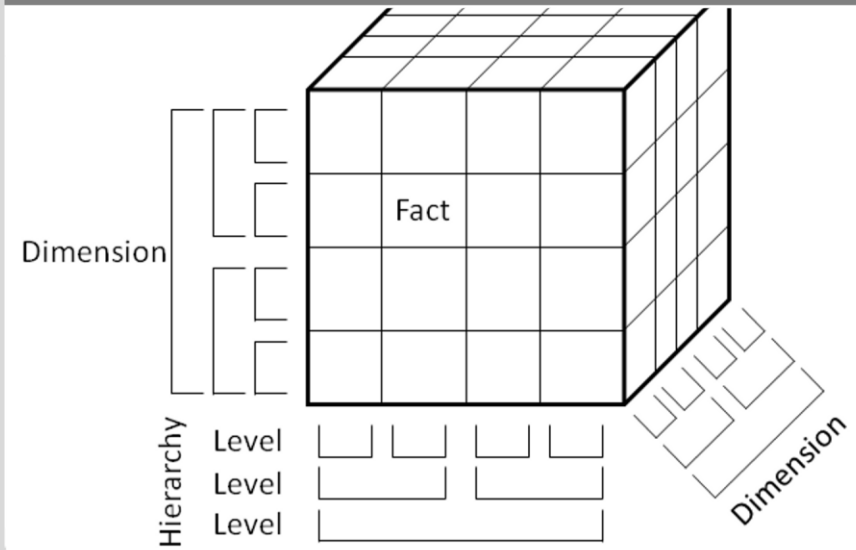
Karlsruhe Institute of Technology

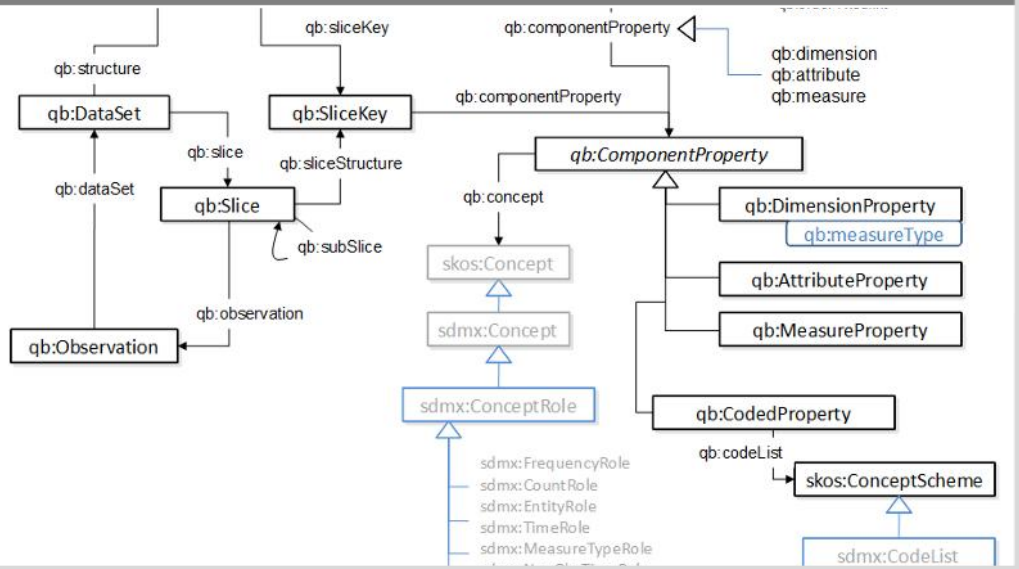
Transforming Statistical Linked Data for Use in OLAP Systems

Benedikt Kämpgen, **Andreas Harth**

7th International Conference on Semantic Systems, Graz
7 September 2011

Institute of Applied Informatics and Formal Description Methods (AIFB)





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

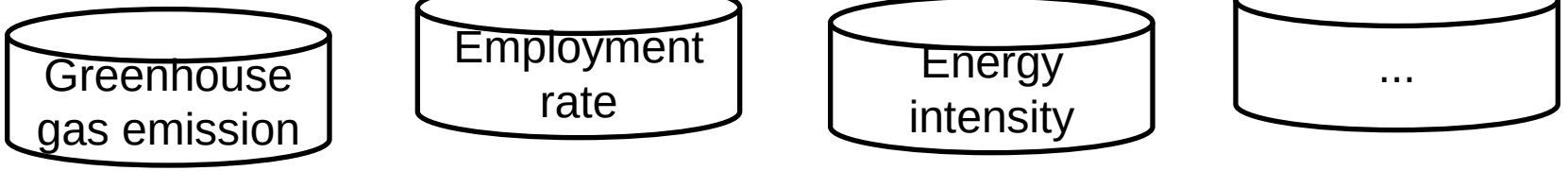
Statistical Linked Data – Scenarios

- **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?



- Semantic Web technologies promise interoperability

SLD - Definition

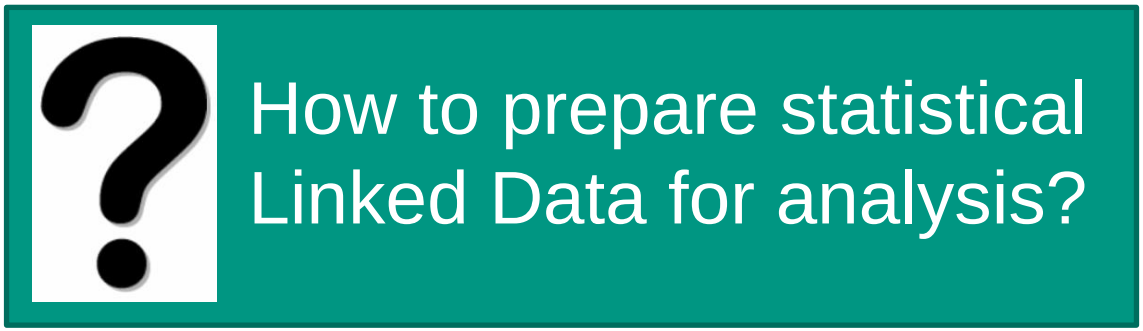
- Statistics
  - “collection, analysis, interpretation, and presentation of masses of numerical data” *Merriam Webster Dictionary*

- 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)

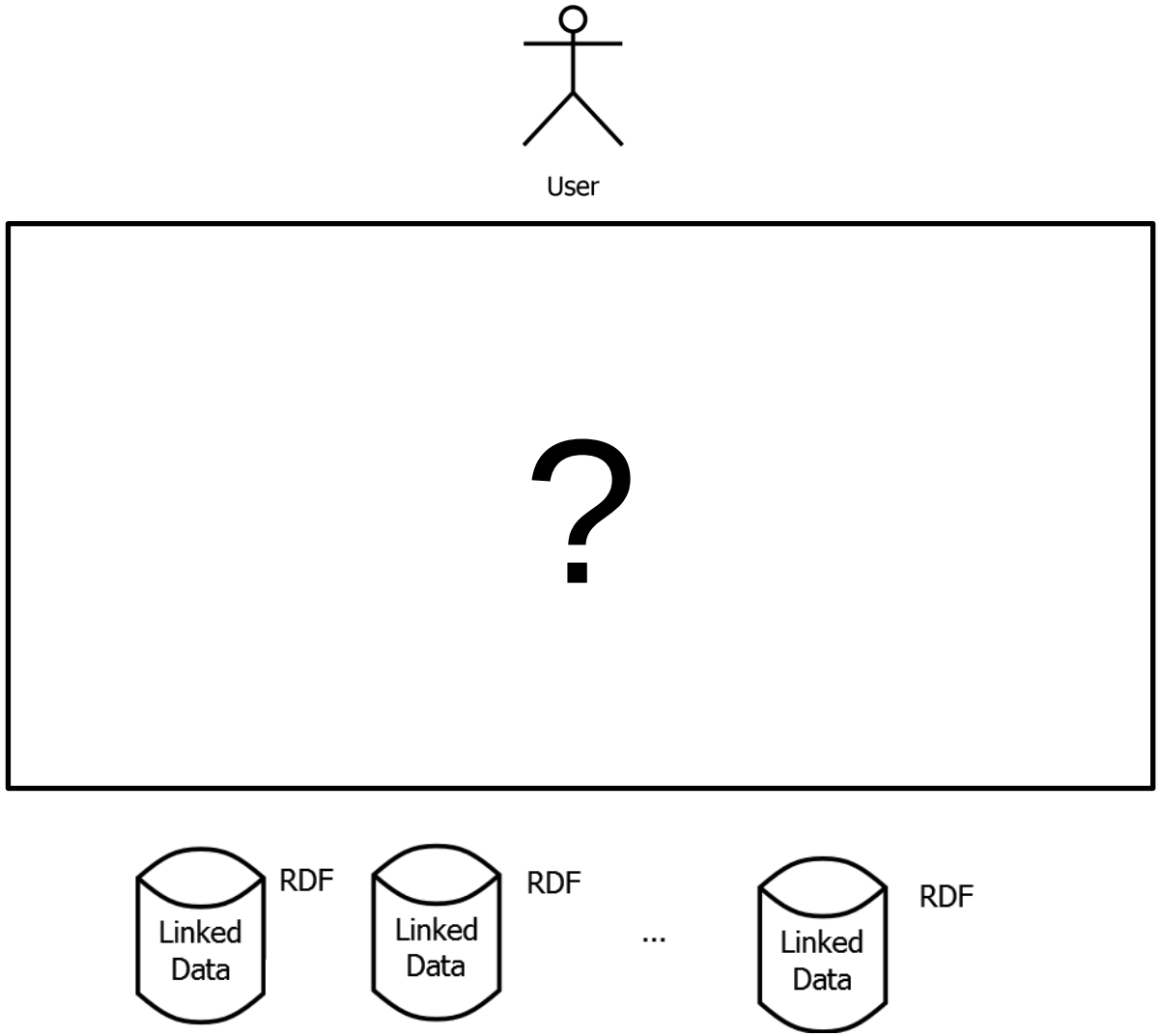
http://www.w3.org/DesignIssues/LinkedData.html

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



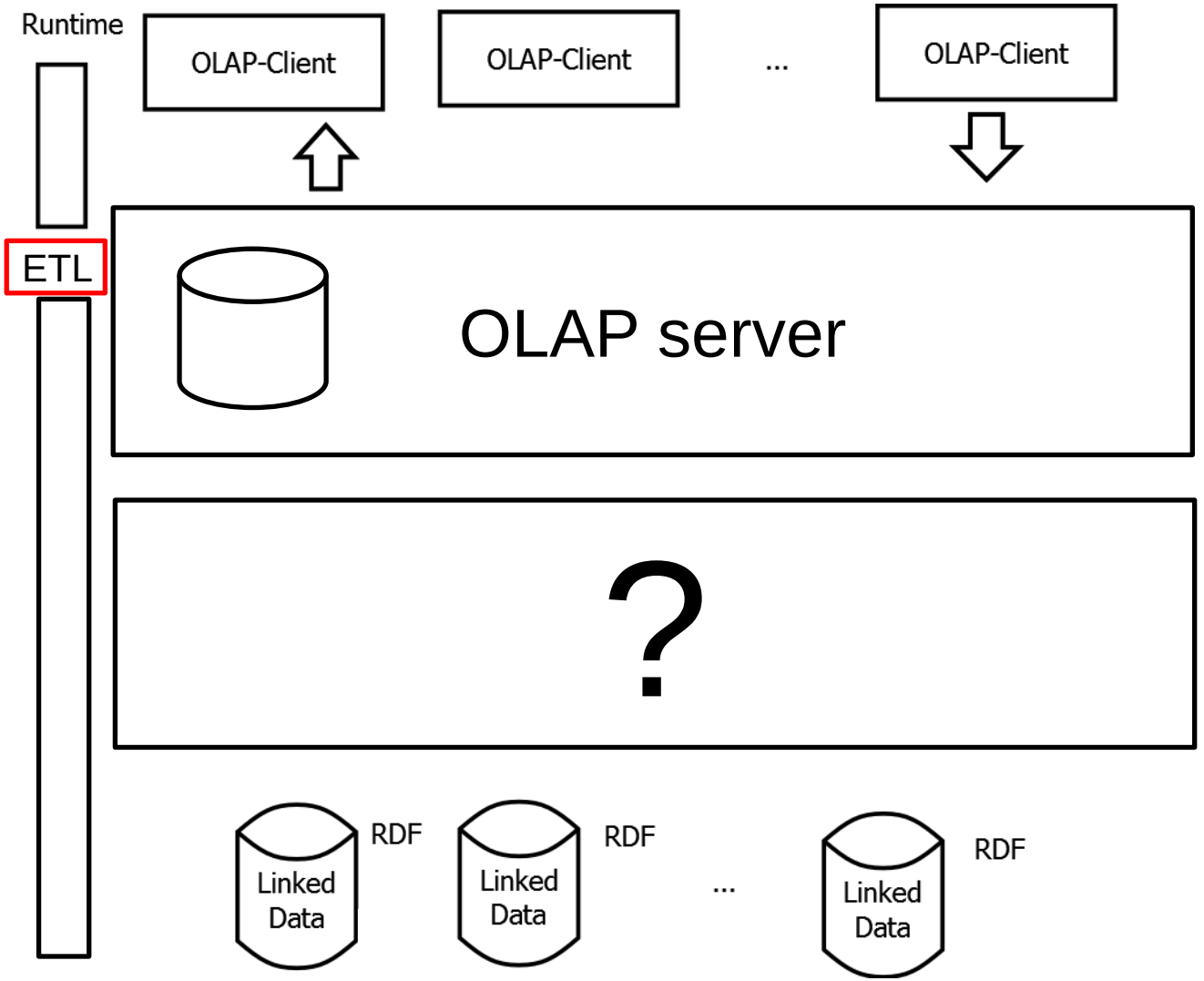
Approach - Overview



OLAP Systems

- 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)

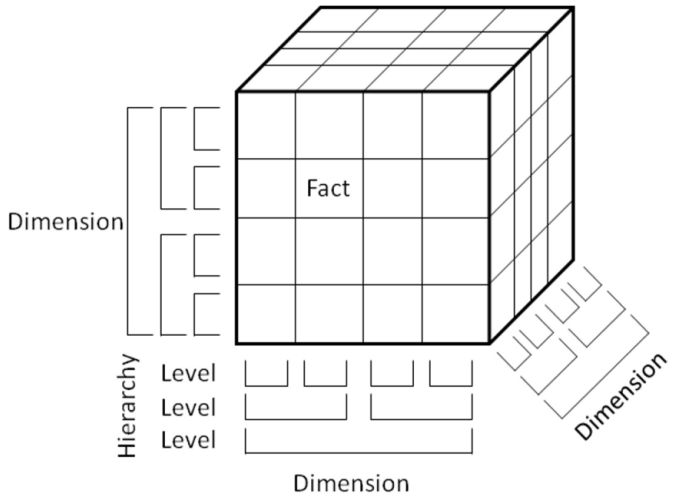


Multidimensional Model and SLD reusing ontology

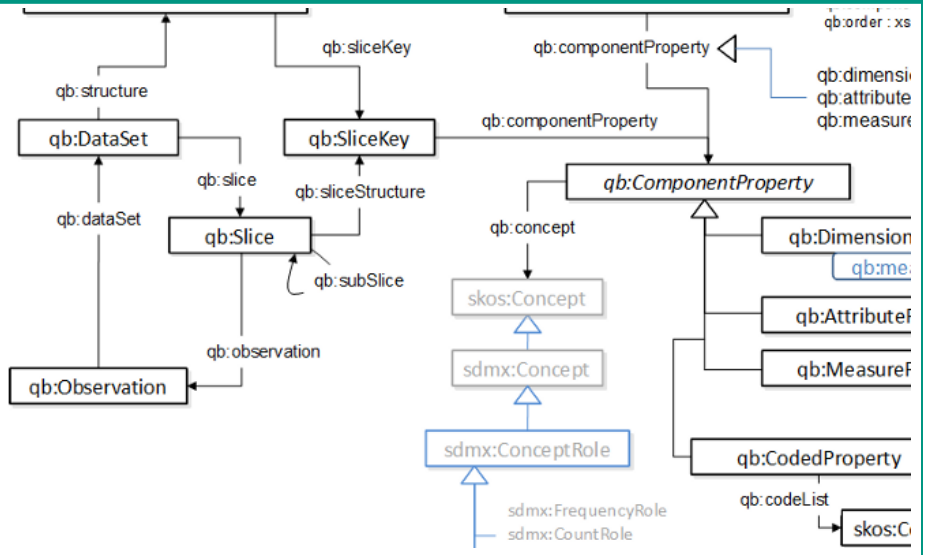
Requirements: Automatically and scalable prepare SLD to answer complex questions using OLAP systems

Approach: Mapping Multidimensional Model and SLD reusing ontology

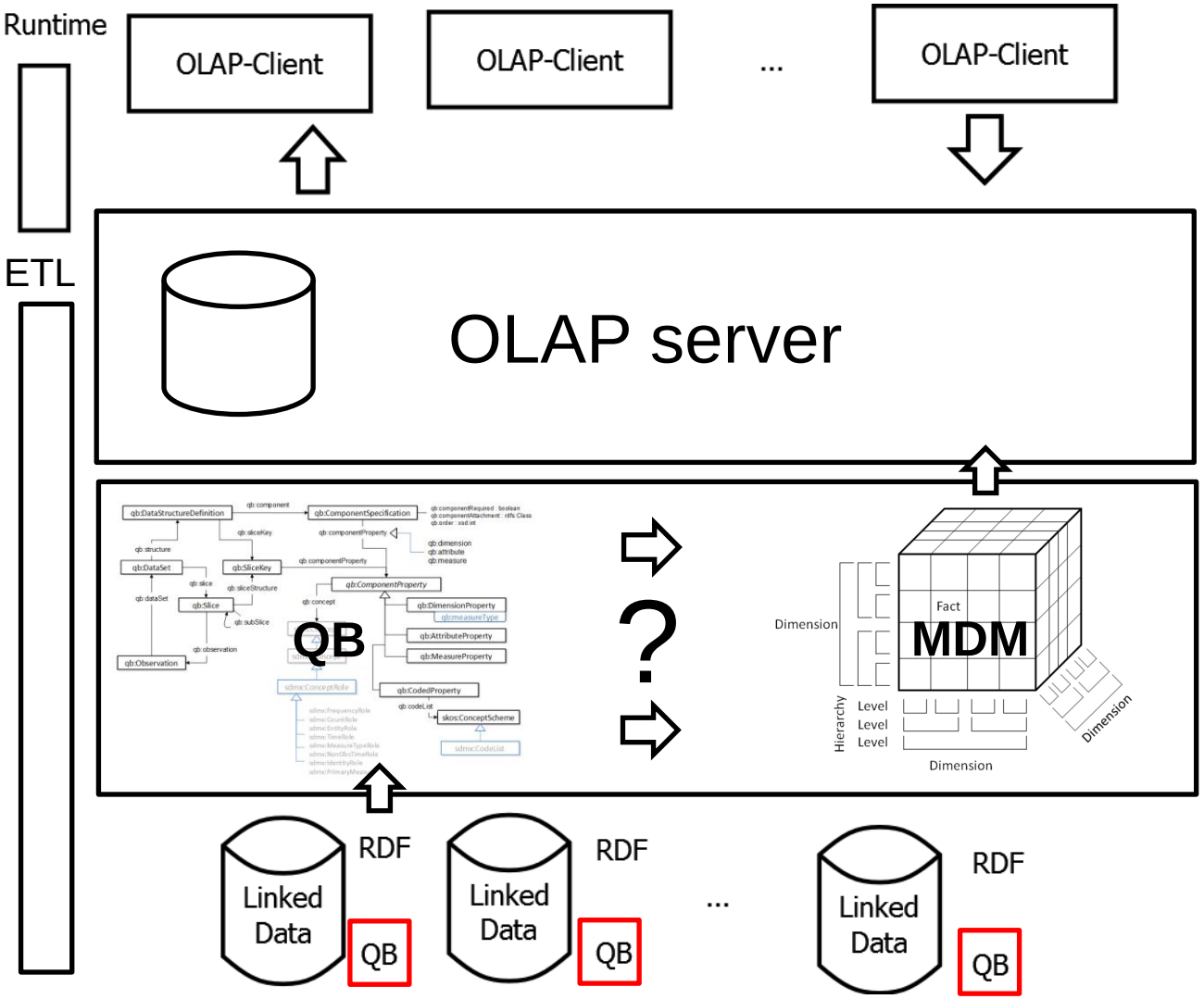
Multidimensional Model (MDM)



RDF Data Cube vocabulary (QB)



Approach – Overview (3)



Multidimensional Model (MDM)

- Advantages
  - No standard but common MDM
  - Expressive enough

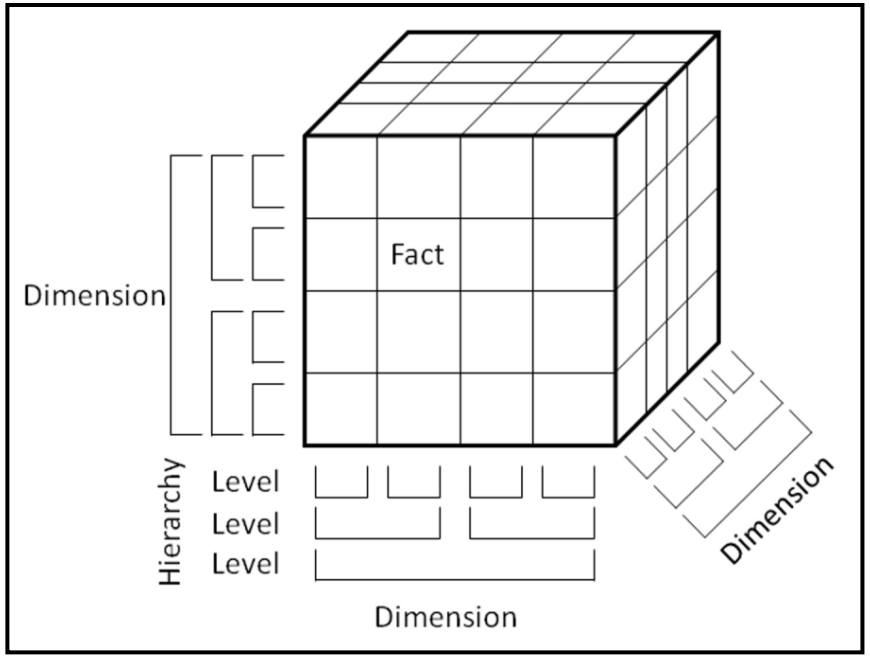
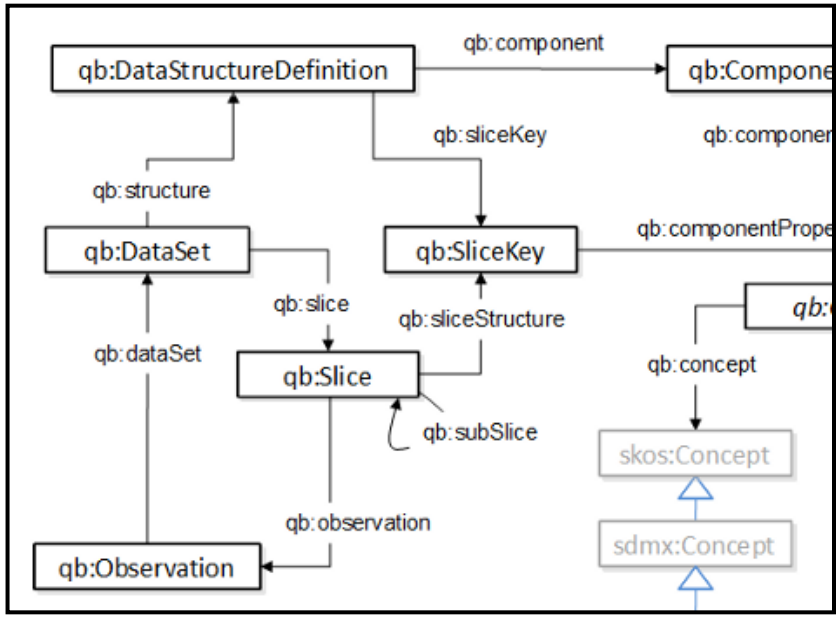


Illustration of common MDM

- 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
- …

RDF Data Cube vocabulary (QB)

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



Cyganiak et al. - The RDF Data Cube vocabulary

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

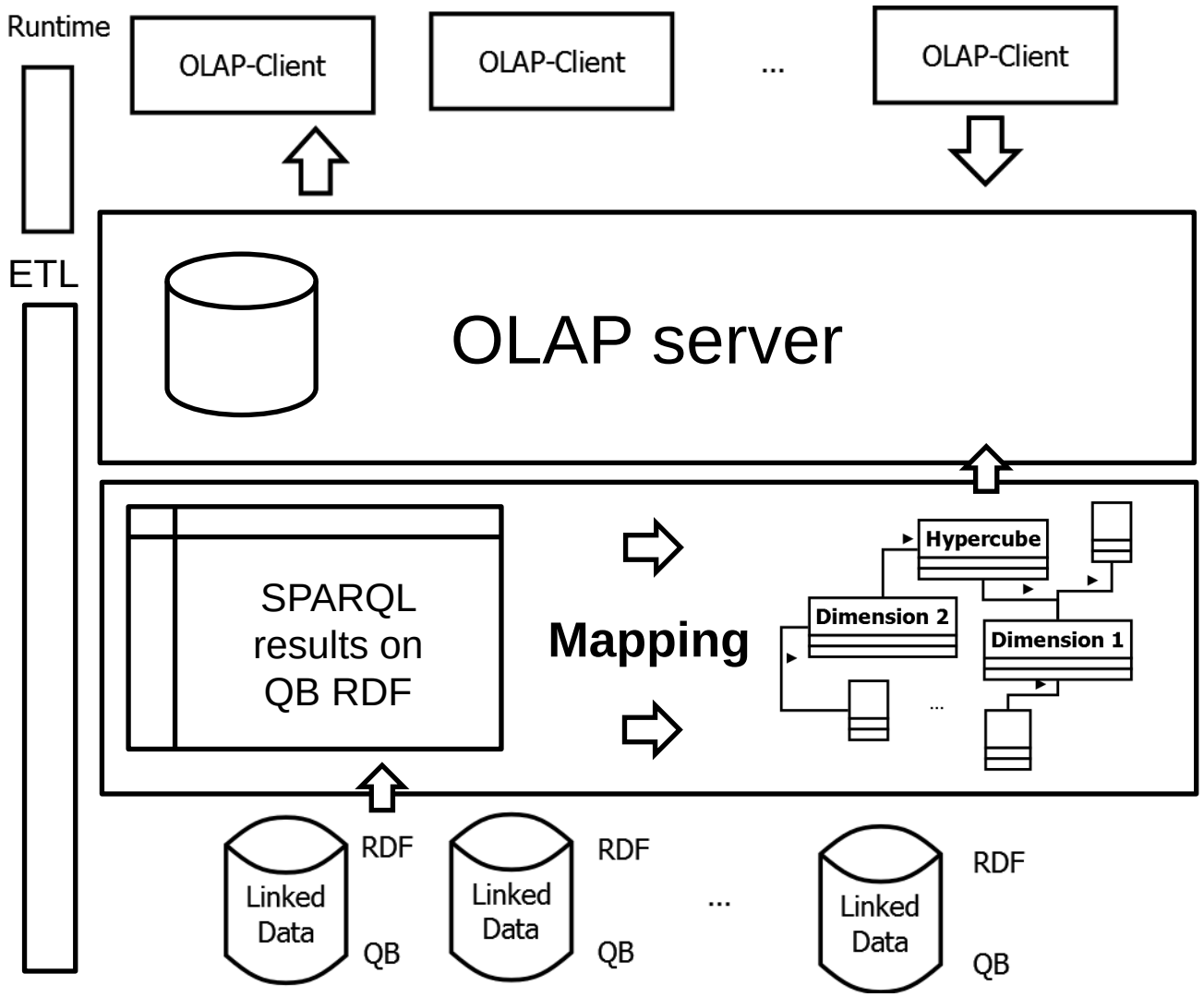
Mapping – Multidimensional Model and RDF Data Cube vocabulary

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

Mapping terms of common MDM to SPARQL queries on RDF using QB

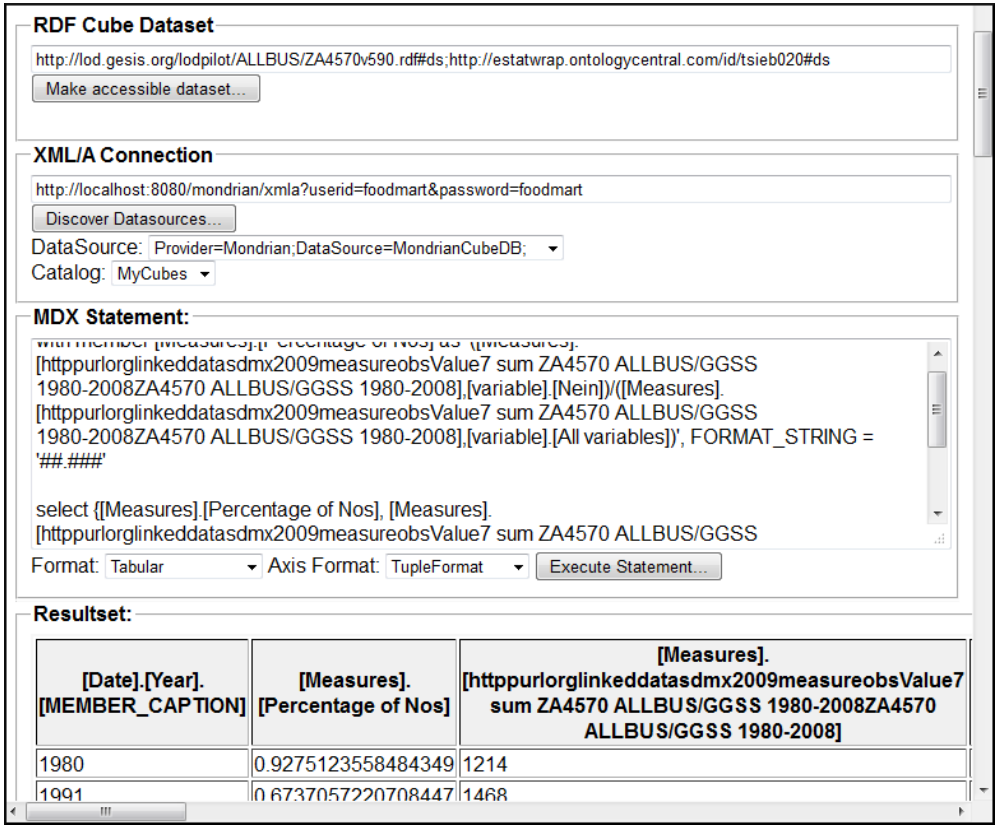
Prefixes, see http://prefix.cc/

Approach – Overview (4)



Experiments – Implementation

- OLAP client: xmla4js
- OLAP server: Mondrian/XMLA
- ETL pipeline: PHP web service
- SPARQL engine: qcrumb.com



Application Interface

Institute of Applied Informatics and Formal Description Methods (AIFB)

Experiments – GDP growth influencing unemployment fear?

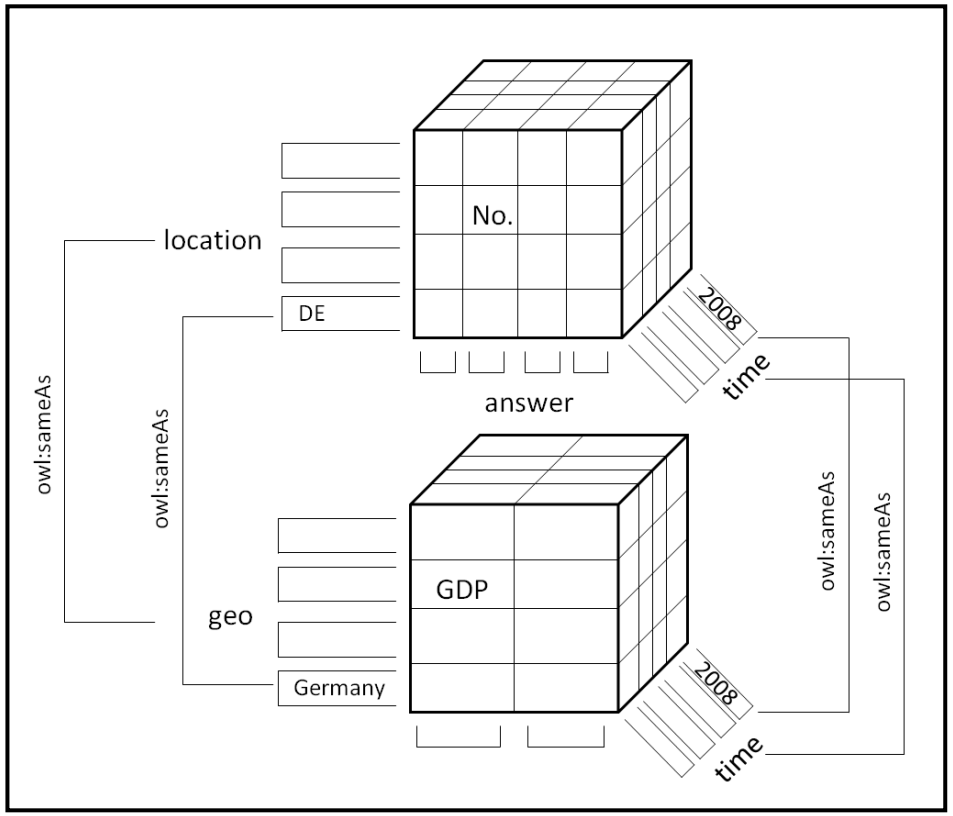
- Datasets
  - Survey data about German employees' fear of unemployment in the last few years (http://lod.gesis.org/lodpilot/ALLBUS/ZA4570v590.rdf#ds)
  - GDP growth of European countries per year as provided by Eurostat (http://estatwrap.ontologycentral.com/id/tsieb020#ds)
- Wanted result

Filter: Germany

Year\Metric	GDP growth	Percentage of people without fear of unemployment
2011	X	X
2010	X	X
...	X	X

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)



MDM of Unemployment fear and GDP growth

Experiments – Trends in Eurostat EU 2020 indicators?

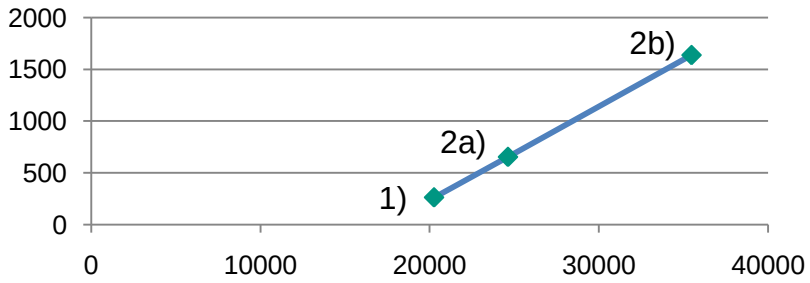
- Datasets
  - Employment rate (http://estatwrap.ontologycentral.com/id/t2020_10#ds)
  - Greenhouse gas emissions (http://estatwrap.ontologycentral.com/id/t2020_30#ds)
  - ...
- Wanted result

Country\Metric	Employment rate	Greenhouse gas emissions	...
Austria	X	X	X
Belgium	X	X	X
...	X	X	X

Experiments – Evaluation

Experiment	Datasets	Triples	ETL pipeline
1) Unemployment fear and GDP growth	2	20 268	4m 23s
2a) EU 2020 indicators	4	24 636	10m 54s
2b) EU 2020 indicators	8	35 482	27m 18s

- 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)

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

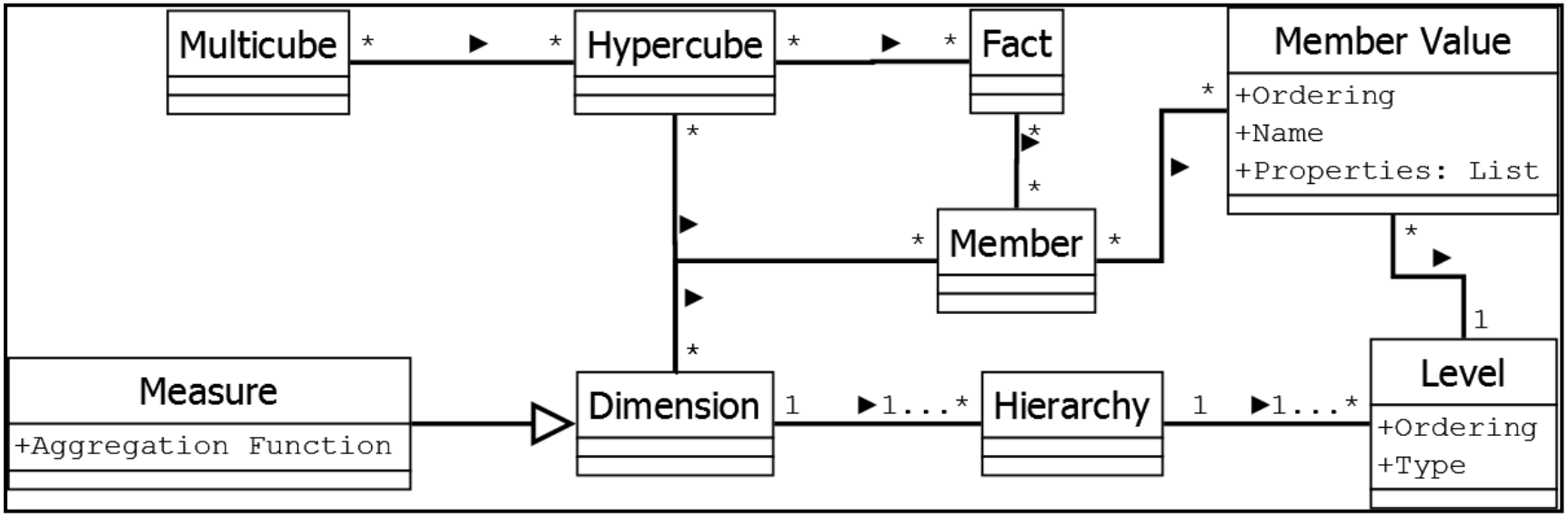
- OLAP-like operations on Web sources **with** Semantic Web technologies
  - E.g., Marko Niinimäki and Tapio Niemi – An ETL Process for OLAP Using RDF/OWL Ontologies, 2009
  - **Problem:** No focus on the problem of semantic heterogeneity in datasources to be integrated in an MDM. No Linked Data

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

[1] http://wiki.planet-data.eu/web/Datasets
[2] Kämpgen, DC Proposal: Online Analytical Processing of Statistical Linked Data. ISWC 2011

Backup: Class Diagram



Class diagram of MDM

Backup: Architecture