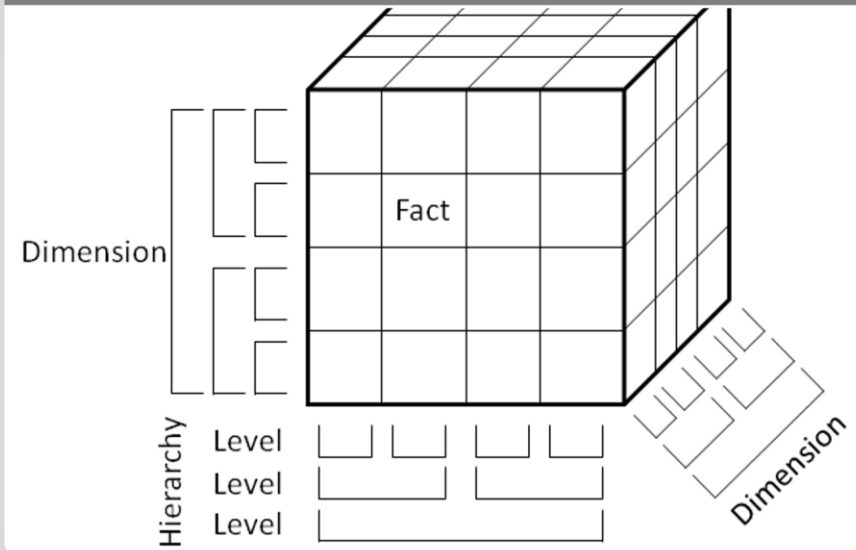
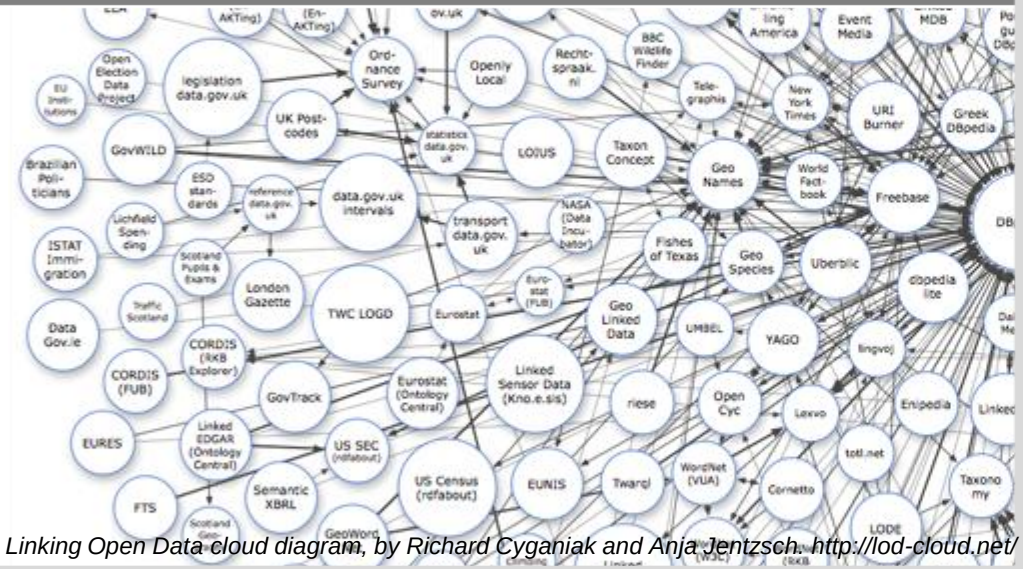
DC Proposal: Online Analytical Processing of Statistical Linked Data

Benedikt Kämpgen

The 10th International Semantic Web Conference (ISWC2011)
24 October 2011

Institute of Applied Informatics and Formal Description Methods (AIFB)





Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Outline

1. Problem: General decision support using Statistical Linked Data (SLD)
2. Approach: Online Analytical Processing (OLAP)
3. Research plan: challenges and evaluation

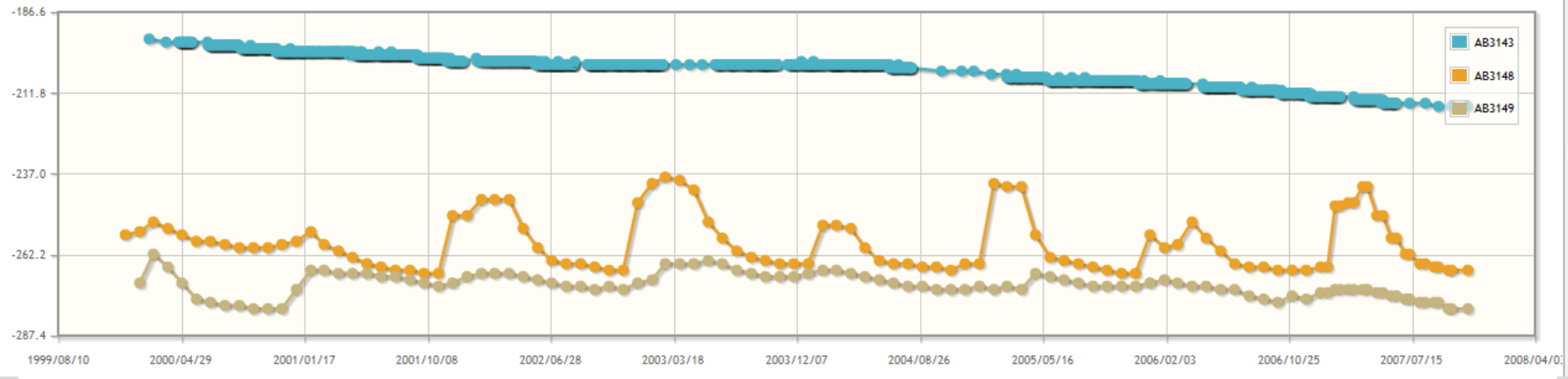
Institute of Applied Informatics and Formal Description Methods (AIFB)

Problem: Decision support using statistical Linked Data (SLD)

page | discussion | edit | history | delete | move | protect | watch | refresh

Level above msl at AB3149, AB3148, AB3143

Datechart [edit]



- Example: Three datasets, time series, visualization
  - Difficult to adapt to other “business questions”
  - E.g., average, per year, integrating other datasources
- Problem: General decision support
  - Approach: Online Analytical Processing (OLAP)

OLAP frontends intuitive, interactive, explorative, fast

- Common front-end: Pivot tables
- Understandable by non-technicians

- Different angles
- Aggregations
- Drill-down / Roll-up
- Filters
- Drill-through
- What-if-analyses

Aggregate function: MEAN
Subtotals aggregate function: COUNT

Filter list

Link: [all] X
StudyArea: [all] X
AnalysisObject: [all] X
Date: [all] X
Day: [all] X
Unit: [all] X
Month: [all] X

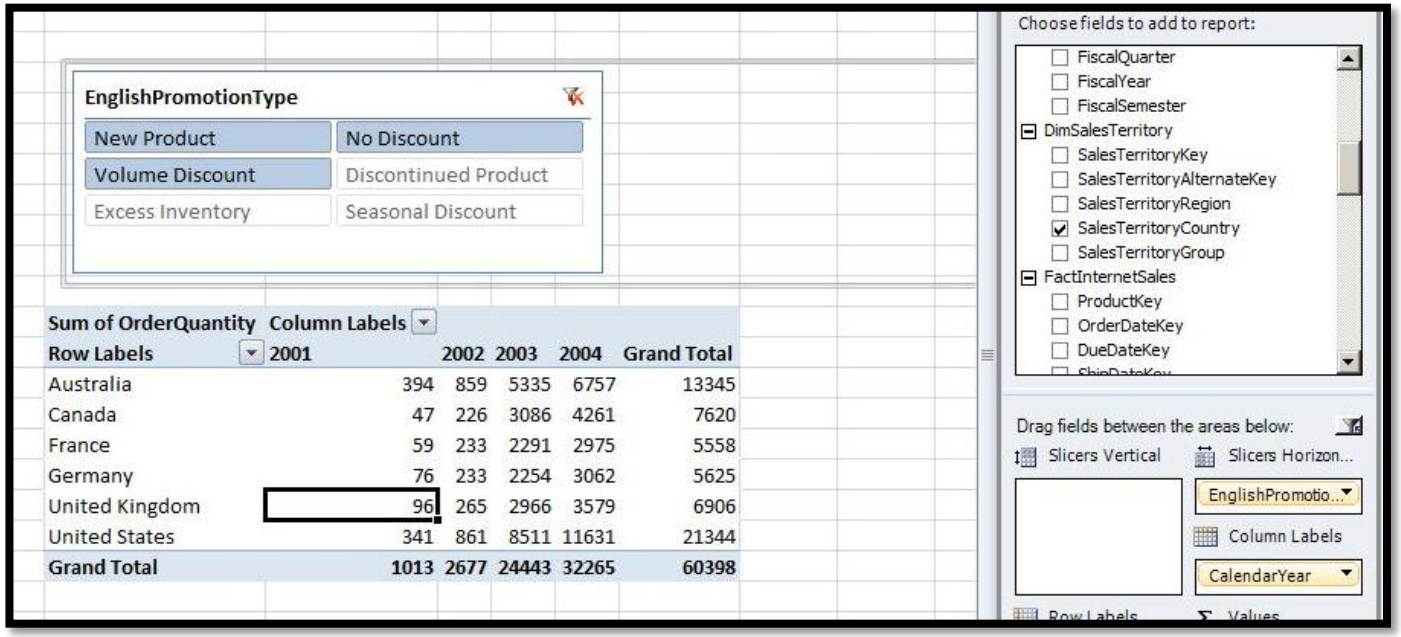
Value	Year	2000	2001	2002	2003	2004	2005	2006	2007	Year
LocationName										
AB3143		-198.00	-200.53	-203.14	-203.55	-203.74	-208.24	-211.04	-214.46	
AB3148		-257.74	-262.70	-258.06	-253.10	-262.27	-259.05	-262.56	-257.64	
AB3149		-274.62	-269.89	-270.98	-267.73	-270.77	-271.80	-273.89	-275.75	
LocationName										

Pivot table

Institute of Applied Informatics and Formal Description Methods (AIFB)

OLAP functionality provided by many tools and may be reused

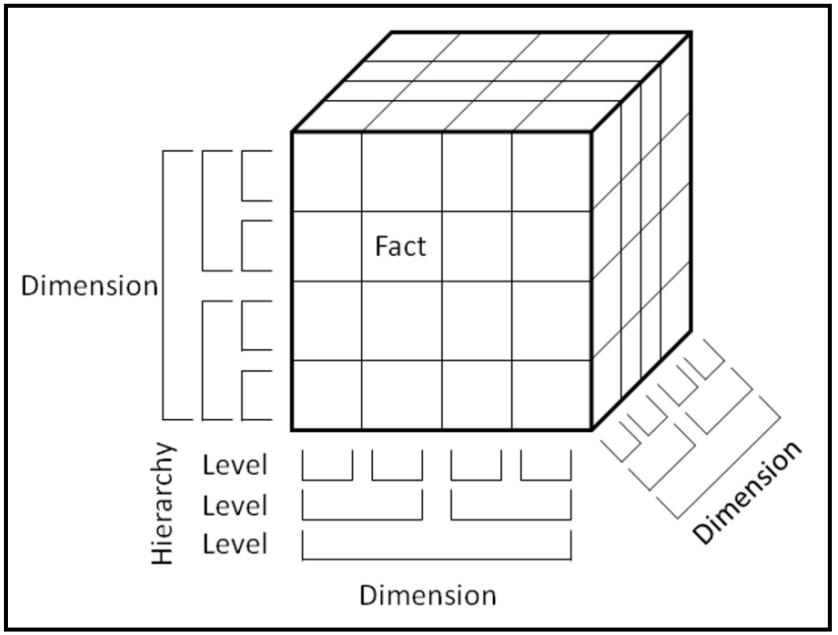
- Client libraries for Java (olap4j), JavaScript (olap4js).
- Microsoft provides OLAP functionality in Excel: PowerPivot



http://www.powerpivotblog.nl/working-with-gemini-and-excel-2010-to-make-a-pivot-table

OLAP features fulfill Statistical Linked Data requirements

Statistical Linked Data	OLAP (ETL / Data Warehousing)
Many, large, and distributed datasources	ETL
Heterogeneous data with variable quality	Multidimensional Model (OLAP model)
Data not available permanently	Data Warehouse
Dynamic changes	Near real-time updates
Confidential information	Access-control

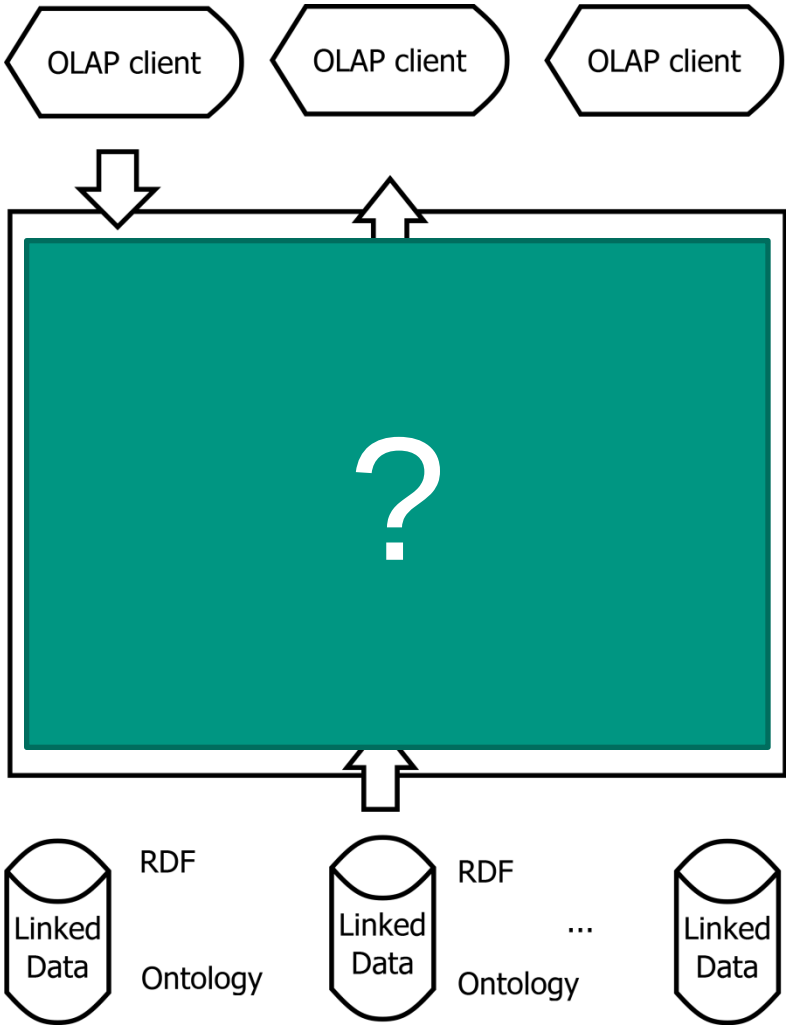


OLAP model

Institute of Applied Informatics and Formal Description Methods (AIFB)

Research question

- **Research question:** How to allow OLAP on Statistical Linked Data?
- **Hypothesis:** Statistical Linked Data self-descriptive enough to be automatically preprocessed for common OLAP systems



Related work

- OLAP-like operations on Web data **without** using Semantic Web technologies
  - Exchange formats often based on XML (e.g., XBRL, SDMX, and DDI)
  - E.g., Google Public Data Explorer
  - **Problem:** Creating XML from datasources manual effort.

- OLAP-like operations on Web data **with** Semantic Web technologies
  - Based on ontologies
  - E.g., Marko Niinimäki and Tapio Niemi – An ETL Process for OLAP Using RDF/OWL Ontologies, 2009
  - **Problem:** Ontology engineering from datasources manual effort.

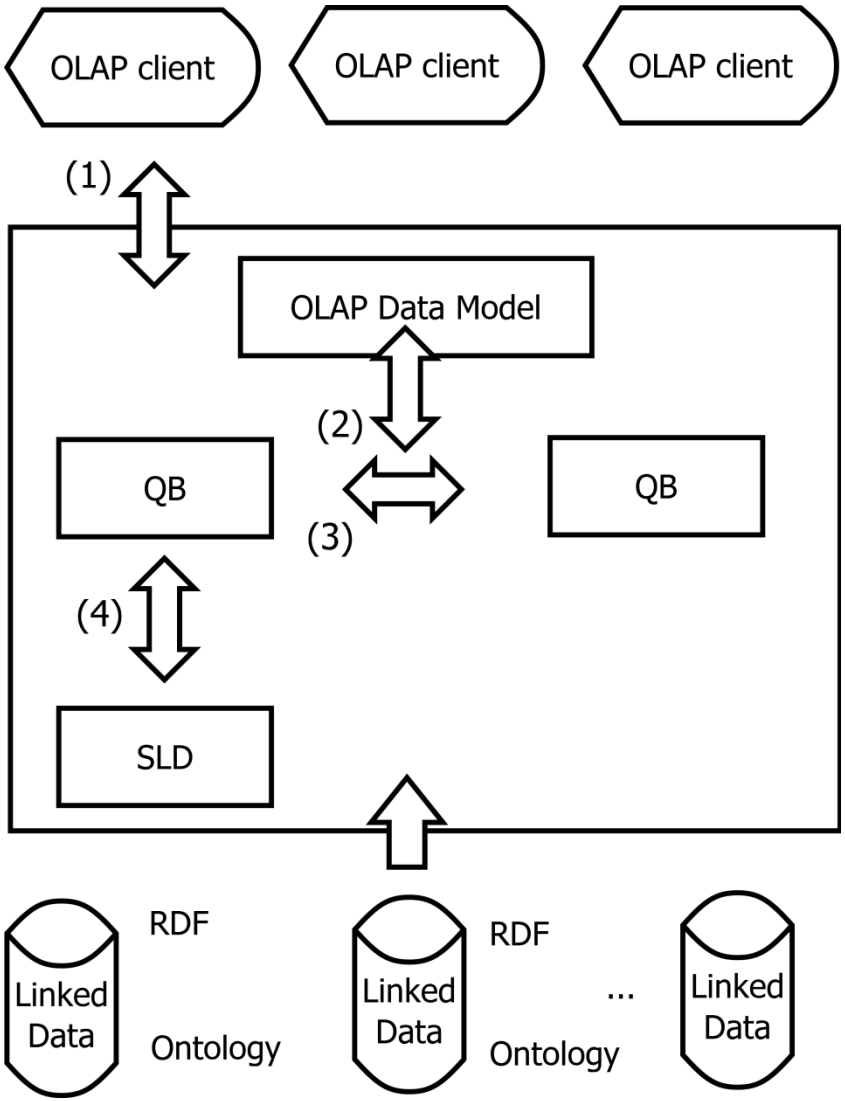
Research plan

(1) Issuing OLAP queries on statistical Linked Data

(2) Transforming statistical Linked Data into OLAP model

(3) Integrating statistical Linked Data for transformation into OLAP model

(4) Matching statistical Linked Data



Institute of Applied Informatics and Formal Description Methods (AIFB)

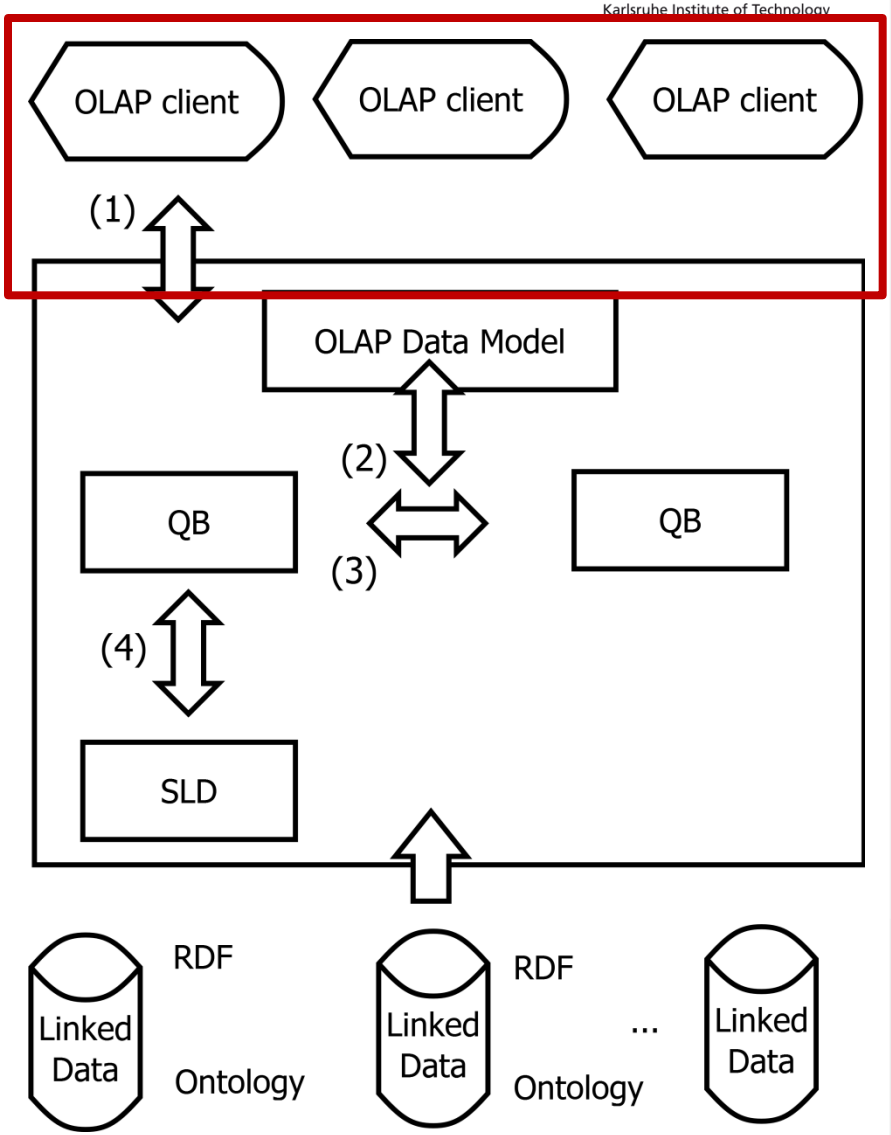
Research plan

(1) Issuing OLAP queries on statistical Linked Data

(2) Transforming statistical Linked Data into OLAP model

(3) Integrating statistical Linked Data for transformation into OLAP model

(4) Matching statistical Linked Data



Institute of Applied Informatics and Formal Description Methods (AIFB)

(1) Issuing OLAP queries on statistical Linked Data

The top 10 most discussed product types of products from a specific country based on number of reviews by reviewers from a certain country.

```
SELECT {[Measures].[Count rev:Review]} ON
COLUMNS,

TopCount({[bsbm:reviewFor].[bsbm:ProductTy
pe].members}
   , 10
   , [Measures].[Count rev:Review]
   ) ON ROWS
FROM [rev:Review]
WHERE
[bsbm:reviewFor].[bsbm:producer].[bsbm:coun
try].[%Country1%] AND
[rev:reviewer].[bsbm:country].[%Country2%]
```

OLAP query (MDX)

```
prefix bsbm: <http://www4.wiwiss.fu-
berlin.de/bizer/bsbm/v01/vocabulary/>
prefix rev: <http://purl.org/stuff/rev#>

Select ?productType ?reviewCount
{
 { Select ?productType (count(?review) As
?reviewCount)
  {
   ?productType a bsbm:ProductType .
   ?product a ?productType .
   ?product bsbm:producer ?producer .
   ?producer bsbm:country %Country1% .
   ?review bsbm:reviewFor ?product .
   ?review rev:reviewer ?reviewer .
   ?reviewer bsbm:country %Country2% .
  }
  Group By ?productType
 }
}
Order By desc(?reviewCount) ?productType
Limit 10
```

SPARQL 1.1

(1) Issuing OLAP queries on statistical Linked Data

The top 10 most discussed product types of products from a specific country based on number of reviews by reviewers from a certain country.

```
SELECT {[Measures].[Count rev:Review]} ON
COLUMNS,

TopCount({[bsbm:reviewFor].[bsbm:ProductTy
pe].members}
   , 10
   , [Measures].[Count rev:Review]
   ) ON ROWS
FROM [rev:Review]
WHERE
[bsbm:reviewFor].[bsbm:producer].[bsbm:coun
try].[%Country1%] AND
[rev:reviewer].[bsbm:country].[%Country2%]
```

OLAP query (MDX)

```
prefix bsbm: <http://www4.wiwiss.fu-
berlin.de/bizer/bsbm/v01/vocabulary/>
prefix rev: <http://purl.org/stuff/rev#>

Select ?productType ?reviewCount
{
 { Select ?productType (count(?review) As
?reviewCount)
  {
   ?productType a bsbm:ProductType .
   ?product a ?productType .
   ?product bsbm:producer ?producer .
   ?producer bsbm:country %Country1% .
   ?review bsbm:reviewFor ?product .
   ?review rev:reviewer ?reviewer .
   ?reviewer bsbm:country %Country2% .
  }
  Group By ?productType
 }
}
Order By desc(?reviewCount) ?productType
Limit 10
```

SPARQL 1.1

(1) Challenge: Optimise Linked Data queries for OLAP

- **Evaluation:** Query performance; Berlin SPARQL benchmark

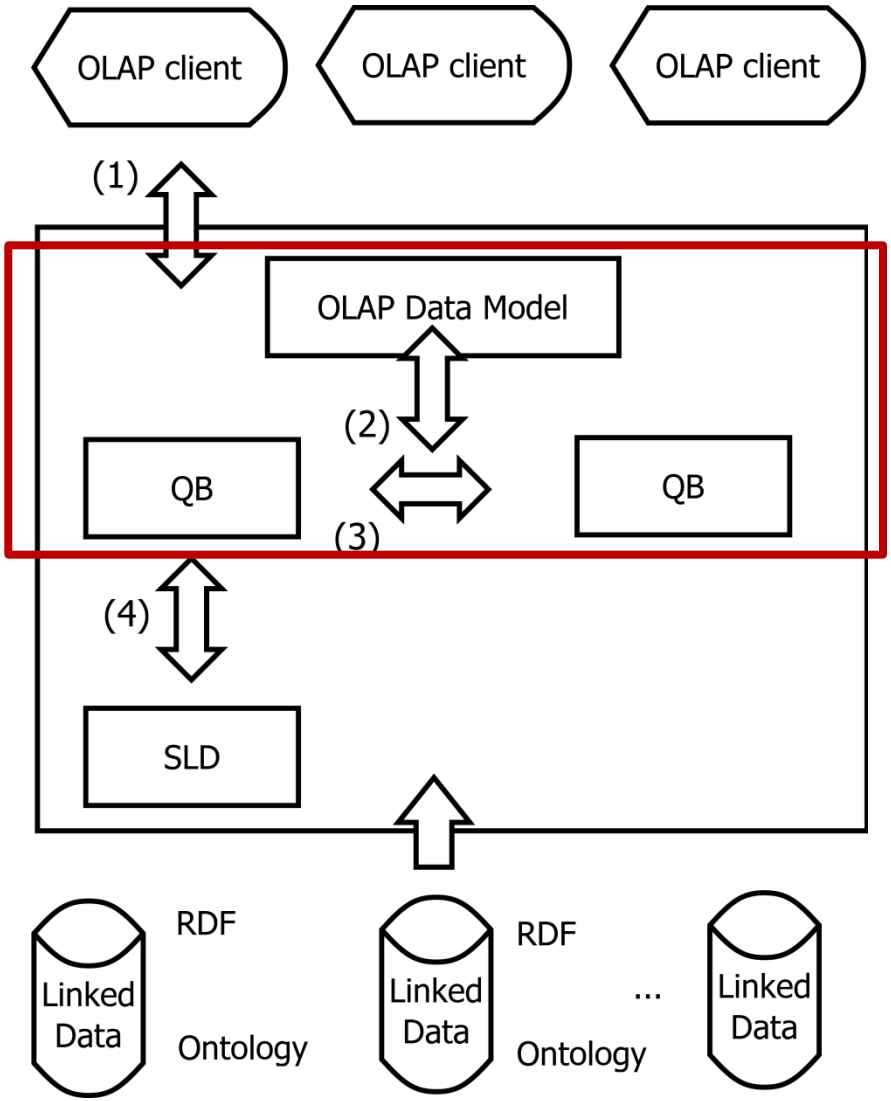
Research plan

(1) Issuing OLAP queries on statistical Linked Data

(2) Transforming statistical Linked Data into OLAP model

(3) Integrating statistical Linked Data for transformation into OLAP model

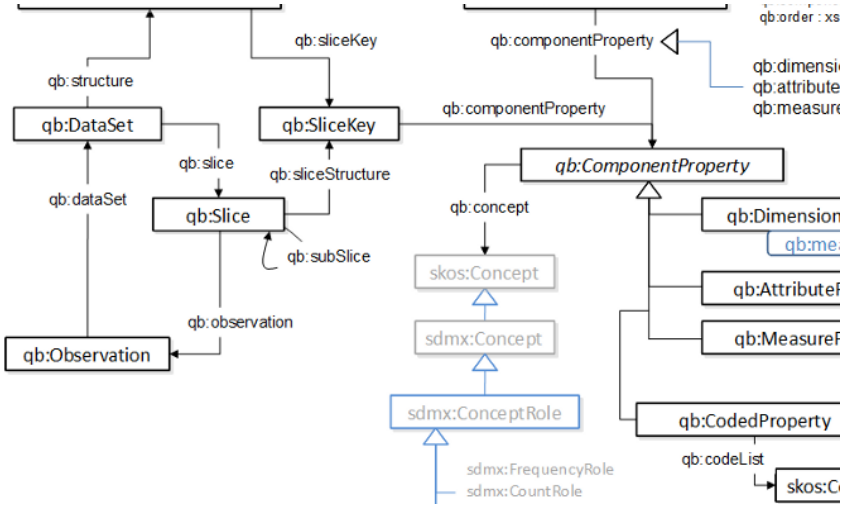
(4) Matching statistical Linked Data



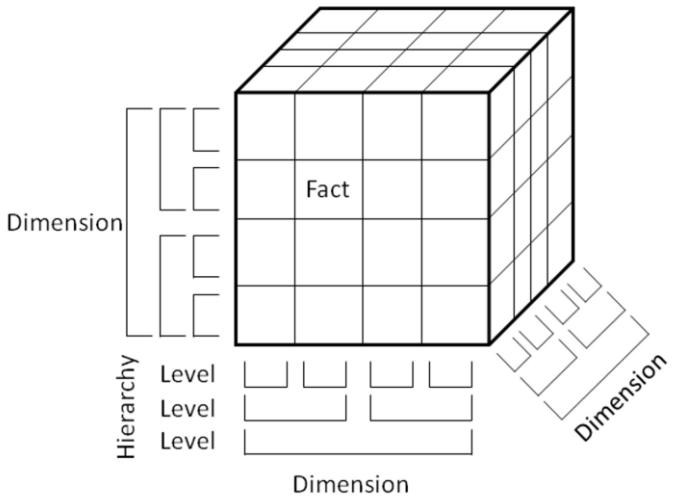
Institute of Applied Informatics and Formal Description Methods (AIFB)

(2) Transforming statistical Linked Data into conceptual model [1]

RDF Data Cube vocabulary (QB)



OLAP model

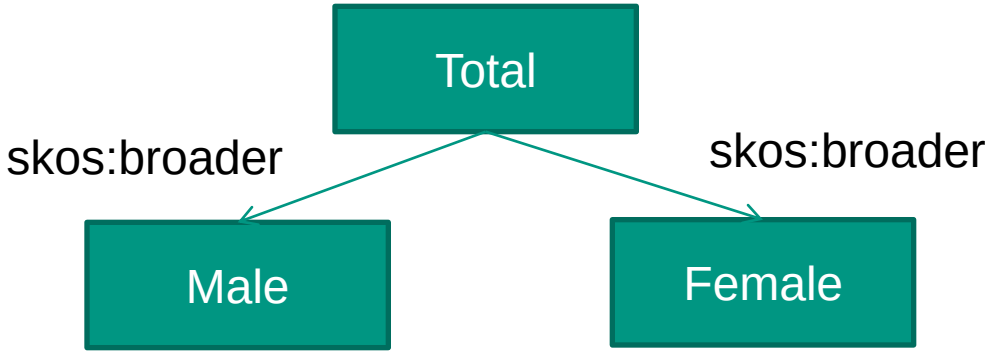


qb:DataSet + qb:DataStructureDefinition	Data Hypercube (Cube)
qb:Observation + qb:DataSet	Fact
qb:ComponentProperty	Dimension
...	

[1] Kämpgen, Harth – Transforming Statistical Linked Data for Use in OLAP Systems, ISEM 2011

(2) Challenge: Construct OLAP hierarchies from statistical Linked Data

- For instance, data with Dimension “gender”



- Result
  - Two Cubes with Dimension gender with 2-Level-Hierarchy
  - Cube 1: Total data; Cube 2: Male/Female data
- **Evaluation:** Total = Male + Female

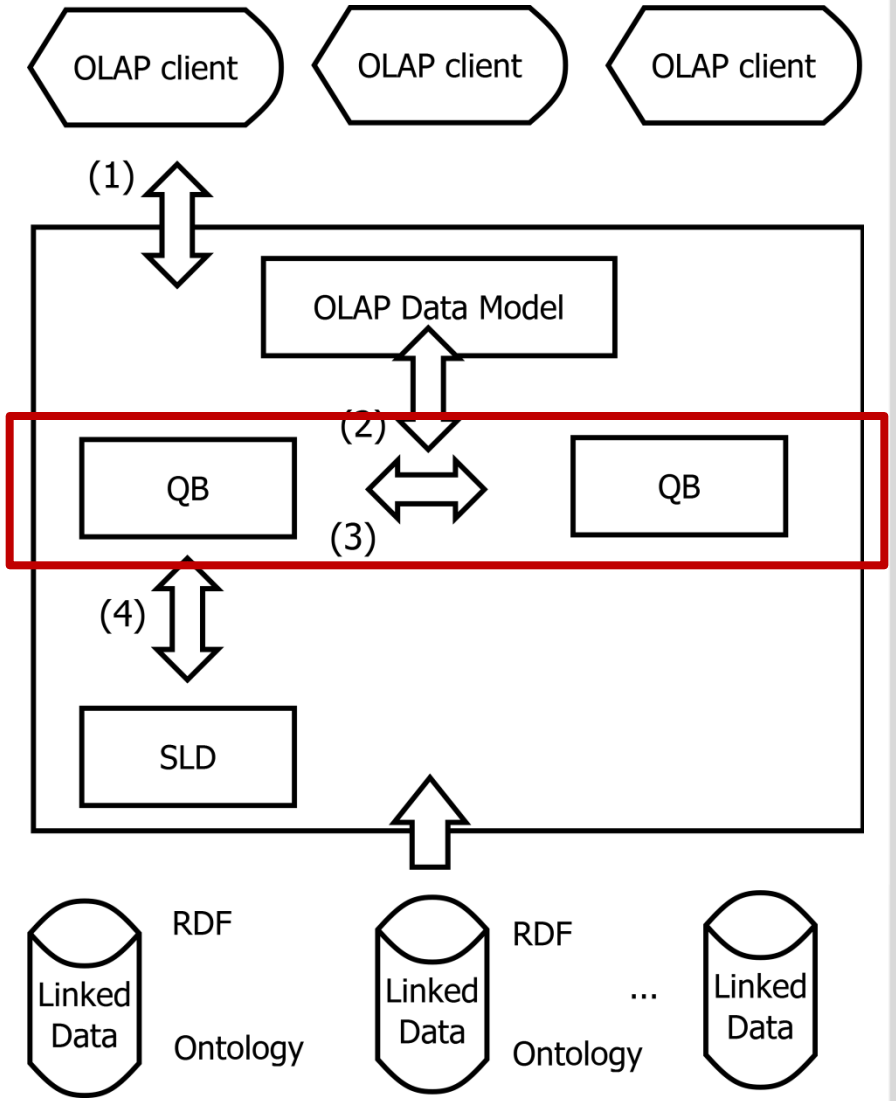
Research plan

(1) Issuing OLAP queries on statistical Linked Data

(2) Transforming statistical Linked Data into OLAP model

(3) Integrating statistical Linked Data for transformation into OLAP model

(4) Matching statistical Linked Data



Institute of Applied Informatics and Formal Description Methods (AIFB)

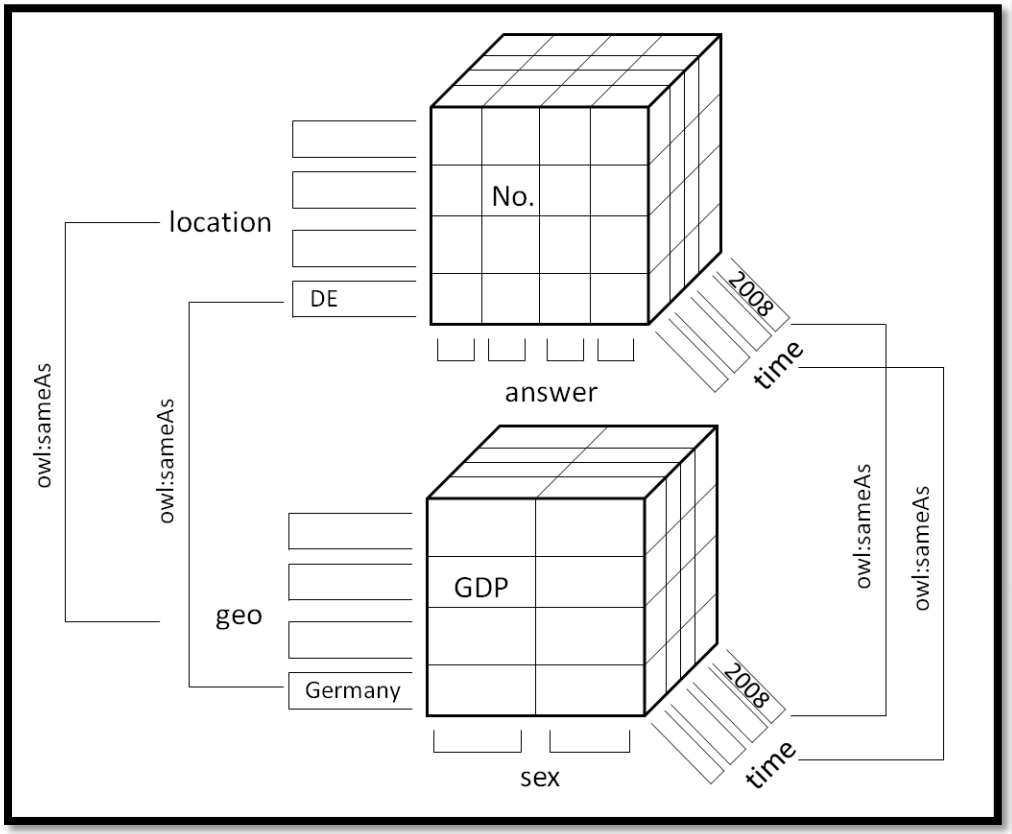
(3) Integrating statistical Linked Data for transformation into OLAP data model

- Statistics may heterogeneously describe related information, e.g.,
  - Dimensions: “Geo” and “Location”
  - Members: “DE” and “Germany”

(3) Challenge: Extract relations between elements in different statistics

- Relating “Geo” and “Location” / “DE” and “Germany”
- One approach: Exploit owl:sameAs links between instances of Dimensions and Members

- **Evaluation:** Possible to compare related statistics



Institute of Applied Informatics and Formal Description Methods (AIFB)

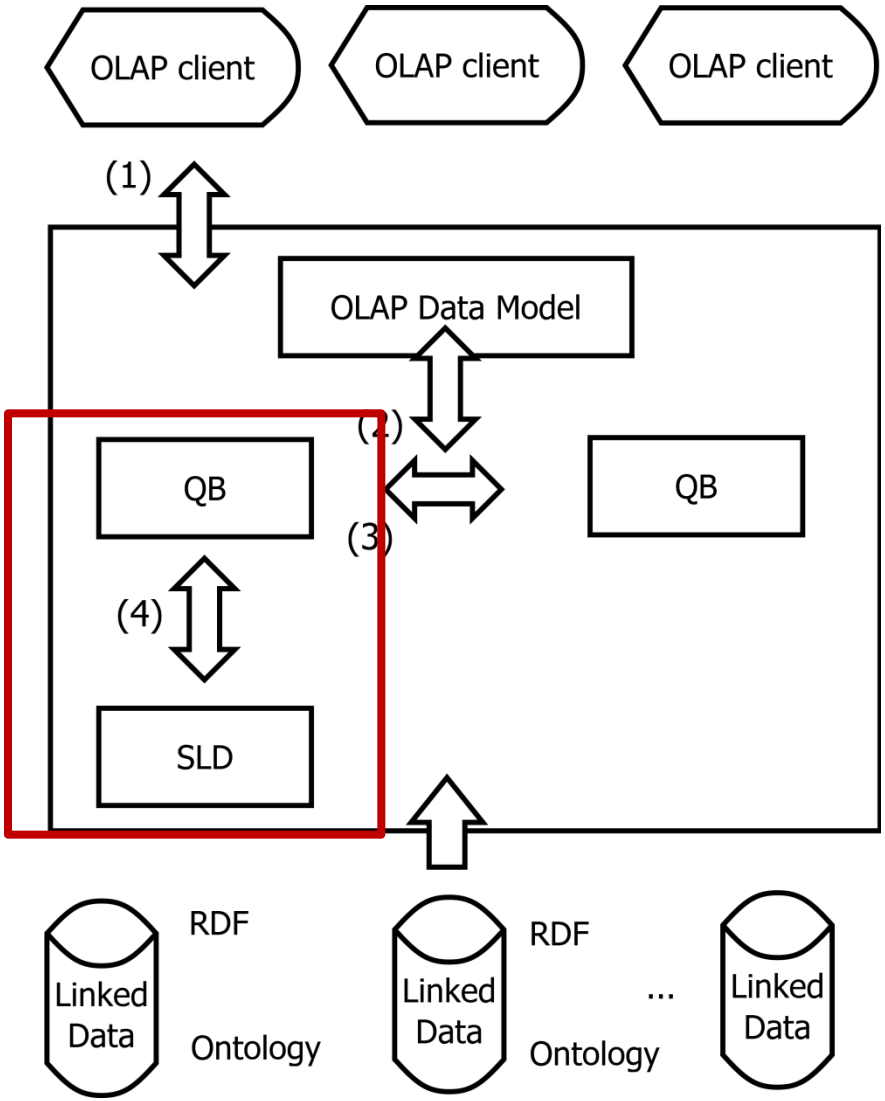
Research plan

(1) Issuing OLAP queries on statistical Linked Data

(2) Transforming statistical Linked Data into OLAP model

(3) Integrating statistical Linked Data for transformation into OLAP model

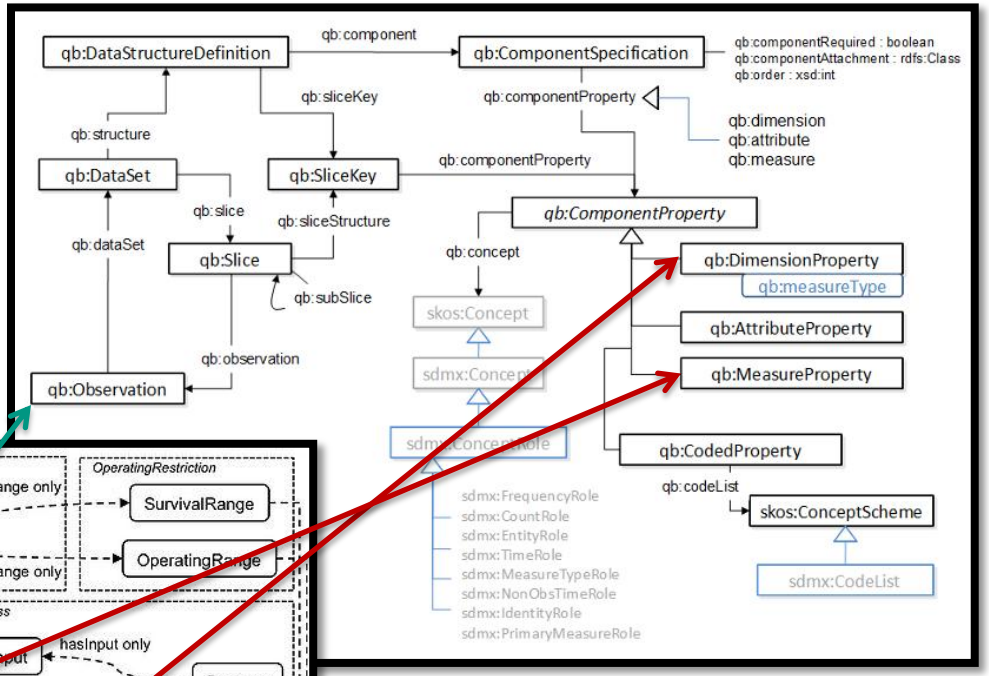
(4) Matching statistical Linked Data

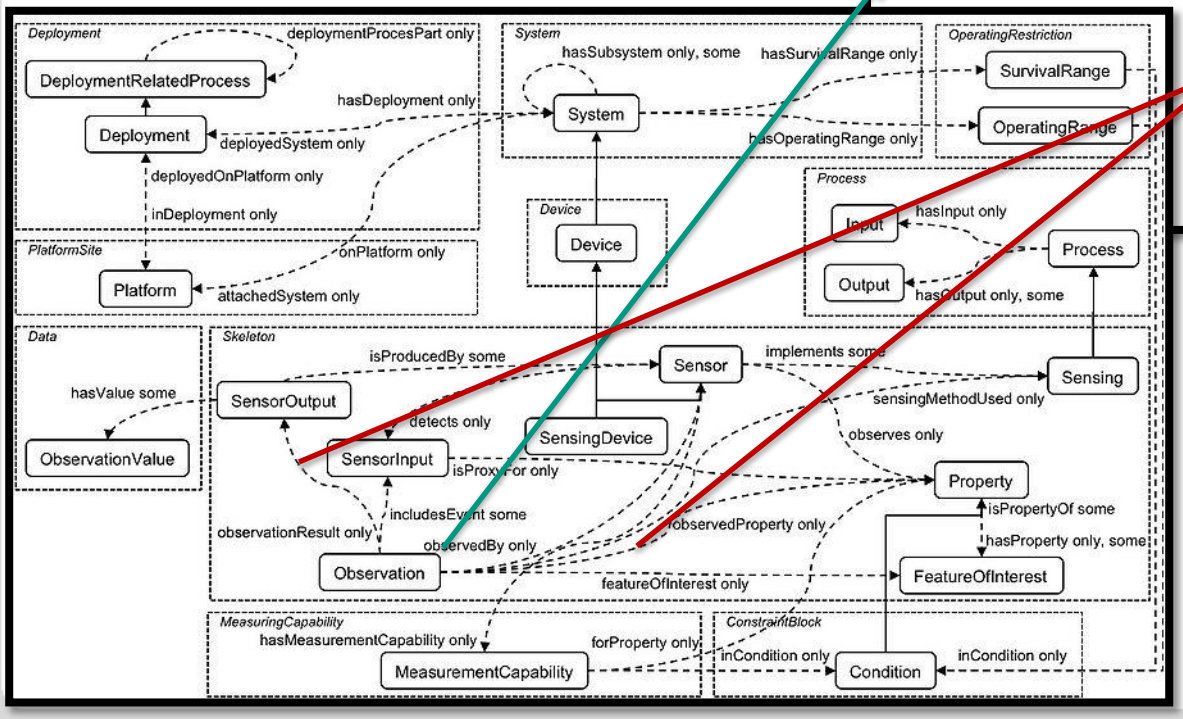


Institute of Applied Informatics and Formal Description Methods (AIFB)

(4) Matching Statistical Linked Data

- Not all Statistical Linked Data is reusing QB
- Possible to match manually





RDF Data Cube vocabulary

rdfs:subClassOf

rdfs:subPropertyOf

Semantic Sensor Network ontology

(4) Challenge: Model statistics automatically

- Find interesting Cubes, Dimensions, Measures…
- In data using various vocabularies
- **Evaluation:** Quality criteria for OLAP models, e.g., analyzability and simplicity

Institute of Applied Informatics and Formal Description Methods (AIFB)

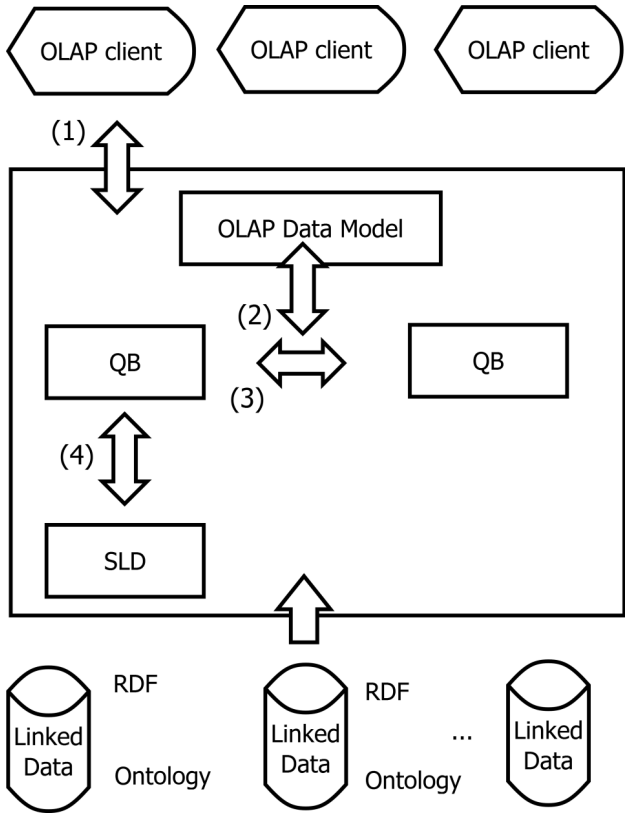
Conclusion

- OLAP for generic decision support using statistical Linked Data

- Status - research plan
  - Preliminary work: B. Kämpgen, A. Harth – Transforming Statistical Linked Data for Use in OLAP Systems, ISEM 2011
  - Future work: Fulfilling the requirements of our use cases, e.g., hierarchies
  - Ongoing work: Java Linked Data driver for various OLAP clients

Q&A

- Questions?
- Feedback?
  - Use cases...
  - Research challenges
  - Evaluation



Thanks!

Institute of Applied Informatics and Formal Description Methods (AIFB)