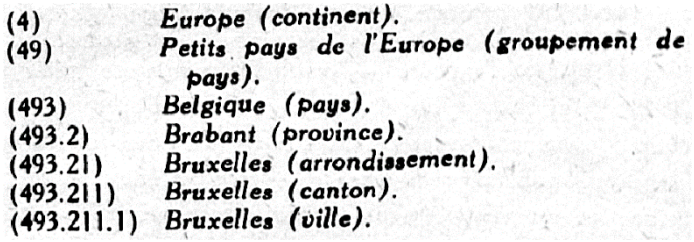
# Outline

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

# Modelling Taxonomies

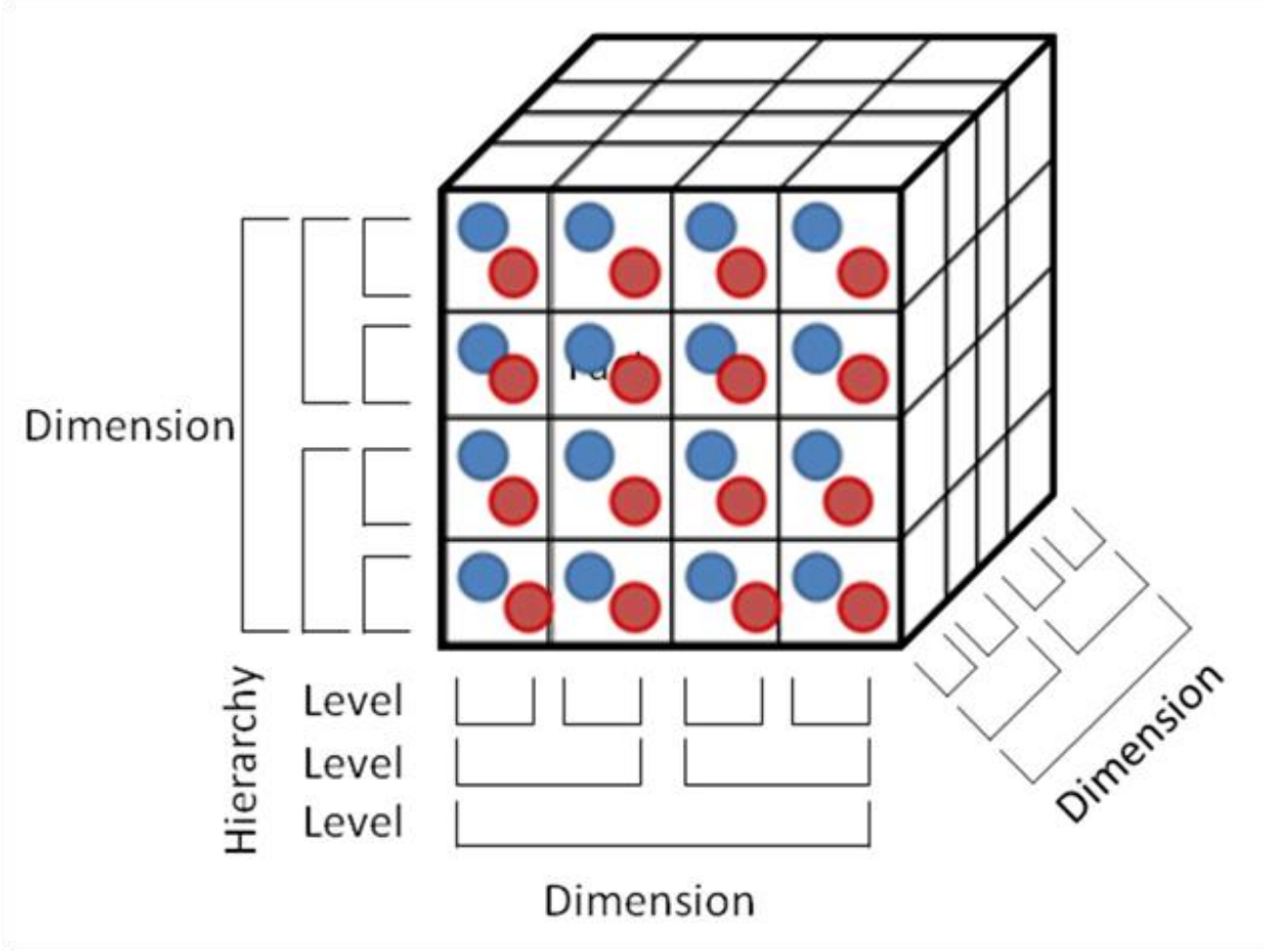
- SKOS – Simple Knowledge Organization System
- Main class is skos:Concept
- Concepts can be organised in a hierarchy via skos:narrower/skos:broader

```
@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 .
```

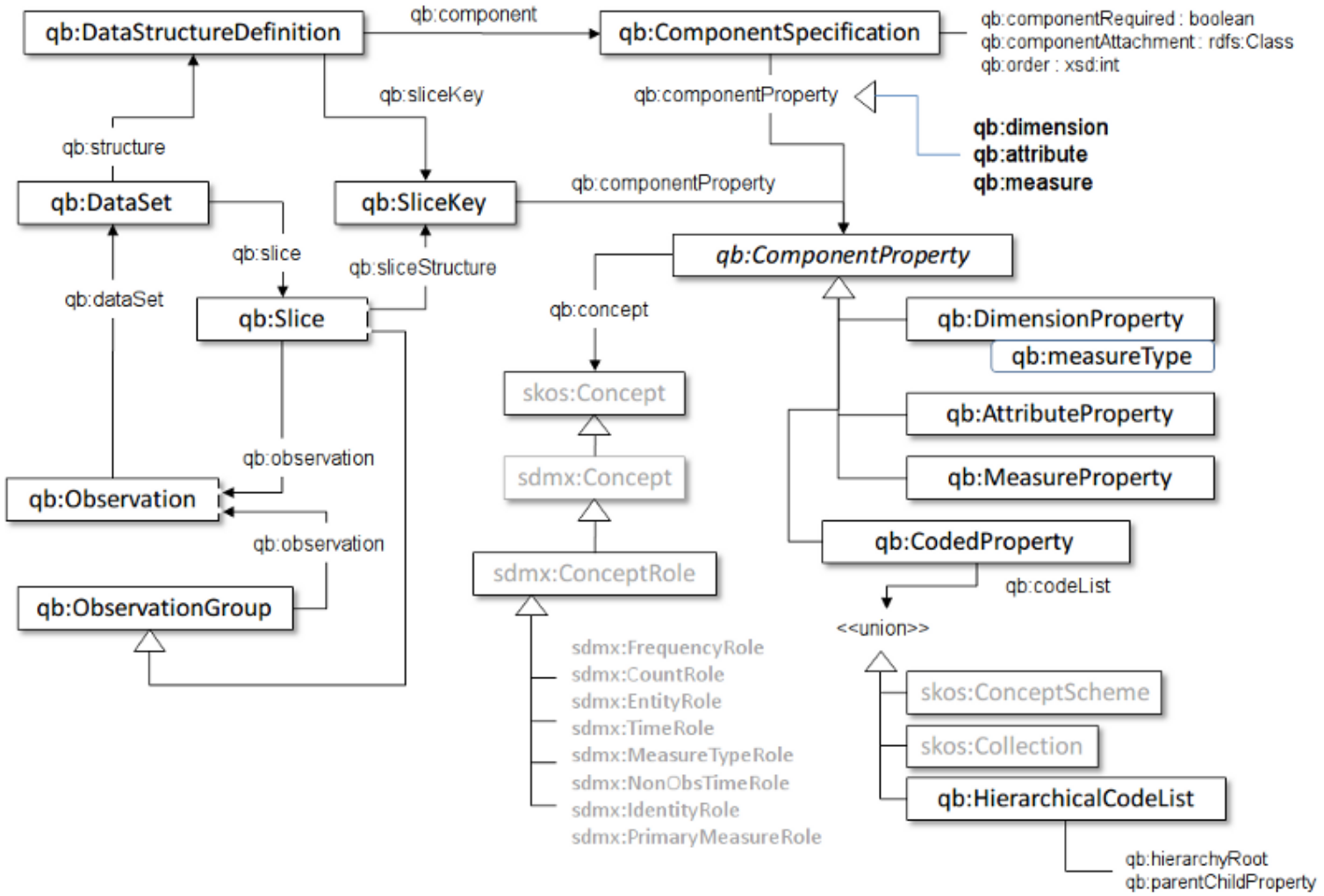


| | |
|---|---|
| (4) | *Europe (continent).* |
| (49) | *Petits pays de l'Europe (groupement de pays).* |
| (493) | *Belgique (pays).* |
| (493.2) | *Brabant (province).* |
| (493.21) | *Bruxelles (arrondissement).* |
| (493.211) | *Bruxelles (canton).* |
| (493.211.1) | *Bruxelles (ville).* |

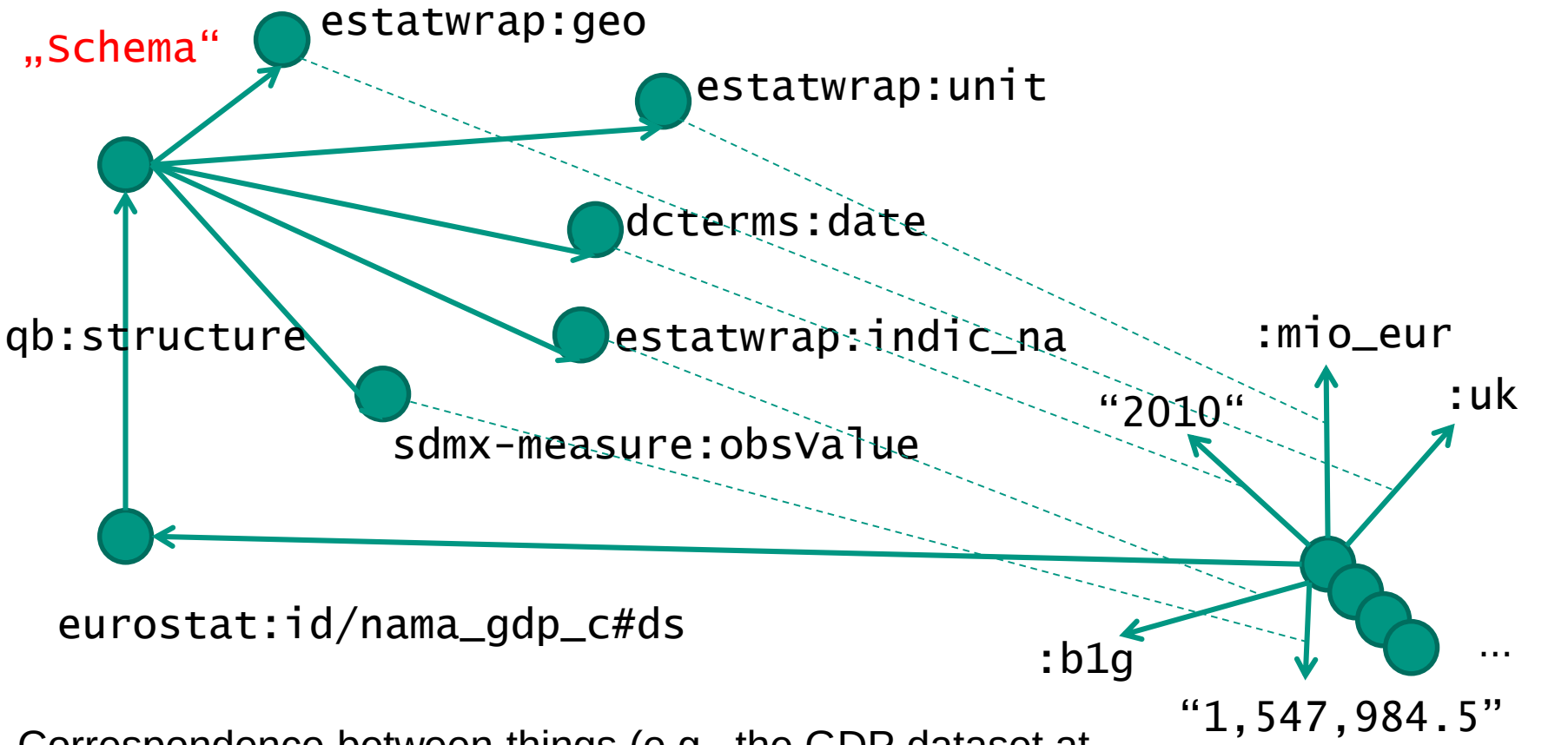
# Multi-Dimensional Data



# RDF Data Cube Vocabulary



# Eurostat Gross Domestic Product

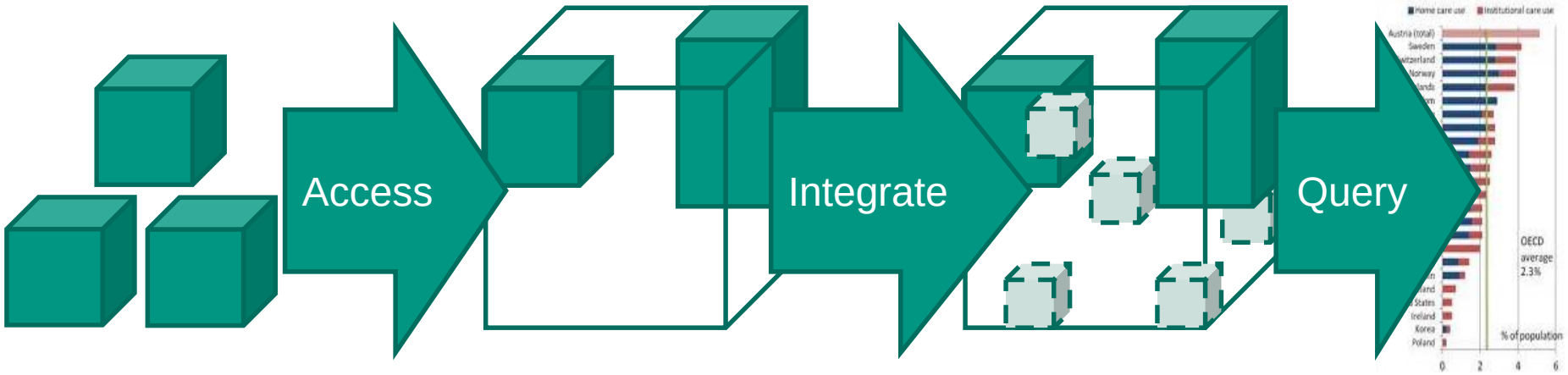



Correspondence between things (e.g., the GDP dataset at http://estatwrap.ontologycentral.com/id/nama_gdp_c#ds) and the source (e.g., the RDF file at http://estatwrap.ontologycentral.com/data/nama_gdp_c)

# 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

| No. | KPI | Data origin(s) | Corresponding URIs |
|---|---|---|---|
| 1 | $\frac{totalTurnover}{numberOfLocations}$ | SEC<br>DBpedia | us-gaap2009:totalTurnover<br>dbpedia-owl:numberOfEmployees |
| 2 | $\frac{Non-interestIncome}{PersonnelCosts}$ | SEC<br>SEC | us-gaap2009:noninterestIncome<br>us-gaap2009:LaborAndRelatedExpense |
| 3 | $\frac{totalTurnover}{numberOfLocations}$ | SEC<br>DBpedia | us-gaap2009:totalTurnover<br>dbpedia-owl:numberOfLocations |
| 4 | $\sum_{t=0}^{n} NetIncomeLoss(t)$ | SEC | us-gaap2009:NetIncomeLoss |
| 5 | Price of shares | SEC<br>Yahoo | –<br>yahoo:Close |

# 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

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

- 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