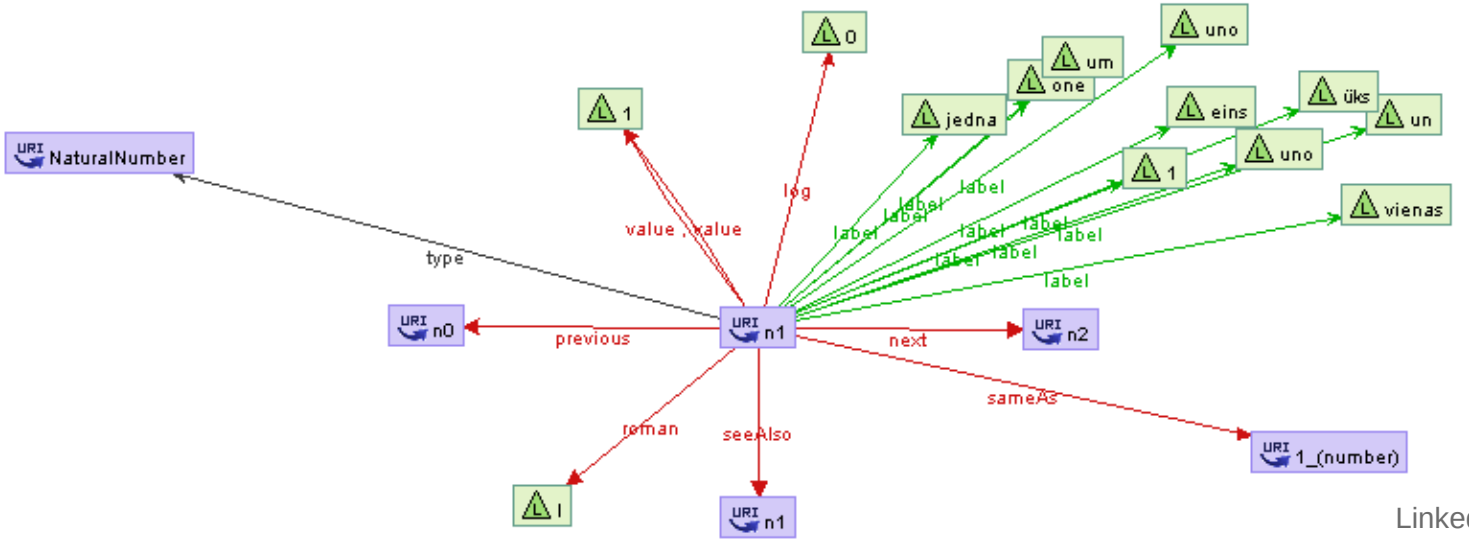
Karlsruhe Institute of Technology

Flexible Integration and Efficient Analysis of Multidimensional Datasets from the Web

Benedikt Kämpgen

Defense Talk, February 25 2015

Institute of Applied Informatics and Formal Description Methods (AIFB)



Linked Open Numbers

Imagine… (1)

- You are a financial analyst
- Assessing performance of Mastercard
- Claims supported with numbers

Indicator	Assets	Share Price	# Employees
Value	?	?	?

Company = Mastercard Inc., Date = 2015

Imagine… (2)

- You are a journalist
- Writing article about financial situation of Greece
- Claims supported with numbers

Indicator	GDP per Capita
Geo	GR
Date	2010
Unit	Euro per inhabitant
Value	?

Imagine… (2)

- You are a journalist
- Writing article about financial situation of Greece
- Claims supported with numbers

Indicator	GDP per Capita
Geo	GR
Date	2010
Unit	Euro per inhabitant
Value	*** 19,900**

* Confirmed by 2 datasets

Imagine… (3)

- You are a scientist
- Identifying correlations between indicators
- Claims supported with numbers

Indicator	Indicator 1	Indicator 2
Date		
2012	?	?
2013	?	?
2014	?	?
2015	?	?
2016	?	?
…	…	…

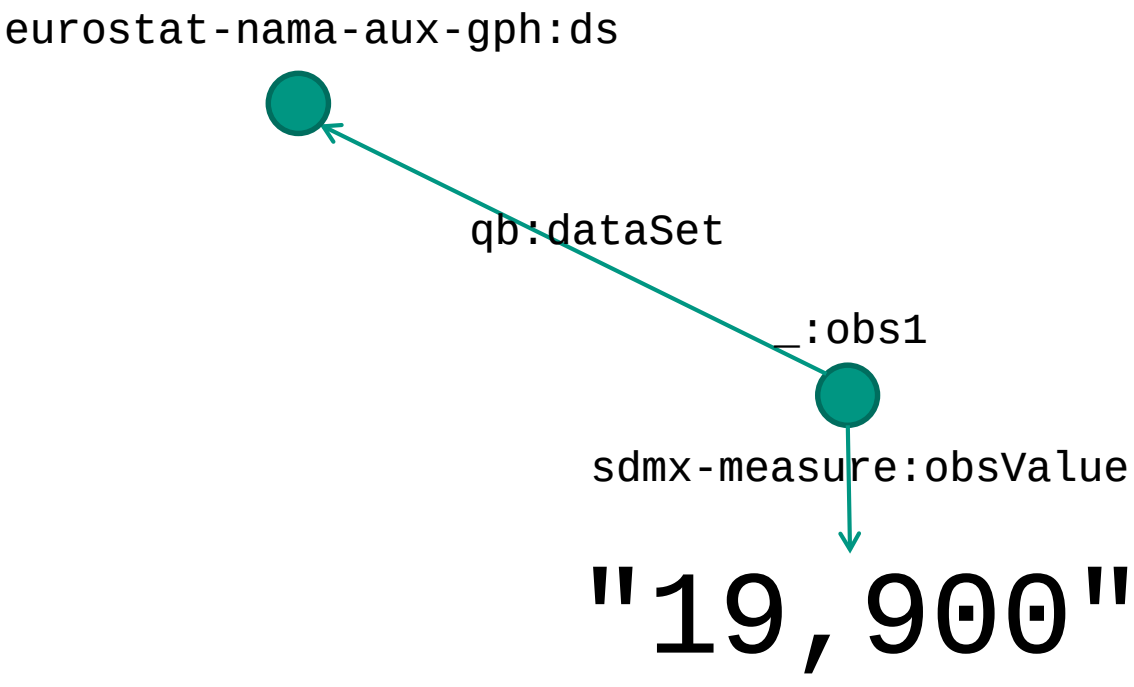
Geo = DE

Results used to import datasets from the Web (Paulheim et al. '14)

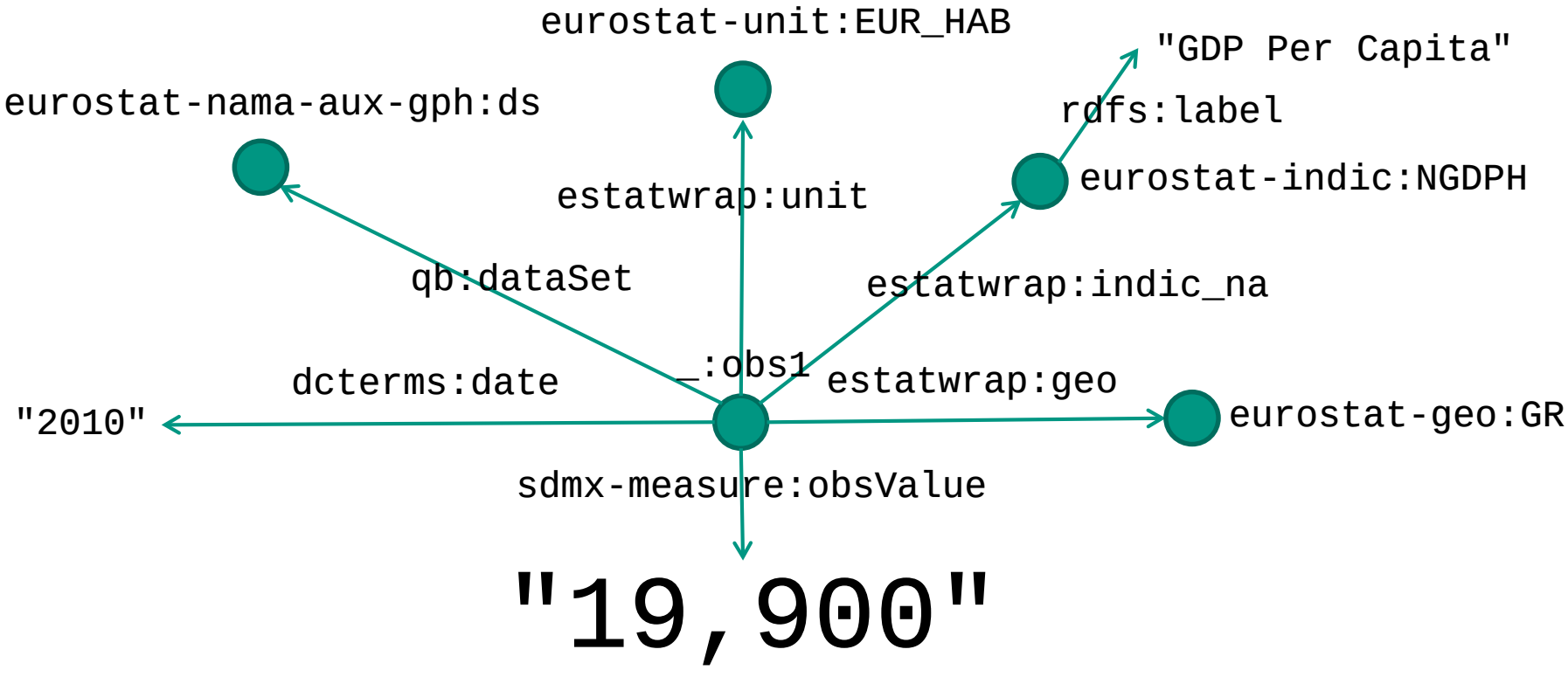
Statistical Linked Data – Number

"19,900"

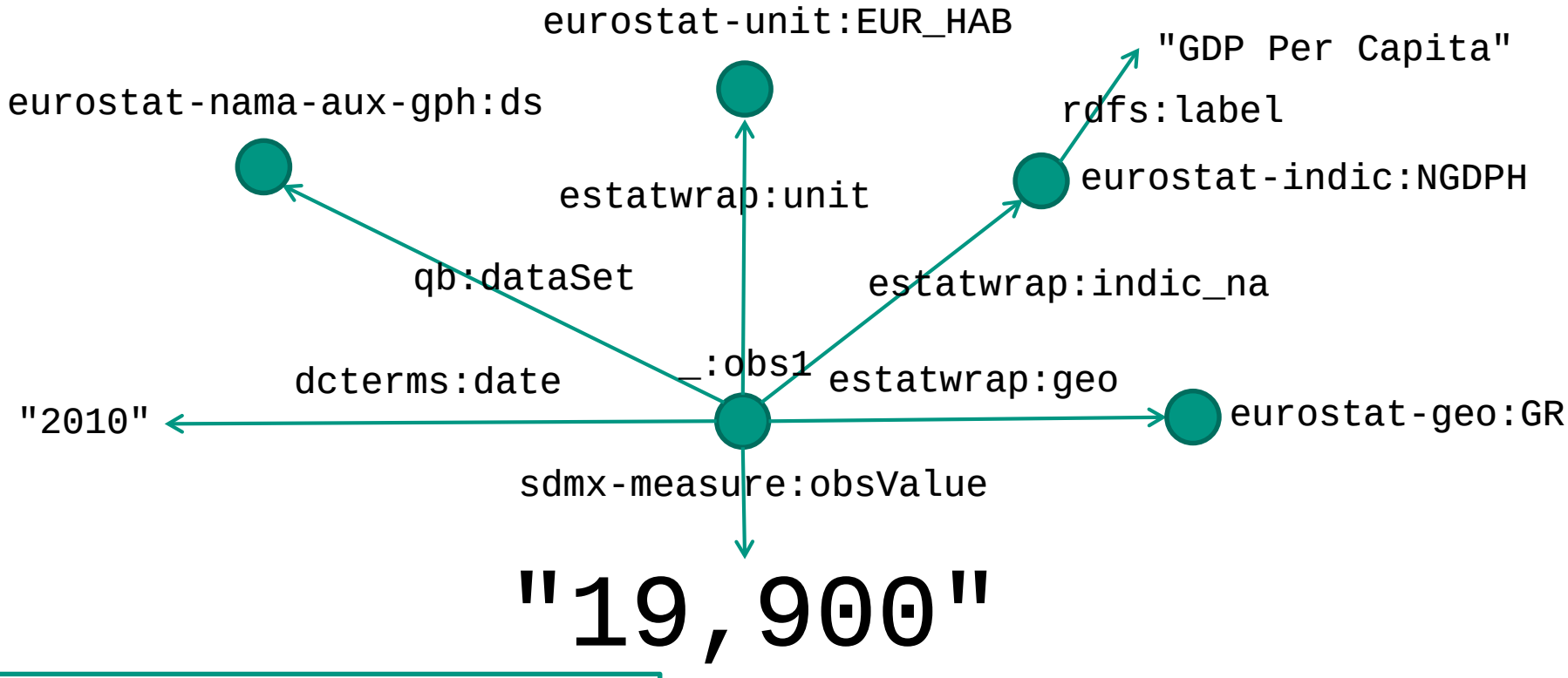
Statistical Linked Data – QB Observation and Dataset



Statistical Linked Data – Dimensions



Statistical Linked Data – Dimensions



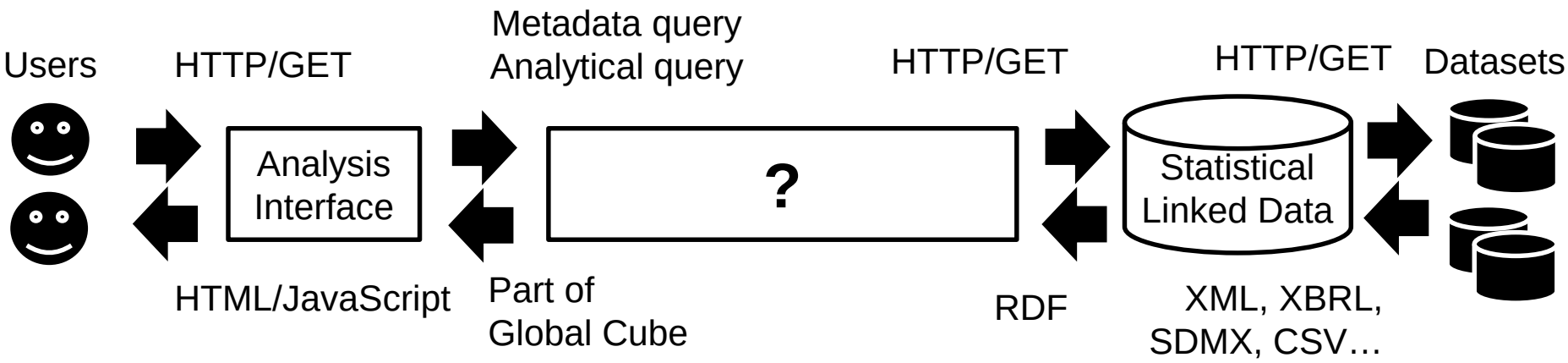
Experiences supported standardisation of QB vocabulary (Kämpgen & Cyganiak '13)

Outline

- Motivation
- **Research Questions**
- Contributions
- Related Work
- Application
- Conclusions

Institute of Applied Informatics and Formal Description Methods (AIFB)

Problem



Institute of Applied Informatics and Formal Description Methods (AIFB)

Research Questions

1. How to integrate datasets?
2. How to evaluate queries using SPARQL?
3. How to pre-aggregate values?
4. How to automatically derive values?

Research Questions

1. **How to integrate datasets?**
2. How to evaluate queries using SPARQL?
3. How to pre-aggregate values?
4. How to automatically derive values?

Integrating Datasets

Geo	Unit	Date	Variable	Value
GR	MIO_EUR	2010	B1G	1,547,984
…	…	…	…	…

GDP Components Dataset

Loc.	Date	Sex	Age	Value
Greece	2010	F	<65	4,183,516
…	…	…	…	…

Population Dataset

Geo	Unit	Date	Indicator	Value
GR	EUR_HAB	2010	NGDPH	19,900
…	…	…	…	…

GDP Per Capita Dataset

Integrating Datasets

Geo / Loc.	Unit	Date	Variable / Indicator	Sex	Age	Value
GR / Greece	MIO_EUR	2010	B1G	ALL	ALL	**1,547,984**
…	…	…	…	ALL	ALL	**…**
GR / Greece	ALL	2010	ALL	F	<65	**4,183,516**
...	ALL	...	ALL	...	...	**...**
GR / Greece	EUR_HAB	2010	NGDPH	ALL	ALL	**19,900**
...	...	...	...	ALL	ALL	**...**

Global Cube

Research Questions

1. How to integrate datasets?
2. **How to evaluate queries using SPARQL?**
3. How to pre-aggregate values?
4. How to automatically derive values?

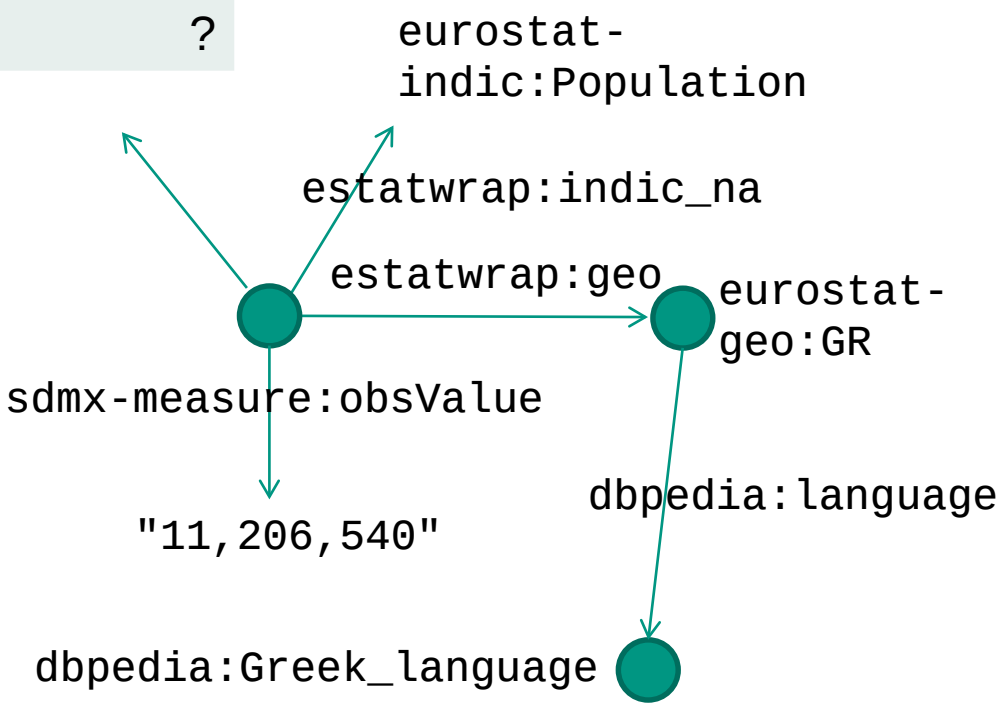
Using SPARQL for Analytical Queries

Analytical Query

Indicator	Population

Date	Geo	
2010	GR	?

Queried RDF Graph



Research Questions

1. How to integrate datasets?
2. How to evaluate queries using SPARQL?
3. **How to pre-aggregate values?**
4. How to automatically derive values?

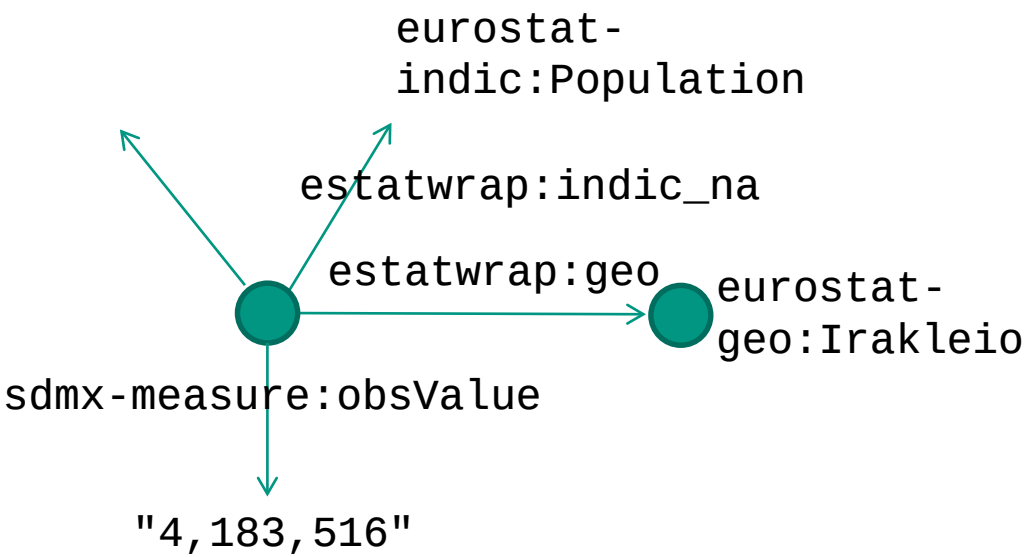
Pre-Aggregating Values

Analytical Query

Indicator	Population

Date	Geo	
2010	GR	?

Queried RDF Graph



Research Questions

1. How to integrate datasets?
2. How to evaluate queries using SPARQL?
3. How to pre-aggregate values?
4. **How to automatically derive values?**

Deriving Values

Geo / Loc.	Unit	Date	Variable / Indicator	Sex	Age	Value
GR / Greece	MIO_EUR	2010	B1G	ALL	ALL	**1,547,984**
…	…	…	…	ALL	ALL	**…**
GR / Greece	ALL	2010	ALL	F	>65	**4,183,516**
…	ALL	…	ALL	…	…	**…**
GR / Greece	EUR_HAB	2010	NGDPH	ALL	ALL	**19,900**
…	…	…	…	ALL	ALL	**…**

Global Cube

Institute of Applied Informatics and Formal Description Methods (AIFB)

Deriving Values

Geo / Loc.	Unit	Date	Variable / Indicator	Sex	Age	Value
GR / Greece	MIO_EUR	2010	B1G	ALL	ALL	**1,547,984**
…	…	…	…	ALL	ALL	**…**
GR / Greece	ALL	2010	ALL	F	>65	**4,183,516**
…	ALL	…	ALL	…	…	**…**
GR / Greece	EUR_HAB	2010	NGDPH	ALL	ALL	**19,900**
…	…	…	…	ALL	ALL	**…**

Outline

- Motivation
- Research Questions
- **Contributions**
- Related Work
- Application
- Conclusions

Institute of Applied Informatics and Formal Description Methods (AIFB)

Contributions

How to integrate datasets?

1 MDM-QB Mapping [ISEM Research 2011]

How to evaluate queries using SPARQL?

2 OLAP-to-SPARQL Algorithm [ESWC Workshop 2012]

How to pre-aggregate values?

3 RDF Aggregate Views [ESWC Research 2013]

How to automatically derive values?

4 Complex Correspondences [EKAW Research 2014]

Contributions

How to integrate datasets?

1 MDM-QB Mapping [ISEM Research 2011]

How to evaluate queries using SPARQL?

2 OLAP-to-SPARQL Algorithm [ESWC Workshop 2012]

How to pre-aggregate values?

3 **RDF Aggregate Views** [ESWC Research 2013]

How to automatically derive values?

4 Complex Correspondences [EKAW Research 2014]

RDF Aggregate Views – Problem & Approach

- Definition
  - Values on same **aggregation level**

View

estatwrap:geo (cities)
estatwrap:indic_na
estatwrap:unit

Institute of Applied Informatics and Formal Description Methods (AIFB)

RDF Aggregate Views – Problem & Approach

- Definition
  - Values on same **aggregation level**

View

estatwrap:geo (cities)
estatwrap:indic_na
estatwrap:unit

- Computation
  - Specific indices require adapted query engines
  - Aggregations with **SPARQL** and materialisation as **RDF**

RDF Aggregate Views – Problem & Approach

- Definition
  - Values on same **aggregation level**

View

estatwrap:geo (cities)
estatwrap:indic_na
estatwrap:unit

- Computation
  - Specific indices require adapted query engines
  - Aggregations with **SPARQL** and materialisation as **RDF**
- Selection
  - Problem with **limited space** is NP-complete (Harinarayan et al. '96)
  - Best case: select **smallest view** for each query

Evaluation Results – SQL vs SPARQL

- Star Schema Benchmark (O'Neil et al. '09)
  - **Data:** 6M lineorders as Star Schema
  - **Queries:** 13 queries in SQL

	No Materialisation	Materialisation
Star Schema (MySQL)	SQL	
RDF / QB (Open Virtuoso)	SPARQL	SPARQL-M

$$\frac{\sum eqt_{sparql}(q_i)}{\sum eqt_{sql}(q_i)} \approx 12$$

Analytical query processing more difficult with RDF than with Star Schema

Evaluation Results – SQL vs SPARQL

- Star Schema Benchmark (O'Neil et al. '09)
  - **Data:** 6M lineorders as Star Schema
  - **Queries:** 13 queries in SQL

	No Materialisation	Materialisation
Star Schema (MySQL)	SQL	
RDF / QB (Open Virtuoso)	SPARQL	SPARQL-M

OPENLINK SOFTWARE
Making Technology Work For You

Benchmark repeated with 300GB confirmed difficulty of analytical queries over RDF (Erling '13)

Analytical query processing more difficult with RDF than with Star Schema

Evaluation Results – SPARQL vs SPARQL-M

- Star Schema Benchmark (O'Neil et al. '09)
  - **Data:** 6M lineorders as Star Schema
  - **Queries:** 13 queries in SQL

	No Materialisation	**Materialisation**
Star Schema (MySQL)	SQL	
RDF / QB (Open Virtuoso)	SPARQL	SPARQL-M

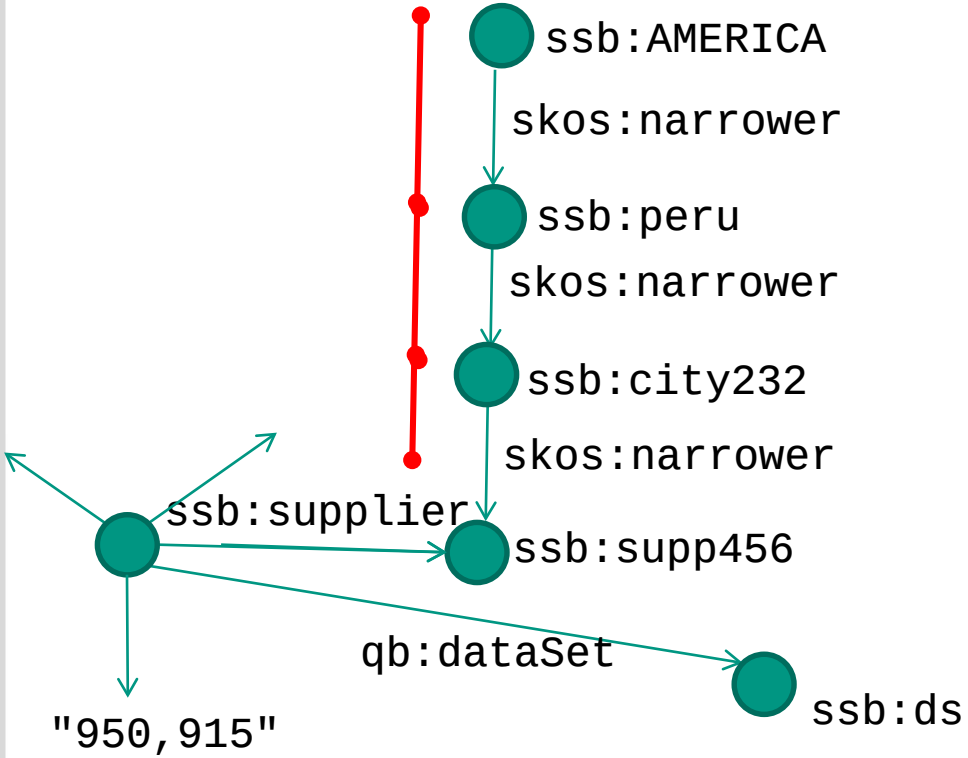
Pre-computing aggregate views in RDF not always more efficient

$$\frac{\sum eqt_{sparql-m}(q_i)}{\sum eqt_{sparql}(q_i)} \approx 2$$

Institute of Applied Informatics and Formal Description Methods (AIFB)

Evaluation Results – SPARQL vs SPARQL-M (2)

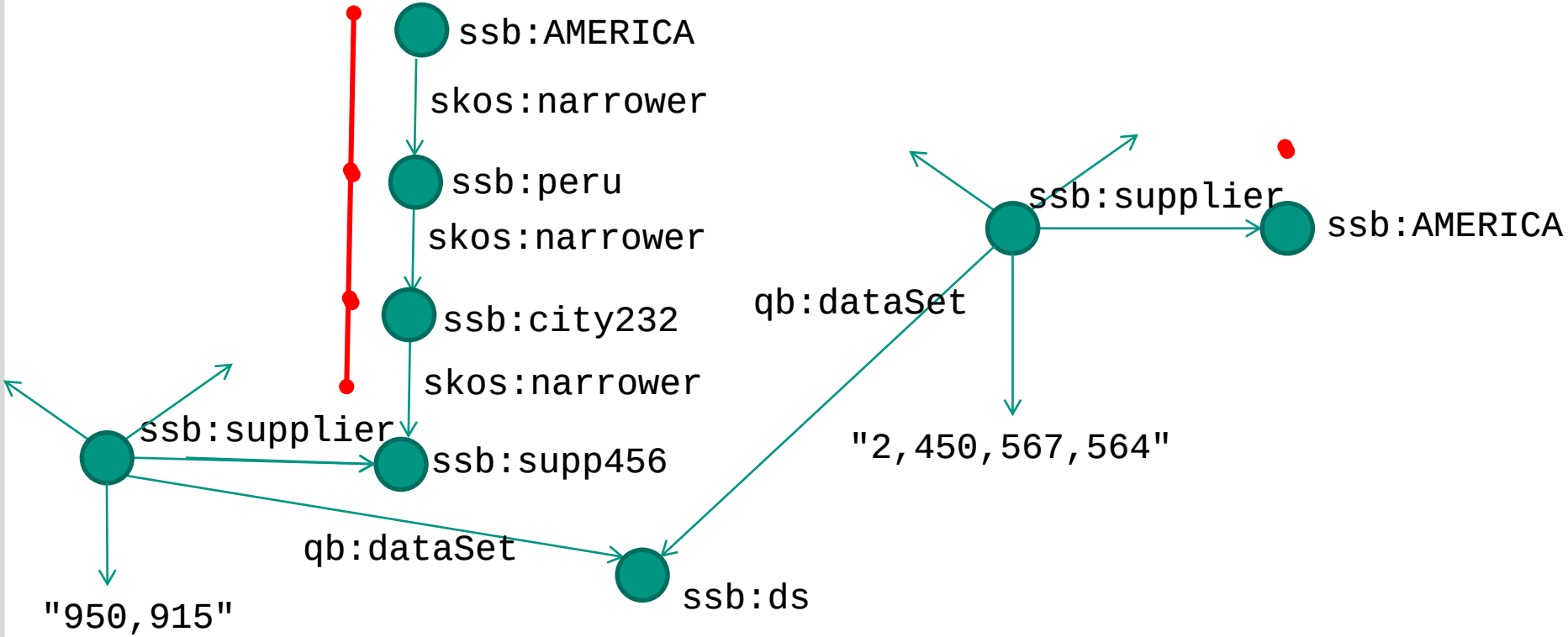
SPARQL



Institute of Applied Informatics and Formal Description Methods (AIFB)

Evaluation Results – SPARQL vs SPARQL-M (2)

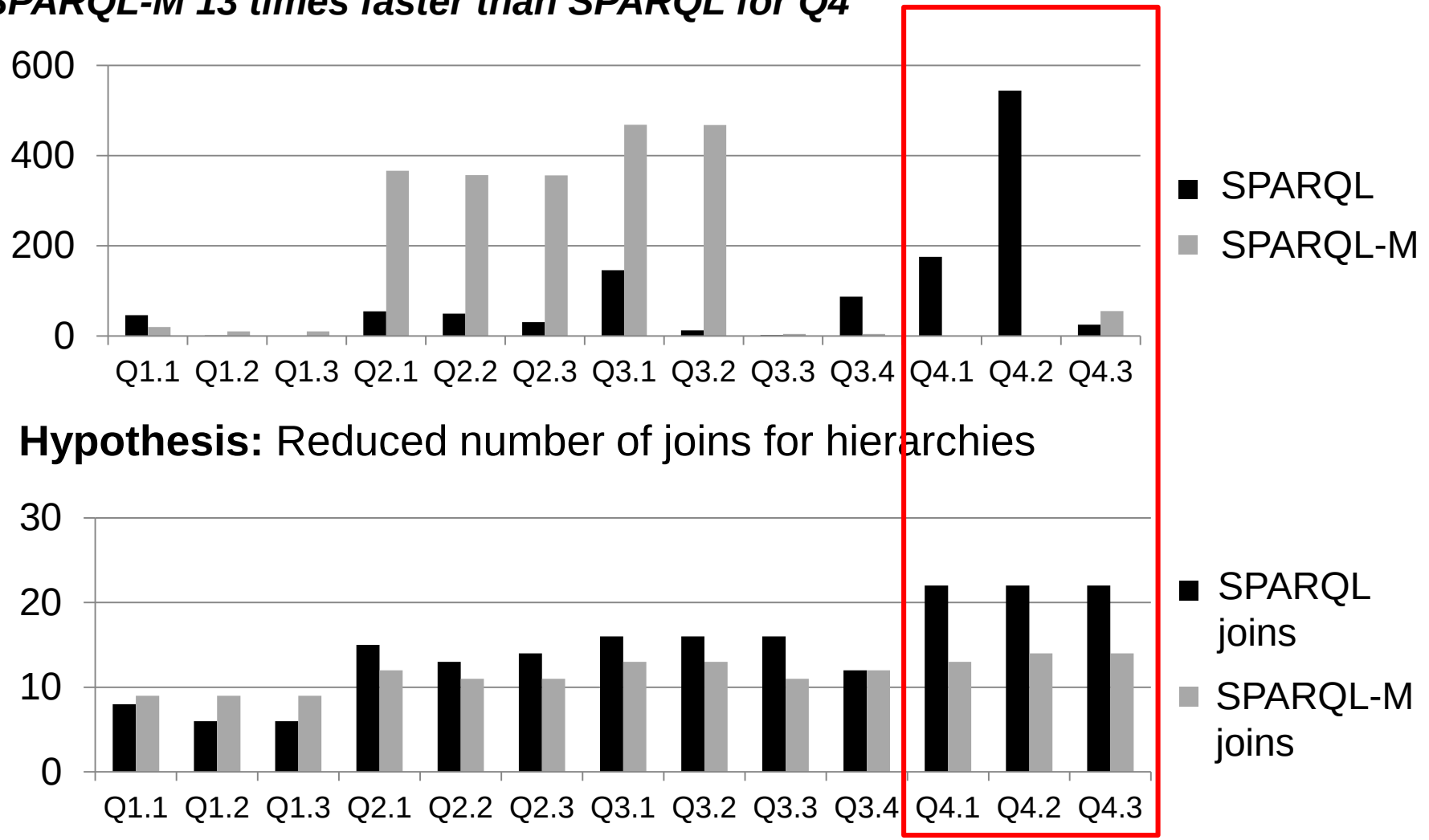
SPARQL-M



- Reduced number of joins
- But: triples are stored in same graph

Evaluation Results – SPARQL vs SPARQL-M (3)

SPARQL-M 13 times faster than SPARQL for Q4



Hypothesis: Reduced number of joins for hierarchies

Contributions

How to integrate datasets?

1 MDM-QB Mapping [ISEM Research 2011]

How to evaluate queries using SPARQL?

2 OLAP-to-SPARQL Algorithm [ESWC Workshop 2012]

How to pre-aggregate values?

3 RDF Aggregate Views [ESWC Research 2013]

How to automatically derive values?

4 **Complex Correspondences** [EKAW Research 2014]

Complex Correspondences – Problem

Geo / Loc.	Unit	Date	Variable / Indicator	Sex	Age	Value
GR / Greece	MIO_EUR	2010	B1G	ALL	ALL	1,547,984
…	…	…	…	ALL	ALL	…
GR / Greece	ALL ?	2010	ALL	F	<65	4,183,516
…	ALL	…	ALL	…	…	…
GR / Greece	EUR_HAB	2010	NGDPH	ALL	ALL	19,900
…	…	…	…	ALL	ALL	…

Complex Correspondences – Approach (2)

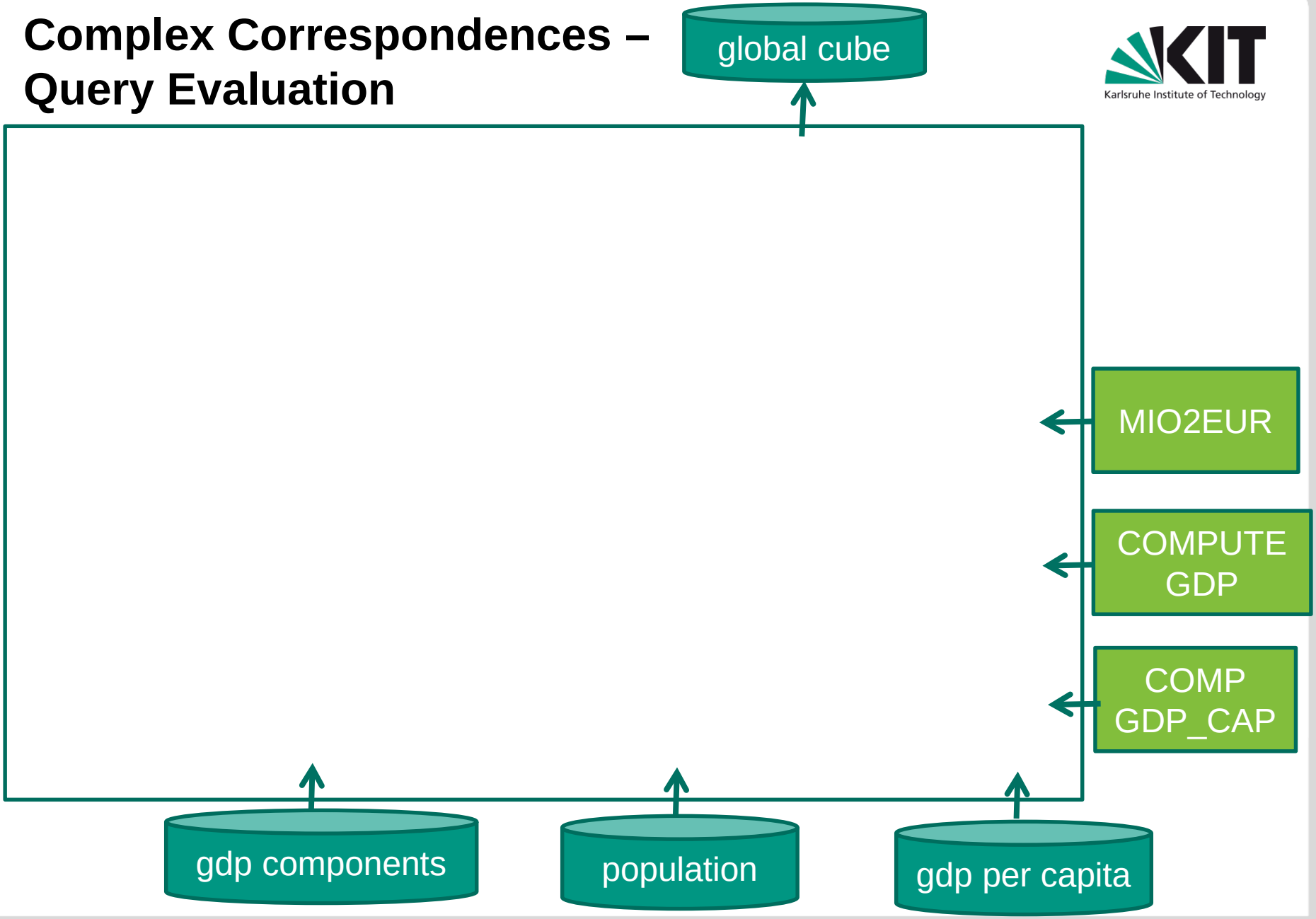
- Syntax (Example)

```
MIO2EUR =
({(unit, MIO_EUR)},          //inputmembers
{(unit, EUR)},               //outputmembers
"f(x) = 1,000,000 * x")      //function
```

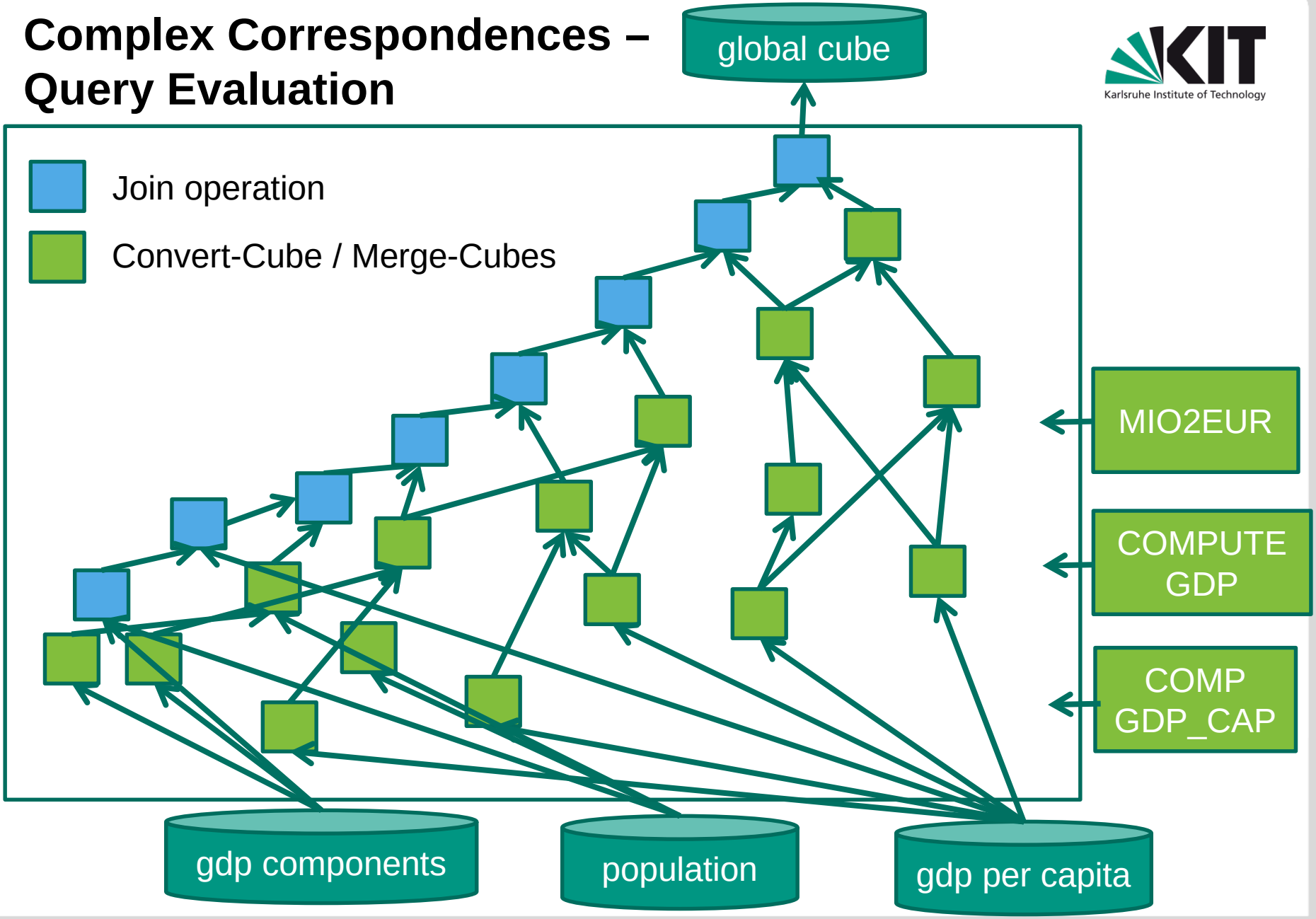
- Semantics (Example)

```
Convert-Cube(gdp_components, MIO2EUR) = gdp_components_mio

gdp_components_mio(Geo, EUR, Date, Indicator, Value2)
:- gdp_components(Geo, MIO_EUR, Date, Indicator, Value1),
Value2 = 1,000,000 * Value1.

                                  //implemented using SPARQL
```

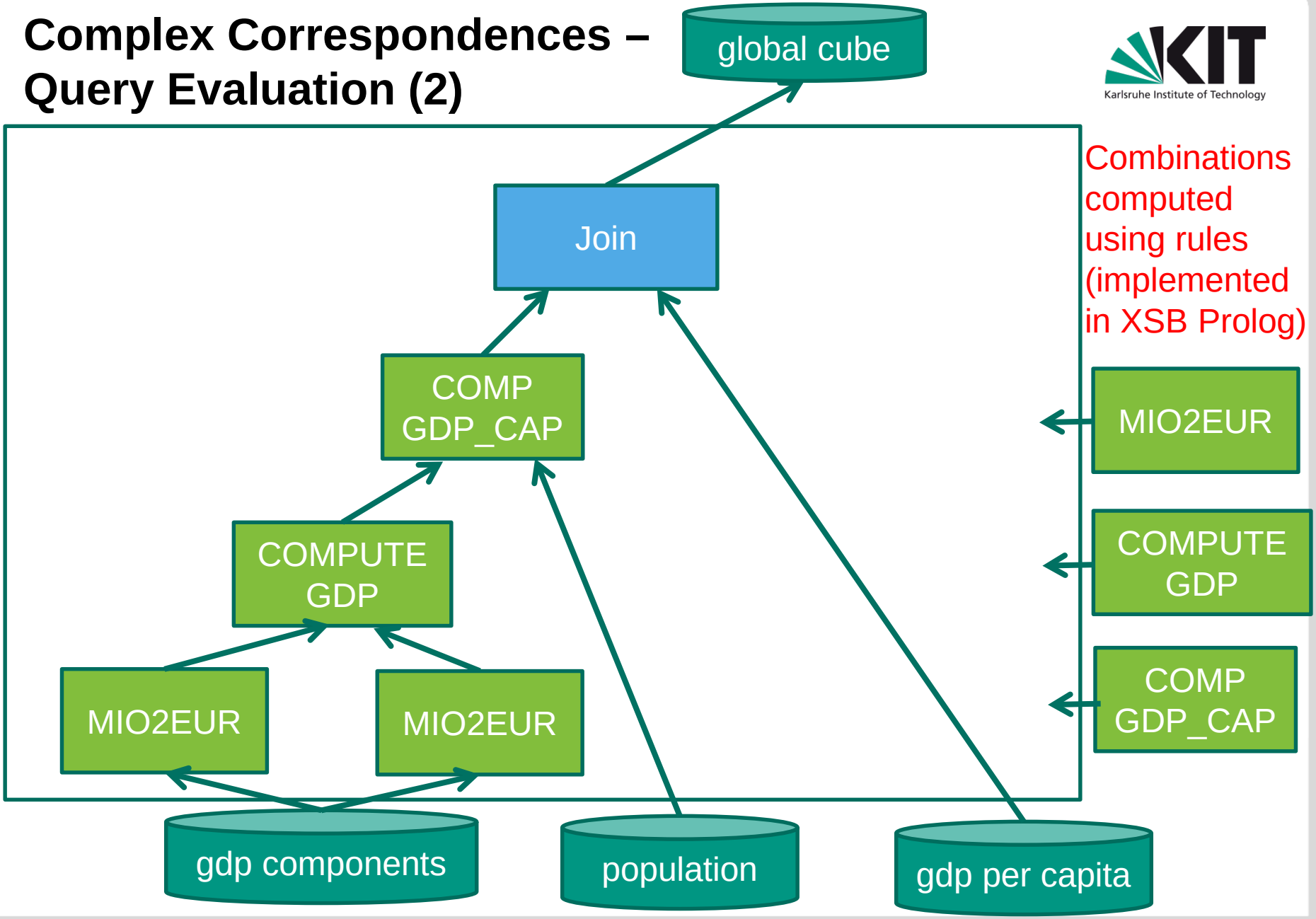
Complex Correspondences – Query Evaluation



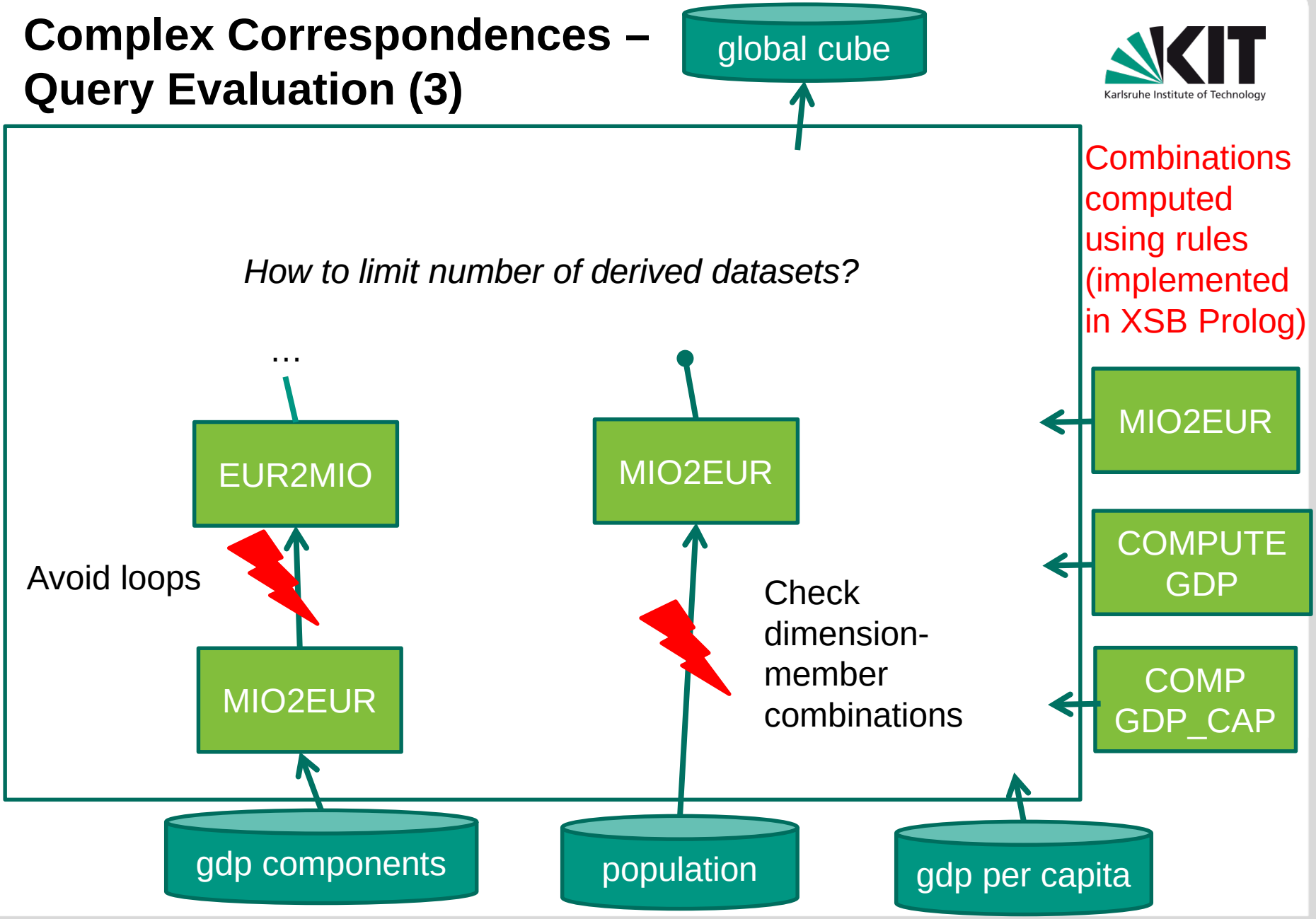
Complex Correspondences – Query Evaluation



Complex Correspondences – Query Evaluation (2)



Complex Correspondences – Query Evaluation (3)

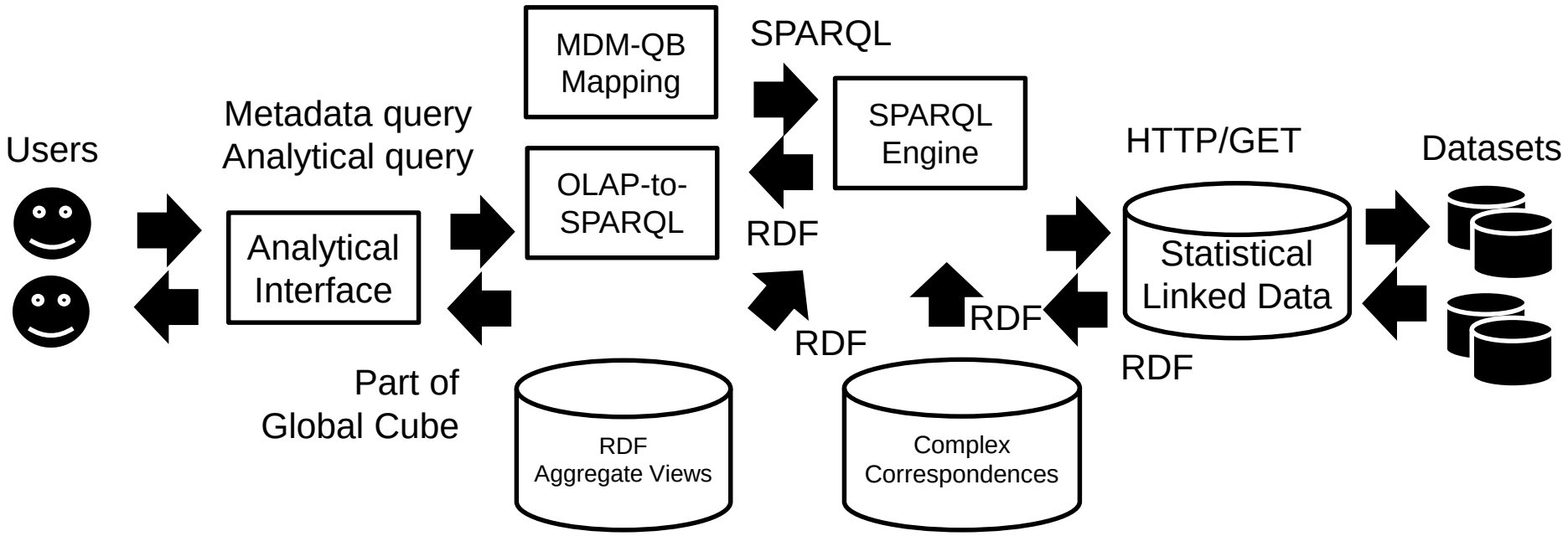


Outline

- Motivation
- Research Questions
- Contributions
- **Related Work**
- Application
- Conclusions

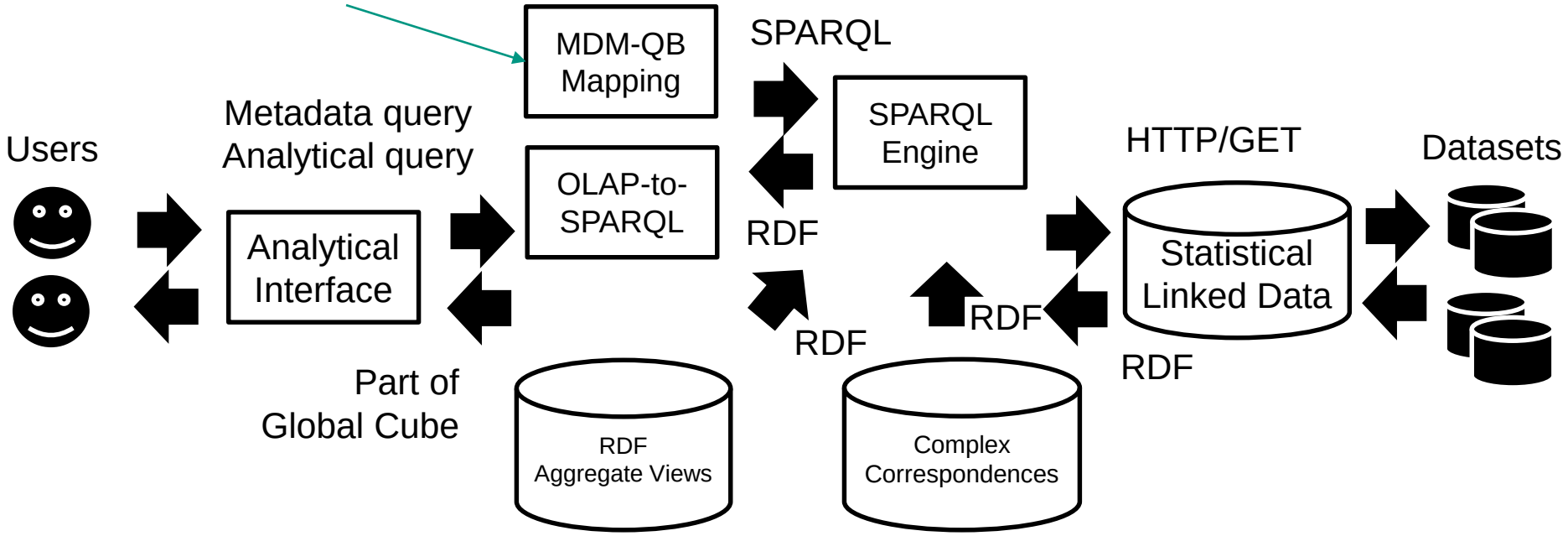
Institute of Applied Informatics and Formal Description Methods (AIFB)

Related Work



Related Work

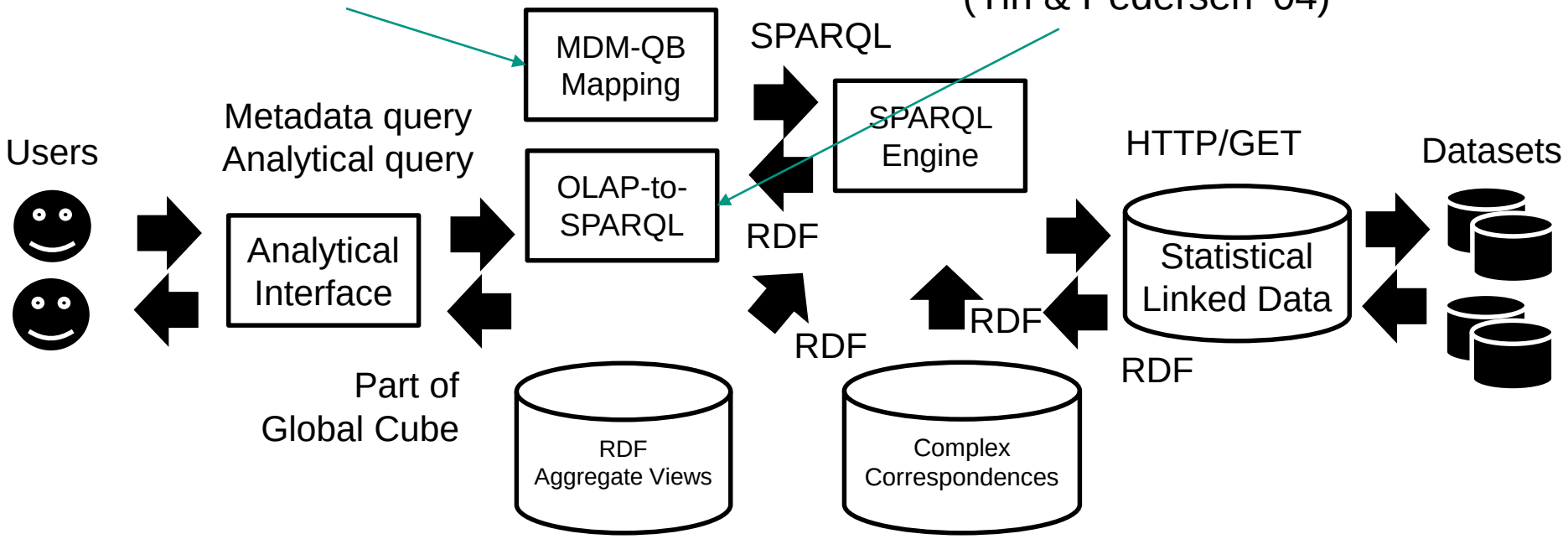
1 **Ontology Matching**
(Zapilko & Mathiak '14)
(Torlone '08)



Related Work

1 **Ontology Matching**
(Zapilko & Mathiak '14)
(Torlone '08)

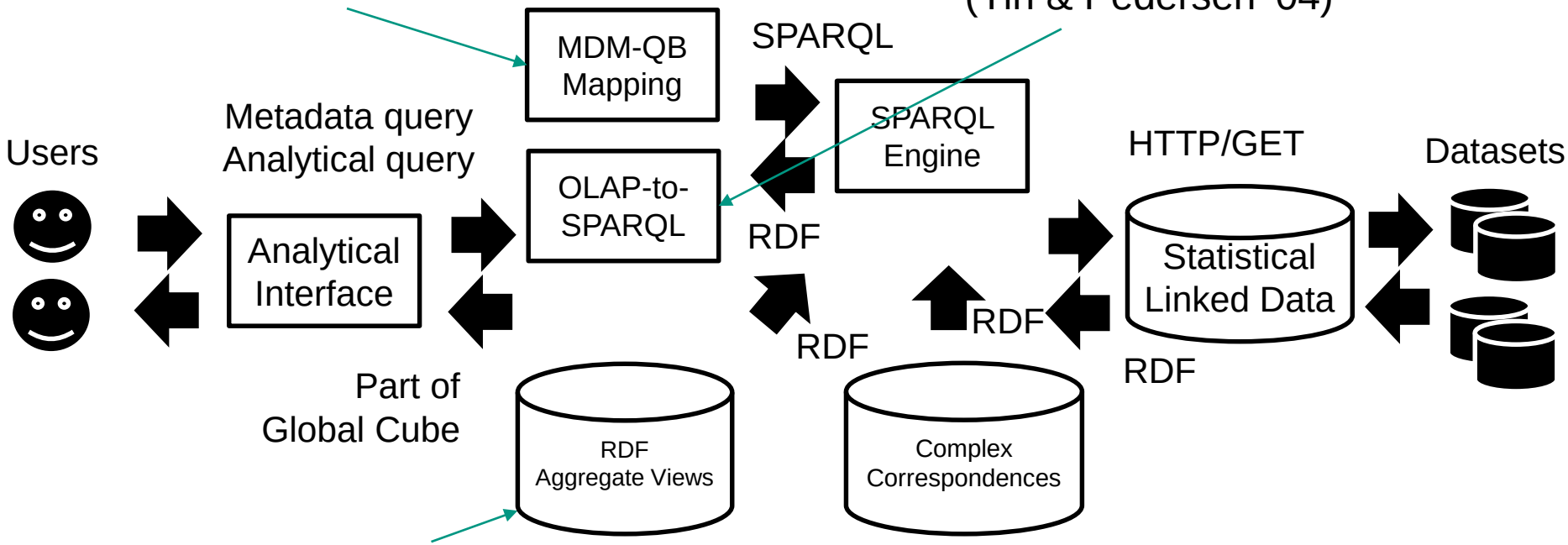
2 **Query Processing over Cubes and External Data**
(Potena & Diamantini '08)
(Yin & Pedersen '04)



Related Work

1 **Ontology Matching**
(Zapilko & Mathiak '14)
(Torlone '08)

2 **Query Processing over Cubes and External Data**
(Potena & Diamantini '08)
(Yin & Pedersen '04)

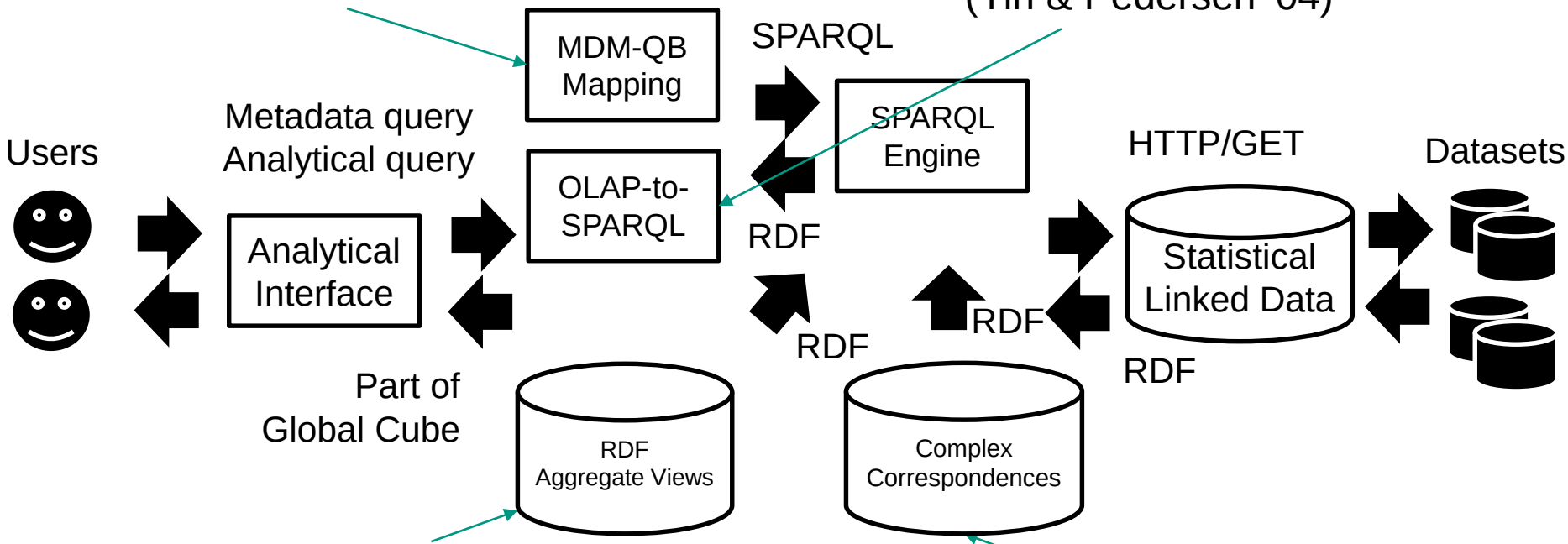


3 **View Definition using SPARQL (Etcheverry & Vaisman '12)**
SPARQL++, Networked Graphs, vSPARQL

Related Work

1 **Ontology Matching**
(Zapilko & Mathiak '14)
(Torlone '08)

2 **Query Processing over Cubes and External Data**
(Potena & Diamantini '08)
(Yin & Pedersen '04)



3 **View Definition using SPARQL (Etcheverry & Vaisman '12)**
SPARQL++, Networked Graphs, vSPARQL

4 **Dataset Integration**
(Ambite & Kapoor '07)
(Bischof & Polleres '13)

Outline

- Motivation
- Research Questions
- Contributions
- Related Work
- **Application**
- Conclusions

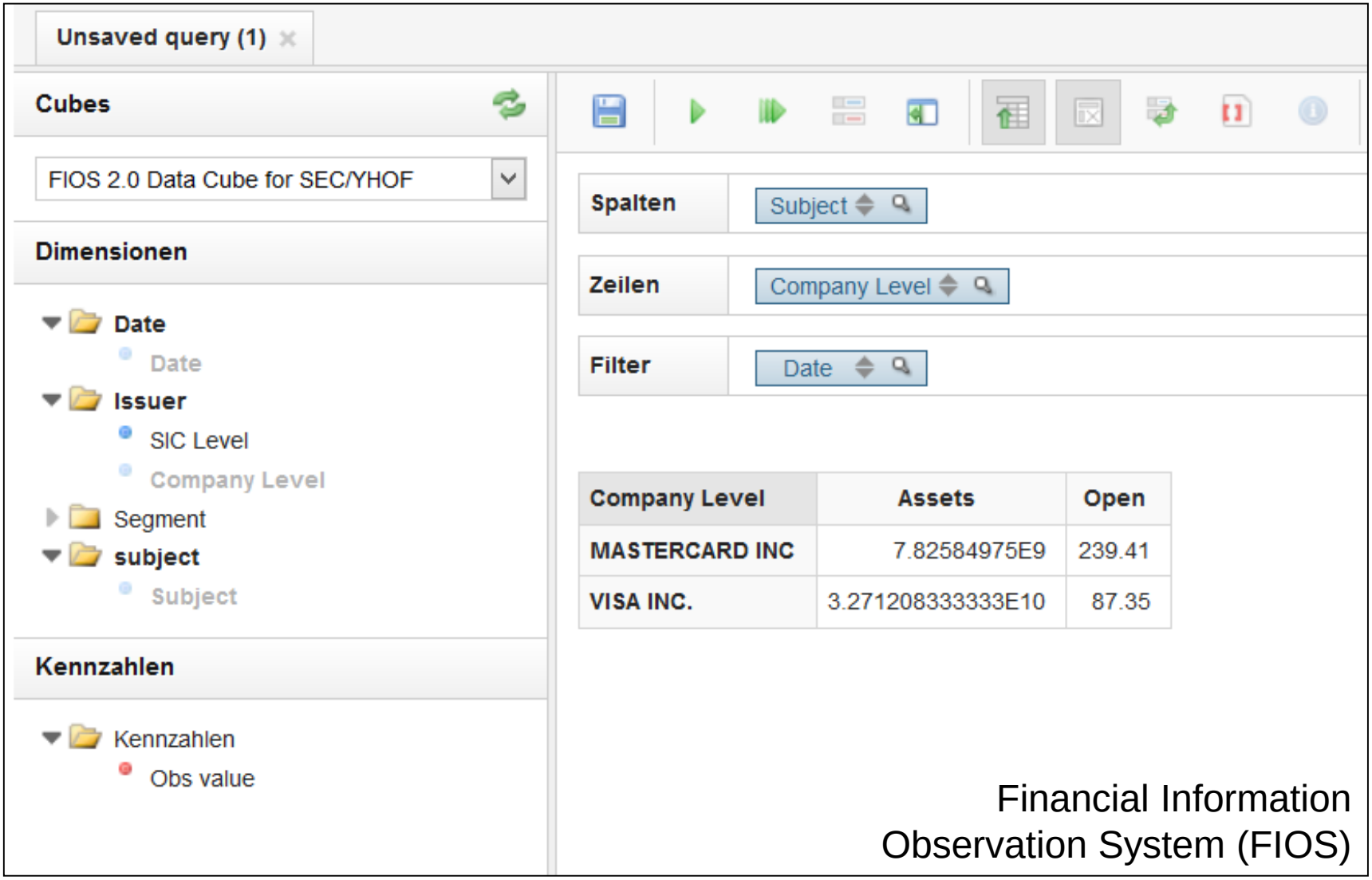
Application

- Natural Science [SePublica Workshop 2014]
- Financial Data Analysis [ESWC In-Use 2014]
- Open Government Data [ESWC Demo 2014]

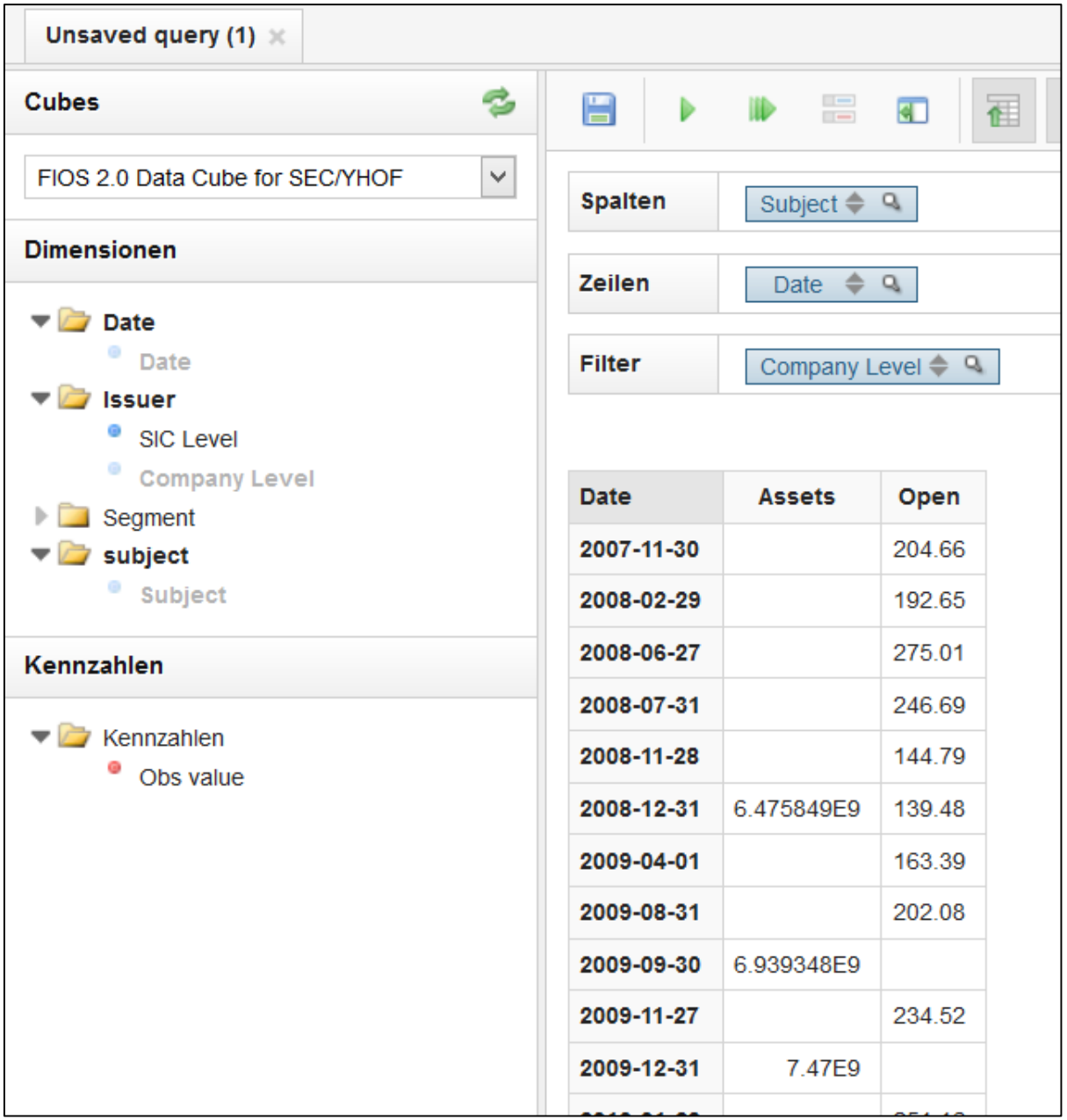
Application

- Natural Science [SePublica Workshop 2014]
- **Financial Data Analysis** [ESWC In-Use 2014]
- Open Government Data [ESWC Demo 2014]

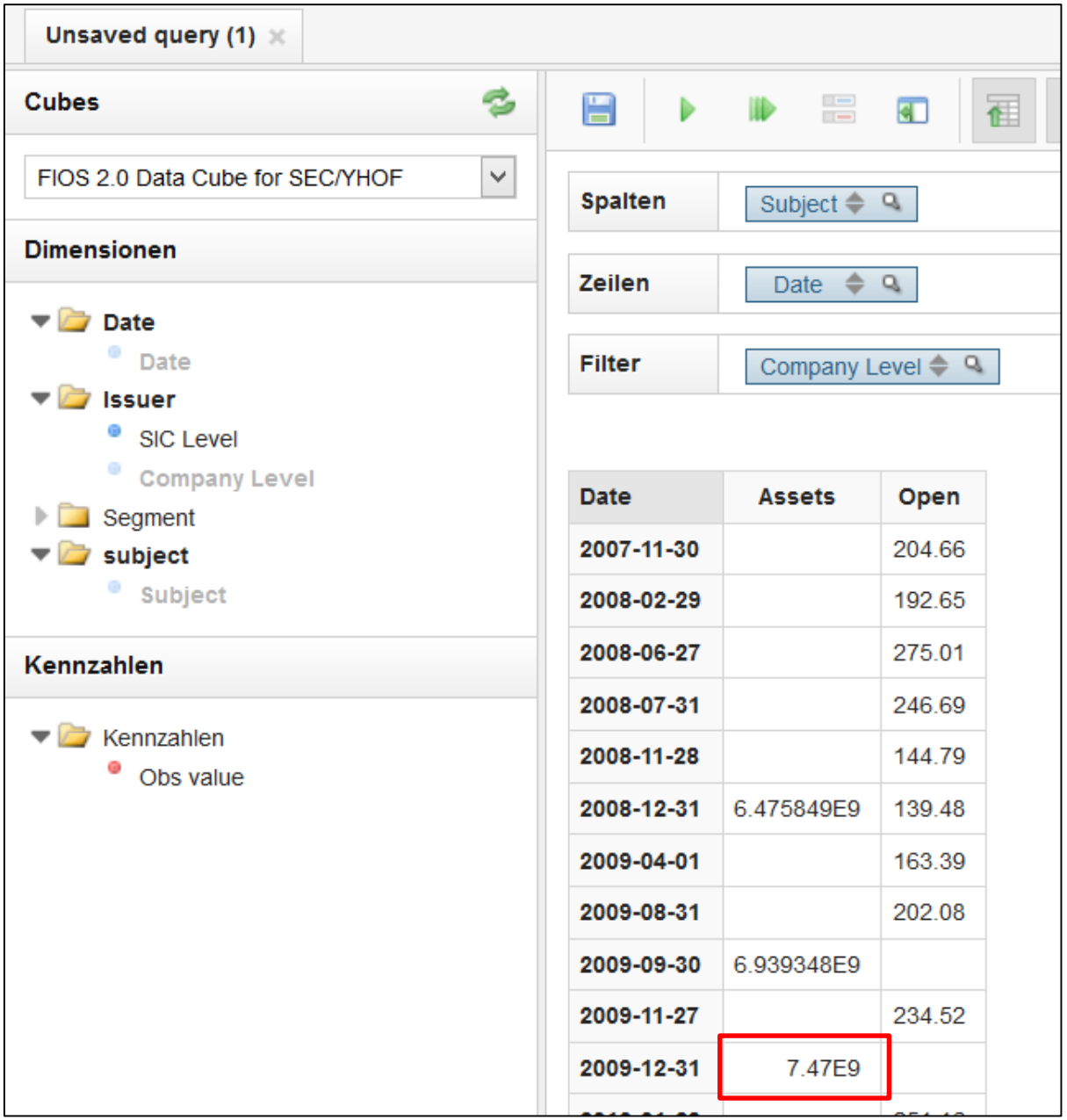
Analyst Assessing Mastercard – Overview



Analyst Assessing Mastercard – Zoom-In



Analyst Assessing Mastercard – Zoom-In



Analyst Assessing Mastercard – Details

- Browsing single Assets value

Property	Value
qb:dataSet	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
dcterms:date	2009-12-31
?:issuer	<http://public.b-kaempgen.de:8080/pubby/cik/1141391%23id>
sdmx-measure:obsValue	7470000000
is rdfs:seeAlso of	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
?:segment	0001141391 2009-12-31
?:subject	<http://public.b-kaempgen.de:8080/pubby/vocab/us-gaap-2009-01-31%23Assets>
rdf:type	qb:Observation

Analyst Assessing Mastercard – Details

- Browsing single Assets value

Property	Value
qb:dataSet	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
dcterms:date	2009-12-31
?:issuer	<http://public.b-kaempgen.de:8080/pubby/cik/1141391%23id>
sdmx-measure:obsValue	7470000000
is rdfs:seeAlso of	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
?:segment	0001141391 2009-12-31
?:subject	<http://public.b-kaempgen.de:8080/pubby/vocab/us-gaap-2009-01-31%23Assets>
rdf:type	qb:Observation

Analyst Assessing Mastercard – Details

- Browsing single Assets value

Property	Value
qb:dataSet	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
dcterms:date	2009-12-31
?:issuer	<http://public.b-kaempgen.de:8080/pubby/cik/1141391%23id>
sdmx-measure:obsValue	7470000000
is rdfs:seeAlso of	<http://public.b-kaempgen.de:8080/pubby/archive/1141391/0001193125-10-243917%23ds>
?:segment	0001141391 2009-12-31
?:subject	<http://public.b-kaempgen.de:8080/pubby/vocab/us-gaap-2009-01-31%23Assets>
rdf:type	qb:Observation

- Browsing Mastercard

?:locationCountry	<http://public.b-kaempgen.de:8080/pubby/dbp/United_States>
?:logo	190 (xsd:int)
foaf:name	MasterCard Incorporated (en)
?:name	MasterCard Incorporated (en)
?:netIncome	US$ 1.846 billion (en)
?:numEmployees	5600 (xsd:int)

Outline

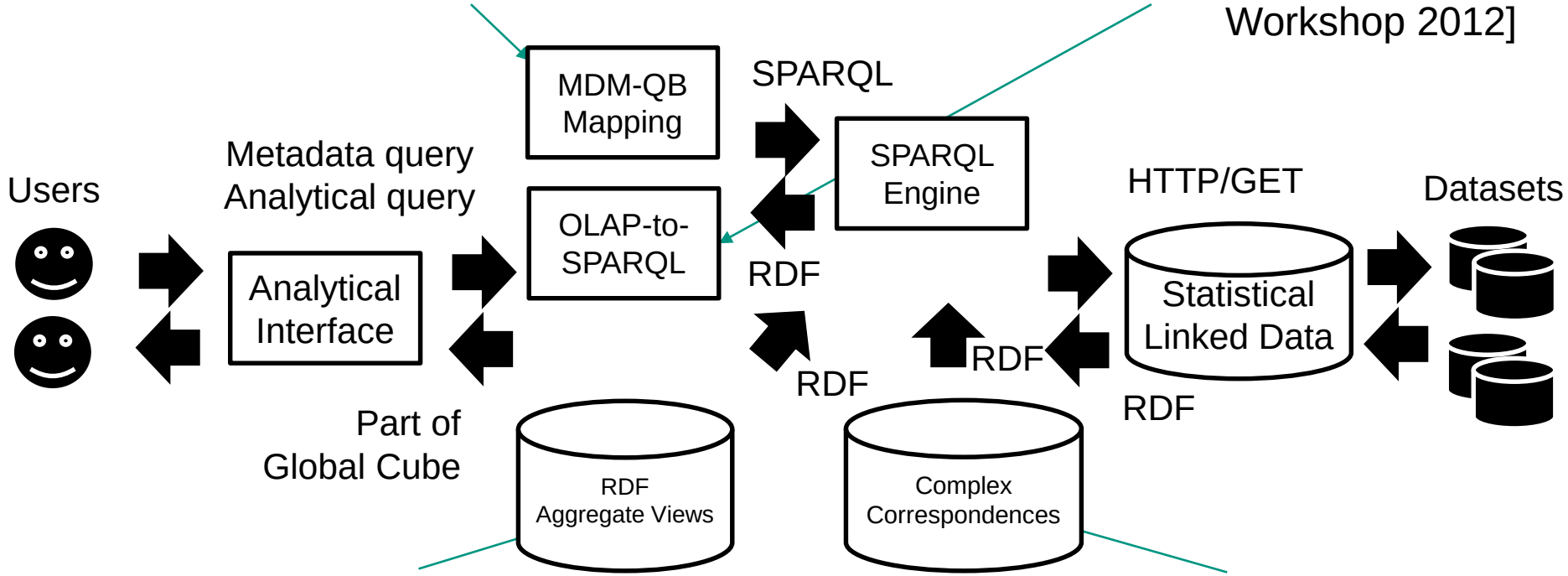
- Motivation
- Research Questions
- Contributions
- Related Work
- Application
- **Conclusions**

Institute of Applied Informatics and Formal Description Methods (AIFB)

Summary of Results

1 Mapping based on simple set notation from Statistical Linked Data to Global Cube [ISEM Research 2011]

2 Algorithm to translate any OLAP query to a SPARQL SELECT query [ESWC Workshop 2012]



3 Evaluation of performance gain of pre-aggregated values for analytical query processing over RDF [ESWC Research 2013]

4 Formalisation of implicit relationships between datasets based on rules for data integration [EKAW Research 2014]

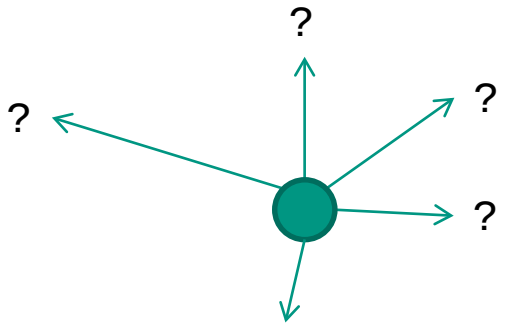
Open Questions

- How to extract more background information?
- How to efficiently query global cube?
- How to model more complex domains?

Institute of Applied Informatics and Formal Description Methods (AIFB)

Open Questions

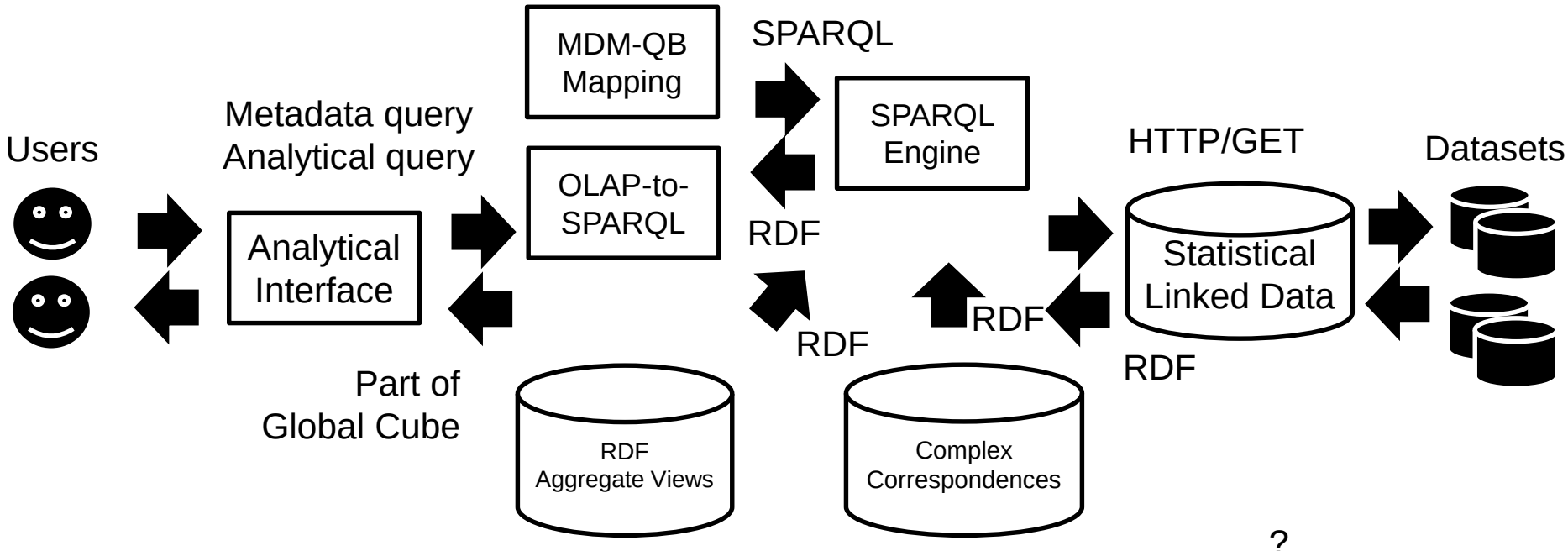
- How to extract more background information?
- How to efficiently query global cube?
- How to model more complex domains?

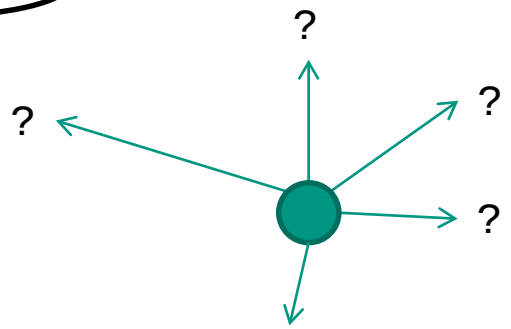


Douglas Adams. The Hitchhiker's Guide to the Galaxy

"42"

Thank you!


Users
Metadata query
Analytical query
Analytical Interface
Part of Global Cube
MDM-QB Mapping
OLAP-to-SPARQL
SPARQL
SPARQL Engine
RDF
RDF
RDF Aggregate Views
Complex Correspondences
RDF
HTTP/GET
Statistical Linked Data
RDF
Datasets


?
?
?
?

"42"

References

Daniel J. Abadi and Samuel R. Madden. Column-Stores vs . Row-Stores: How Different Are They Really? In ACM SIGMOD International Conference on Management of Data (SIGMOD), 2008.

José Luis Ambite and Dipsy Kapoor. Automatically Composing Data Workflows with Relational Descriptions and Shim Services. In International Semantic Web Conference (ISWC), 2007.

Stefan Bischof and Axel Polleres. RDFS with Attribute Equations via SPARQL Rewriting. In 10th Extended Semantic Web Conference (ESWC), 2013.

Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. Data Integration in Data Warehousing. International Journal of Cooperative Information Systems, 10(3), 2001.

Roger Castillo and Ulf Leser. Selecting Materialized Views for RDF Data. In 10th International Conference on Current Trends in Web Engineering, 2010.

Claudia Diamantini and Domenico Potena. Semantic Enrichment of Strategic Datacubes. In 11th ACM International Workshop on Data Warehousing and OLAP (DOLAP), 2008.

Orri Erling. Blog entry: ESWC 2013 Panel on Semantic Technologies for Big Data Analytics, http://www.openlinksw.com/dataspace/oerling/weblog/Orri%20Erling%27s%20Blog/1730

Lorena Etcheverry and Alejandro A. Vaisman. Views over RDF Datasets: A State-of-the-Art and Open Challenges. The Computing Research Repository (CoRR), Nov 2012.

References (2)

Francois Goasdoué, Konstantinos Karanasos, Julien Leblay, and Ioana Manolescu. View Selection in Semantic Web Databases. VLDB Endowment, 5(2), 2011.

Venky Harinarayan, Anand Rajaraman, and Jeffrey D. Ullman. Implementing Data Cubes Efficiently. In ACM International Conference on Management of Data (SIGMOD), 1996.

Benedikt Kämpgen and Richard Cyganiak. Use Cases and Lessons for the Data Cube Vocabulary. Working Group Note – http://www.w3.org/TR/2013/NOTE-vocab-data-cube-use-cases-20130801/, W3C, USA, Aug 2013.

Victoria Nebot and Rafael Berlanga. Building data warehouses with semantic web data. Decision Support Systems, 52(4), 2012.

Marko Niinimäki and Tapio Niemi. An ETL Process for OLAP Using RDF/OWL Ontologies. Data Semantics XIII, 5530, 2009.

Pat O'Neil, Betty O'Neil, and Xuedong Chen. Star Schema Benchmark – Revision 3. Technical report, UMass, Boston (USA), June 2009.

Heiko Paulheim, Petar Ristoski, Evgeny Mitichkin, Christian Bizer. "Accessing RDF Data Cubes" in RapidMiner Linked Open Data Extension, Manual, Version 1.5, 09/19/14, http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/research/RapidMinerLODExtension/RapidMinerLODExtensionManual.pdf

References (3)

Michael Siegel, Edward Sciore, and Arnon Rosenthal. Using Semantic Values to Facilitate Interoperability Among Heterogeneous Information Systems. ACM Transactions on Database Systems (TODS), 19(2), 1994.

Riccardo Torlone. Two approaches to the integration of heterogeneous data warehouses. Distributed and Parallel Databases, 23(1), 2008.

Elena Vasilyeva, Maik Thiele, Christof Bornhövd, and Wolfgang Lehner. Leveraging Flexible Data Management with Graph Databases. 1st International Workshop on Graph Data Management Experiences and Systems (GRADES), 2013.

Marcin Wylot, Jigé Pont, Mariusz Wisniewski, and Philippe Cudré-Mauroux. dipLODocus – Short and Long-Tail RDF Analytics for Massive Webs of Data. In 10th International Semantic Web Conference (ISWC), 2011.

Xuepeng Yin and Torben Bach Pedersen. Evaluating XML-extended OLAP Queries Based on a Physical Algebra. 7th ACM International Workshop on Data Warehousing and OLAP (DOLAP), 2004.

Benjamin Zapilko and Brigitte Mathiak. Object Property Matching Utilizing the Overlap between Imported Ontologies. In 11th Extended Semantic Web Conference (ESWC), 2014.

Own Publications

[ISEM Research 2011] Benedikt Kämpgen and Andreas Harth. Transforming Statistical Linked Data for Use in OLAP Systems. In 7th International Conference on Semantic Systems (ISEMANTICS), 2011.

[ESWC Workshop 2012] Benedikt Kämpgen and Séan O'Riain and Andreas Harth. Interacting with Statistical Linked Data via OLAP Operations. In 1st ESWC Workshop on Interacting with Linked Data (ILD), 2012.

[ESWC Research 2013] Benedikt Kämpgen and Andreas Harth. No Size Fits All – Running the Star Schema Benchmark with SPARQL and RDF Aggregate Views. In 10th Extended Semantic Web Conference (ESWC), 2013.

[SePublica Workshop 2014] Benedikt Kämpgen and David Riepl and Jochen Klinger. SMART Research using Linked Data – Sharing Research Data for Integrated Water Resources Management in the Lower Jordan Valley. In ESWC Workshop on Semantic Publishing (SePublica), 2014.

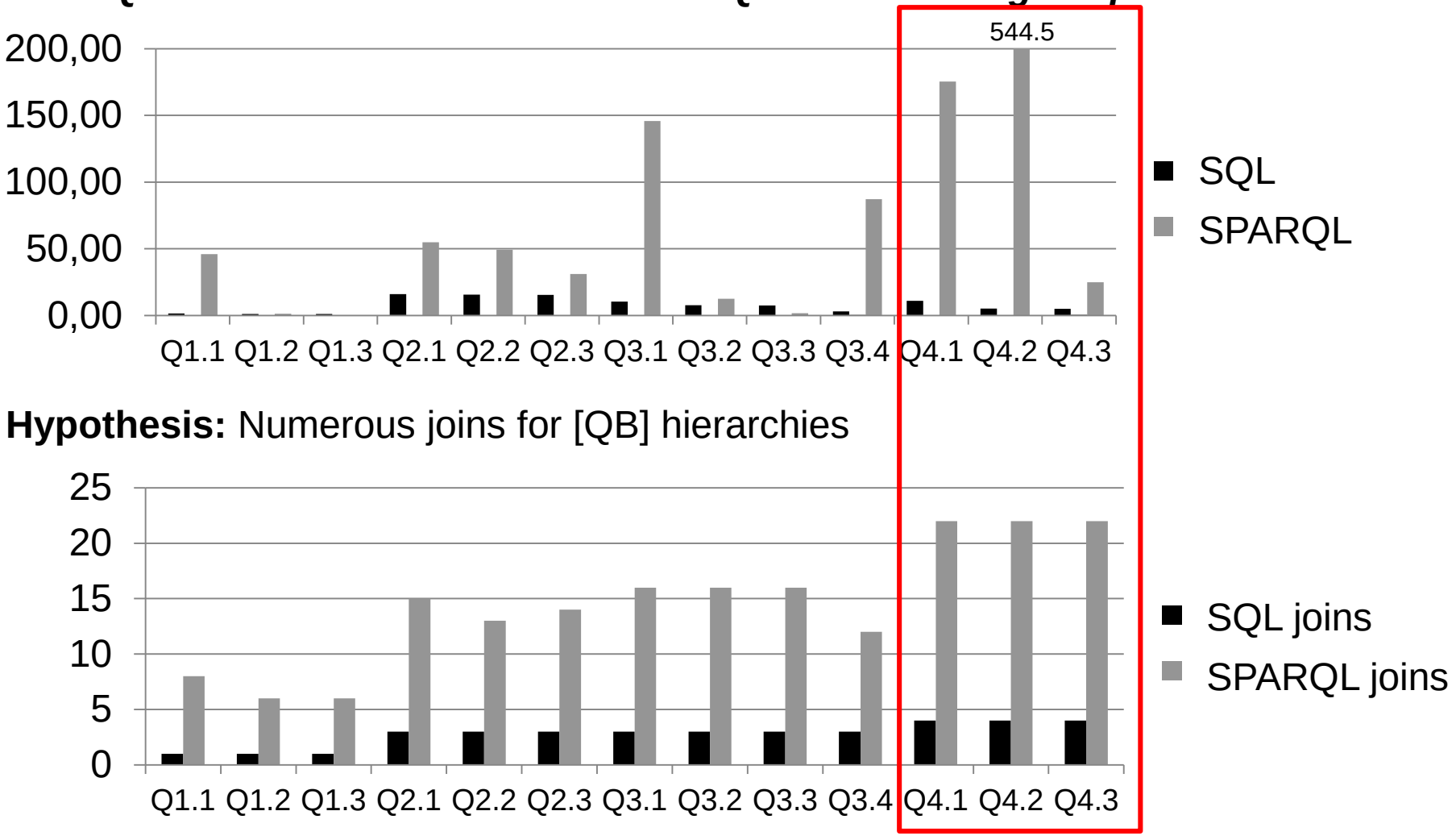
[ESWC In-Use 2014] Benedikt Kämpgen and Tobias Weller and Séan O'Riain and Craig Weber and Andreas Harth. Accepting the XBRL Challenge with Linked Data for Financial Data Integration. In 11th Extended Semantic Web Conference (ESWC), 2014.

[ESWC Demo 2014] Benedikt Kämpgen and Andreas Harth. OLAP4LD – A Framework for Building Analysis Applications over Governmental Statistics. In 11th Extended Semantic Web Conference (ESWC) Satellite Events, 2014.

[EKAW Research 2014] Benedikt Kämpgen and Steffen Stadtmüller and Andreas Harth. Querying the Global Cube: Integration of Multidimensional Datasets from the Web. In 19th International Conference on Knowledge Engineering and Knowledge Management (EKAW), 2014.

Backup: Evaluation Results (1)

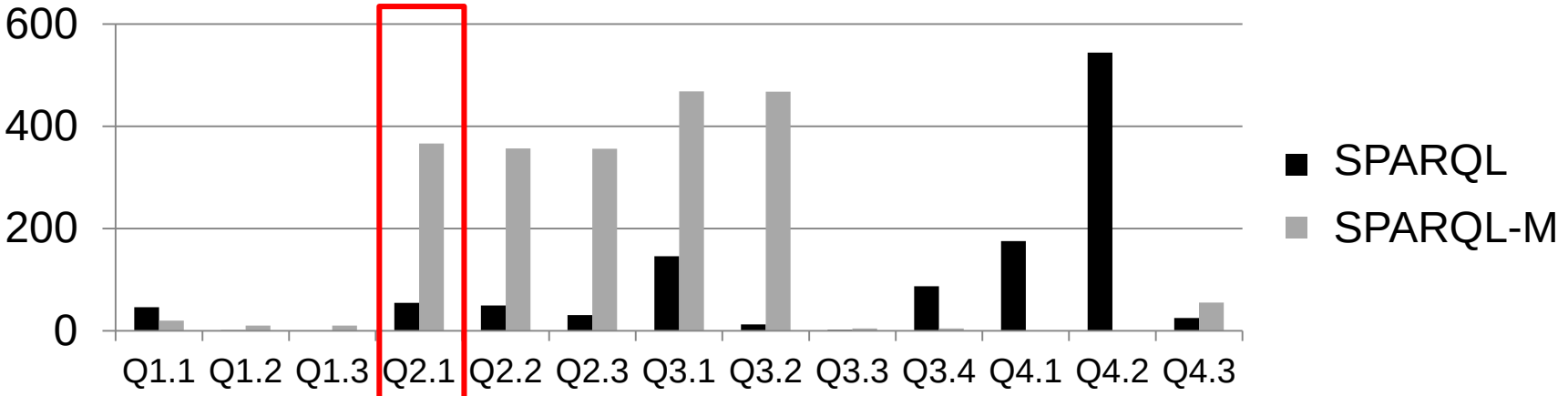
SPARQL overall 12 times slower than SQL for executing all queries?



Hypothesis: Numerous joins for [QB] hierarchies

Backup: Evaluation Results (2)

SPARQL-M overall 2 times slower than SPARQL



Hypothesis: SPARQL-M scans all lineorders although view contains fraction