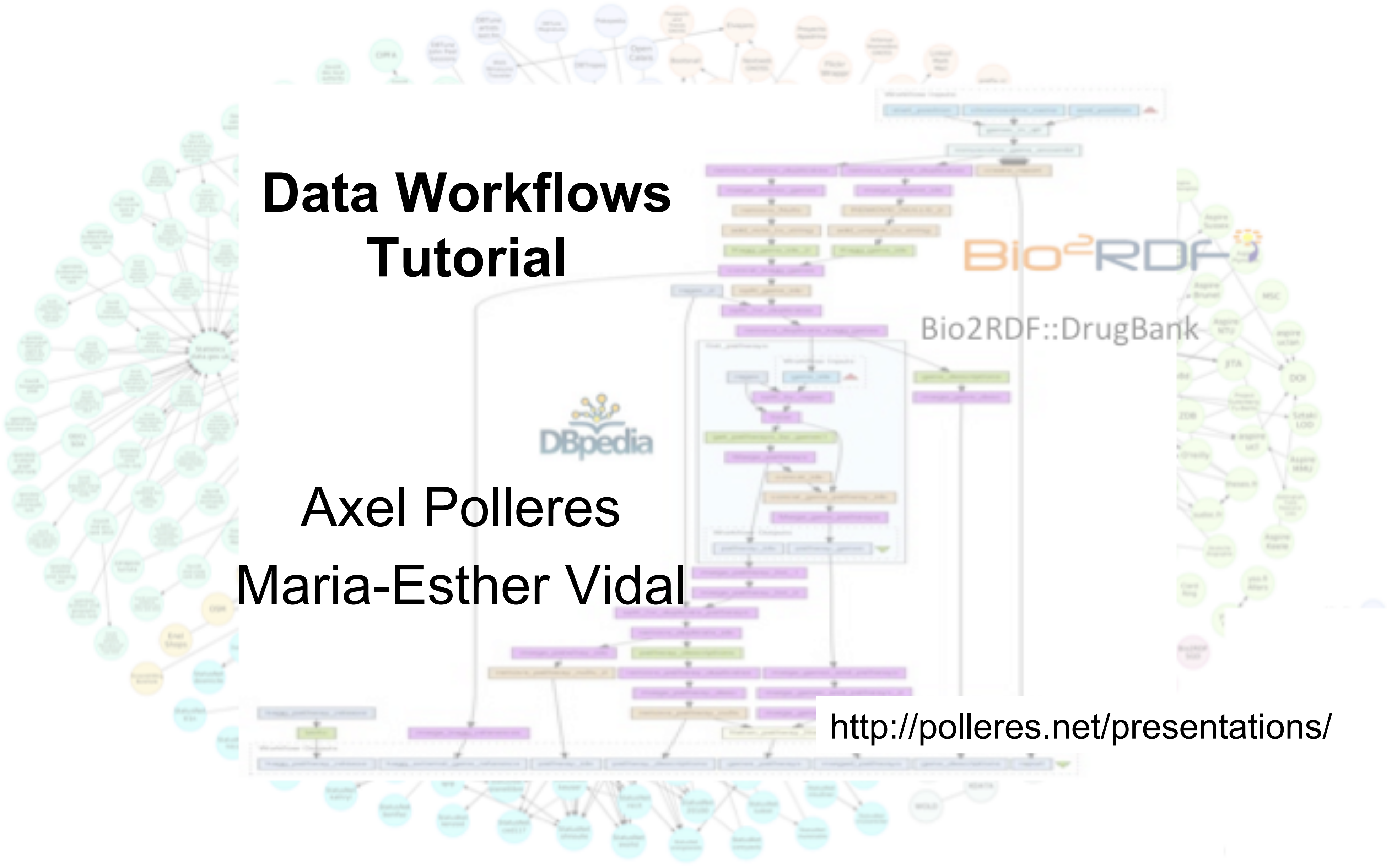


Data Workflows Tutorial

Axel Polleres

Maria-Esther Vidal

http://polleres.net/presentations/

Outline

- Motivation
  - Integrating (Open) Data from different sources
  - Not only Linked Data (NoLD)
  - Data workflows and Open data in the context of rise of Big Data
- What is a "Data Workflow"?
  - Different Views of Data Workflows in the context of the Semantic Web
  - Key steps involved
  - Tools?
- Data Integration Systems
  - Wrappers vs. Mediators
  - GAV vs. LAV
  - Query rewriting vs. Materialisation
  - Data Integration using Ontologies
- Challenges:
  - How to find Rules and ontologies?
  - Data Quality & Incomplete Data
  - Maintainance/Evolution/Sustainability of Data Workflows
  - **Break?**
- Open Problems – Research Tasks

Motivation

- Integrating (Open) Data from different sources

Open Data is a global trend – Good for us!

- Cities, International Organizations, National and European **portals**, etc.:

UNdata A world of information

THE WORLD BANK Open Data

DBpedia

wien.at

NYC OpenData

london.gov.uk

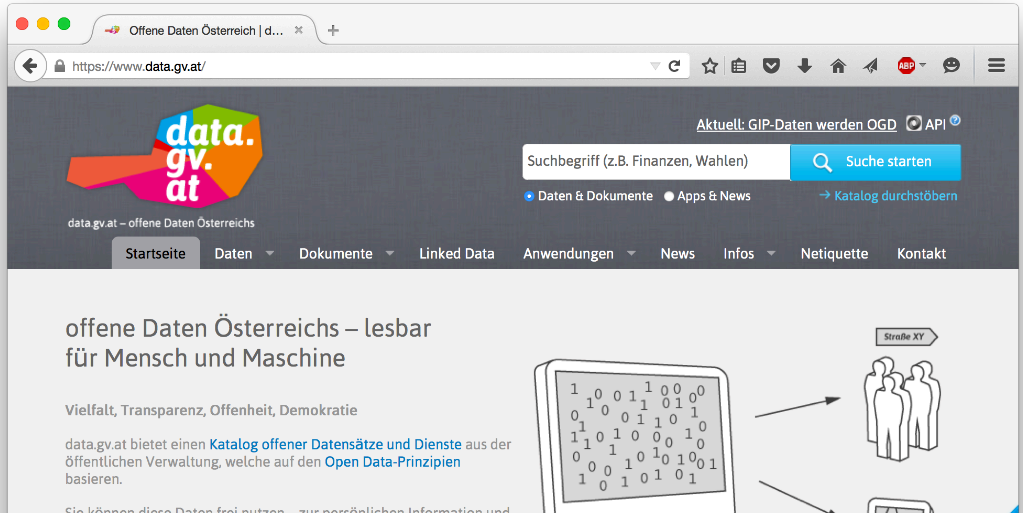
European Union Open Data Portal

DATA.GOV.UK Beta Opening up Government

DATA.GOV

data.gv.at

odp BETA All you can Data OPEN DATA PORTAL ÖSTERREICH


Offene Daten Österreich | d...

https://www.data.gv.at/

data.gv.at – offene Daten Österreichs

Aktuell: GIP-Daten werden OGD API

Suchbegriff (z.B. Finanzen, Wahlen) Suche starten

● Daten & Dokumente ○ Apps & News → Katalog durchstöbern

Startseite | Daten | Dokumente | Linked Data | Anwendungen | News | Infos | Netiquette | Kontakt

offene Daten Österreichs – lesbar für Mensch und Maschine

Vielfalt, Transparenz, Offenheit, Demokratie

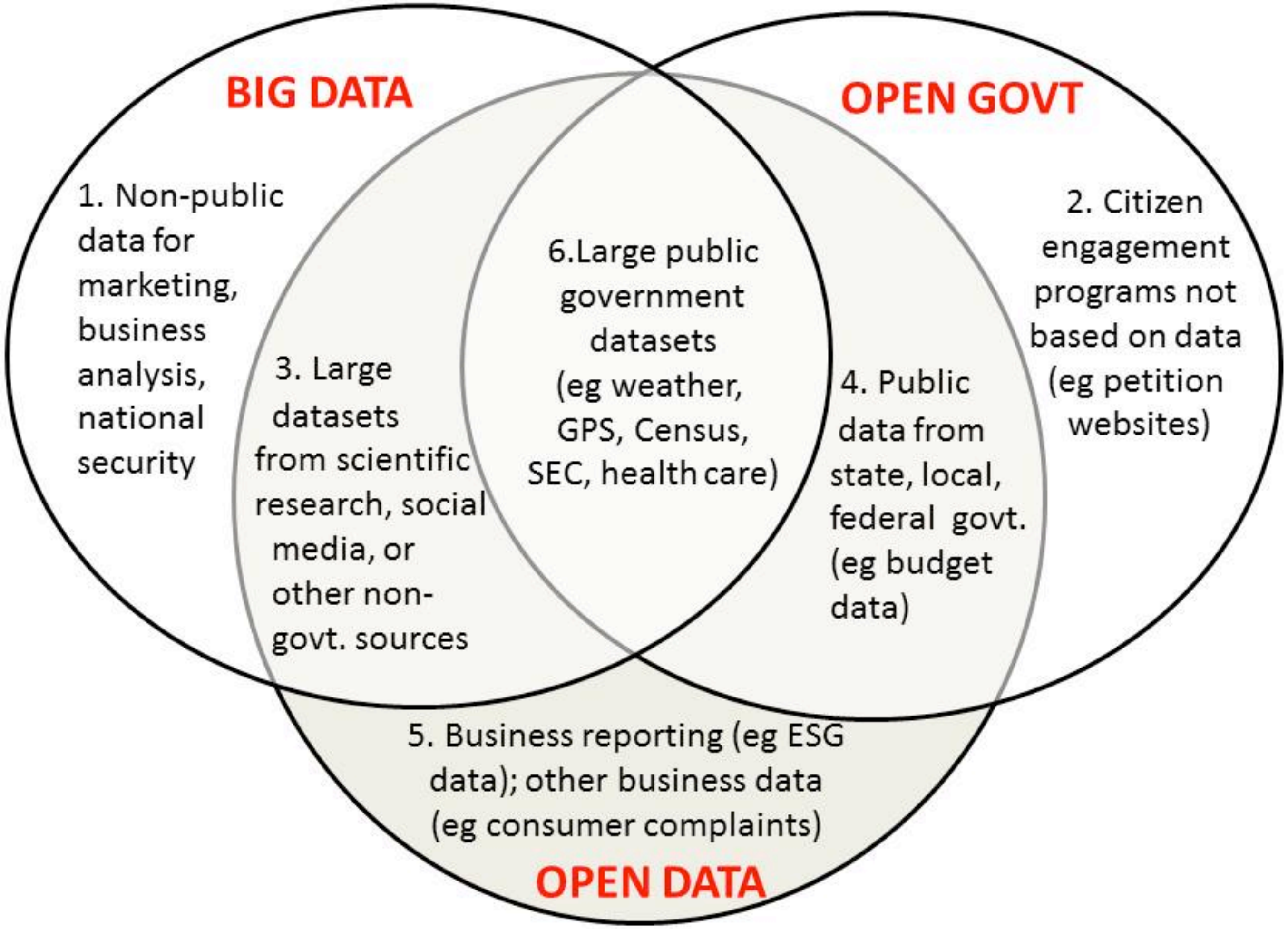
data.gv.at bietet einen Katalog offener Datensätze und Dienste aus der öffentlichen Verwaltung, welche auf den Open Data-Prinzipien basieren.

Sie können diese Daten frei nutzen – zur persönlichen Information und

- In general: more and more structured data available at our fingertipps
- It's on the Web
- It's open

  → no restrictions w.r.t. re-use

Buzzword Bingo 1/3: Open Data vs. Big Data vs. Open Government



BIG DATA

OPEN GOVT

OPEN DATA

1. Non-public data for marketing, business analysis, national security

2. Citizen engagement programs not based on data (eg petition websites)

3. Large datasets from scientific research, social media, or other non-govt. sources

4. Public data from state, local, federal govt. (eg budget data)

5. Business reporting (eg ESG data); other business data (eg consumer complaints)

6.Large public government datasets (eg weather, GPS, Census, SEC, health care)

- http://www.opendatanow.com/2013/11/new-big-data-vs-open-data-mapping-it-out/ 1

Buzzword Bingo 2/3:
Open Data vs. Big Data

- **Volume:**
  - It's growing! (we currently monitor 90 CKAN portals, 512543 resources/ 160069 datasets,

  at the moment (statically) ~1TB only CSV files...

- **Variety:**
  - different datasets (from different cities, countries, etc.), only partially comparable, partially not.
  - Different metadata to describe datasets
  - Different data formats

- **Velocity:**
  - Open Data changes regularly (fast and slow)
  - New datasets appear, old ones disappear

- **Value:**
  - building ecosystems ("Data value chain") around Open Data is a key priority of the EC

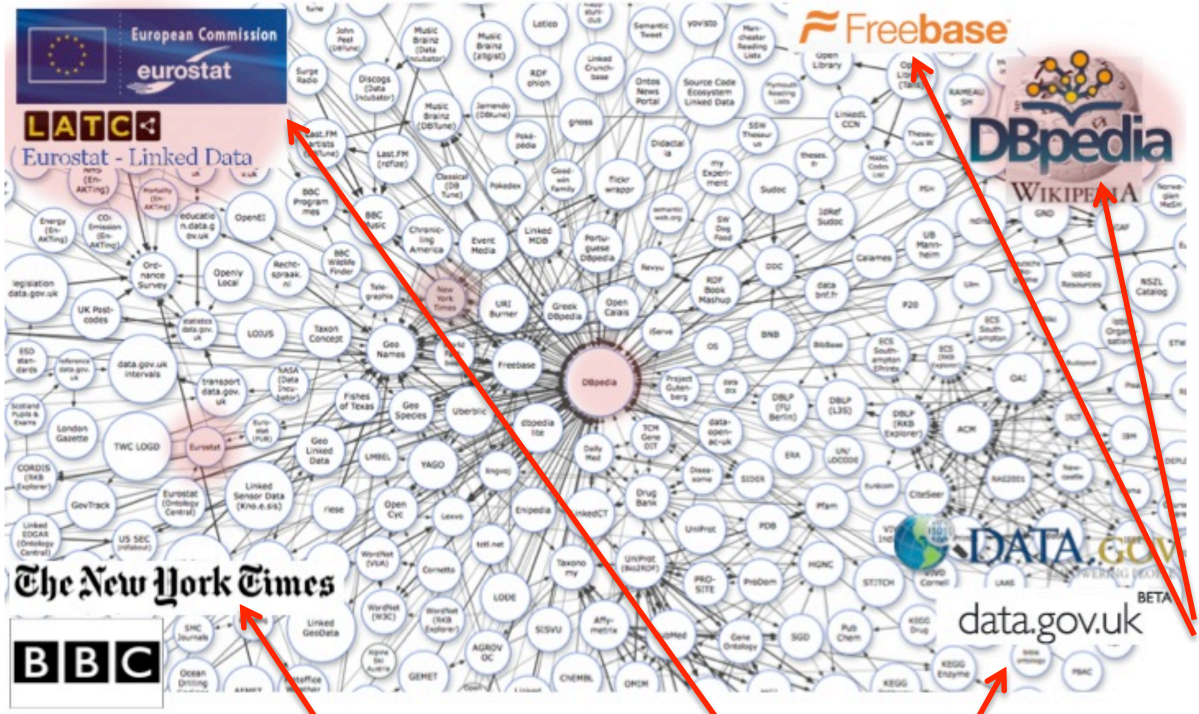
- **Veracity:**
  - quality, trust

Buzzword Bingo 3/3: Open Data vs. Linked Data

This talk is NOT about DL Reasoning over Linked Data:

cf.: [Polleres OWLED2013], [Polleres et al. Reasoning Web 2013]

Linked Data on the Web: Adoption



LOD is till growing, but OD is growing faster and challenges aren't necessarily the exactly same…

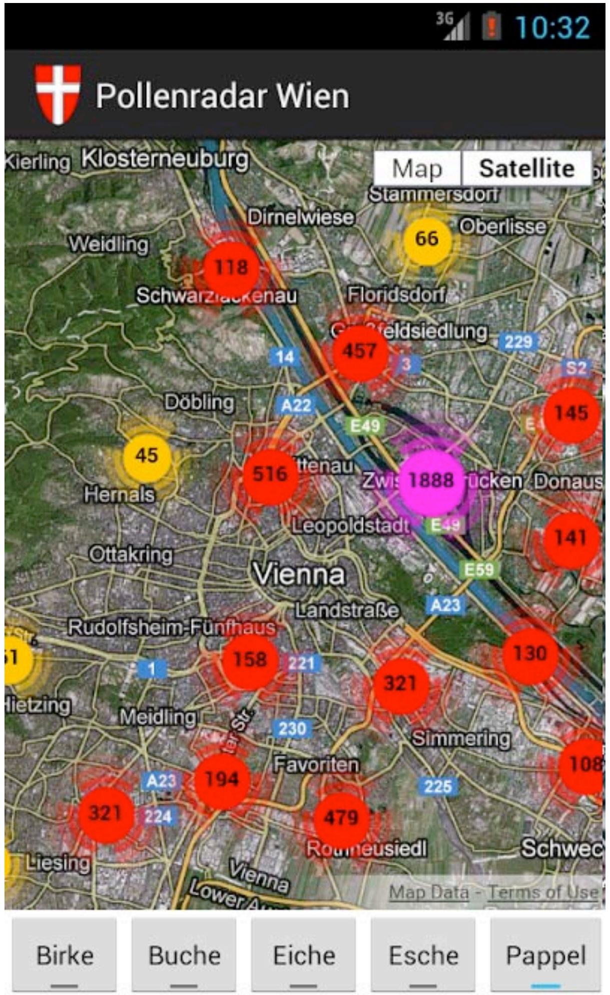
So. let's focus on Open Data in general…

Alternatives in the meantime: (wikidata...)

LD efforts discontinued?!

LOD in OGD growing, but slowly

What makes Open Data useful beyond "single dataset Apps"...

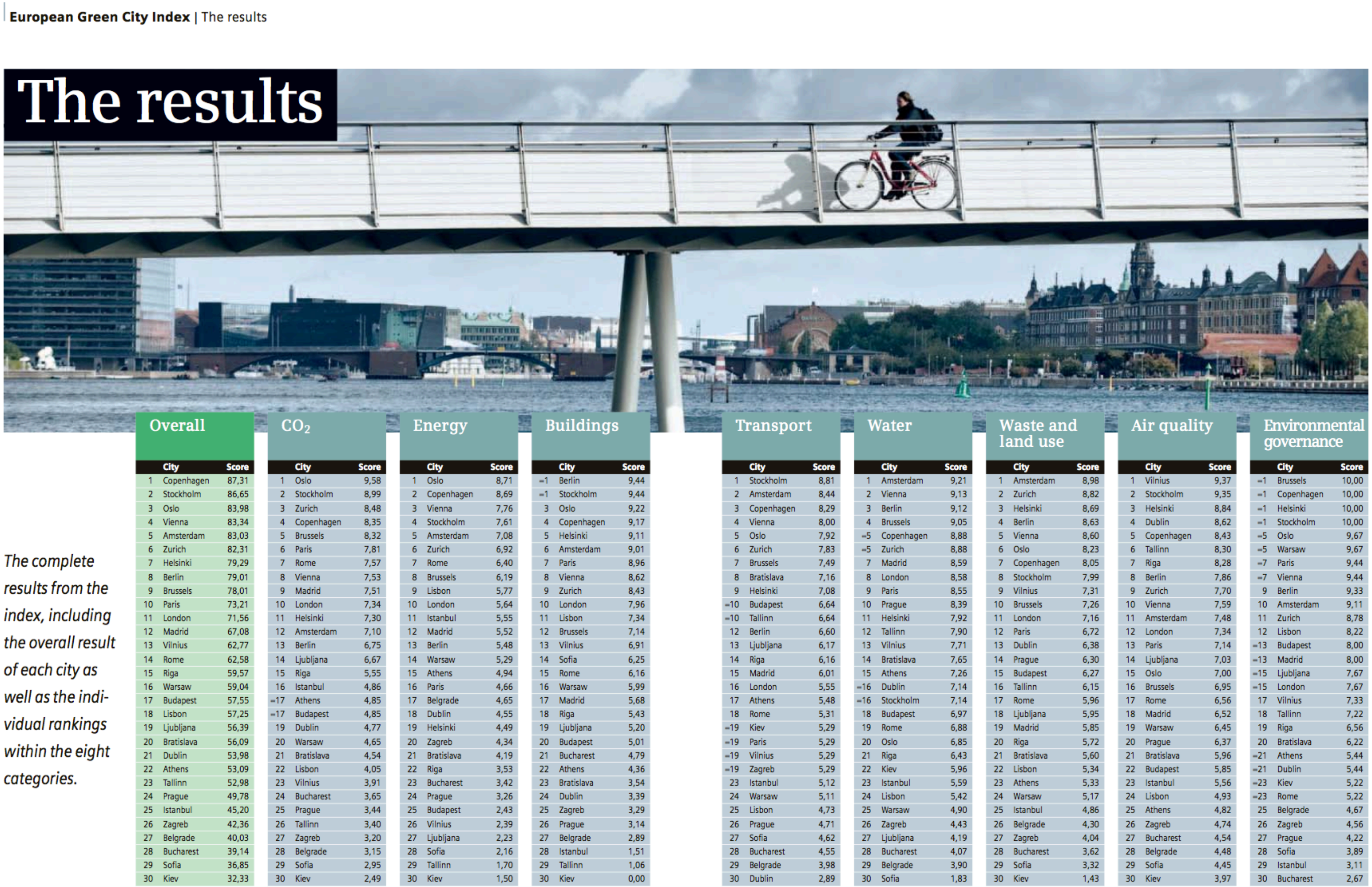


Great stuff, but limited potential...

More interesting:

- **Data Integration** & building **Data Workflows** from different Open Data sources!!!

Open Data Integration - A concrete use case:



European Green City Index | The results

The results

The complete results from the index, including the overall result of each city as well as the individual rankings within the eight categories.

Overall

	City	Score
1	Copenhagen	87,31
2	Stockholm	86,65
3	Oslo	83,98
4	Vienna	83,34
5	Amsterdam	83,03
6	Zurich	82,31
7	Helsinki	79,29
8	Berlin	79,01
9	Brussels	78,01
10	Paris	73,21
11	London	71,56
12	Madrid	67,08
13	Vilnius	62,77
14	Rome	62,58
15	Riga	59,57
16	Warsaw	59,04
17	Budapest	57,55
18	Lisbon	57,25
19	Ljubljana	56,39
20	Bratislava	56,09
21	Dublin	53,98
22	Athens	53,09
23	Tallinn	52,98
24	Prague	49,78
25	Istanbul	45,20
26	Zagreb	42,36
27	Belgrade	40,03
28	Bucharest	39,14
29	Sofia	36,85
30	Kiev	32,33

CO_2

	City	Score
1	Oslo	9,58
2	Stockholm	8,99
3	Zurich	8,48
4	Copenhagen	8,35
5	Brussels	8,32
6	Paris	7,81
7	Rome	7,57
8	Vienna	7,53
9	Madrid	7,51
10	London	7,34
11	Helsinki	7,30
12	Amsterdam	7,10
13	Berlin	6,75
14	Ljubljana	6,67
15	Riga	5,55
16	Istanbul	4,86
=17	Athens	4,85
=17	Budapest	4,85
19	Dublin	4,77
20	Warsaw	4,65
21	Bratislava	4,54
22	Lisbon	4,05
23	Vilnius	3,91
24	Bucharest	3,65
25	Prague	3,44
26	Tallinn	3,40
27	Zagreb	3,20
28	Belgrade	3,15
29	Sofia	2,95
30	Kiev	2,49

Energy

	City	Score
1	Oslo	8,71
2	Copenhagen	8,69
3	Vienna	7,76
4	Stockholm	7,61
5	Amsterdam	7,08
6	Zurich	6,92
7	Rome	6,40
8	Brussels	6,19
9	Lisbon	5,77
10	London	5,64
11	Istanbul	5,55
12	Madrid	5,52
13	Berlin	5,48
14	Warsaw	5,29
15	Athens	4,94
16	Paris	4,66
17	Belgrade	4,65
18	Dublin	4,55
19	Helsinki	4,49
20	Zagreb	4,34
21	Bratislava	4,19
22	Riga	3,53
23	Bucharest	3,42
24	Prague	3,26
25	Budapest	2,43
26	Vilnius	2,39
27	Ljubljana	2,23
28	Sofia	2,16
29	Tallinn	1,70
30	Kiev	1,50

Buildings

	City	Score
=1	Berlin	9,44
=1	Stockholm	9,44
3	Oslo	9,22
4	Copenhagen	9,17
5	Helsinki	9,11
6	Amsterdam	9,01
7	Paris	8,96
8	Vienna	8,62
9	Zurich	8,43
10	London	7,96
11	Lisbon	7,34
12	Brussels	7,14
13	Vilnius	6,91
14	Sofia	6,25
15	Rome	6,16
16	Warsaw	5,99
17	Madrid	5,68
18	Riga	5,43
19	Ljubljana	5,20
20	Budapest	5,01
21	Bucharest	4,79
22	Athens	4,36
23	Bratislava	3,54
24	Dublin	3,39
25	Zagreb	3,29
26	Prague	3,14
27	Belgrade	2,89
28	Istanbul	1,51
29	Tallinn	1,06
30	Kiev	0,00

Transport

	City	Score
1	Stockholm	8,81
2	Amsterdam	8,44
3	Copenhagen	8,29
4	Vienna	8,00
5	Oslo	7,92
6	Zurich	7,83
7	Brussels	7,49
8	Bratislava	7,16
9	Helsinki	7,08
=10	Budapest	6,64
=10	Tallinn	6,64
12	Berlin	6,60
13	Ljubljana	6,17
14	Riga	6,16
15	Madrid	6,01
16	London	5,55
17	Athens	5,48
18	Rome	5,31
=19	Kiev	5,29
=19	Paris	5,29
=19	Vilnius	5,29
=19	Zagreb	5,29
23	Istanbul	5,12
24	Warsaw	5,11
25	Lisbon	4,73
26	Prague	4,71
27	Sofia	4,62
28	Bucharest	4,55
29	Belgrade	3,98
30	Dublin	2,89

Water

	City	Score
1	Amsterdam	9,21
2	Vienna	9,13
3	Berlin	9,12
4	Brussels	9,05
=5	Copenhagen	8,88
=5	Zurich	8,88
7	Madrid	8,59
8	London	8,58
9	Paris	8,55
10	Prague	8,39
11	Helsinki	7,92
12	Tallinn	7,90
13	Vilnius	7,71
14	Bratislava	7,65
15	Athens	7,26
=16	Dublin	7,14
=16	Stockholm	7,14
18	Budapest	6,97
19	Rome	6,88
20	Oslo	6,85
21	Riga	6,43
22	Kiev	5,96
23	Istanbul	5,59
24	Lisbon	5,42
25	Warsaw	4,90
26	Zagreb	4,43
27	Ljubljana	4,19
28	Bucharest	4,07
29	Belgrade	3,90
30	Sofia	1,83

Waste and land use

	City	Score
1	Amsterdam	8,98
2	Zurich	8,82
3	Helsinki	8,69
4	Berlin	8,63
5	Vienna	8,60
6	Oslo	8,23
7	Copenhagen	8,05
8	Stockholm	7,99
9	Vilnius	7,31
10	Brussels	7,26
11	London	7,16
12	Paris	6,72
13	Dublin	6,38
14	Prague	6,30
15	Budapest	6,27
16	Tallinn	6,15
17	Rome	5,96
18	Ljubljana	5,95
19	Madrid	5,85
20	Riga	5,72
21	Bratislava	5,60
22	Lisbon	5,34
23	Athens	5,33
24	Warsaw	5,17
25	Istanbul	4,86
26	Belgrade	4,30
27	Zagreb	4,04
28	Bucharest	3,62
29	Sofia	3,32
30	Kiev	1,43

Air quality

	City	Score
1	Vilnius	9,37
2	Stockholm	9,35
3	Helsinki	8,84
4	Dublin	8,62
5	Copenhagen	8,43
6	Tallinn	8,30
7	Riga	8,28
8	Berlin	7,86
9	Zurich	7,70
10	Vienna	7,59
11	Amsterdam	7,48
12	London	7,34
13	Paris	7,14
14	Ljubljana	7,03
15	Oslo	7,00
16	Brussels	6,95
17	Rome	6,56
18	Madrid	6,52
19	Warsaw	6,45
20	Prague	6,37
21	Bratislava	5,96
22	Budapest	5,85
23	Istanbul	5,56
24	Lisbon	4,93
25	Athens	4,82
26	Zagreb	4,74
27	Bucharest	4,54
28	Belgrade	4,48
29	Sofia	4,45
30	Kiev	3,97

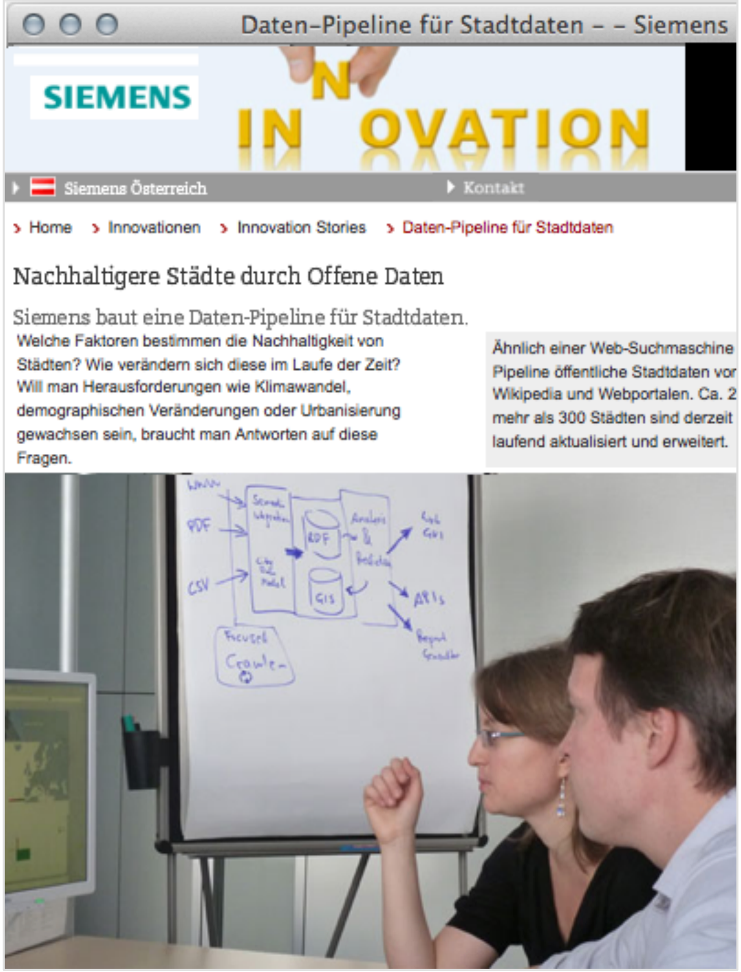
Environmental governance

	City	Score
=1	Brussels	10,00
=1	Copenhagen	10,00
=1	Helsinki	10,00
=1	Stockholm	10,00
=5	Oslo	9,67
=5	Warsaw	9,67
=7	Paris	9,44
=7	Vienna	9,44
9	Berlin	9,33
10	Amsterdam	9,11
11	Zurich	8,78
12	Lisbon	8,22
=13	Budapest	8,00
=13	Madrid	8,00
=15	Ljubljana	7,67
=15	London	7,67
17	Vilnius	7,33
18	Tallinn	7,22
19	Riga	6,56
20	Bratislava	6,22
=21	Athens	5,44
=21	Dublin	5,44
=23	Kiev	5,22
=23	Rome	5,22
25	Belgrade	4,67
26	Zagreb	4,56
27	Prague	4,22
28	Sofia	3,89
29	Istanbul	3,11
30	Bucharest	2,67

A concrete use case/running example: The "City Data Pipeline"

Idea – a "classic" Semantic **Web** use case!

- Regularly integrate **various** relevant Open Data **sources** (e.g. eurostat, UNData, ...)
- Make **integrated data** available for re-use



Daten–Pipeline für Stadtdaten – – Siemens

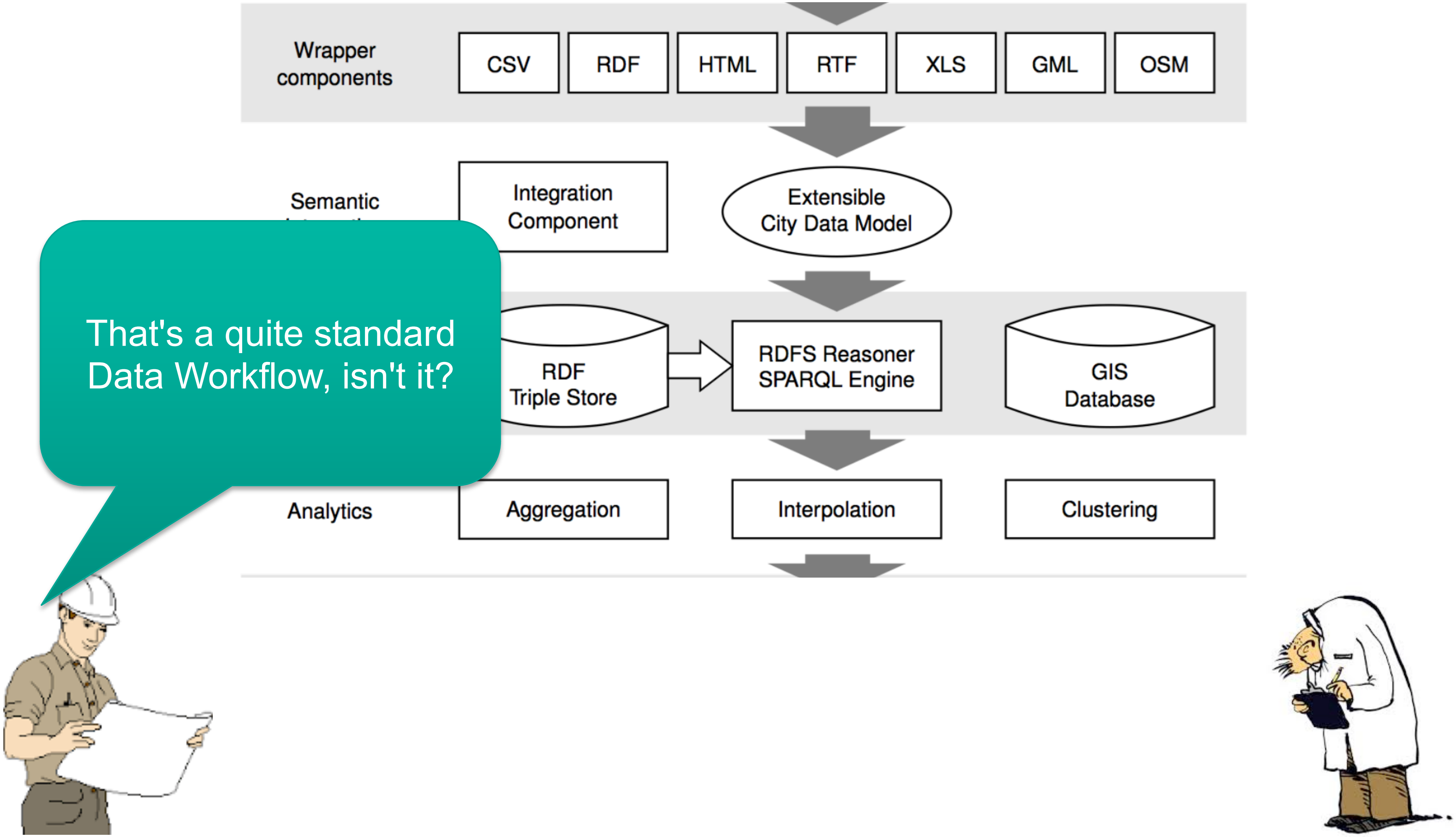
Nachhaltigere Städte durch Offene Daten

Siemens baut eine Daten-Pipeline für Stadtdaten.

Welche Faktoren bestimmen die Nachhaltigkeit von Städten? Wie verändern sich diese im Laufe der Zeit? Will man Herausforderungen wie Klimawandel, demographischen Veränderungen oder Urbanisierung gewachsen sein, braucht man Antworten auf diese Fragen.

Ähnlich einer Web-Suchmaschine
Pipeline öffentliche Stadtdaten vor
Wikipedia und Webportalen. Ca. 2
mehr als 300 Städten sind derzeit
laufend aktualisiert und erweitert.

A concrete use case:
The "City Data Pipeline" – a "fairly standard" data workflow



Wrapper components

CSV | RDF | HTML | RTF | XLS | GML | OSM

Semantic

Integration Component

Extensible City Data Model

RDF Triple Store

RDFS Reasoner SPARQL Engine

GIS Database

Analytics

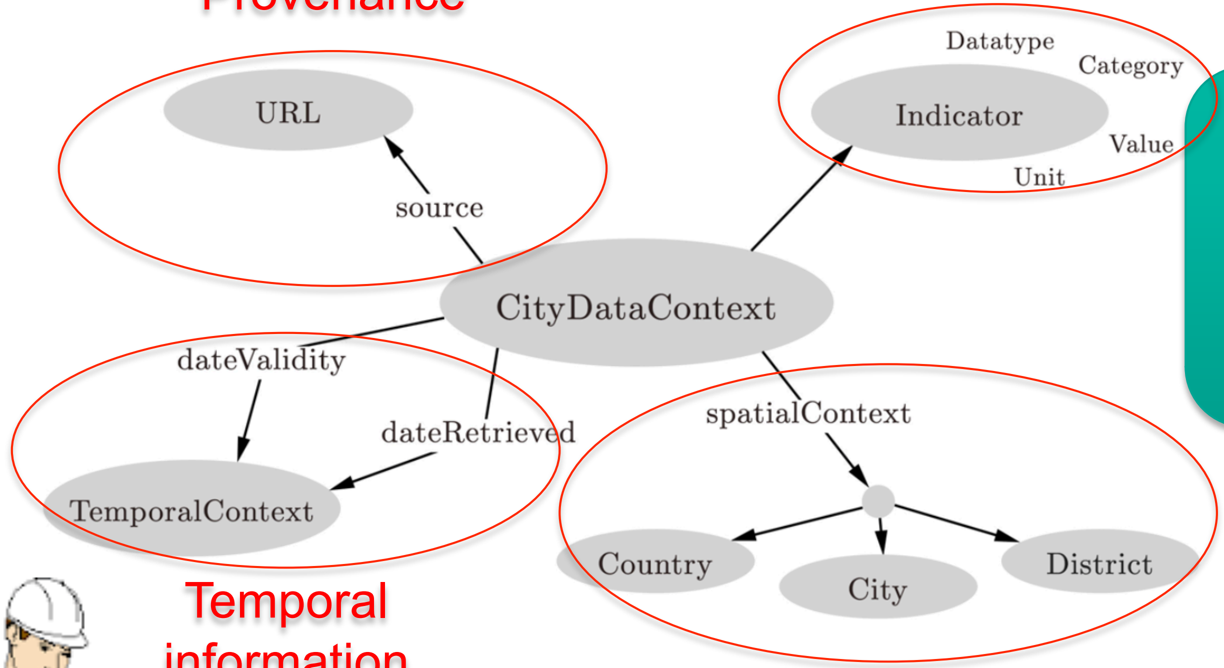
Aggregation

Interpolation

Clustering

A concrete use case:
The "City Data Pipeline" – a "fairly standard" data workflow

City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:



Provenance

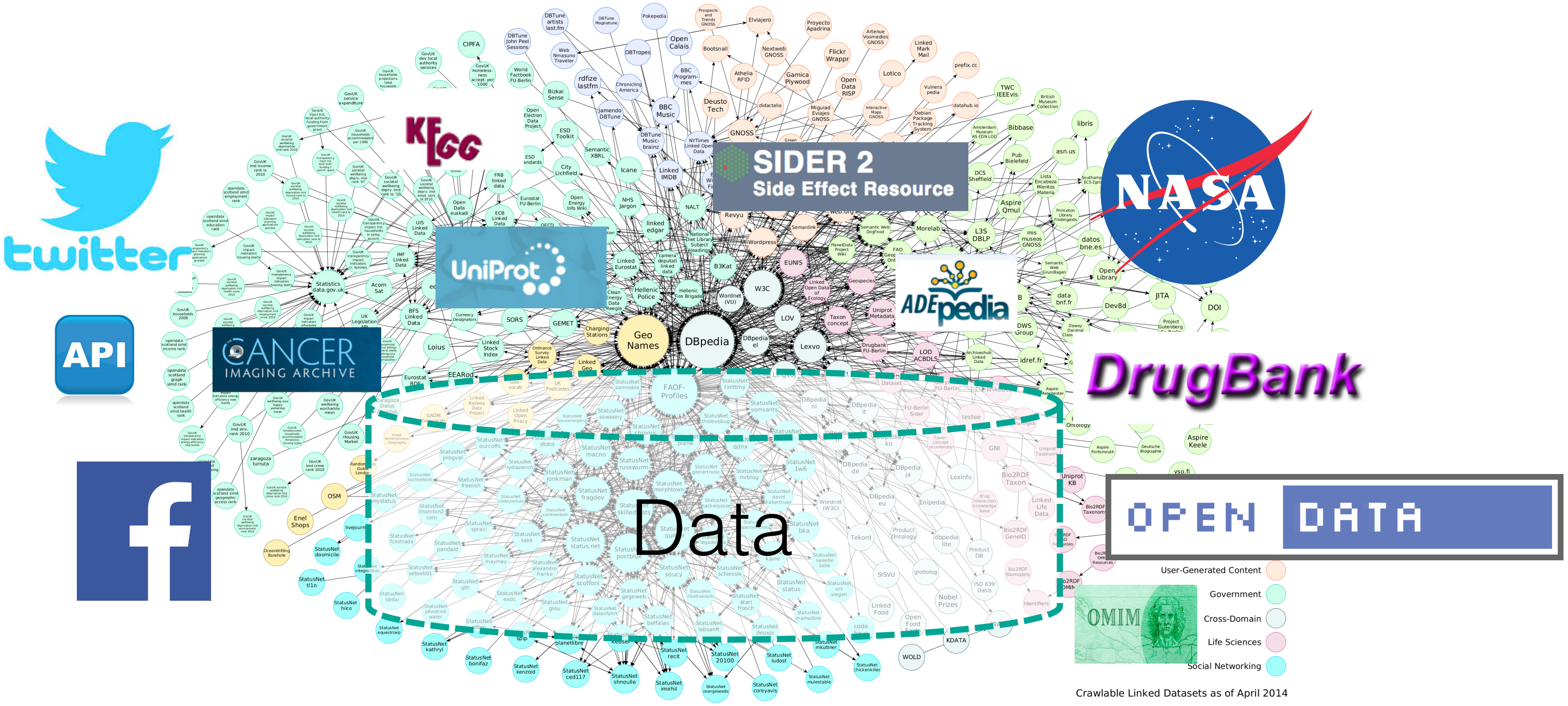
Indicators,
e.g. area in km2,
tons CO2/capita

Temporal
information

Spatial context

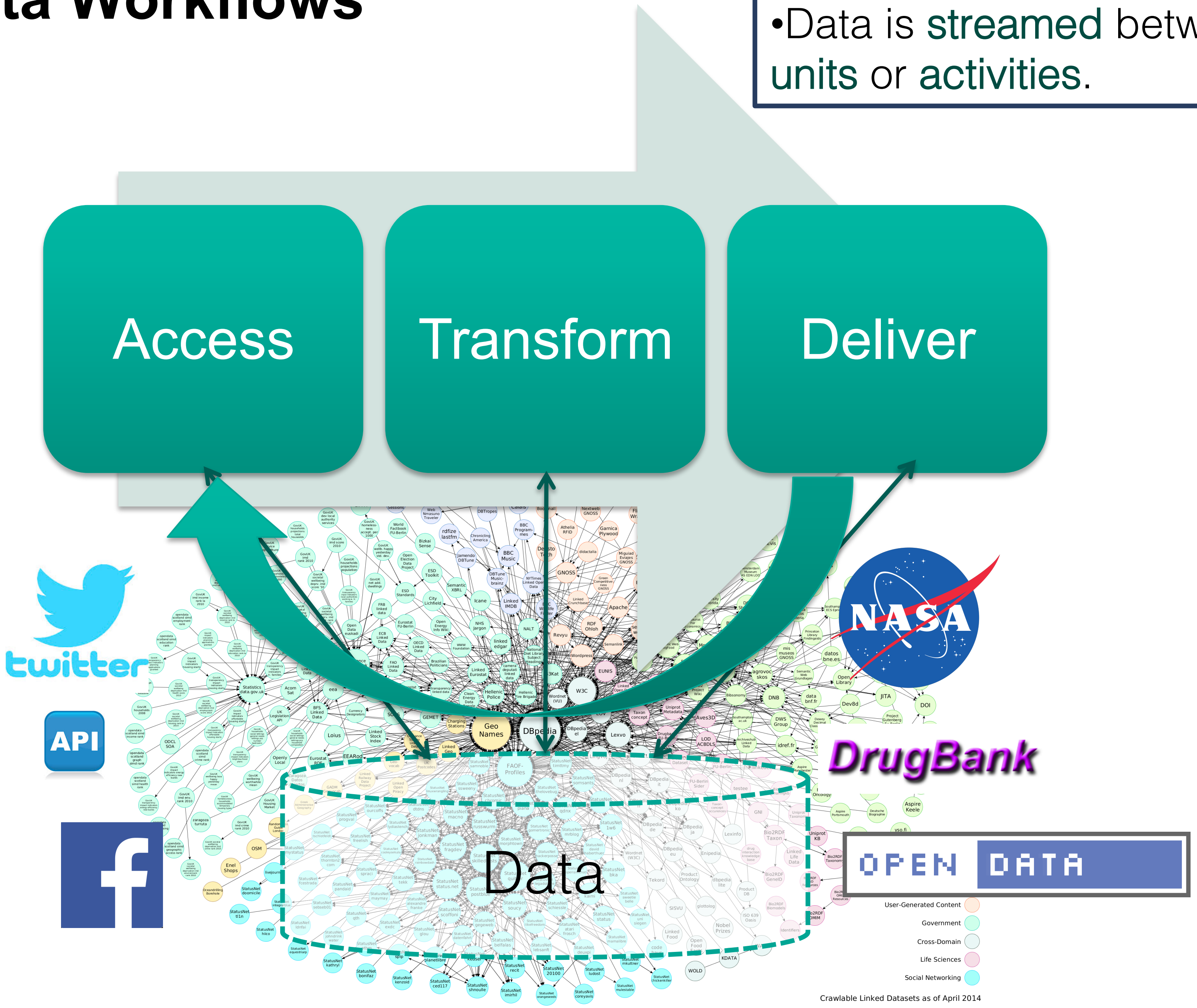
But we use and flexible Semantic integration using ontologies and reasoning!

So: What is a "standard data workflow"?



Data Workflows

- Well-defined functional units.
- Data is streamed between units or activities.



Different Views & Examples of "What is a Data Workflow:

Different Views & Examples:
1/5 „Classic" ETL-Process in Datawarehousing

Wikipedia:

- *In computing, **Extract, Transform and Load** (**ETL**) refers to a process in database usage and especially in data warehousing that:*
  - *Extracts data from homogeneous or heterogeneous data sources*
  - ***Cleansing**: deduplication, inconsistencies, missing data,...*
  - *Transforms the data for storing it in proper format or structure for querying and analysis purpose*
  - *Loads it into the final target (database, more specifically, operational data store, data mart, or data warehouse)*

"*Hard-wired*" Data integration

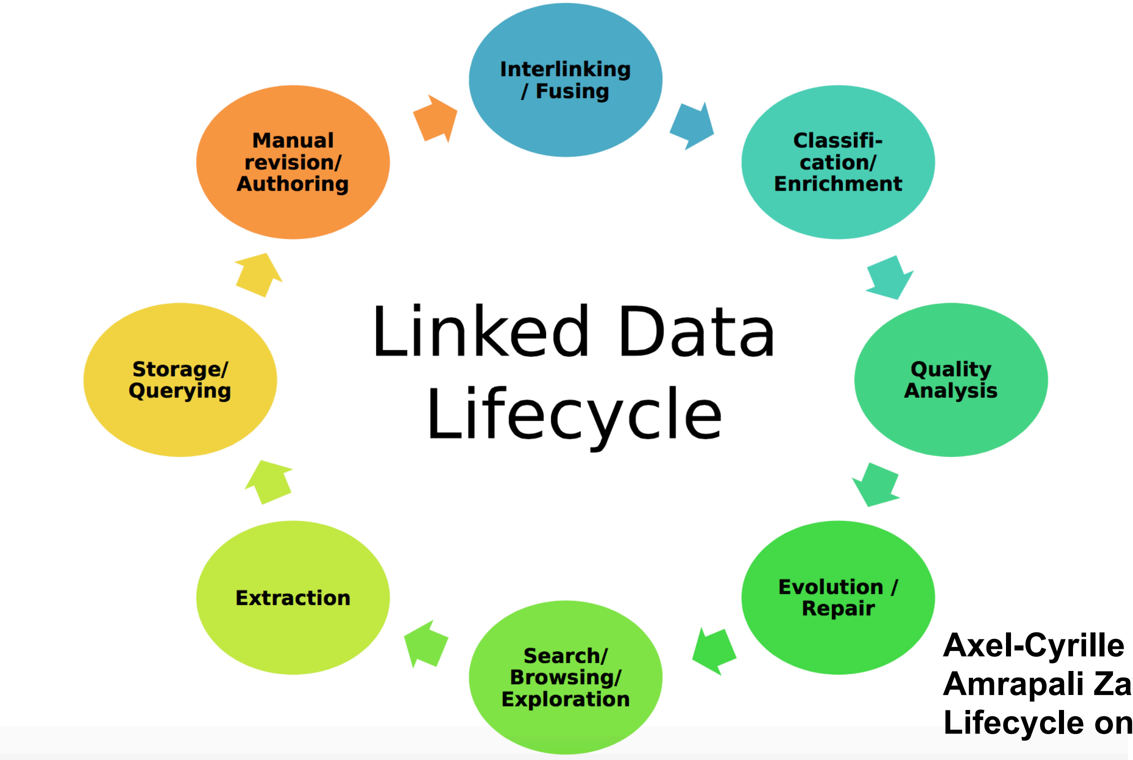
- Typically assumes: fixed, static pipeline, fixed final schema in the final DB/DW
- Cleansing sometimes viewed as a part of Transform, sometimes not.
- Typically assumes complete/clean data at the "Load" stage
- Aggregation sometimes viewed as a part of tranformation, sometimes higher up in the Datawarehouse access layer (OLAP)

- WARNING: At each stage, things can go wrong! Filtering/aggregation may bias the data!

- References:[Golfarelli, Rizzi, 2009]
  - https://en.wikipedia.org/wiki/Extract,_transform,_load
  - https://en.wikipedia.org/wiki/Staging_%28data%29#Functions

Different Views & Examples:
2/5 Or is it rather a Lifecycle...

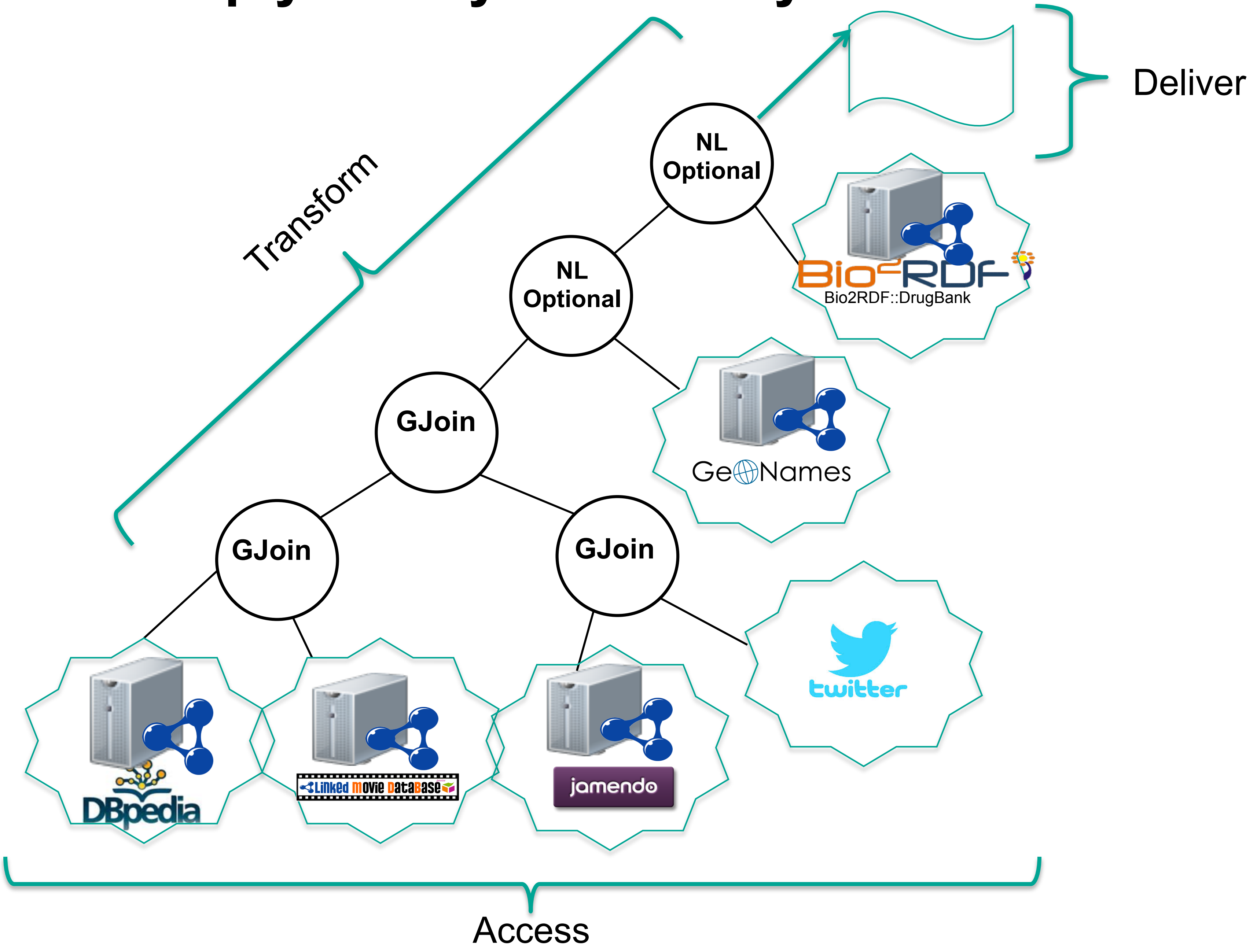
- E.g. good example: Linked Data Lifecycle



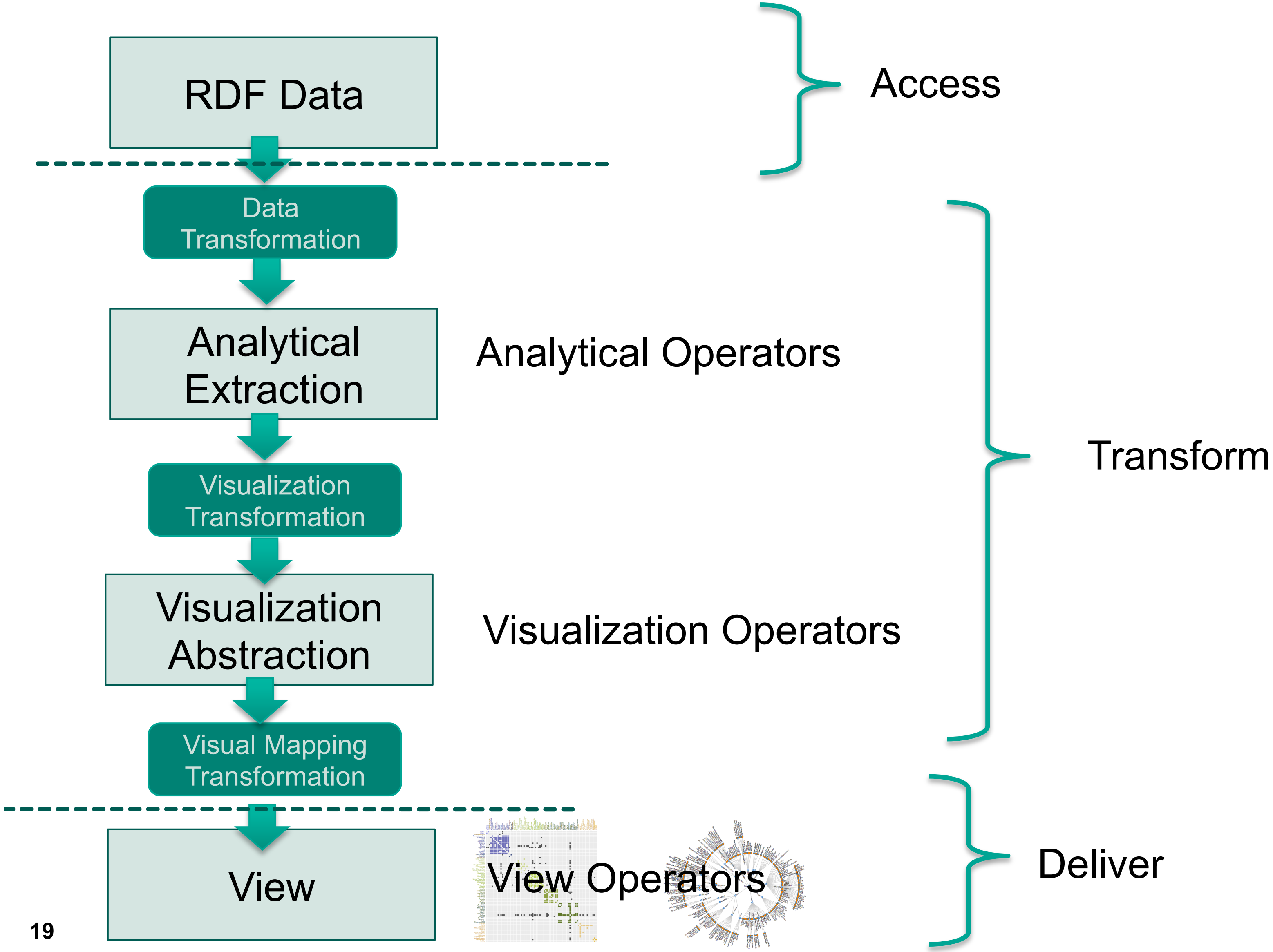
Axel-Cyrille Ngonga Ngomo, Sören Auer, Jens Lehmann, Amrapali Zaveri. Introduction to Linked Data and Its Lifecycle on the Web. ReasoningWeb. 2014

- **NOTE:** Independent of whether Linked Data or other sources, you need to revisit/revalidate your workflow, either for improving it or for maintainance (sources changing, source formats changing, etc.)

Different views & examples:
3/5 Or Simply a "Physical Query Plan"?



Different Views & Examples:
4/5 Linked Data Visualization Model



RDF Data

Access

Data Transformation

Analytical Extraction

Analytical Operators

Transform

Visualization Transformation

Visualization Abstraction

Visualization Operators

Visual Mapping Transformation

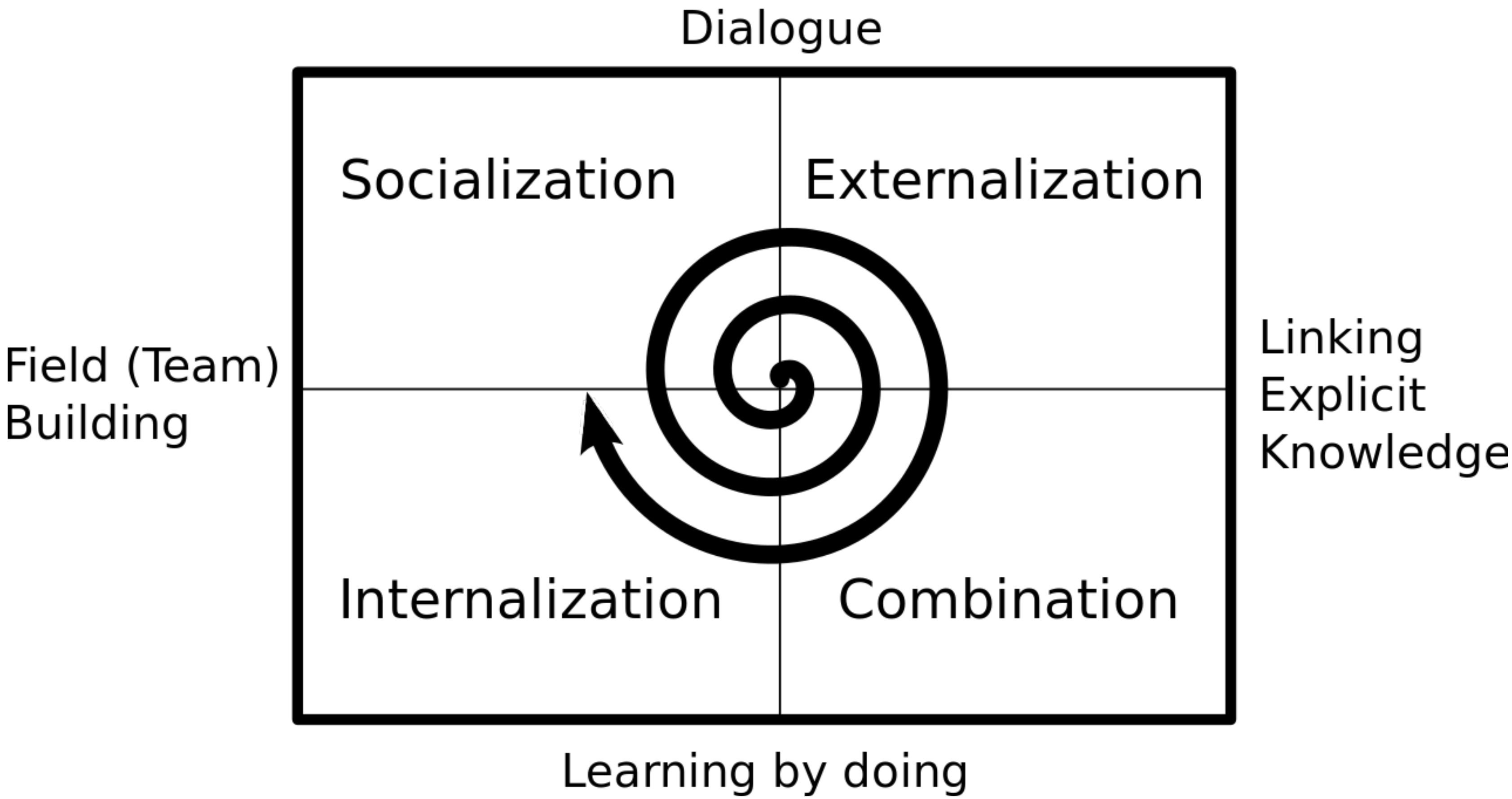
View

View Operators

Deliver

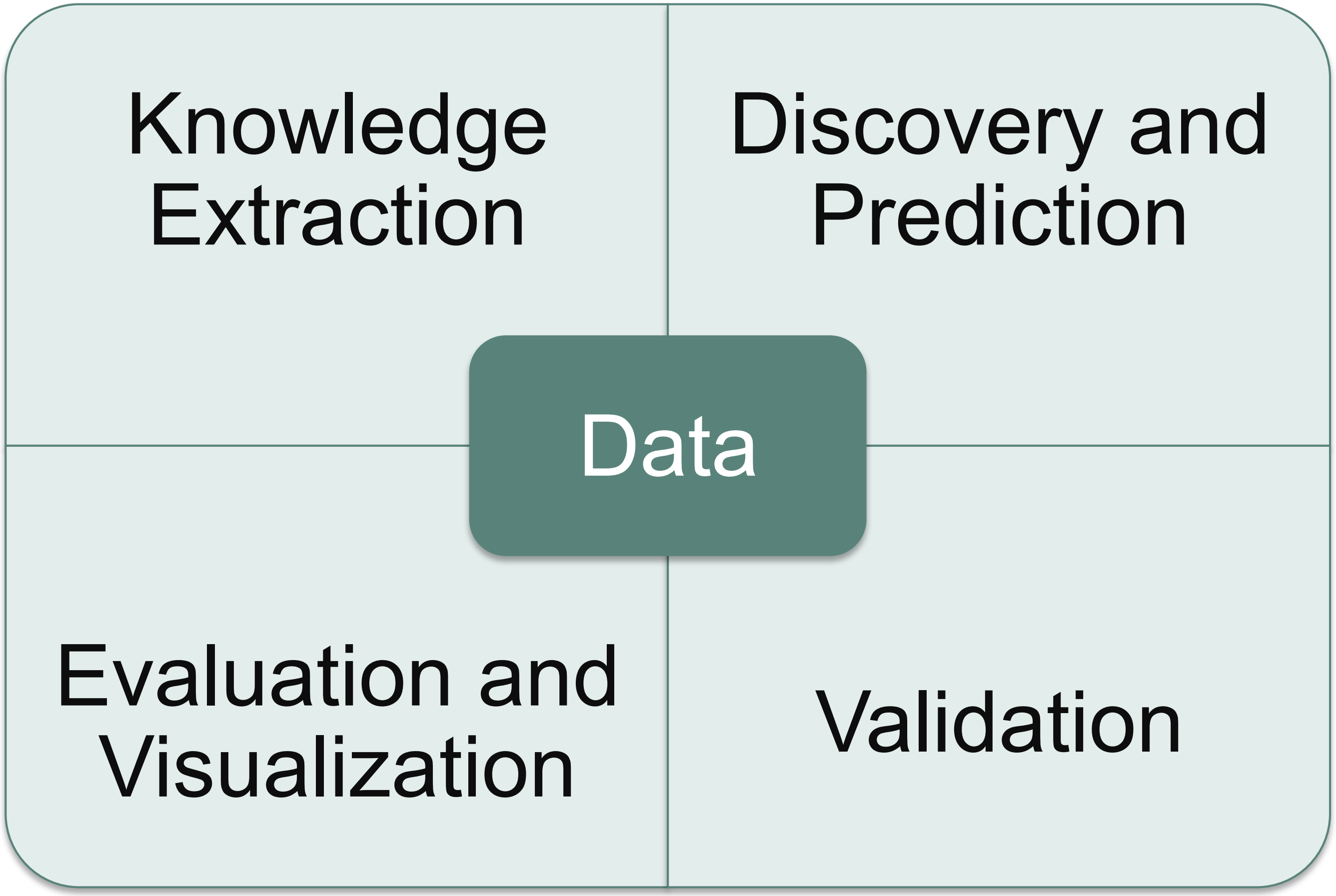
J. Brunetti , S. Auer, R. García,"The Linked Data Visualization Model'.

Different Views & Examples:
5/5 From a Knowledge-centric Approach…



[Nonaka & Takeuchi, 1995]

Different Views & Examples:
5/5 ... towards a Data-centric Approach



General challenges to be addressed

Distributed data sources

Non-standard processing

Semantic heterogeneity

Naming ambiguity

Uncertainty and evolving concepts

Specific Steps (non-exhaustive, overlapping!)

- Extraction
- Inconsistency handling
- Incompleteness handling (sometimes called "Enrichment")
- Data Integration (alignment, source reconciliation)
- Aggregation
- Cleansing (removing outliers)
- Deduplication/Interlinking (could involve "triplification")
- Change dedection (Maintainance/Evolution)
- Validation (quality anaysis)
- Visualization

Tools and current approaches support you **partially** in different parts of these steps.... Bad news: there is no "one-size-fits-all" solution.

Some Tools (again, exemplary):

- Linux-commandline Tools: curl, sed, awk, + postgresql does a good job in many cases...
- LOD2 stack, stack of tools for integrating and generating Linked Data, http://stack.lod2.eu/
  - e.g., SILK http://silk-framework.com/ (Interlinking)
- KARMA (extraction, data integration) http://usc-isi-i2.github.io/karma/
- XSPARQL (extraction from XML and JSON/triplicifation) http://sourceforge.net/projects/xsparql/
  - Seel also: https://ai.wu.ac.at/~polleres/20140826xsparql_st.etienne/

Outline

- Motivation
  - Integrating (Open) Data from different sources
  - Not only Linked Data
  - Data workflows and Open data in the context of rise of Big Data
- What is a "Data Workflow"?
  - Diefferent Views of Data Workflows in the context of the Semantic Web
  - Key steps involved
  - Tools?
- **Data Integration Systems**
  - Wrappers vs. Mediators
  - LAV vs. GAV
  - Data Integration using Ontologies – Query rewriting vs. Materialisation
- Challenges:
  - How to find Rules and ontologies?
  - Data Quality & Incomplete Data
  - Maintainance/evolution/sustainability of Data Workflows

Data Integration Systems[Lenzerini2002]

- IS=<O,S,M>
- Let O be a set of general concepts in a general schema (virtual).
- Let S={S1,..,Sn} be a set of data sources.
- Let M be a set of mappings between sources in S and general concepts in O.

cf. [Lenzerini 2002]

Global Schema

(**financial** rdf:type rdf:Property).

(**climate** rdf:type rdf:Property).

(**rating** rdf:type rdf:Property).

(**financial** rdfs:subPropertyOf **rating**).

(**climate** rdfs:subPropertyOf **rating**).

(**euroCity** rdf:type rdfs:Class).

(**amCity** rdf:type rdfs:Class)

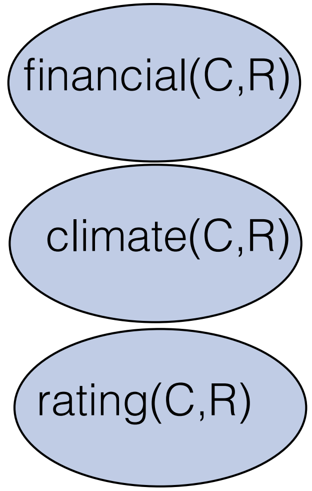
(**afCity** rdf:type rdfs:Class)

Global Schema

(**financial** rdf:type rdf:Property).

(**climate** rdf:type rdf:Property).

(**rating** rdf:type rdf:Property).

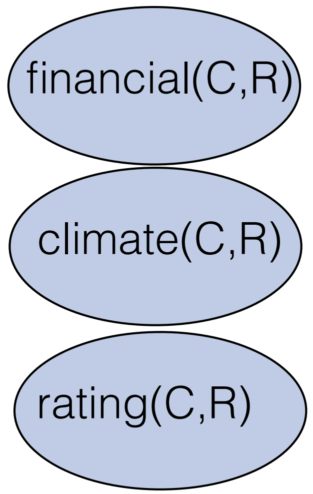


Global Schema

(**financial** rdf:type rdf:Property).

(**climate** rdf:type rdf:Property).

(**rating** rdf:type rdf:Property).



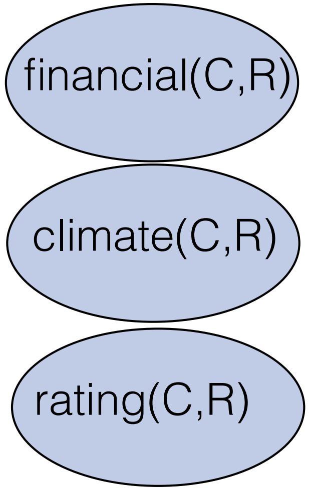
(**financial** rdfs:subPropertyOf **rating**).
(**climate** rdfs:subPropertyOf **rating**).

Global Schema

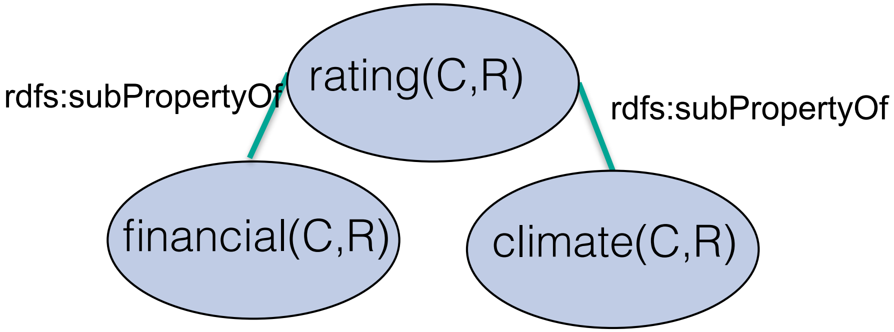
(**financial** rdf:type rdf:Property).

(**climate** rdf:type rdf:Property).

(**rating** rdf:type rdf:Property).



(**financial** rdfs:subPropertyOf **rating**).
(**climate** rdfs:subPropertyOf **rating**).

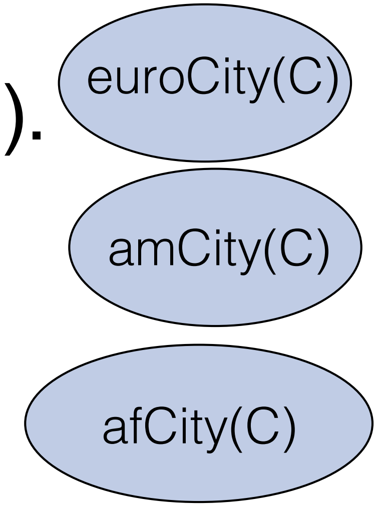


Global Schema

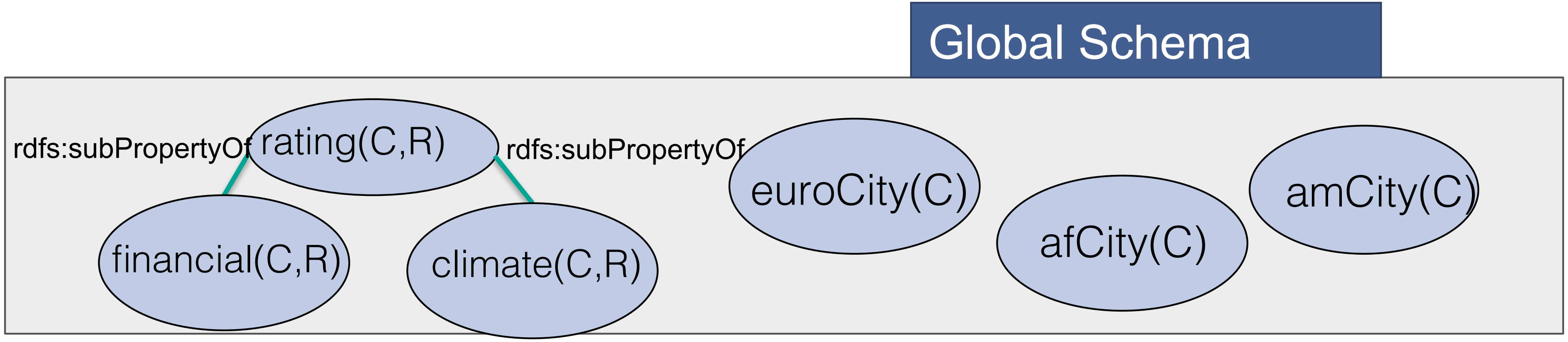
(**euroCity** rdf:type rdfs:Class).

(**amCity** rdf:type rdfs:Class)

(**afCity** rdf:type rdfs:Class)



Integration Systems



Source Schema

(***amFinancial*** rdf:type rdf:Property).

(***euClimate*** rdf:type rdf:Property).

(***tunisRating*** rdf:type rdf:Property).

(***similarFinancial*** rdf:type rdf:Property).

IDB
Inter-American Development Bank

eurostat

The World Bank

amFinancial(C,R) provides the financial rating R of an American city C.
euClimate(C,R) provides the climate rating R of an European city C.
tunisRating(T,R) tells the ratings R (T is climate and financial) of Tunis.
similarFinancial(C1,C1) relates two American cities C1 and C2 that have the same financial rating.

Integration Systems

euClimate(C,R)

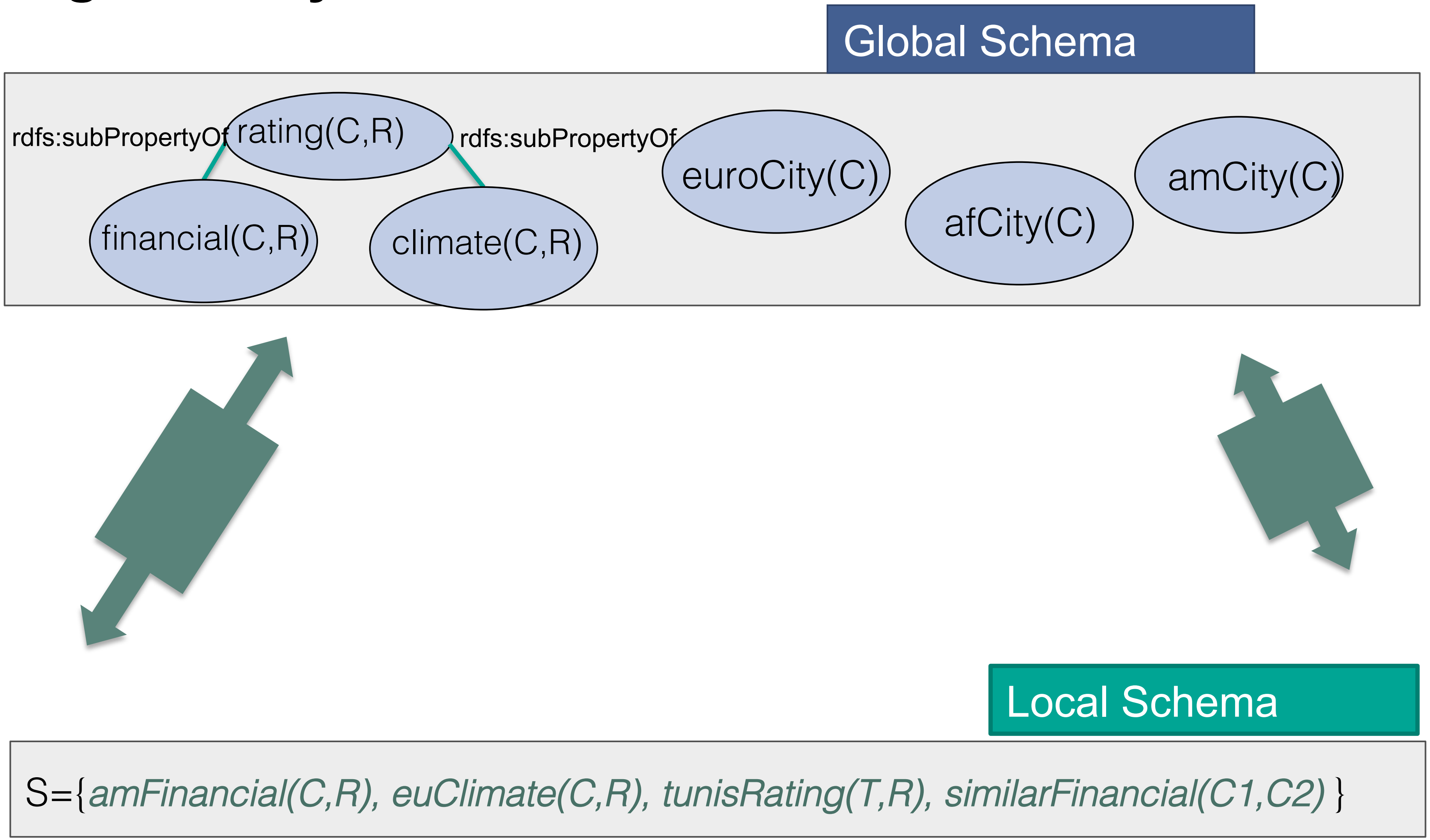
similarFinancial(C1,C2)

amFinancial(C,R)

tunisRating(T,R)

amFinancial(C,R) provides the financial rating R of an American city C.
euClimate(C,R) provides the climate rating R of an European city C.
tunisRating(T,R) tells the ratings R (T is climate and financial) of Tunis.
similarFinancial(C1,C1) relates two American cities C1 and C2 that have the same financial rating.

Integration Systems



Global Schema

rdfs:subPropertyOf rating(C,R) rdfs:subPropertyOf

financial(C,R)

climate(C,R)

euroCity(C)

afCity(C)

amCity(C)

Local Schema

S={*amFinancial(C,R), euClimate(C,R), tunisRating(T,R), similarFinancial(C1,C2)* }

Integration Systems

IS=<O,S,M>

Global-as-View (GAV):

- Concepts in the Global Schema (O) are defined in terms of combinations of Sources (S).

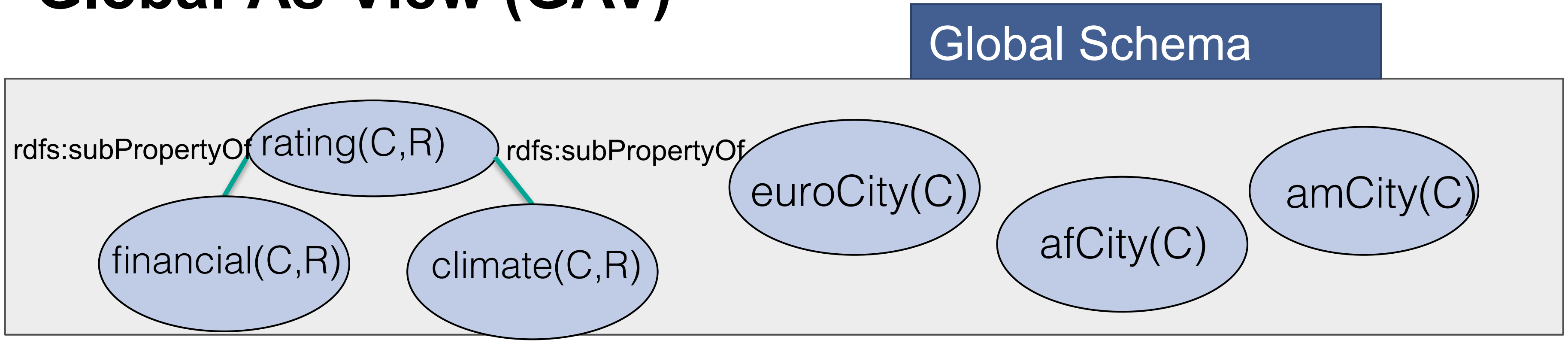
Local-As-View (LAV):

- Sources in S are defined in terms of combinations of Concepts in O.

Global- & Local-As-View (GLAV):

- Combinations of concepts in the Global Schema (O) are defined in combinations of Sources (S).

Global-As-View (GAV)

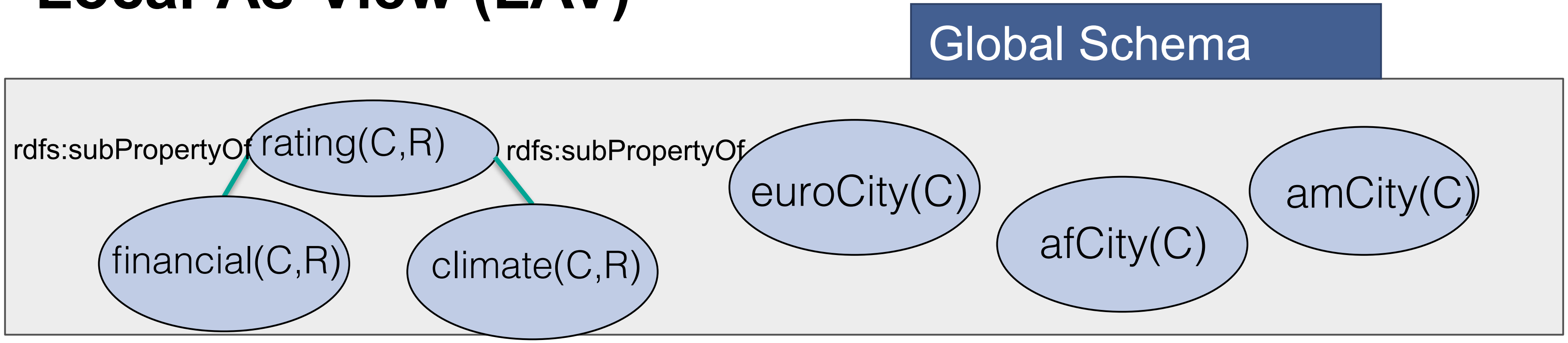


Local Schema

S={*amFinancial(C,R), euClimate(C,R), tunisRating(T,R), similarFinancial(C1,C2)* }

α0: amCity(C):-*amFinancial(C,R).*
α1: financial(C,R):-*amFinancial(C,R)*.
α2: euroCity(C):-*euClimate(C,R)*.
α3: climate(C,R):-*euClimate(C,R)*.
α4: financial("Tunis",R):-*tunisRating("financial",R).*
α5: climate("Tunis",R):-*tunisRating("climate",R)*
α6: afCity("Tunis").
α7: amCity(C1):-*similarFinancial(C1,C2).*
α8: amCity(C2):-*similarFinancial(C1,C2).*
α9: financial(C1,R):-*similarFinancial(C1,C2)*, *amFinancial(C2,R)*.

Local-As-View (LAV)

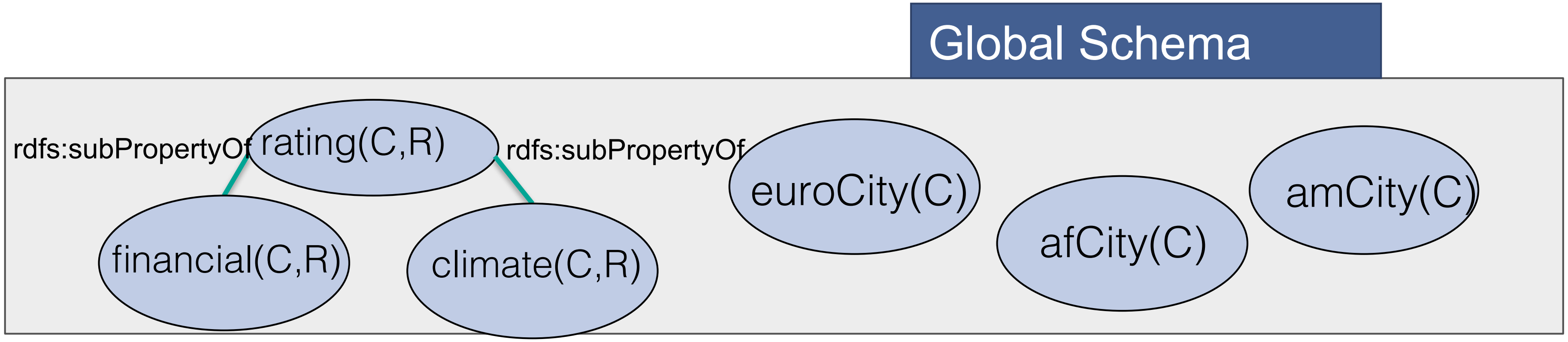


Local Schema

S={*amFinancial(C,R), euClimate(C,R), tunisRating(T,R), similarFinancial(C1,C2)* }

α0:*amFinancial(C,R)*:-amCity(C),financial(C,R).
α1:*euClimate(C,R)*:-euCity(C),climate(C,R).
α2:*tunisRating(“financial”,R)*:-afCity(“Tunis”),financial(“Tunis”,R).
α3:*tunisRating(“climate”,R)*:-afCity(“Tunis”),climate(“Tunis”,R).
α4:*similarFinancial(C1,C2)*:-amCity(C1),amCity(C2),
financial(C1,R),financial(C2,R).

Global and Local-As-View (GLAV)



Local Schema

S={*amFinancial(C,R), euClimate(C,R), tunisRating(T,R), similarFinancial(C1,C2)* }

α0: *amFinancial(C1,R)*,*similarFinancial(C1,C2):-*
amCity(C1),amCity(C2),financial(C1,R),financial(C2,R).

Query Rewriting GAV

- A query Q in terms of the global schema elements in O.
- **Problem:** Rewrite Q into a query Q' expressed in sources in S.

Example GAV:

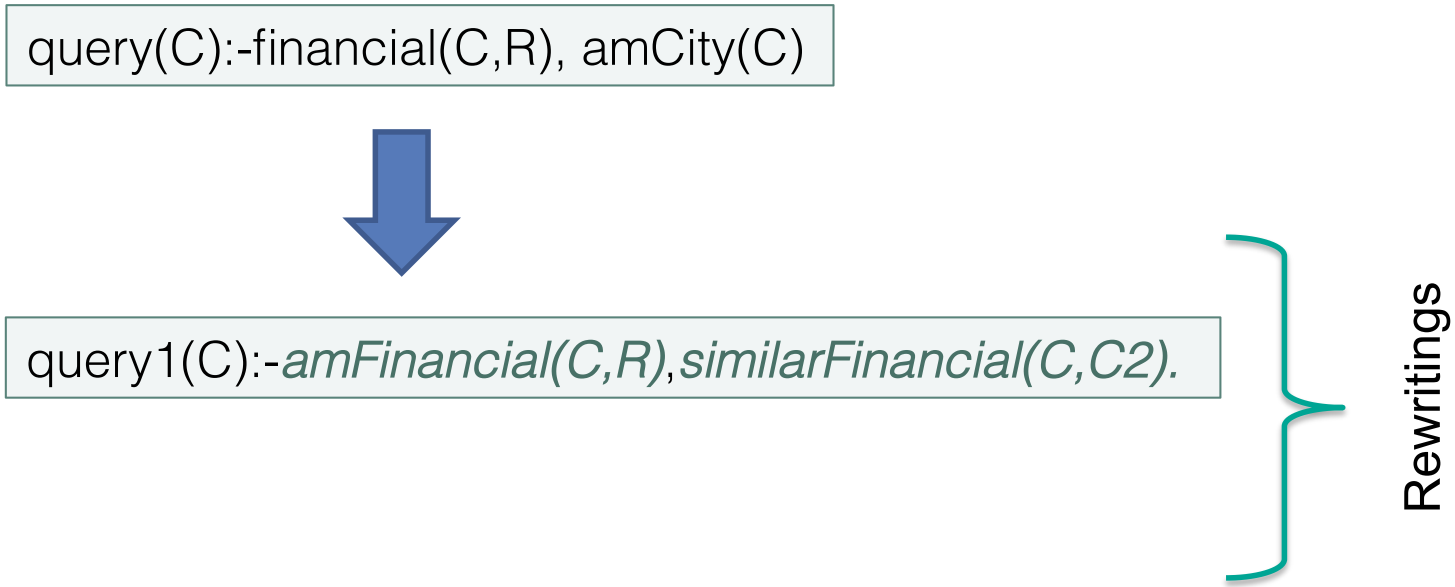
```
query(C):-financial(C,R), amCity(C)
```

α0: amCity(C):-*amFinancial(C,R).*
α1: financial(C,R):-*amFinancial(C,R)*.
α2: euroCity(C):-*euClimate(C,R)*.
α3: climate(C,R):-*euClimate(C,R)*.
α4: financial("Tunis",R):-*tunisRating("financial",R).*
α5: climate("Tunis",R):-*tunisRating("climate",R)*
α6: afCity("Tunis").
α7: amCity(C1):-*similarFinancial(C1,C2).*
α8: amCity(C2):-*similarFinancial(C1,C2).*
α9: financial(C1,R):-*similarFinancial(C1,C2)*, *amFinancial(C2,R)*.

Query Rewriting GAV

- A query Q in terms of the global schema elements in O.
- **Problem:** Rewrite Q into a query Q’ expressed in sources in S.

Example GAV:



query(C):-financial(C,R), amCity(C)

query1(C):-*amFinancial(C,R)*,*similarFinancial(C,C2).*

Rewritings

Query Rewriting GAV

- A query Q in terms of the global schema elements in O.
- **Problem:** Rewrite Q into a query Q’ expressed in sources in S.

Example GAV:

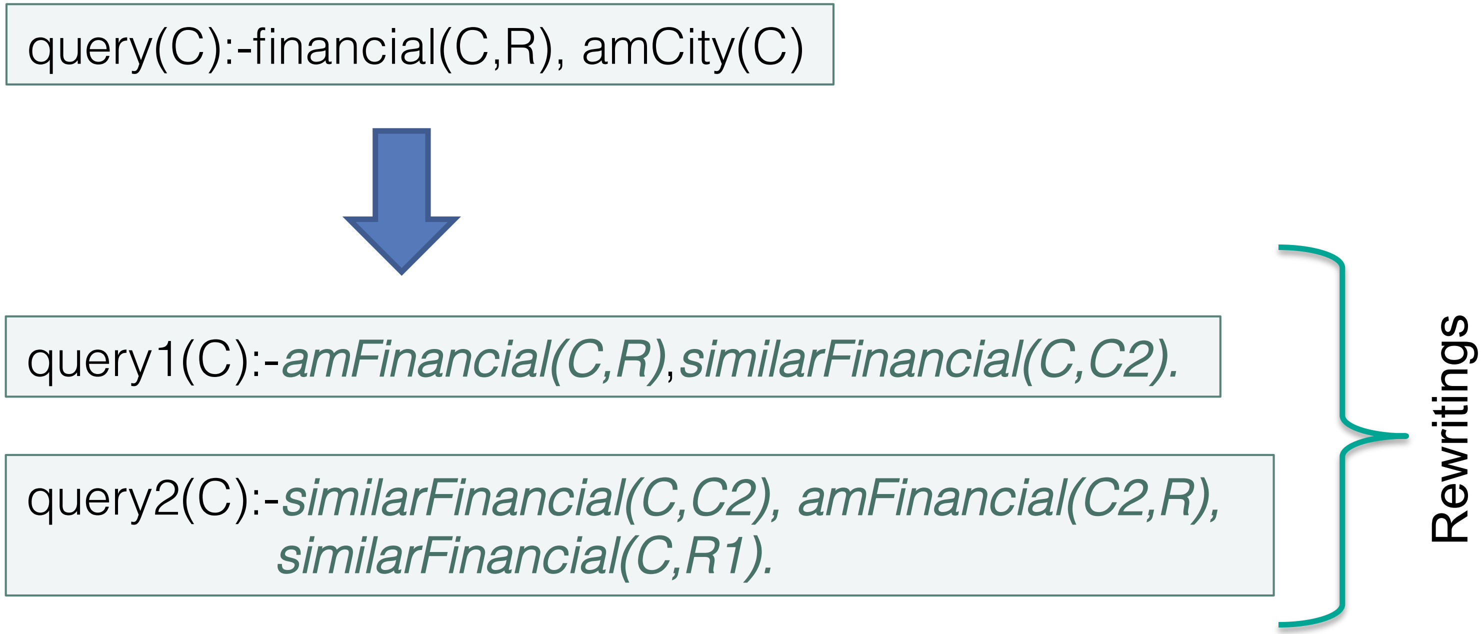
query(C):-financial(C,R), amCity(C)

α0: amCity(C):-*amFinancial(C,R).*
α1: financial(C,R):-*amFinancial(C,R)*.
α2: euroCity(C):-*euClimate(C,R)*.
α3: climate(C,R):-*euClimate(C,R)*.
α4: financial(“Tunis”,R):-*tunisRating(“financial”,R).*
α5: climate(“Tunis”,R):-*tunisRating(“climate”,R)*
α6: afCity(“Tunis”).
α7: amCity(C1):-*similarFinancial(C1,C2).*
α8: amCity(C2):-*similarFinancial(C1,C2).*
α9: financial(C1,R):-*similarFinancial(C1,C2)*, *amFinancial(C2,R)*.

Query Rewriting GAV

- A query Q in terms of the global schema elements in O.
- **Problem:** Rewrite Q into a query Q' expressed in sources in S.

Example GAV:



query(C):-financial(C,R), amCity(C)

query1(C):-*amFinancial(C,R)*,*similarFinancial(C,C2).*

query2(C):-*similarFinancial(C,C2), amFinancial(C2,R), similarFinancial(C,R1).*

Rewritings

Query Rewriting LAV

α0:*amFinancial(C,R)*:-amCity(C),financial(C,R).
α1:*euClimate(C,R)*:-euCity(C),climate(C,R).
α2:*tunisRating("financial",R)*:-afCity("Tunis"),financial("Tunis",R).
α3:*tunisRating("climate",R)*:-afCity("Tunis"),climate("Tunis",R).
α4:*similarFinancial(C1,C2)*:-amCity(C1),amCity(C2),
financial(C1,R),financial(C2,R).

Example LAV:

query(C):-financial(C,R), amCity(C)

query1(C):-*amFinancial(C,R).*

query2(C):-*similarFinancial(C,C2).*

query3(C):-*similarFinancial(C1,C).*

Rewritings

Query Rewriting GLAV

α0: *amFinancial(C1,R)*,*similarFinancial(C1,C2):-*
amCity(C1),amCity(C2),financial(C1,R),financial(C2,R).

Example GLAV:

query(C):-financial(C,R), amCity(C)

query1(C):-: *amFinancial(C,R)*,*similarFinancial(C,C2)*

Rewritings

Query Rewriting

DB is a Virtual Database with the instances of the elements in O.

Query Containment: Q’ ⊆Q ←→∀DB Q’(DB) ⊆Q(DB)

query1(C):-*amFinancial(C,R)*,*similarFinancial(C,C2).* ⊆ query(C):-financial(C,R), amCity(C)

Query Rewriting

DB is a Virtual Database with the instances of the elements in O.

Query Containment: $Q' \subseteq Q \leftarrow\rightarrow \forall DB\ Q'(DB) \subseteq Q(DB)$

query1(C):-*amFinancial(C,R)*,*similarFinancial(C,C2).* $\subseteq$ query(C):-financial(C,R), amCity(C)

Washington
NYC
Miami
Caracas
Lima

Source Database

$\subseteq$

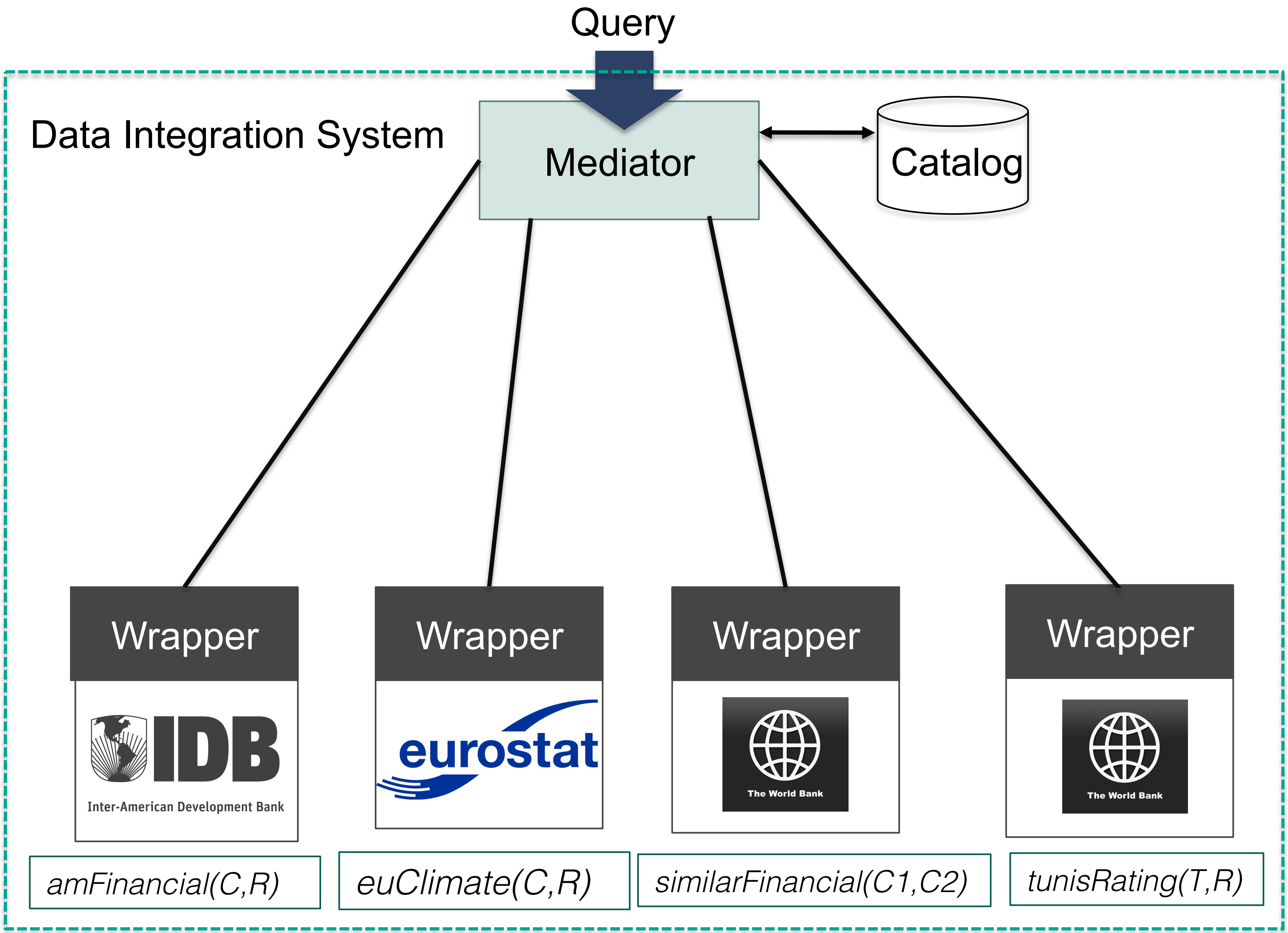
Washington
NYC
Miami
Caracas
Lima
……
Bogota
College Park
Quito

Virtual Database

Existing Approaches for LAV Query Rewriting

- Bucket Algorithm [Levy & Rajaraman & Ullman 1996]
- Inverse Rules Algorithm [Duscka & Genesereth 1997]
- MiniCom Algorithm [Pottinger & Halevy 2001]
- MDCSAT [Arvelo & Bonet & Vidal 2006]
- SSSAT [Izquierdo & Vidal & Bonet 2011]
- GQR [Konstantinidis & Ambite, 2011]
- IQR [Vidal & Castillo 2015]

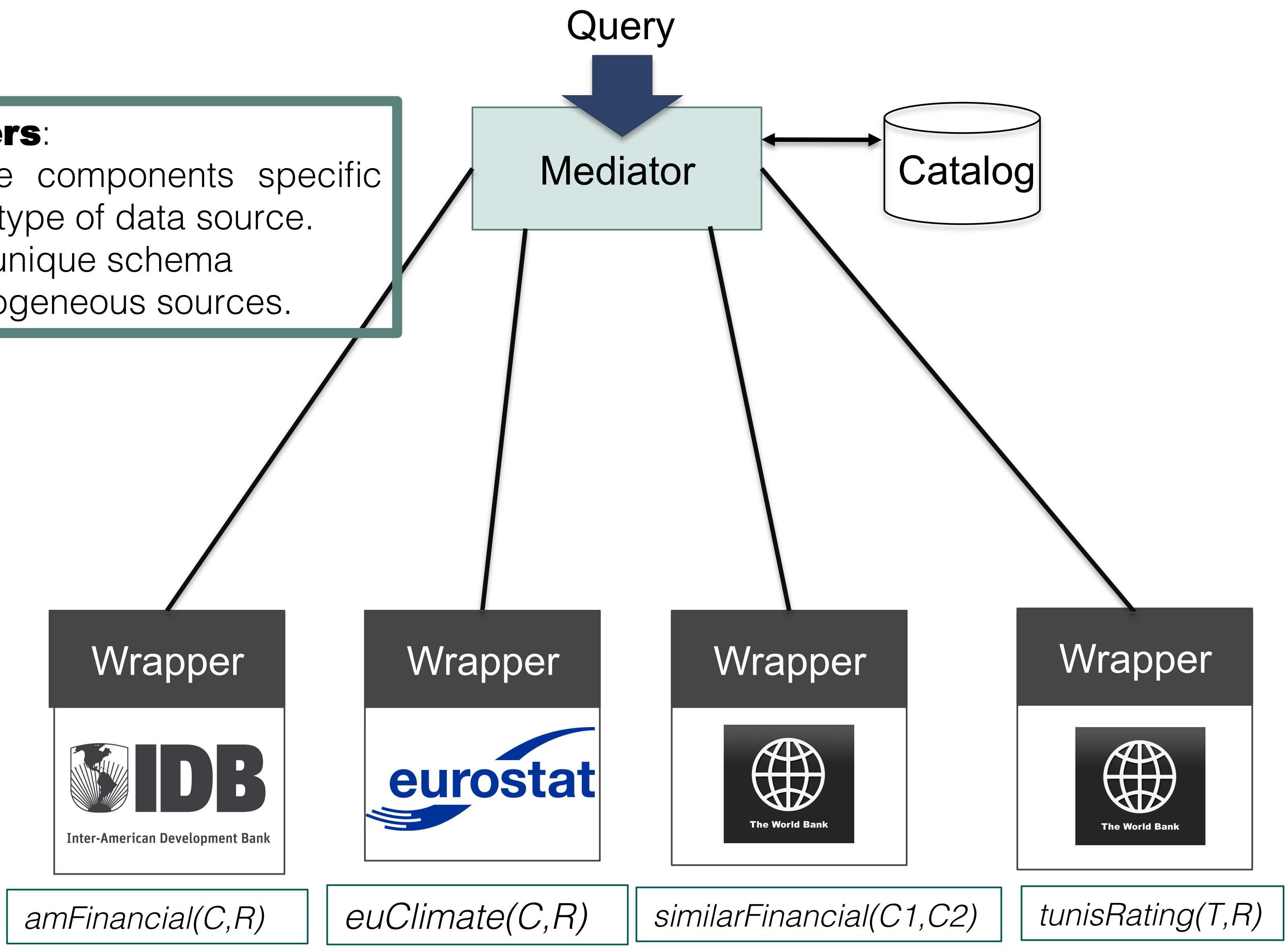
The Mediator and Wrapper Architecture [Wiederhold92]



[Wiederhold92]Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)

The Mediator and Wrapper Architecture [Wiederhold92]

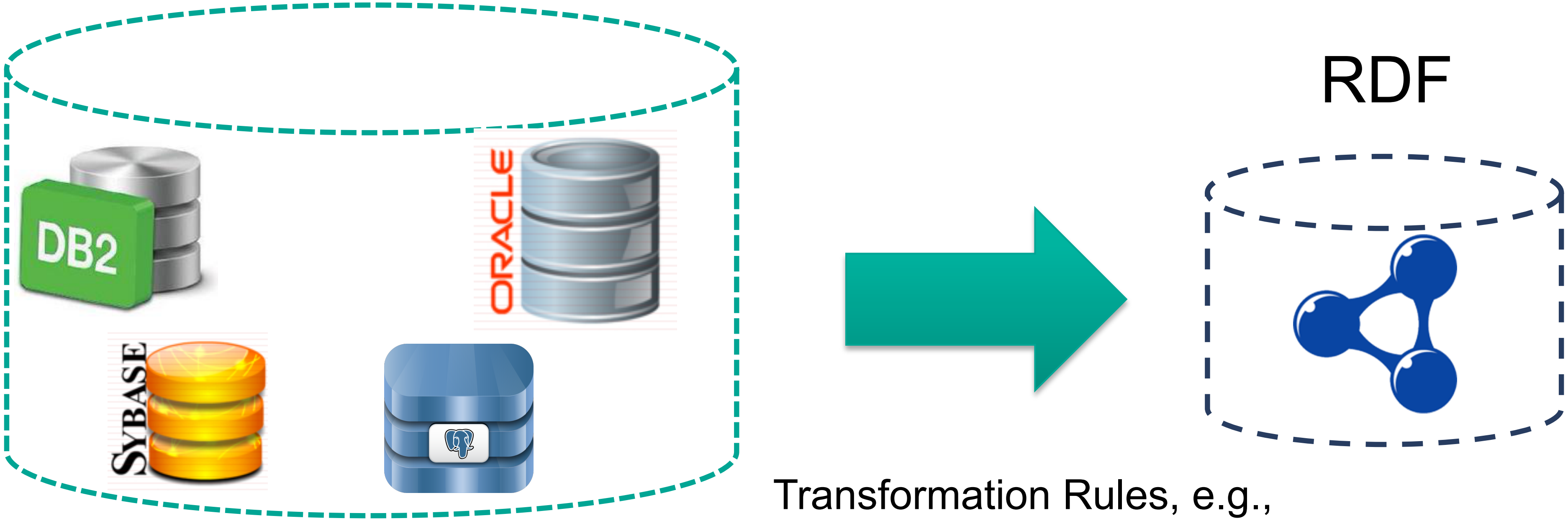
Wrappers:
- Software components specific for each type of data source.
- Export unique schema for heterogeneous sources.



[Wiederhold92]Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)

Wrappers for RDF Data

RDB2RDF Systems

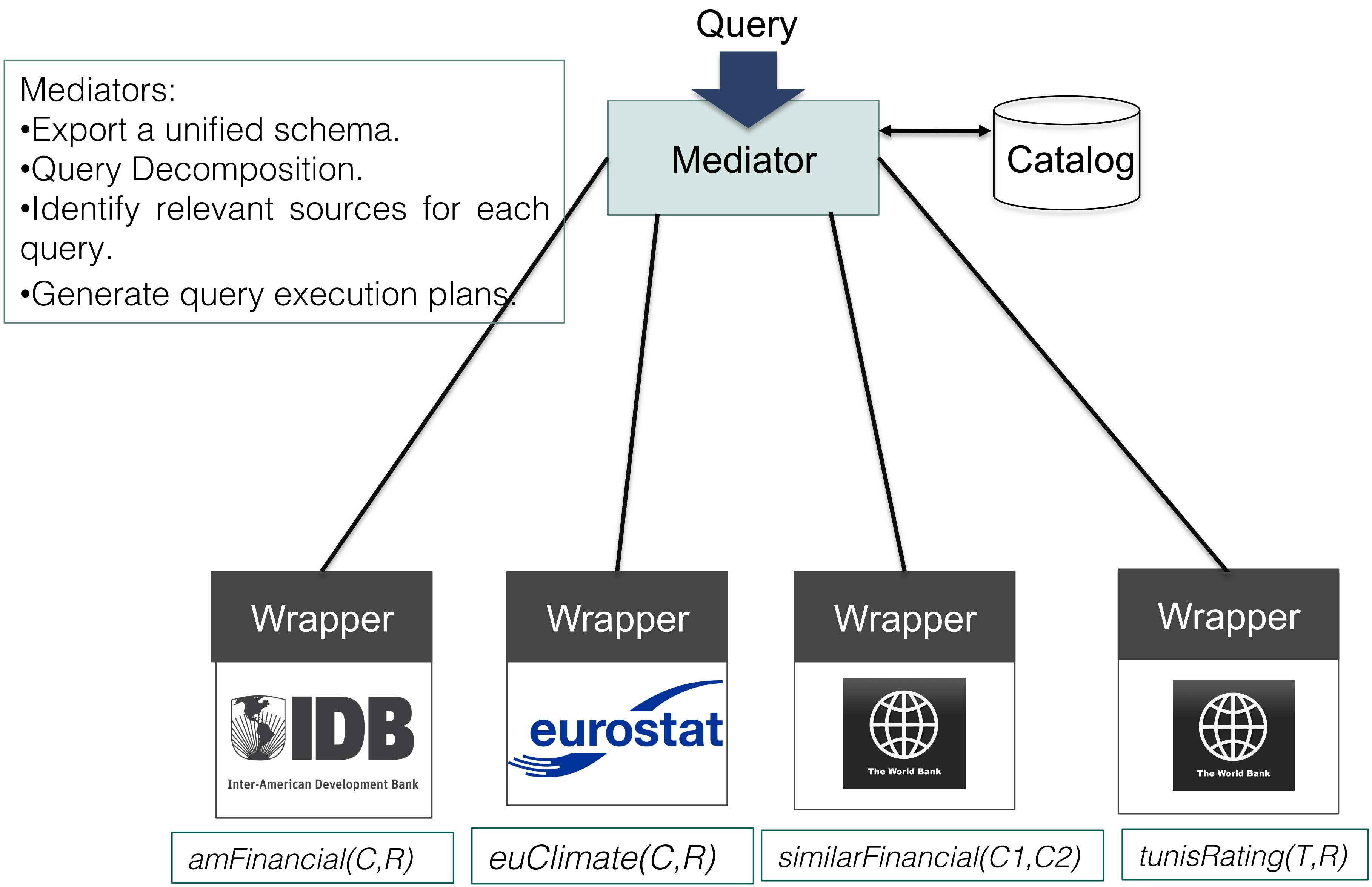


RDF

Transformation Rules, e.g.,
R2RML

Cf. R2RML W3C standard: http://www.w3.org/TR/r2rml/ see also [Priyatna 2014]]
UltraWrap http://capsenta.com/ultrawrap/ [Sequeda & Miranker 2013],
D2RQ http://d2rq.org/

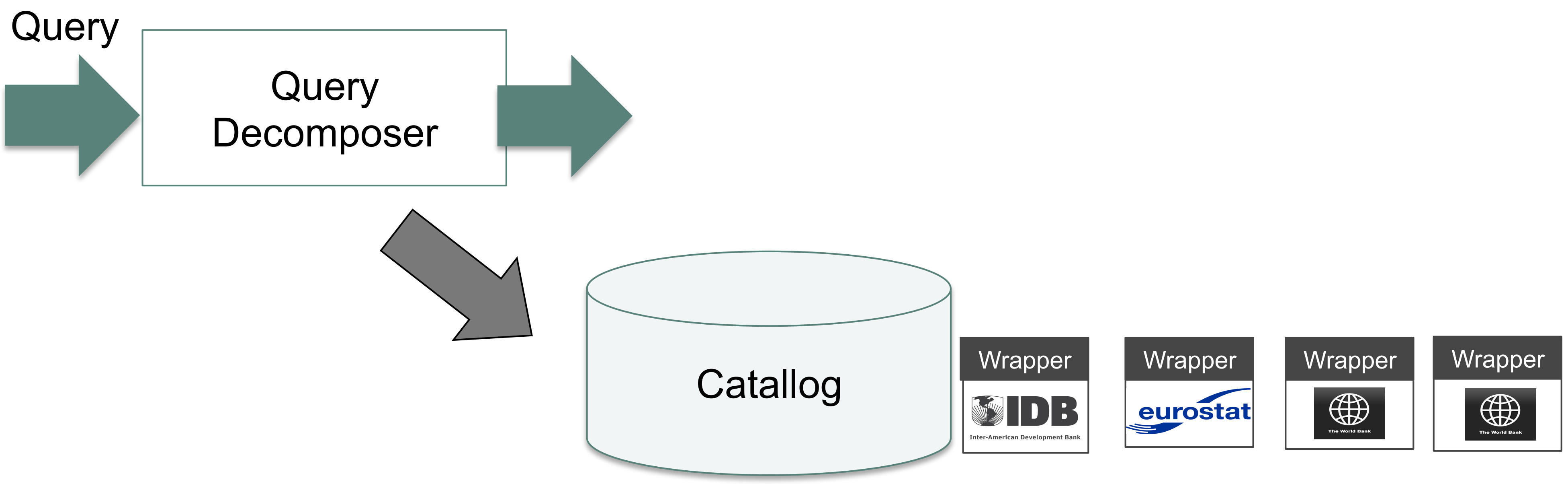
The Mediator and Wrapper Architecture [Wiederhold92]



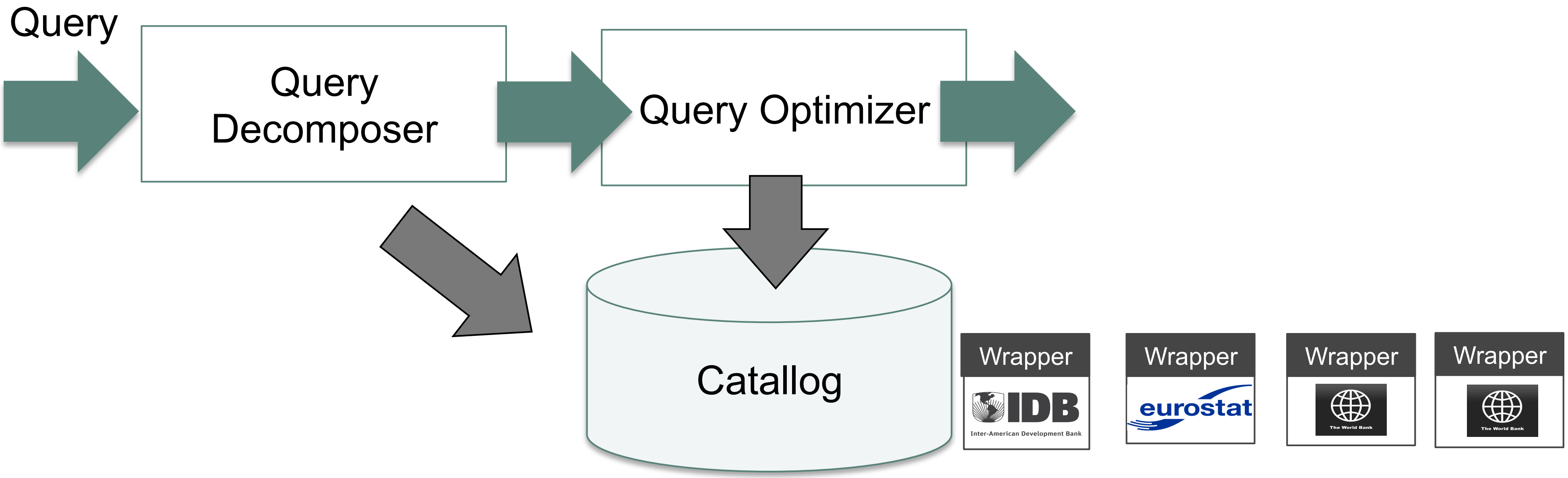
Mediators:
- Export a unified schema.
- Query Decomposition.
- Identify relevant sources for each query.
- Generate query execution plans.

[Wiederhold92]Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)

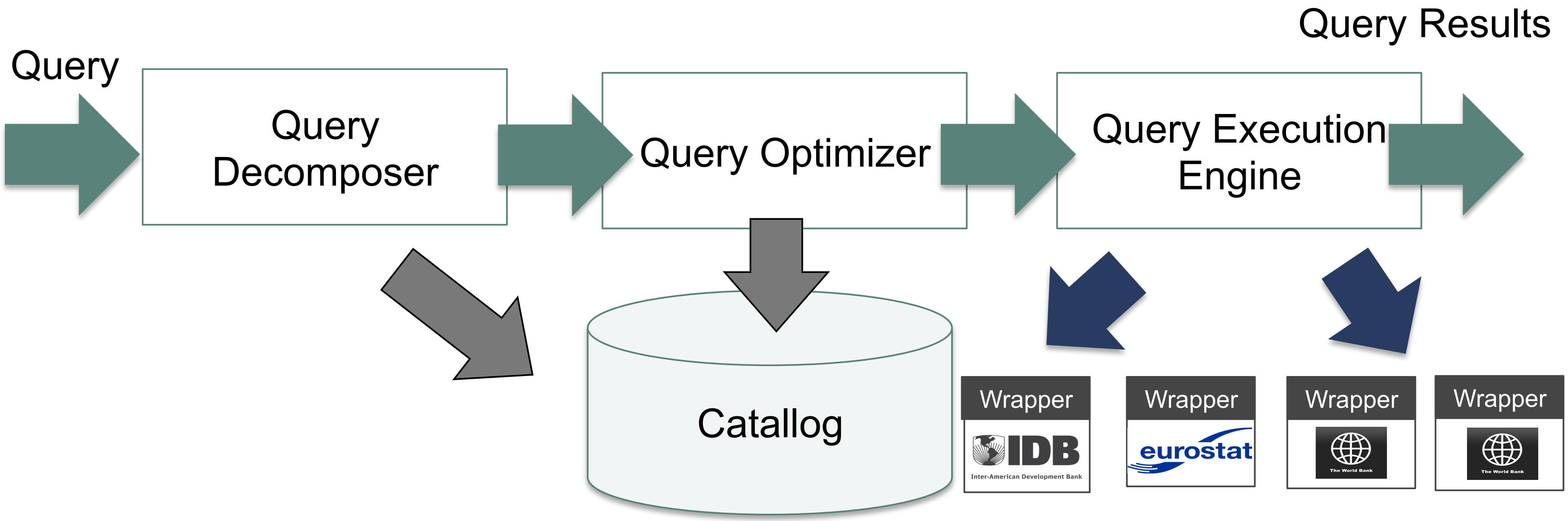
Mediator


Query
Query Decomposer
Catallog
Wrapper
IDB
Inter-American Development Bank
Wrapper
eurostat
Wrapper
The World Bank
Wrapper
The World Bank

Mediator


Query
Query Decomposer
Query Optimizer
Catallog
Wrapper
IDB
Inter-American Development Bank
Wrapper
eurostat
Wrapper
The World Bank
Wrapper
The World Bank

Mediator

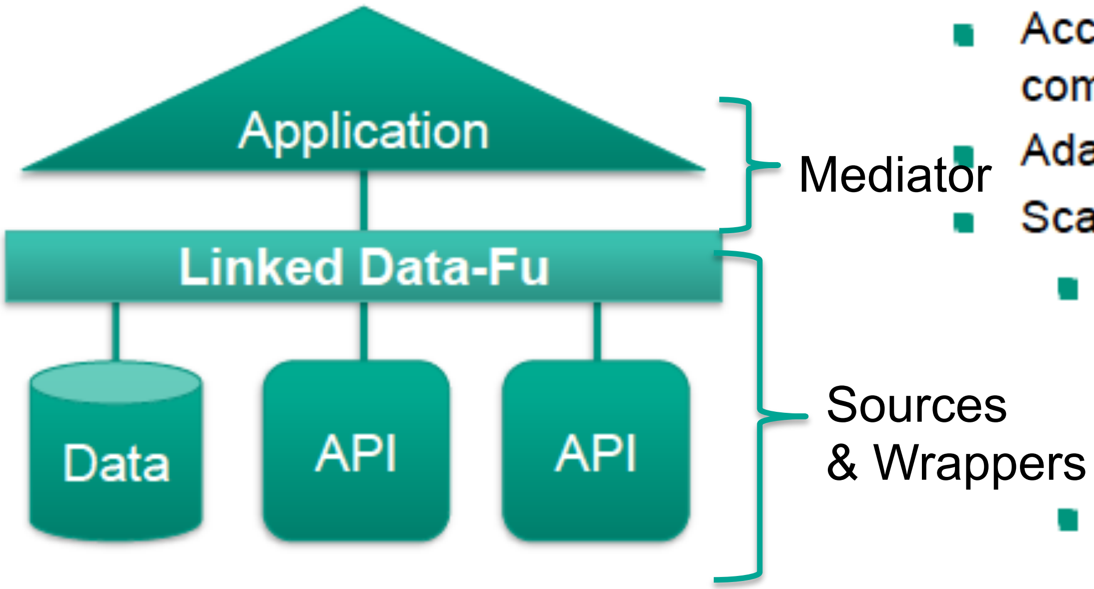

Query Results
Query
Query Decomposer
Query Optimizer
Query Execution Engine
Catallog
Wrapper
IDB
Inter-American Development Bank
Wrapper
eurostat
Wrapper
The World Bank
Wrapper
The World Bank

Linked Data Mediators

[Stadtmüller et al. 2013]

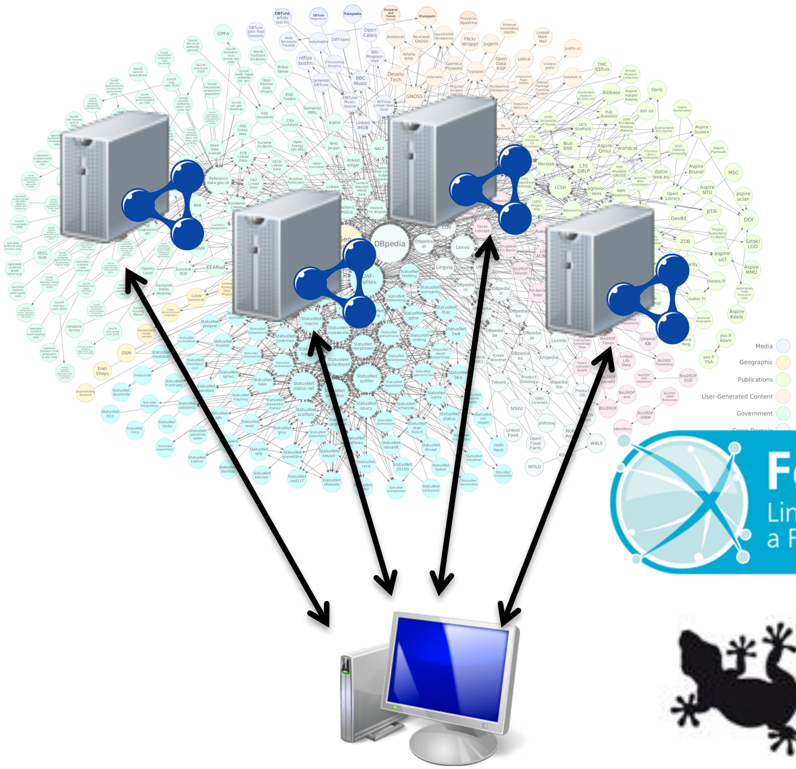
- Linked Data-Fu is a declarative layer
  - Between application and the APIs, components and data sources
  - Data-driven specification of intents via rules
  - Execution heavily utilizes hardware parallelization
  - Combination of RESTful services and Linked Data

- Benefits
  - Bridge heterogeneity
  - Access to distributed components
  - Adaptive behavior
  - Scalability achieved by
    - Rulesets of different expressivity (RDF, RDFS, RDFS Plus, OWL LD…)
    - Parallel execution model



Linked Data Mediators: Federated Query Processing

Publicly available SPARQL endpoints



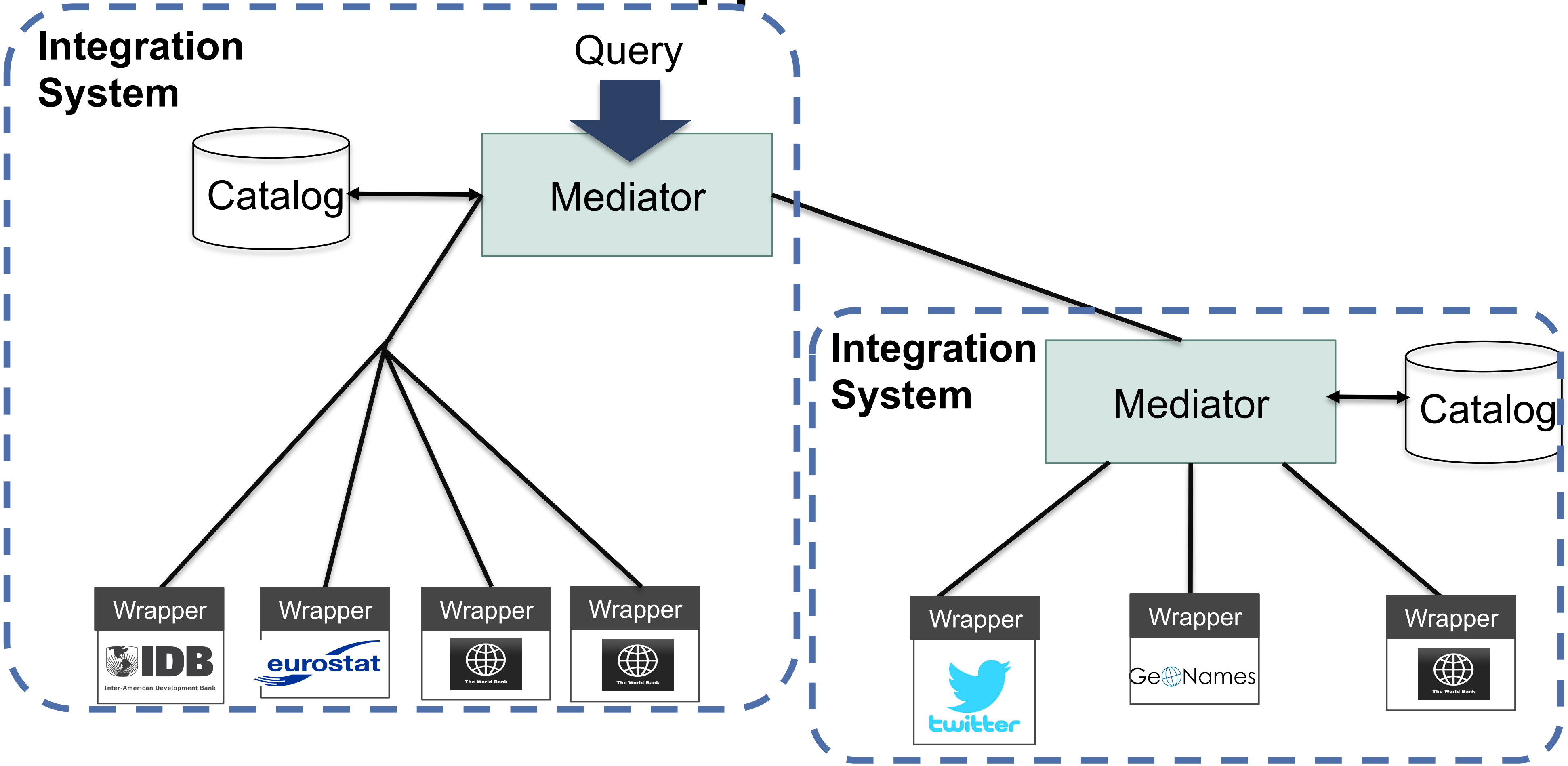
FedX
Linked Data in a Federation

Federated
query processing

ANAPSID

SPLENDID

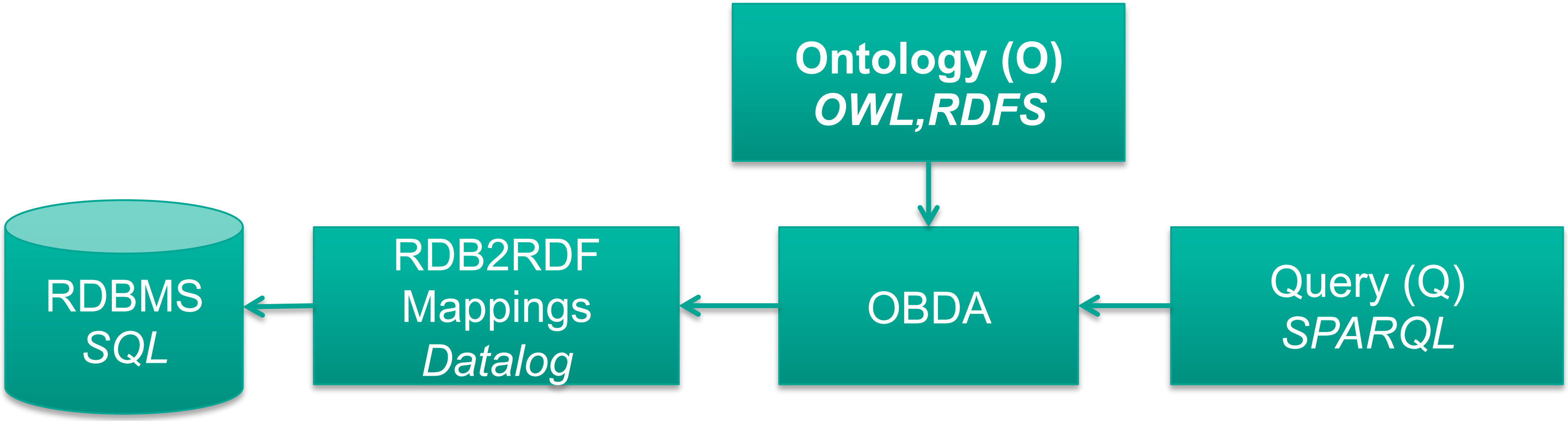
The Mediator and Wrapper Architecture



What is the role of ontologies here?

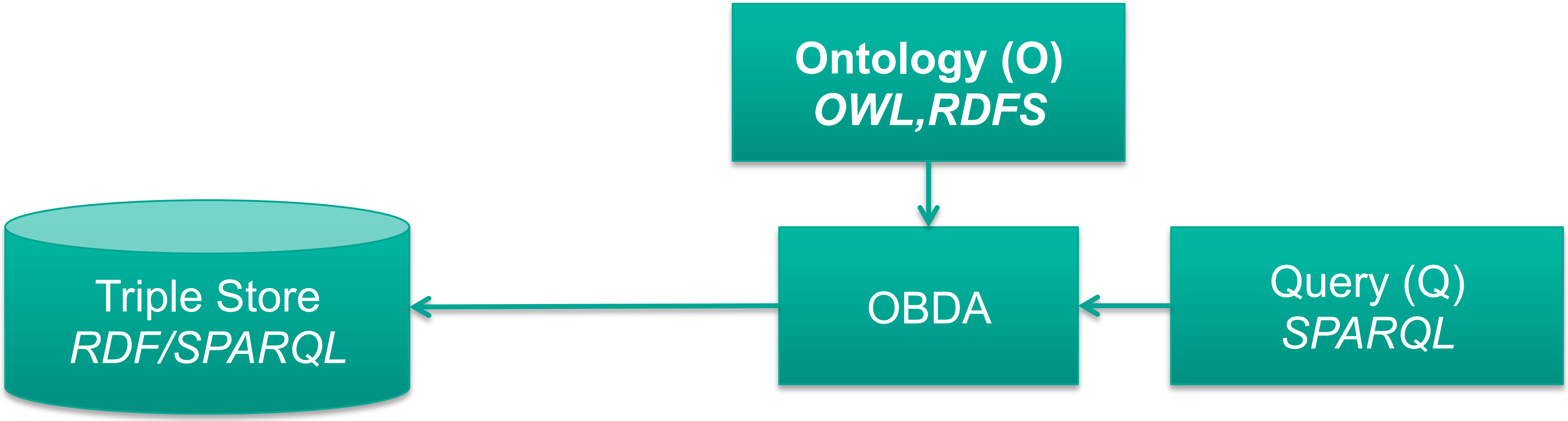
Linked Data integration using ontologies:

- Also popular under the term Ontology-based data-access (**OBDA**) [Kontchakov et al. 2013]:
  - Typically conisders a relational DB, mappings (rules), an ontology Tbox (typically OWL QL (DL-Lite), or OWL RL (rules))



Linked Data integration using ontologies:

- Also popular under the term Ontology-based data-access (**OBDA**) [Kontchakov et al. 2013]:
  - Typically conisders a relational DB, mappings (rules), an ontology Tbox (typically OWL QL (DL-Lite), or OWL RL (rules))

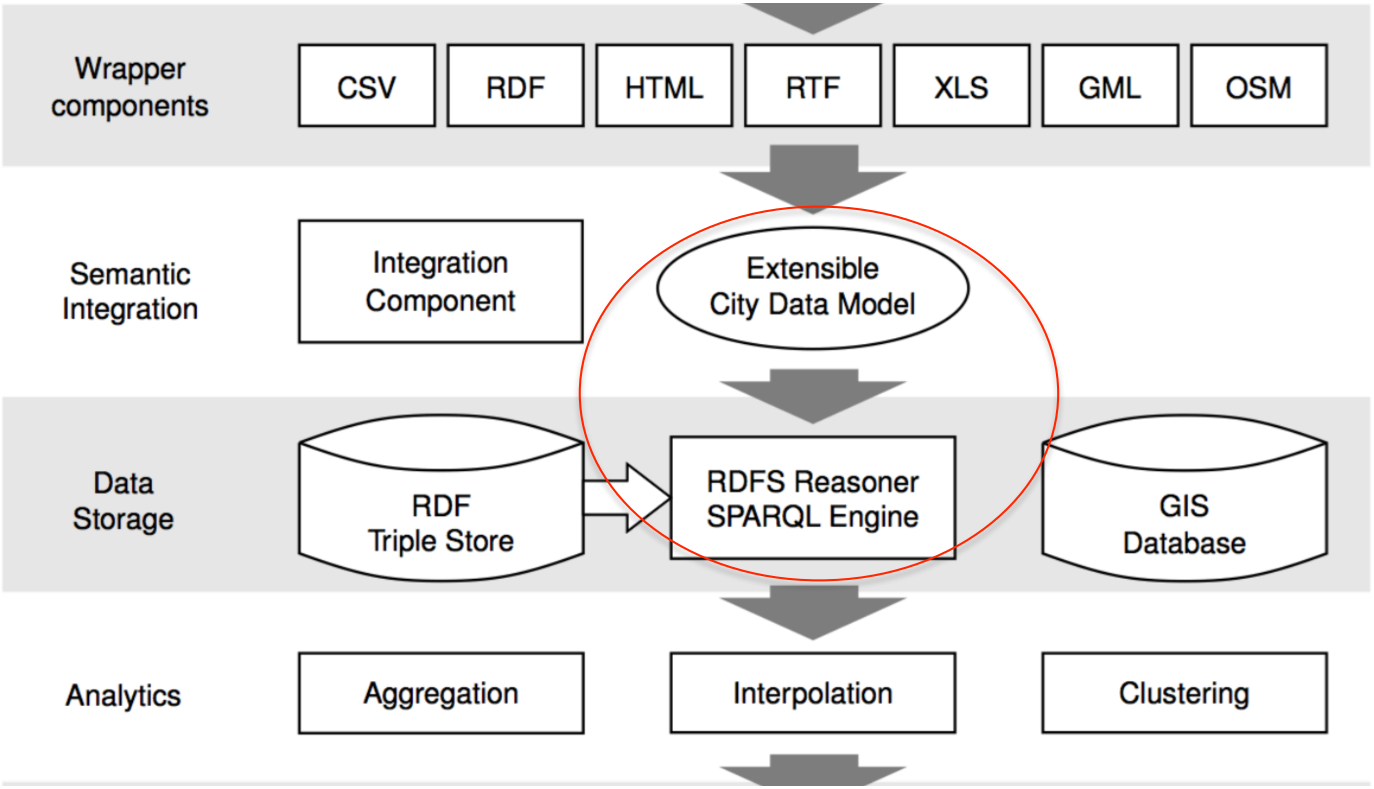


- For simplicity, let's leave out the Relational DB part, assuming Data is already in RDF...

Linked Data integration using ontologies (example)

"Places with a Population Density below 5000/km2"?

A concrete use case:
The "City Data Pipeline"



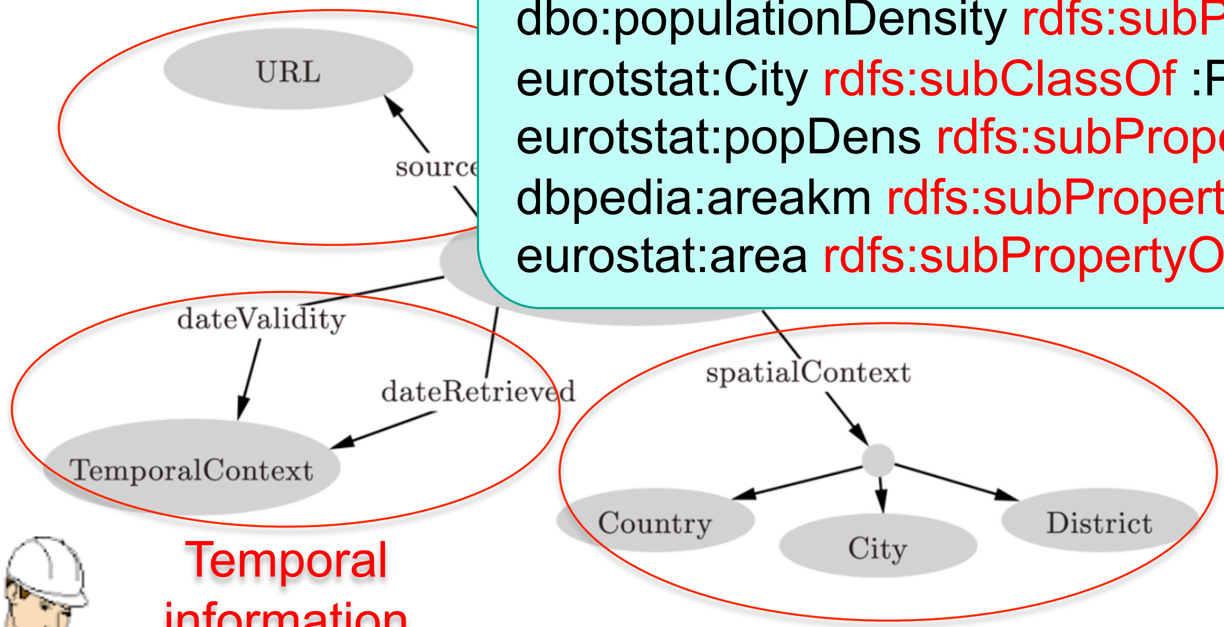
A concrete use case:
The "City Data Pipeline"

City Data Model: extensible
$\mathcal{ALH}(\mathbf{D})$ ontology:

Indicators,
e.g. area in km2,
tons CO2/capita

Provenance

dbo:PopulatedPlace rdfs:subClassOf :Place.
dbo:populationDensity rdfs:subPropertyOf :populationDensity.
eurotstat:City rdfs:subClassOf :Place.
eurotstat:popDens rdfs:subPropertyOf :populationDensity.
dbpedia:areakm rdfs:subPropertyOf :area
eurostat:area rdfs:subPropertyOf :area



Temporal
information

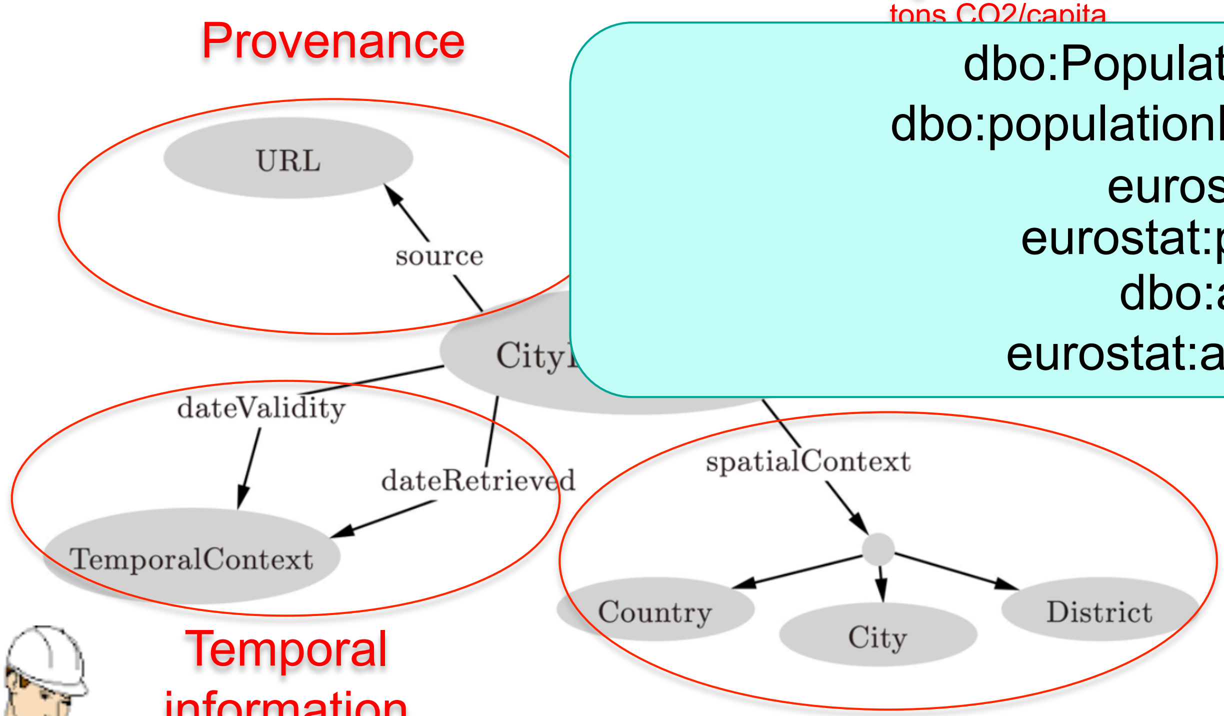
Spatial context

A concrete use case:
The "City Data Pipeline"

City Data Model: extensible
$\mathcal{ALH}(\mathbf{D})$ ontology:

Indicators,
e.g. area in km2,
tons CO2/capita

dbo:PopulatedPlac $\sqsubseteq$:Place
dbo:populationDensity $\sqsubseteq$:populationDensity
eurostat:City $\sqsubseteq$:Place
eurostat:popDen $\sqsubseteq$:populationDensity
dbo:areakm $\sqsubseteq$:area
eurostat:area $\sqsubseteq$:area

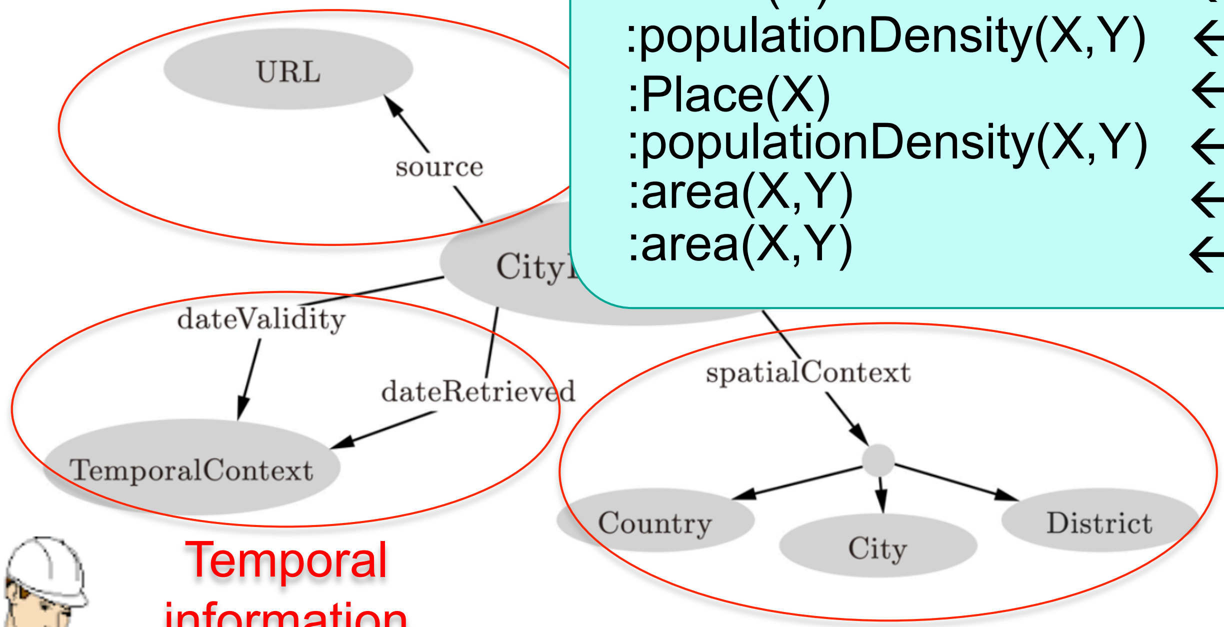


Temporal information

Spatial context

A concrete use case:
The "City Data Pipeline"

City Data Model: extensible
$\mathcal{ALH}$(**D**) ontology:



Provenance

Indicators,
e.g. area in km2,
tons CO2/capita

:Place(X) ← dbo:PopulatedPlace(X)
:populationDensity(X,Y) ← dbo:populationDensity(X,Y)
:Place(X) ← eurostat:City(X)
:populationDensity(X,Y) ← eurostat:popDens(X)
:area(X,Y) ← dbo:areakm(X,Y)
:area(X,Y) ← eurostat:area(X,Y)

Temporal information

Spatial context

A concrete use case: The "City Data Pipeline"

"Places with a Population Density below 5000/km2"?

```
SELECT ?X WHERE { ?X a :Place . ?X :populationDensity ?Y .
                                          FILTER(?Y < 5000) }
```

:Place(X) ← dbo:PopulatedPlace(X)
:populationDensity(X,Y) ← dbo:populationDensity(X,Y)
:Place(X) ← eurostat:City(X)
:populationDensity(X,Y) ← eurostat:popDens(X)
:area(X,Y) ← dbo:areakm(X,Y)
:area(X,Y) ← eurostat:area(X,Y)

Approach 1: Materialization (input: triple store + Ontology output: materialized triple store)

```
SELECT ?X WHERE { ?X a :Place . ?X :populationDensity ?Y .
                                              FILTER(?Y < 5000) }
```

:Vienna a dbo:PopulatedPlace.
:Vienna dbo:populationDensity 4326.1 .
:Vienna dbo:areaKm 414.65 .
:Vienna dbo:populationTotal 1805681 .
:Vienna a :Place.
:Vienna :populationDensity 4326.1 .
:Vienna :area 414.65

:Place(X) ← dbo:PopulatedPlace(X)
:populationDensity(X,Y) ← dbo:populationDensity(X,Y)
:Place(X) ← eurostat:City(X)
:populationDensity(X,Y) ← eurostat:popDens(X)
:area(X,Y) ← dbo:areakm(X,Y)
:area(X,Y) ← eurostat:area(X,Y)

- RDF triple stores implement it naitively (OWLIM, Jena Rules, Sesame)
- Can handle a large part of OWL: OWL 2 RL [Krötzsch, 2012]
- OWL 2 RL covers most RDF/OWL usage on the Web in Linked Data! [Glimm et al. 2012]

Approach 2: Query rewriting (input: conjunctive query (CQ) + Ontology output: UCQ)

```
SELECT ?X WHERE { ?X a :Place . ?X :populationDensity ?Y .
                                          FILTER(?Y < 5000) }
```

:Vienna a dbo:PopulatedPlace.
:Vienna dbo:populationDensity 4326.1 .
:Vienna dbo:areaKm 414.65 .
:Vienna dbo:populationTotal 1805681 .

:Place(X) ← dbo:PopulatedPlace(X)
:populationDensity(X,Y) ← dbo:populationDensity(X,Y)
:Place(X) ← eurostat:City(X)
:populationDensity(X,Y) ← eurostat:popDens(X)
:area(X,Y) ← dbo:areakm(X,Y)
:area(X,Y) ← eurostat:area(X,Y)

```
SELECT ?X WHERE { { {?X a :Place . ?X :populationDensity ?Y . }
        UNION {?X a dbo:Place . ?X :populationDensity ?Y . }
        UNION {?X a :Place . ?X dbo:populationDensity ?Y . }
        UNION {?X a dbo:Place . ?X dbo:populationDensity ?Y . }
        UNION {?X a dbo:Place . ?X dbo:populationDensity ?Y . }
        ... }
                                         FILTER(?Y < 5000) }
```

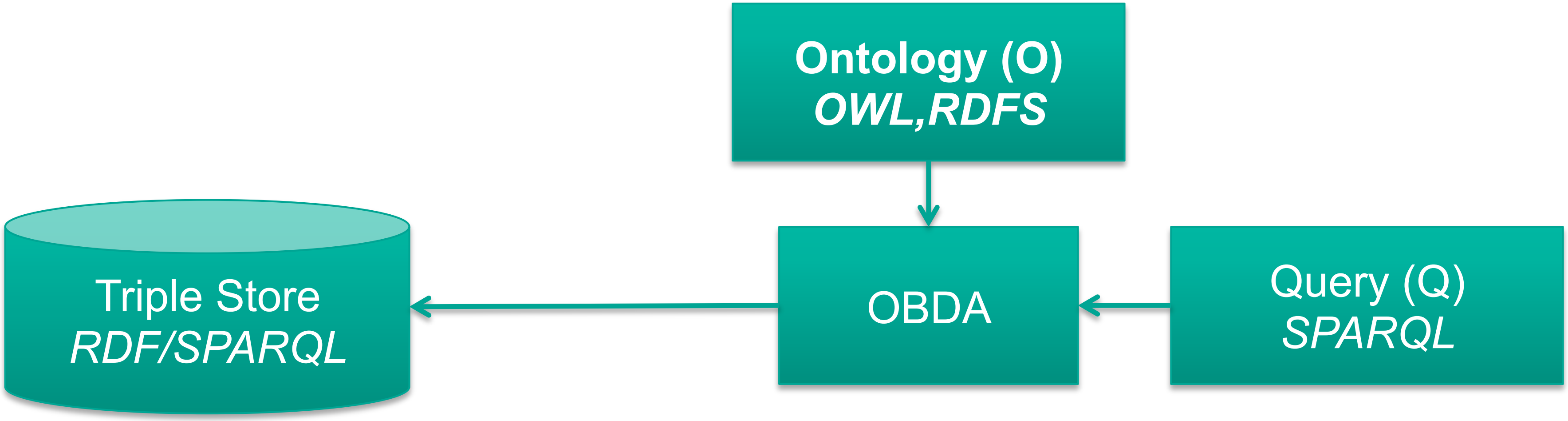
Approach 2: Query rewriting (input: conjunctive query (CQ) + Ontology output: UCQ)

```
SELECT ?X WHERE { ?X a :Place . ?X :populationDensity ?Y .
                                          FILTER(?Y < 5000) }
```

- Observation: essentially, **GAV-style rewriting**
- Can handle a large part of OWL (corresponding to DL-Lite [Calvanese et al. 2007]): OWL 2 QL
- Query-rewriting- based tools and systems available, many optimizations to naive rewritings, e.g. taking into account mappings to a DB:
  - REQUIEM [Perez-Urbina et al., 2009]
  - Quest [Rodriguez-Muro, et al. 2012]
  - ONTOP [Rodriguez-Muro, et al. 2013]
  - Mastro [Calvanese et al. 2011]
  - Presto [Rosati et al. 2010]
  - KYRIE2 [Mora & Corcho, 2014]
- Rewriting vs. Materialization – tradeoff: [Sequeda et al. 2014]

- OBDA is a booming field of research!

Where to find suitable ontologies?



Ontology (O)
OWL,RDFS

Triple Store
RDF/SPARQL

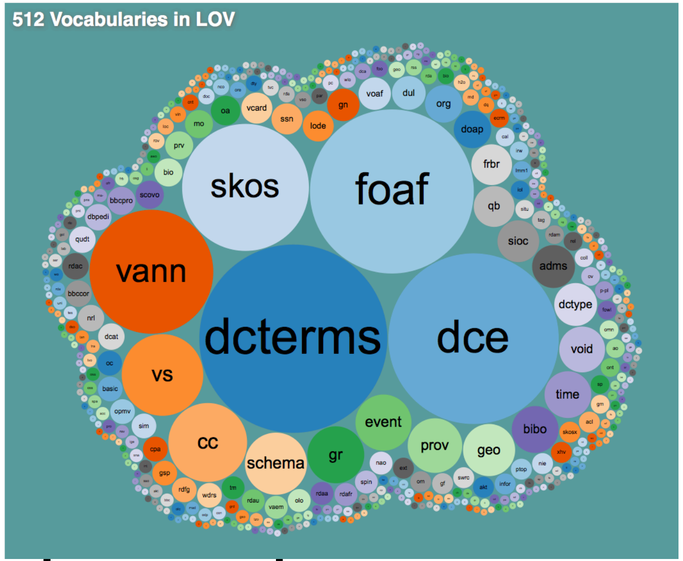
OBDA

Query (Q)
SPARQL

Ontologies and mapping between Linked Data Vocabularies

- Good Starting point: Linked Open Vocabularies

  http://lov.okfn.org/dataset/lov/



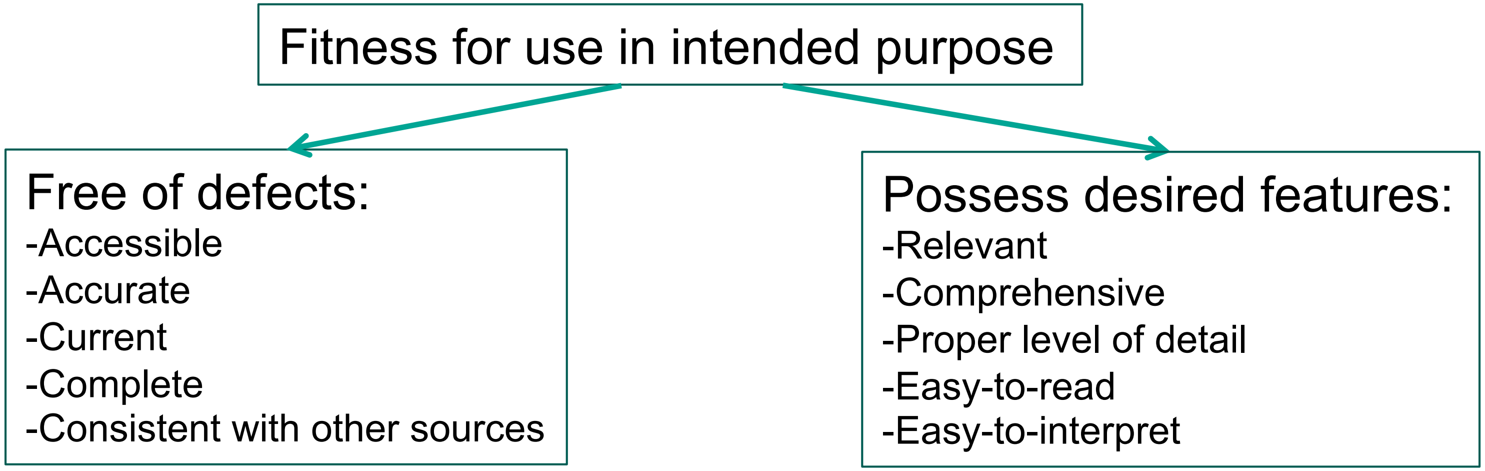
- Still, probably a lot of manual mapping...
  - Literature search for suitable ontologies → don't re-invent the wheel, re-use where possible
  - Crawl
  - Ontology learning, i.e. learn mappings?
    - e.g. using Ontology matching [Shvaiko&Euzenat, 2013]

DATA QUALITY

Data Quality [Lenz 2007]

- Quality reflects the **ability** of an object to meet a **purpose**.
- ISO Norm: **Suitability** for use relative to a given objective of **usage**.
- Industry Quality: is the **conformance** to requirements.

Data is considered high quality if “they are fit for their intended uses in operations, decision making, and planning” (J.M. Juran).

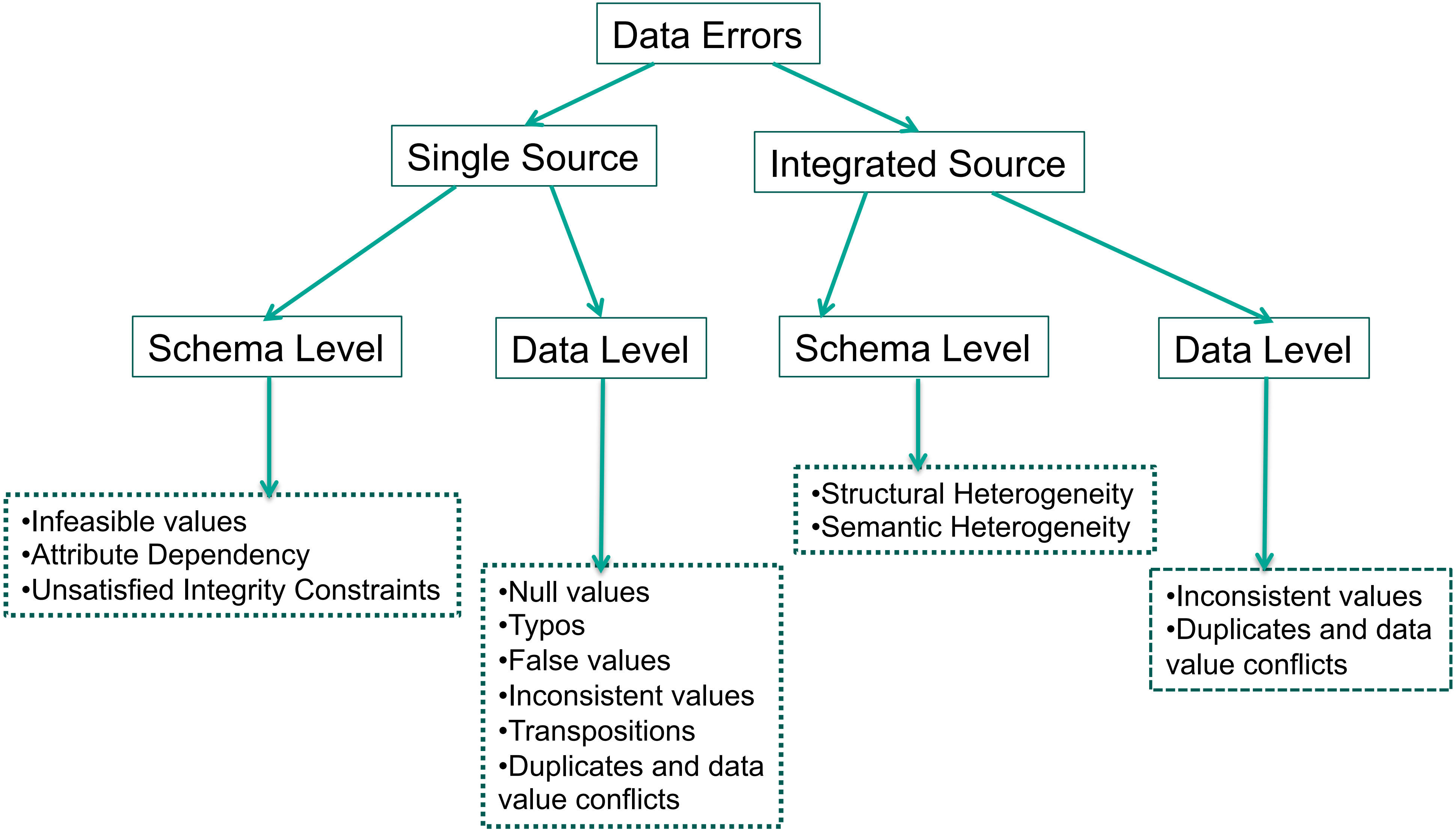


Fitness for use in intended purpose

Free of defects:
-Accessible
-Accurate
-Current
-Complete
-Consistent with other sources

Possess desired features:
-Relevant
-Comprehensive
-Proper level of detail
-Easy-to-read
-Easy-to-interpret

Data Quality Issues [Naumann02]



Data Errors

Single Source

Integrated Source

Schema Level

Data Level

Schema Level

Data Level

- Infeasible values
- Attribute Dependency
- Unsatisfied Integrity Constraints

- Null values
- Typos
- False values
- Inconsistent values
- Transpositions
- Duplicates and data value conflicts

- Structural Heterogeneity
- Semantic Heterogeneity

- Inconsistent values
- Duplicates and data value conflicts

Data Quality-Duplicated Resources

Venezuela

<http://worldbank.270a.info/classification/country/VE>

The World Bank

<http://eurostat.linked-statistics.org/dic/geo#VE>

eurostat

<http://sws.geonames.org/3625428/>

GeoNames

<http://dbpedia.org/resource/Venezuela>

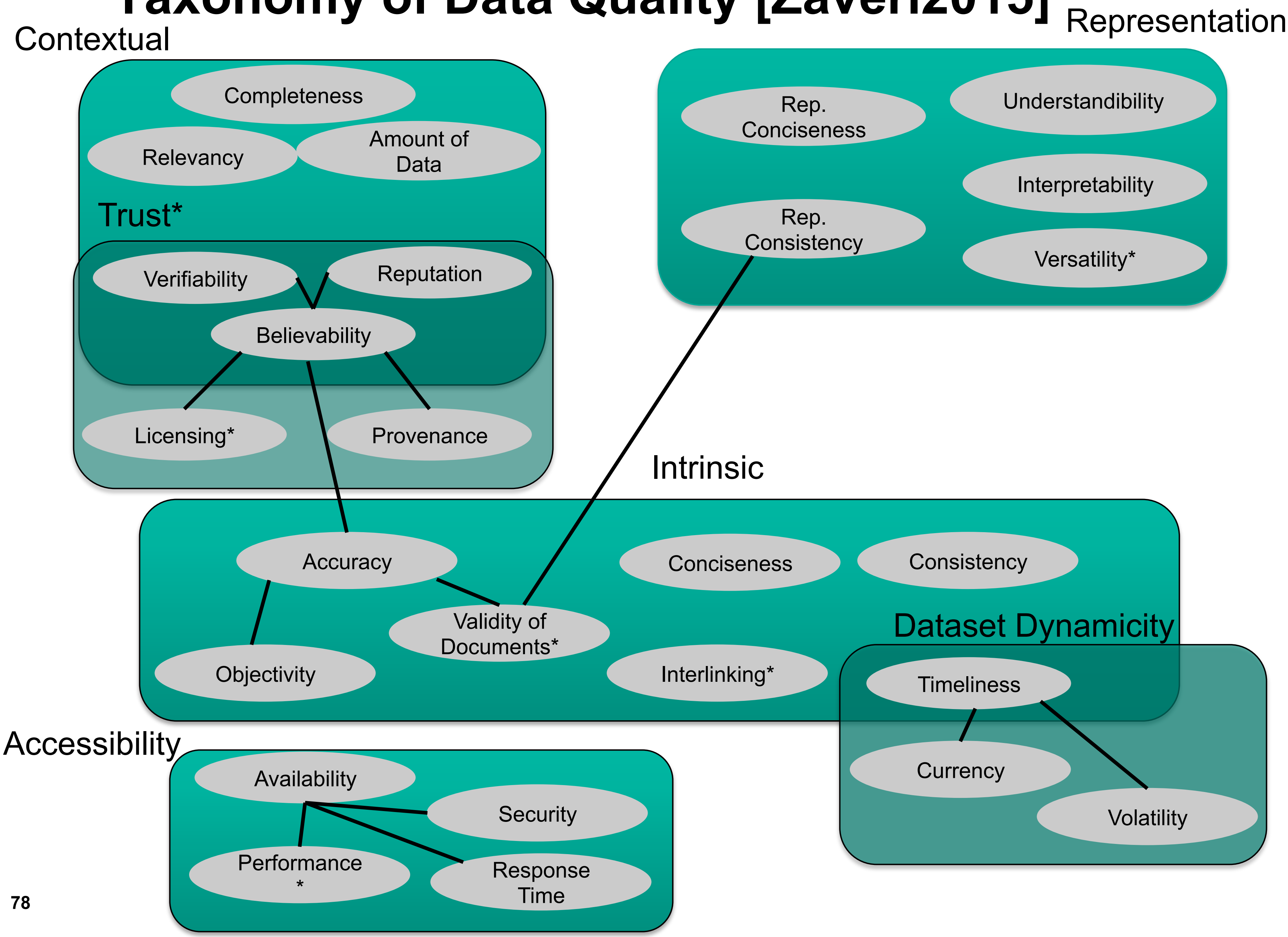
DBpedia

Taxonomy of Data Quality [Naumann02]

Class	Dimension
Intrinsic Data Quality	Believability **Accuracy** Objectivity Reputation
Contextual Data Quality	Value-added Relevancy **Timeliness** **Completeness** Amount of Data
Representation of Data Quality	Interpretability Understandibility Representational **Consistency** Conciseness
Accessibility	Accessibility

[Naumann02] Felix Naumann: Quality-Driven Query Answering for Integrated Information Systems. LNCS 2261, Springer 2002

Taxonomy of Data Quality [Zaveri2015]



Contextual

Completeness

Relevancy

Amount of Data

Trust*

Verifiability

Reputation

Believability

Licensing*

Provenance

Representation

Rep. Conciseness

Understandibility

Interpretability

Rep. Consistency

Versatility*

Intrinsic

Accuracy

Conciseness

Consistency

Validity of Documents*

Objectivity

Interlinking*

Dataset Dynamicity

Timeliness

Currency

Volatility

Accessibility

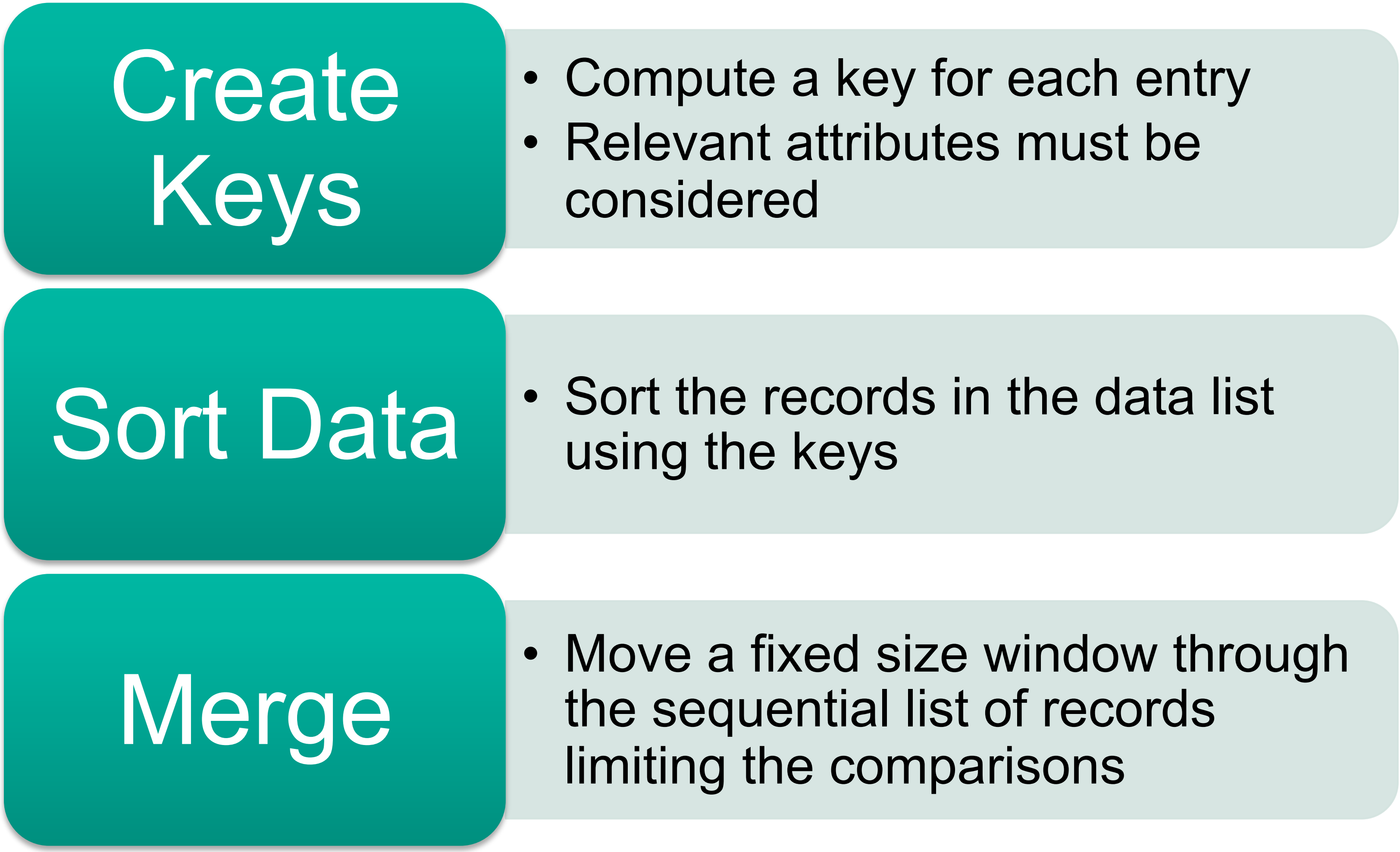
Availability

Security

Performance *

Response Time

Duplicate Detection-Sorted Neighborhood Method [Hernandez&Stolfo, 1998]



Create Keys

- Compute a key for each entry
- Relevant attributes must be considered

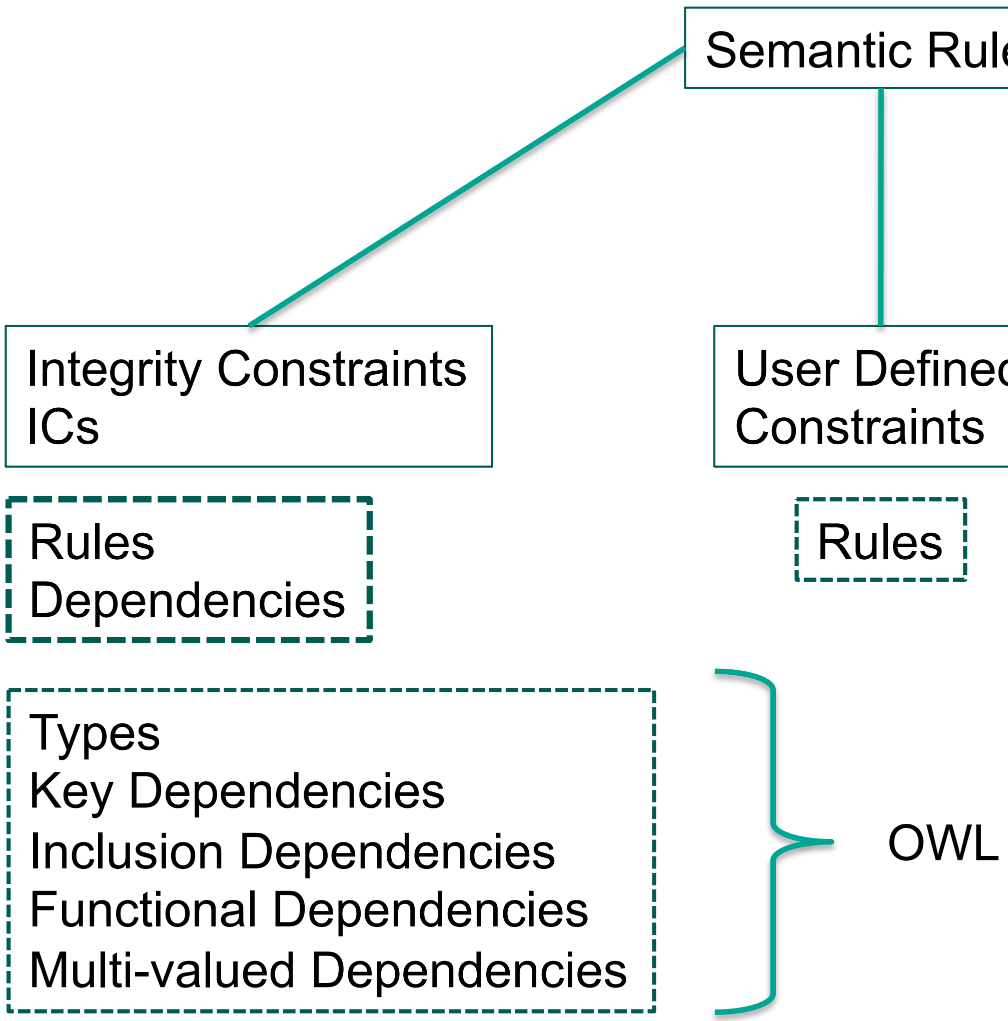
Sort Data

- Sort the records in the data list using the keys

Merge

- Move a fixed size window through the sequential list of records limiting the comparisons

Consistency Detection



Semantic Rules

Integrity Constraints ICs

Rules
Dependencies

Types
Key Dependencies
Inclusion Dependencies
Functional Dependencies
Multi-valued Dependencies

OWL

User Defined Constraints

Rules

Edits

Rules
Equations

Or, alternatively, use statistics?

[Naumann02] Felix Naumann: Quality-Driven Query Answering for Integrated Information Systems. LNCS 2261, Springer 2002

Data Quality (back to our example)

- Duplicates
- Incomplete values (partially solved by inferences/ OBDA)
  - Are OWL+RDFS actually enough?
    - Equations
    - Statistics
- Ambiguous/inconsistent values
  - actually, by inferences and OBDA, even more duplicates values → more possible inconsistencies

Back to the example:
Are RDFS and OWL2 (RL/QL) enough?

```
SELECT ?X WHERE { ?X a :Place . ?X :populationDensity ?Y .
                                     FILTER(?Y < 5000) }
```

:Vienna a dbo:PopulatedPlace.
:Vienna dbo:populationDensity 4326.1 .
:Vienna dbo:areaKm 414.65 .
:Vienna dbo:populationTotal 1805681 .
:Bologna a dbo:PopulatedPlace.
:Bologna dbo:areaKm 140.7 .
:Bologna dbo:populationTotal 386298 .

:Place(X) ← dbo:PopulatedPlace(X)
:populationDensity(X,Y) ← dbo:populationDensity(X,Y)
:Place(X) ← eurostat:City(X)
:populationDensity(X,Y) ← eurostat:popDens(X)
:area(X,Y) ← dbo:areakm(X,Y)
:area(X,Y) ← eurostat:area(X,Y)

? :populationDensity = :population/:area
:area = 0,386102 * dbpedia:areaMi2

Probably not...

A possible solution: [Bischof & Polleres, 2013]

- [Bischof&Polleres 2013] Basic Idea: Consider clausal form of all variants of equations and use Query rewriting with "blocking":

$$(S, \text{popDensity}, PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A),\ PD := P/A$$

$$(S, \text{area}, PD) \leftarrow (S, \text{population}, P), (S, \text{popDensity}, PD),\ A := P/PD$$

$$(S, \text{population}, P) \leftarrow (S, \text{area}, A), (S, \text{popDensity}, PD),\ P := A * PD$$

:Bologna dbo:population 386298 .
:Bologna dbo:areaKm 140.7 .

Finally, the resulting UCQs with assignments can be rewritten back to SPARQL using BIND

```
SELECT ?PD WHERE { :Bologna dbo:popDensity ?PD}
```

$$q(PD) \leftarrow \boxed{(S, \text{popDensity}, PD)}$$

$$q(PD) \leftarrow \boxed{(S, \text{population}, P)}, (S, \text{area}, A), PD := P/A$$

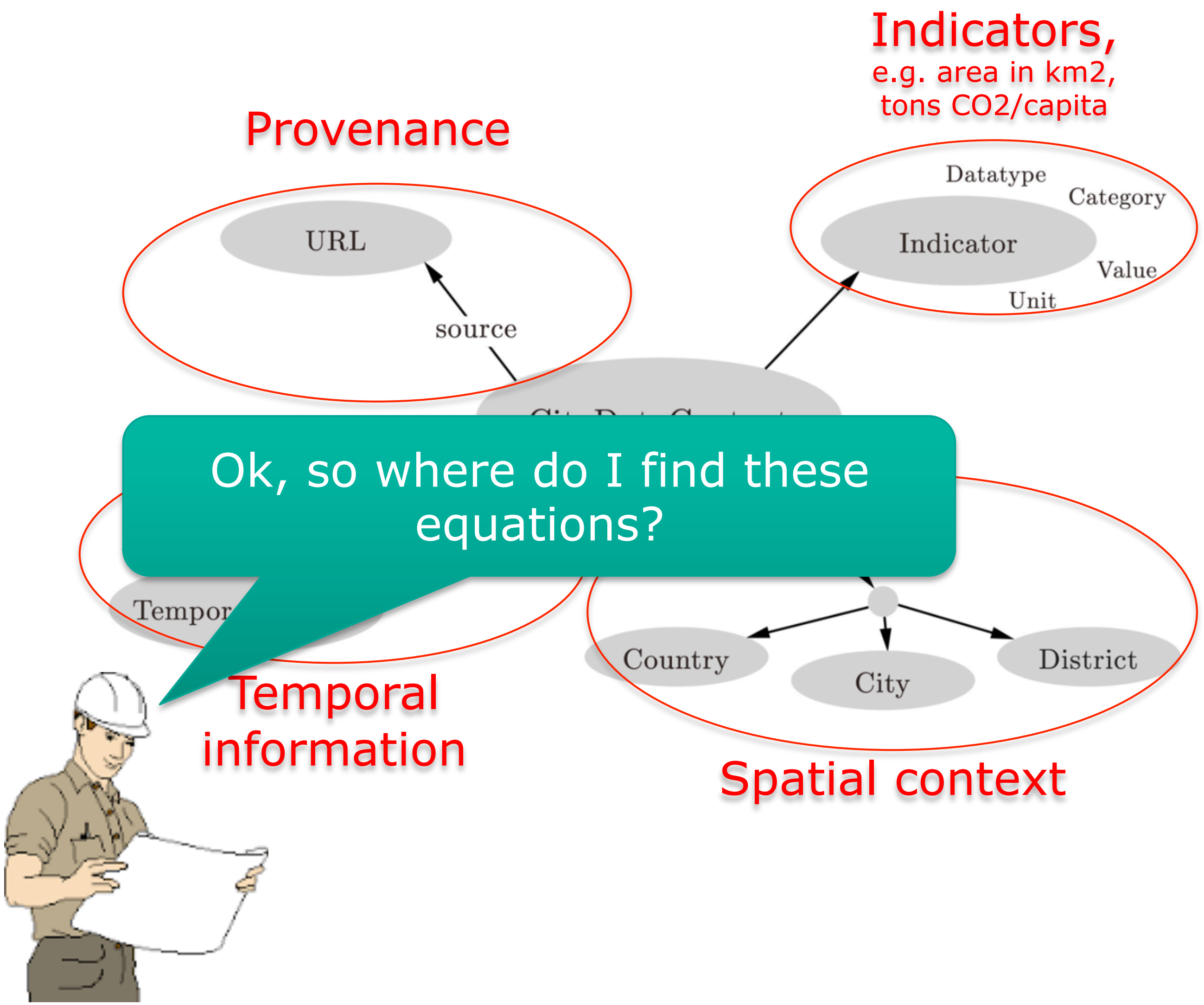
~~$q(PD) \leftarrow \boxed{(S, \text{popDensity}, PD')}, (S, \text{area}, A'), (S, \text{area}, A), PD := P/A, P := PD' * A'$~~

.. infinite expansion even if only 1 equation is considered.

Solution: “blocking” recursive expansion of the same equation for the same value.

```
SELECT ?PD WHERE { {:Athens dbo:popDensity ?PD }
                   UNION
                   { :Athens dbo:population ?P ; dbo:area ?A .
                     BIND (?P/?A AS ?PD )}
                 }
```

A concrete use case: The "City Data Pipeline"



Indicators, e.g. area in km2, tons CO2/capita

Provenance

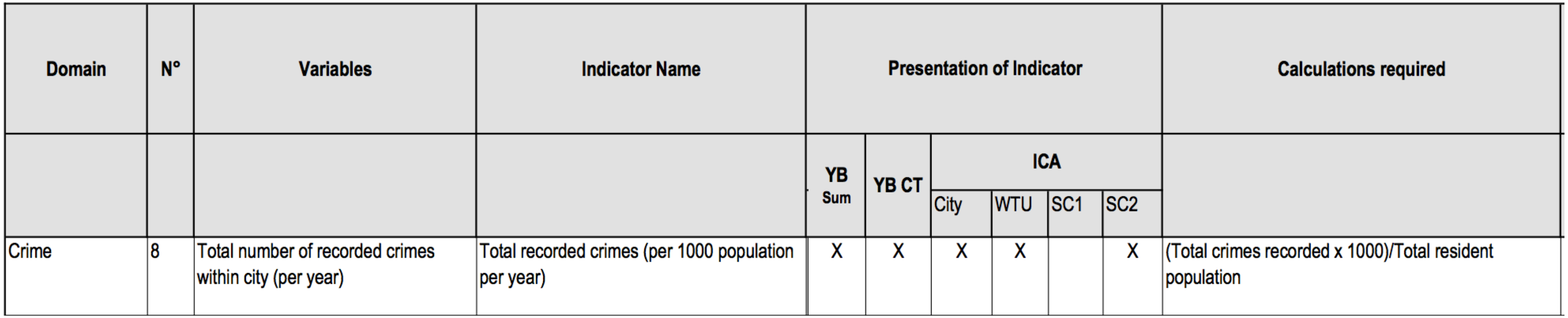
Ok, so where do I find these equations?

Temporal information

Spatial context

Equational knowledge:

- Eurostat/Urbanaudit:
  - http://ec.europa.eu/regional_policy/archive/urban2/urban/audit/ftp/vol3.pdf

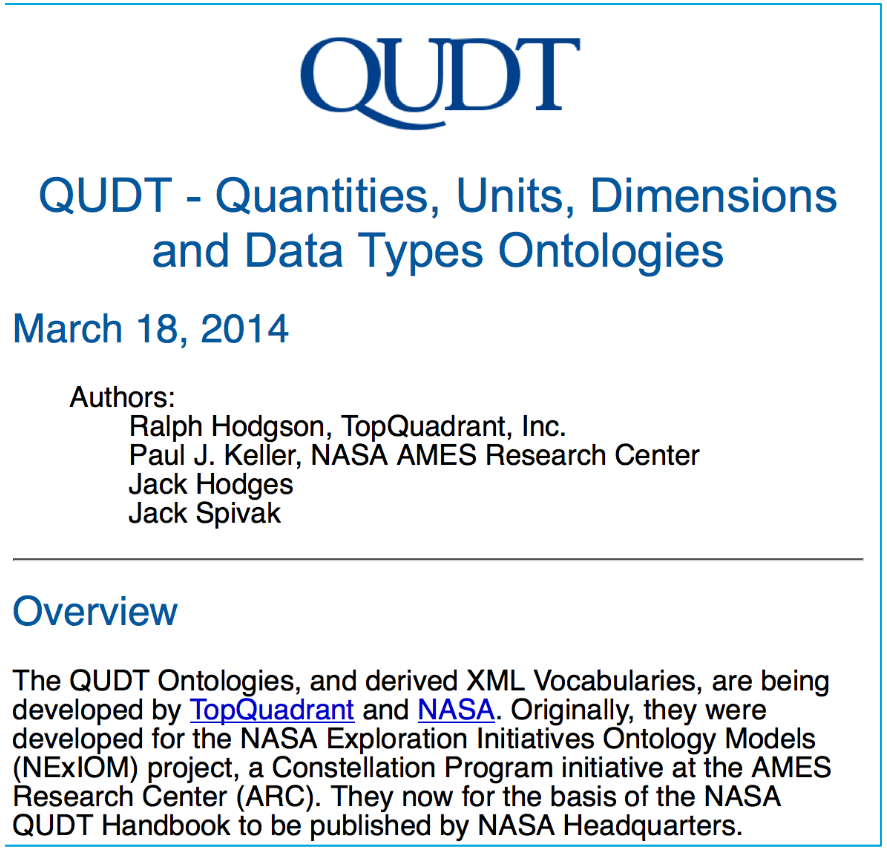


Domain	N°	Variables	Indicator Name	Presentation of Indicator						Calculations required
				YB Sum	YB CT	ICA				
						City	WTU	SC1	SC2	
Crime	8	Total number of recorded crimes within city (per year)	Total recorded crimes (per 1000 population per year)	X	X	X	X		X	(Total crimes recorded x 1000)/Total resident population

Equational knowledge:
Unit conversion

http://qudt.org/

http://www.wurvoc.org/vocabularies/om-1.8/



QUDT

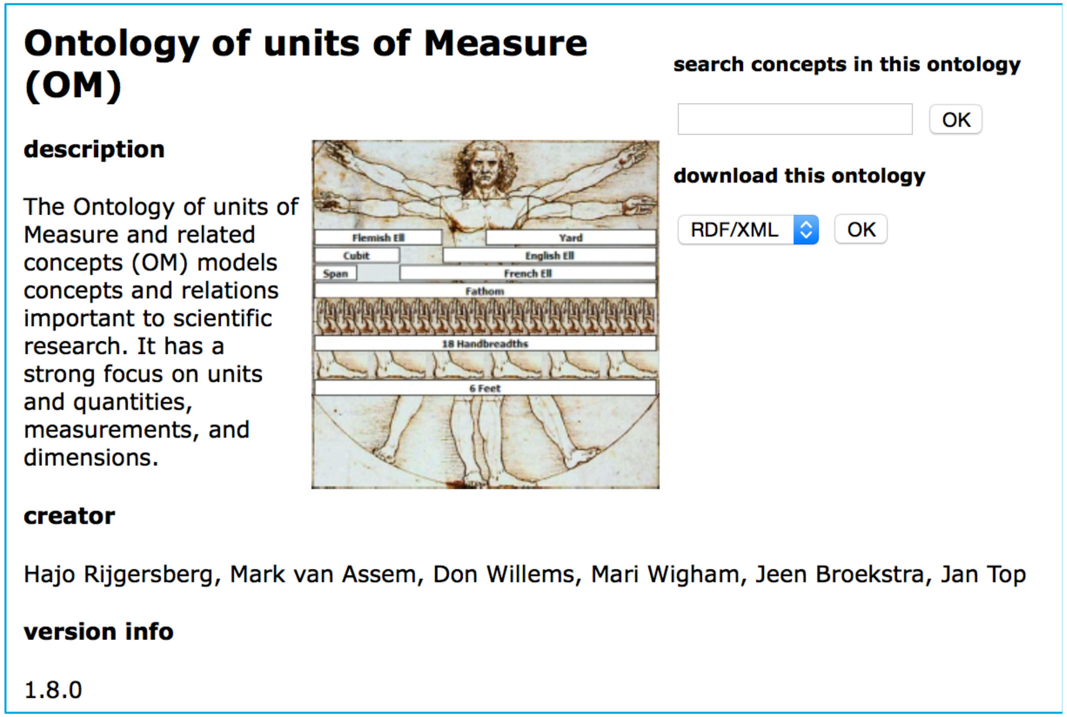
QUDT - Quantities, Units, Dimensions and Data Types Ontologies

March 18, 2014

Authors:
Ralph Hodgson, TopQuadrant, Inc.
Paul J. Keller, NASA AMES Research Center
Jack Hodges
Jack Spivak

Overview

The QUDT Ontologies, and derived XML Vocabularies, are being developed by TopQuadrant and NASA. Originally, they were developed for the NASA Exploration Initiatives Ontology Models (NExIOM) project, a Constellation Program initiative at the AMES Research Center (ARC). They now for the basis of the NASA QUDT Handbook to be published by NASA Headquarters.



Ontology of units of Measure (OM)

description

The Ontology of units of Measure and related concepts (OM) models concepts and relations important to scientific research. It has a strong focus on units and quantities, measurements, and dimensions.

creator

Hajo Rijgersberg, Mark van Assem, Don Willems, Mari Wigham, Jeen Broekstra, Jan Top

version info

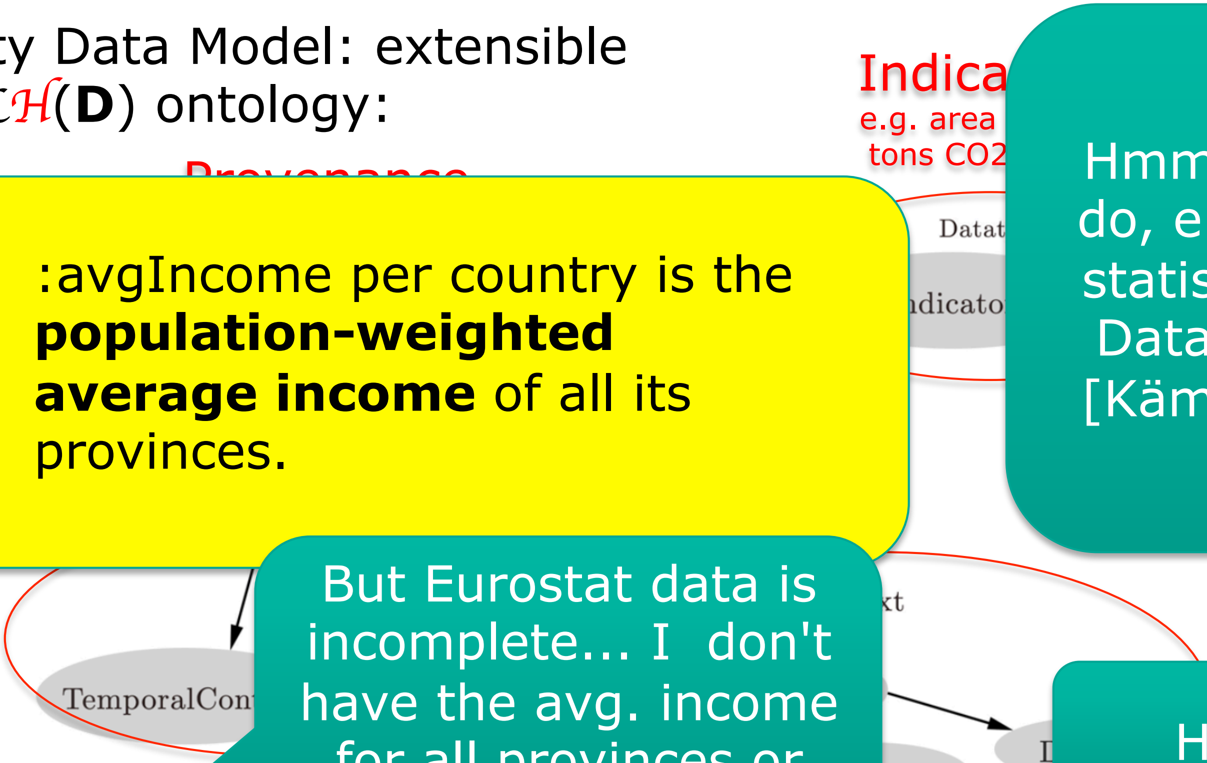
1.8.0

Data Quality Data Quality (back to our example)

- Duplicates
- Incomplete values (partially solved by inferences/ OBDA)
    - Are OWL+RDFS actually enough?
        - Equations
        - Statistics
- Ambiguous/inconsistent values
    - actually, by inferences and OBDA, even more duplicates values → more possible inconsistencies

A concrete use case:
The "City Data Pipeline"

City Data Model: extensible $\mathcal{ALH}(\mathbf{D})$ ontology:



:avgIncome per country is the **population-weighted average income** of all its provinces.

But Eurostat data is incomplete... I don't have the avg. income for all provinces or countries in the EU!

Hmmm... Still a lot of work to do, e.g. adding aggregates for statistical data (Eurostat, RDF Data Cube Vocabulary) ... cf. [Kämpgen, 2014, PhD Thesis]

Hmmm...

Challenge – Missing values [Bischof et al. 2015]

- WARNING: In Open Data we find huge amounts of missing values
- Two Reasons:
  - Incomplete data published by providers (Tables 1+2)
  - The combination of different data sets with disjoint cities and indicators (later)

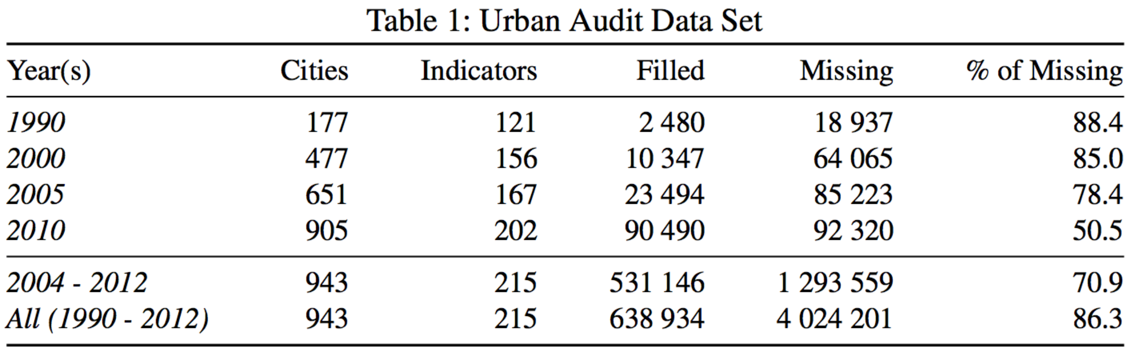


Table 1: Urban Audit Data Set

Year(s)	Cities	Indicators	Filled	Missing	% of Missing
1990	177	121	2 480	18 937	88.4
2000	477	156	10 347	64 065	85.0
2005	651	167	23 494	85 223	78.4
2010	905	202	90 490	92 320	50.5
2004 - 2012	943	215	531 146	1 293 559	70.9
All (1990 - 2012)	943	215	638 934	4 024 201	86.3

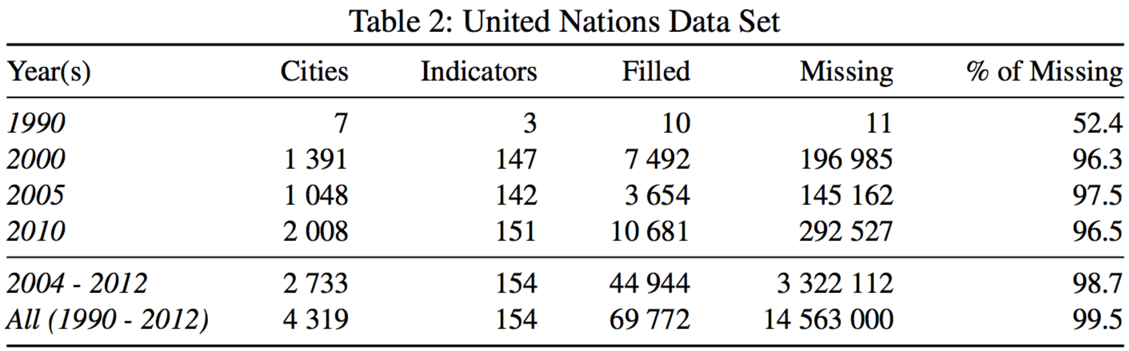
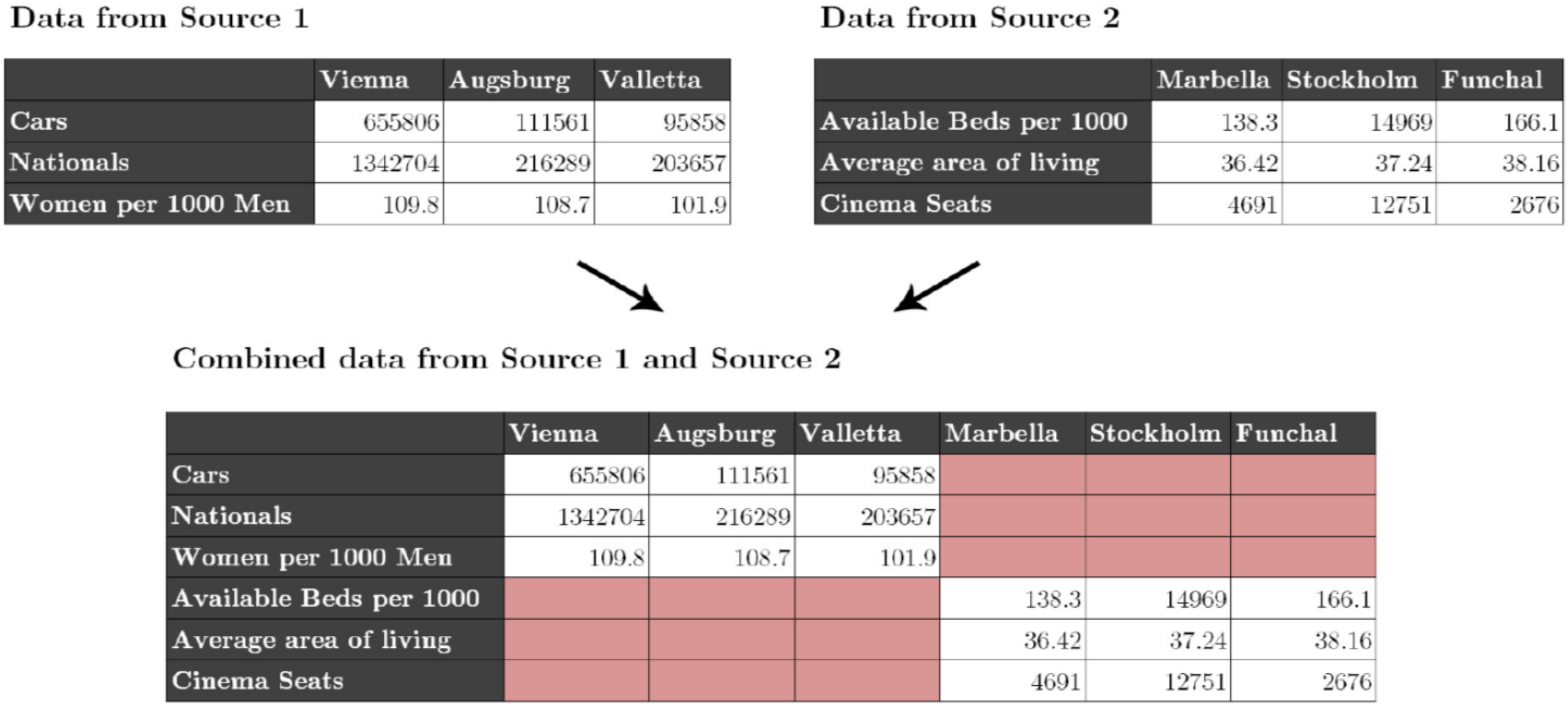


Table 2: United Nations Data Set

Year(s)	Cities	Indicators	Filled	Missing	% of Missing
1990	7	3	10	11	52.4
2000	1 391	147	7 492	196 985	96.3
2005	1 048	142	3 654	145 162	97.5
2010	2 008	151	10 681	292 527	96.5
2004 - 2012	2 733	154	44 944	3 322 112	98.7
All (1990 - 2012)	4 319	154	69 772	14 563 000	99.5

Challenges – Missing values

- Individual datasets (e.g. from Eurostat) have missing values
- **Merging together datasets** with different indicators/cities adds sparsity



Data from Source 1

	Vienna	Augsburg	Valletta
Cars	655806	111561	95858
Nationals	1342704	216289	203657
Women per 1000 Men	109.8	108.7	101.9

Data from Source 2

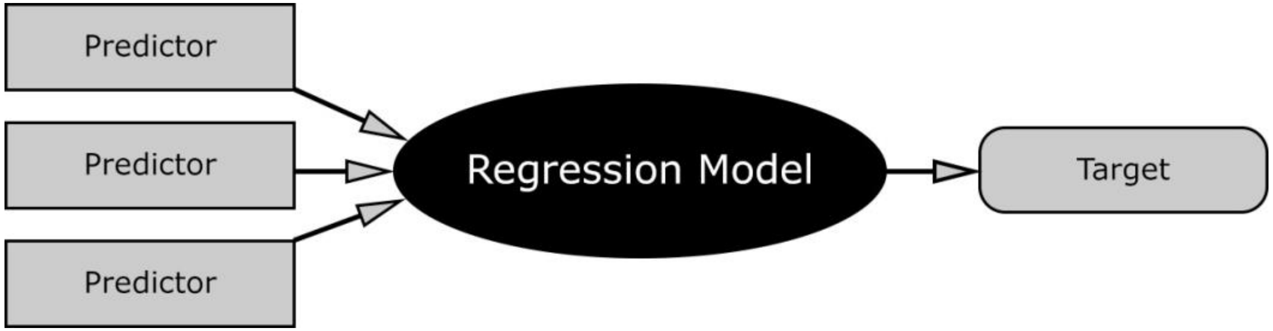
	Marbella	Stockholm	Funchal
Available Beds per 1000	138.3	14969	166.1
Average area of living	36.42	37.24	38.16
Cinema Seats	4691	12751	2676

Combined data from Source 1 and Source 2

	Vienna	Augsburg	Valletta	Marbella	Stockholm	Funchal
Cars	655806	111561	95858	🟥	🟥	🟥
Nationals	1342704	216289	203657	🟥	🟥	🟥
Women per 1000 Men	109.8	108.7	101.9	🟥	🟥	🟥
Available Beds per 1000	🟥	🟥	🟥	138.3	14969	166.1
Average area of living	🟥	🟥	🟥	36.42	37.24	38.16
Cinema Seats	🟥	🟥	🟥	4691	12751	2676

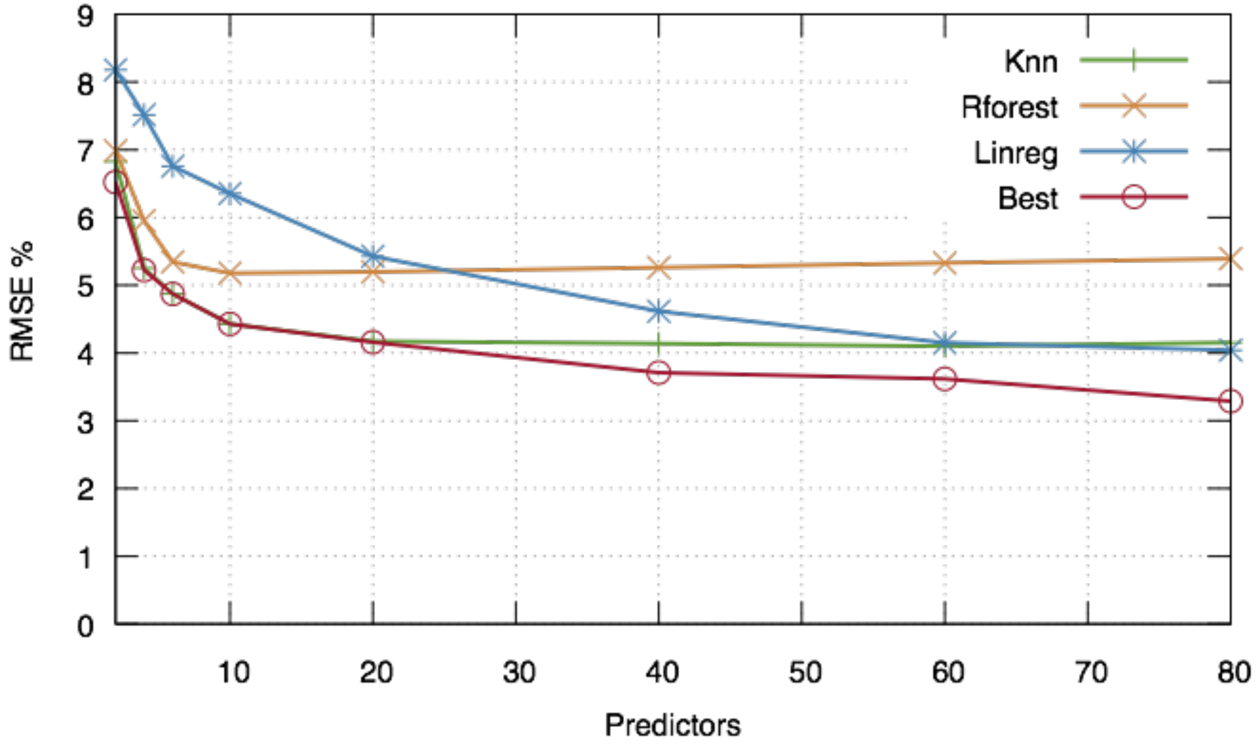
Missing Values – Hybrid approach choose best prediction method per indicator:

- Our assumption: every indicator has its own distribution and relationship to others.
- Basket of „standard" regression methods:
  - K-Nearest Neighbour Regression (KNN)
  - Multiple Linear Regression (MLR)
  - Random Forest Decision Trees (RFD)



Missing Values – Hybrid approach choose best prediction method per indicator:

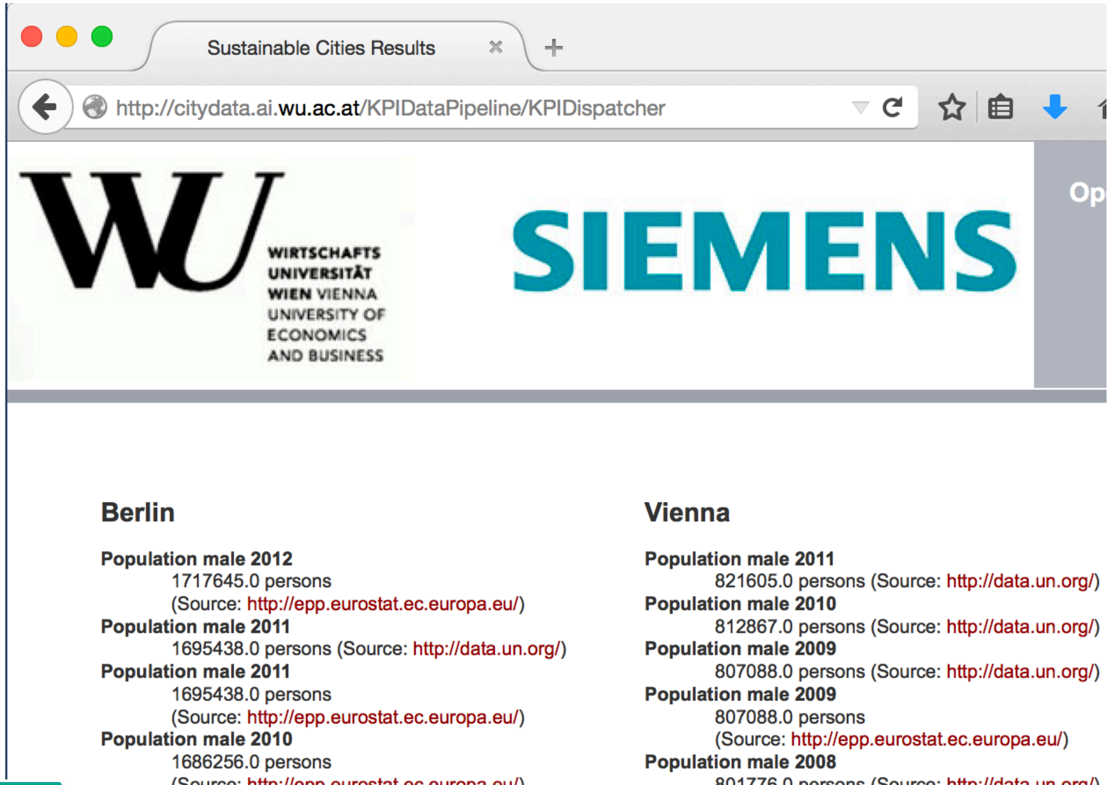
- Instead of using indicators directly we use Principle Components, built from the indicators
- For buidling the PCs, fill in missing data points with neutral values → predict all rows



City Data Pipeline

citydata.wu.ac.at

- Search for indicators & cities
- obtain results incl. sources
- Integrated data served as Linked Data
- Predicted values AND estimated error (RMSE) for missing data...

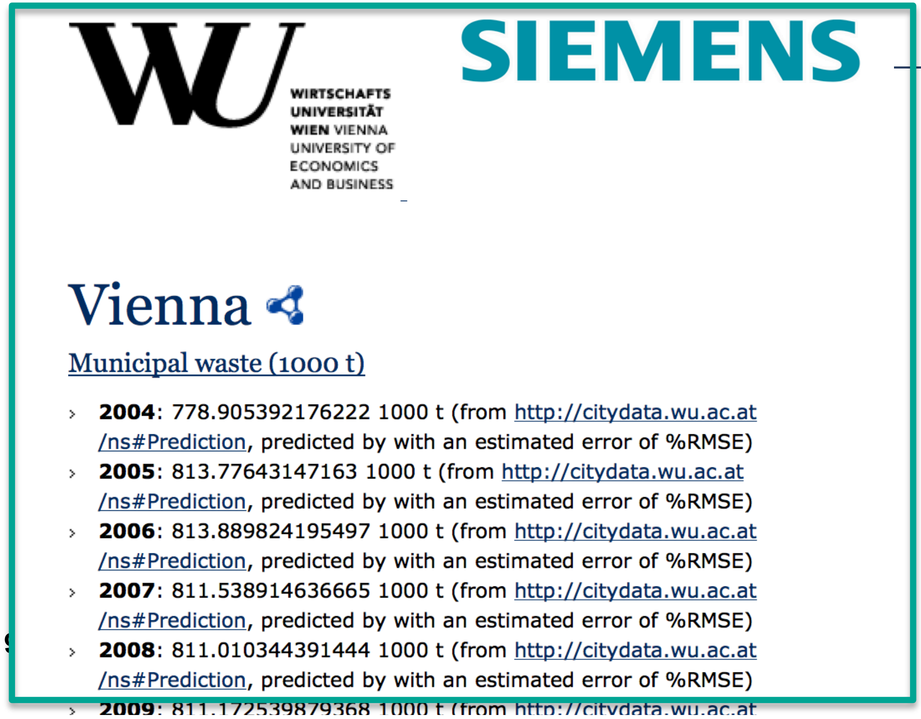

Sustainable Cities Results
http://citydata.ai.wu.ac.at/KPIDataPipeline/KPIDispatcher
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS
Op

Berlin

Population male 2012
1717645.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2011
1695438.0 persons (Source: http://data.un.org/)
Population male 2011
1695438.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2010
1686256.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2009
1686256.0 persons

Vienna

Population male 2011
821605.0 persons (Source: http://data.un.org/)
Population male 2010
812867.0 persons (Source: http://data.un.org/)
Population male 2009
807088.0 persons (Source: http://data.un.org/)
Population male 2009
807088.0 persons
(Source: http://epp.eurostat.ec.europa.eu/)
Population male 2008
801776.0 persons (Source: http://data.un.org/)
Population male 2008
800361.0 persons


WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

Vienna

Municipal waste (1000 t)

- **2004**: 778.905392176222 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by with an estimated error of %RMSE)
- **2005**: 813.77643147163 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by with an estimated error of %RMSE)
- **2006**: 813.889824195497 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by with an estimated error of %RMSE)
- **2007**: 811.538914636665 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by with an estimated error of %RMSE)
- **2008**: 811.010344391444 1000 t (from http://citydata.wu.ac.at/ns#Prediction, predicted by with an estimated error of %RMSE)
- **2009**: 811.172539879368 1000 t (from http://citydata.wu.ac.at

...assumption: Predictions get better, the more Open data we integrate...

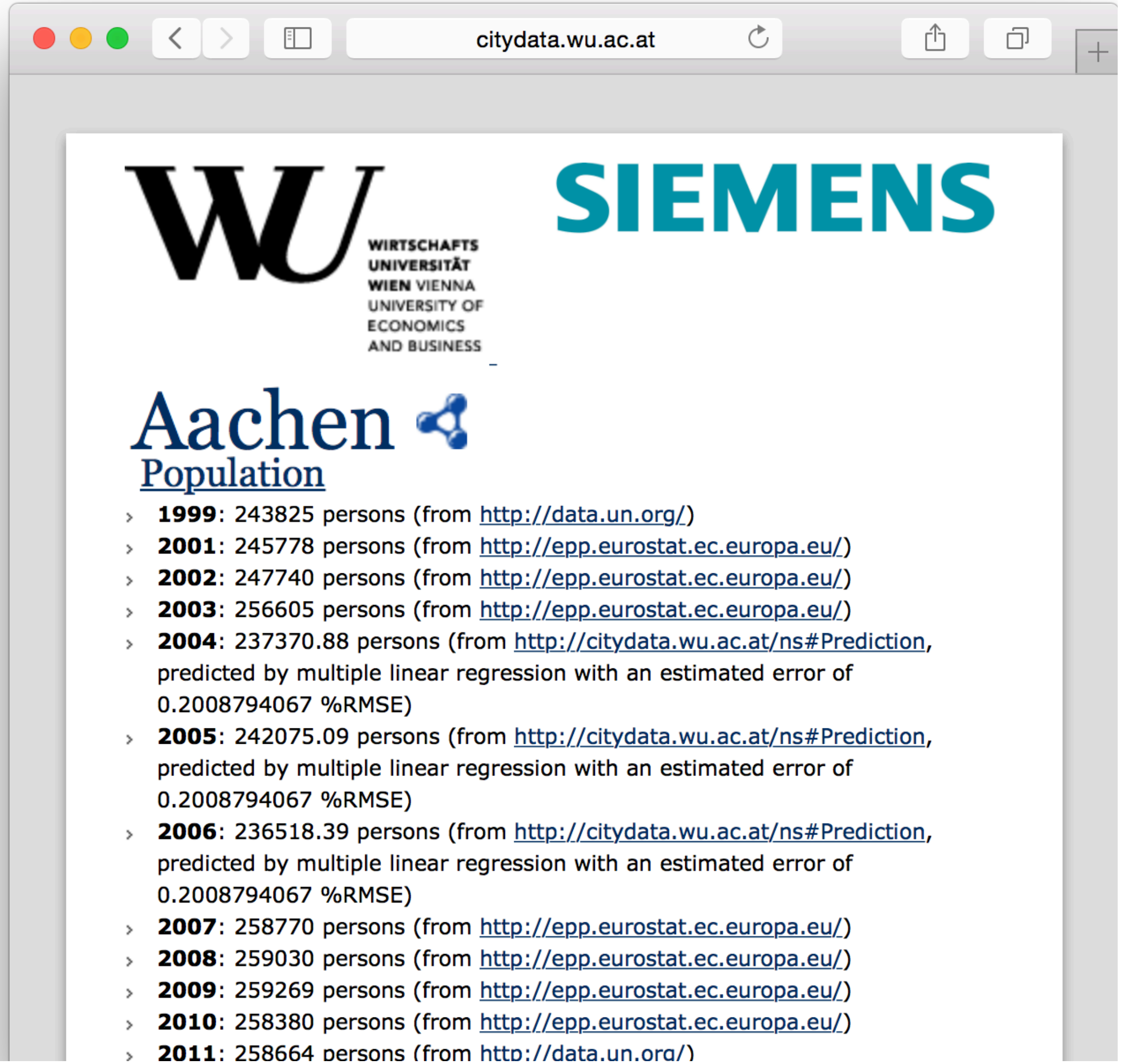
Data Quality

- Duplicates
- Incomplete values (partially solved by inferences/ OBDA)
    - Are OWL+RDFS actually enough?
        - Equations
        - Statistics
- Ambiguous/inconsistent values
    - actually, by inferences and OBDA, even more duplicates values → more possible inconsistencies

Still a lot to be done:
Time series analysis shows obvious inconsistencies

- Predictions on time series are partially very bad at the moment:

- Most of the data we look at is **time series data**/data changing over time.



citydata.wu.ac.at
WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

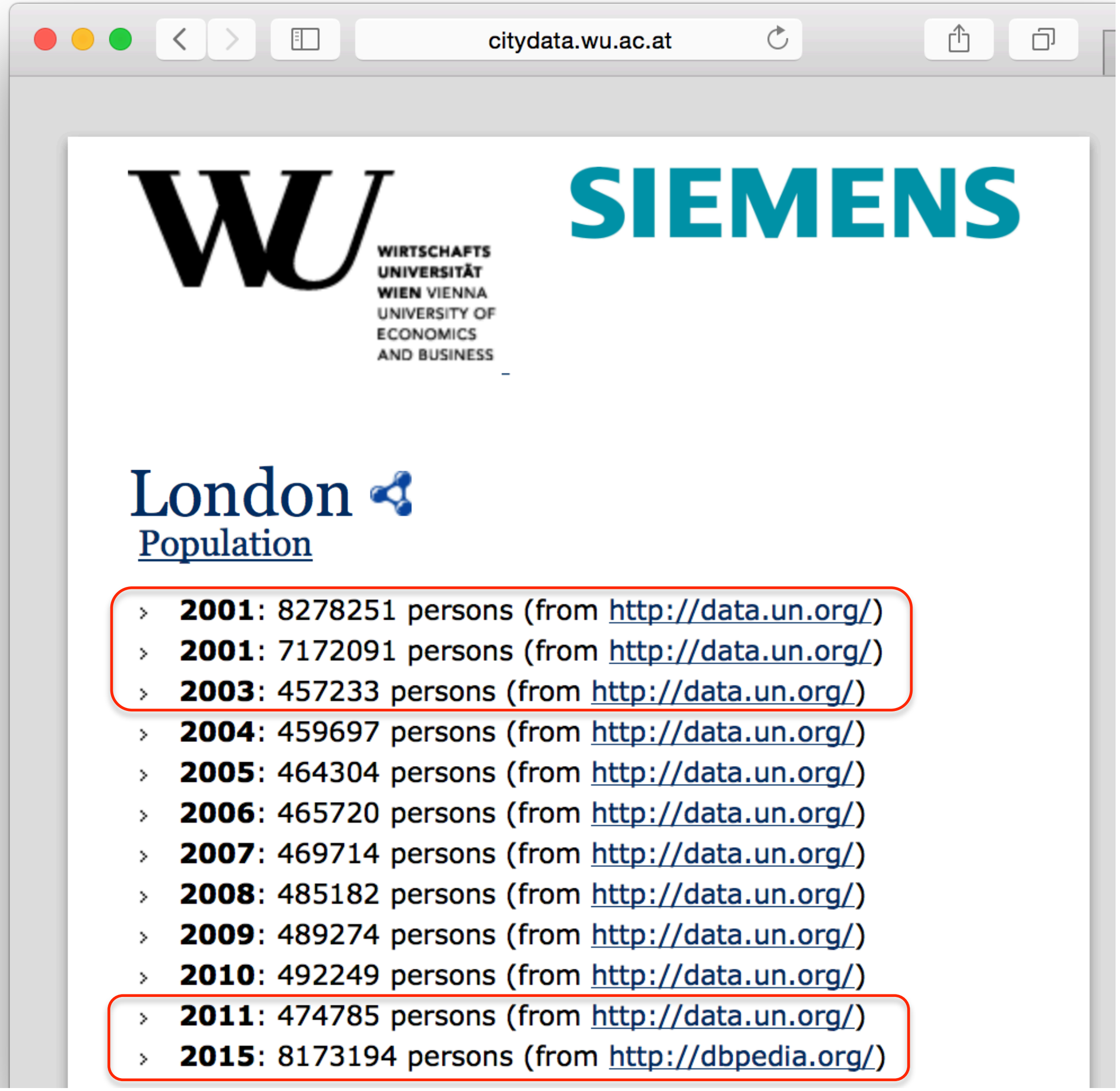
Aachen

Population

- **1999**: 243825 persons (from http://data.un.org/)
- **2001**: 245778 persons (from http://epp.eurostat.ec.europa.eu/)
- **2002**: 247740 persons (from http://epp.eurostat.ec.europa.eu/)
- **2003**: 256605 persons (from http://epp.eurostat.ec.europa.eu/)
- **2004**: 237370.88 persons (from http://citydata.wu.ac.at/ns#Prediction, predicted by multiple linear regression with an estimated error of 0.2008794067 %RMSE)
- **2005**: 242075.09 persons (from http://citydata.wu.ac.at/ns#Prediction, predicted by multiple linear regression with an estimated error of 0.2008794067 %RMSE)
- **2006**: 236518.39 persons (from http://citydata.wu.ac.at/ns#Prediction, predicted by multiple linear regression with an estimated error of 0.2008794067 %RMSE)
- **2007**: 258770 persons (from http://epp.eurostat.ec.europa.eu/)
- **2008**: 259030 persons (from http://epp.eurostat.ec.europa.eu/)
- **2009**: 259269 persons (from http://epp.eurostat.ec.europa.eu/)
- **2010**: 258380 persons (from http://epp.eurostat.ec.europa.eu/)
- **2011**: 258664 persons (from http://data.un.org/)

Still a lot to be done:
Open Data is incomparable/inconsistent in itself

- Surprising maybe, how much obviously weird data you find:
  - Inconsistencies **across** and **within** datasets



citydata.wu.ac.at
WU WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

London

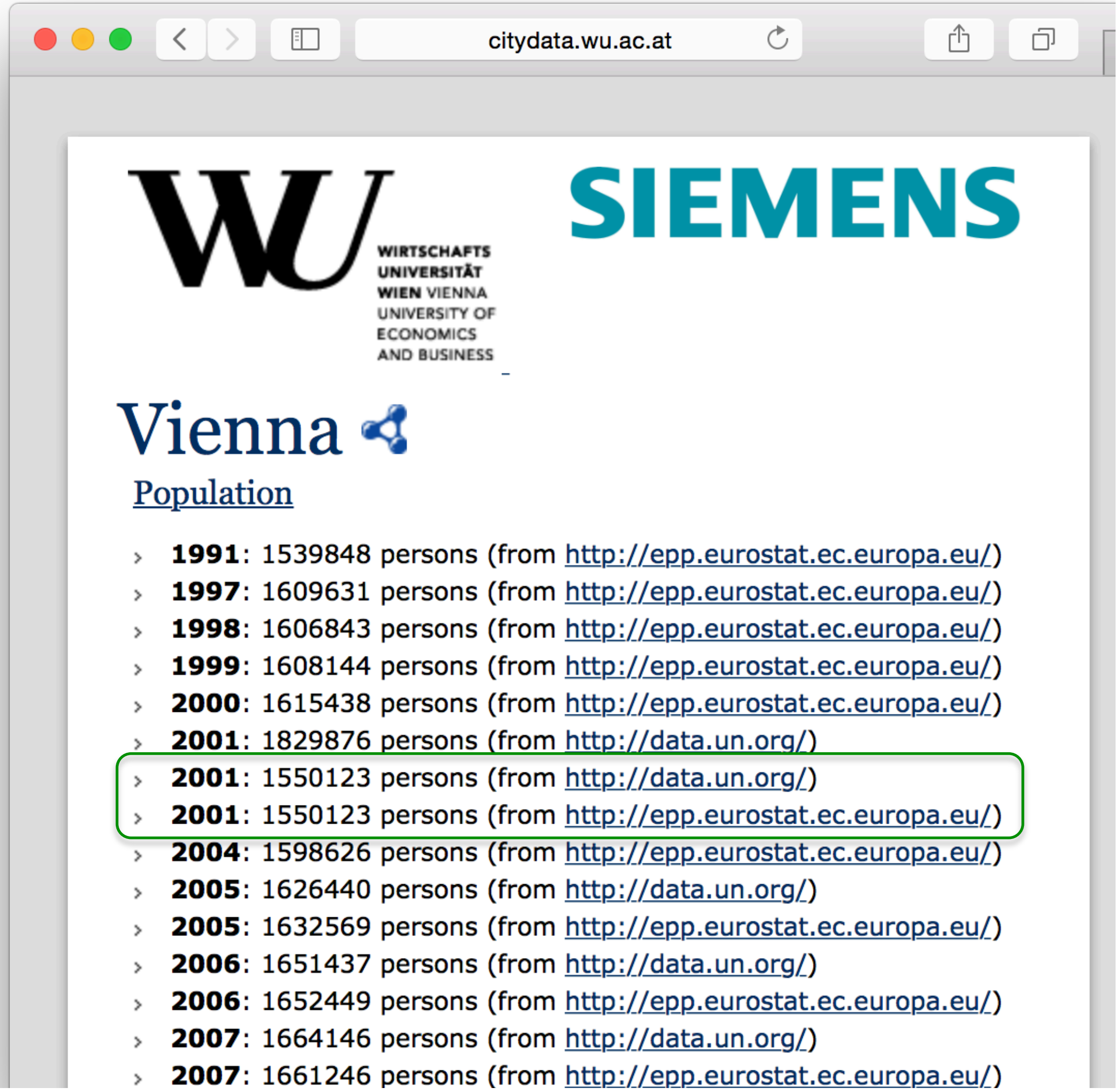
Population

- **2001**: 8278251 persons (from http://data.un.org/)
- **2001**: 7172091 persons (from http://data.un.org/)
- **2003**: 457233 persons (from http://data.un.org/)
- **2004**: 459697 persons (from http://data.un.org/)
- **2005**: 464304 persons (from http://data.un.org/)
- **2006**: 465720 persons (from http://data.un.org/)
- **2007**: 469714 persons (from http://data.un.org/)
- **2008**: 485182 persons (from http://data.un.org/)
- **2009**: 489274 persons (from http://data.un.org/)
- **2010**: 492249 persons (from http://data.un.org/)
- **2011**: 474785 persons (from http://data.un.org/)
- **2015**: 8173194 persons (from http://dbpedia.org/)

Still a lot to be done:
Open Data is incomparable/inconsistent in itself

- Surprising maybe, how much obviously weird data you find:
  - Inconsistencies across and within datasets
  - Still, some datasets match quite well on certain indicators
  - Open: (How) can we exploit this?

→ *Ontology learning? Predicting values from one dataset into the other?*


citydata.wu.ac.at
WU
WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS
SIEMENS

Vienna

Population

- **1991**: 1539848 persons (from http://epp.eurostat.ec.europa.eu/)
- **1997**: 1609631 persons (from http://epp.eurostat.ec.europa.eu/)
- **1998**: 1606843 persons (from http://epp.eurostat.ec.europa.eu/)
- **1999**: 1608144 persons (from http://epp.eurostat.ec.europa.eu/)
- **2000**: 1615438 persons (from http://epp.eurostat.ec.europa.eu/)
- **2001**: 1829876 persons (from http://data.un.org/)
- **2001**: 1550123 persons (from http://data.un.org/)
- **2001**: 1550123 persons (from http://epp.eurostat.ec.europa.eu/)
- **2004**: 1598626 persons (from http://epp.eurostat.ec.europa.eu/)
- **2005**: 1626440 persons (from http://data.un.org/)
- **2005**: 1632569 persons (from http://epp.eurostat.ec.europa.eu/)
- **2006**: 1651437 persons (from http://data.un.org/)
- **2006**: 1652449 persons (from http://epp.eurostat.ec.europa.eu/)
- **2007**: 1664146 persons (from http://data.un.org/)
- **2007**: 1661246 persons (from http://epp.eurostat.ec.europa.eu/)

CONCLUSIONS

Conclusions

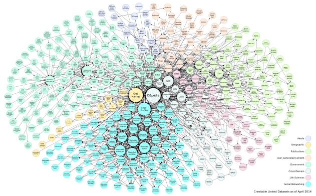
Heterogeneous Web Sources

eurostat

IDB

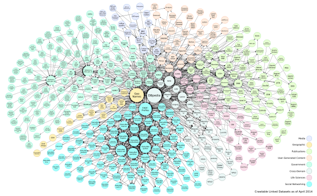
Inter-American Development Bank

The World Bank

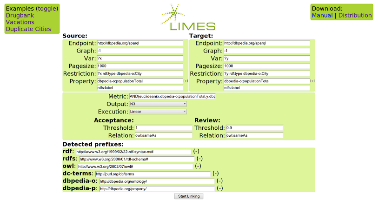


Conclusions

Heterogeneous Web Sources



Tools to Access/Integrate Web Sources



Examples (toggle)
Drugbank
Vacations
Duplicate Cities

LIMES

Download:
Manual | Distribution

Source:
Endpoint: http://dbpedia.org/sparql
Graph: -1
Var: ?x
Pagesize: 1000
Restriction: ?x rdf:type dbpedia-o:City
Property: dbpedia-o:populationTotal [+]
rdfs:label

Target:
Endpoint: http://dbpedia.org/sparql
Graph: -1
Var: ?y
Pagesize: 1000
Restriction: ?y rdf:type dbpedia-o:City
Property: dbpedia-o:populationTotal [+]
rdfs:label

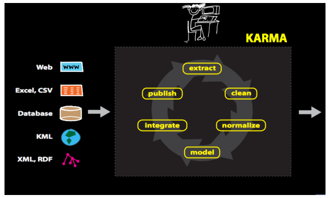
Metric: AND(euclidean(x.dbpedia-o:populationTotal,y.dbp
Output: N3
Execution: Linear

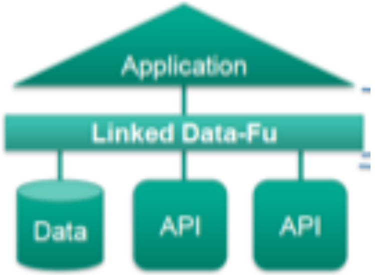
Acceptance:
Threshold: 1
Relation: owl:sameAs

Review:
Threshold: 0.9
Relation: owl:sameAs

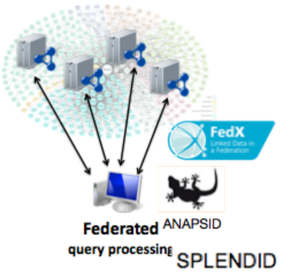
Detected prefixes:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# (-)
rdfs: http://www.w3.org/2000/01/rdf-schema# (-)
owl: http://www.w3.org/2002/07/owl# (-)
dc-terms: http://purl.org/dc/terms (-)
dbpedia-o: http://dbpedia.org/ontology/ (-)
dbpedia-p: http://dbpedia.org/property/ (-)

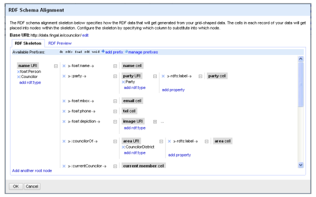
Start Linking





RDB2RDF Systems





RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://data.fingal.ie/councilor/ edit

RDF Skeleton | RDF Preview

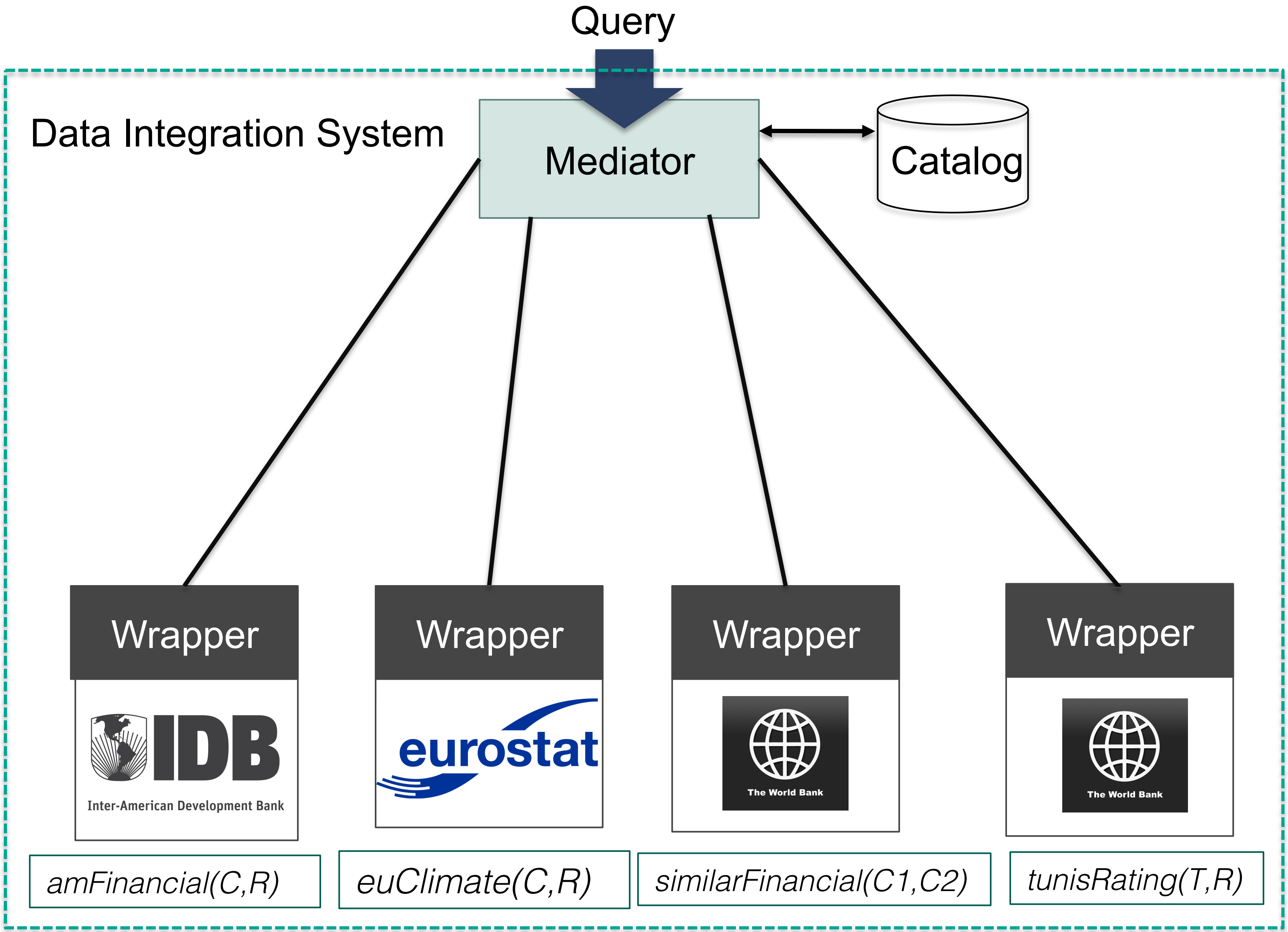
Available Prefixes: dc rdfs foaf rdf void + add prefix manage prefixes

name URI — foaf:Person, :Councilor, add rdf:type
- foaf:name → **name cell**
- :party → **party** URI (:Party, add rdf:type) — rdfs:label → **party cell**; add property
- foaf:mbox → **email cell**
- foaf:phone → **tel cell**
- foaf:depiction → **image** URI (add rdf:type) ...
- :councilorOf → **area** URI (:CouncilorDistrict, add rdf:type) — rdfs:label → **area cell**; add property
- :currentCouncilor → **current member cell**

Add another root node

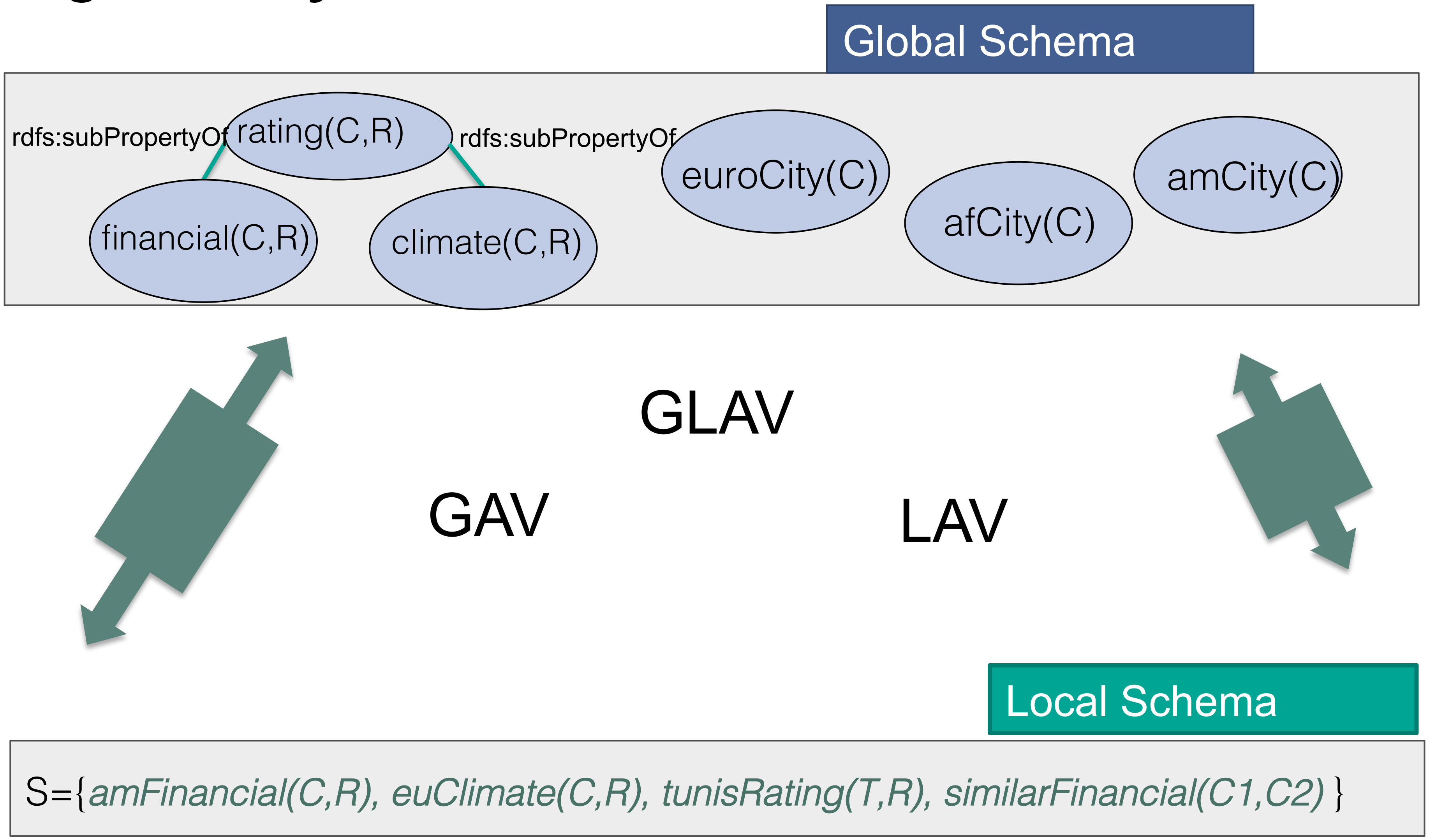
OK Cancel

Conclusions



[Wiederhold92]Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)

Integration Systems



Global Schema

rdfs:subPropertyOf rating(C,R) rdfs:subPropertyOf

financial(C,R)

climate(C,R)

euroCity(C)

afCity(C)

amCity(C)

GLAV

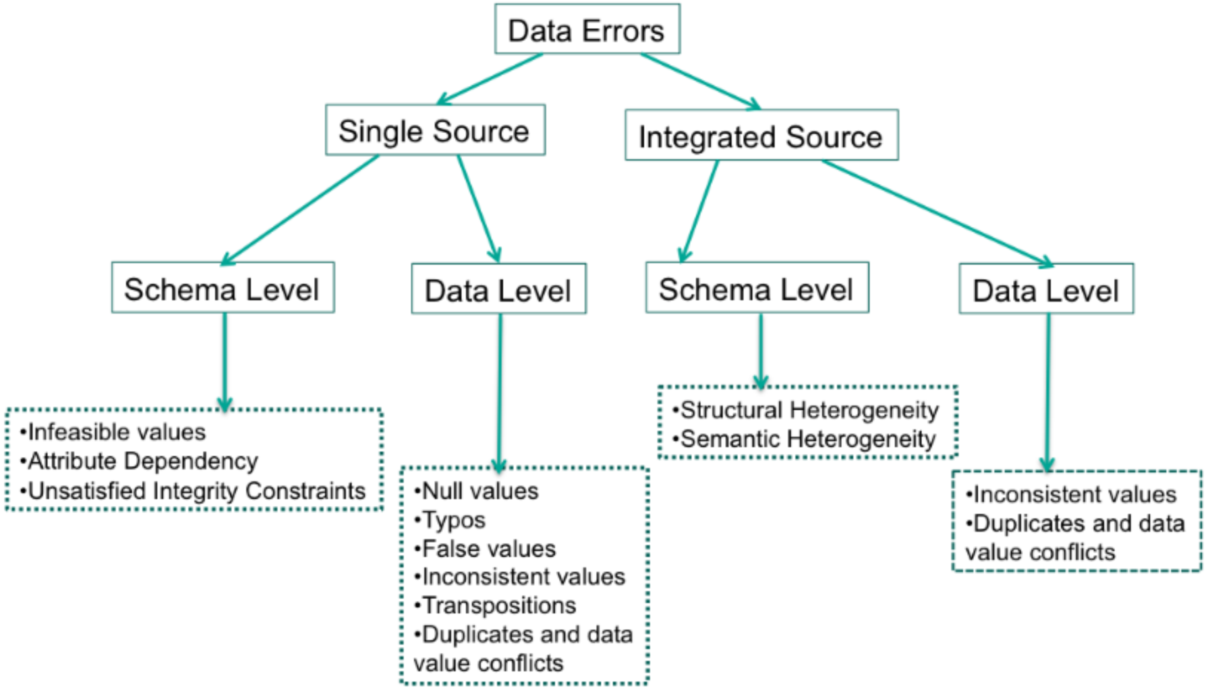
GAV

LAV

Local Schema

S={*amFinancial(C,R), euClimate(C,R), tunisRating(T,R), similarFinancial(C1,C2)* }

Data Quality Issues



Data Errors

Single Source

Integrated Source

Schema Level

Data Level

Schema Level

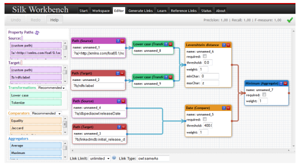
Data Level

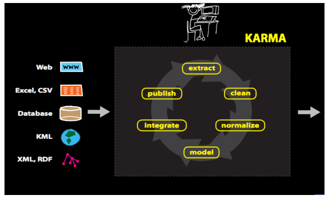
- Infeasible values
- Attribute Dependency
- Unsatisfied Integrity Constraints

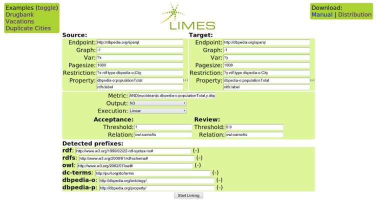
- Null values
- Typos
- False values
- Inconsistent values
- Transpositions
- Duplicates and data value conflicts

- Structural Heterogeneity
- Semantic Heterogeneity

- Inconsistent values
- Duplicates and data value conflicts






Examples (toggle)
Drugbank
Vacations
Duplicate Cities

LIMES

Download:
Manual | Distribution

Source:
Endpoint: http://dbpedia.org/sparql
Graph: -1
Var: ?x
Pagesize: 1000
Restriction: ?x rdf:type dbpedia-o:City
Property: dbpedia-o:populationTotal [+]
rdfs:label

Target:
Endpoint: http://dbpedia.org/sparql
Graph: -1
Var: ?y
Pagesize: 1000
Restriction: ?y rdf:type dbpedia-o:City
Property: dbpedia-o:populationTotal [+]
rdfs:label

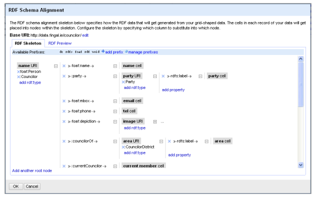
Metric: AND(euclidean(x.dbpedia-o:populationTotal,y.dbp
Output: N3
Execution: Linear

Acceptance:
Threshold: 1
Relation: owl:sameAs

Review:
Threshold: 0.9
Relation: owl:sameAs

Detected prefixes:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns# (-)
rdfs: http://www.w3.org/2000/01/rdf-schema# (-)
owl: http://www.w3.org/2002/07/owl# (-)
dc-terms: http://purl.org/dc/terms (-)
dbpedia-o: http://dbpedia.org/ontology/ (-)
dbpedia-p: http://dbpedia.org/property/ (-)

Start Linking


RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: http://data.fingal.ie/councilor/ edit

Take-home messages:

- Semantic Web technologies help in Open Data Integration workflows and can add flexibility
- It's worthwhile to consider traditional "Data Integration" approaches & literature!
- Non-Clean Data requires: Statistics & machine learning (outlier detection, imputing missing values, resolving inconsistencies, etc.)

Many Thanks!
Questions

References

References 1

- [Polleres 2013] Axel Polleres. Tutorial "OWL vs. Linked Data: Experiences and Directions" OWLED2013. http://polleres.net/presentations/20130527OWLED2013_Invited_talk.pdf
- [Polleres et al. 2013] Axel Polleres, Aidan Hogan, Renaud Delbru, Jürgen Umbrich: RDFS and OWL Reasoning for Linked Data. Reasoning Web 2013: 91-149
- [Golfarelli,Rizzi,2009] Matteo Golfarelli, Stefano Rizzi. Data Warehouse Design: Modern Principles and Methodologies. McGraw-Hill, 2009.
- [Lenzerini2002] Maurizio Lenzerini: Data Integration: A Theoretical Perspective. PODS 2002: 233-246
- [Auer et al. 2012] Sören Auer, Lorenz Bühmann, Christian Dirschl, Orri Erling, Michael Hausenblas, Robert Isele, Jens Lehmann, Michael Martin, Pablo N. Mendes, Bert Van Nuffelen, Claus Stadler, Sebastian Tramp, Hugh Williams: Managing the Life-Cycle of Linked Data with the LOD2 Stack. International Semantic Web Conference (2) 2012: 1-16 see also http://stack.lod2.eu/
- [Taheriyan et al. 2012] Mohsen Taheriyan, Craig A. Knoblock, Pedro A. Szekely, José Luis Ambite: Rapidly Integrating Services into the Linked Data Cloud. International Semantic Web Conference (1) 2012: 559-574
- [Bischof et al. 2012] Stefan Bischof, Stefan Decker, Thomas Kr ennwallner, Nuno Lopes, Axel Polleres: Mapping between RDF and XML with XSPARQL. J. Data Semantics 1(3): 147-185 (2012)
- [Nonaka & Takeuchi, 1995] "The Knowledge-Creating Company - How Japanese Companies Create the Dynamics of Innovation" (Nonaka, Takeuchi, New York Oxford 1995)
- [Bischof et al. 2015] Stefan Bischof, Christoph Martin, Axel Polleres, and Patrik Schneider. Open City Data Pipeline: Collecting, Integrating, and Predicting Open City Data. In *4th Workshop on Knowledge Discovery and Data Mining Meets Linked Open Data (Know@LOD)*, Portoroz, Slovenia, May 2015.

References 2

- [Levy & Rajaraman & Ullman 1996] Alon Y. Levy, Anand Rajaraman, Jeffrey D. Ullman: Answering Queries Using Limited External Processors. PODS 1996: 227-237
- [Duscka & Genesereth 1997]
- [Pottinger & Halevy 2001] Rachel Pottinger, Alon Y. Halevy: MiniCon: A scalable algorithm for answering queries using views. VLDB J. 10(2-3): 182-198 (2001)
- [Arvelo & Bonet & Vidal 2006] Yolifé Arvelo, Blai Bonet, Maria-Esther Vidal: Compilation of Query-Rewriting Problems into Tractable Fragments of Propositional Logic. AAAI 2006: 225-230
- [Konstantinidis & Ambite, 2011] George Konstantinidis, José Luis Ambite: Scalable query rewriting: a graph-based approach. SIGMOD Conference 2011: 97-108
- [Izquierdo & Vidal & Bonet 2011] Daniel Izquierdo, Maria-Esther Vidal, Blai Bonet: An Expressive and Efficient Solution to the Service Selection Problem. International Semantic Web Conference (1) 2010: 386-401
- [Wiederhold92] Gio Wiederhold: Mediators in the Architecture of Future Information Systems. IEEE Computer 25(3): 38-49 (1992)
- [Stadtmüller et al. 2013] Steffen Stadtmüller, Sebastian Speiser, Andreas Harth, Rudi Studer: Data-Fu: a language and an interpreter for interaction with read/write linked data. WWW 2013: 1225-1236

References 3

- [Priyatna et al. 2014] Freddy Priyatna, Óscar Corcho, Juan Sequeda: Formalisation and experiences of R2RML-based SPARQL to SQL query translation using morph. WWW 2014: 479-490
- [Sequeda & Miranker 2013] Juan Sequeda, Daniel P. Miranker. Ultrawrap: SPARQL execution on relational data. J. Web Sem. 22: 19-39 (2013)
- [Krötzsch 2012] Markus Krötzsch: OWL 2 Profiles: An Introduction to Lightweight Ontology Languages. Reasoning Web 2012: 112-183
- [Glimm et al. 2012] Birte Glimm, Aidan Hogan, Markus Krötzsch, Axel Polleres: OWL: Yet to arrive on the Web of Data? LDOW 2012
- [Kontchakov et al. 2013] Roman Kontchakov, Mariano Rodriguez-Muro, Michael Zakharyaschev: Ontology-Based Data Access with Databases: A Short Course. Reasoning Web 2013: 194-229
- [Calvanese et al. 2007] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Riccardo Rosati: Tractable Reasoning and Efficient Query Answering in Description Logics: The DL-Lite Family. J. Autom. Reasoning 39(3): 385-429 (2007)
- [Perez-Urbina et al., 2009] Héctor Pérez-Urbina, Boris Motik and Ian Horrocks, A Comparison of Query Rewriting Techniques for DL-Lite, In Proc. of the Int. Workshop on Description Logics (DL 2009), Oxford, UK, July 2009.
- [Rodriguez-Muro, et al. 2012] Mariano Rodriguez-Muro, Diego Calvanese: Quest, an OWL 2 QL Reasoner for Ontology-based Data Access. OWLED 2012
- [Rodriguez-Muro, et al. 2013] Mariano Rodriguez-Muro, Roman Kontchakov, Michael Zakharyaschev: Ontology-Based Data Access: Ontop of Databases. International Semantic Web Conference (1) 2013: 558-573
- [Calvanese et al. 2011] Diego Calvanese, Giuseppe De Giacomo, Domenico Lembo, Maurizio Lenzerini, Antonella Poggi, Mariano Rodriguez-Muro, Riccardo Rosati, Marco Ruzzi, Domenico Fabio Savo: The MASTRO system for ontology-based data access. Semantic Web 2(1): 43-53 (2011)
- [Rosati et al. 2010] Riccardo Rosati, Alessandro Almatelli: Improving Query Answering over DL-Lite Ontologies. KR 2010
- [Mora & Corcho, 2014] José Mora, Riccardo Rosati, Óscar Corcho: kyrie2: Query Rewriting under Extensional Constraints in ELHIO. Semantic Web Conference (1) 2014: 568-583
- [Sequeda et al. 2014] Juan F. Sequeda, Marcelo Arenas, Daniel P. Miranker: OBDA: Query Rewriting or Materialization? In Practice, Both! Semantic Web Conference (1) 2014: 535-551

References 4

- [Acosta et al 2011] M. Acosta, M.-E. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. Anapsid: an adaptive query processing engine for sparql endpoints. ISWC 2011.
- [Basca and Bernstein 2014] C. Basca and A. Bernstein. Querying a messy web of data with avalanche. In Journal of Web Semantics, 2014.
- [Cohen-Boalaki and . Leser. 2013] S. Cohen-Boalakia, U. Leser. Next Generation Data Integration for the Life Sciences. Tutorial at ICDE 2013. https://www2.informatik.hu-berlin.de/~leser/icde_tutorial_final_public.pdf
- [Doan et el. 2012] A. Doan, A. Halevy, Z. Ives, Data Integration. Morgan Kaukman 2012.
- [Halevy et al 2006] A. Y. Halevy, A. Rajaraman, J. Ordille: Data Integration: The Teenage Years. VLDB 2006: 9-16.
- [Halevy et al 2001] A. Y. Halevy. Answering queries using views: A survey. VLDB J., 2001.
- [Hassanzadeh et al. 2013] Oktie Hassanzadeh, Ken Q. Pu, Soheil Hassas Yeganeh, Renée J. Miller, Lucian Popa, Mauricio A. Hernández, Howard Ho: Discovering Linkage Points over Web Data. PVLDB 2013
- [Gorlitz and Staab 2011] O. Gorlitz and S. Staab. SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions. In Proceedings of the 2nd International Workshop on Consuming Linked Data, 2011.
- [Schwarte et al. 2011] A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. Fedx: Optimization techniques for federated query processing on linked data. ISWC 2011.
- [Verborgh et al. 2014] Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, Rik Van de Walle: Querying Datasets o the Web with High Availability. ISWC2014

References 5

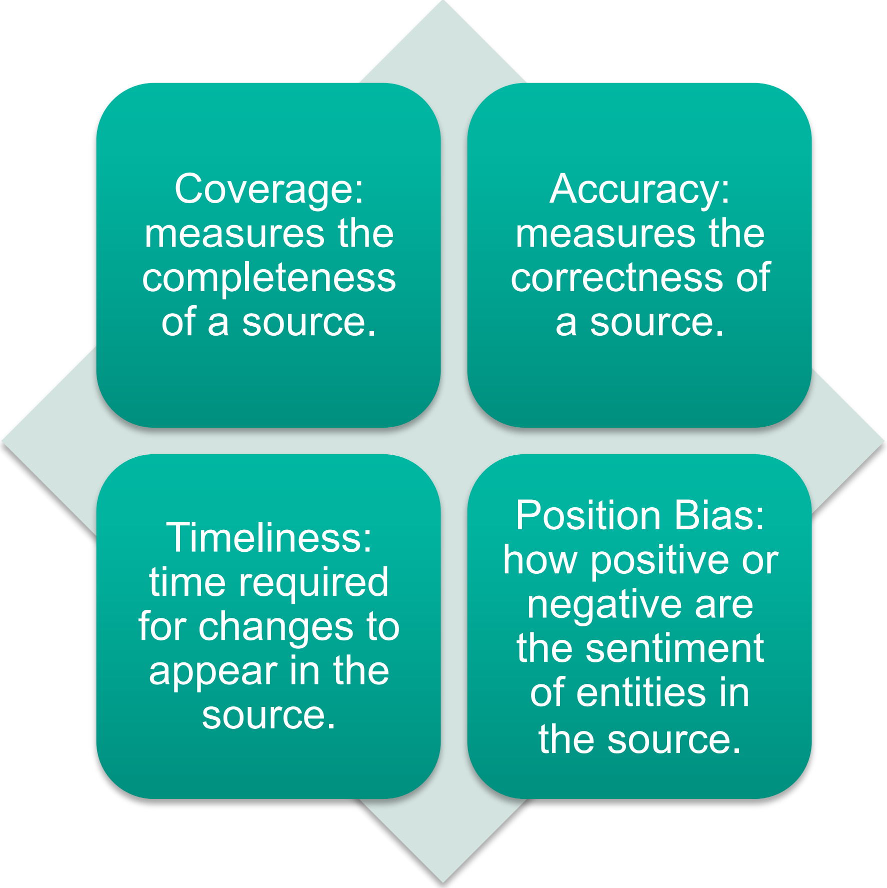
- [Acosta et al. 2015] Maribel Acosta, Amrapali Zaveri, Elena Simperl, Dimitris Kontokostas, Sören Auer, Jens Lehmann: Crowdsourcing Linked Data Quality Assessment. ISWC 2013
- [Lenz 2007] Hans - J. Lenz. Data Quality Defining, Measuring and Improving. Tutorial at IDA 2007.
- [Naumann02] Felix Naumann: Quality-Driven Query Answering for Integrated Information Systems. LNCS 2261, Springer 2002
- [Ngonga et al. 2011] Axel-Cyrille Ngonga Ngomo, Sören Auer: LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. IJCAI 2011
- [Saleem et al 2014] Muhammad Saleem, Maulik R. Kamdar, Aftab Iqbal, Shanmukha Sampath, Helena F. Deus, Axel-Cyrille Ngonga Ngomo: Big linked cancer data: Integrating linked TCGA and PubMed. J. Web Sem. 2014
- [Soru et al. 2015] Tommaso Soru, Edgard Marx, Axel-Cyrille Ngonga Ngomo: ROCKER: A Refinement Operator for Key Discovery. WWW 2015
- [Volz et al 2009] Julius Volz, Christian Bizer, Martin Gaedke, Georgi Kobilarov: Discovering and Maintaining Links on the Web of Data. ISWC 2009
- [Zaveri,et al 2015] Amrapali J. Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality Assessment for Linked Data: A Survey. Semantic Web Journal 2015
- [Hernandez&Stolfo, 1998] M. A. Hernández, S. J. Stolfo: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem.
- Data Min. Knowl. Discov. 2(1): 9-37 (1998)

TRENDS & OPEN RESEARCH PROBLEMS (SOME)

Data Source Quality in Integration Systems

- IS=<O,S,M>

Sources in S are described in terms of Quality Metrics:


Coverage: measures the completeness of a source.

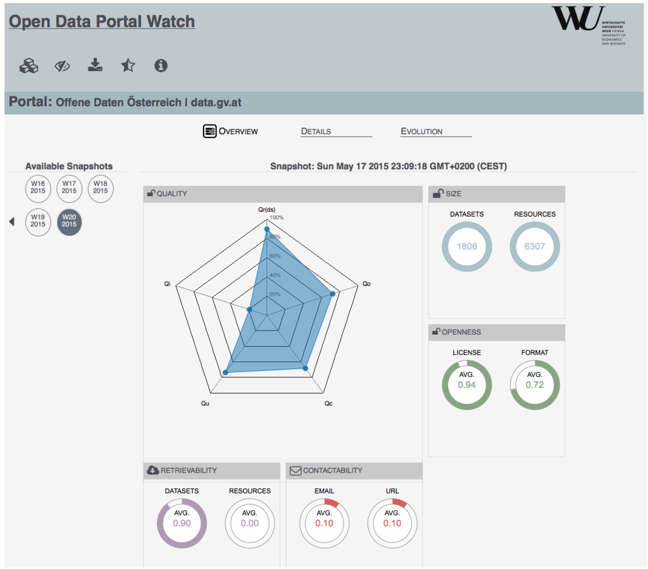
Accuracy: measures the correctness of a source.

Timeliness: time required for changes to appear in the source.

Position Bias: how positive or negative are the sentiment of entities in the source.

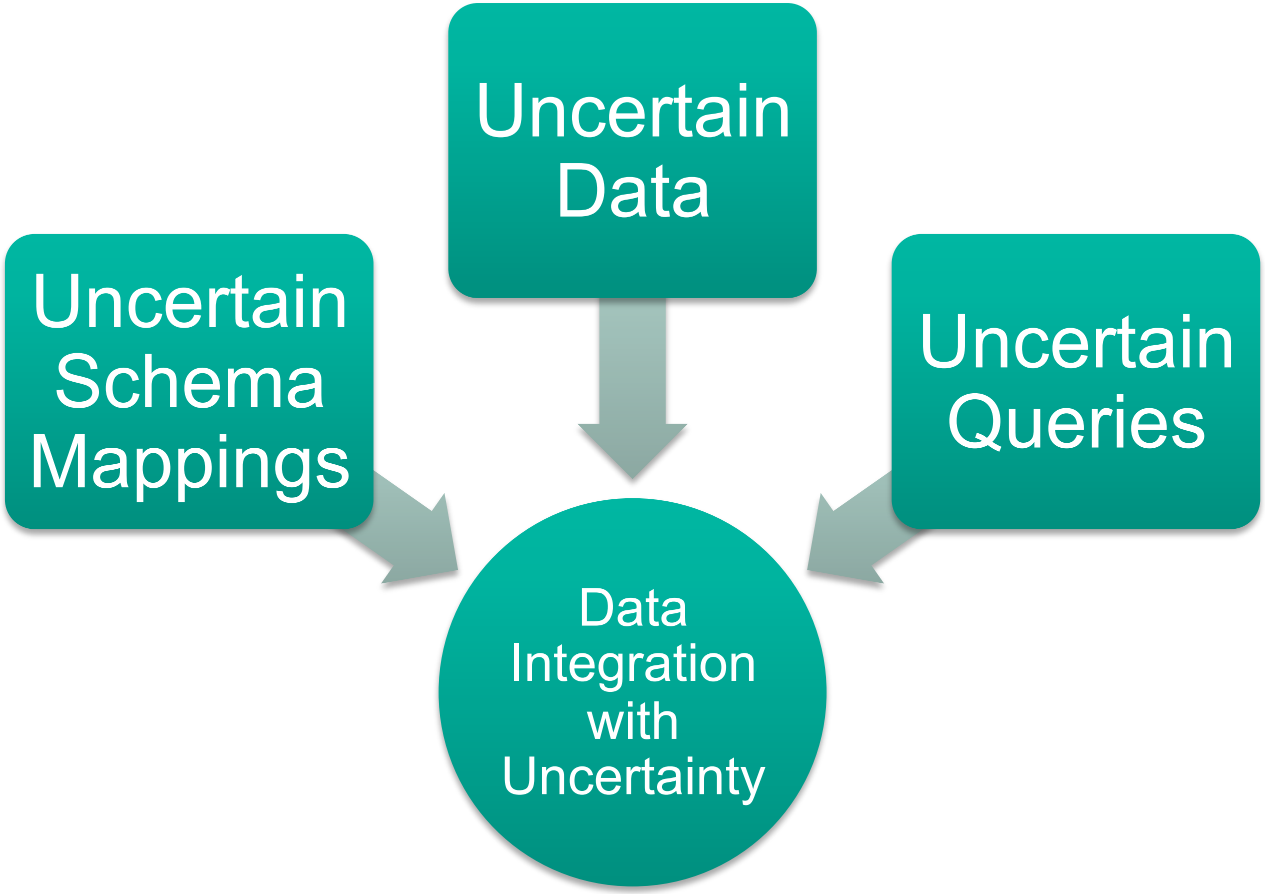
Possible use case:

http://data.wu.ac.at/portalwatch



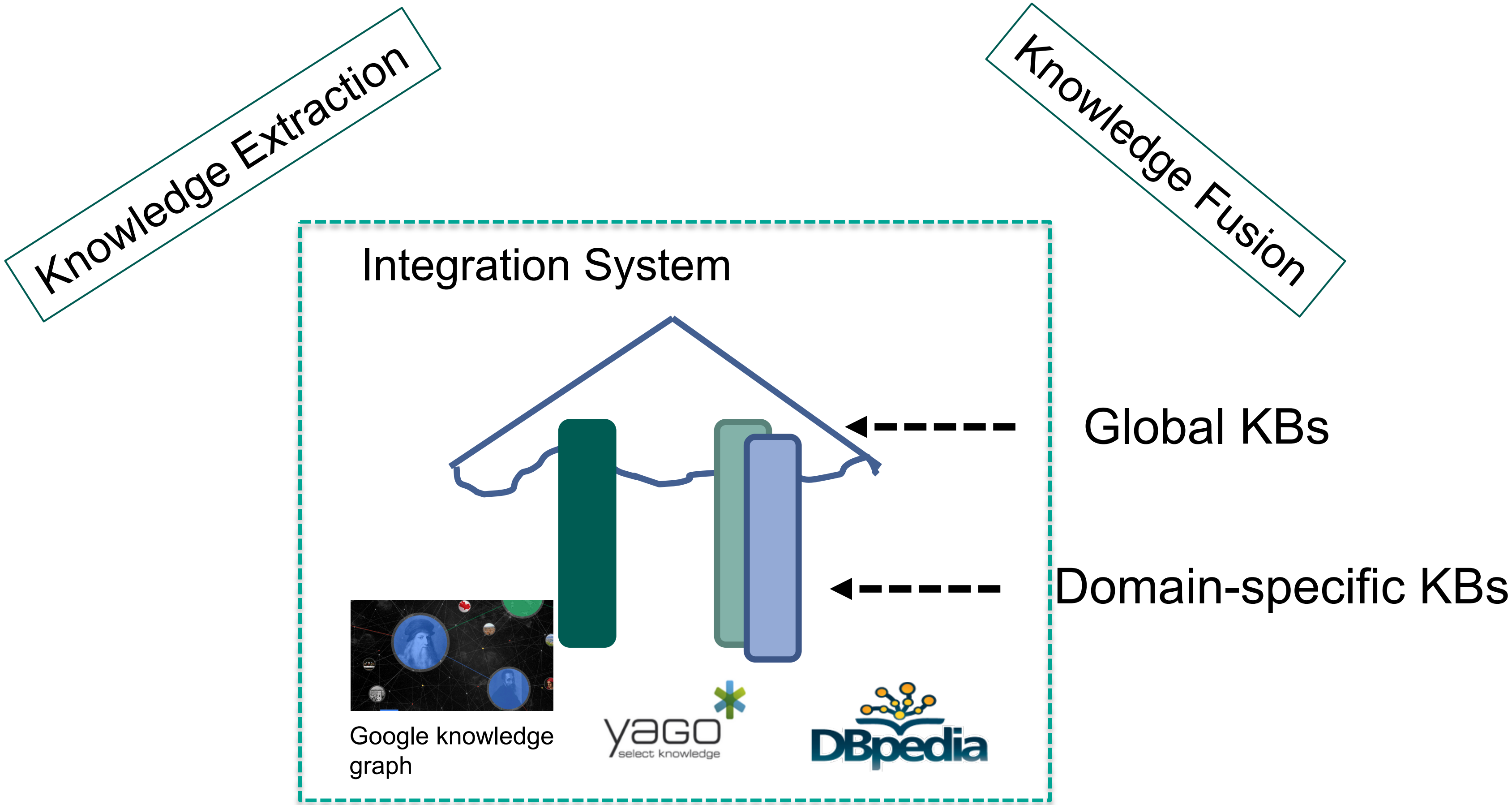
Theodoros Rekatsinas, Xin Luna Dong, Lise Getoor and Divesh Srivastava. Finding quality in quantity: the challenge of discovering valuable sources for integration.CIDR 2015

Uncertainty in Integration Systems



Xin Luna Dong, Alon Y. Halevy, Cong Yu: Data integration with uncertainty. VLDB J. 18(2): 469-500 (2009)

Knowledge Integration Systems



Knowledge Extraction

Knowledge Fusion

Integration System

Global KBs

Domain-specific KBs

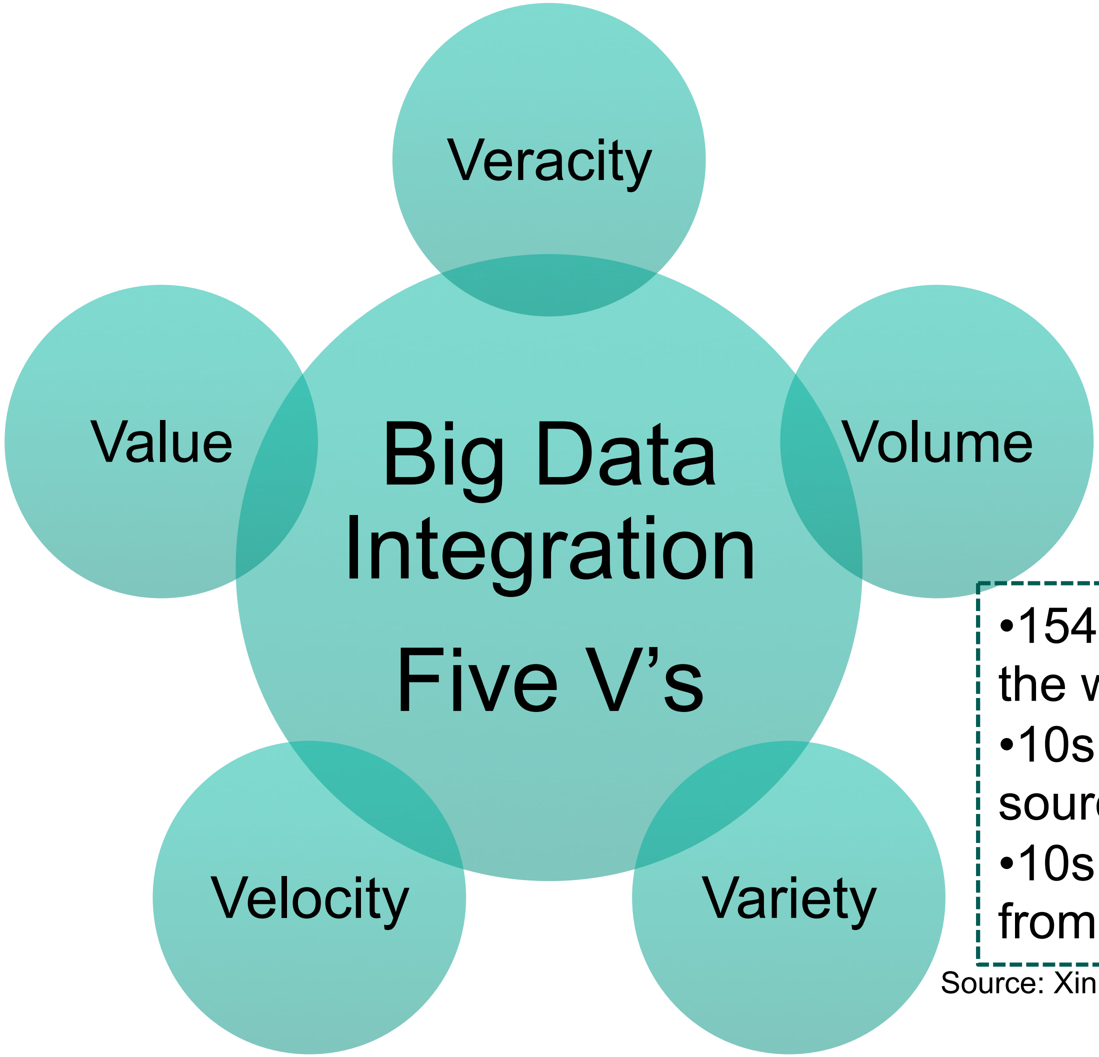
Google knowledge graph

Global KBs:

- Covers a variety of knowledge across domains
- Intensional: Cyc, WordNet
- Extensional: Freebase, Knowledge Graph, Yago/Yago2, DeepDive, NELL, Prospera, ReVerb, Knowledge Vault

Knowledge Curation and Knowledge Fusion: challenges, models and applications. Xin Luna Dong and Divesh Srivastava. Tutorial in Proceedings of the ACM SIGMOD 2015.

Big Data Integration Systems


Big Data Integration

Five V’s

Veracity

Value

Volume

Velocity

Variety

- 154 million high quality relational tables on the web
- 10s of millions of high quality deep web sources
- 10s of millions of useful relational tables from web lists

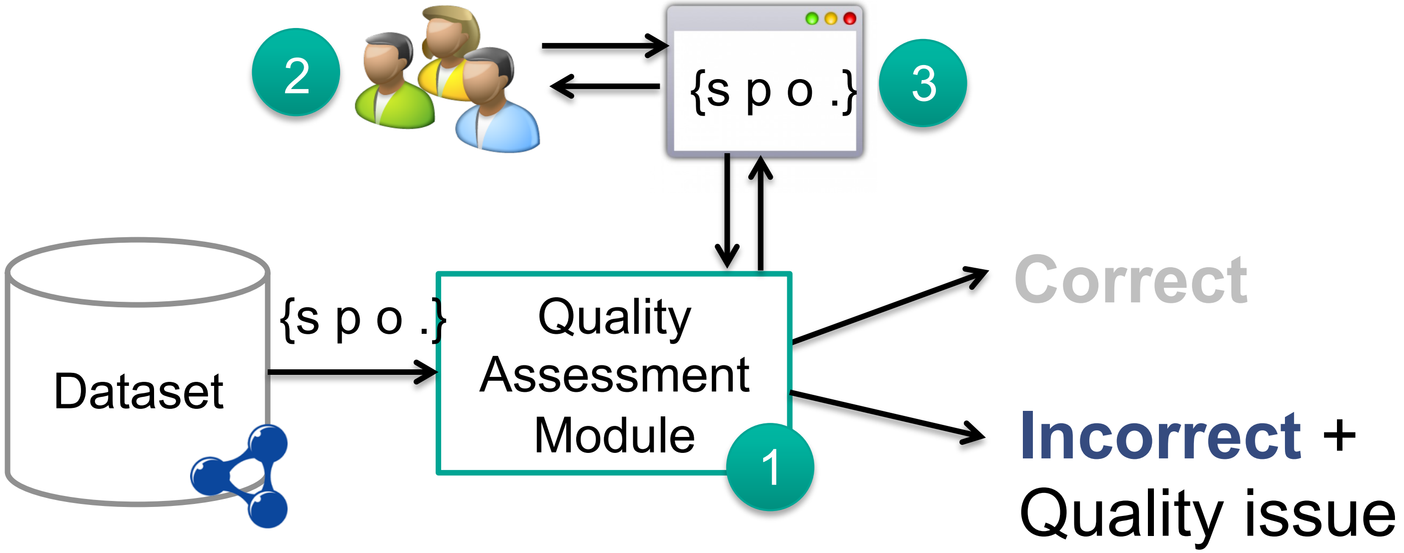
Source: Xin Luna Dong, Divesh Srivastava. Big Data Integration. PVLDB2013

Serge Abiteboul, Xin Luna Dong, Oren Etzioni, Divesh Srivastava, Gerhard Weikum, Julia Stoyanovich and Fabian M. Suchanek. The elephant in the room: getting value from big data. WebDB 2015

Xin Luna Dong, Divesh Srivastava. Big Data Integration. PVLDB2013

. Barna Saha and Divesh Srivastava. Data quality: the other face of big data Tutorial in ICDE 2014

Crowdsourcing LD Quality Assessment



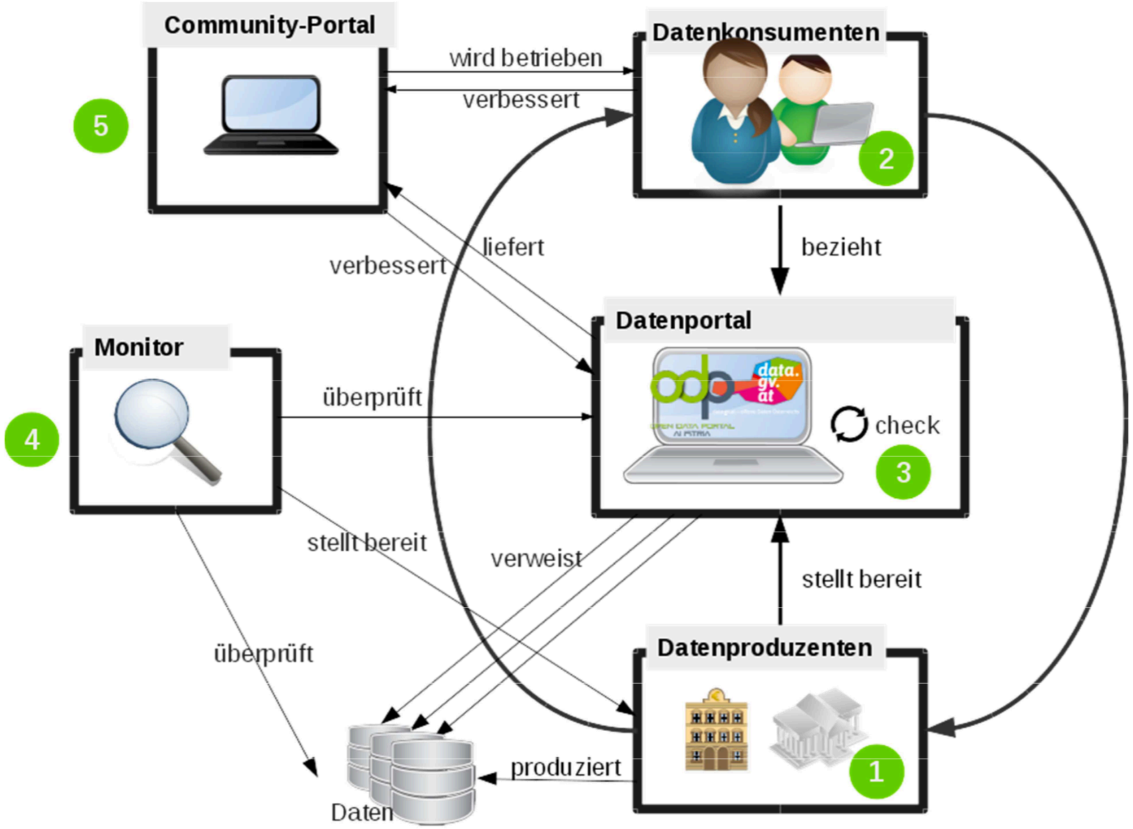
Challenges

1. Selecting **LD quality issues** to crowdsource
2. Selecting the appropriate **crowdsourcing approaches**
3. Generating the **interfaces** to present the data to the crowd

Acosta et al. – *Crowdsourcing Linked Data Quality Assessment*. ISWC, 2013.

Crowdsourcing Open Quality Assessment

How to involve the users in CKAN portals?

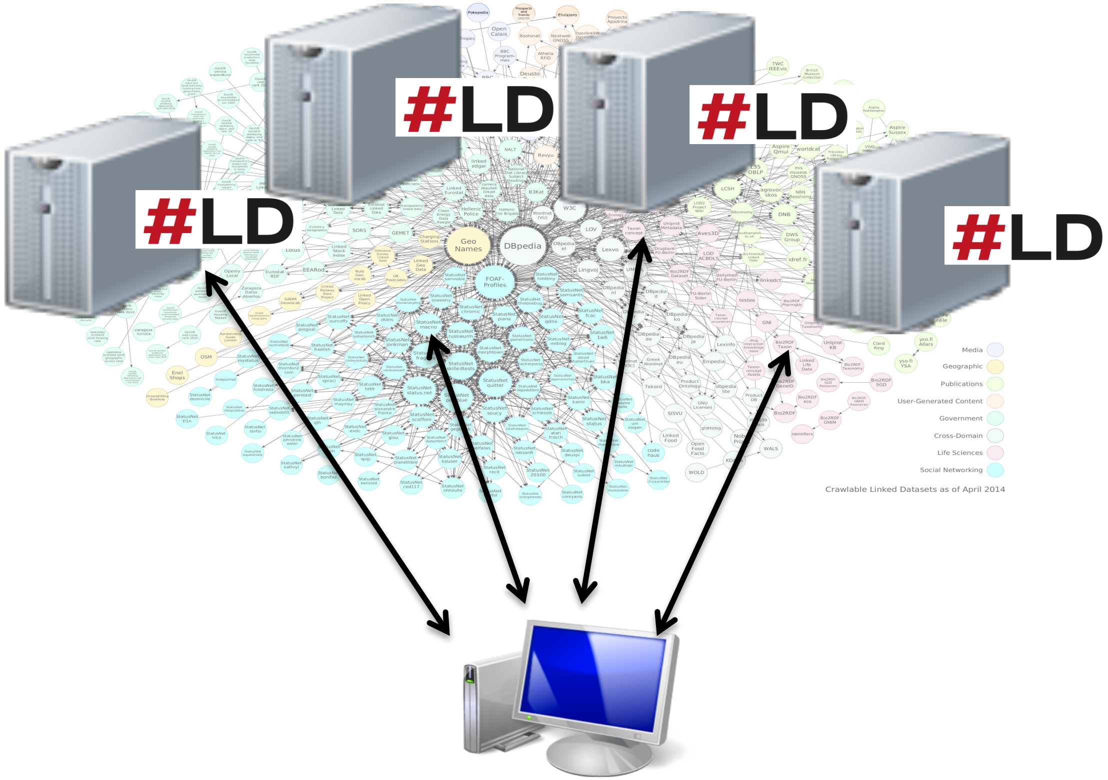


COOPERATION OGD ÖSTERREICH

SPARQL Query Execution using LAV views

Publicly available Linked Data Fragments (LAV views)

Servers



Mediator

Lower Bounds for the Space of Query Rewritings

- CQs and OWL2QL-ontologies [Gottlob14]
  - Exponential and Superpolynomial lower bounds on the size of pure rewritings.
  - Polynomial-size under some restrictions.

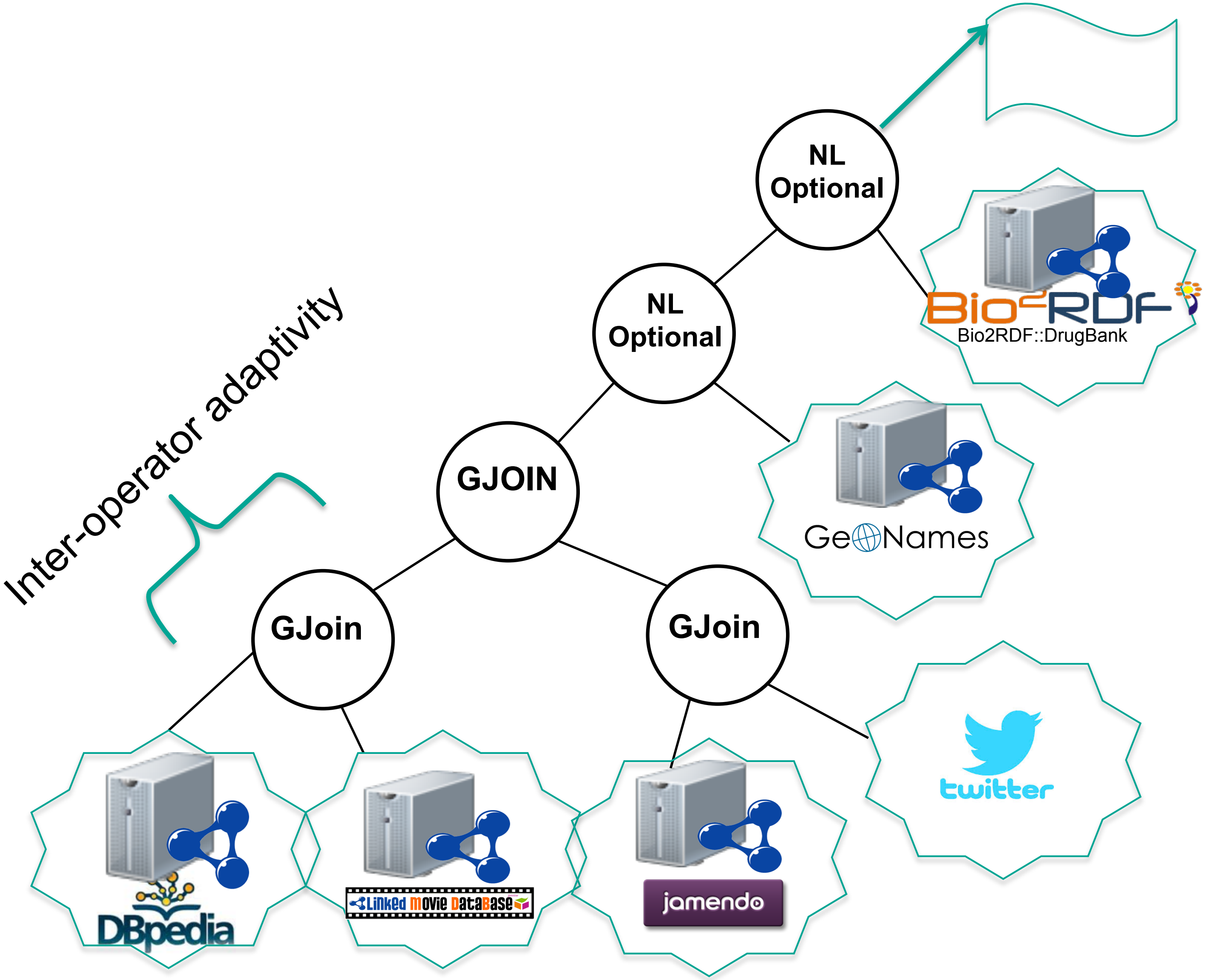
[Gottlob14]

Georg Gottlob, Stanislav Kikot, Roman Kontchakov, Vladimir V. Podolskii, Thomas Schwentick, Michael Zakharyaschev: The price of query rewriting in ontology-based data access. Artif. Intell. 213: 42-59 (2014)

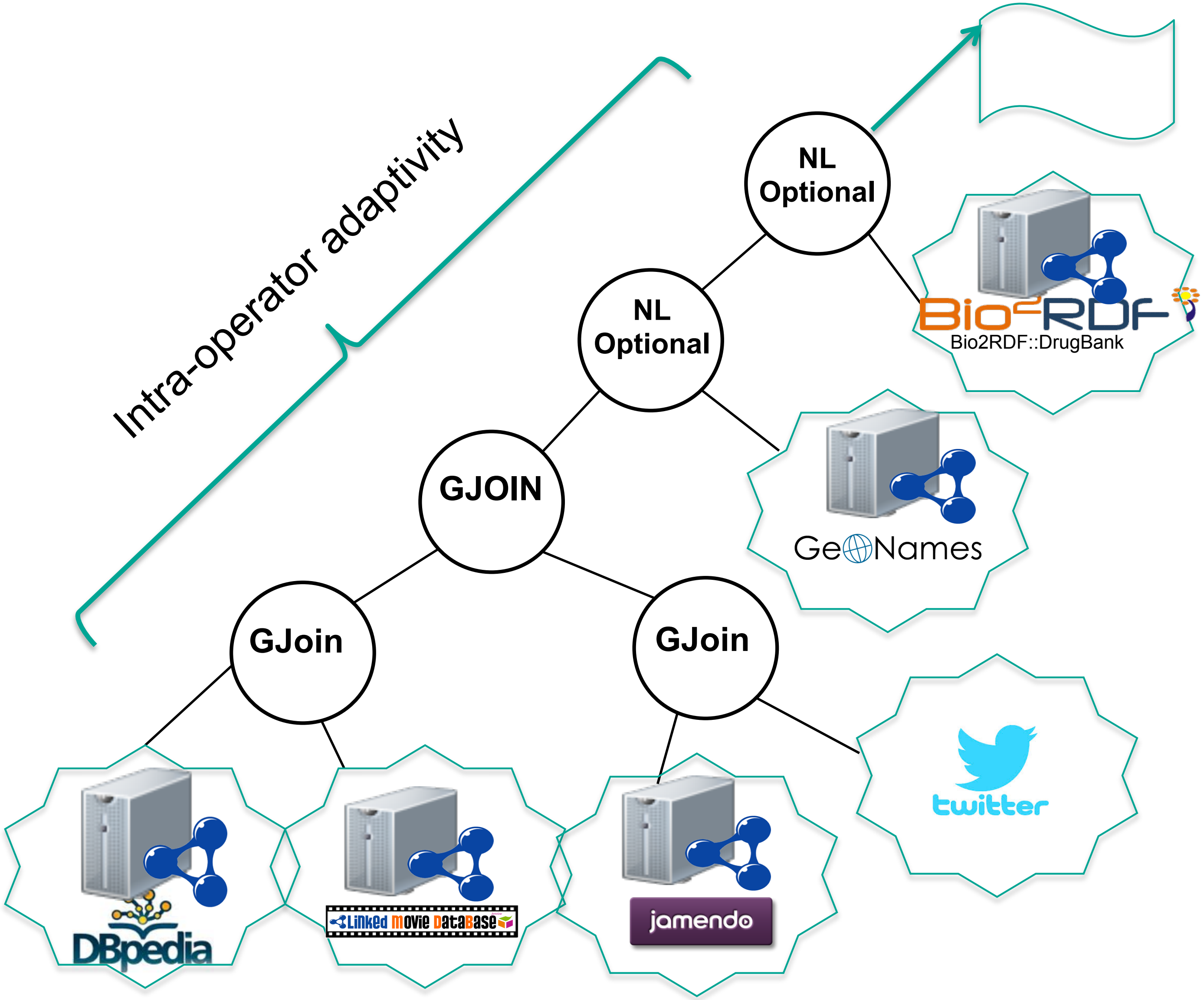
SPARQL Query Execution using LAV views.

- Gabriela Montoya, Luis Daniel Ibáñez, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal. SemLAV: Local-As-View Mediation for SPARQL Queries. T. Large-Scale Data- and Knowledge-Centered Systems 13: 33-58 (2014).
- Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle. Querying Datasets on the Web with High Availability. ISWC 2014.
- Gabriela Montoya, Hala Skaf-Molli, Pascal Molli, Maria-Esther Vidal. Federated SPARQL Queries Processing with Replicated Fragments. Accepted at ISWC 2015.

Adaptive Execution of SPARQL Queries



Adaptive Execution of SPARQL Queries

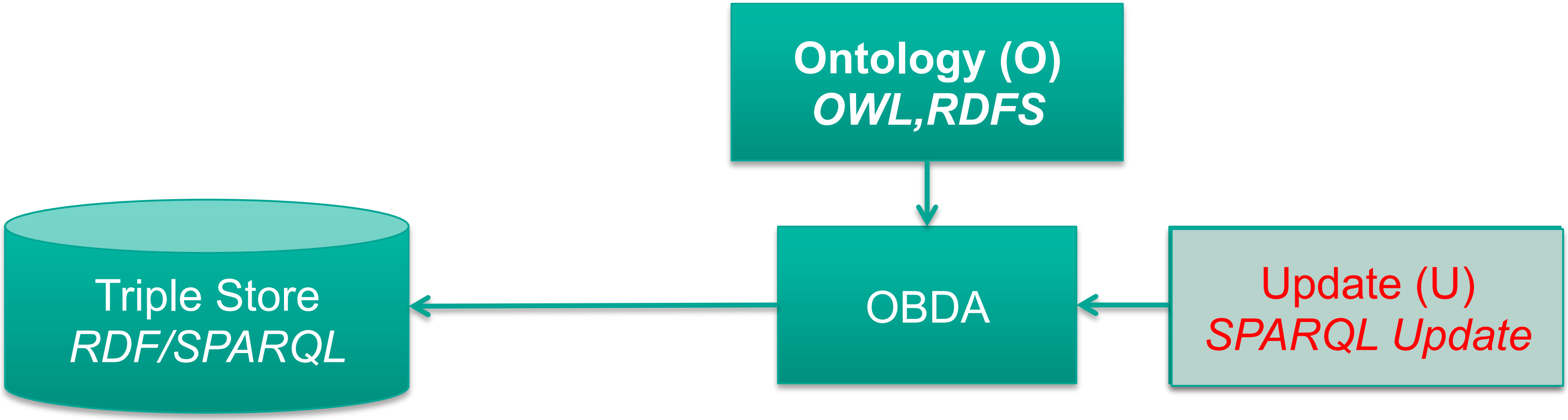


Adaptive Execution of SPARQL Queries

- Maribel Acosta, Maria-Esther Vidal, Tomas Lampo, Julio Castillo, Edna Ruckhaus: ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. International Semantic Web Conference (1) 2011: 18-34.
- Cosmin Basca, Abraham Bernstein: Querying a messy web of data with Avalanche. J. Web Sem. 26: 1-28 (2014)
- Maribel Acosta and Maria-Esther Vidal. Networks of Linked Data Eddies: An Adaptive Web Query Processing Engine for RDF Data. Accepted at ISWC 2015.

Updates in OBDA

- How to do updates in such a setting?



- So far, we only scratched the surface:

Albin Ahmeti, Diego Calvanese, and Axel Polleres. Updating RDFS ABoxes and TBoxes in SPARQL. In *Proceedings of the 13th International Semantic Web Conference (ISWC 2014)*, Lecture Notes in Computer Science (LNCS). Springer, October 2014.

Albin Ahmeti, Diego Calvanese, Vadim Savenkov, and Axel Polleres. Dealing with Inconsistencies due to Class Disjointness in SPARQL Update. In *28th International Workshop on Description Logics (DL2015)*, Athens, Greece, June 2015.

- For details, cf.:
- http://polleres.net/presentations/20150226SPARQL_Update_Entailment_Karlsruhe_Service_Summit.pptx

Your Research Task(s) for the rest of the day:

- Work on those in your mini-project groups!

Some of the overall Research Questions (**too generic on purpose!!!!**) from the slides before:

- *Quality:* Handling Data Quality Issues in (Linked) (Open) Data Integration Systems
- *Uncertainty:* Handling Uncertainty in (Linked) (Open) Data Integration Systems
- *Big Data:* Handling Scalable Processing of Rapidly growing data in (Linked) (Open) Data Integration Systems
- *LAV vs. GAV for the Semantic Web*: OBDA for SPARQL using LAV (SPARQL Query Execution using LAV)
- *Updates:* Handling Updates in OBDA

For each problem you work on:

1) **Problems**: Why is this a difficult problem? Find obstacles, find literature. Define concrete (sub-)research questions!

2) **Solutions**: What could be strategies to overcome these obstacles?

mandatory

3) **Systems**: What could be a strategy/roadmap/method to implement these strategies?

4) **Benchmarks**: What could be a strategy/roadmap/method to evaluate a solution?

optional

Result: **short** written report per group addressing these 4 questions and findings.

Tips:

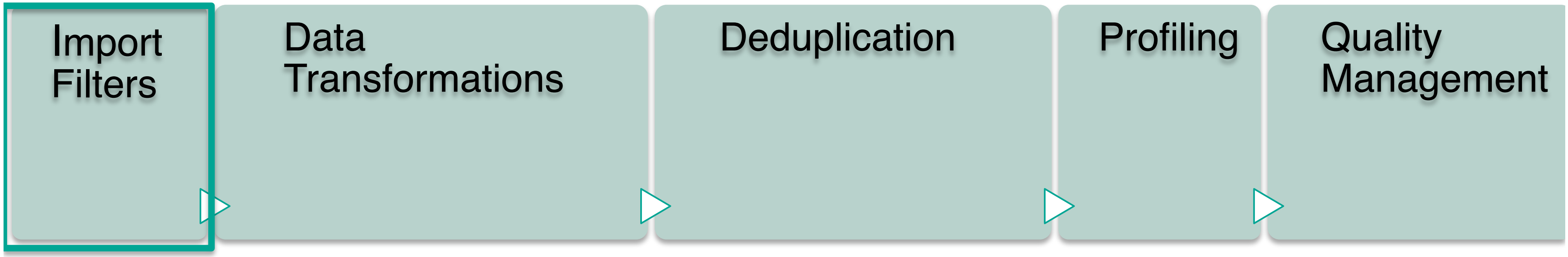
→ *Please email reports to axel[at]polleres.net*

- Think about how much time you dedicate to which of these four questions
- **Don't** start with 3)
- Prepare some answers or discussions for the final plenary session!

Backup slides

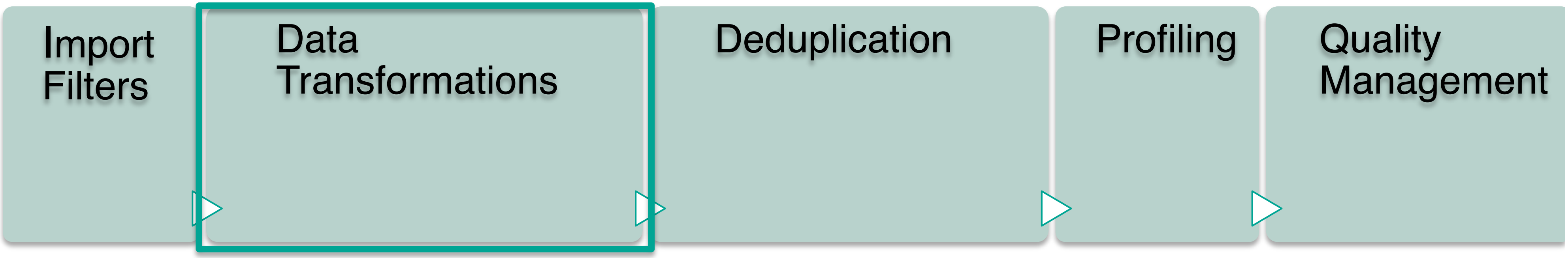
There are tons of things we did NOT talk about: e.g. stream data management, ontology evolution, federated query processing, etc. ...

Extraction-Transform-Load (ETL) Tools



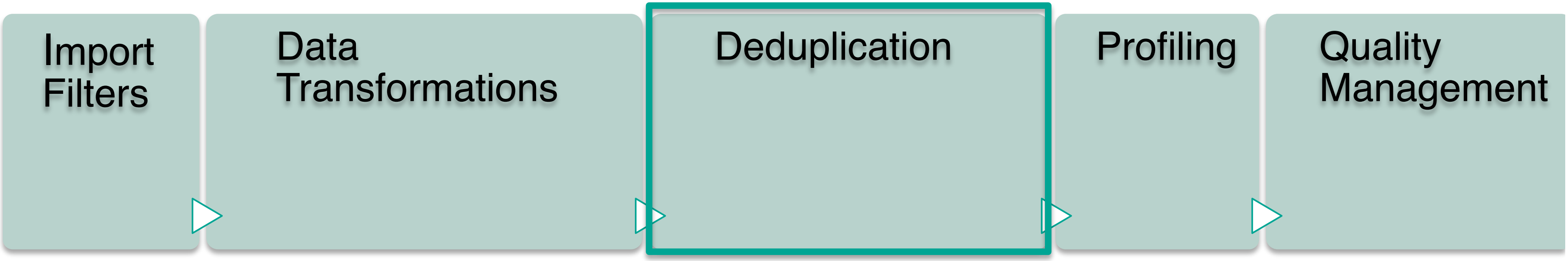
Parsers for external file formats, or drivers to interact with third-party systems.

Extraction-Transform-Load (ETL) Tools



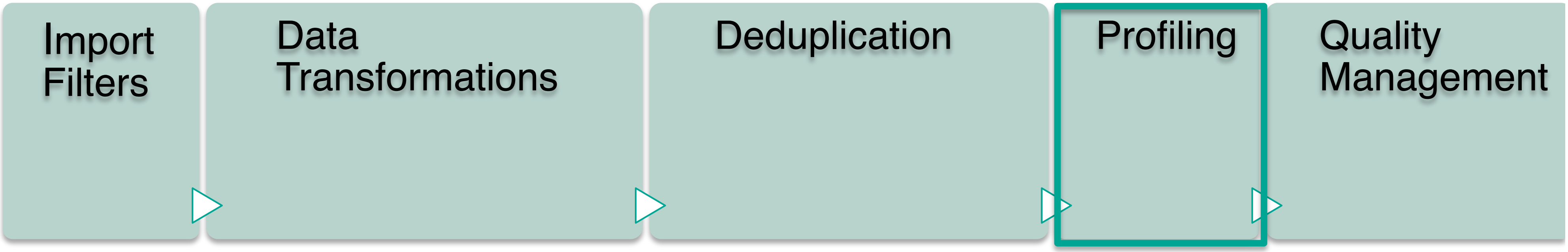
Schema mappings. Data may be joined, aggregated, and filtered.

Extraction-Transform-Load (ETL) Tools



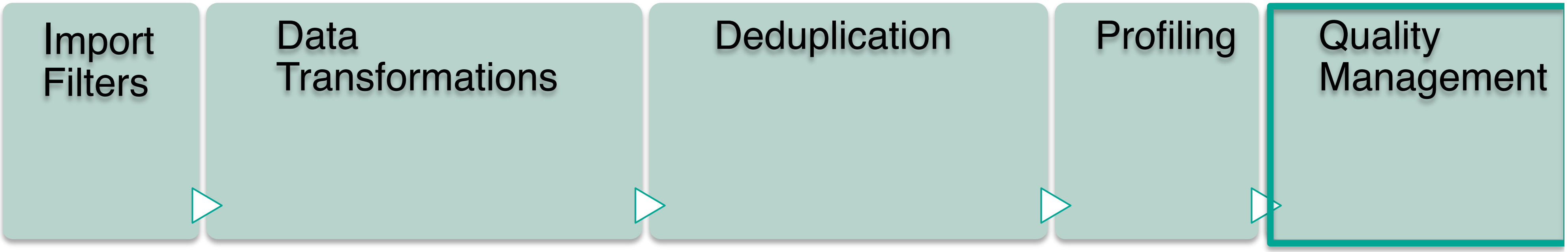
Tools that detect when multiple records refer to the same entity.

Extraction-Transform-Load (ETL) Tools



Tools that characterize and describe data in the data warehouse, e.g., histograms.

Extraction-Transform-Load (ETL) Tools



Tools that enhance data quality, e.g., testing against a master list, validating known business rules, record merging.

Wrappers for RDF Data

SPARQL Endpoints

Web services that implement the SPARQL protocol, and enable users to **query particular datasets**.

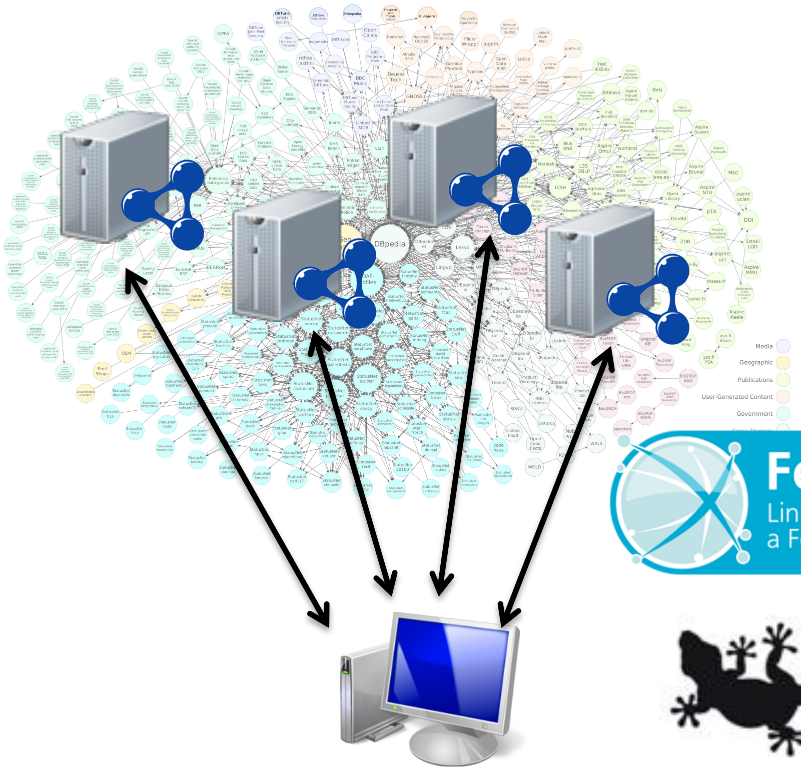
Linked Data Fragments[Verborgh2014]

Web services that access views of **triple patterns.**

[Verborgh2014]Ruben Verborgh, Olaf Hartig, Ben De Meester, Gerald Haesendonck, Laurens De Vocht, Miel Vander Sande, Richard Cyganiak, Pieter Colpaert, Erik Mannens, and Rik Van de Walle. Querying Datasets on the Web with High Availability. ISWC 2014.

Linked Data Mediators: Federated Query Processing

Publicly available SPARQL endpoints



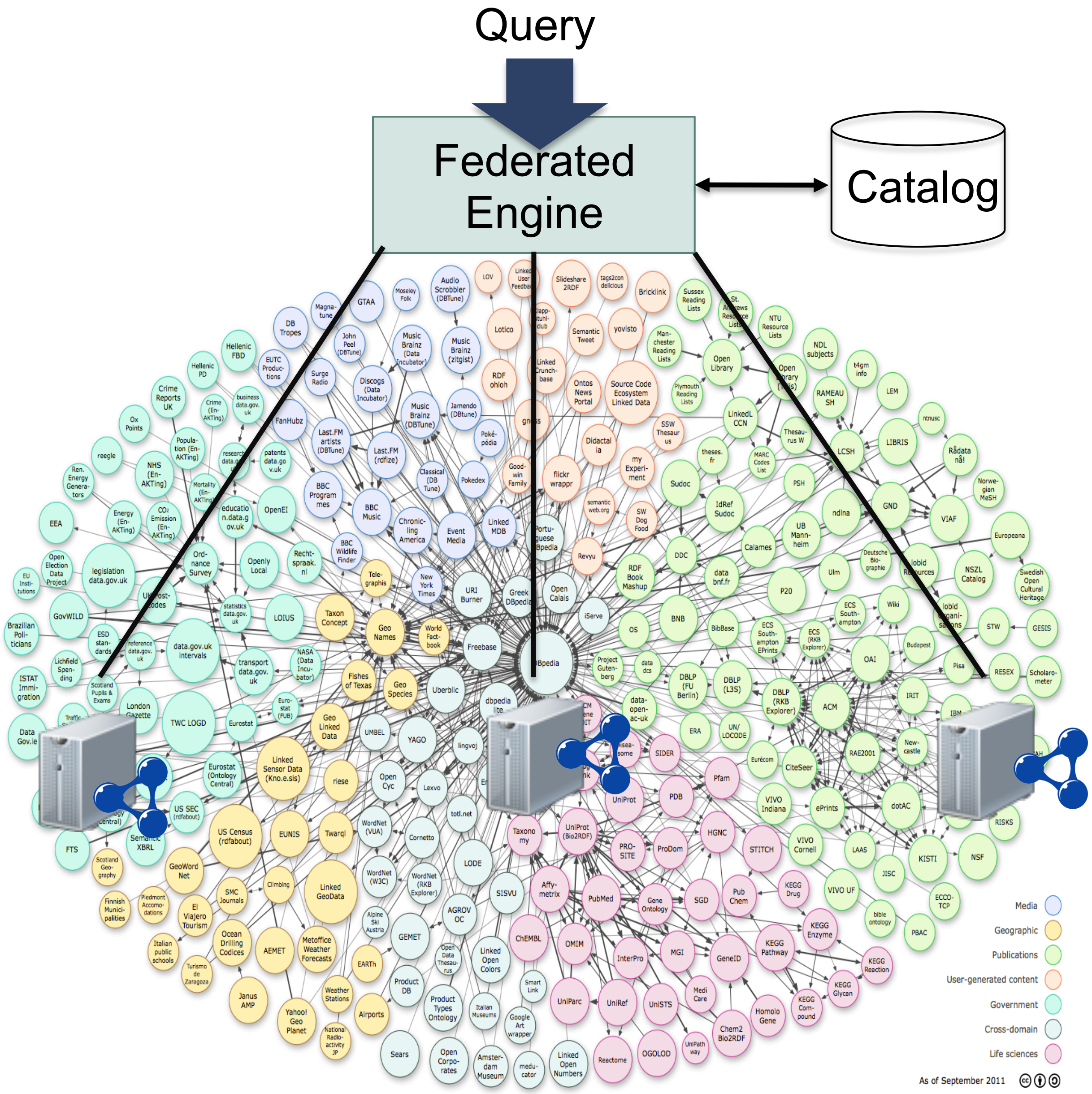
Federated
query processing

FedX
Linked Data in
a Federation

ANAPSID

SPLENDID

Federated Query Engine



Federation of SPARQL Endpoints

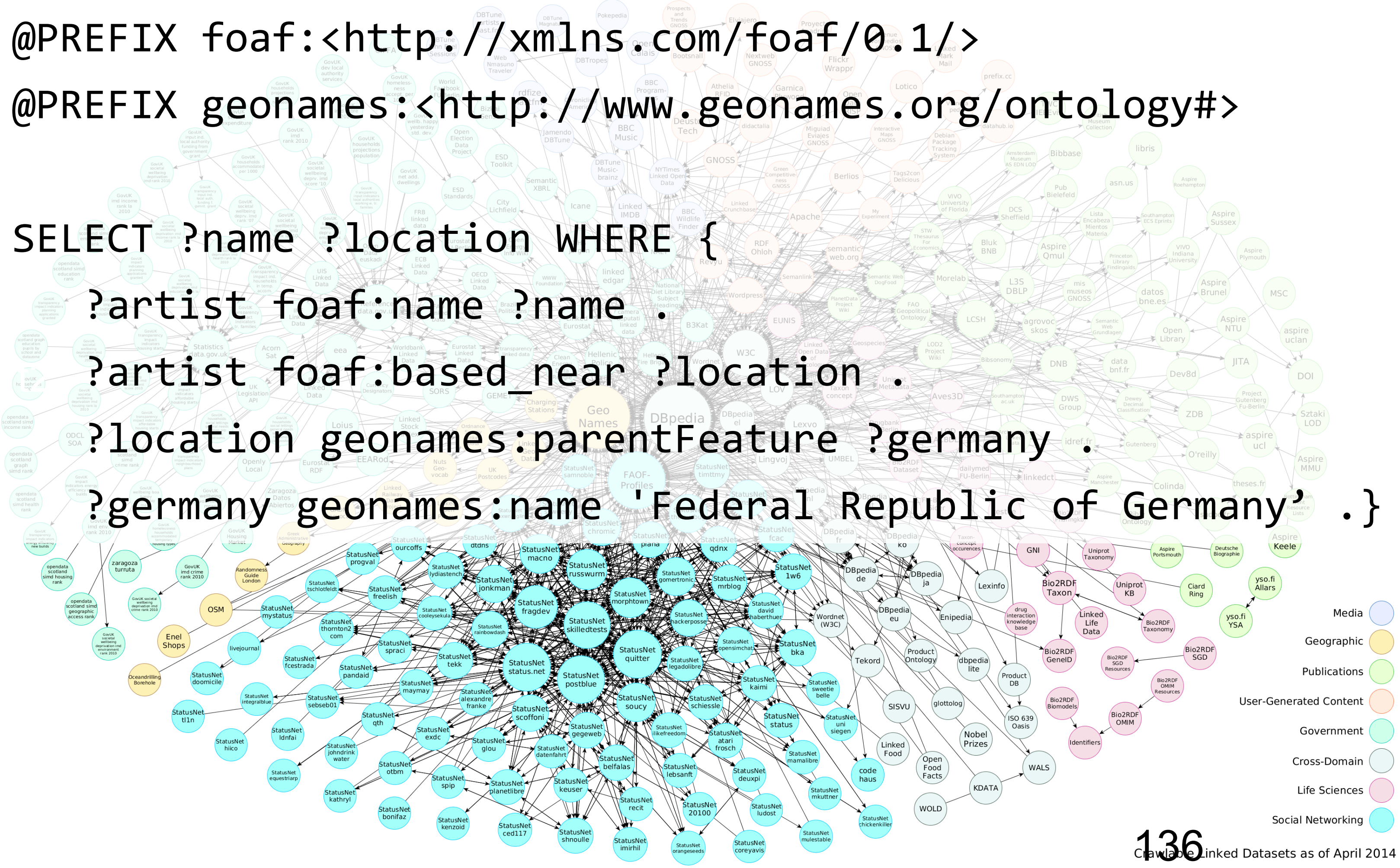
http://data.linkedmdb.org/sparql := http://data.linkedmdb.org/resource/movie/personal_film_appearance;
http://www.w3.org/2002/07/owl#sameAs;
http://www.w3.org/1999/02/22-rdf-syntax-ns#type;
http://xmlns.com/foaf/0.1/based_near;
http://xmlns.com/foaf/0.1/name;
…. .

http://dbtune.org/jamendo/sparql := http://www.w3.org/1999/02/22-rdf-syntax-ns#type;
http://purl.org/dc/elements/1.1/title;
http://xmlns.com/foaf/0.1/based_near;
http://xmlns.com/foaf/0.1/homepage;
http://purl.org/ontology/mo/biography;
….

http://dbpedia.org/sparql := http://xmlns.com/foaf/0.1/name;
http://dbpedia.org/ontology/award;
http://dbpedia.org/ontology/almaMater;
http://www.geonames.org/ontology#name;
http://www.geonames.org/ontology#parentFeatures;
….

http://www.lotico.com:3030/lotico/sparq := http://www.geonames.org/ontology#name;
http://www.geonames.org/ontology#parentFeatures;
http://www.geonames.org/ontology#officialName;
http://www.geonames.org/ontology#postalCode;
…. .

Executing a Federated Query

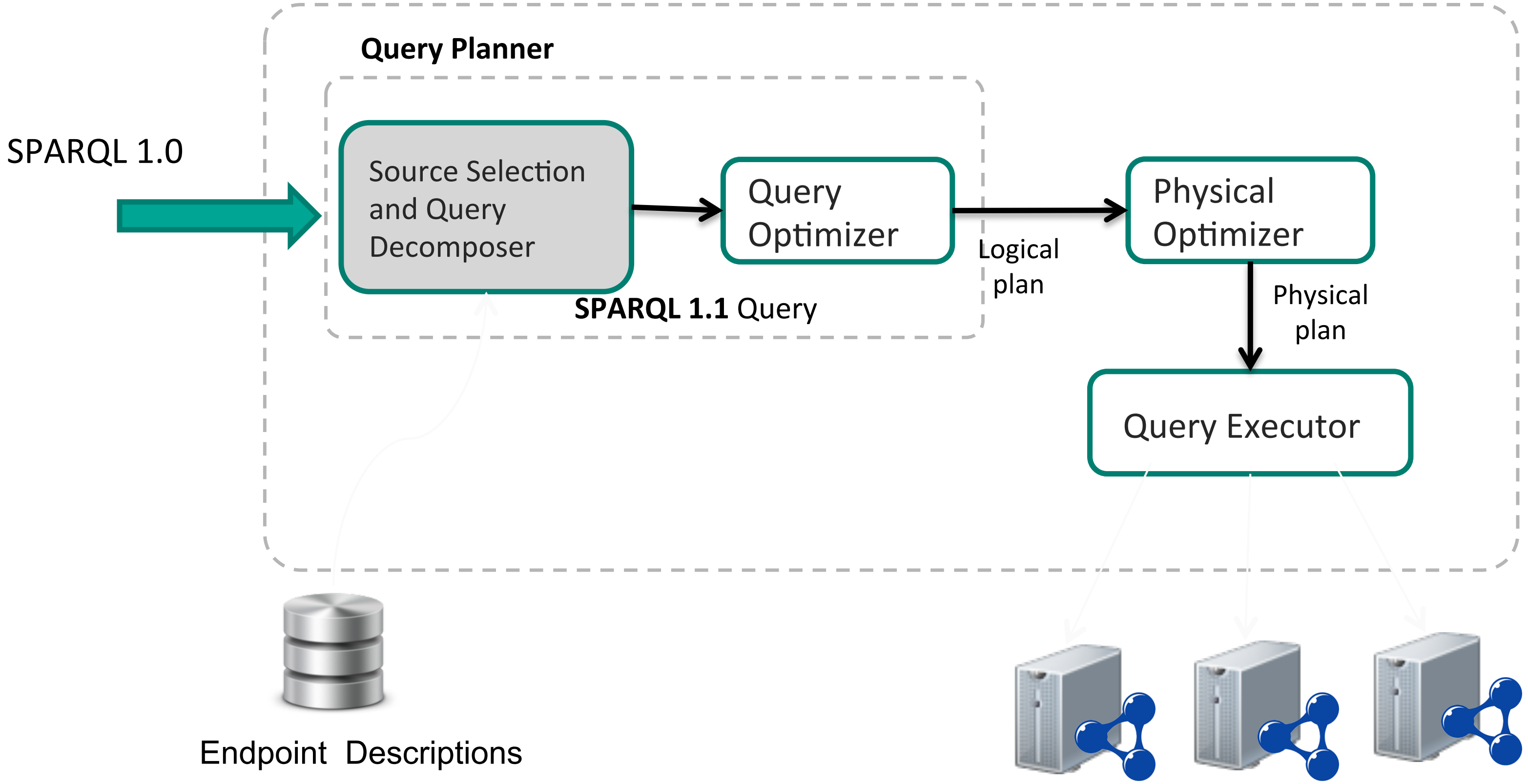


```
@PREFIX foaf:<http://xmlns.com/foaf/0.1/>
@PREFIX geonames:<http://www.geonames.org/ontology#>

SELECT ?name ?location WHERE {
   ?artist foaf:name ?name .
   ?artist foaf:based_near ?location .
   ?location geonames:parentFeature ?germany .
   ?germany geonames:name 'Federal Republic of Germany’ .}
```

Crawlable Linked Datasets as of April 2014

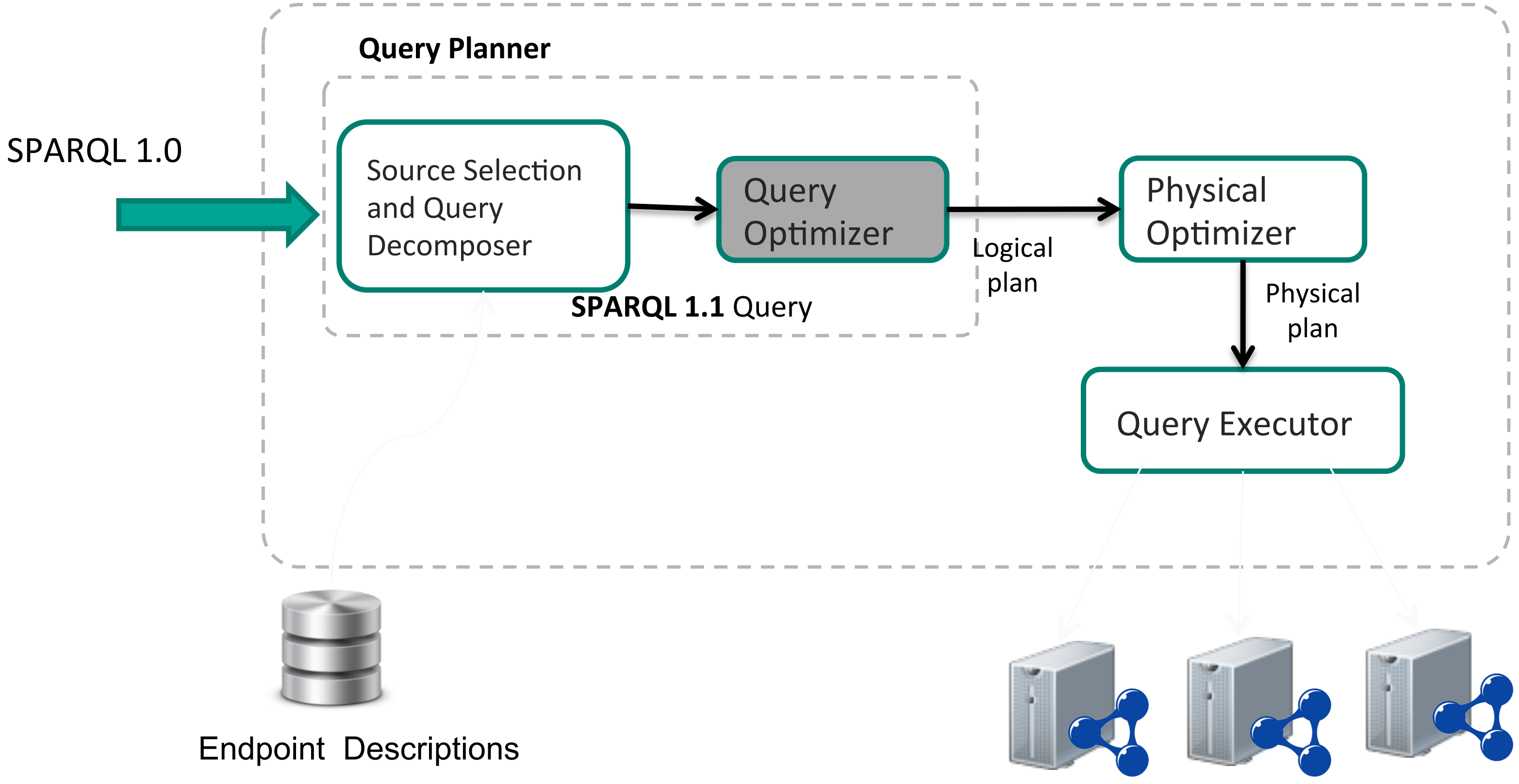
Federated Engines: Architecture



Federated Query SPARQL 1.1

```
@PREFIX foaf:<http://xmlns.com/foaf/0.1/>
@PREFIX geonames:http://www.geonames.org/ontology#
SELECT ?name ?location WHERE {
  {  SERVICE <http://data.linkedmdb.org/sparql> {
t1 ?artist foaf:name ?name . } }.
   { SERVICE <http://dbtune.org/jamendo/sparql>{
t2 ?artist foaf:based_near ?location .} }.
   { SERVICE <http://dbpedia.org/sparql> {
t3 ?location geonames:parentFeature ?germany } }.
   { SERVICE <http://www.lotico.com:3030/lotico/sparq> {
t4 ?germany geonames:name 'Federal Republic of Germany’ } }
}
```

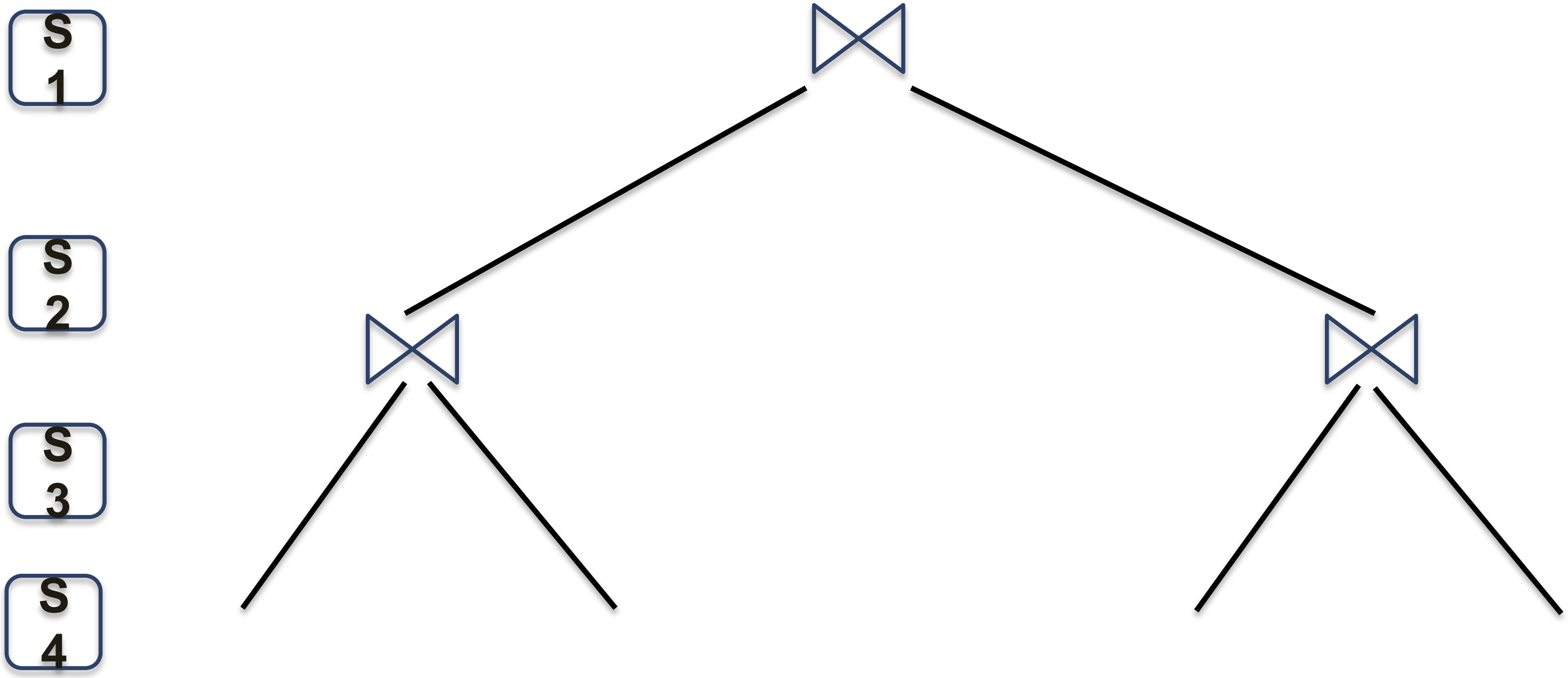
Federated Engines: Architecture



Executing a Federated Query
SPARQL 1.1

```
@PREFIX foaf:<http://xmlns.com/foaf/0.1/>
@PREFIX geonames:http://www.geonames.org/ontology#
SELECT ?name ?location WHERE {
    {  SERVICE <http://data.linkedmdb.org/sparql> {                S1
t1 ?artist foaf:name ?name . } }.
     { SERVICE <http://dbtune.org/jamendo/sparql>{                 S2
t2 ?artist foaf:based_near ?location .} }.
     { SERVICE <http://dbpedia.org/sparql> {                       S3
t3 ?location geonames:parentFeature ?germany } }.
     { SERVICE <http://www.lotico.com:3030/lotico/sparq> {              S4
t4 ?germany geonames:name 'Federal Republic of Germany’ } }
}
```

Query Rewriting Plan



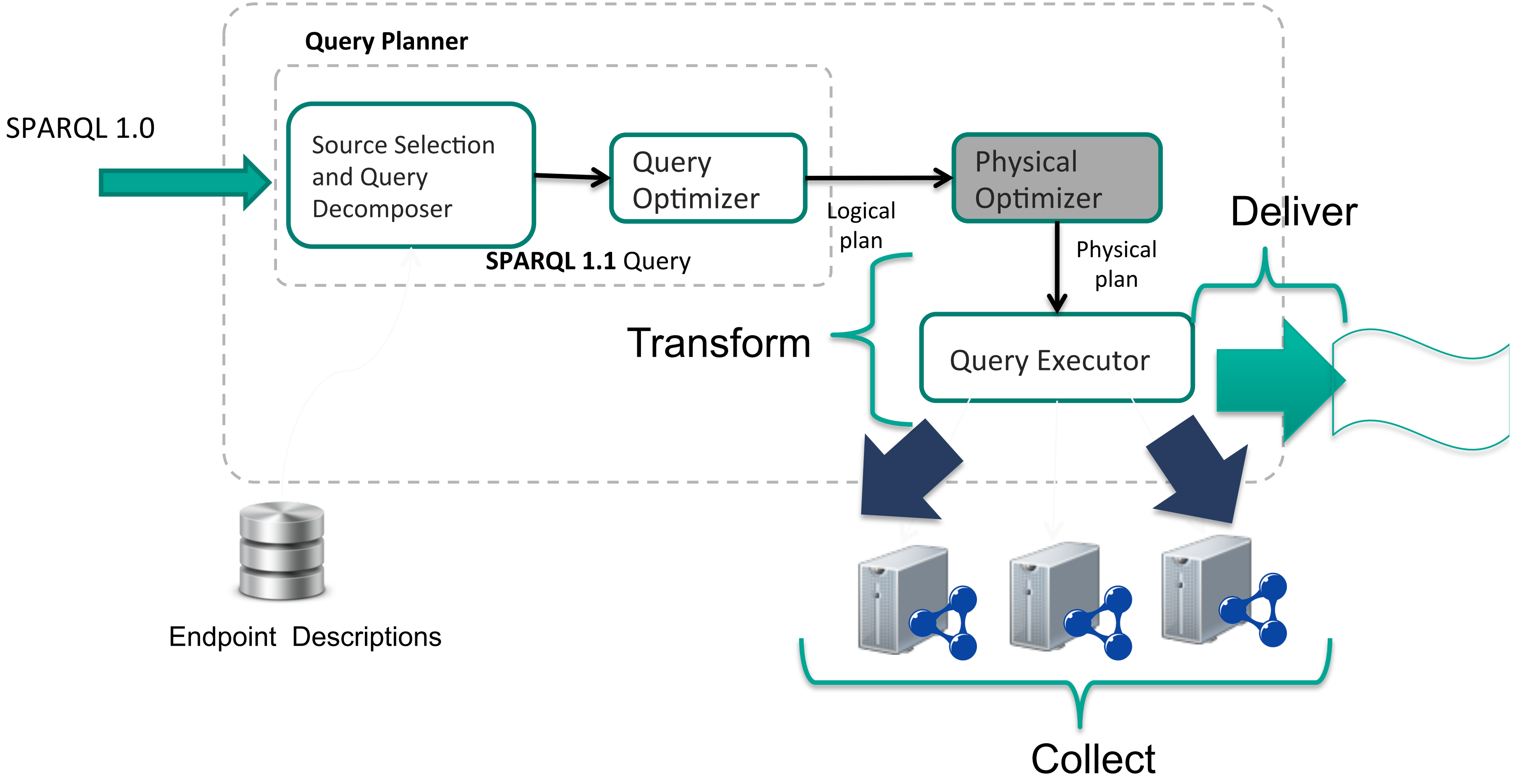
S1

S2

S3

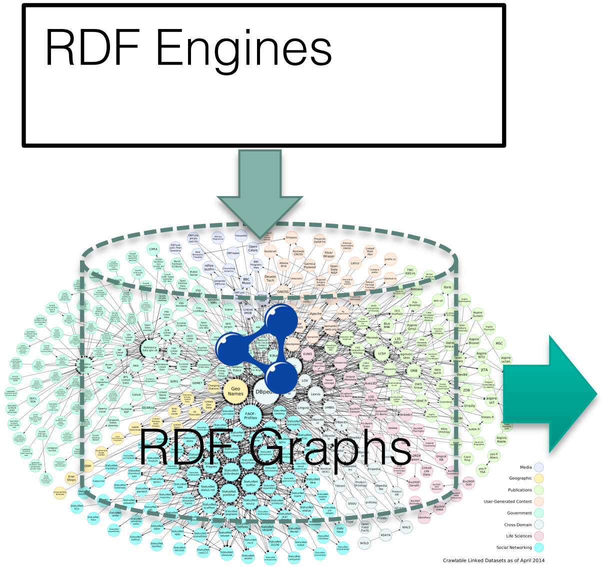
S4

Federated Engines: Architecture



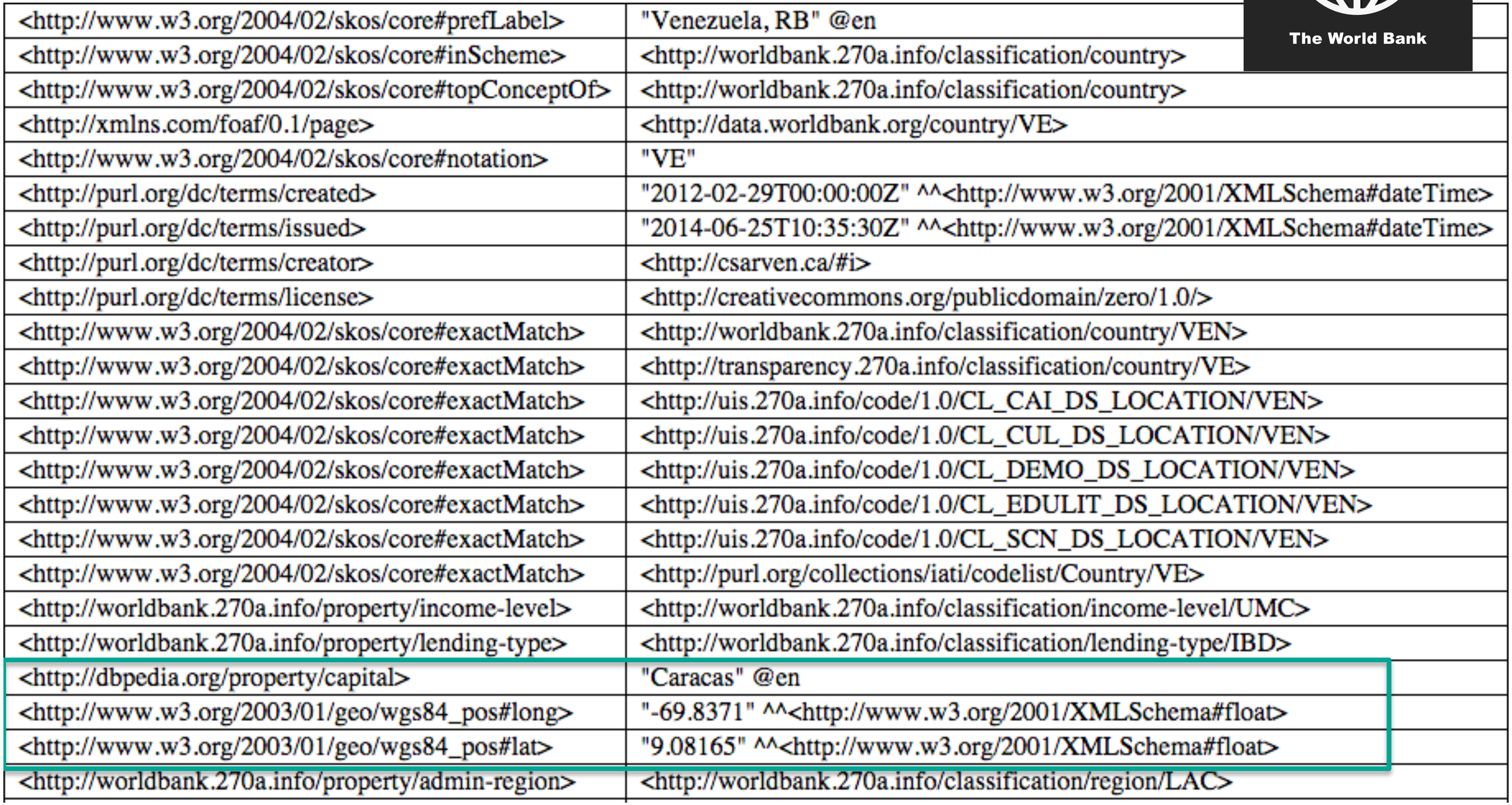
SPARQL Query Processing

```
@PREFIX foaf:<http://xmlns.com/foaf/0.1/>
@PREFIX geonames:http://www.geonames.org/ontology#
SELECT ?name ?location WHERE {
   ?artist foaf:name ?name .
   ?artist foaf:based_near ?location .
   ?location geonames:parentFeature ?germany .
   ?germany geonames:name 'Federal Republic of Germany’ .}
```



```
{'news': '', 'name': 'Michael Bartels^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Melophon^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Remote Controlled^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Arne Pahlke^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Superdefekt^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Chaos^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'The Gay Romeos^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'Der tollw\\u00FCtige Kasper^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'the ad.kowas^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'herr gau^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
{'news': '', 'name': 'The Rodeo Five^^<http://www.w3.org/2001/XMLSchema#string>', 'location': 'http://sws.geonames.org/2911297/'}
```

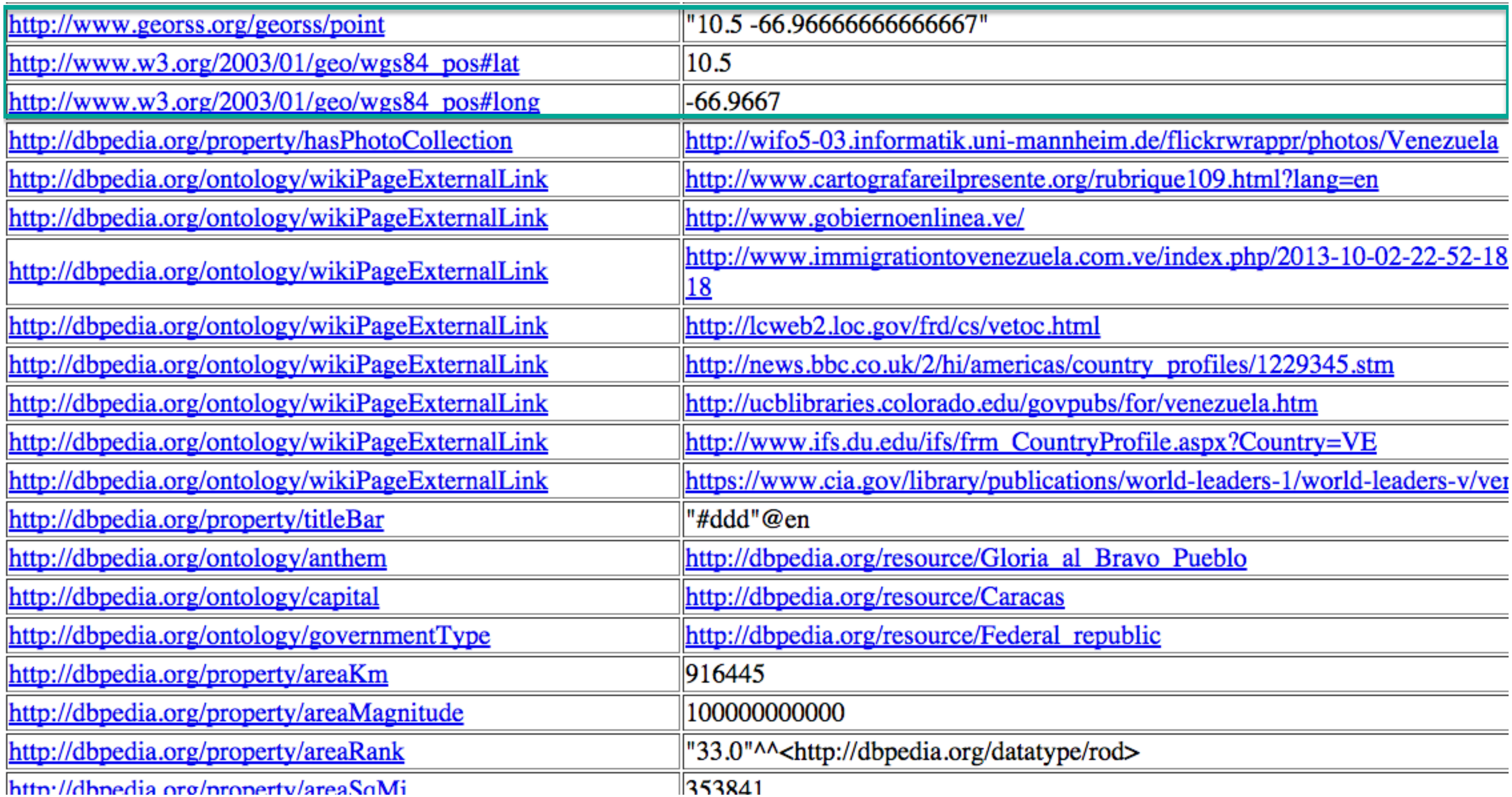
Data Quality-Duplicated Resources



<http://www.w3.org/2004/02/skos/core#prefLabel>	"Venezuela, RB" @en
<http://www.w3.org/2004/02/skos/core#inScheme>	<http://worldbank.270a.info/classification/country>
<http://www.w3.org/2004/02/skos/core#topConceptOf>	<http://worldbank.270a.info/classification/country>
<http://xmlns.com/foaf/0.1/page>	<http://data.worldbank.org/country/VE>
<http://www.w3.org/2004/02/skos/core#notation>	"VE"
<http://purl.org/dc/terms/created>	"2012-02-29T00:00:00Z" ^^<http://www.w3.org/2001/XMLSchema#dateTime>
<http://purl.org/dc/terms/issued>	"2014-06-25T10:35:30Z" ^^<http://www.w3.org/2001/XMLSchema#dateTime>
<http://purl.org/dc/terms/creator>	<http://csarven.ca/#i>
<http://purl.org/dc/terms/license>	<http://creativecommons.org/publicdomain/zero/1.0/>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://worldbank.270a.info/classification/country/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://transparency.270a.info/classification/country/VE>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://uis.270a.info/code/1.0/CL_CAI_DS_LOCATION/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://uis.270a.info/code/1.0/CL_CUL_DS_LOCATION/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://uis.270a.info/code/1.0/CL_DEMO_DS_LOCATION/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://uis.270a.info/code/1.0/CL_EDULIT_DS_LOCATION/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://uis.270a.info/code/1.0/CL_SCN_DS_LOCATION/VEN>
<http://www.w3.org/2004/02/skos/core#exactMatch>	<http://purl.org/collections/iati/codelist/Country/VE>
<http://worldbank.270a.info/property/income-level>	<http://worldbank.270a.info/classification/income-level/UMC>
<http://worldbank.270a.info/property/lending-type>	<http://worldbank.270a.info/classification/lending-type/IBD>
<http://dbpedia.org/property/capital>	"Caracas" @en
<http://www.w3.org/2003/01/geo/wgs84_pos#long>	"-69.8371" ^^<http://www.w3.org/2001/XMLSchema#float>
<http://www.w3.org/2003/01/geo/wgs84_pos#lat>	"9.08165" ^^<http://www.w3.org/2001/XMLSchema#float>
<http://worldbank.270a.info/property/admin-region>	<http://worldbank.270a.info/classification/region/LAC>

<http://worldbank.270a.info/classification/country/VE>

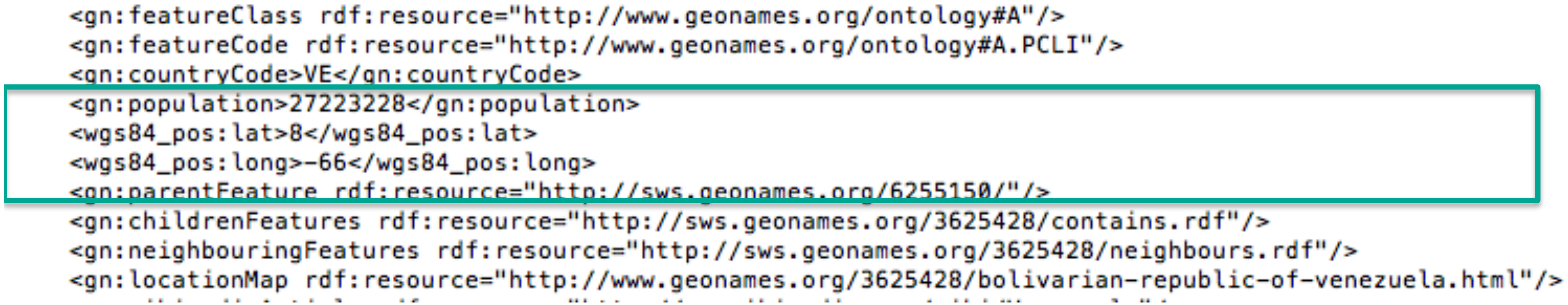
Data Quality-Duplicated Resources



http://www.georss.org/georss/point	"10.5 -66.96666666666667"
http://www.w3.org/2003/01/geo/wgs84_pos#lat	10.5
http://www.w3.org/2003/01/geo/wgs84_pos#long	-66.9667
http://dbpedia.org/property/hasPhotoCollection	http://wifo5-03.informatik.uni-mannheim.de/flickrwrappr/photos/Venezuela
http://dbpedia.org/ontology/wikiPageExternalLink	http://www.cartografareilpresente.org/rubrique109.html?lang=en
http://dbpedia.org/ontology/wikiPageExternalLink	http://www.gobiernoenlinea.ve/
http://dbpedia.org/ontology/wikiPageExternalLink	http://www.immigrationtovenezuela.com.ve/index.php/2013-10-02-22-52-18
http://dbpedia.org/ontology/wikiPageExternalLink	http://lcweb2.loc.gov/frd/cs/vetoc.html
http://dbpedia.org/ontology/wikiPageExternalLink	http://news.bbc.co.uk/2/hi/americas/country_profiles/1229345.stm
http://dbpedia.org/ontology/wikiPageExternalLink	http://ucblibraries.colorado.edu/govpubs/for/venezuela.htm
http://dbpedia.org/ontology/wikiPageExternalLink	http://www.ifs.du.edu/ifs/frm_CountryProfile.aspx?Country=VE
http://dbpedia.org/ontology/wikiPageExternalLink	https://www.cia.gov/library/publications/world-leaders-1/world-leaders-v/ver
http://dbpedia.org/property/titleBar	"#ddd"@en
http://dbpedia.org/ontology/anthem	http://dbpedia.org/resource/Gloria_al_Bravo_Pueblo
http://dbpedia.org/ontology/capital	http://dbpedia.org/resource/Caracas
http://dbpedia.org/ontology/governmentType	http://dbpedia.org/resource/Federal_republic
http://dbpedia.org/property/areaKm	916445
http://dbpedia.org/property/areaMagnitude	100000000000
http://dbpedia.org/property/areaRank	"33.0"^^<http://dbpedia.org/datatype/rod>
http://dbpedia.org/property/areaSqMi	353841

<http://dbpedia.org/resource/Venezuela>

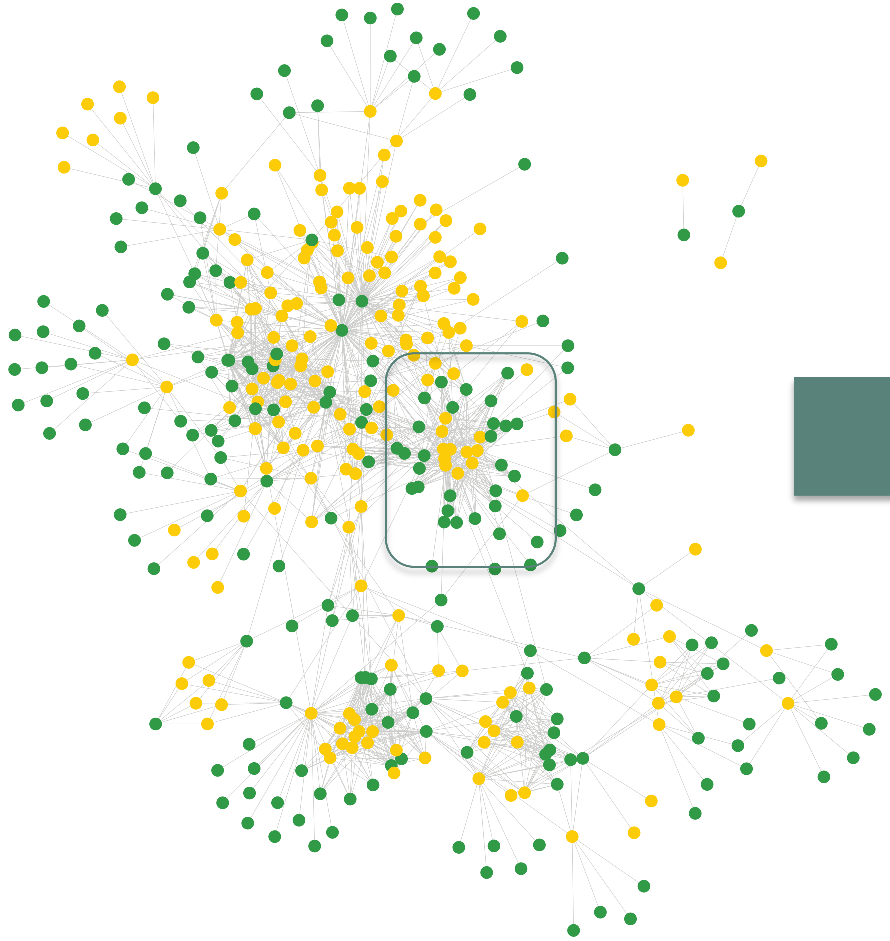
Data Quality-Duplicated Resources

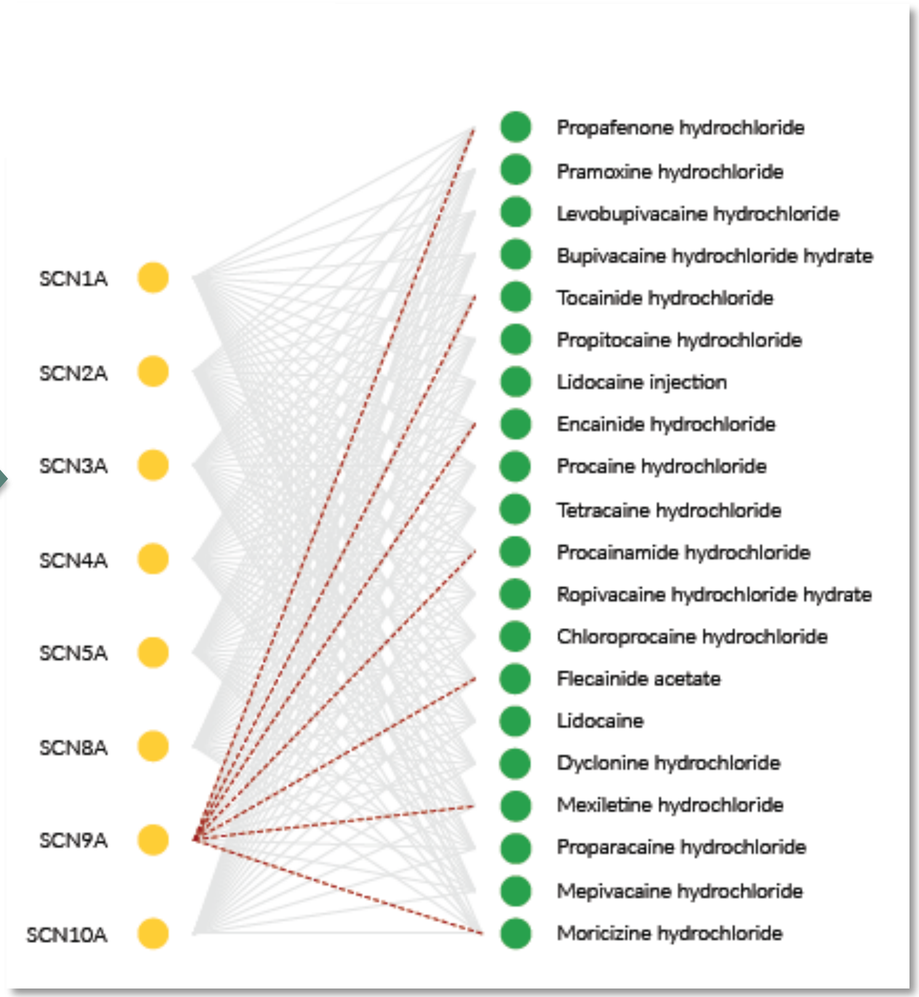


```
<gn:featureClass rdf:resource="http://www.geonames.org/ontology#A"/>
<gn:featureCode rdf:resource="http://www.geonames.org/ontology#A.PCLI"/>
<gn:countryCode>VE</gn:countryCode>
<gn:population>27223228</gn:population>
<wgs84_pos:lat>8</wgs84_pos:lat>
<wgs84_pos:long>-66</wgs84_pos:long>
<gn:parentFeature rdf:resource="http://sws.geonames.org/6255150/"/>
<gn:childrenFeatures rdf:resource="http://sws.geonames.org/3625428/contains.rdf"/>
<gn:neighbouringFeatures rdf:resource="http://sws.geonames.org/3625428/neighbours.rdf"/>
<gn:locationMap rdf:resource="http://www.geonames.org/3625428/bolivarian-republic-of-venezuela.html"/>
```

<http://sws.geonames.org/3625428/>

Mining Techniques to Predict Links and Discover Patterns



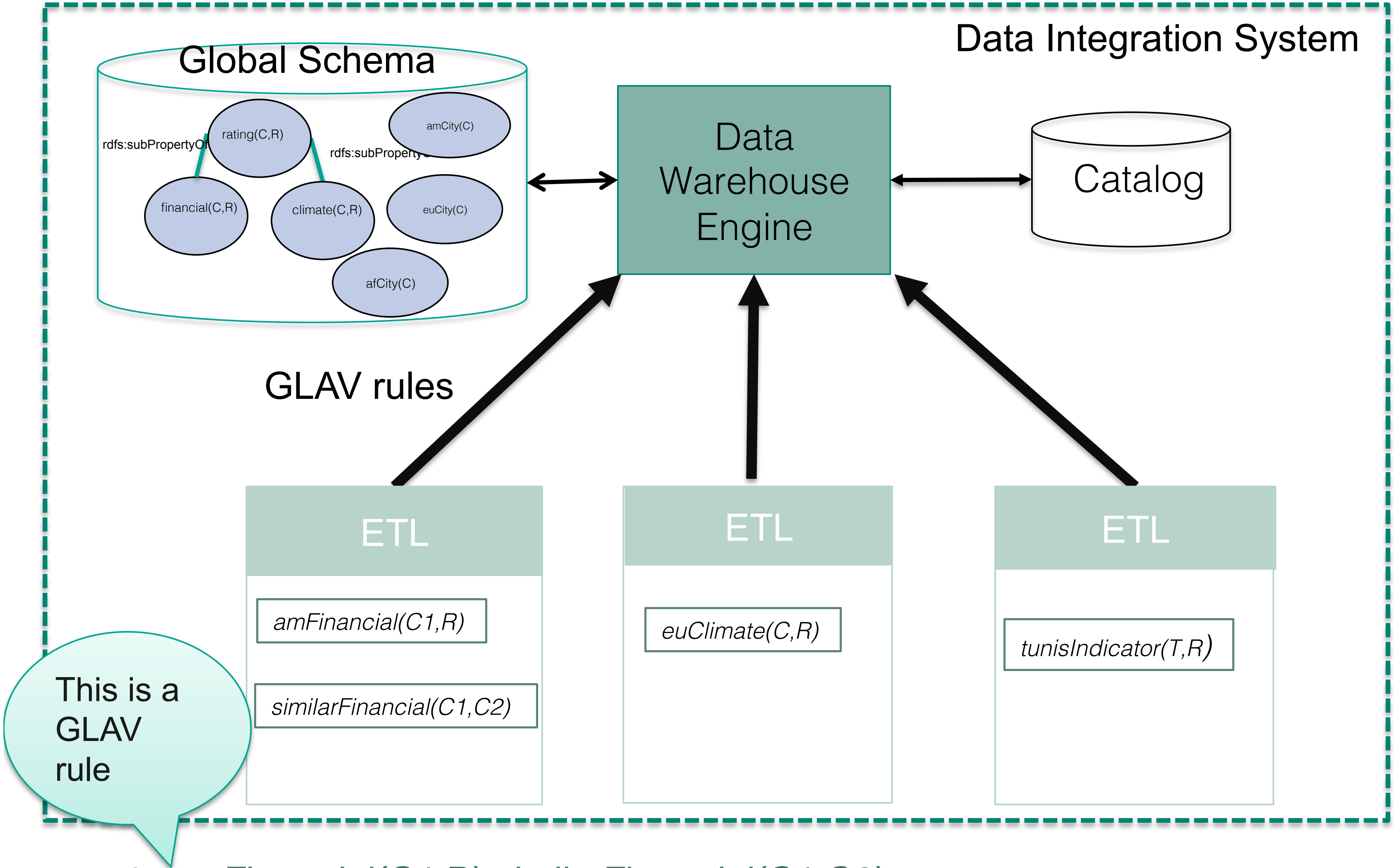


Mining Techniques to Predict Links and Discover Patterns

- Ding, H., Takigawa, I., Mamitsuka, H., Zhu, S.: Similarity-based machine learning methods for predicting drug-target interactions: A brief review. Briefings in Bioinformatics (2013)
- Fakhraei, S., Huang, B., Raschid, L., Getoor, L.: Network-based drug-target interaction prediction with probabilistic soft logic. In: IEEE/ACM Transactions on Computational Biology and Bioinformatics (2014)
- Flores A., Vidal M.E., Palma G.: Exploiting Semantics to Predict Potential Novel Links from Dense Subgraphs. AMW 2015
- Palma G., Vidal M.E., Raschid L.: Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. Semantic Web Conference (1) 2014: 131-146

MATERIALIZED GLOBAL SCHEMA-DATA WAREHOUSE

Data Warehouse-Materialized Global Schema



Data Integration System

Global Schema

rdfs:subPropertyOf

rating(C,R)

rdfs:subPropert

amCity(C)

financial(C,R)

climate(C,R)

euCity(C)

afCity(C)

Data Warehouse Engine

Catalog

GLAV rules

ETL

amFinancial(C1,R)

similarFinancial(C1,C2)

ETL

euClimate(C,R)

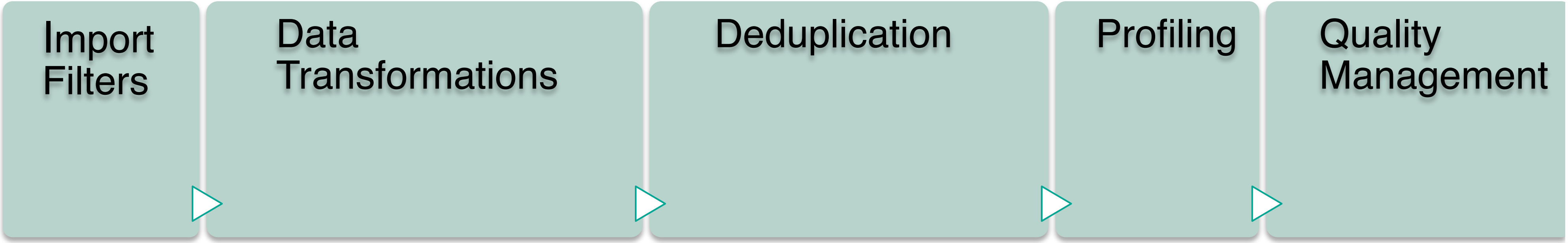
ETL

tunisIndicator(T,R)

This is a GLAV rule

α0: *amFinancial(C1,R),similarFinancial(C1,C2):-*
amCity(C1),amCity(C2),financial(C1,R),financial(C2,R).

Extraction-Transform-Load (ETL) Tools



Import Filters
Data Transformations
Deduplication
Profiling
Quality Management