Karlsruhe Institute of Technology

supported by:

Share-PSI 2.0

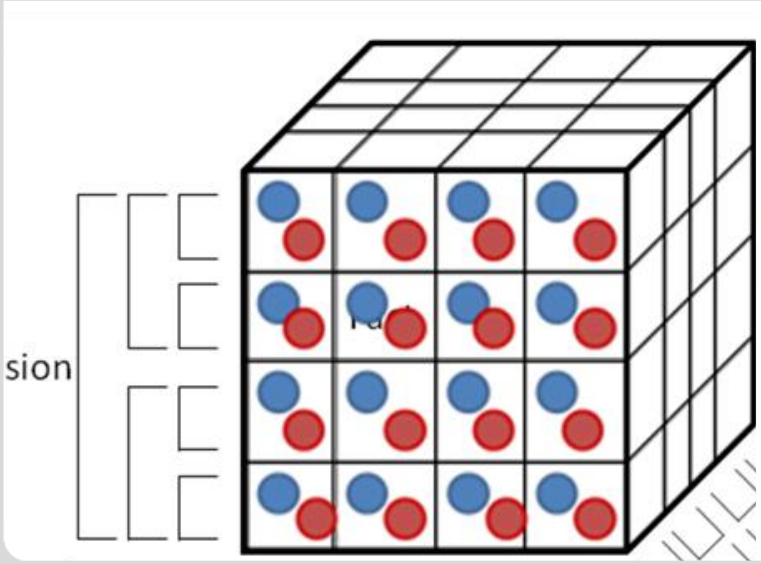
software campus

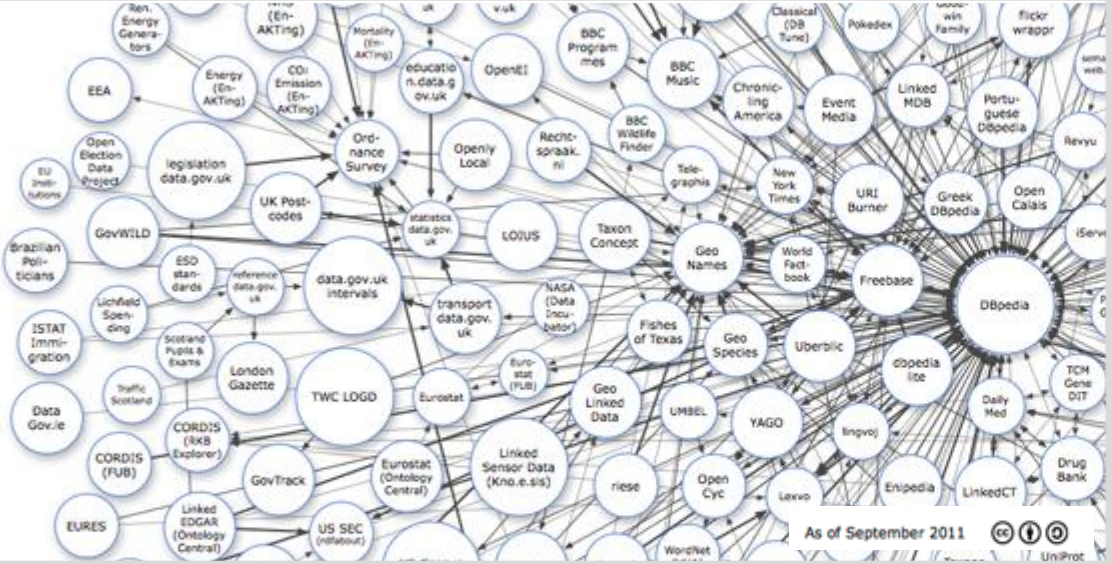
Querying the Global Cube: Integration of Multidimensional Datasets from the Web

Benedikt Kämpgen, Steffen Stadtmüller, Andreas Harth

EKAW 2014

Institute of Applied Informatics and Formal Description Methods (AIFB)

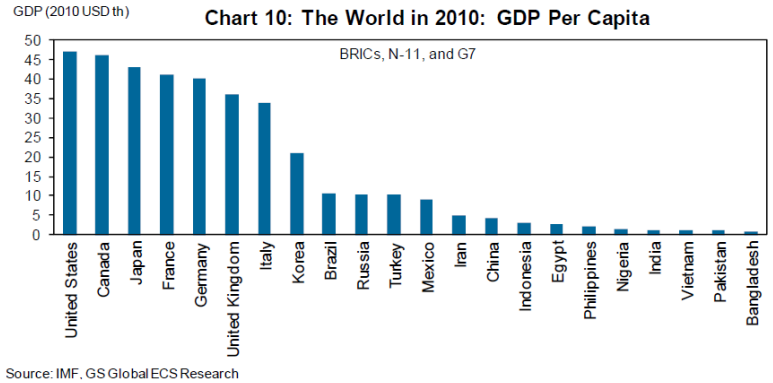


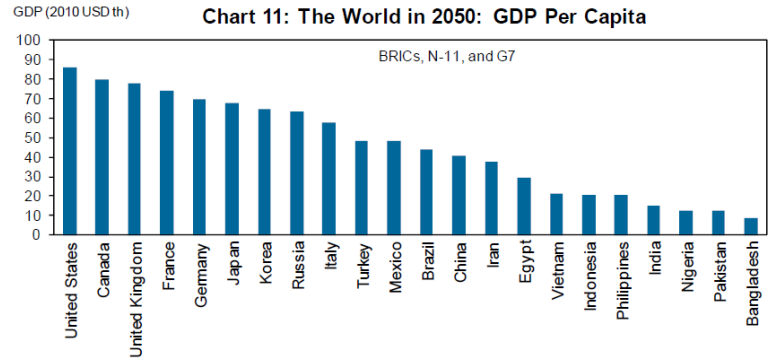


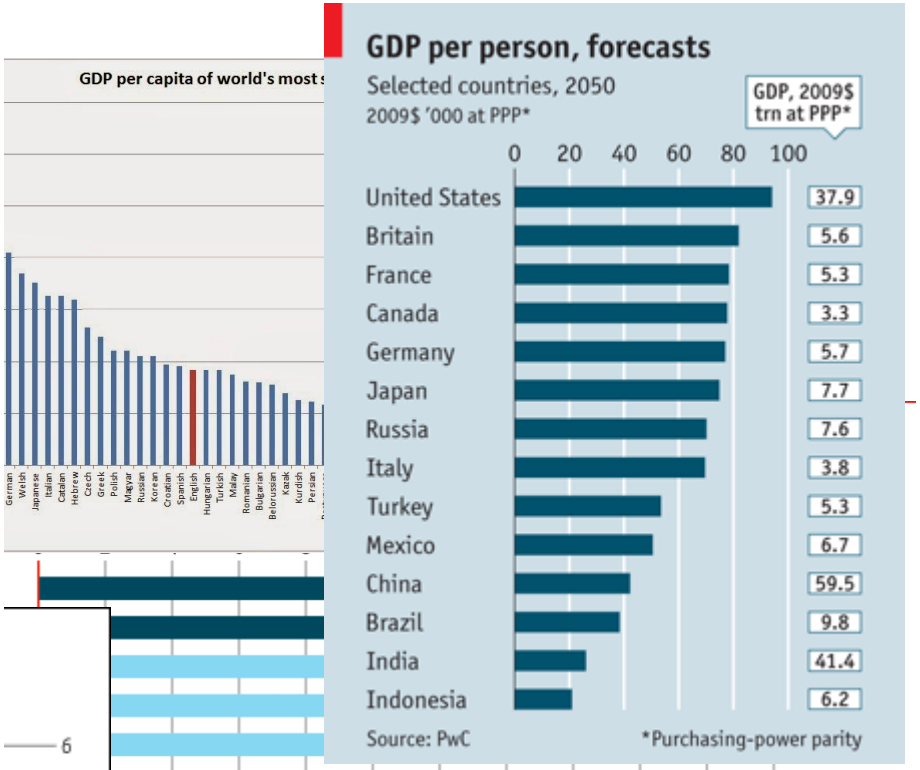
G8 Open Data Charter

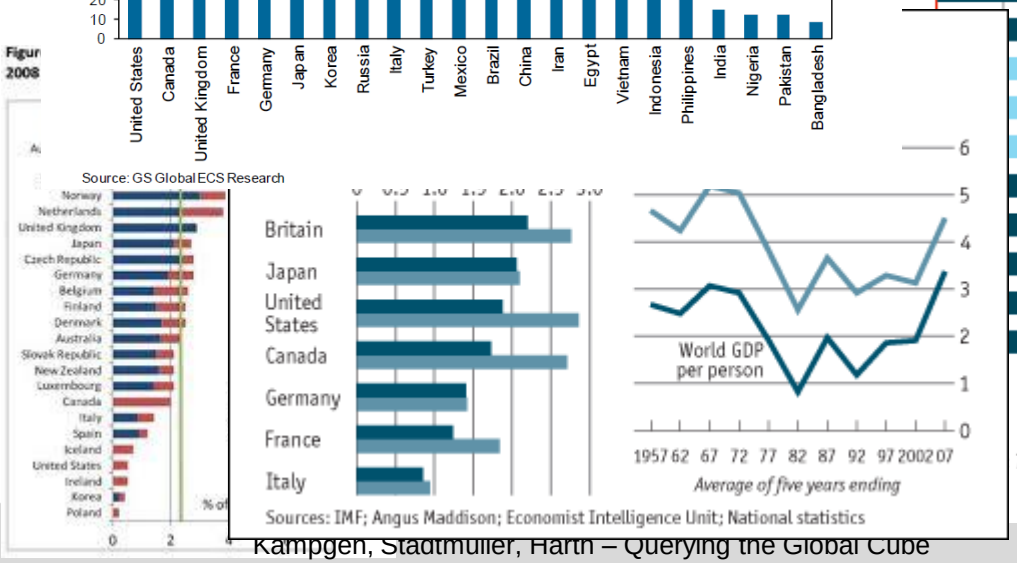
http://www.telegraph.co.uk/news/worldnews/g8/10128266/G8-Open-Data-Charter-why-it-matters.html

GDP Per Capita









Requirements – GDP Per Capita Example

- Answered by datasets from the Web
- Confirmed by as many sources as possible

Problem:
Query in Terms of the Global Cube

Karlsruhe Institute of Technology

Data Structure:

Geo	Unit	Date	Indicna	Sex	...	Value
de, uk, ...	eur, eur_hab, ...	2001, ...	b1g, ngdp,...	f, m,...	...	**Decimal**

Query:

Geo	Unit	Date	Indicna	Sex	...	Value
uk	eur_hab	2010	ngdph	ALL	ALL	**?**

Institute of Applied Informatics and Formal Description Methods (AIFB)

Problem:
Query in Terms of the Global Cube

Data Structure:

Geo	Unit	Date	Indicna	Sex	...	Value
de, uk, ...	mio_eur, eur_hab, ...	2001, ...	b1g, d21_m_d31, ngdp,...	f, m, t,...	...	**Decimal**

Data:

?

Query:

Geo	Unit	Date	Indicna	Sex	...	Value
uk	eur_hab	2010	ngdph	ALL	ALL	**?**

Institute of Applied Informatics and Formal Description Methods (AIFB)

Outline

- Global Cube Definition
- Reducing Heterogeneity in Global Cube
- Analysis of Size of Global Cube

Institute of Applied Informatics and Formal Description Methods (AIFB)

Outline

- **Global Cube Definition**
- Reducing Heterogeneity in Global Cube
- Analysis of Size of Global Cube

Institute of Applied Informatics and Formal Description Methods (AIFB)

Real-World Dataset
Eurostat GDP Components (readable)

Karlsruhe Institute of Technology

Data Structure:

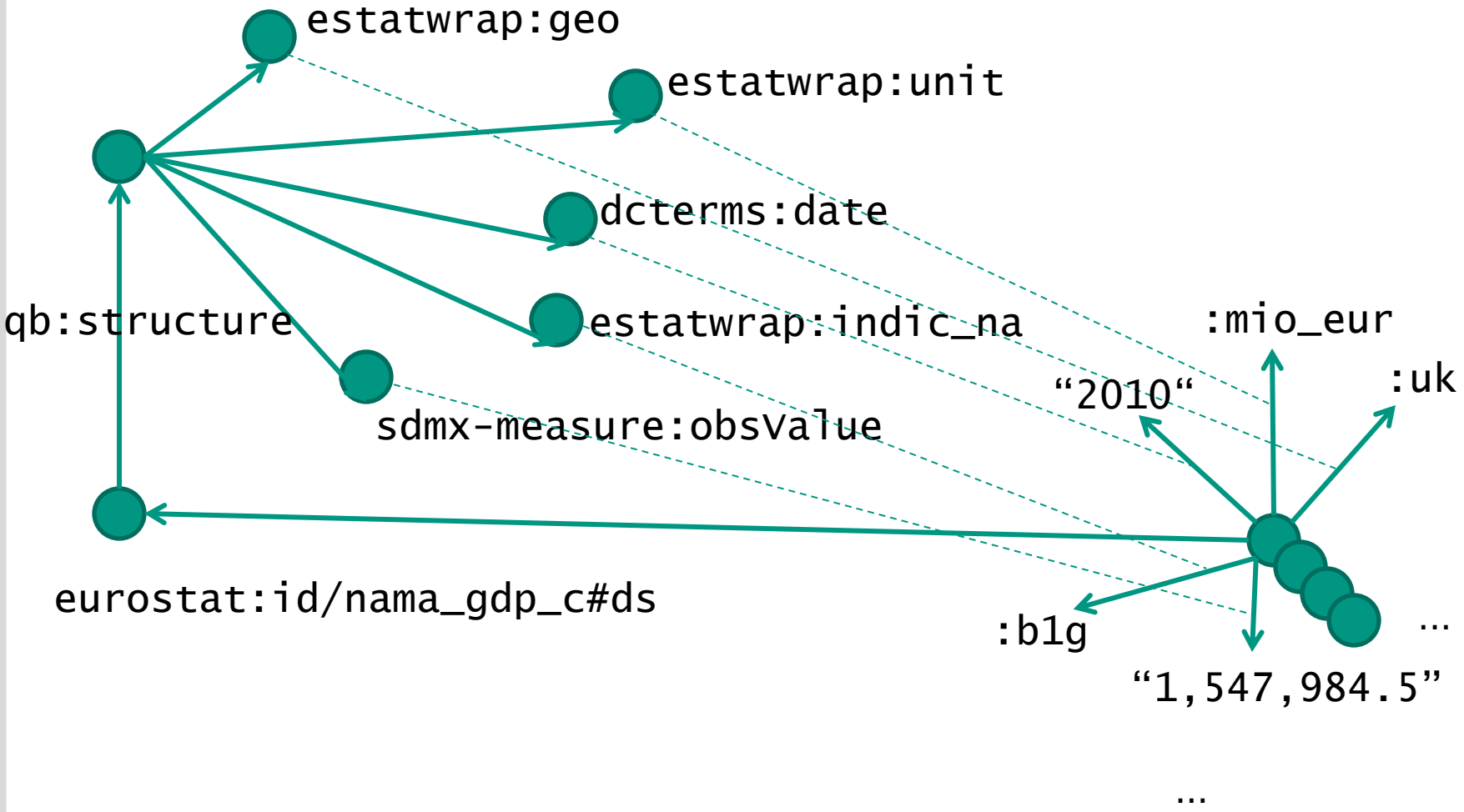
Geo	Unit	Date	Indicna	Value
de, uk, ...	mio_eur	2001, ...	b1g, d21_m_d31, ...	**Decimal**

Data:

Geo	Unit	Date	Indicna	Value
uk	mio_eur	2010	b1g	**1,547,984.5**
...	...	...	...	**...**

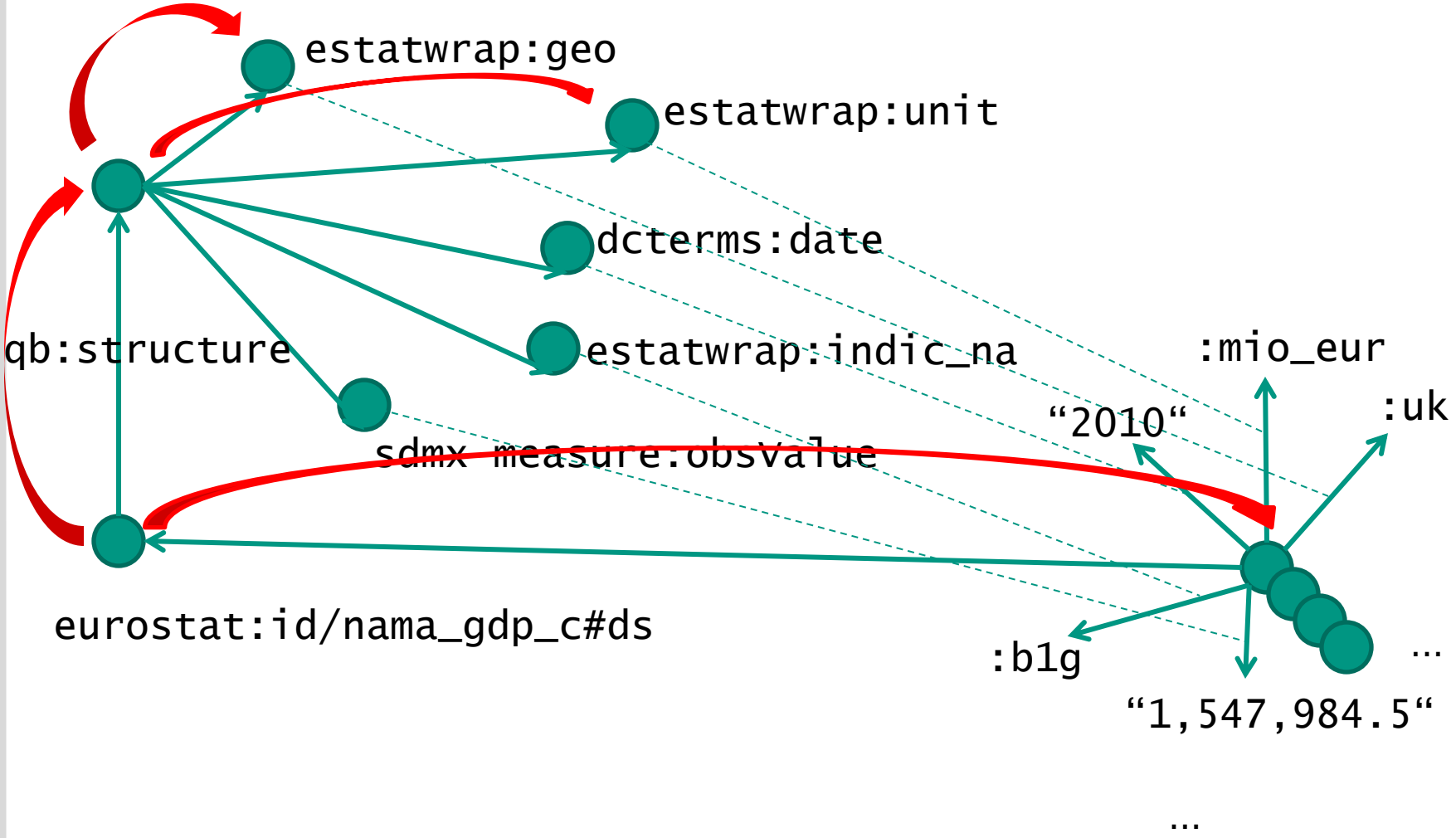
Institute of Applied Informatics and Formal Description Methods (AIFB)

Real-World Dataset
Eurostat GDP Components (Linked Data)



Linked Data URIs serving RDF, e.g., eurostat:id/nama_gdp_c#ds = http://estatwrap.ontologycentral.com/id/nama_gdp_c#ds

Real-World Dataset
Eurostat GDP Components (Linked Data)



Linked Data URIs serving RDF, e.g., eurostat:id/nama_gdp_c#ds = http://estatwrap.ontologycentral.com/id/nama_gdp_c#ds

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Cube\Dimension	Geo	Unit	Date	Indicna	Sex	Age	Value
gdp-components	de, uk, ...	mio_eur, ...	2001, ...	b1g, d21_m_d31	-	-	**Decimal**
population	de, uk, ...	-	2001, ...	-	f, m, t	y18,..., total	**Decimal**
gdp-per-capita	de, uk, ...	eur_hab, ...	2001, ...	ngdph	-	-	**Decimal**
...	...	...	...	...	...	...	**...**

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Cube\Dimension	Geo	Unit	Date	Indicna	Sex	Age	Value
gdp-components	de, uk, ...	mio_eur, ...	2001, ...	b1g, d21_m_d31	-	-	**Decimal**
population	de, uk, ...	-	2001, ...	-	f, m, t	y18,..., total	**Decimal**
gdp-per-capita	de, uk, ...	eur_hab, ...	2001, ...	ngdph	-	-	**Decimal**
...	...	...	...	...	...	...	**...**

- Example Data Sources
  - Eurostat as Linked Data (>5,000 datasets) http://eurostat.linked-statistics.org/
  - The World Bank as Linked Data (>8,000 datasets) http://worldbank.270a.info/
  - U.S. SEC as Linked Data (>1,000 datasets) http://edgarwrap.ontologycentral.com/
  - OECD as Linked Data (>100 datasets) http://oecd.270a.info/
  - ...

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Cube\Dimension	Geo	Unit	Date	Indicna	Sex	Age	Value
gdp-components	de, uk, ...	mio_eur, ...	2001, ...	b1g, d21_m_d31	-	-	Decimal
population	de, uk, ...	-	2001, ...	-	f, m, t	y18,..., total	Decimal
gdp-per-capita	de, uk, ...	eur_hab, ...	2001, ...	ngdph	-	-	Decimal

Data:

Cube\Dimension	Geo	Unit	Date	Indicna	Sex	Age	Value
gdp-components	uk	mio_eur	2010	b1g	-	-	1547984
...	...	...	...	...	-	-	...
population	uk	-	2010	-	t	total	62510197
...	...	-	...	-	...	...	...
gdp-per-capita	uk	eur_hab	2010	ngdph	-	-	27800
...	...	...	...	...	-	-	

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Geo	Unit	Date	Indicna	Sex	Age	Value
de, uk, ...	mio_eur, eur_hab, ...	2001, ...	b1g, d21_m_d31, ngdph	f, m, t	y18,..., total	**Decimal**

Data:

Cube\Dimension	Geo	Unit	Date	Indicna	Sex	Age	Value
gdp-components	uk	mio_eur	2010	b1g	-	-	1547984
...	...	...	...	...	-	-	...
population	uk	-	2010	-	t	total	62510197
...	...	-	...	-	...	...	...
gdp-per-capita	uk	eur_hab	2010	ngdph	-	-	27800
...	...	...	...	...	-	-	

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Geo	Unit	Date	Indicna	Sex	Age	Value
de, uk, ...	mio_eur, eur_hab, ...	2001, ...	b1g, d21_m_d31, ngdph	f, m, t	y18,..., total	Decimal

Data:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	mio_eur	2010	b1g	ALL	ALL	1547984
...	...	...	...	ALL	ALL	...
uk	ALL	2010	ALL	t	total	62510197
...	ALL	...	ALL	...	...	...
uk	eur_hab	2010	ngdph	ALL	ALL	27800
...	...	...	...	ALL	ALL	

Global Cube Definition: GDP Per Capita Example

Karlsruhe Institute of Technology

Data Structures:

Geo	Unit	Date	Indicna	Sex	Age	Value
de, uk, ...	mio_eur, eur_hab, ...	2001, ...	b1g, d21_m_d31, ngdph	f, m, t	y18,..., total	Decimal

Data:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	mio_eur	2010	b1g	ALL	ALL	1547984
...	...	...	...	ALL	ALL	...
uk	ALL	2010	ALL	t	total	62510197
...	ALL	...	ALL	...	...	...
uk	eur_hab	2010	ngdph	ALL	ALL	27800
...	...	...	...	ALL	ALL	

Related Work

- We define a unified view over datasets from the Web.
  - Ontology matching approaches are less suitable for finding relationships between multidimensional datasets [Zapilko].
- We solve heterogeneities with conversion (combination) relationships between datasets in Linked Data.
  - Conversion functions and combinations have been defined in a relational setting [Siegel, Calvanese] but not in Linked Data.
- We estimate the size of Global Cube with number of derived datasets.
  - Semantic integration systems [Bressan, Ambite] focus on fast query execution based on input-/output descriptions.

Outline

- Global Cube Definition
- Related Work
- **Reducing Heterogeneity in Global Cube**
- Analysis of Size of Global Cube

Heterogeneity Problems

Data:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	mio_eur	2010	b1g	ALL	ALL	1547984
...	...	...	...	ALL	ALL	...
uk	ALL	2010	ALL	t	total	62510197
...	ALL	...	ALL	...	...	...
uk	eur_hab	2010	ngdph	ALL	ALL	27800
...	...	...	...	ALL	ALL	

Query:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	eur_hab	2010	ngdph	ALL	ALL	?

Institute of Applied Informatics and Formal Description Methods (AIFB)

Heterogeneity Problems

Data:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	mio_eur	2010	b1g	ALL	ALL	1547984
...	...	...	...	ALL	ALL	...
uk	ALL	2010	ALL	t	total	62510197
...	ALL	...	ALL	...	...	...
uk	eur_hab	2010	ngdph	ALL	ALL	27800
...	...	...	...	ALL	ALL	

Query:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	eur_hab	2010	ngdph	ALL	ALL	?

Institute of Applied Informatics and Formal Description Methods (AIFB)

Heterogeneity Problems

Data:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	mio_eur	2010	b1g	ALL	ALL	1547984
...	...	...	...	ALL	ALL	...
uk	ALL	2010	ALL	t	total	62510197
...	ALL	...	ALL	...	...	...
uk	eur_hab	2010	ngdph	ALL	ALL	27800
...	...	...	...	ALL	ALL	

Query:

Geo	Unit	Date	Indicna	Sex	Age	Value
uk	eur_hab	2010	ngdph	ALL	ALL	?

Institute of Applied Informatics and Formal Description Methods (AIFB)

Solution – Complex Relationships: GDP Per Capita Example

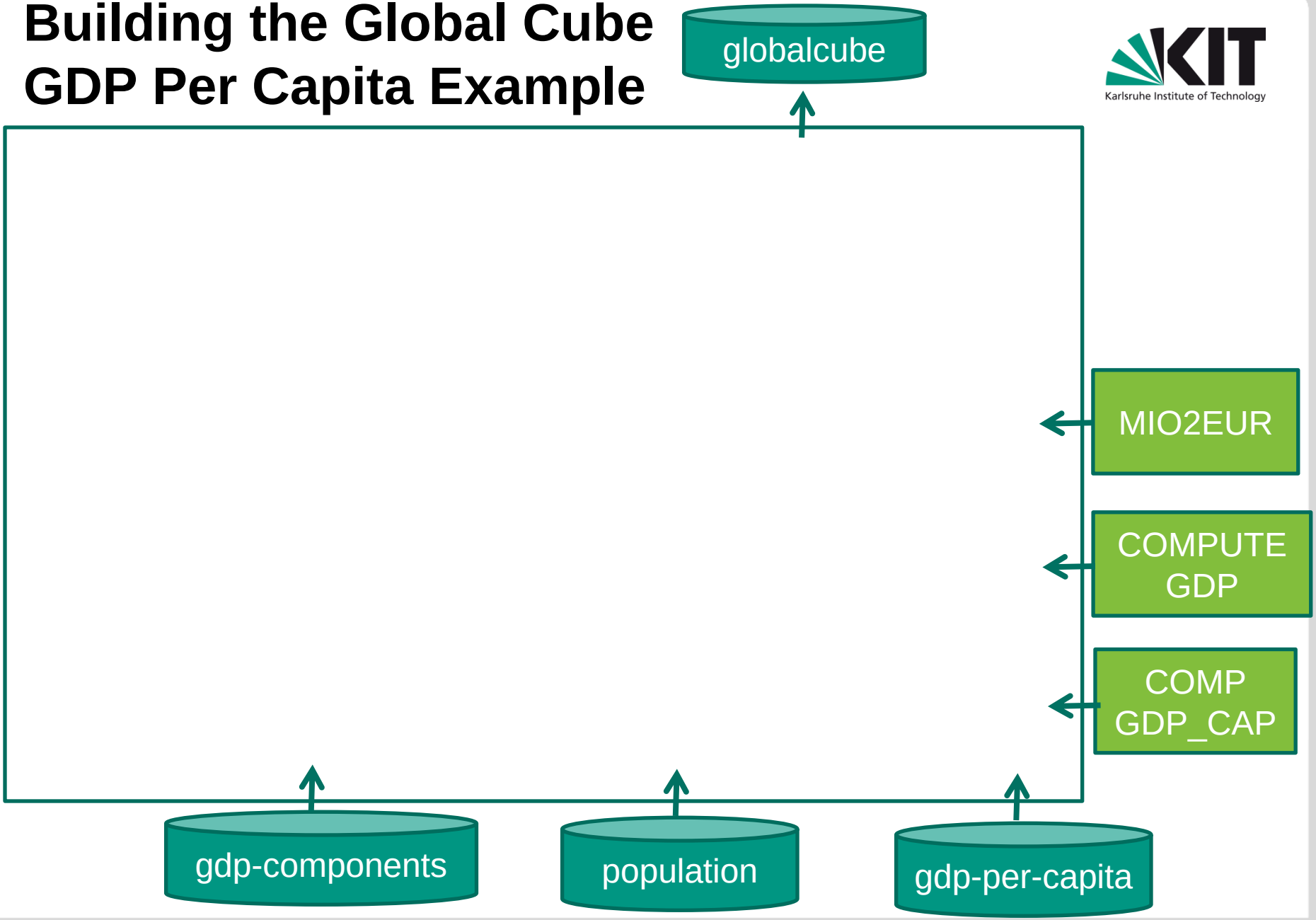
Conversion (and Merging) Relationships

MIO2EUR

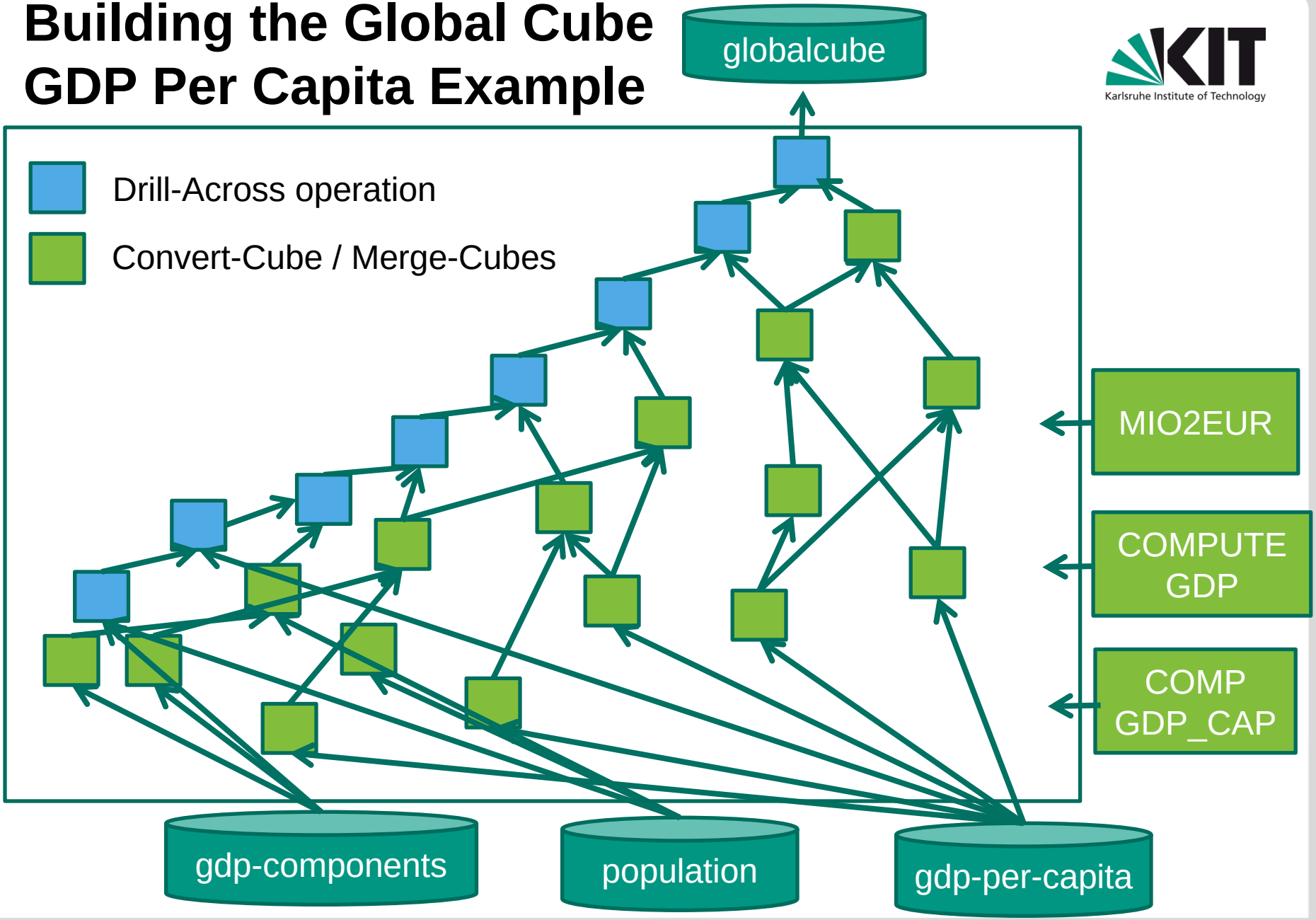
Unit	Value
mio_eur	value
eur	1,000,000 * value

Relationships can be implemented as Convert-Cube (Merge-Cubes) operation using Datalog (and SPARQL)

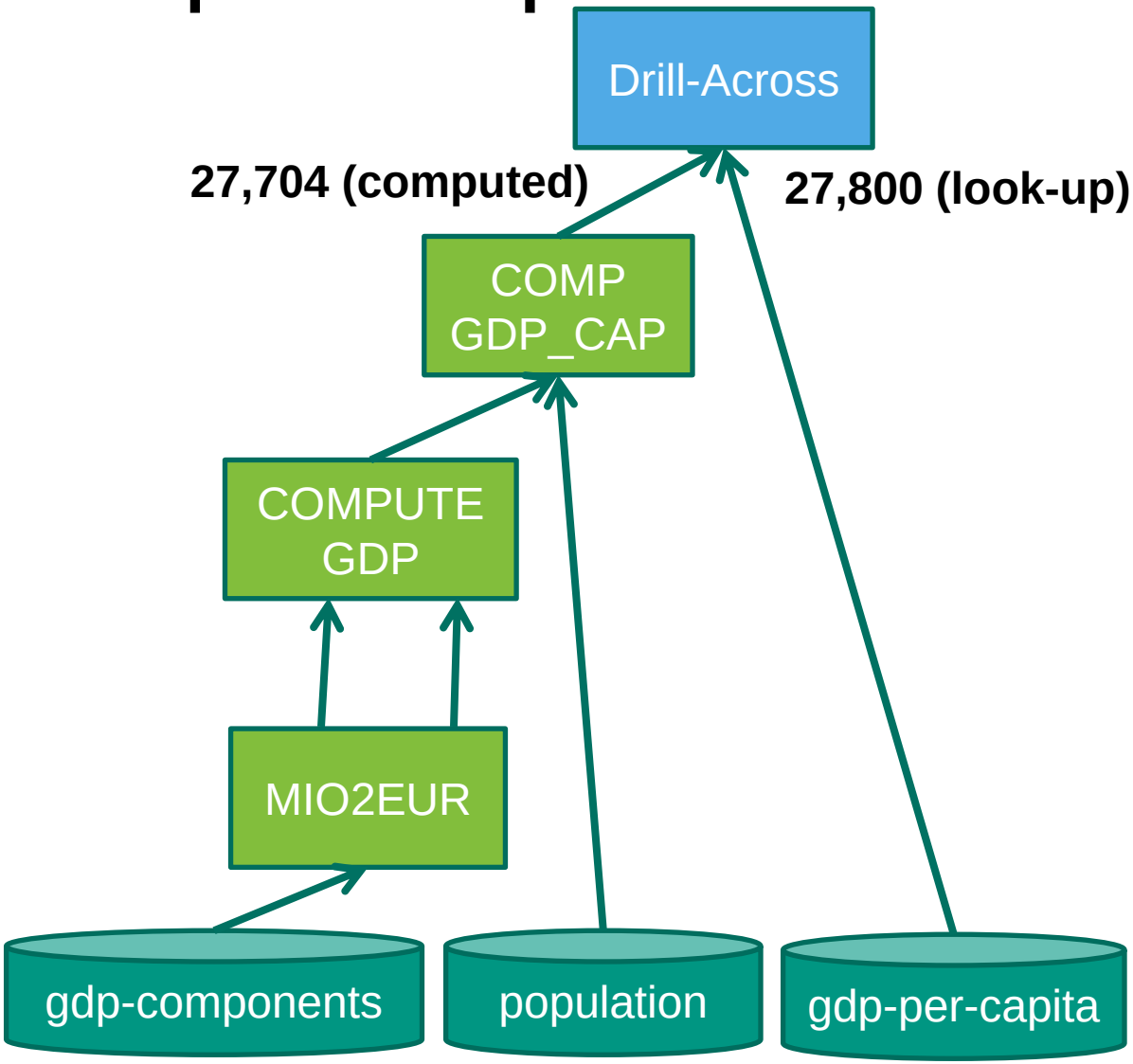
Building the Global Cube
GDP Per Capita Example



Building the Global Cube
GDP Per Capita Example



Building the Global Cube
GDP Per Capita Example



Outline

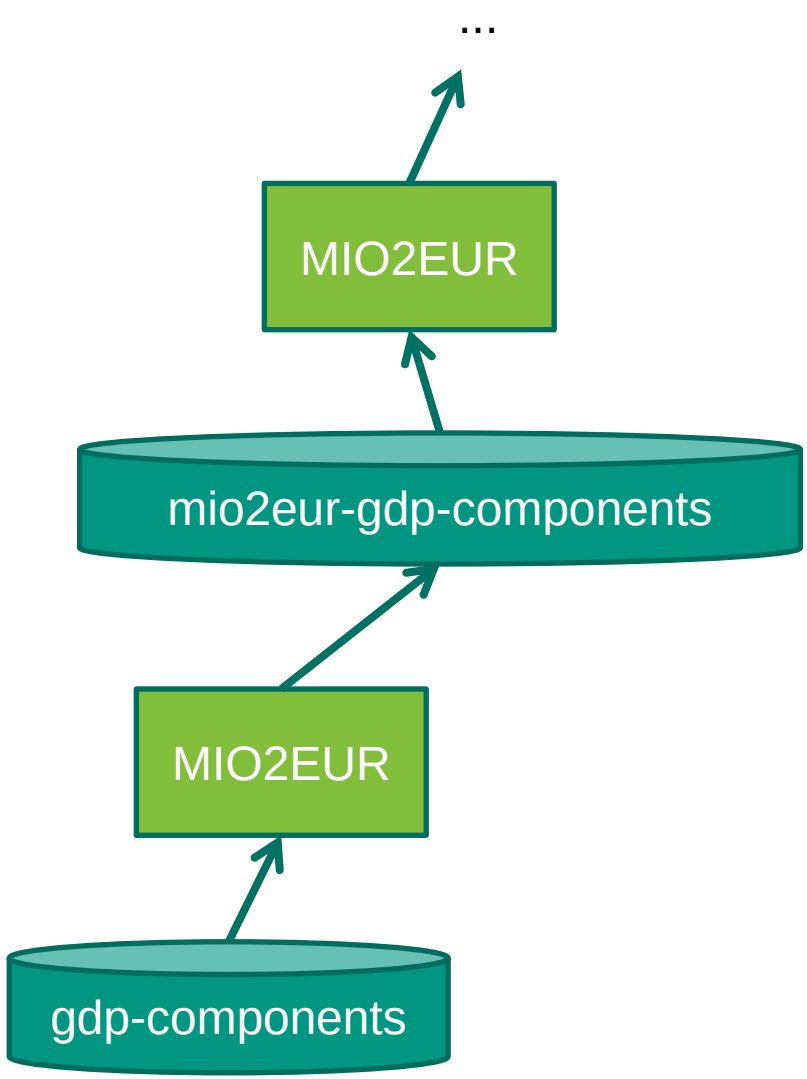
- Global Cube Definition
- Reducing Heterogeneity in Global Cube
- **Analysis of Size of Global Cube**

Institute of Applied Informatics and Formal Description Methods (AIFB)

How to restrict the number of derived datasets?

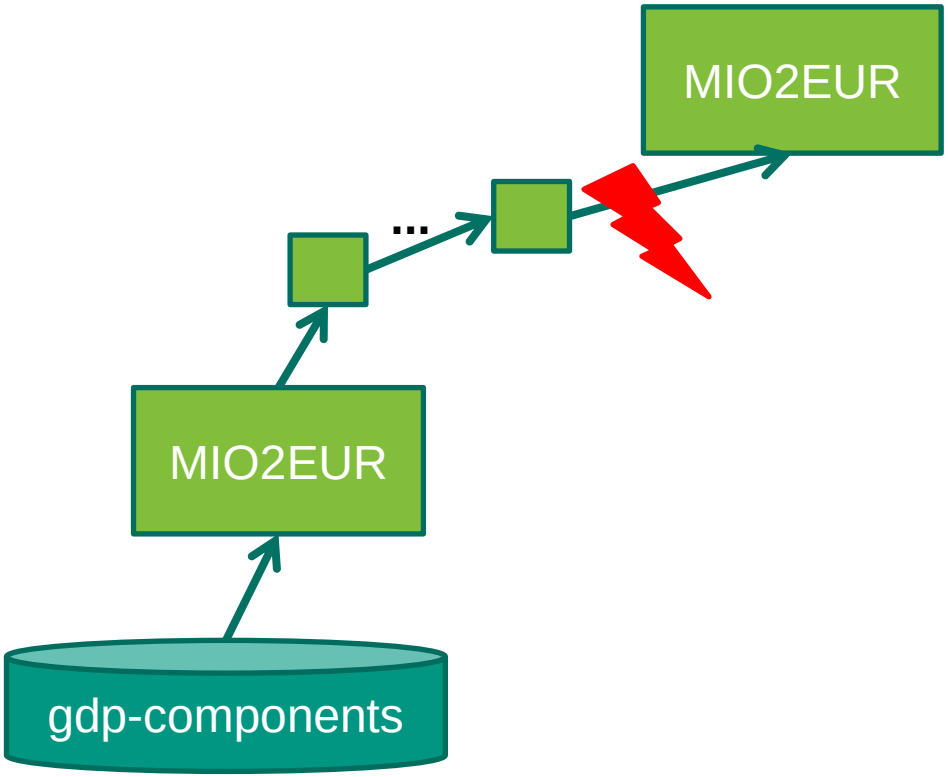
1) No restrictions

Unlimited



How to restrict the possible number of conversion and merging operations? (2)

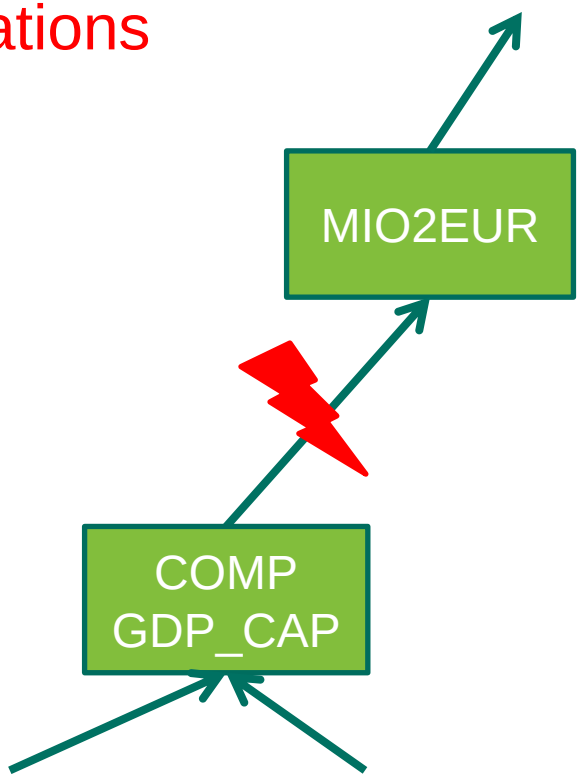
2) No repeated applications / cycles



Estimated 3,017,172 derived datasets

How to restrict the possible number of conversion and merging operations? (3)

3) Only matching dimension-member combinations



Estimated 54 derived datasets

Conclusions

- Interesting numeric datasets in **Linked Data**
  - **Global Cube** as a unified view over all datasets
  - Reducing heterogeneity with **conversion (merging) relationships**

- Reliable statistical indicators from the Web if
  - we can find relationships „in the wild“
  - we can materialise the Global Cube
  - we find out why Open Government Data sources...

A different topic...

27,704 (computed)

27,800 (look-up)

eurostat

Thanks!

27,704 (computed)

27,800 (look-up)

eurostat

References

- Ambite, J. L., & Kapoor, D. (2007). Automatically Composing Data Workflows with Relational Descriptions and Shim Services. In ISWC.
- Bressan, S., & Goh, C. (1997). Semantic Integration of Disparate Information Sources over the Internet Using Constraint Propagation Semantic Integration of Disparate Information Sources over the Internet using Constraint Propagation, (August).
- Zapilko, B., & Mathiak, B. (2014). Object Property Matching Utilizing the Overlap between Imported Ontologies, 737–751.
- Siegel, M., Sciore, E., & Rosenthal, A. (1994). Using semantic values to facilitate interoperability among heterogeneous information systems.
- Calvanese, D., De Giacomo, G., Lenzerini, M., Nardi, D., & Rosati, R. (2001). Data Integration in Data Warehousing. International Journal of Cooperative Information Systems.