Semantics in RDF Data Cubes

Speaker: Benedikt Kämpgen (FZI)

Location: WU Vienna, Group of Axel Polleres

Date: 22.07.2015

FZI FORSCHUNGSZENTRUM INFORMATIK

About FZI

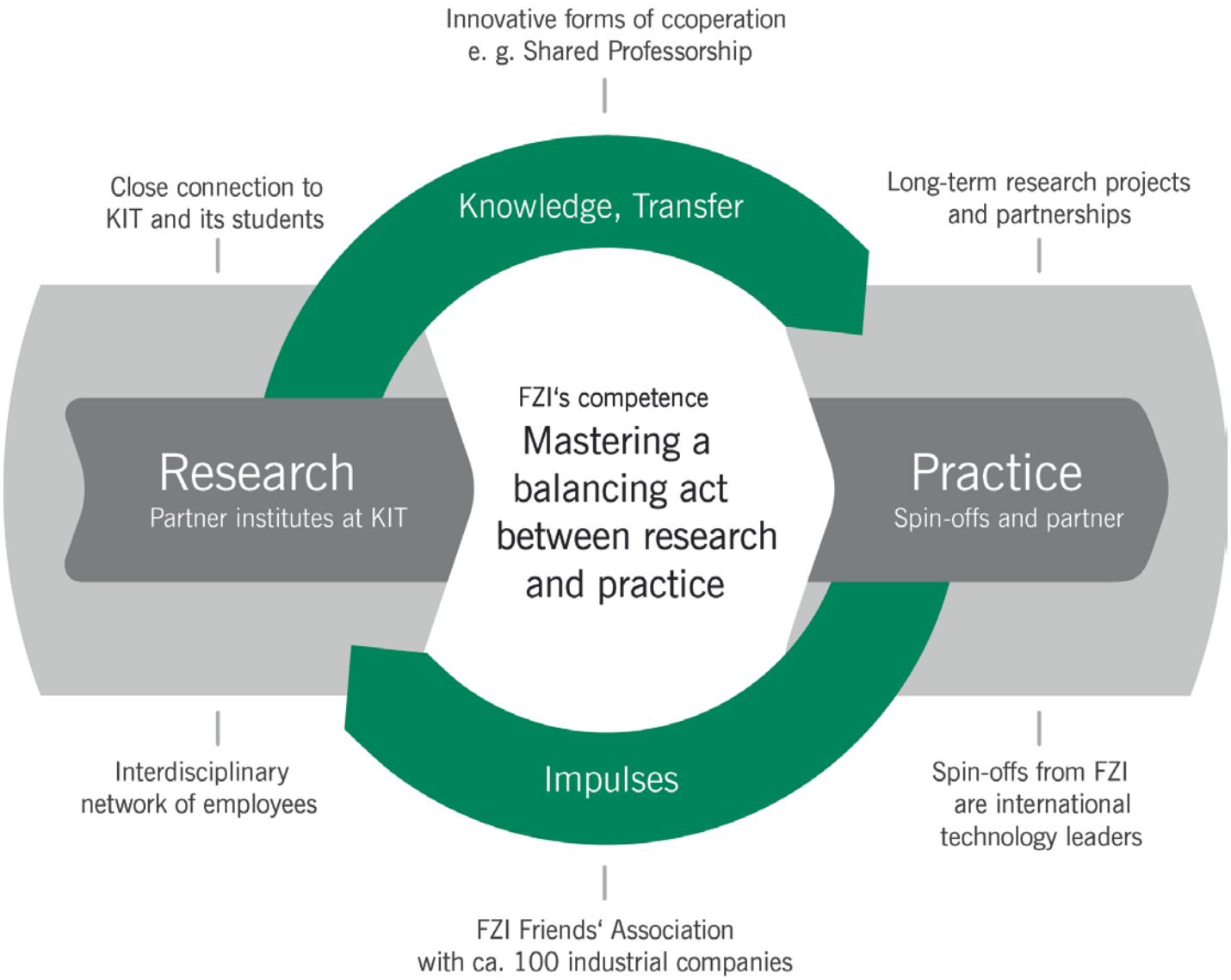
- Foundation of the federal state Baden-Wuerttemberg
- Member of Innovationsallianz Baden-Württemberg and Technologieregion Karlsruhe
- Innovation hub in Baden-Wuerttemberg for information technology
- Innovation partner of the Karlsruhe Institute of Technology
- Medium-sized provider of research services

Figures and Facts 2014

- Overall budget: 14 M €
- Researchers: 135
- More than 200 projects – commissioned directly from industry and through publicly funded joint projects
- Certified Quality Management System according to ISO 9001
- Accredited PROFIBUS Competence Center and Test Lab
- Accredited KNX Test Lab

LGA InterCert
Zertifiziertes QM-System
DIN EN ISO 9001:2008

FZI's formula for success: a strong network



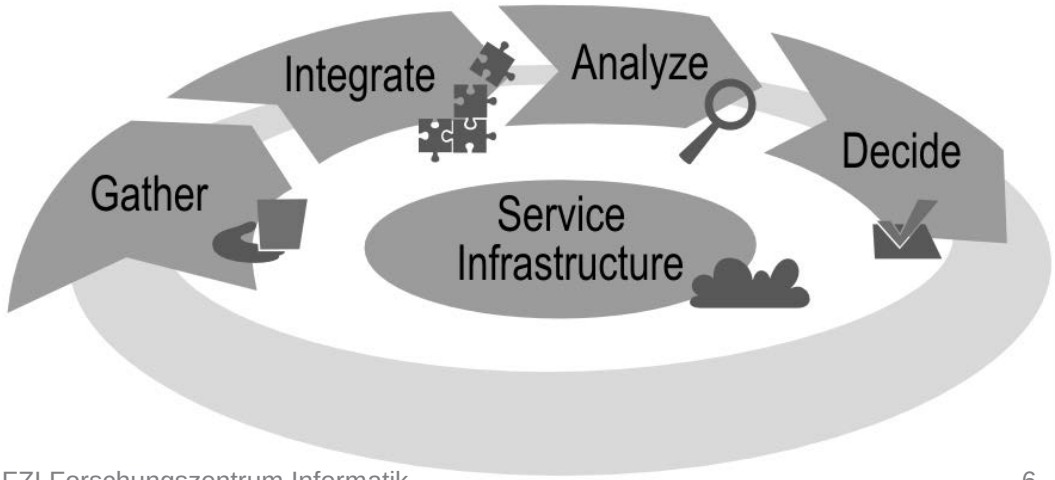
Our strength: Interdisciplinary research

Information Process Engineering (IPE) Division

- Novel methodologies and technologies for continuous information-driven entrepreneurial decisions …
  - to immediately **respond to changes**
  - to utilize **huge data volumes** together with the **collective intelligence** of an organization‘s employees
  - to reliably **discover coherences** in continuously growing data volumes
  - to **systematically combine** and utilize information on the basis of mathematical and economic models
  - to **initiate purposeful actions** at the right time.




WIM Group
Knowledge Management Based on the Web of Things



WIM

Knowledge Management in the Web of Things

Rudi Studer

Stefan Zander

Ljiljana Stojanovic

Benedikt Kämpgen

Dominik Riemer

Ignacio Traverso Ribon

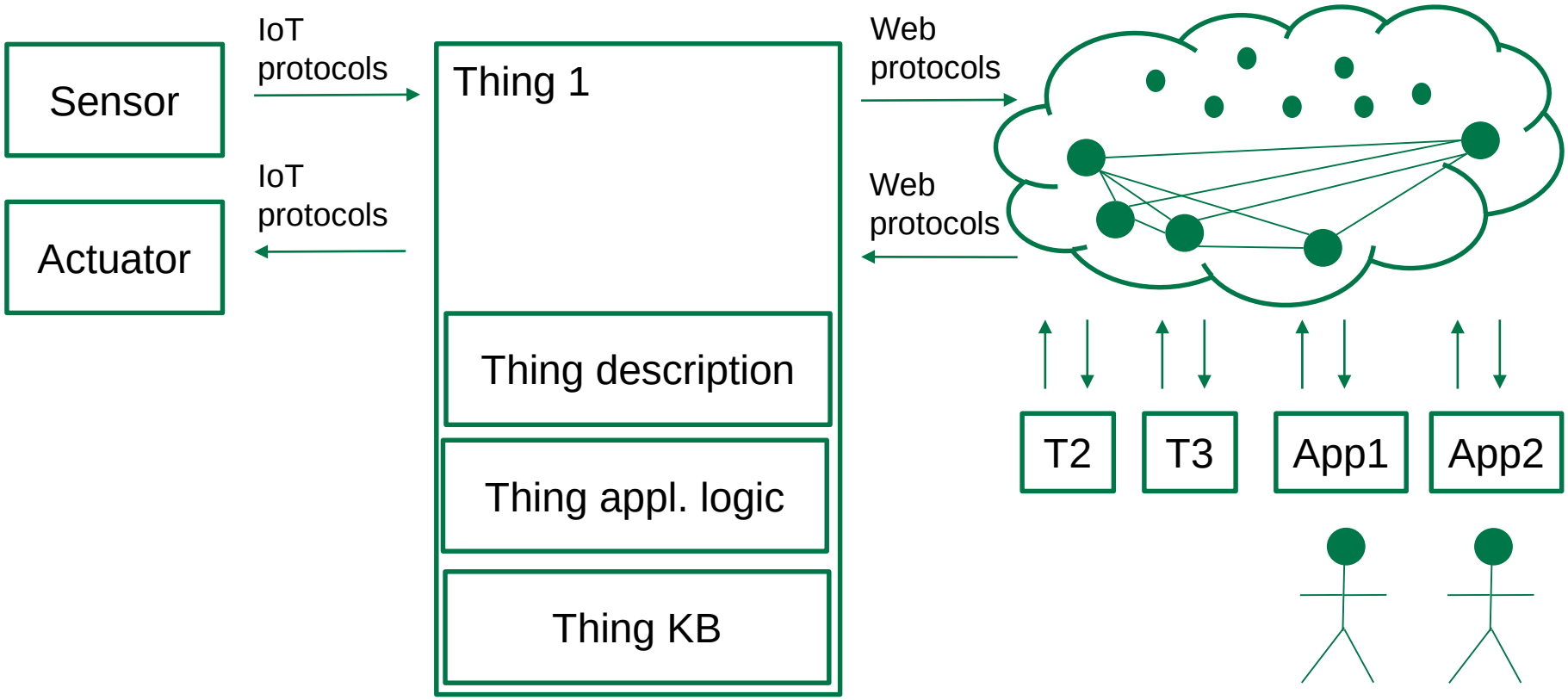
Nicole Merkle

Suad Sejdovic

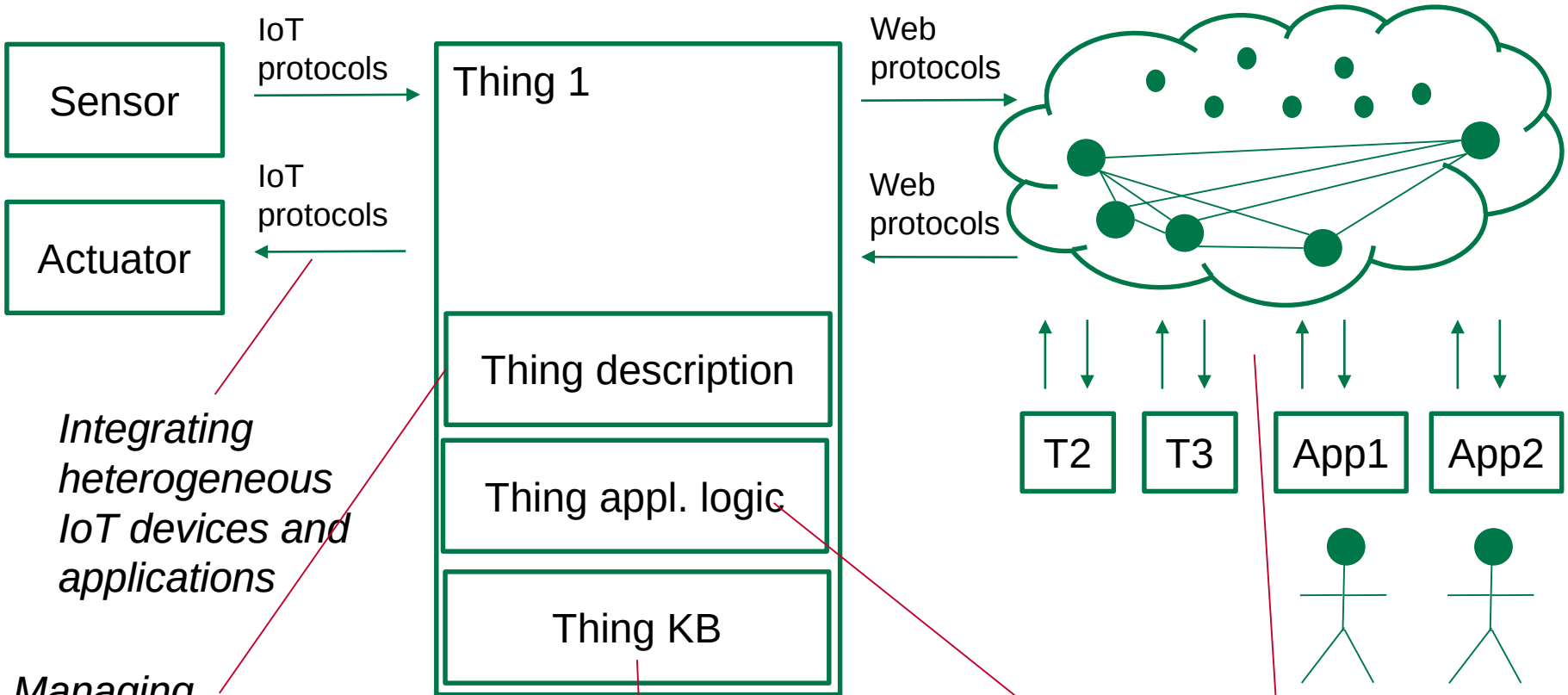
Nadia Ahmed

Matthias Frank

Web of Things



Web of Things: Challenges



Integrating heterogeneous IoT devices and applications

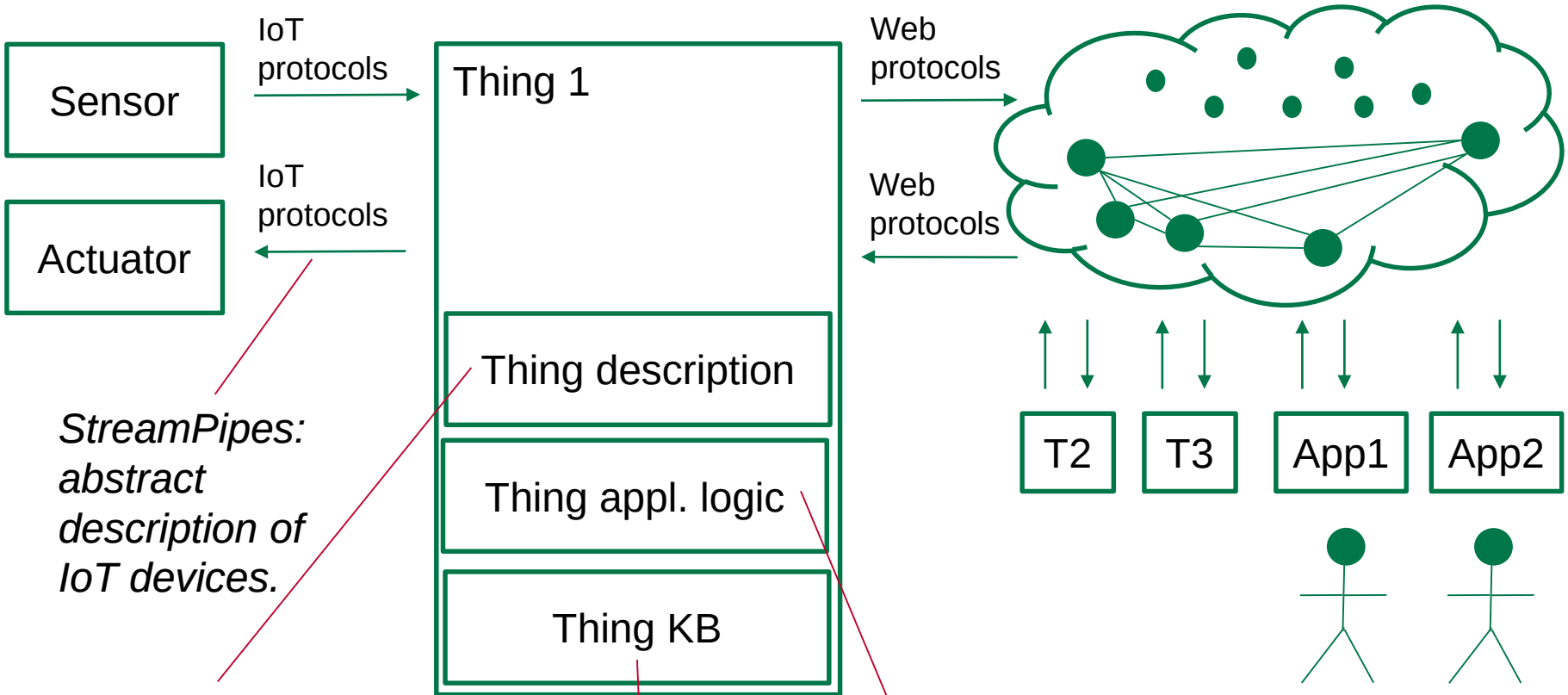
Managing heterogeneous and dynamic thing descriptions

Reasoning over arbitrary background information

Near-real-time reaction on streams of data.

Discovering things, ensuring data privacy

Web of Things: Solutions



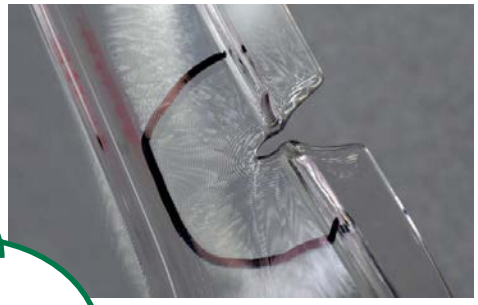
StreamPipes: abstract description of IoT devices.

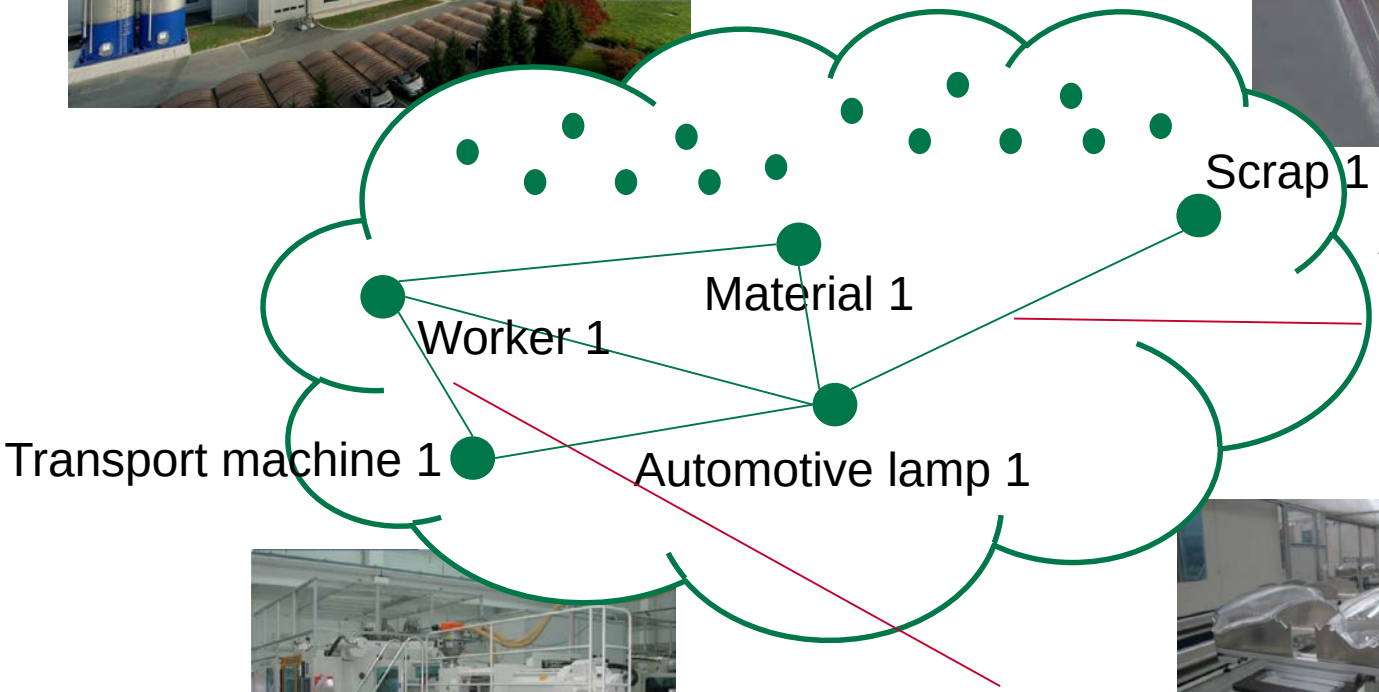
Semantic Sensor Network ontology, RDF Data Cube vocabulary, ReApp ontology, Linked APIs

openAAL: Context management with OWL reasoner.

ETALIS, Linked-Data-Fu: Rule-based interaction and complex event processing languages

Industry 4.0: EU Project ProaSense

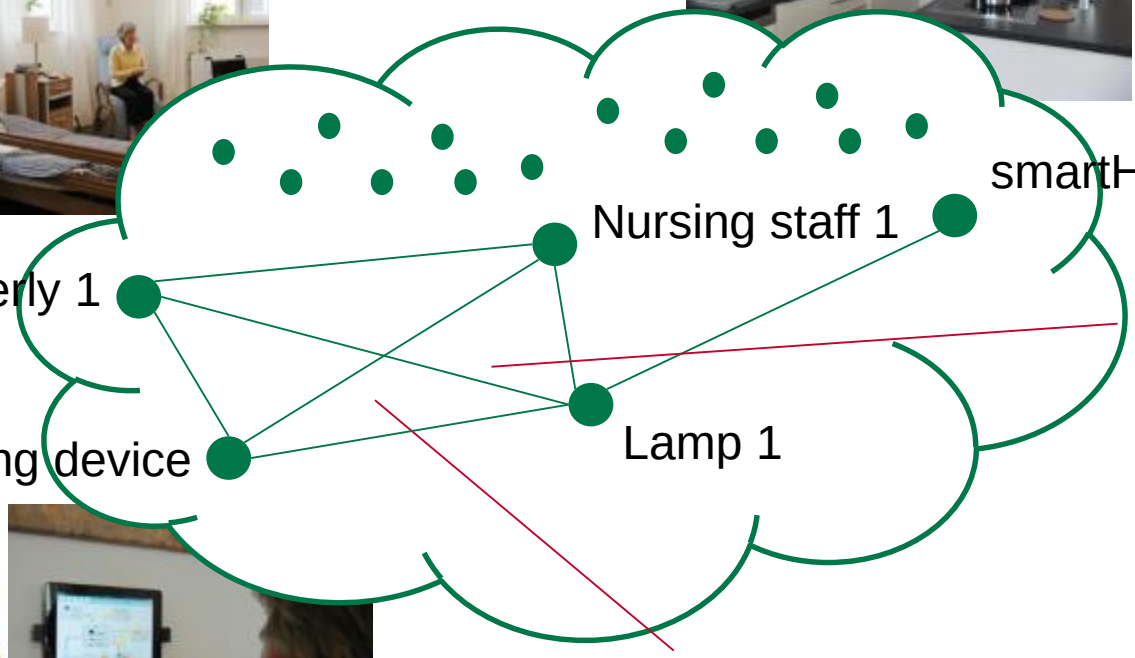




Automotive lamp brought by Transport machine 1 increases scrap rate

Transport machine communicates problems to right person

Ambient Assisted Living: BMBF Project AICASys



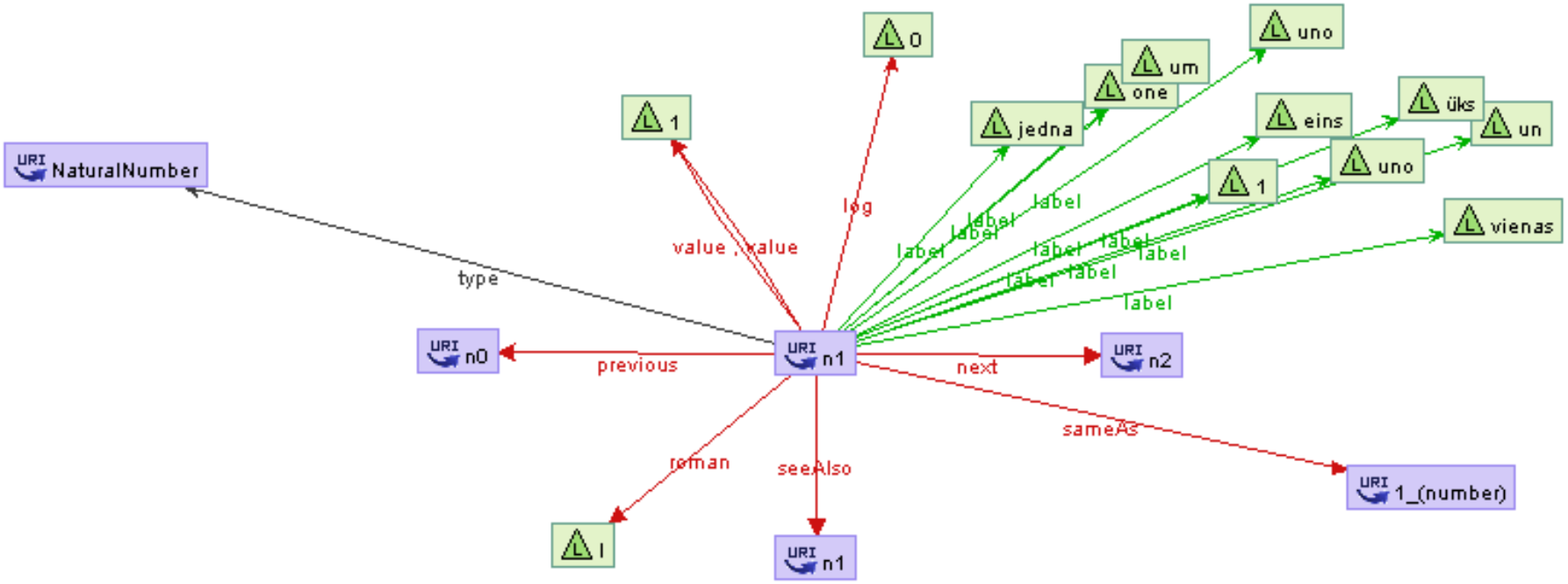
Behaviour linked to intention, e.g., gaze to lamp, to switching on lamp.

Recognising unconsciousness of user, eye tracking device can contact nursing staff

Outline

- FZI / WIM
- Semantics in RDF Data Cubes

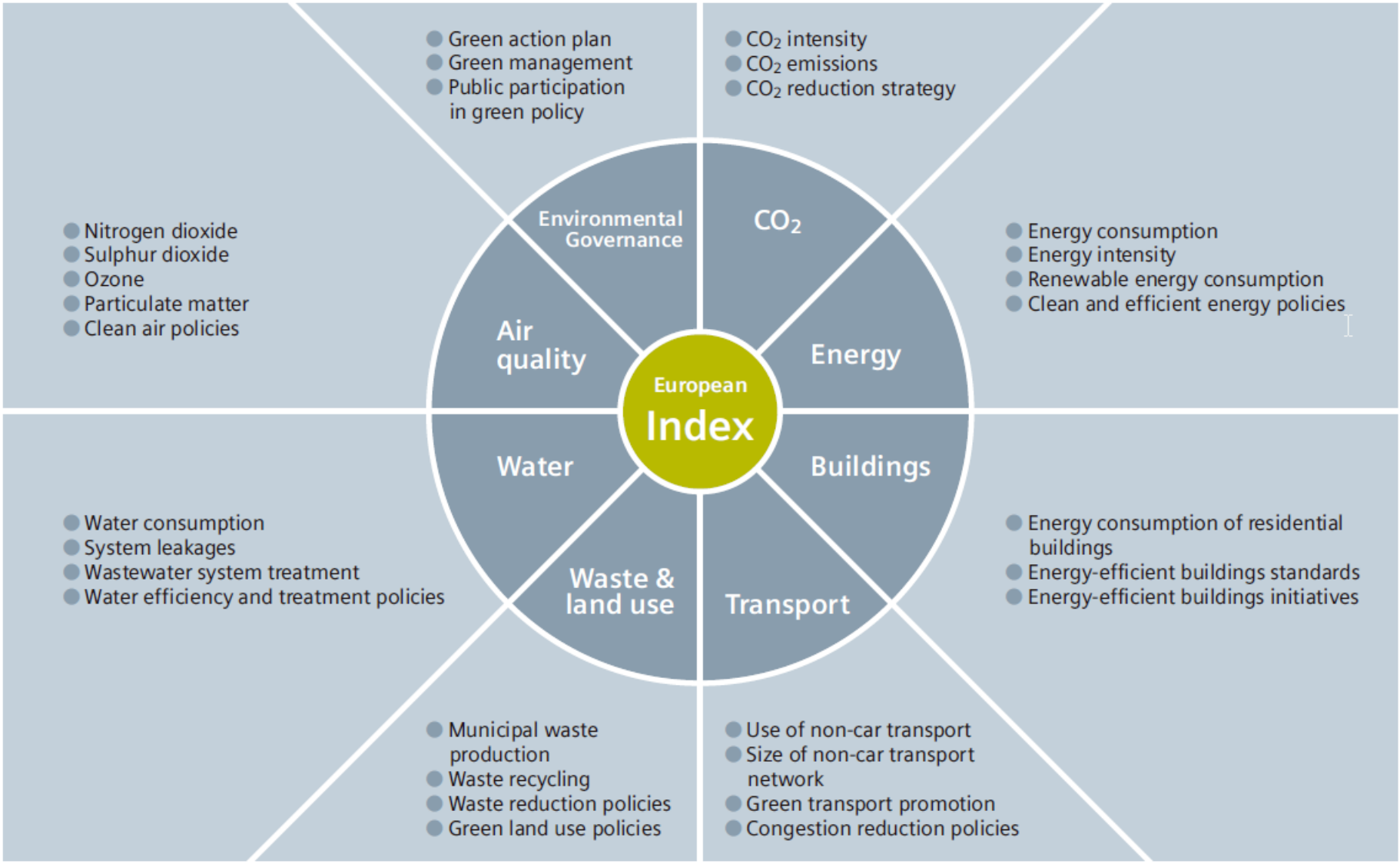
Numeric data



Vrandecic et al.: Linked Open Numbers.1st of April, 2010.

Use Case: Green City Index

(Siemens, GCI, 2014)



Green City Index: Purpose

Oslo uses the highest share of renewable energy at 65%. The Index average is 7%.

Copenhagen's and Berlin's residential buildings consume almost 40% less energy than the Index average.

In Stockholm, 68% of people cycle or walk to work, the highest percentage in the European Index. In contrast, in Helsinki, another Scandinavian city of similar size, only 16% do so.

Riga offers the longest public transport network at 8.6 km per km², almost four times the Index average of 2.3 km per km².

In Kiev, 74% of the population uses public transport to get to work. This is the highest figure in the European Index and the best result for Kiev, which ranks 30th overall.

Tallinn consumes the least amount of water – only 138 litres per person per day, compared with the Index average of 288 litres.

Amsterdam has the lowest water leakage rate of 4%, in Sofia this is 61%.

Helsinki recycles 58% of its waste, compared with the Index average of only 18%.

Green City Index: Purpose

Oslo uses the highest share of renewable energy at 65%. The Index average is 7%.

Copenhagen's and Berlin's residential buildings consume almost 40% less energy than the Index average.

In Stockholm, 68% of people cycle or walk to work, the highest percentage in the European Index. In contrast, in Helsinki, another Scandinavian city of similar size, only 16% do so.

Riga offers the longest public transport network at 8.6 km per km², almost four times the Index average of 2.3 km per km².

Vienna?

Karlsruhe?

In Kiev, 74% of the population uses public transport to get to work. This is the highest figure in the European Index and the best result for Kiev, which ranks 30th overall.

Tallinn consumes the least amount of water – only 138 litres per person per day, compared with the Index average of 288 litres.

Amsterdam has the lowest water leakage rate of 4%, in Sofia this is 61%.

Helsinki recycles 58% of its waste, compared with the Index average of only 18%.

Open City Data Pipeline

(Bischof, Martin, Polleres & Schneider 2015)

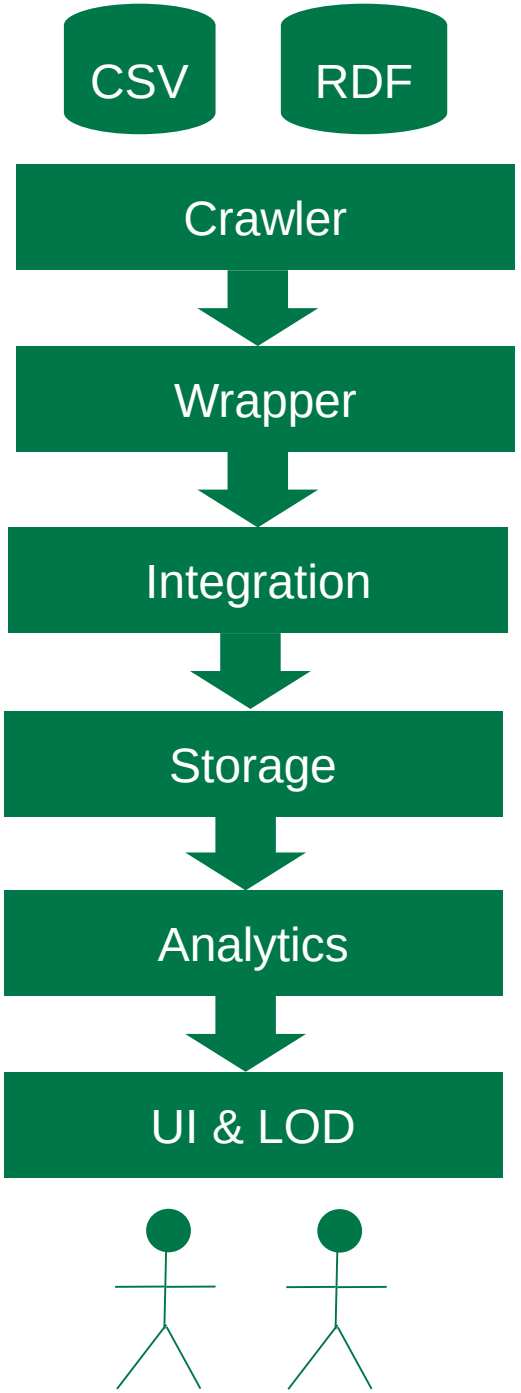
How to add new data sources / datasets?

How to find equivalent instances/properties?

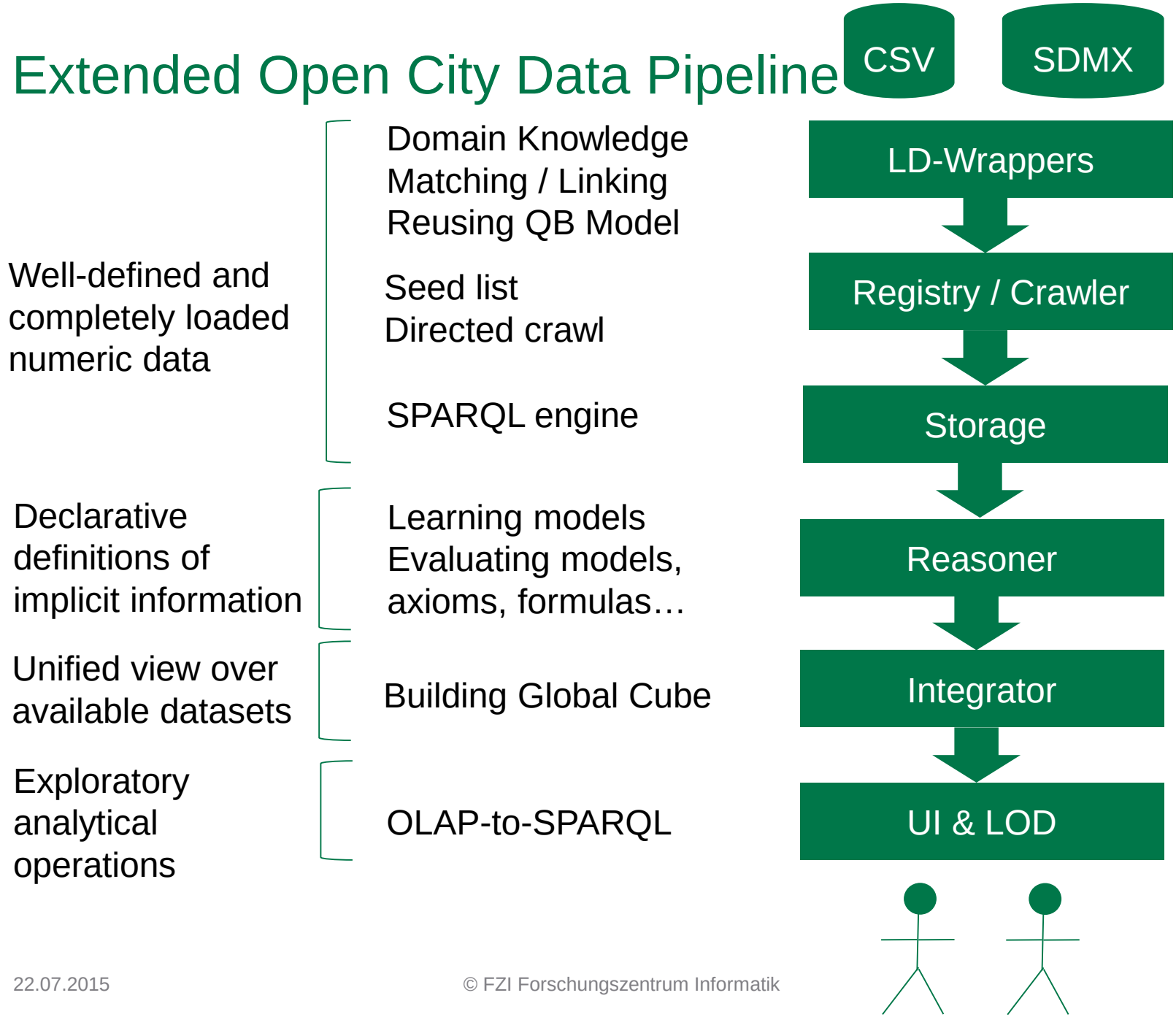
How to discover possible interdependencies between numeric data?

How to integrate all datasets to one dataset with most likely true values?

How to generate useful visualisations?



Extended Open City Data Pipeline

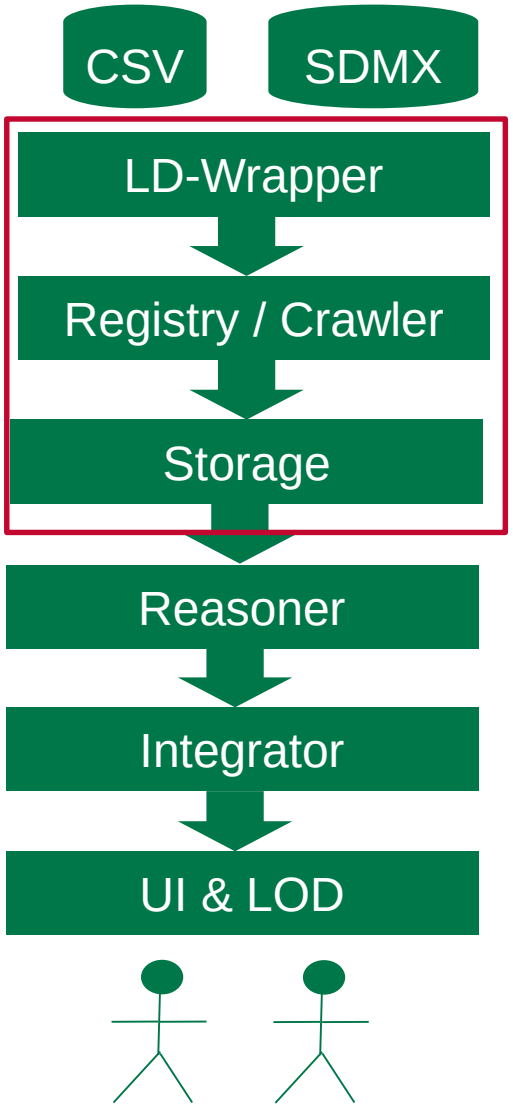


Outline

- FZI / WIM
- Semantics in RDF Data Cubes
  - Well-defined and completely loaded numeric data
  - Declarative definitions of implicit information
  - Unified view over available datasets
  - Exploratory analytical operations

Outline

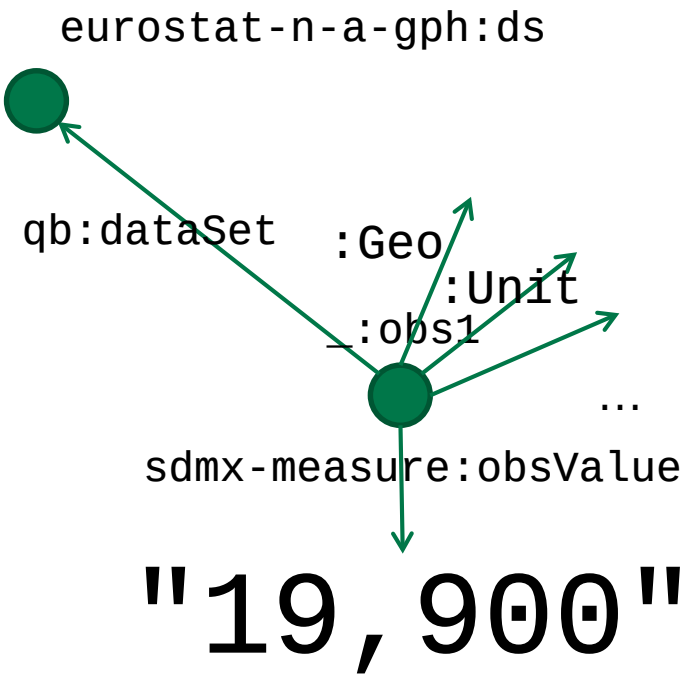
- FZI / WIM
- Semantics in RDF Data Cubes
  - **Well-defined and completely loaded numeric data**
  - Declarative definitions of implicit information
  - Unified view over available datasets
  - Exploratory analytical operations



Well-Defined Numeric Data

- The RDF Data Cube Vocabulary [QB]
  - *W3C Recommendation*
  - Integrity constraints
    - *"Every numeric value is in a dataset"*
    - *"Every dataset has a data structure definition with dimensions"*
    - *"Every numeric value is uniquely defined by its dimension values"*

\implemented using SPARQL



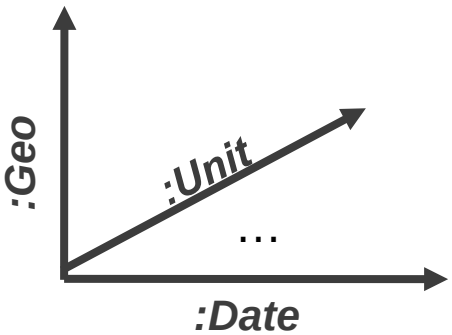
[QB] http://www.w3.org/TR/vocab-data-cube/

- Relational description

```
eurostat-n-a-gph:ds(:Geo, :Unit,   :Date, :Indicator,:Value)
eurostat-n-a-gph:ds(:gr,  :eur_hab,"2010",:ngdph,    "19,900")
```

Well-Defined Numeric Data: Why Cube?

- N-Dimensional Space (coordination system)



:Geo	:Unit	:Date	:Indicator	:Value
:gr	:eur_hab	"2010"	:ngdph	19,900
…	…	…	…	…

GDP Per Capita Dataset

- [CUBE] operator

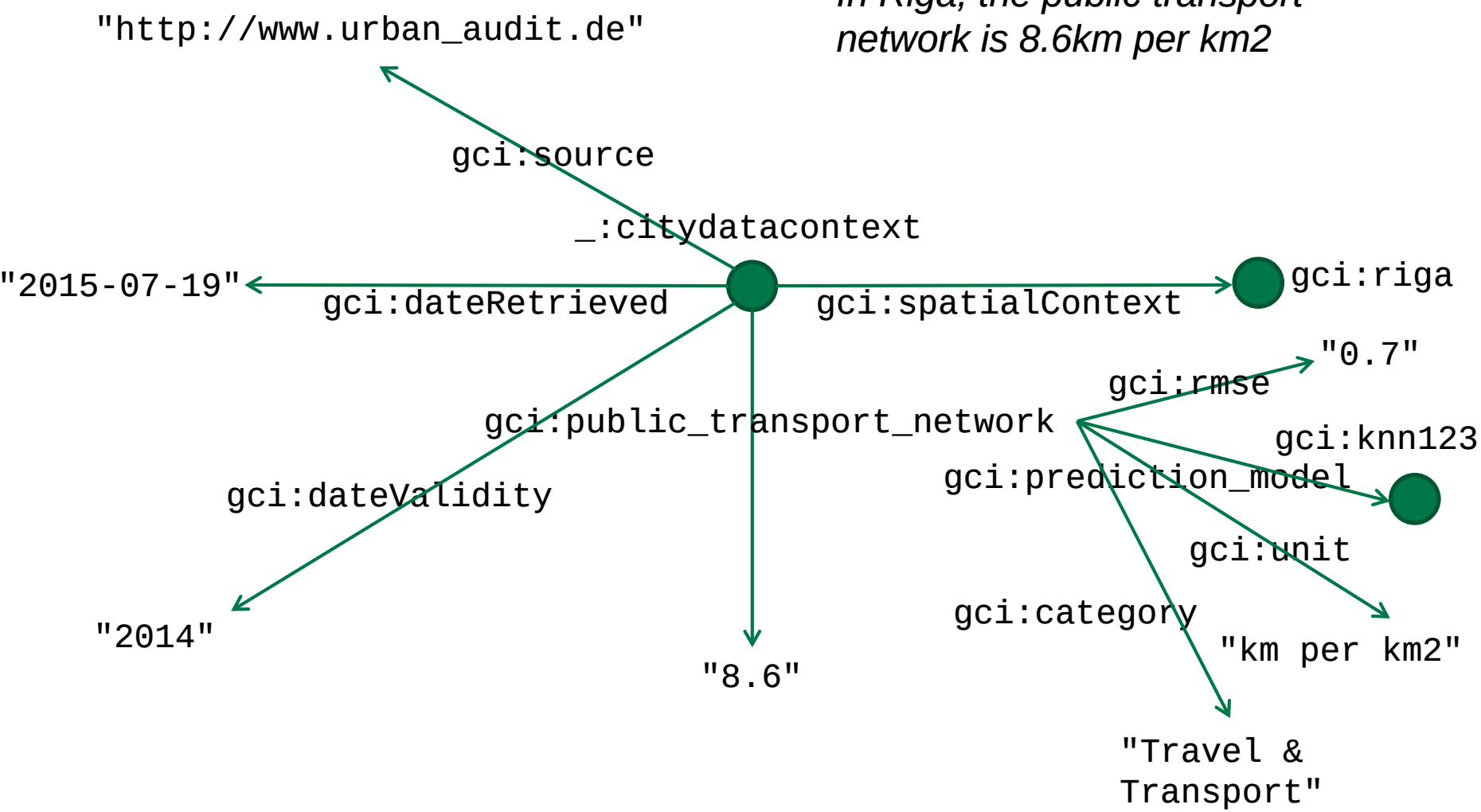
Gray, Chaudhuri, Bosworth: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals. *Data Mining and Knowledge Discovery*. 1997.

:Geo	:Unit	:Date	:Indicator	:Value
:gr	:eur_hab	"2010"	:ngdph	19,900
ALL	:eur_hab	"2010"	:ngdph	?
…	…	…	…	…

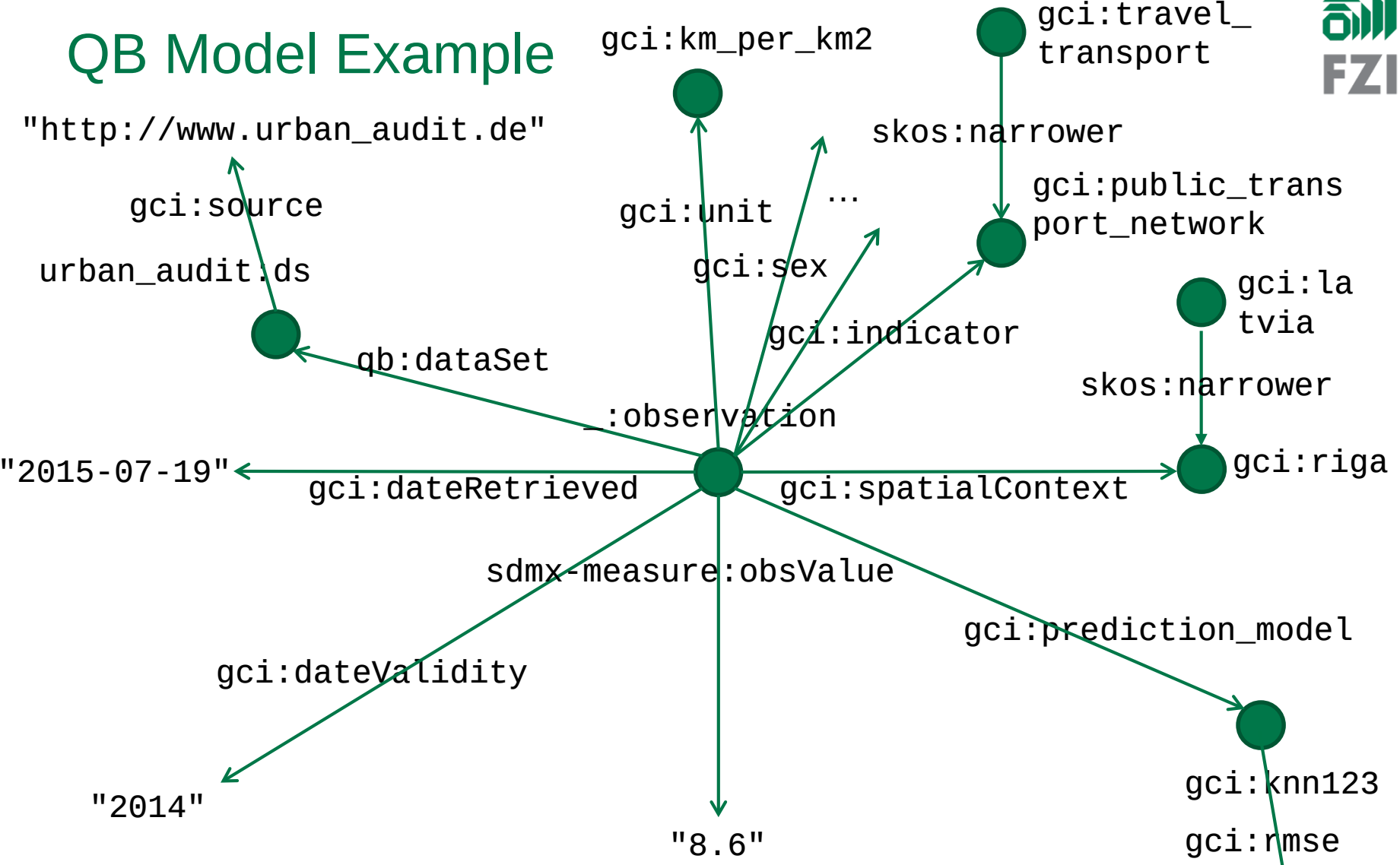
Materialised GDP Per Capita Cube

City Data Model Example

In Riga, the public transport network is 8.6km per km2



QB Model Example



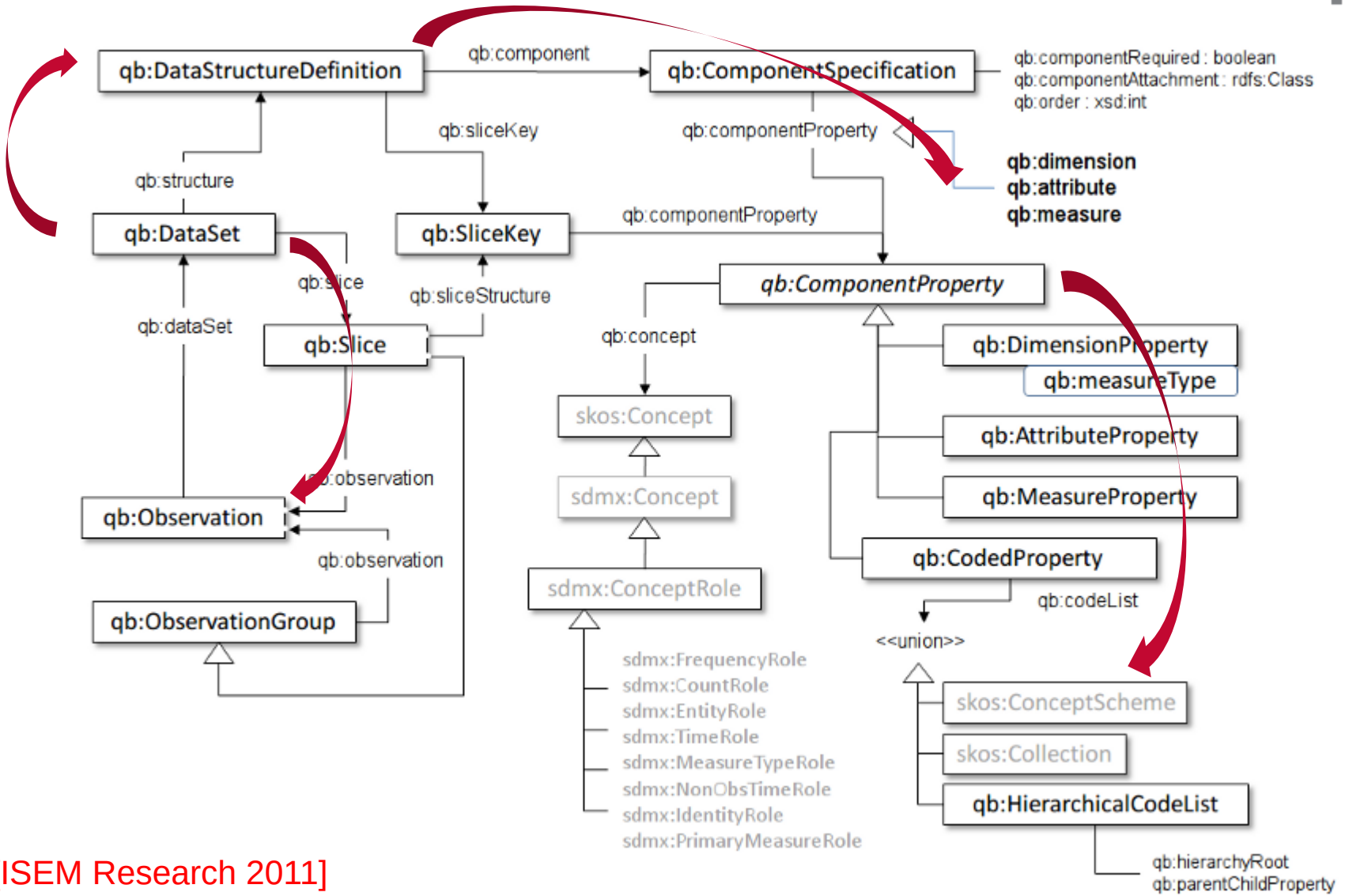
urban_audit:ds(Gci:dateRetrieved, Gci:dateValidity, Gci:unit, ..., Gci:indicator, Gci:spatialContext, Gci:prediction_model, Sdmx-measure:obsValue)

Mapping City Data Model and QB

```
@prefix gci: <http://citydata.wu.ac.at/ns#> .
@prefix qb: <http://purl.org/linked-data/cube#> .

{ ?x a gci:CityDataContext } => { ?x a qb:Observation }.
{ ?x :source ?y } => {concat(?y,"ds") a qb:DataSet. ?x qb:dataSet concat(?y,"ds").
concat(?y,"ds") :source ?y. concat(?y,"ds") qb:structure concat(?y,"dsd"). }.
{ ?x :source ?y } => { concat(?y,"dsd") qb:component concat(?x,"dsd-component-spatialContext").
concat(?x,"dsd-component-spatialContext") qb:dimension gci:spatialContext. }.
gci:spatialContext a qb:DimensionProperty.
# One could explicitly state all possible dimension values if needed.
...
gci:dateRetrieved a qb:DimensionProperty.
{ ?x :source ?y } => { concat(?x,"dsd") qb:component concat(?x,"dsd-component-dateRetrieved").
concat(?x,"dsd-component-dateRetrieved") qb:dimension gci:dateRetrieved. }.
...
gci:dateValidity a qb:DimensionProperty.
{ ?x :source ?y } => { concat(?x,"dsd") qb:component concat(?x,"dsd-component-dateValidity").
concat(?x,"dsd-component-dateValidity") qb:dimension gci:dateValidity. }.
...
{ ?x :source ?y } => { concat(?x,"dsd") qb:component concat(?x,"dsd-component-
indicatorContext"). concat(?x,"dsd-component-indicatorContext") qb:dimension
gci:indicatorContext. }.
gci:indicatorContext a qb:DimensionProperty.
{ ?x :source ?y } => { concat(?x,"dsd") qb:component concat(?x,"dsd-component-unitContext").
concat(?x,"dsd-component-unitContext") qb:dimension gci:unitContext. }.
gci:unitContext a qb:DimensionProperty.
# Assumption: for each CityDataContext, there is exactly one indicator value.
{ ?x a gci:CityDataContext. ?x ?p ?y. ?p a gci:Indicator. ?p gci:unit ?u. ?p gci:category ?c. }
=> { ?x gci:indicatorContext ?p. ?c skos:narrower ?p. ?x gci:unitContext ?u. }.
gci:estimatedRMSEForPredictedValues a qb:DimensionProperty.
gci:predictorForPredictedValues a qb:DimensionProperty.
```

\\implementeable using SPARQL

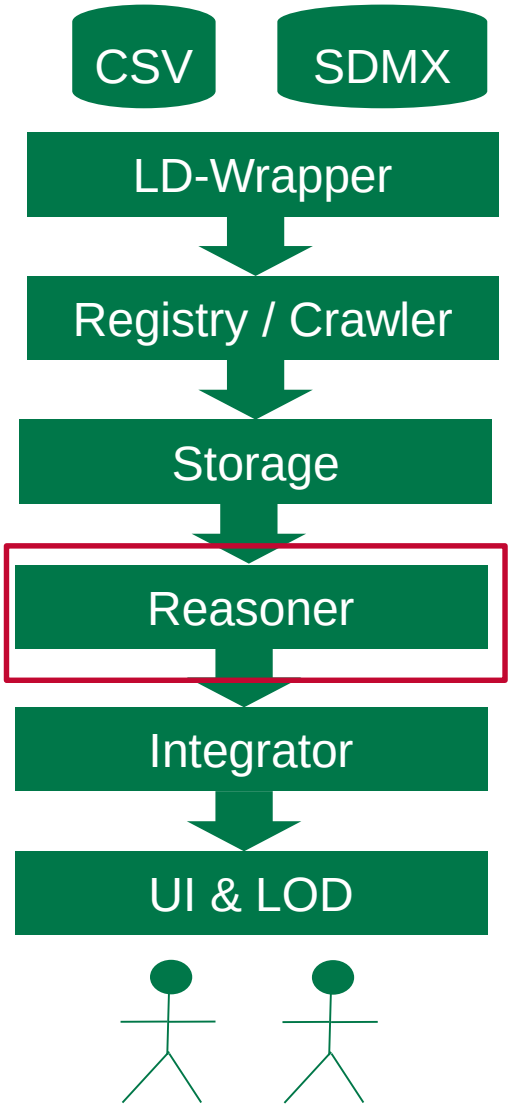
Completely Loaded Numeric Data



[ISEM Research 2011]

Outline

- FZI / WIM
- Semantics in RDF Data Cubes
  - Well-defined and completely loaded numeric data
  - **Declarative definitions of implicit information**
  - Unified view over available datasets
  - Exploratory analytical operations



Declarative Definitions of Implicit Information

- **RDFS / OWL Axioms (e.g., equivalence)**
- QB normalisation of abbreviated datasets
- Derived Datasets

RDFS / OWL Axioms (e.g., equivalence)

gci:Geo	gci:Unit	Dcterms:date	gci:Indicator	gci:Value
gci:greece	:mio_eur	"2010"	:b1g	1,547,984
...		...	...	

`owl:sameAs`

`owl:equivalentPropery`

eurostat:Geo	eurostat:Unit	Dcterms:date	eurostat: Indic_na	Sdmx-measure: obsValue
eurostat:gr	:eur_hab	"2010"	:ngdph	19,900
...	...	...	...	...

- Mapping dimensions and dimension values (Zapilko & Mathiak '14)

- Evaluation
  - Offline Materialisation //implemented using Entity Consolidation
  - Offline Materialisation //implemented using OWL LD Profile rules
  - Online Evaluation //implemented using Open Virtuoso Reasoning

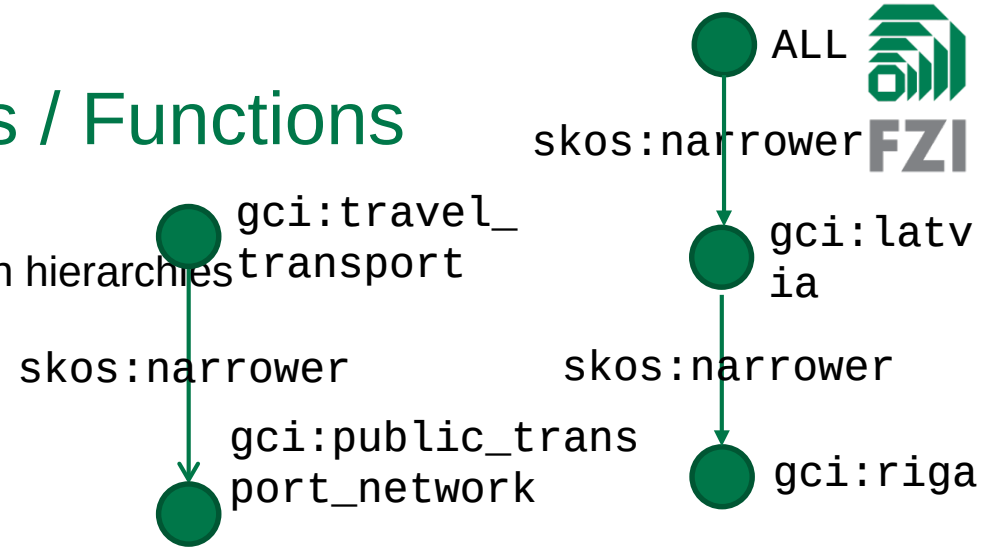
Declarative Definitions of Implicit Information

- RDFS / OWL Axioms (e.g., equivalence)
- **QB normalisation of abbreviated datasets**
  - Preprocessing <= //implemented using QB normalisation algorithm
- Derived Datasets

Declarative Definitions of Implicit Information

- RDFS / OWL Axioms (e.g., equivalence)
- QB normalisation of abbreviated datasets
- Derived Datasets
  - **Aggregation hierarchies / functions**
  - Complex Correspondences
  - Prediction models

Aggregation Hierarchies / Functions



- QB defines implicit / explicit dimension hierarchies

- QB does not define how to aggregate from lower-level to higher-level values.
  - Summarisability: Summing up the public transport network over time is meaningless. Averaging up the public transport network for all cities would give the Index average.
  - Redundancy: When aggregating, values should not be aggregated over several times. E.g., if we have the values male, female, total, computing ALL would be meaningless.

- *"Riga offers the longest public transport network at 8.6 km per km2, almost four times the Index average of 2.3 km per km2."*

```
urban_audit:ds_view("2015-07-19", "2014", gci:km_per_km2,
gci:public_transport_network, ALL, gci:none, avg(?)).
```

- Aggregation function can be explicitly stated for an indicator.
- CUBE operator

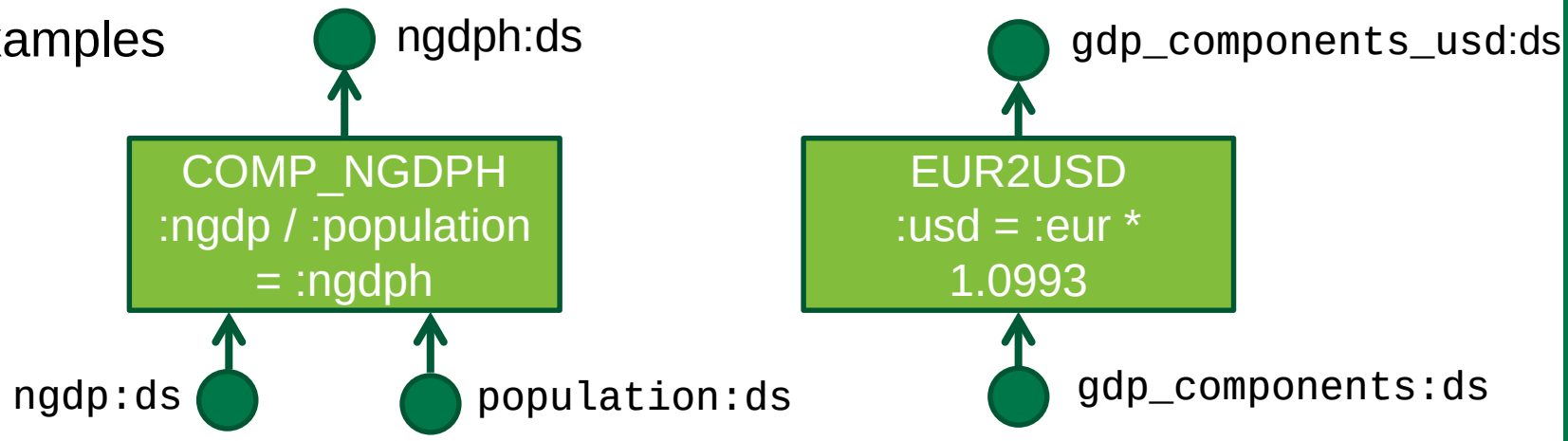
//implemented using Views over RDF (SPARQL) [ESWC Research 2013] (Etcheverry & Vaisman '12)

Declarative Definitions of Implicit Information

- RDFS / OWL Axioms (e.g., equivalence)
- QB normalisation of abbreviated datasets
- Derived Datasets
  - Aggregation hierarchies / functions
  - **Complex Correspondences**
  - Prediction models

Complex Correspondences

- Examples



  - How to describe and efficiently evaluate?

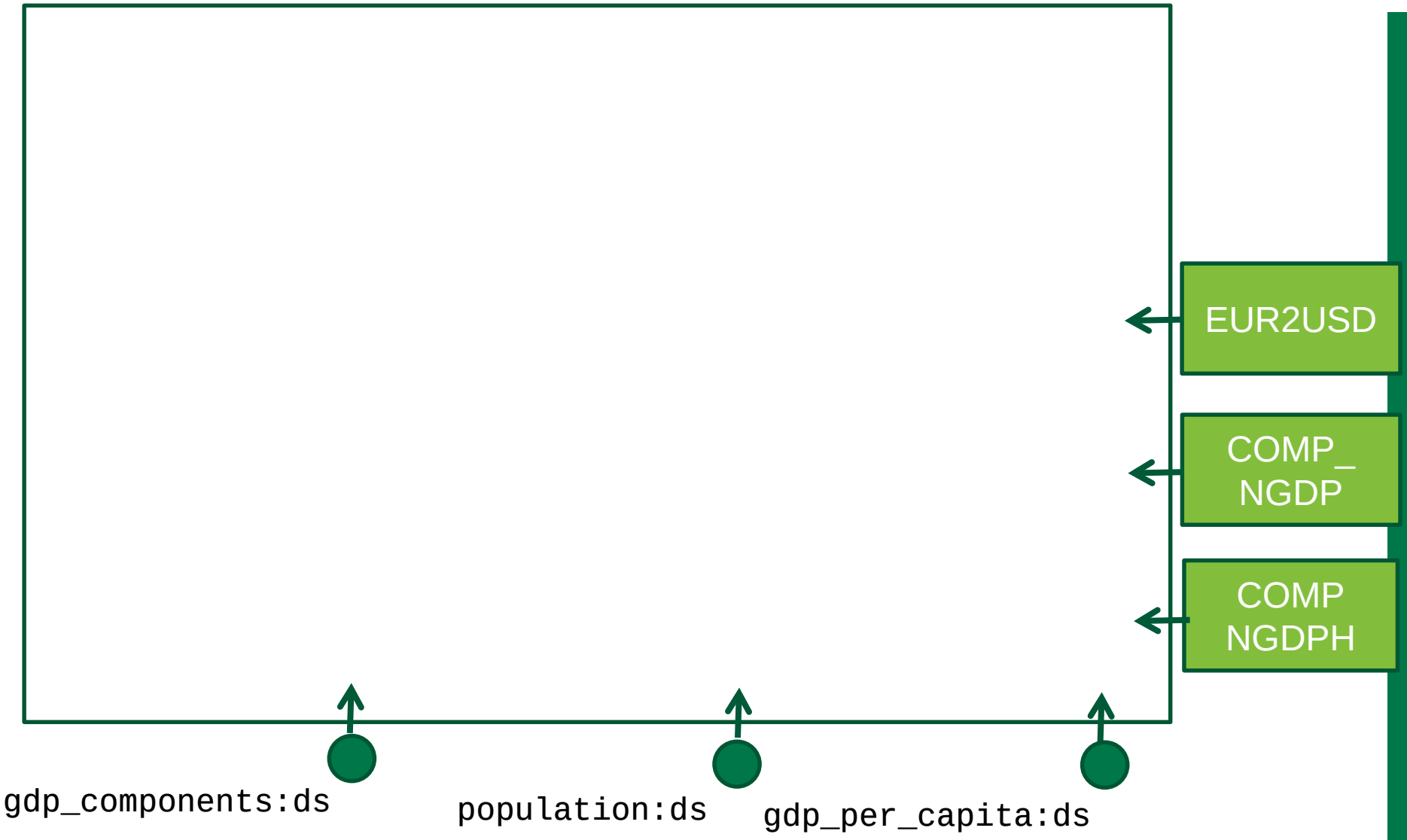
//related to attribute equations (Bischof & Polleres '13)

- Semantics (Example)

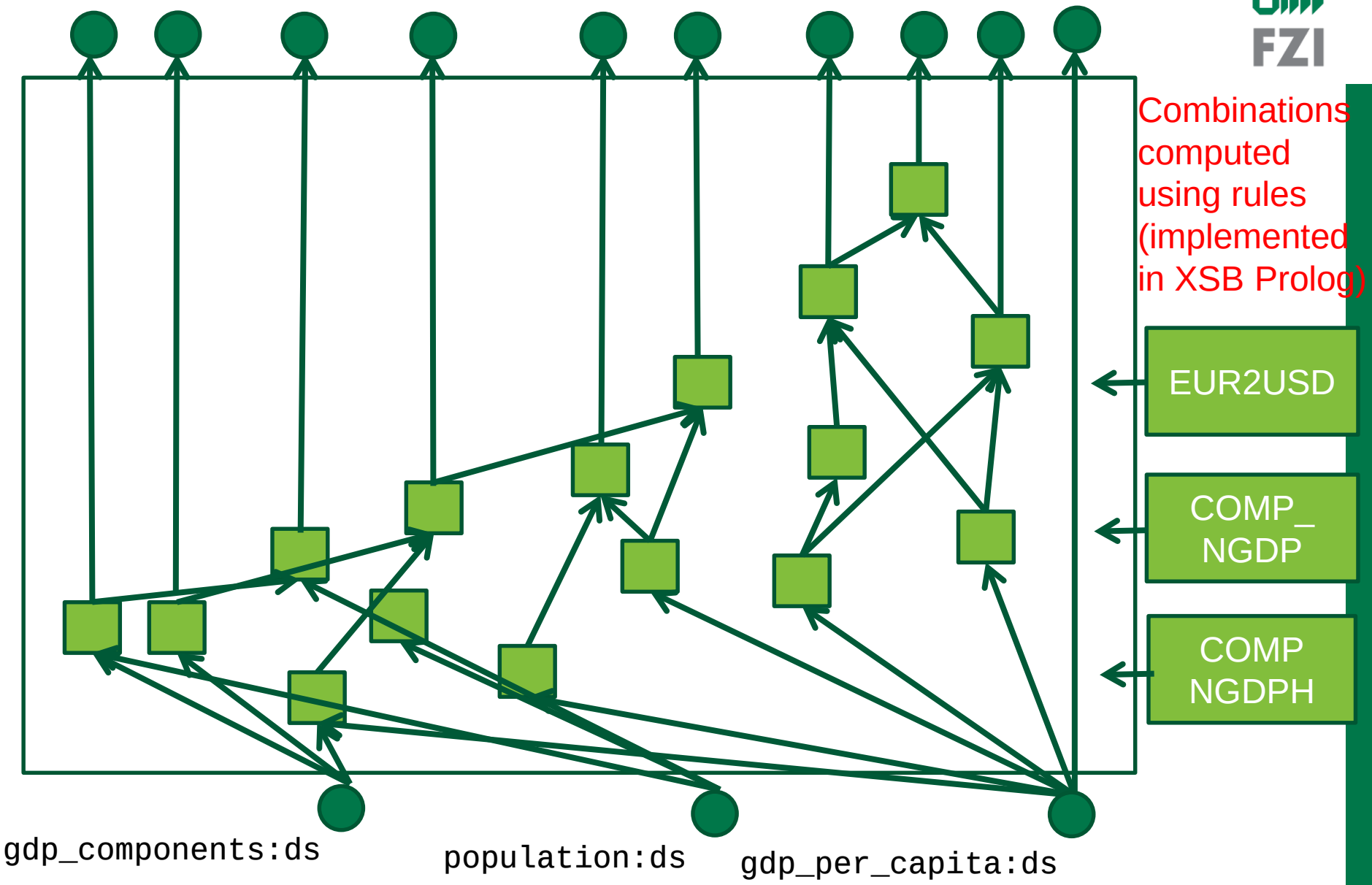
```
Convert-Cube(gdp_components:ds, EUR2USD) = gdp_components_usd

gdp_components_usd:ds(Geo, :usd, Date, Indicator, Value2)
:- gdp_components(Geo, :eur, Date, Indicator, Value1), Value2
= 1,0993 * Value1.
```

//implemented using SPARQL [EKAW Research 2014]

Complex Correspondences – Computing All Derived Datasets



Complex Correspondences – Computing All Derived Datasets

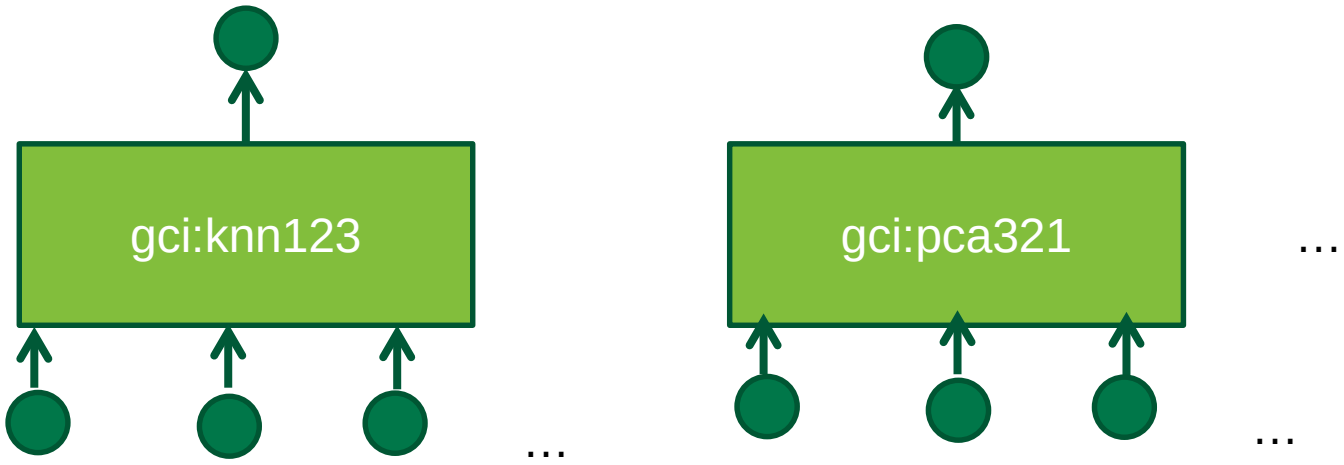


Declarative Definitions of Implicit Information

- RDFS / OWL Axioms (e.g., equivalence)
- QB normalisation of abbreviated datasets
- Derived Datasets
  - Aggregation hierarchies / functions
  - Complex Correspondences
  - **Prediction models**

Prediction models

- "Prediction Models"
  as first-class citizens

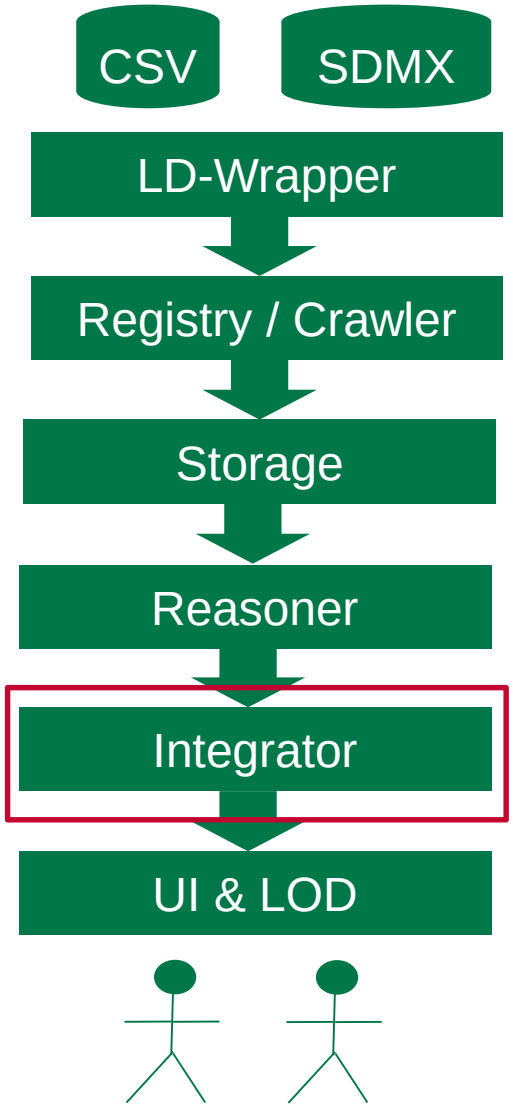



```
urban_audit:ds_knn("2015-07-19", "2014",
gci:km_per_km2, gci:public_transport_network,
gci:vienna, gci:knn123, f(?)).
```

//Open Work

Outline

- FZI / WIM
- Semantics in RDF Data Cubes
  - Well-defined and completely loaded numeric data
  - Declarative definitions of implicit information
  - **Unified view over available datasets**
  - Exploratory analytical operations



Unified View Over Available Datasets

- Query in terms of a unified view – Global Cube:

```
global_cube(ALL, "2014", gci:km_per_km2, ...,
gci:public_transport_network, gci:vienna, ALL, f(?)).
```

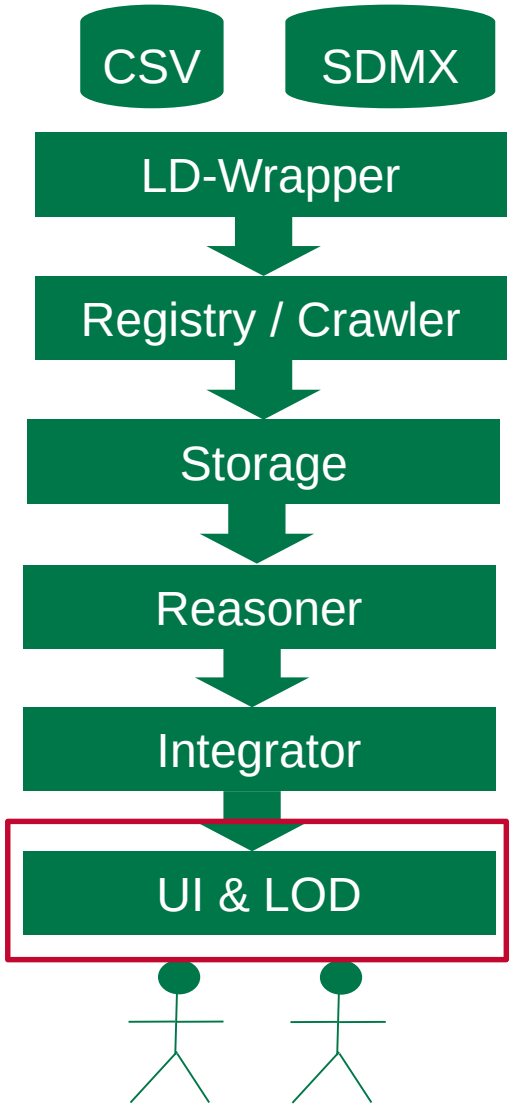
- Global Cube in terms of available datasets:

```
global_cube(Gci:dateRetrieved, Gci:dateValidity,
Gci:unit, ALL, ..., Gci:indicator, Gci:spatialContext,
gci:none, "1", Sdmx-measure:obsValue) <=
urban_audit:ds(Gci:dateRetrieved, Gci:dateValidity,
Gci:unit, Gci:indicator, Gci:spatialContext, gci:none,
"0", Sdmx-measure:obsValue).
```

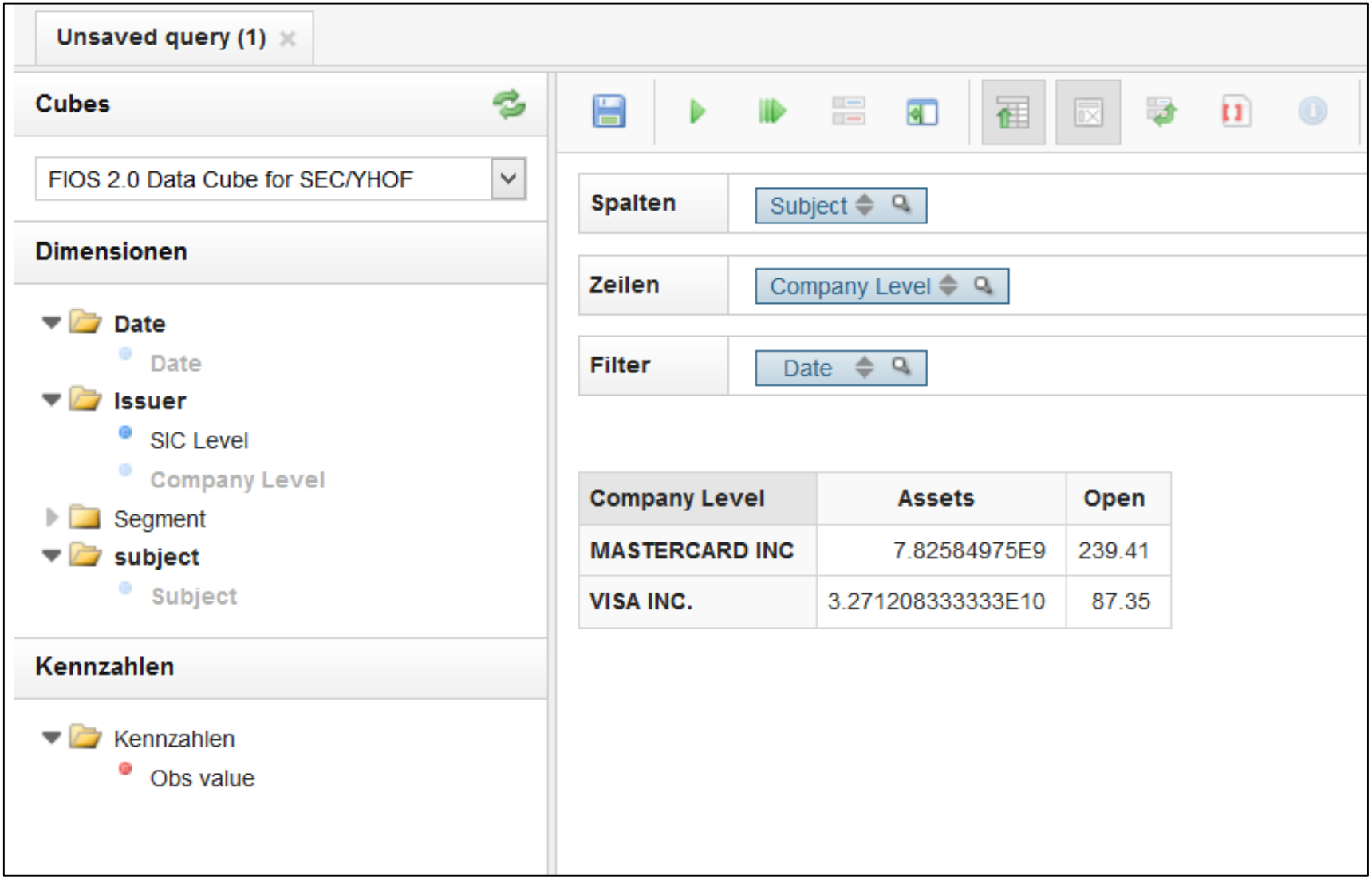
- Automatic "aggregation" to most likely value //Open Work
  - For `Gci:dateRetrieved`, take the most up-to-date one.
  - For `Gci:prediction_model`, take the "best" one, with lowest `gci:rmse`

Outline

- FZI / WIM
- Semantics in RDF Data Cubes
  - Well-defined and completely loaded numeric data
  - Declarative definitions of implicit information
  - Unified view over available datasets
  - **Exploratory analytical operations**

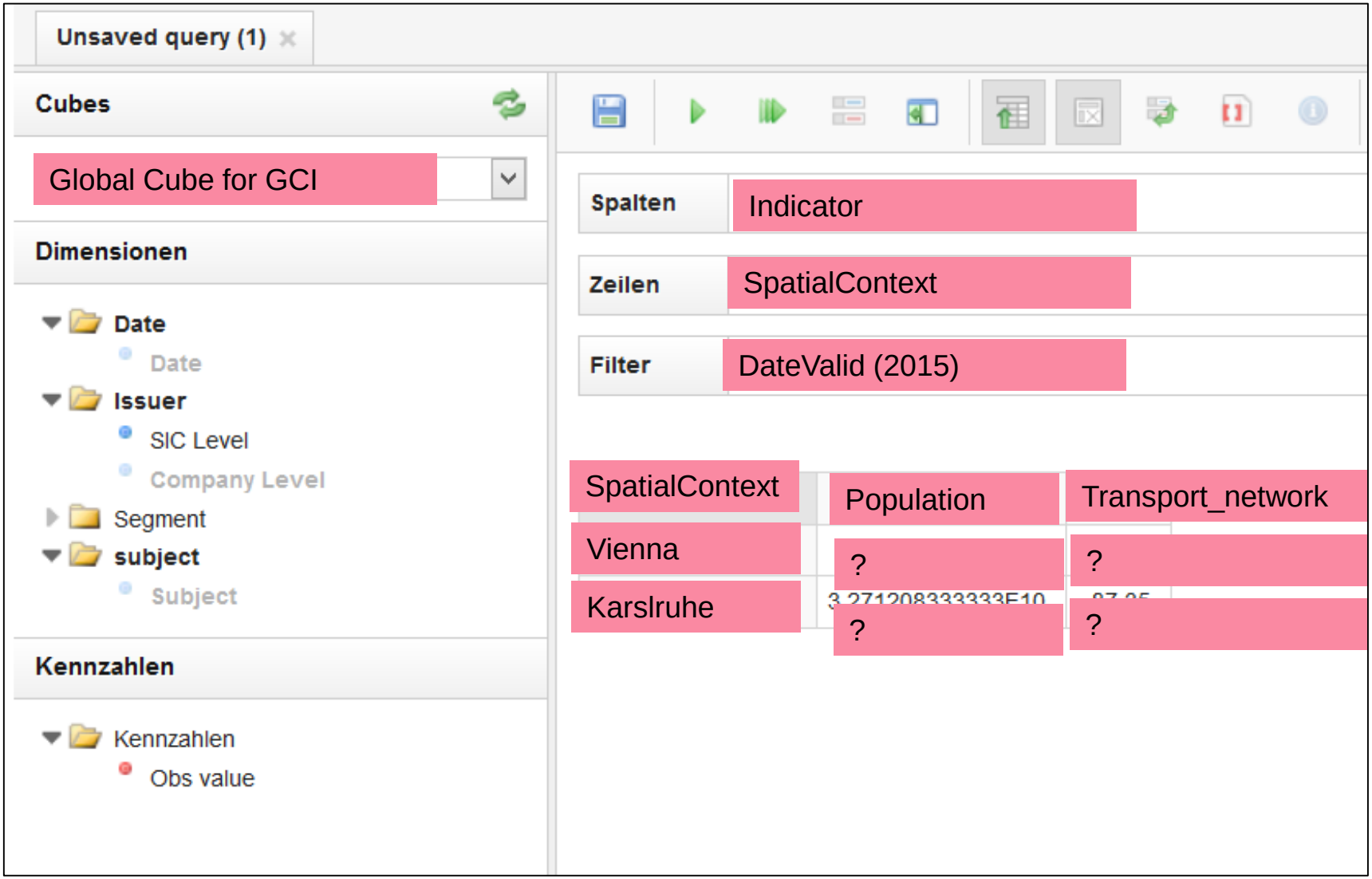


Exploratory Analytical Operations



[ESWC In-Use 2014]

Exploratory Analytical Operations



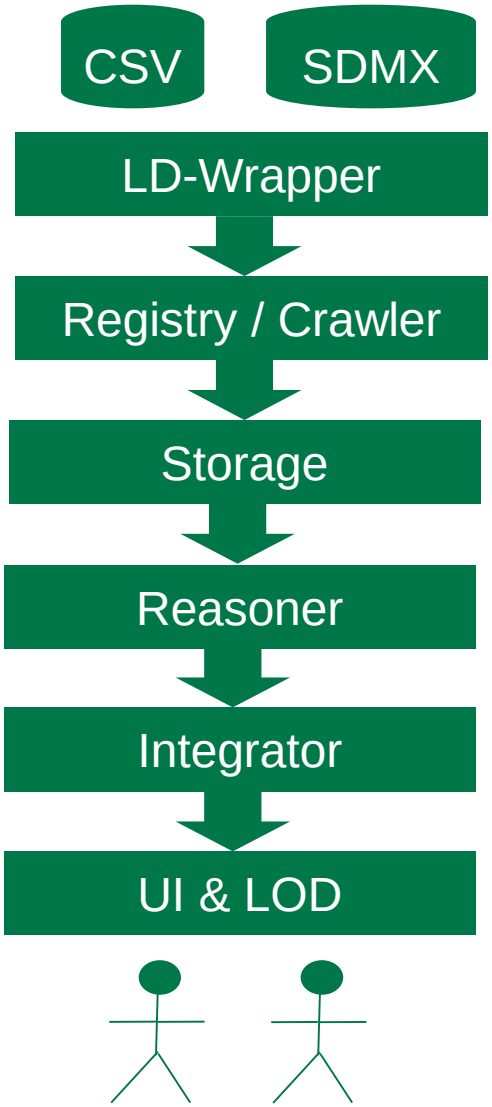
[ESWC In-Use 2014]

Conclusions

- Open City Data Pipeline promising
- Extended approach allows to focus on the important aspects of analytics:
  - Where to get the data from?
  - What does the data mean (from numeric data, over equivalent entities, to interdependencies)?
  - How to discover possible interdependencies between numeric data?
  - How to select between different predictions?
  - How to generate useful visualisations?

Thanks!

- Open City Data Pipeline promising
- Extended approach allows to focus on the important aspects of analytics:
  - Where to get the data from?
  - What does the data mean (from numeric data, over equivalent entities, to interdependencies)?
  - How to discover possible interdependencies between numeric data?
  - How to select between different predictions?
  - How to generate useful visualisations?



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.

Bischof, Martin, Polleres, Schneider: Collecting, Integrating, Enriching and Re-publishing Open City Data as Linked Data. ISWC In-Use Track 2015.

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.