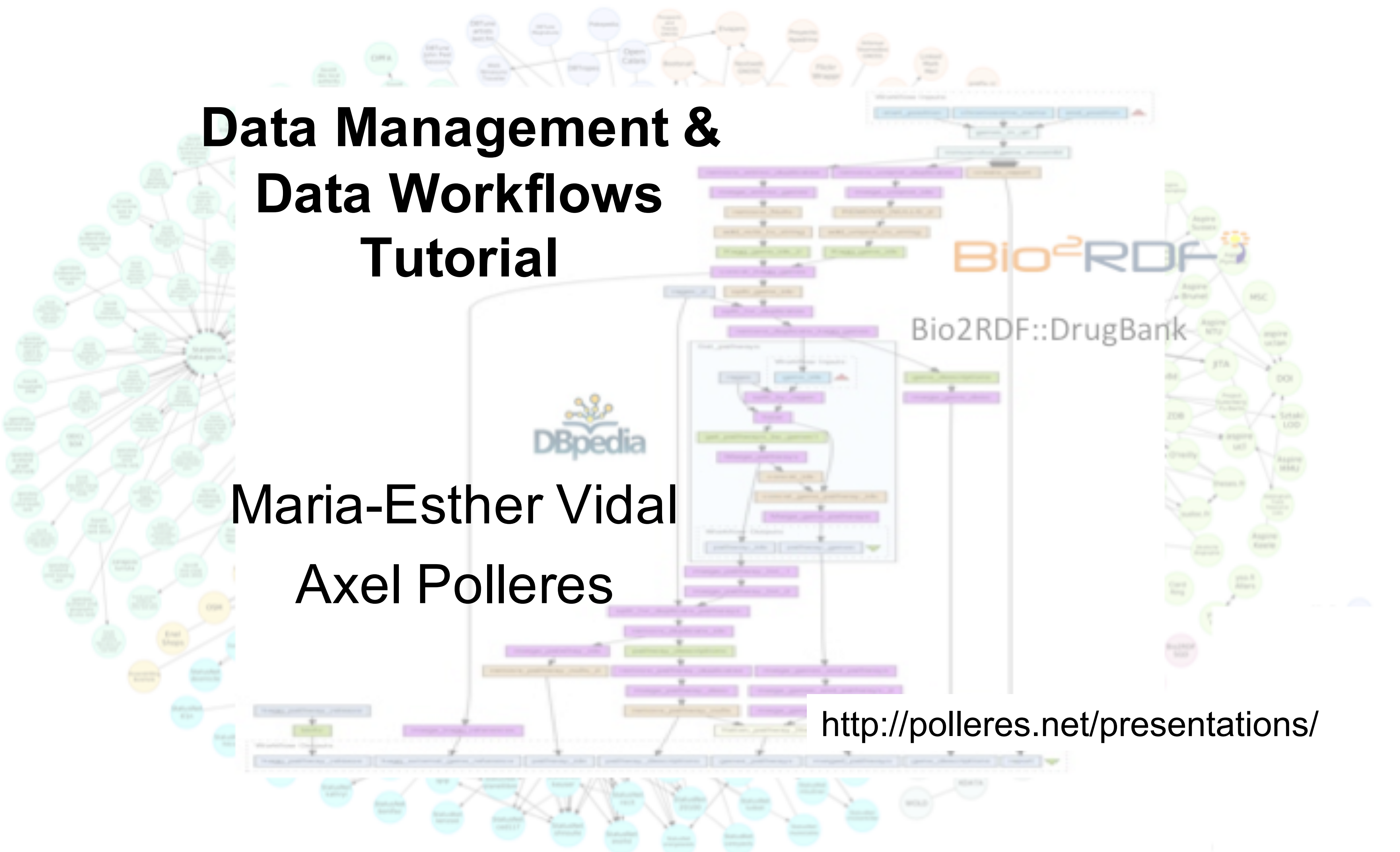
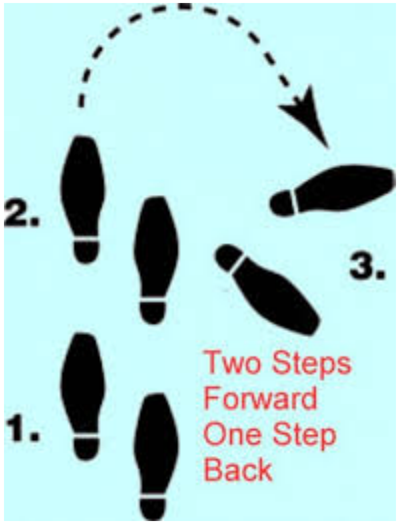
Data Management & Data Workflows Tutorial



Maria-Esther Vidal

Axel Polleres

http://polleres.net/presentations/



Data Management (today)
vs.
Knowledge discovery/modeling (yesterday)

Outline

- Motivation
  - Integrating (Open) Data from different sources
  - Not only Linked Data (NoLD)
  - Data workflows and Data Management 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
  - GAV vs. LAV
  - The Mediator and Wrapper Architecture
  - Query rewriting vs. Materialisation
  - Data Integration using Ontologies
- Challenges:
  - How to find Rules and ontologies?
  - Handling Incompleteness
  - How to find the data?
- 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.:

DATA.GOV.UK Beta Opening up Government

DATA.GOV

london.gov.uk

UNdata A world of information

THE WORLD BANK Open Data

DBpedia

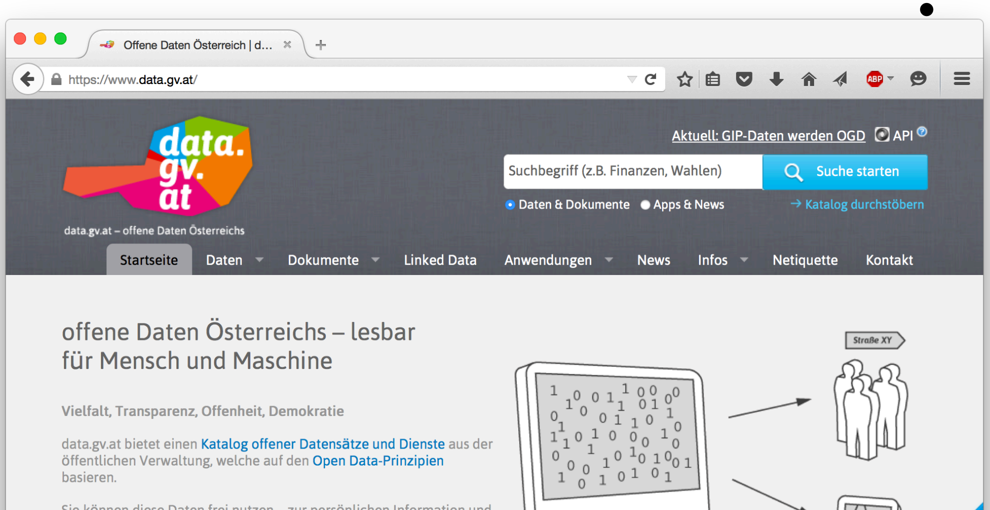
wien.at

NYC OpenData

European Union Open Data Portal

data.gv.at

odp BETA All you can Data OPEN DATA PORTAL ÖSTERREICH



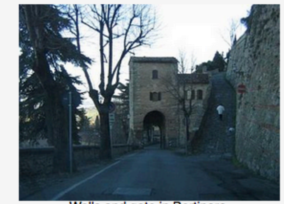
- In general: more and more structured data available at our fingertipps
- It's on the Web
- It's open
  → no restrictions w.r.t. re-use

This image cannot currently be displayed.

Bertinoro

Comune

Comune di Bertinoro



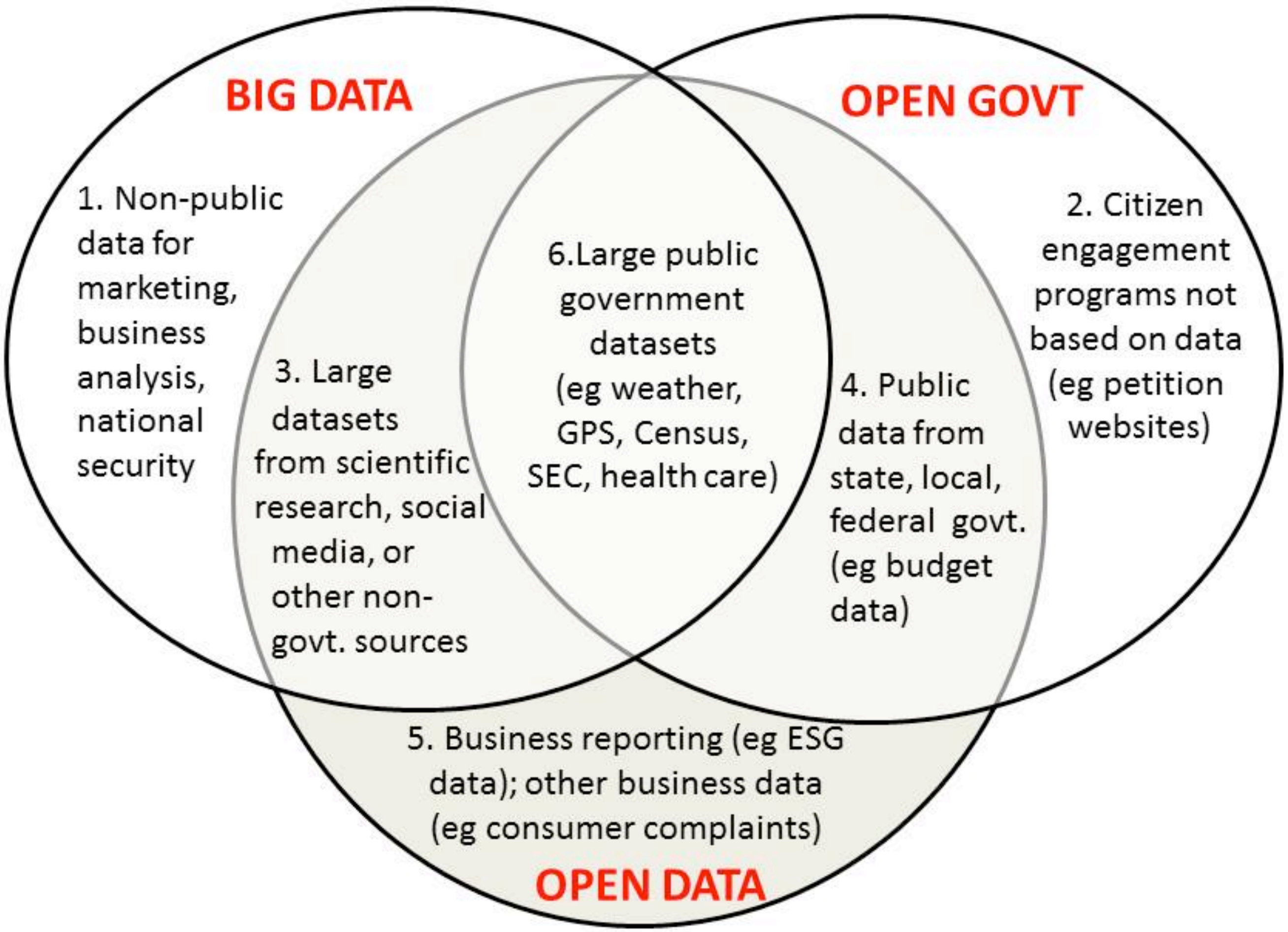
Walls and gate in Bertinoro.

Location of Bertinoro in Italy

Coordinates: 44°09′N 12°08′E

Country	Italy
Region	Emilia-Romagna
Province / Metropolitan city	Forlì-Cesena (FC)
Frazioni	Bracciano, Capocolle, Collinello, Fratta Terme, Ospedaletto, Panighina, Polenta, San Pietro in Guardiano, Santa Croce, Santa Maria Nuova Spallicci
Government	
• Mayor	Nevio Zaccarelli
Area	
• Total	56 km² (22 sq mi)
Elevation	220 m (720 ft)
Population (31 March 2008)	
• Total	10,353
• Density	180/km² (480/sq mi)
Demonym(s)	Bertinoresi
Time zone	CET (UTC+1)
• Summer (DST)	CEST (UTC+2)
Postal code	47032
Dialing code	0543
Patron saint	St. Catherine of Alexandria
Saint day	November 25

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



- 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

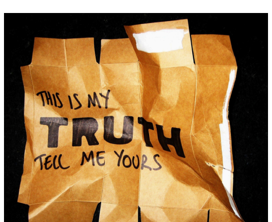
- **Volume:**
  - It's growing! (we currently monitor 90 CKAN portals, 512543 resources/ 160069 datasets,

  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



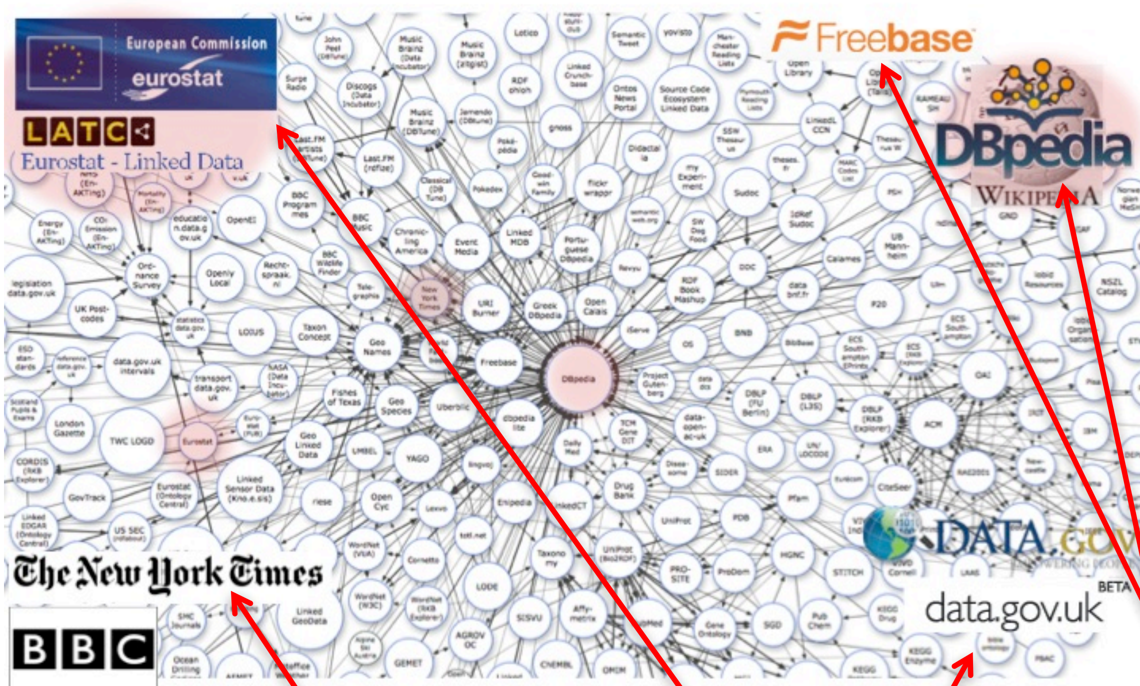
- **Veracity:**
  - quality, trust

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

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

… more specifically on

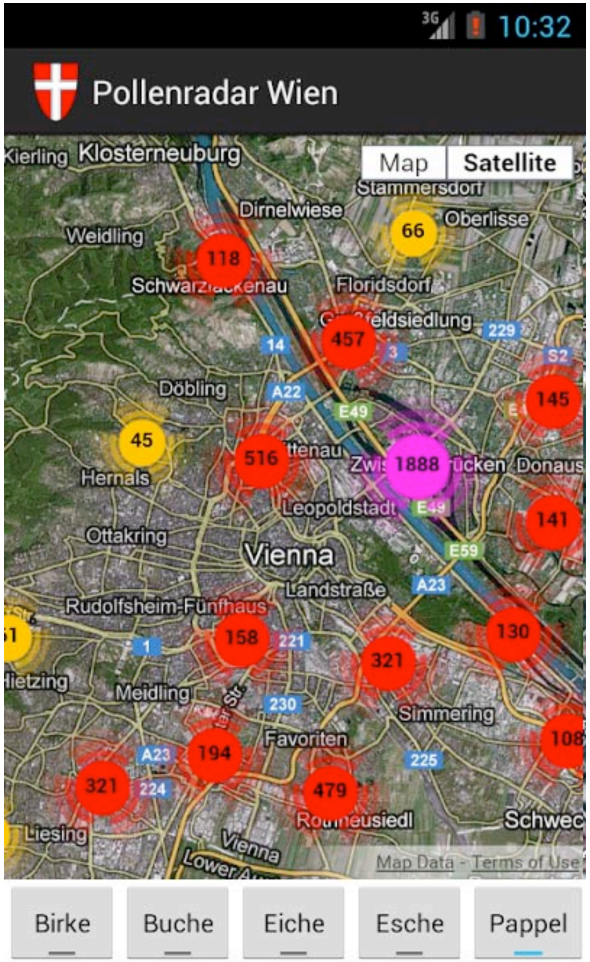
Open **Structured** Data

Alternatives in the meantime: (wikidata...)

LD efforts discontinued?!

LOD in OGD growing, but slowly

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

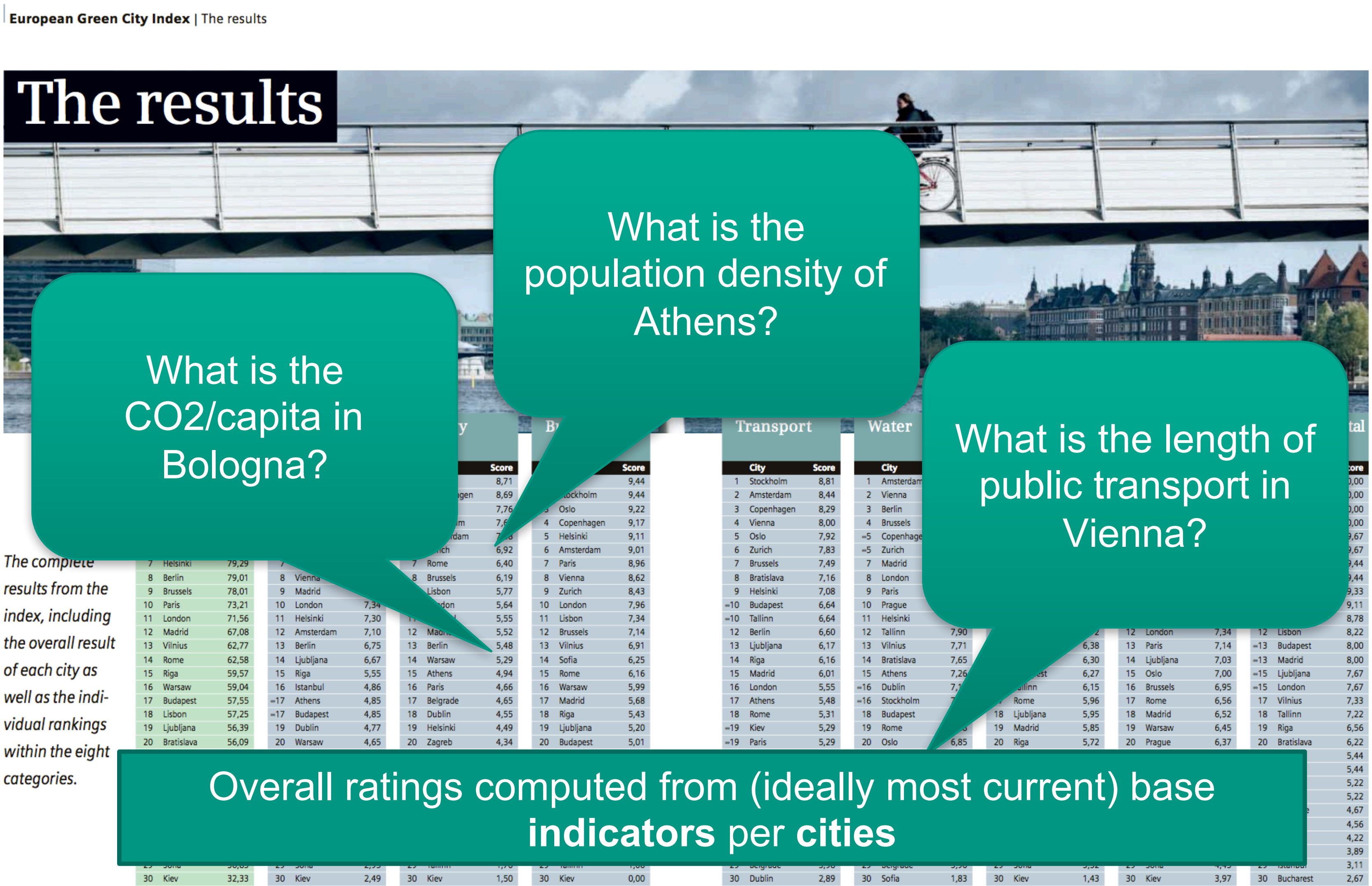


Great stuff, but limited potential...

More interesting:

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

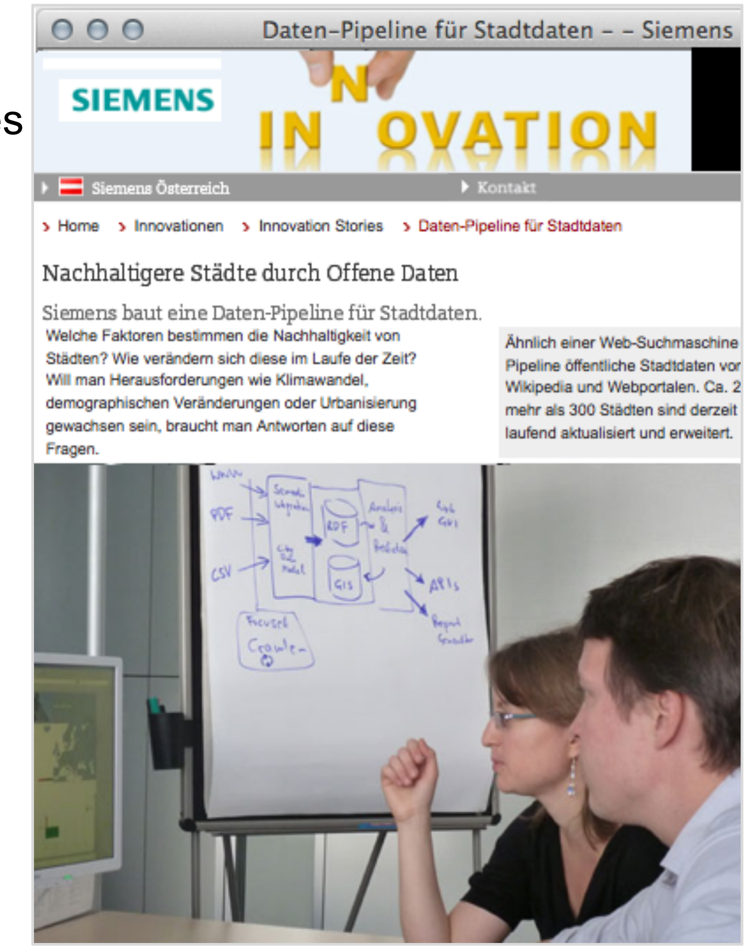
Is Open Data useful at all? A concrete use case:



A concrete use case: The "City Data Pipeline"

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

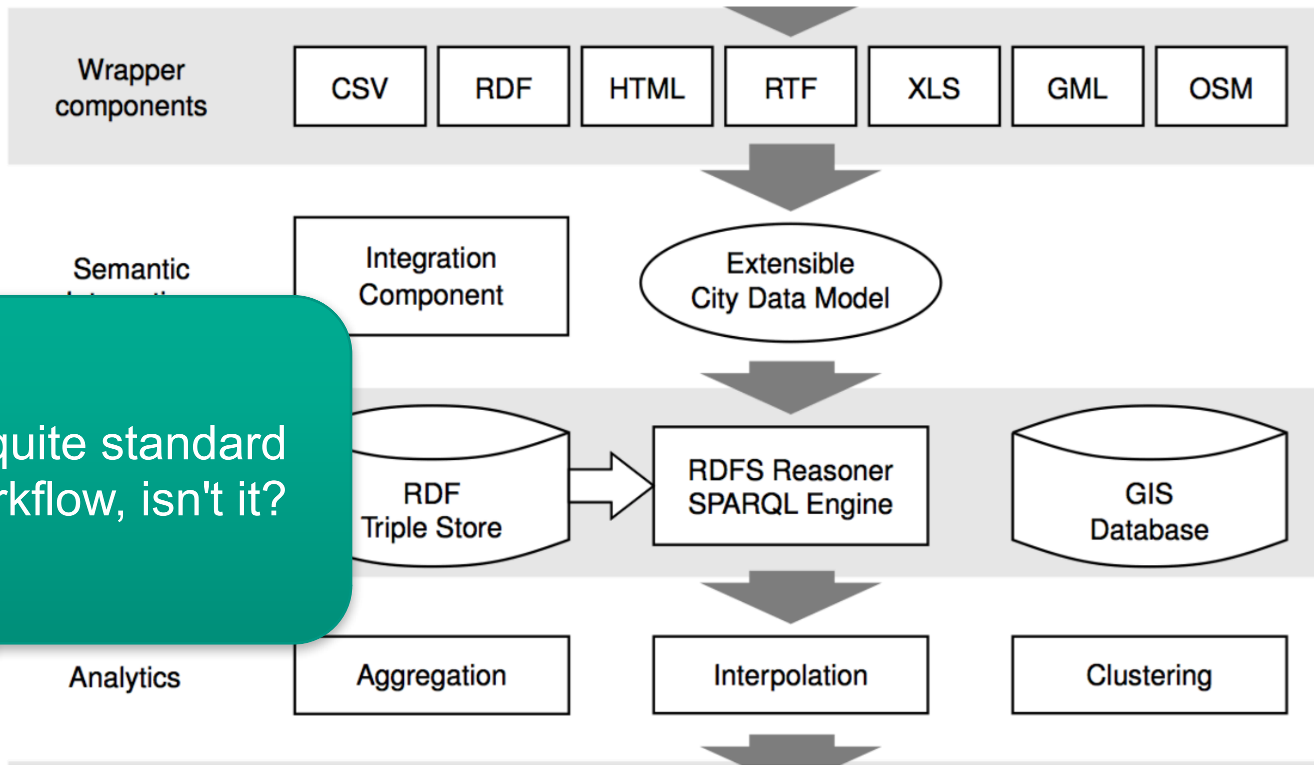
- Regularly integrate various relevant Open Data sources (e.g. eurostat, UNData, ...)
- Make integrated data available for re-use



(How) can ontologies help me?

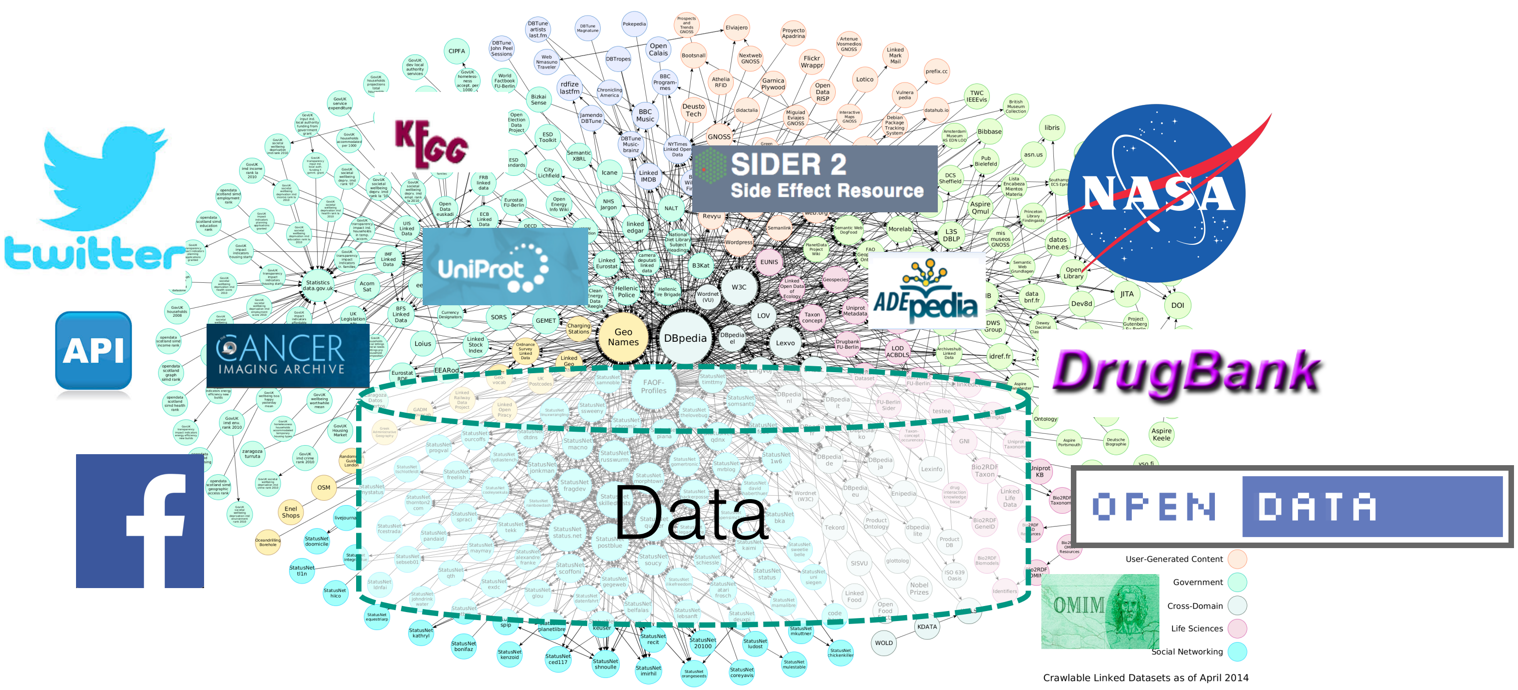
- Are ontology languages expressive enough?
- Which ontologies could I (re-)use?
- Is there enough data at all?
- Where to find the right data?
- Where to find the right ontologies?
- How to tackle inconsistencies?

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



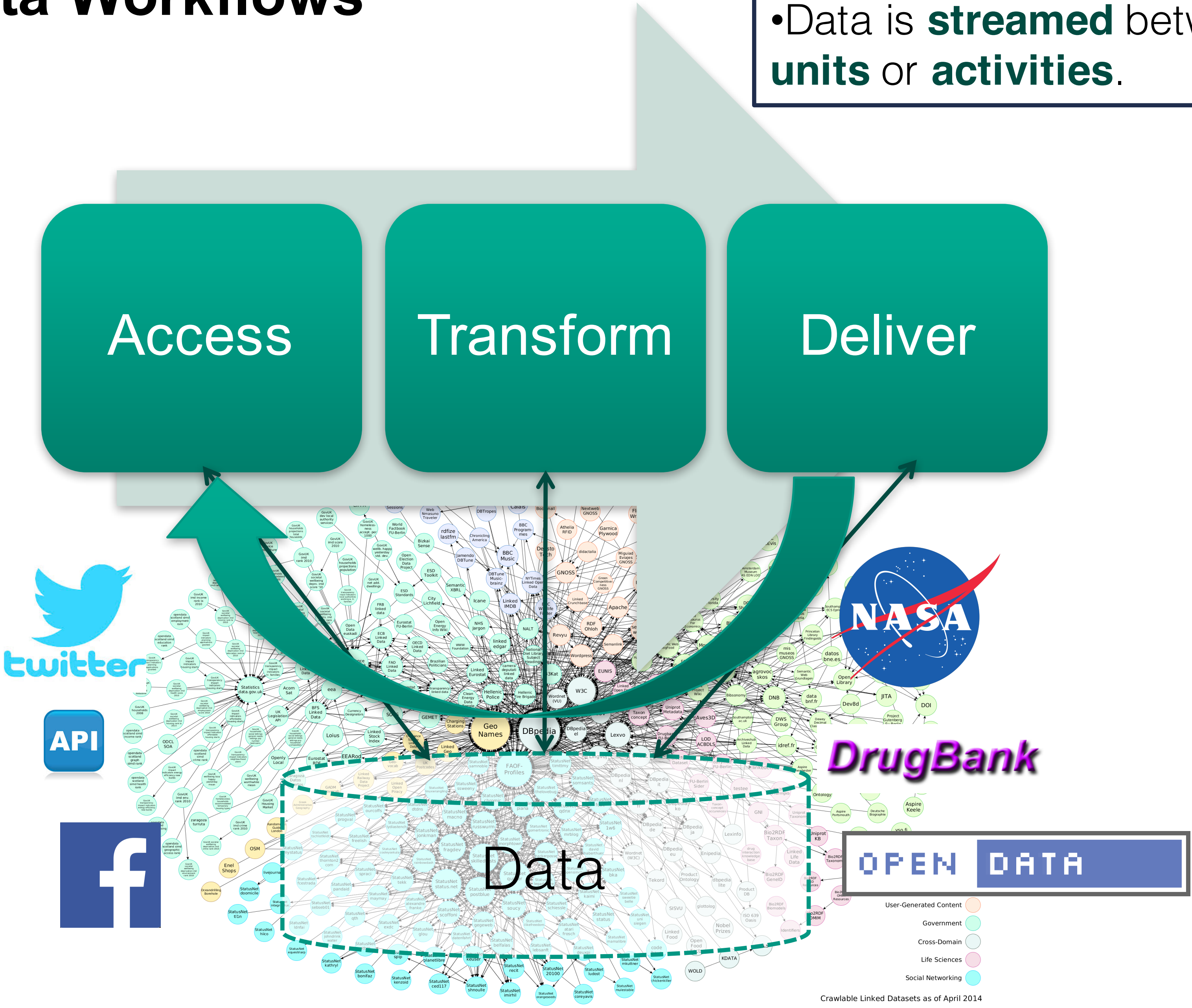
That's a quite standard Data Workflow, isn't it?

So:
a) What is a "standard data workflow"?
b) Where can/shall Semantic Technologies, but also traditional Data Integration technologies be used to build such workflows?



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:
1/7 „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)*

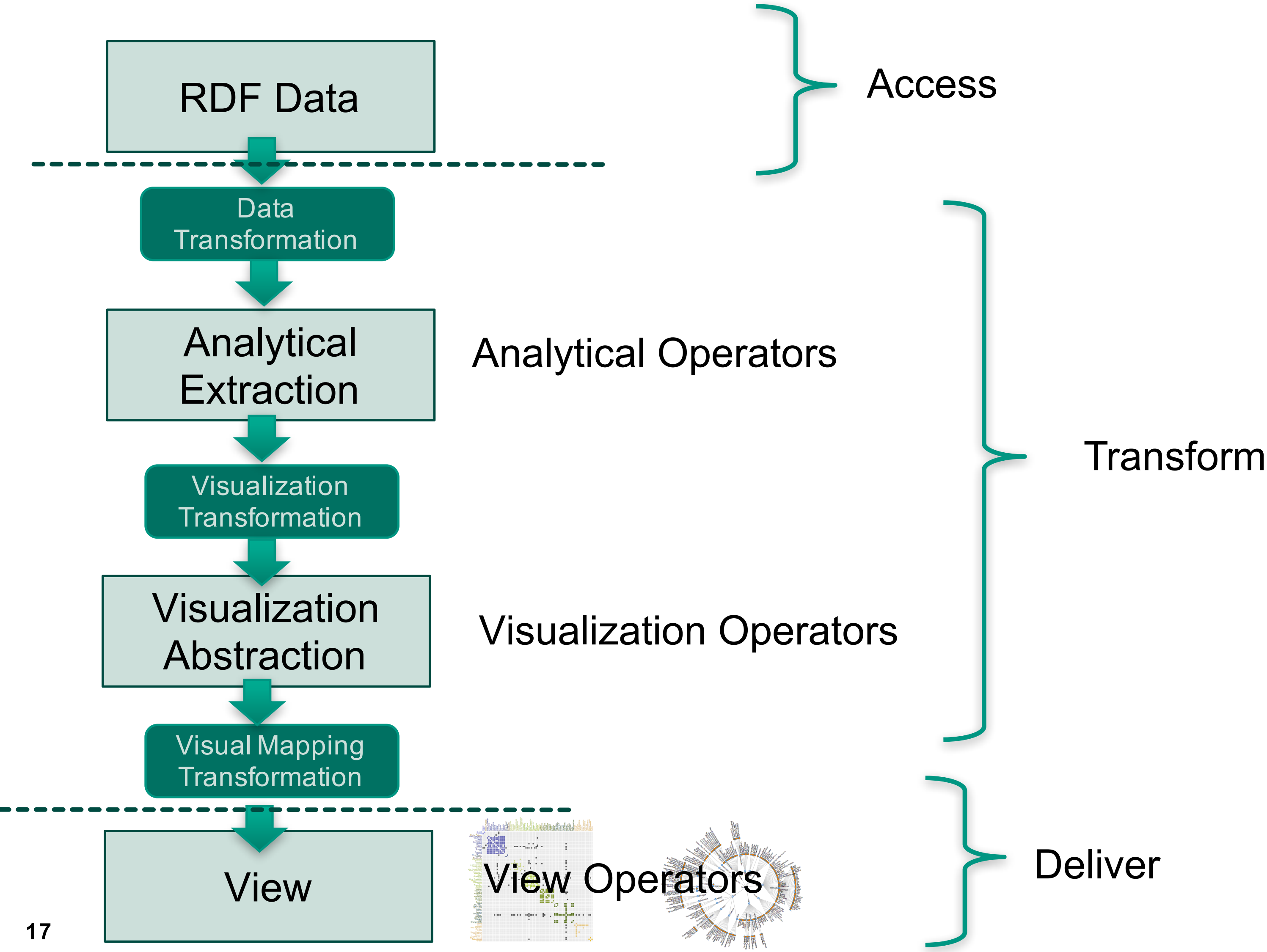
"*Hard-wired*" Data integration

- 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

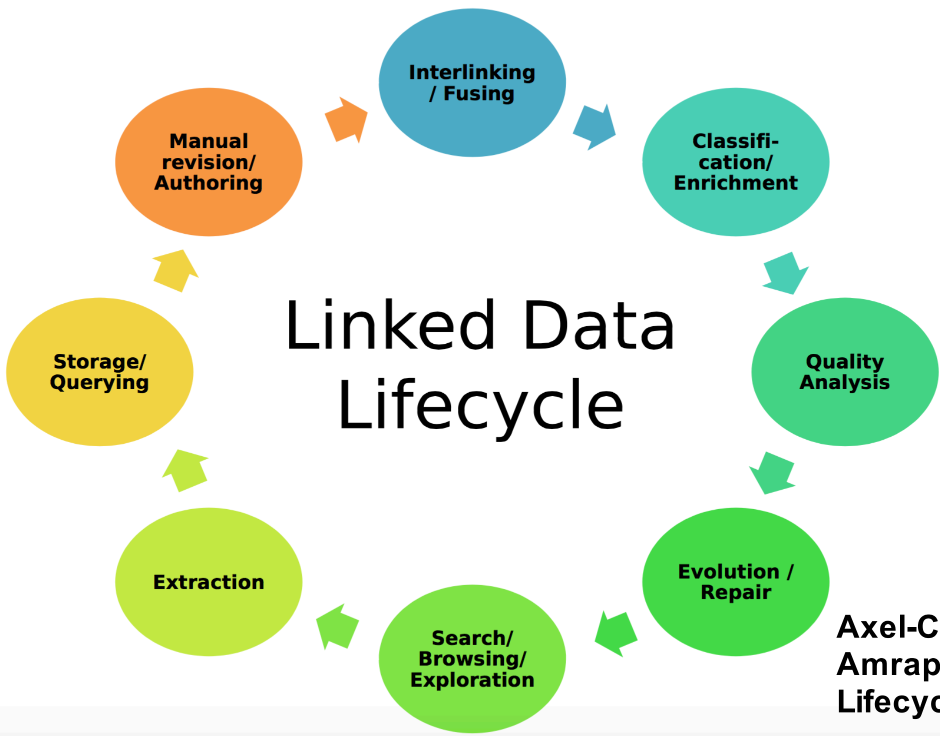
Different Views & Examples:
2/7 Linked Data Visualization Model



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

Different Views & Examples:
3/7 Or is it rather a Lifecycle...

- E.g. good example: Linked Data Lifecycle

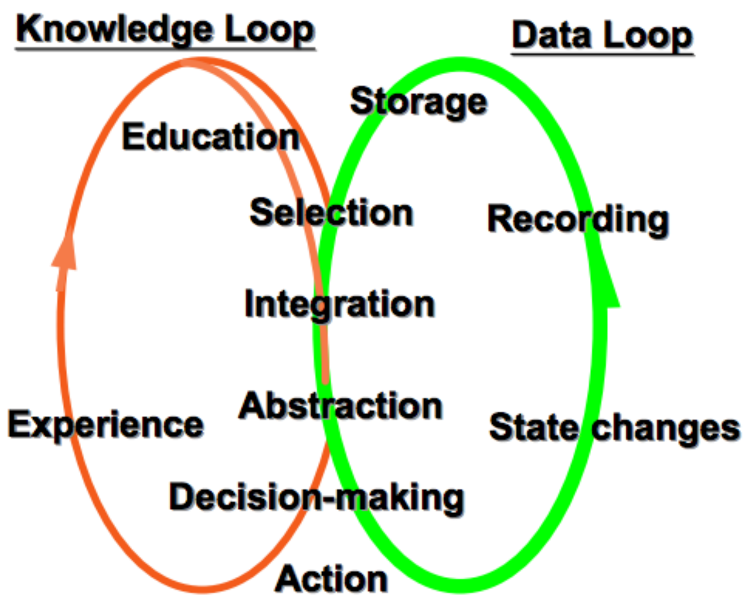


Axel-Cyrille Ngonga Ngomo, Sören Auer, Jens Lehmann, Amrapali Zaveri. Introduction to Linked Data and Its Lifecycle on the Web. ReasoningWeb. 2014

- **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:
4/7 We‘re not the first ones to recognize this is actually a lifecycle… [Wiederhold92]

Data and Knowledge



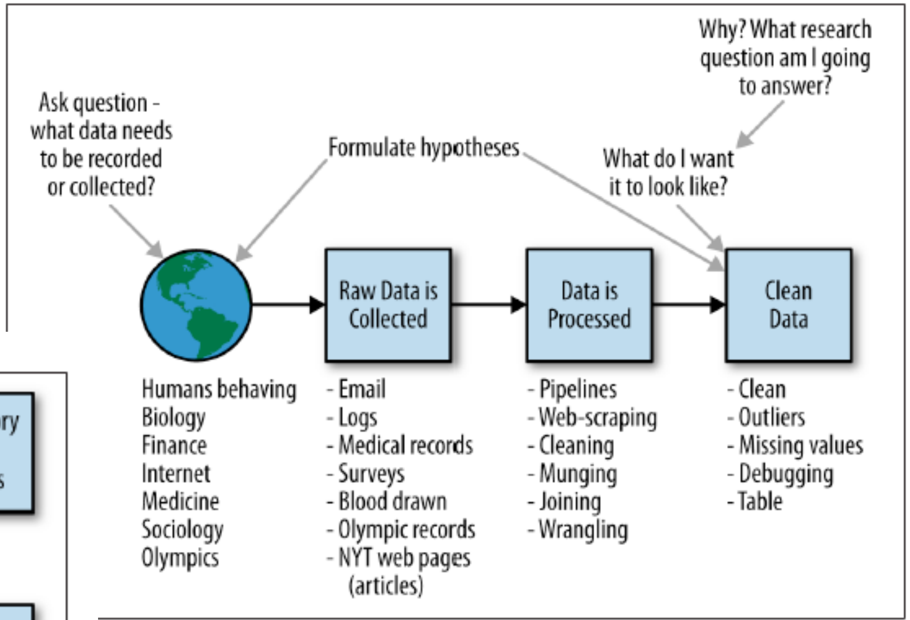
Information is created at the confluence of data -- the state & knowledge -- the ability to select and project the state into the future

Data describes specific instances and events. Data may gathered automatically or clerically. The correctness of data can be verified vis-a-vis the real world.

Knowledge describes abstract classes. Each class typically covers many instances. Experts are needed to gather and formalize knowledge. Data can be used to disprove knowledge.

Different Views & Examples:
5/6 The "Data Science" Process:

What Would a Next-Gen Data Scientist Do?
"[…] data scientists […] ***spend a lot more time trying to get data into shape than anyone cares to admit—maybe up to 90% of their time.*** Finally, they don't find religion in tools, methods, or academic departments. They are versatile and interdisciplinary"



2-3. The data scientist is involved in every part of this process

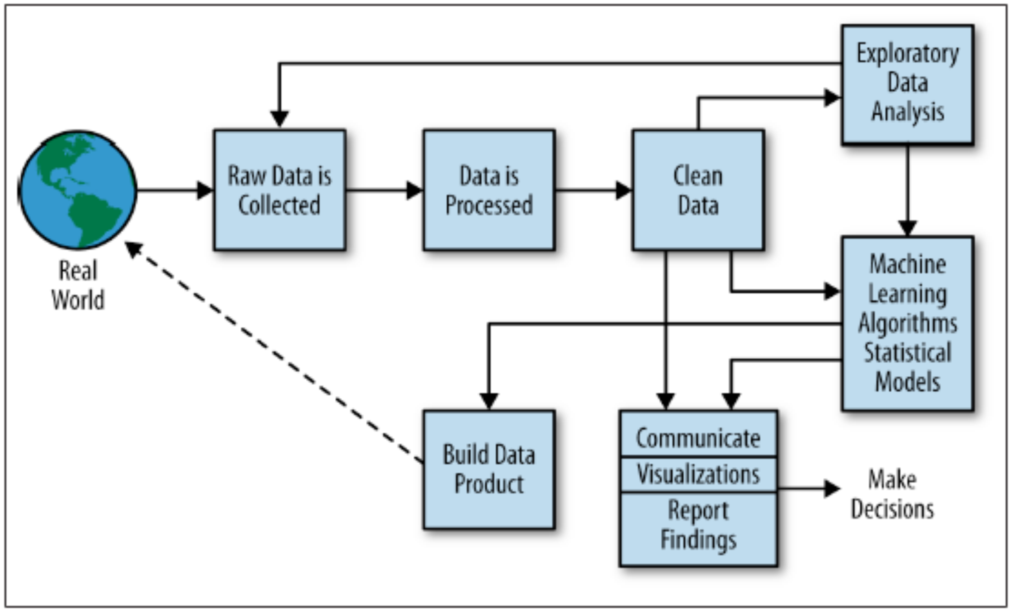
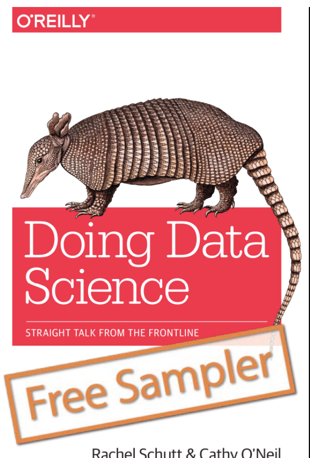
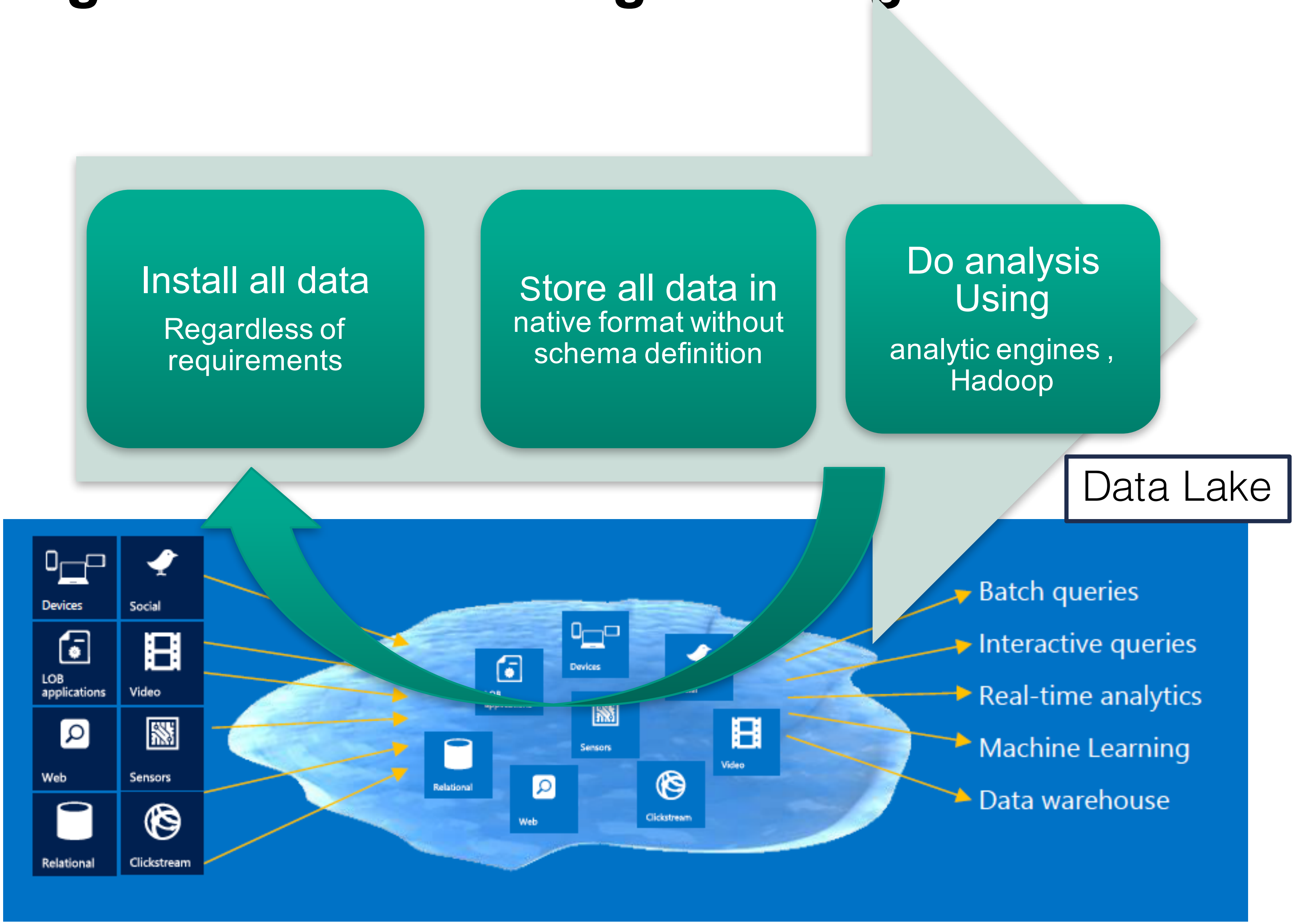


Figure 2-2. The data science process



http://semanticommunity.info/Data_Science/Doing_Data_Science

Different Views & Examples:
7/7 Big Data & Data Management against "Data Lake"



Raghu Ramakrishnan, Big Data @ Microsoft

General challenges to be addressed

Syntactic heterogeneity (different formats)

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“, sometimes imputation of missing values…)
- Data Integration (alignment, source reconciliation)
- Aggregation
- Cleansing (removing outliers)
- Deduplication/Interlinking (could involve "triplification")
- Analytics
- Enrichment
- Change dedection (Maintainance/Evolution)
- Validation (quality anaysis)
- Efficient, sometimes distributed (query) processing
- 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 and SW-biased!):

- 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/objec consolidation)
- KARMA (extraction, data integration) http://usc-isi-i2.github.io/karma/
- RapidMiner Linked Data extension http://dws.informatik.uni-mannheim.de/en/research/rapidminerlodextension/ [Gentile, Paulheim, et al. 2016]

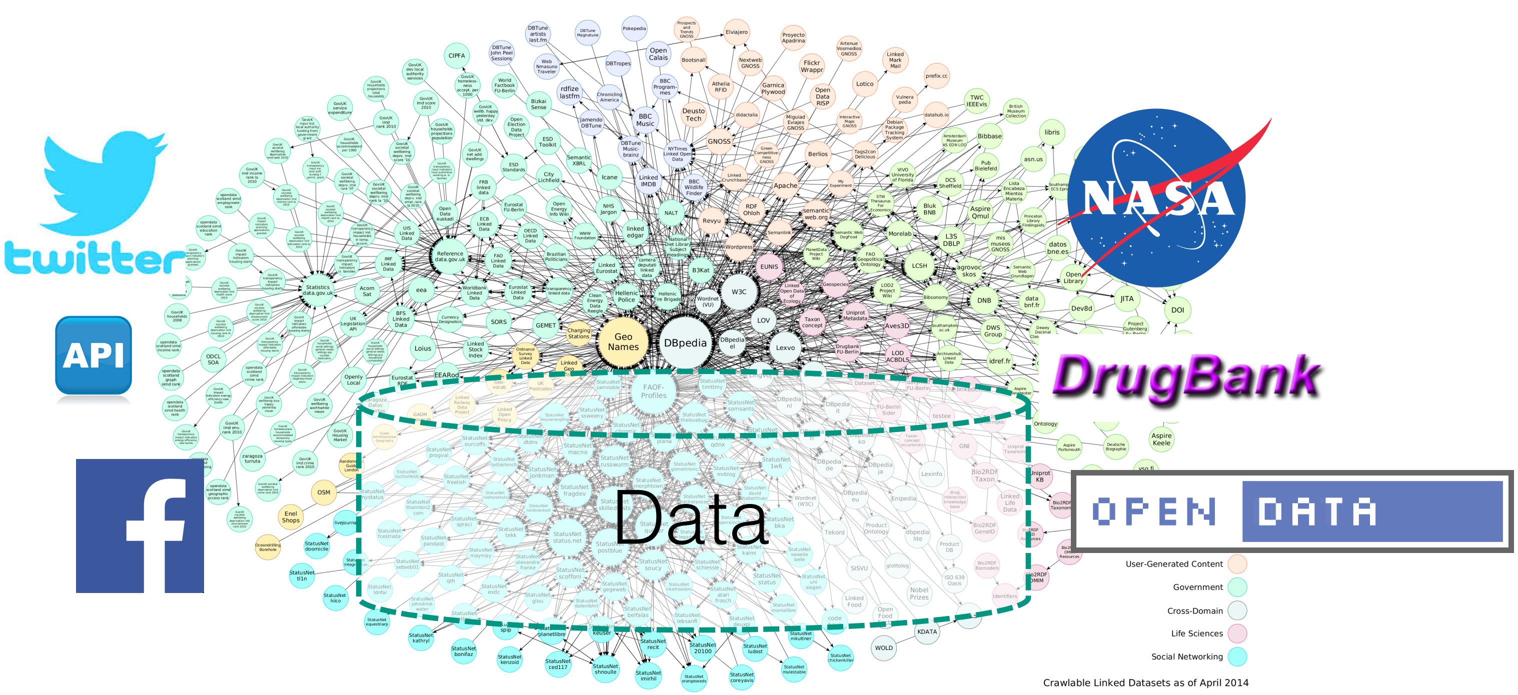
- XSPARQL (extraction from XML and JSON/triplicifation) http://sourceforge.net/projects/xsparql/ [Bischof et al. 2012]
  - Seel also: https://ai.wu.ac.at/~polleres/20140826xsparql_st.etienne/
- STTL: A SPARQL-based Transformation Language for RDF
  - See also: https://hal.inria.fr/hal-01150623 [Corby et al. 2015]

Outline

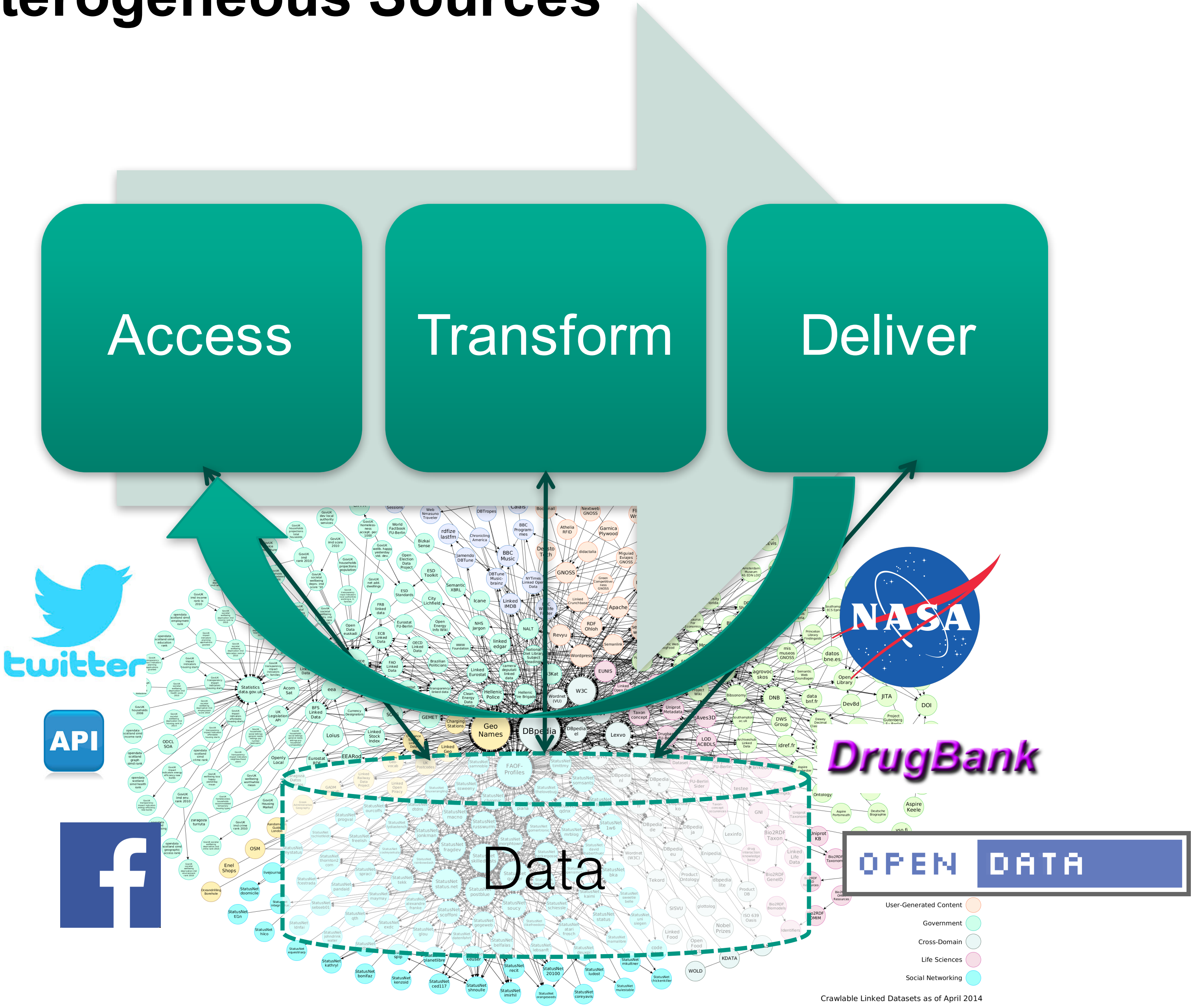
- Motivation
  - Integrating (Open) Data from different sources
  - Not only Linked Data
  - Data workflows and Open data in the context of rise of Big Data
- What is a "Data Workflow"?
  - Different Views of Data Workflows in the context of the Semantic Web
  - Key steps involved
  - Tools?
- **Data Integration Systems & Query Processing**
  - Data Integration Systems - GAV vs. LAV
  - The Mediator and Wrapper Architecture
  - Query rewriting vs. Materialisation
- Challenges:
  - How to find Rules and ontologies?
  - Incomplete Data
  - How to find the data?

DATA INTEGRATION SYSTEMS

Heterogeneous Sources



Data Workflows accessing Heterogeneous Sources



Mediators in the Architecture of Future Information Systems

3072

Gio Wiederhold
Stanford University

September 1991
An edited version of this report was published in
The IEEE Computer Magazine, March 1992

Abstract

The installation of high-speed networks using optical fiber and high bandwidth messsage forwarding gateways is changing the physical capabilties of information systems. These capabilities must be complemented with corresponding software systems advances to obtain a real benefit. Without smart software we will gain access to more data, but not improve access to the type and quality of information needed for decision making.

To develop the concepts needed for future information systems we model information processing as an interaction of data and knowledge. This model provides criteria for a high-level functional partitioning. These partitions are mapped into information processing modules. The modules are assigned to nodes of the distributed information systems. A central role is assigned to modules that *mediate* between the users' workstations and data resources. Mediators contain the administrative and technical knowledge to create information needed for decision-making. Software which mediates is common today, but the structure, the interfaces, and implementations vary greatly, so that automation of integration is awkward.

By formalizing and implementing mediation we establish a partitioned information systems architecture, which is of managable complexity and can deliver much of the power that technology puts into our reach. The partitions and modules map into the powerful distributed hardware that is becoming available. We refer to the modules that perform these services in a sharable and composable way as *mediators*.

We will present conceptual requirements that must be placed on mediators to assure effective large-scale information systems. The modularity in this architecture is not only a goal, but also enables the goal to be reached, since these systems will need autonomous modules to permit growth and enable them to survive in a rapidly changing world.

The intent of this paper is to provide a conceptual framework for many distinct efforts. The concepts provide a direction for an information processing systems in the foreseeable

Data Integration: A Theoretical Perspective

2623

Maurizio Lenzerini
Dipartimento di Informatica e Sistemistica
Università di Roma "La Sapienza"
Via Salaria 113, I-00198 Roma, Italy
lenzerini@dis.uniroma1.it

ABSTRACT

Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data. The problem of designing data integration systems is important in current real world applications, and is characterized by a number of issues that are interesting from a theoretical point of view. This document presents on overview of the material to be presented in a tutorial on data integration. The tutorial is focused on some of the theoretical issues that are relevant for data integration. Special attention will be devoted to the following aspects: modeling a data integration application, processing queries in data integration, dealing with inconsistent data sources, and reasoning on queries.

1. INTRODUCTION

Data integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data [60, 61, 89]. The problem of designing data integration systems is important in current real world applications, and is characterized by a number of issues that are interesting from a theoretical point of view. This tutorial is focused on some of these theoretical issues, with special emphasis on the following topics.

The data integration systems we are interested in this work are characterized by an architecture based on a global schema and a set of sources. The sources contain the real data, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. Modeling the relation between the sources and the global schema is therefore a crucial aspect. Two basic approaches have been proposed to this purpose. The first approach, called global-as-view, requires that the global schema is expressed in terms of the data sources. The second approach, called local-as-view, requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema. Our goal is to discuss the characteristics of the two modeling mechanisms, and to mention other possible approaches.

Irrespectively of the method used for the specification of the mapping between the global schema and the sources, one basic service provided by the data integration system is to answer queries posed in terms of the global schema. Given the architecture of the system, query processing in data integration requires a reformulation step: the query over the global schema has to be reformulated in terms of a set of queries over the sources. In this tutorial, such a reformulation problem will be analyzed for both the case of local-as-view, and the case of global-as-view mappings. A main theme will be the strong relationship between query processing in data integration and the problem of query answering with incomplete information.

Since sources are in general autonomous, in many real-world applications the problem arises of mutually inconsistent data sources. In practice, this problem is generally dealt with by means of suitable transformation and cleaning procedures applied to data retrieved from the sources. In this tutorial, we address this issue from a more theoretical perspective.

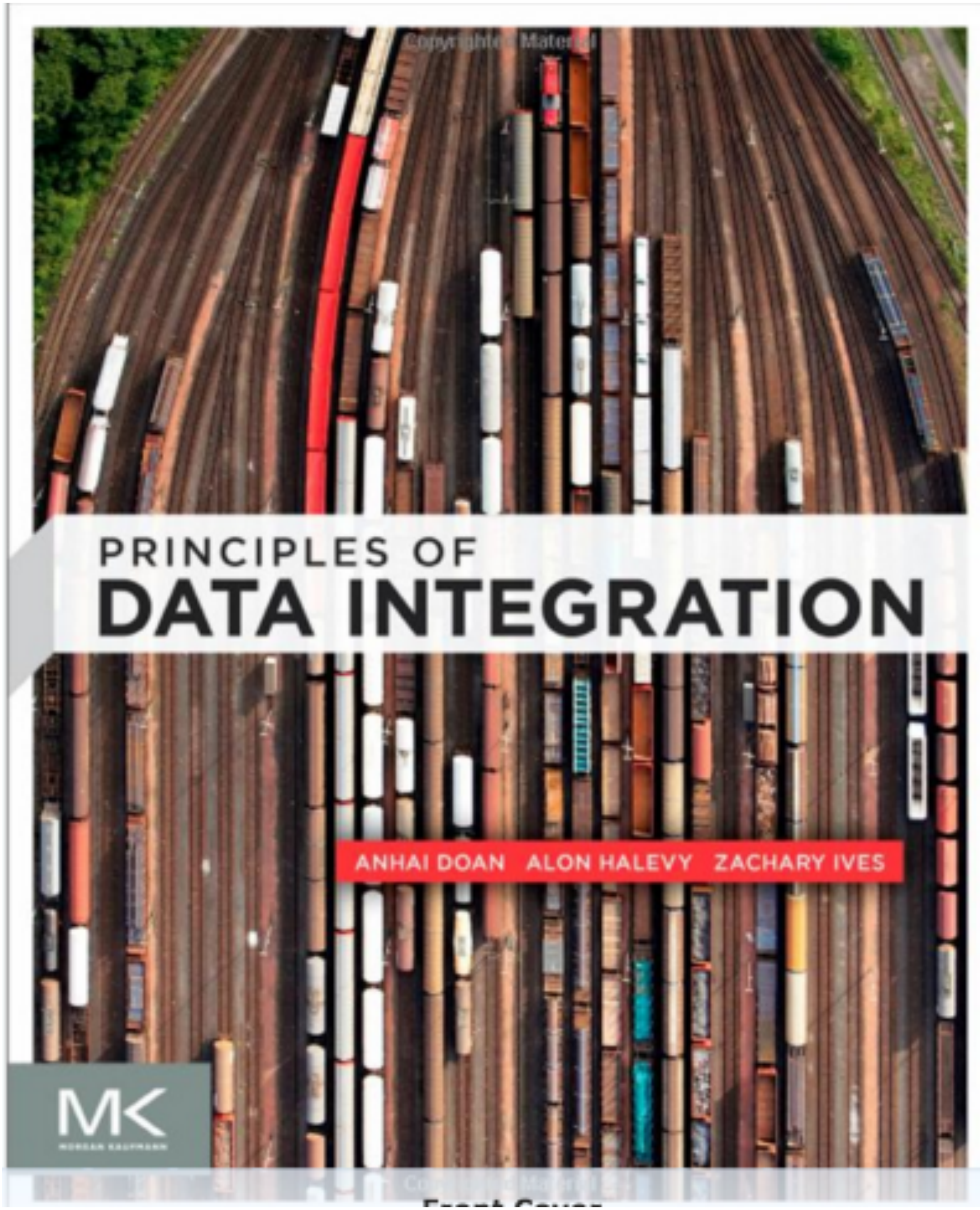
Finally, there are several tasks in the operation of a data integration system where the problem of reasoning on queries (e.g., checking whether two queries are equivalent) is relevant. Indeed, query containment is one of the basic problems in database theory, and we will discuss several notions generalizing this problem to a data integration setting.

The paper is organized as follows. Section 2 presents our formalization of a data integration system. In Section 3 we discuss the various approaches to modeling. Sections 4 and 5 present an overview of the methods for processing queries in the local-as-view and in the global-as-view approach, respectively. Section 6 discusses the problem of dealing with inconsistent sources. Section 7 provides an overview on the problem of reasoning on queries. Finally, Section 8 concludes the paper by mentioning some open problems, and several research issues related to data integration that are not addressed in the tutorial.

Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART. Symposium on Principles of database systems. 2002

2. DATA INTEGRATION FRAMEWORK

In this section we set up a logical framework for data integration. We restrict our attention to data integration systems



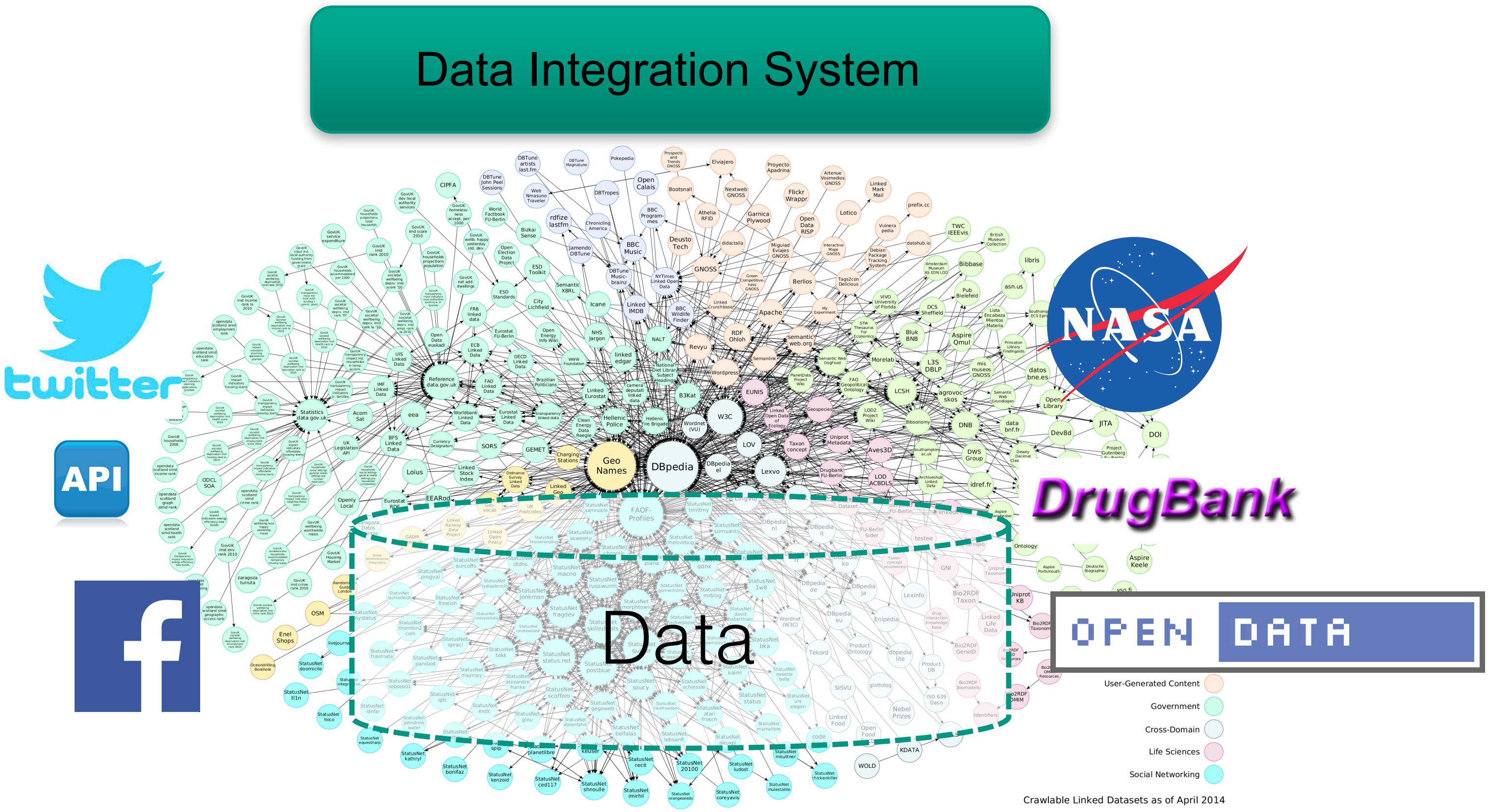
[Doan et al. 2012]

Data Integration Systems[Lenzerini2002]

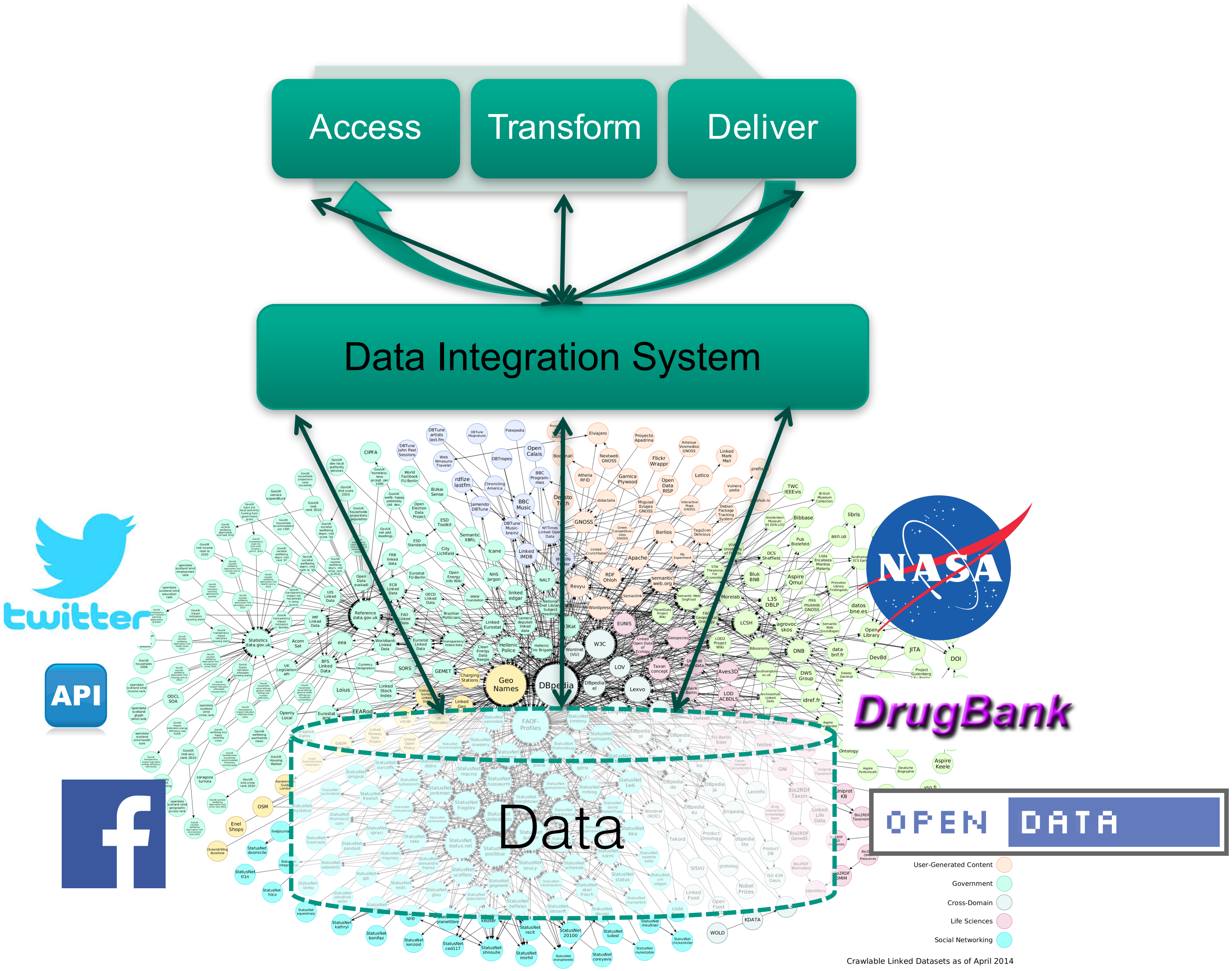
- IS=<O,S,M>
- Let O be a set of general concepts in a general schema (virtual).
- Let S={S1,..,Sn} be a set of data sources.
- Let M be a set of mappings between sources in S and general concepts in O.

cf. [Lenzerini 2002]

Heterogeneous Sources



Heterogeneous Sources



Global Schema

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

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

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

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

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

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

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

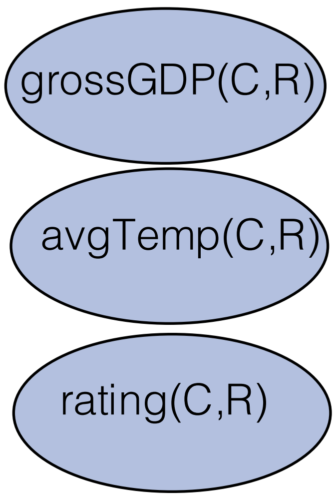
(**afCity** rdf:type rdfs:Class)

Global Schema

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

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

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

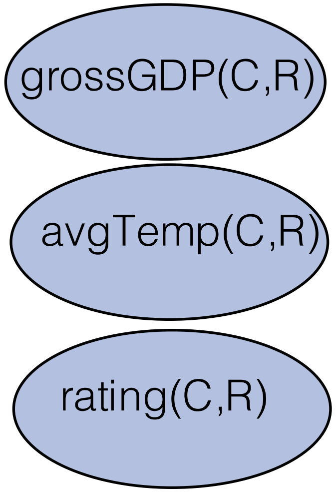


Global Schema

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

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

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



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

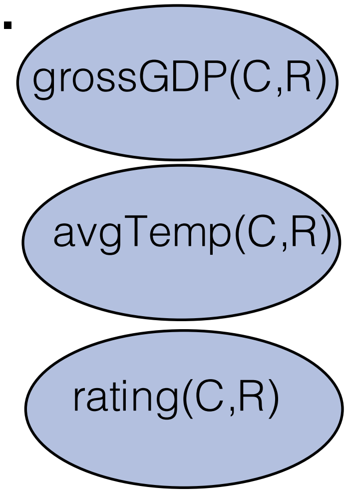
(**avgTemp** rdfs:subPropertyOf **rating**).

Global Schema

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

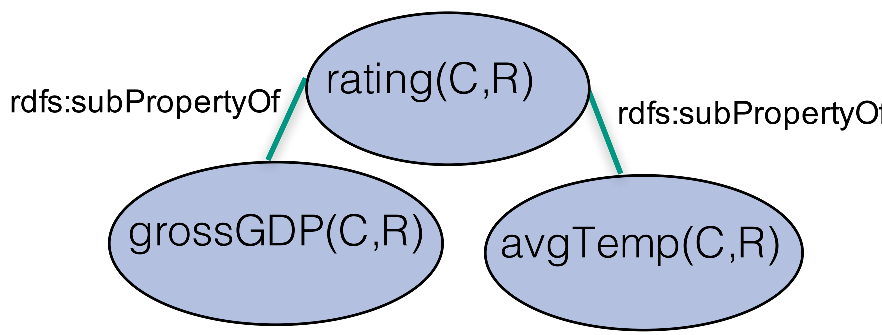
(**avgTemp** rdf:type rdf:Property).

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



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

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

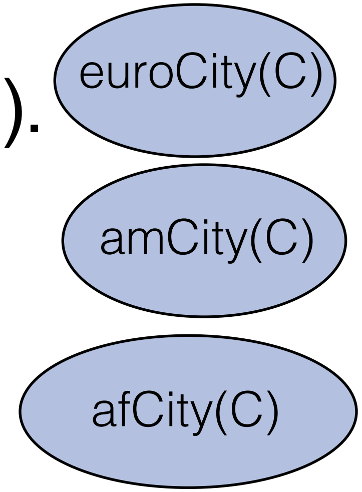


Global Schema

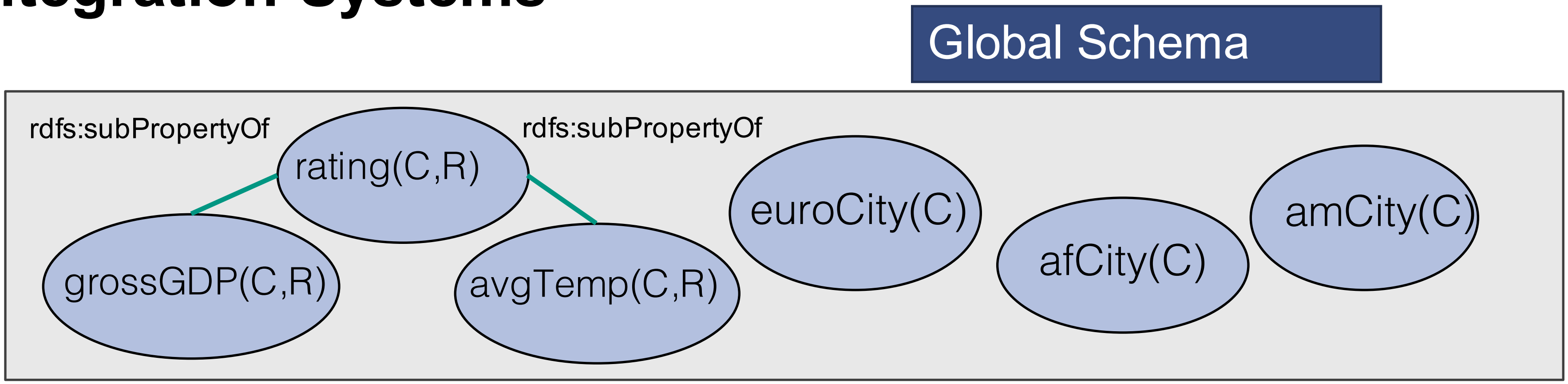
(**euroCity** rdf:type rdfs:Class).

(**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).

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)

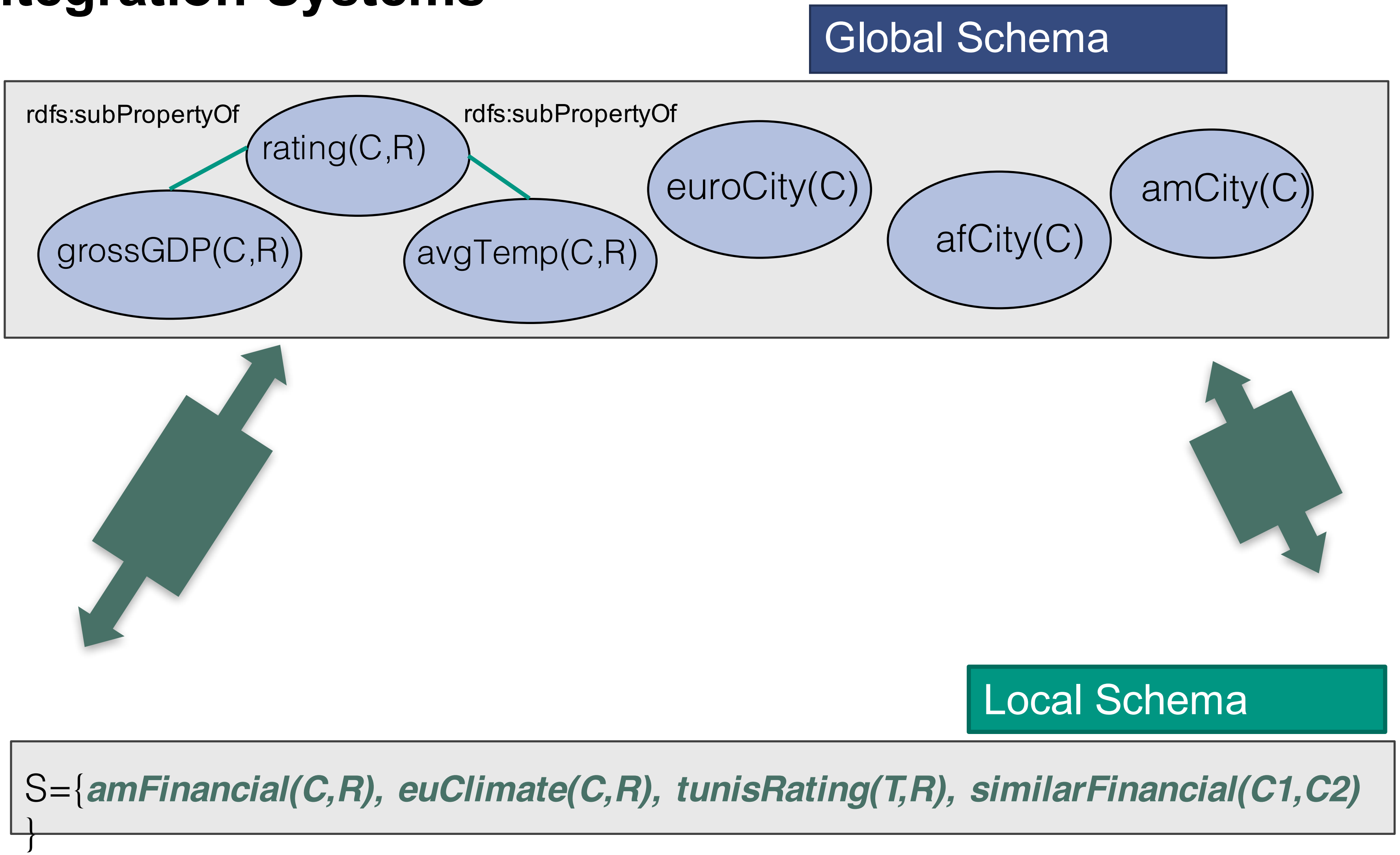
similarFinancial(C1,C2)

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



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).

Conjunctive Rules

- Q(X1,X2,..,Xn):-P1(Y11,..,Y1m),P2(Y21,…,Y2k),..,
  Pt(Yt1,…,Ytl),
  X1=Y1m, X2=Y2k, Xn=,Ytl.

Conjunctive Rules

Head of the Rule

- Q(X1,X2,..,Xn):-P1(Y11,..,Y1m),P2(Y21,…,Y2k),..,
  Pt(Yt1,…,Ytl),
  X1=Y1m, X2=Y2k, Xn=,Ytl.

Conjunctive Rules

Body of the Rule

- Q(X1,X2,..,Xn):-P1(Y11,..,Y1m),P2(Y21,…,Y2k),..,
  Pt(Yt1,…,Ytl),
  X1=Y1m, X2=Y2k, Xn=,Ytl.

Conjunctive Rules

Body of the Rule

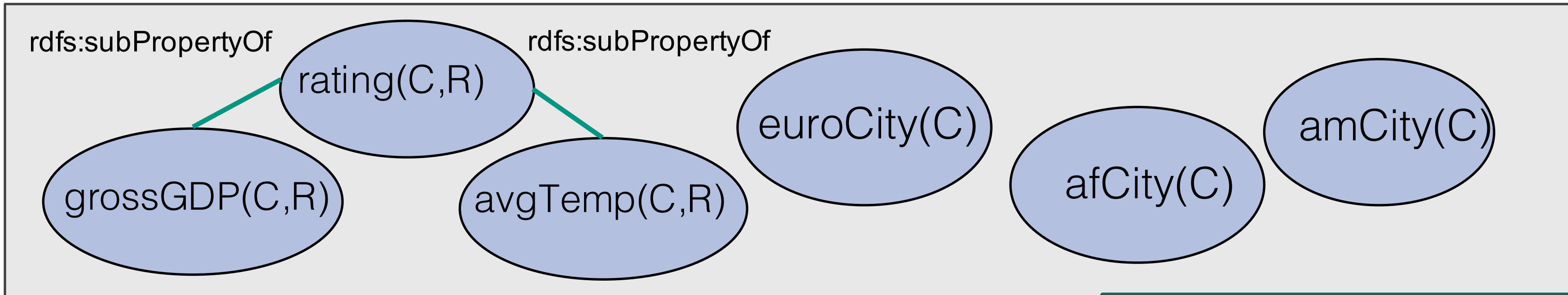
- Q(X1,X2,..,Xn):-P1(Y11,..,Y1m),P2(Y21,…,Y2k),..,
  Pt(Yt1,…,Ytl),
  X1=Y1m, X2=Y2k, Xn=,Ytl.

P1(Y11,..,Y1m),P2(Y21,…,Y2k),..,Pt(Yt1,…,Ytl),, X1=Y1m, X2=Y2k, Xn=,Ytl.

Q(X1,X2,..,Xn)

Global-As-View (GAV)

Global Schema



Local Schema

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

α0: amCity(C):-***amFinancial(C,R).***

α1: grossGDP(C,R):-***amFinancial(C,R)***.

α2: euroCity(C):-***euClimate(C,R)***.

α3: avgTemp(C,R):-***euClimate(C,R)***.

α4: grossGDP("Tunis",R):-***tunisRating("financial",R).***

α5: avgTemp("Tunis",R):-***tunisRating("climate",R)***

α6: afCity("Tunis").

α7: amCity(C1):-***similarFinancial(C1,C2).***

α8: amCity(C2):-***similarFinancial(C1,C2).***

α9: grossGDP(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):-grossGDP(C,R), amCity(C)

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

α1: grossGDP(C,R):-***amFinancial(C,R)***.

α7: amCity(C1):-***similarFinancial(C1,C2).***

Example GAV:

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

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

Rewritings

Global-As-View (GAV)- Query Unfolding

- query(X):-p1(Y1),p2(Y2),….,pn(Yn).

  p1(Y1):-q11(Y11),…,q1m(Y1m)

  p2(Y2):-q21(Y22),…,q2l(Y2l)

  …

  pn(Yn):-qn1(Yn1),…,qnk(Yn1)

Global-As-View (GAV)- Query Unfolding

- query(X):-p1(Y1),p2(Y2),….,pn(Yn).

  p1(Y1):-q11(Y11),…,q1m(Y1m)

  p2(Y2):-q21(Y22),…,q2l(Y2l)

  …

  pn(Yn):-qn1(Yn1),…,qnk(Yn1)

Global-As-View (GAV)- Query Unfolding

- query(X):-p1(Y1),p2(Y2),….,pn(Yn).

  p1(Y1):-q11(Y11),…,q1m(Y1m)

  p2(Y2):-q21(Y22),…,q2l(Y2l)

  …

  pn(Yn):-qn1(Yn1),…,qnk(Yn1)

Global-As-View (GAV)- Query Unfolding

- query(X):-p1(Y1),p2(Y2),….,pn(Yn).

  p1(Y1):-q11(Y11),…,q1m(Y1m)

  p2(Y2):-q21(Y22),…,q2l(Y2l)

  …

  pn(Yn):-qn1(Yn1),…,qnk(Yn1)

query(X):-q11(Y11),…,q1m(Y1m) , q21(Y22),…,q2l(Y2l),..,
qn1(Yn1),…,qnk(Ynk).

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):-grossGDP(C,R), amCity(C)

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

α9:grossGDP(C1,R):-***similarFinancial(C1,C2)***, ***amFinancial(C2,R)***.

α7: amCity(C1):-***similarFinancial(C1,C2).***

Example GAV:

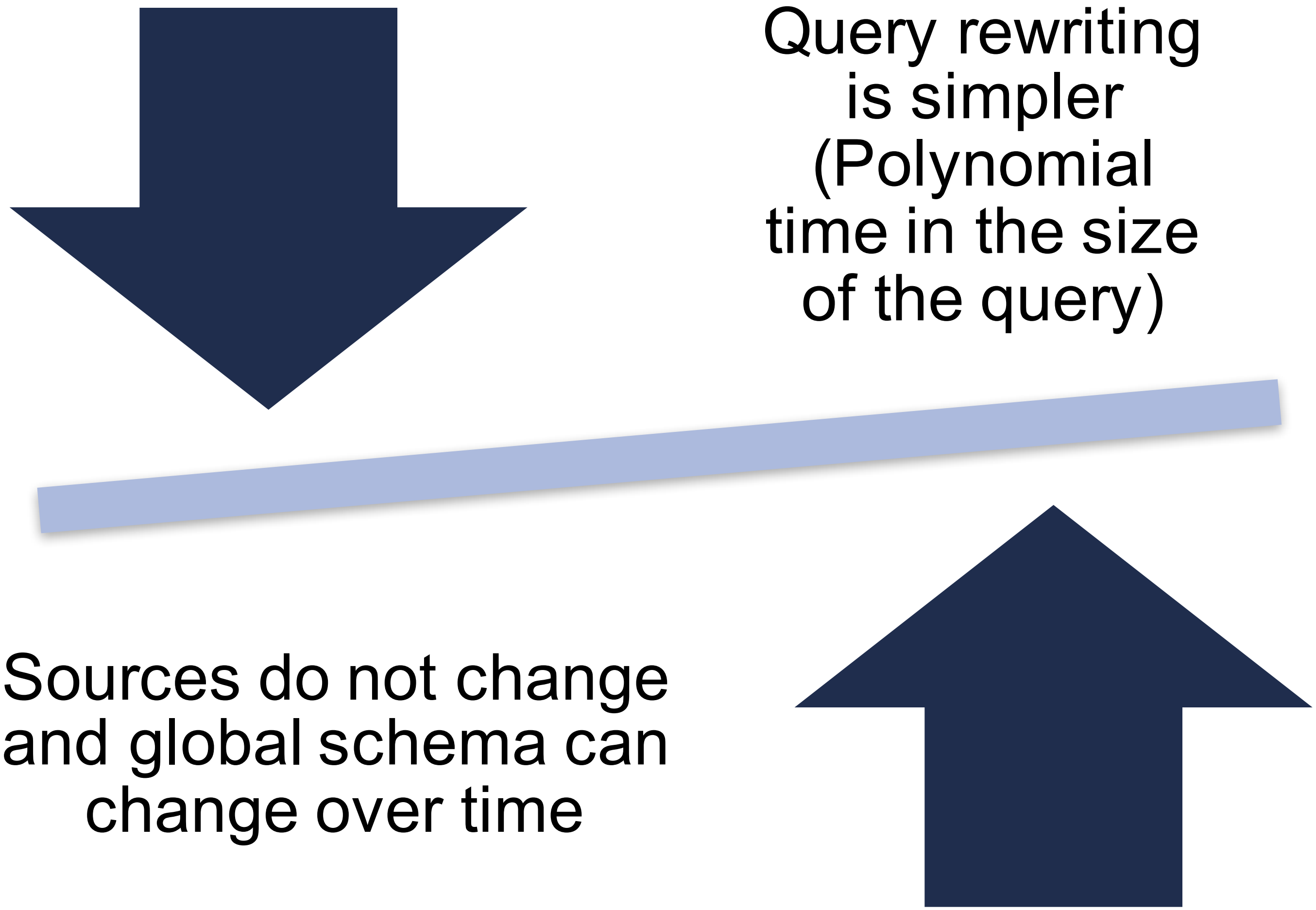
query(C):-grossGDP(C,R), amCity(C)

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

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

Rewritings

When to use GAV



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)

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).

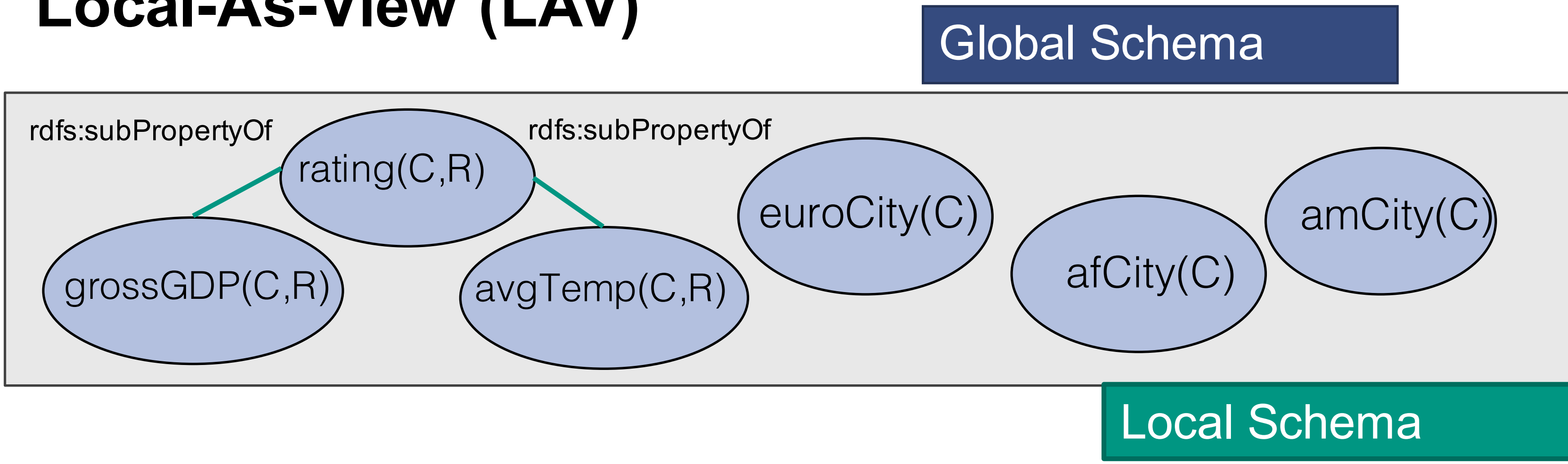
Local-As-View (LAV):

- 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).

Local-As-View (LAV)



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

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

Query Rewriting LAV

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

Example LAV:

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

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

Rewritings

Local As View-Query Rewriting

query(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

S1(X1,X2,X3):-C1(X1,X2),C2(X2,X3).

S2(X3,X4,X5):-C3(X3,X4),C4(X4,X5).

S3(X2,X3,X4):-C2(X2,X3),C3(X3,X4).

S4(X1,X2):-C1(X1,X2).

Local As View-Query Rewriting

query(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

S1(X1,X2,X3):-C1(X1,X2),C2(X2,X3).

S2(X3,X4,X5):-C3(X3,X4),C4(X4,X5).

S3(X2,X3,X4):-C2(X2,X3),C3(X3,X4).

S4(X1,X2):-C1(X1,X2).

S1(X1,X2,X3) S2(X3,X4,X5)

query'(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

Local As View-Query Rewriting

query(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

S1(X1,X2,X3):-C1(X1,X2),C2(X2,X3).

S2(X3,X4,X5):-C3(X3,X4),C4(X4,X5).

S3(X2,X3,X4):-C2(X2,X3),C3(X3,X4).

S4(X1,X2):-C1(X1,X2).

S1(X1,X2,X3) S2(X3,X4,X5)

query'(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

S4(X1,X2) S3(X2,X3,X4) S4(X3,X4,X5)

query''(X1,X5):-C1(X1,X2),C2(X2,X3),C3(X3,X4),C4(X4,X5)

Query Rewriting

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

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

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

Source Database		Virtual Database
Washington NYC Miami Caracas Lima	$\subseteq$	Washington NYC Miami Caracas Lima …… Bogota College Park Quito

Query Rewriting LAV

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

Example LAV:

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

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

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

} Rewritings

Query Rewriting LAV

α0:***amFinancial(C,R)***:-amCity(C),grossGDP(C,R).

α1:***euClimate(C,R)***:-euCity(C),avgTemp(C,R).

α2:***tunisRating("financial",R)***:-afCity("Tunis"),grosGDP("Tunis",R).

α3:***tunisRating("climate",R)***:-afCity("Tunis"),avgTemp("Tunis",R).

α4:***similarFinancial(C1,C2)***:-amCity(C1),amCity(C2),
grossGDP(C1,R),grossGDP(C2,R).

Example LAV:

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

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

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

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

Rewritings

Time Complexity

To check whether there is a valid rewriting R of Q with at most the same number of goals as Q is an NP-complete problem.

Levy, A.; Mendelzon, A.; Sagiv, Y.; and Srivastava, D. 1995. Answering queries using views. In Proc. of PODS, 95–104.

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]

When to use LAV

A GAV catalog cannot be easily adapted to changes in the data sources

LAV views can be easily adapted to changes in the data sources

Data Sources can be easily described

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).

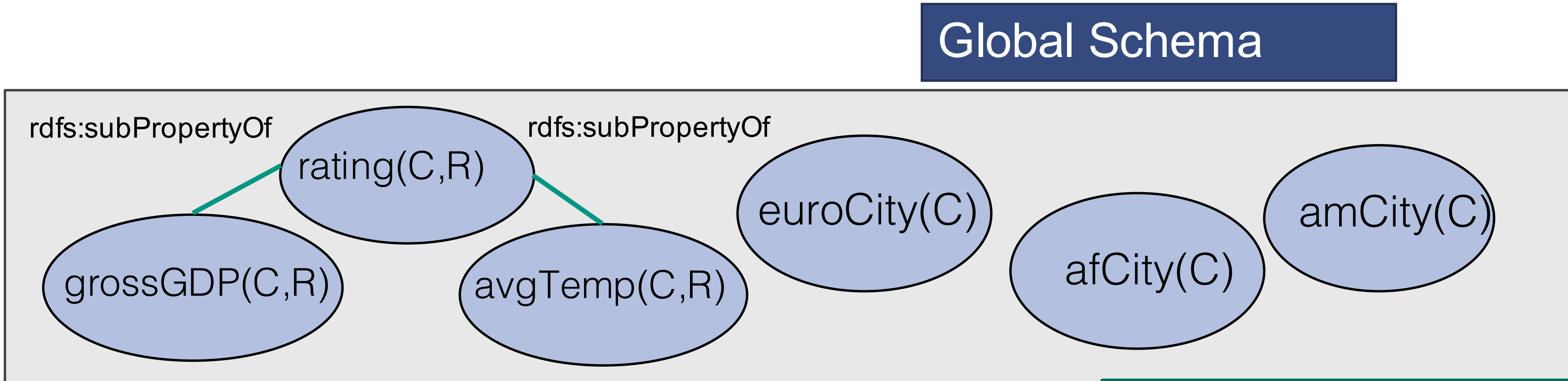
Local-As-View (LAV):

- 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-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 GLAV

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

Example GLAV:

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

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

Rewritings

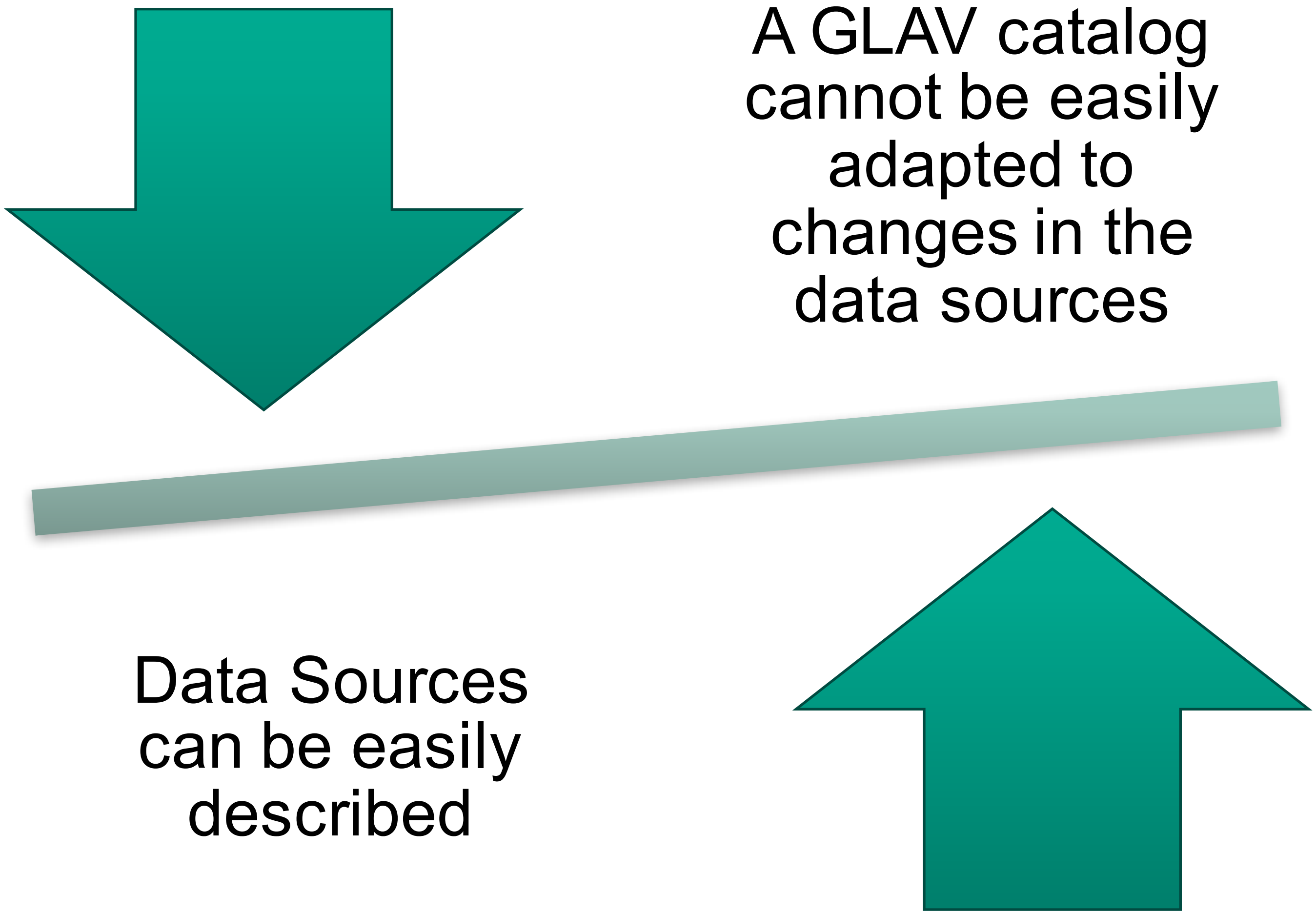
Query Rewriting

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

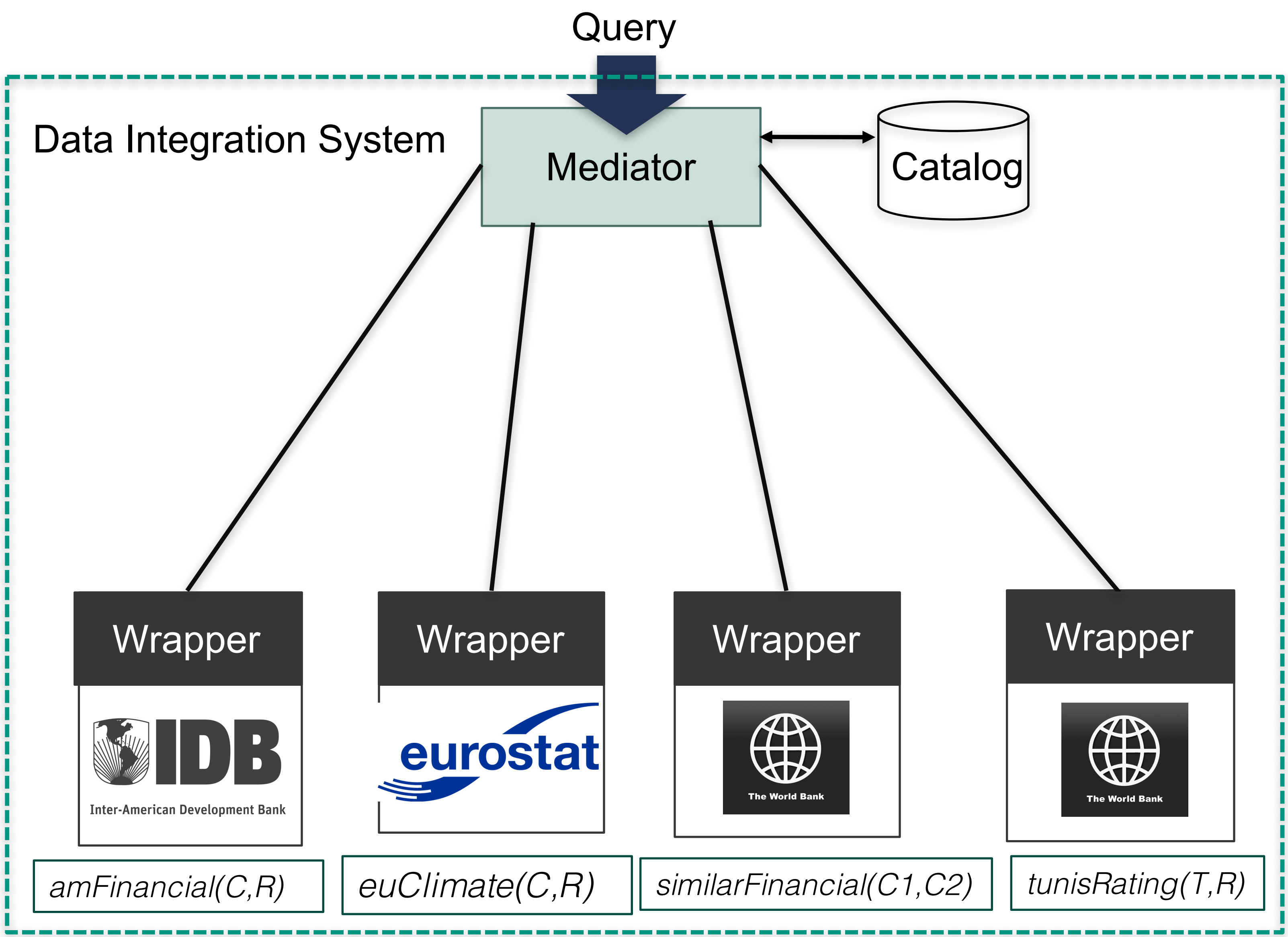
Query Containment: $Q' \subseteq Q \leftrightarrow \forall DB\ Q'(DB) \subseteq Q(DB)$

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

When to use GLAV



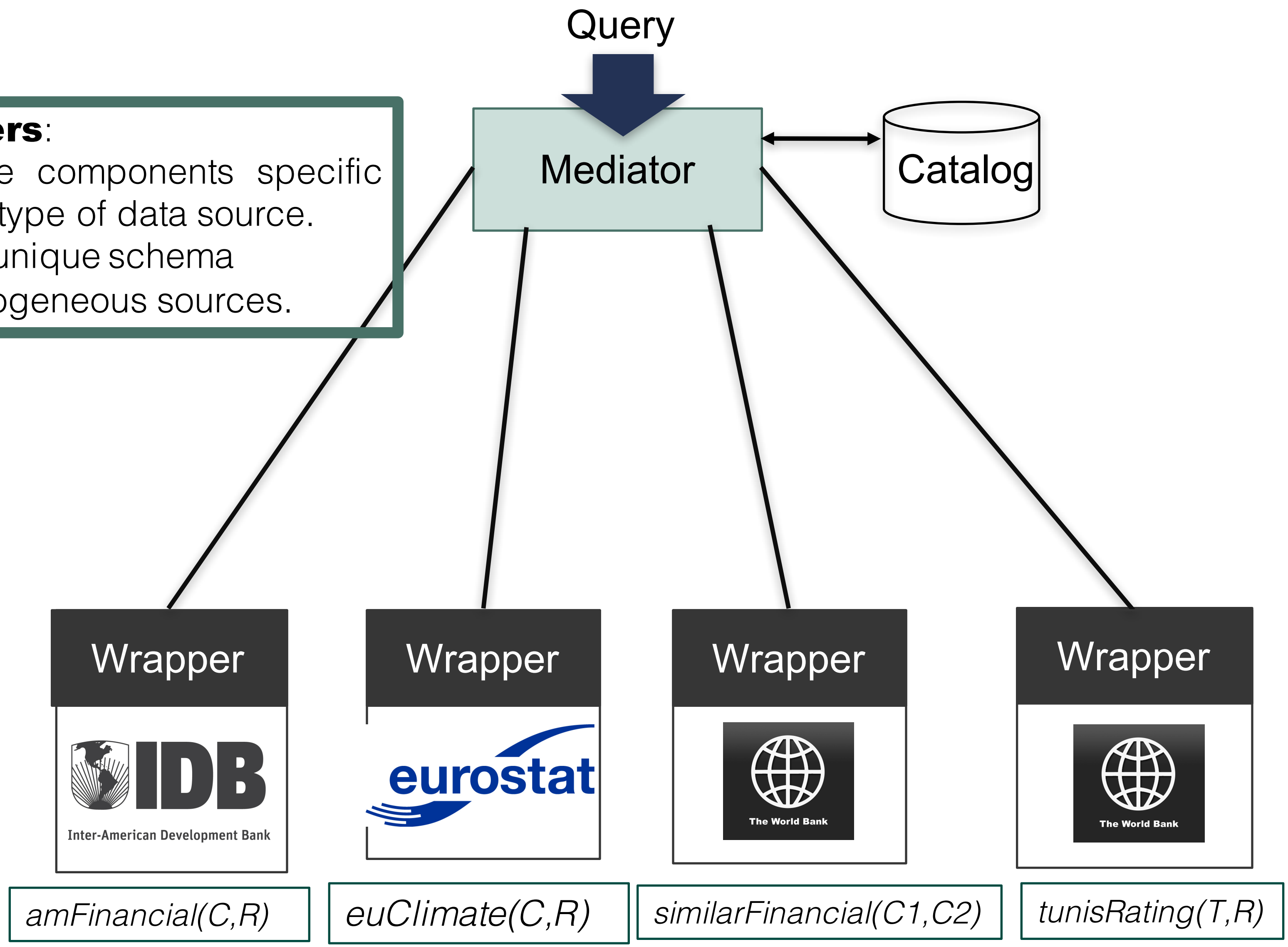
The Mediator and Wrapper Architecture [Wiederhold92]



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

The Mediator and Wrapper Architecture [Wiederhold92]

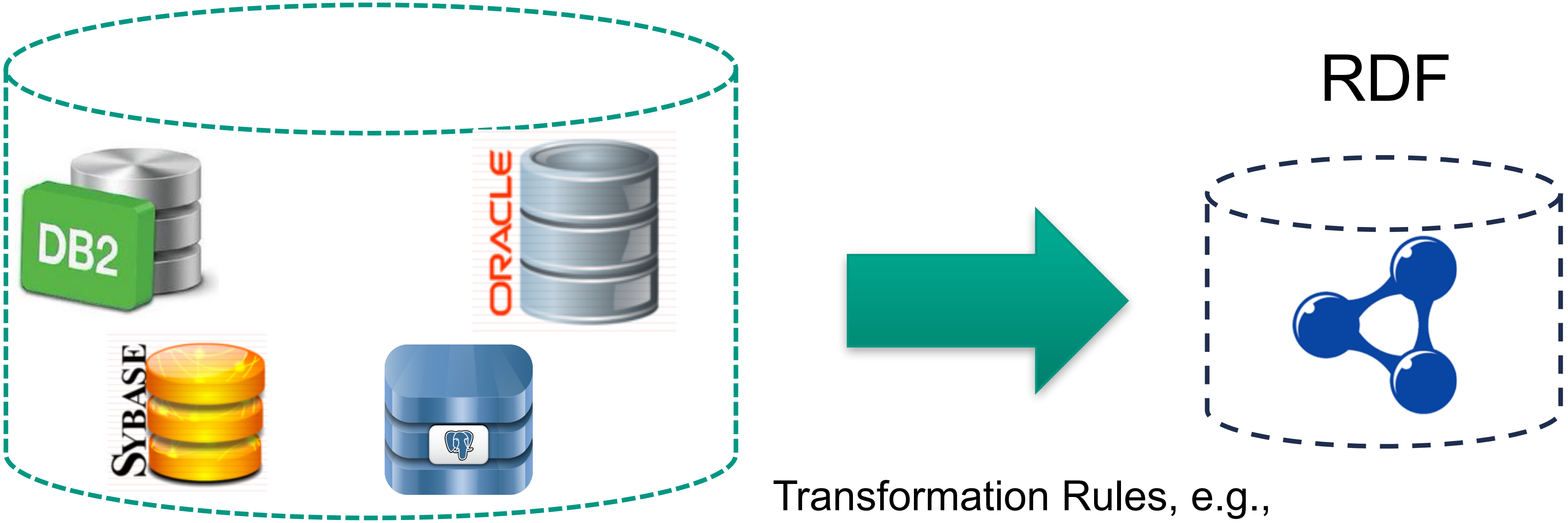
Wrappers:
- Software components specific for each type of data source.
- Export unique schema for heterogeneous sources.



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

Wrappers in the context of RDF Data:

e.g. RDB2RDF Systems

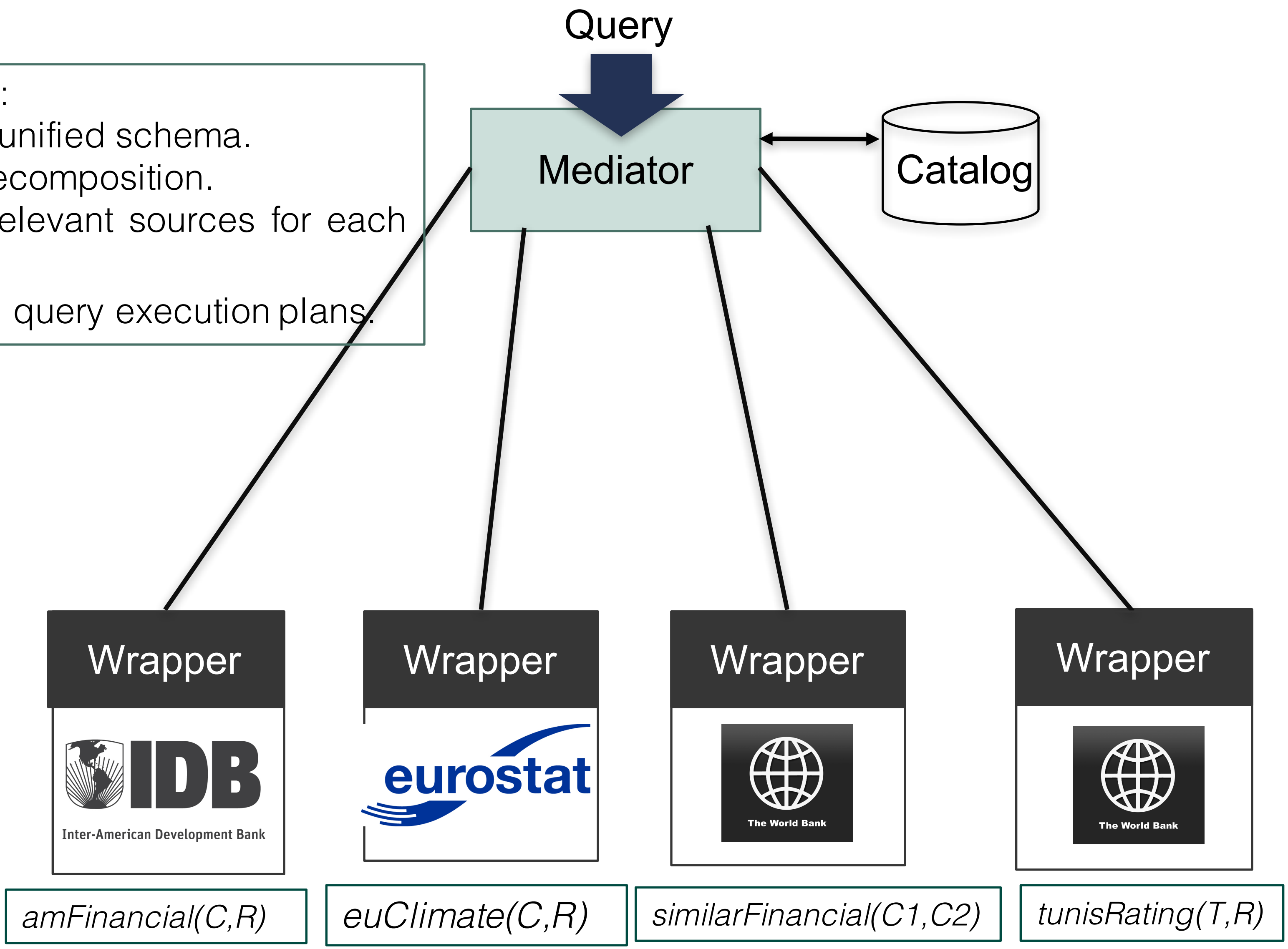


Transformation Rules, e.g., R2RML

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)

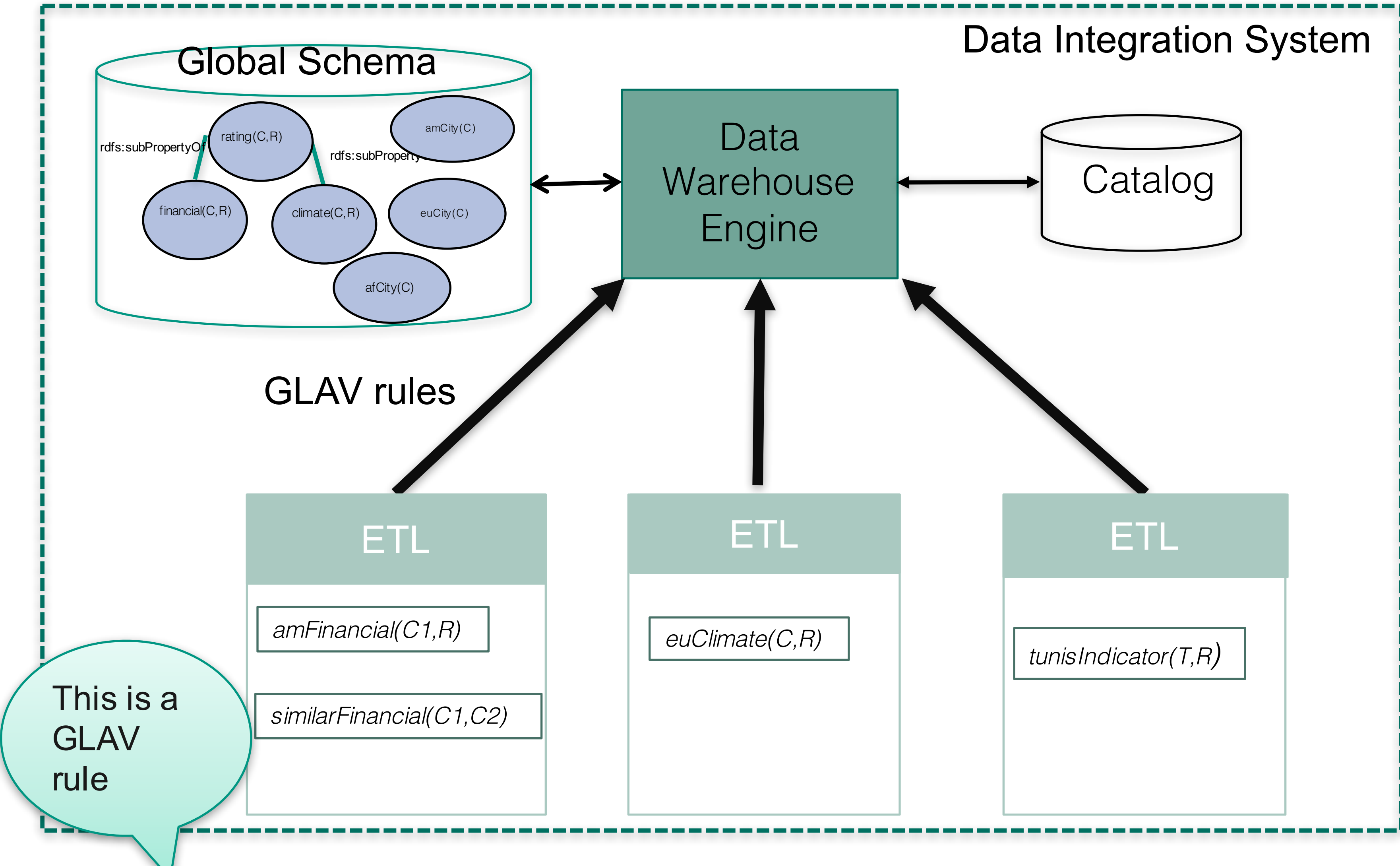
Some recent works which implement Wiederhold's mediator/wrapper architecture in the SW:

Linked Data-Fu [Stadtmüller et al. 2013]
SemLAV [Montoya et al. 2014]

… both LAV-inspired.

MATERIALIZED GLOBAL SCHEMA-DATA WAREHOUSE

Data Warehouse-Materialized Global Schema



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

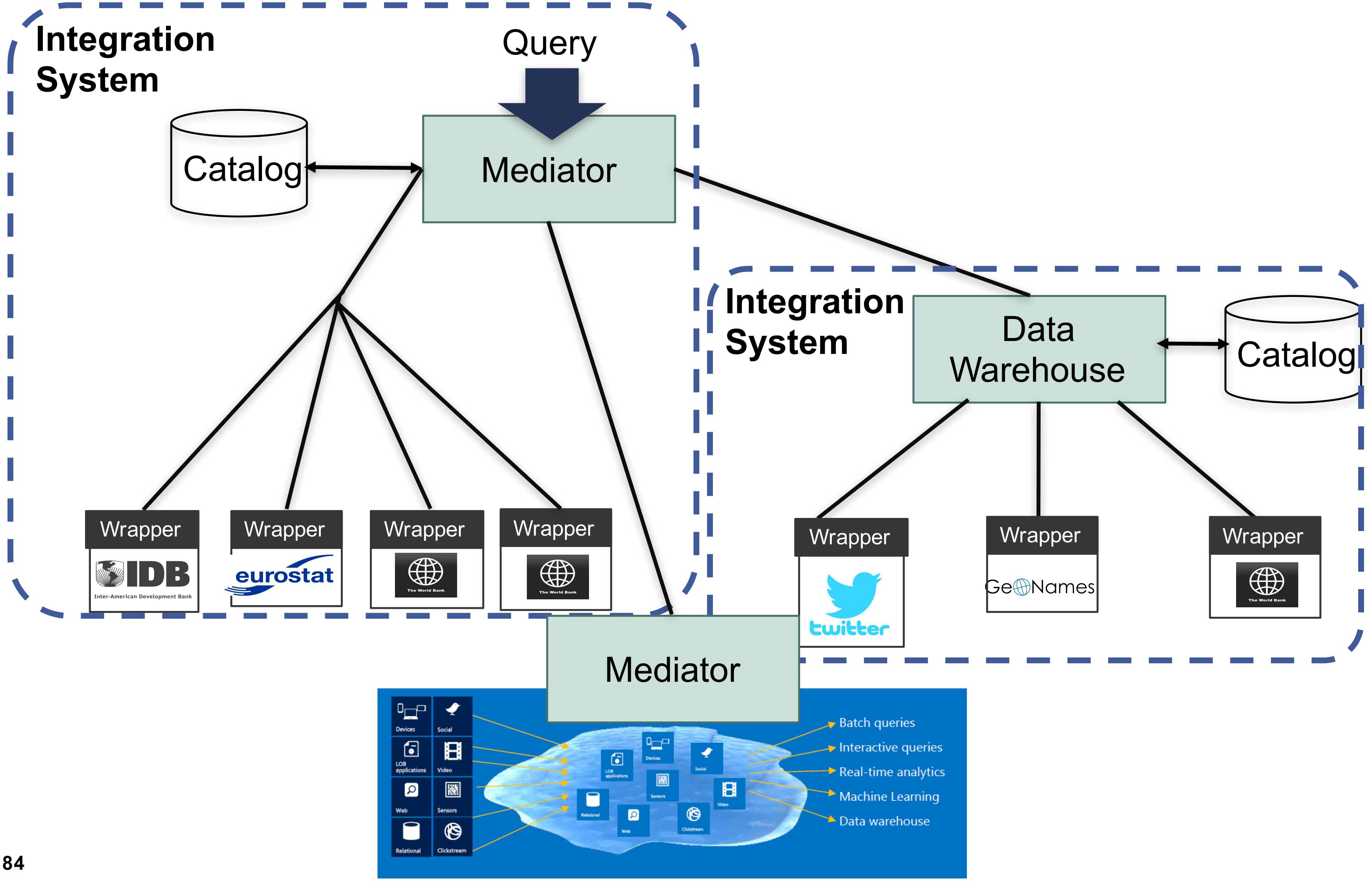
Materialized versus Virtual Access

The Mediator and Wrapper Architecture requires to access remote data sources on the fly

Materialized Data can be locally accessed.

Convenient whenever data is static

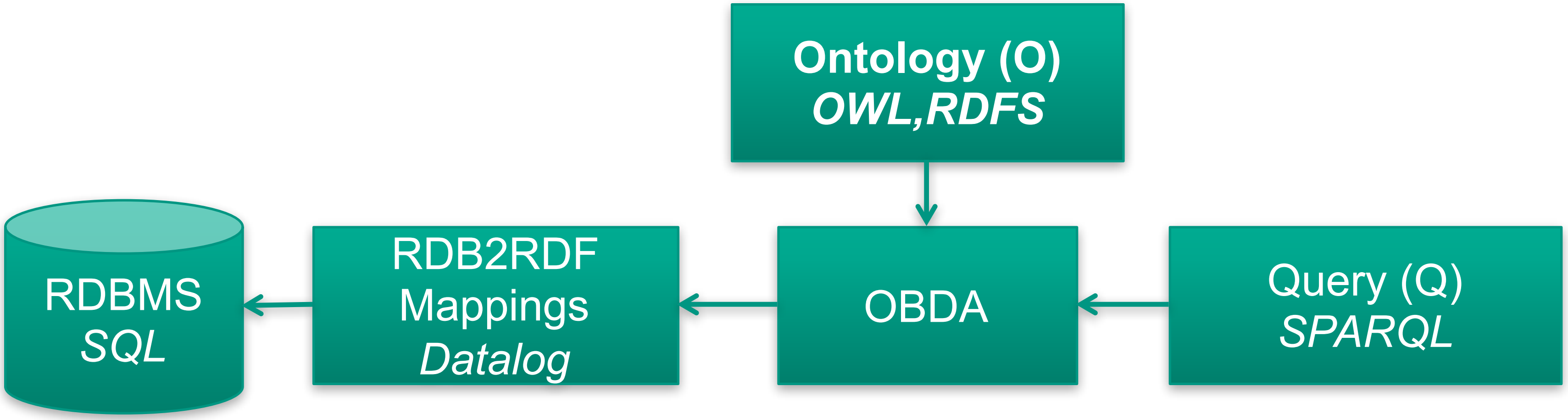
Hybrid Architectures



What is the role of Ontologies in Data Workflows/Data Integration Systems?

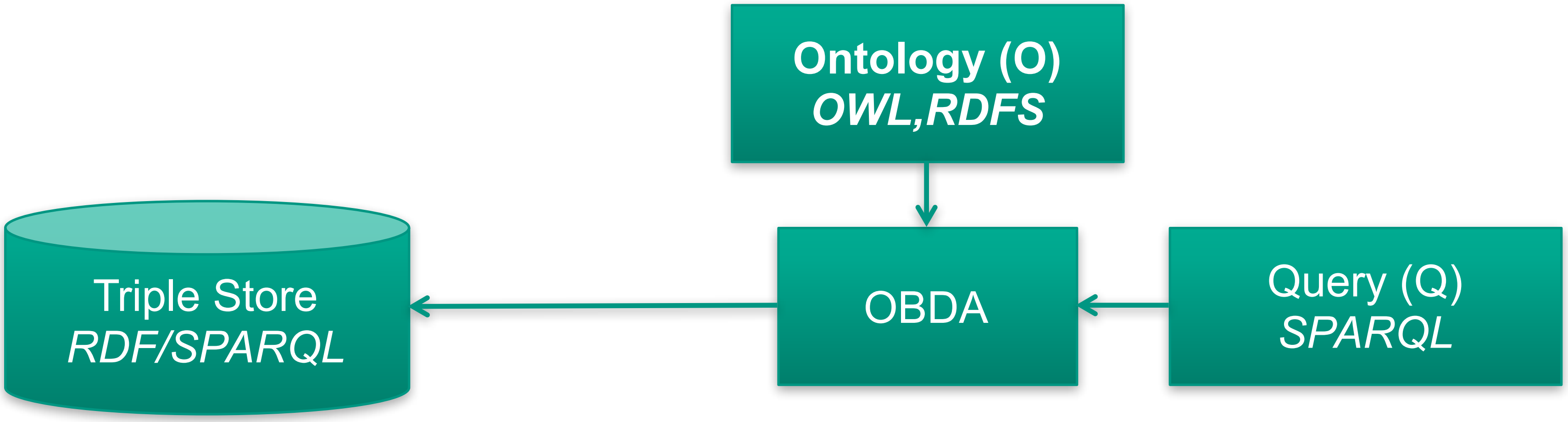
Linked Data integration using ontologies:

- 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))

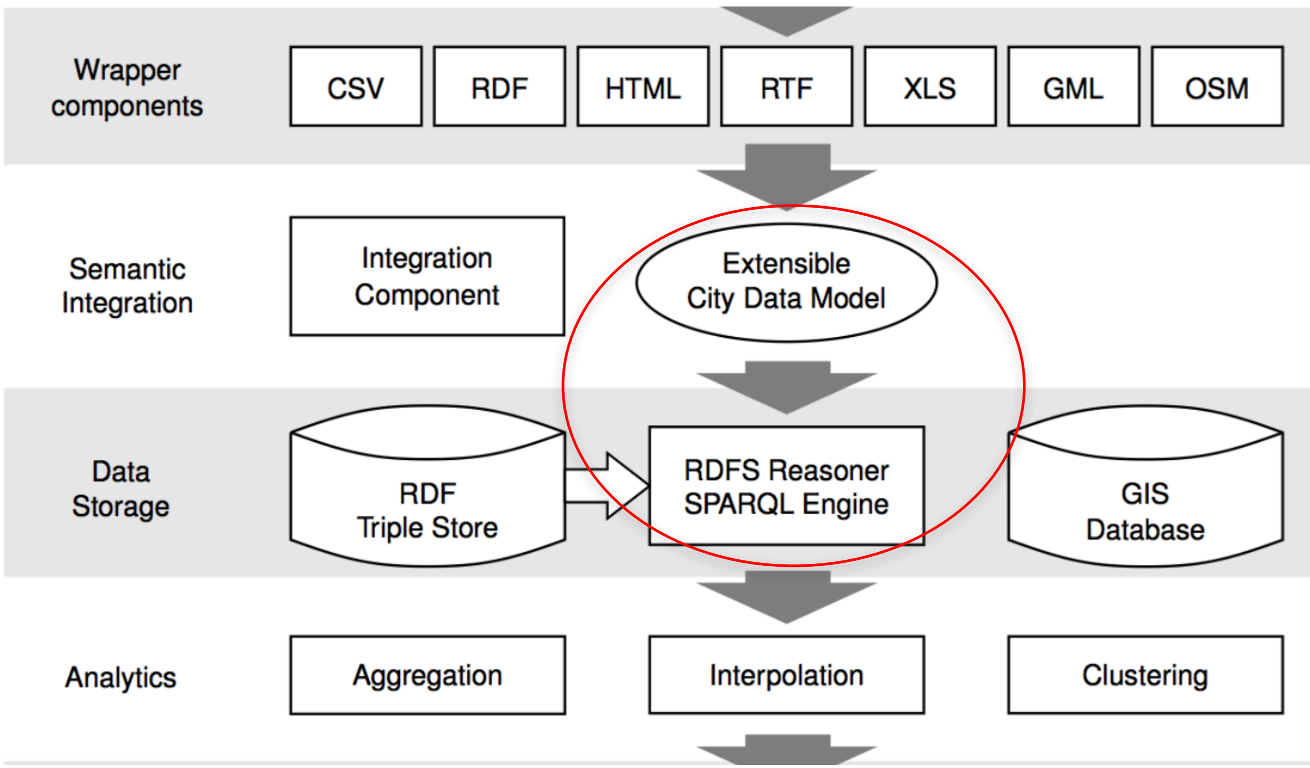


- For simplicity, let's leave out the Relational DB part, assuming Data is already in RDF...

Linked Data integration using ontologies (example)

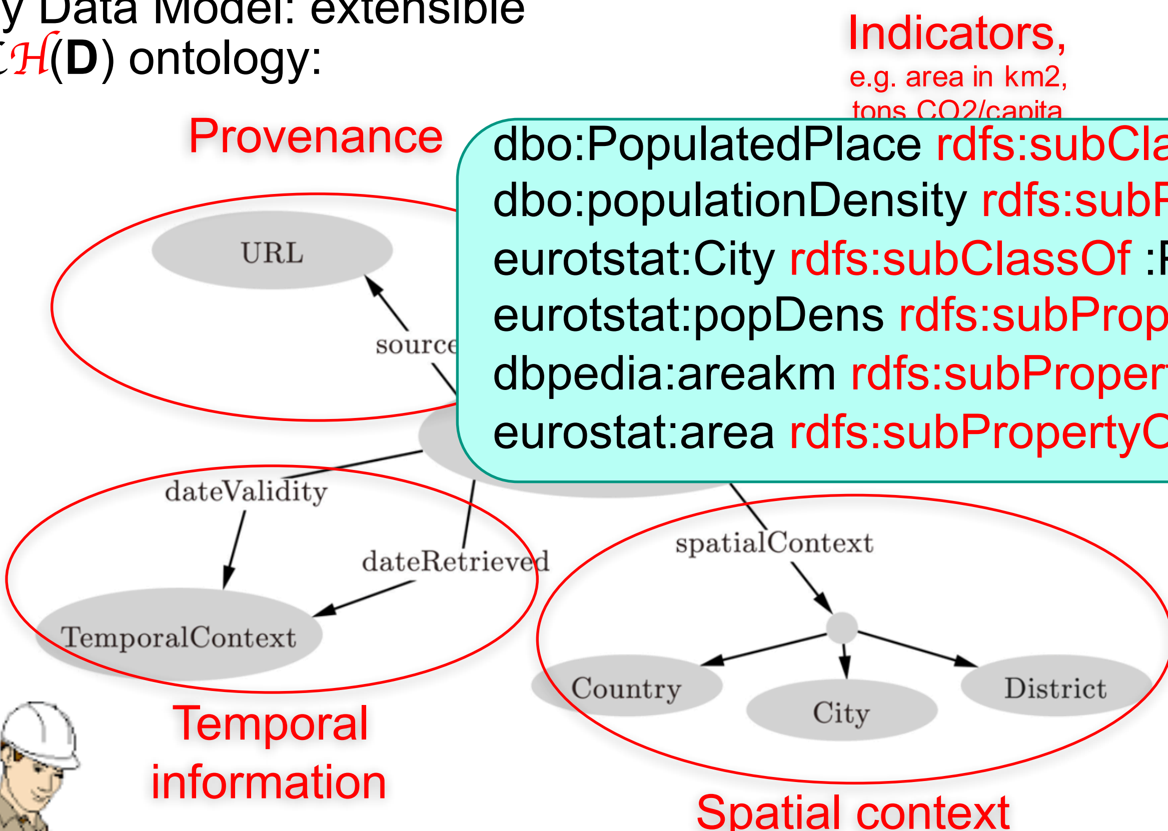
"Places with a Population Density below 5000/km2"?

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:



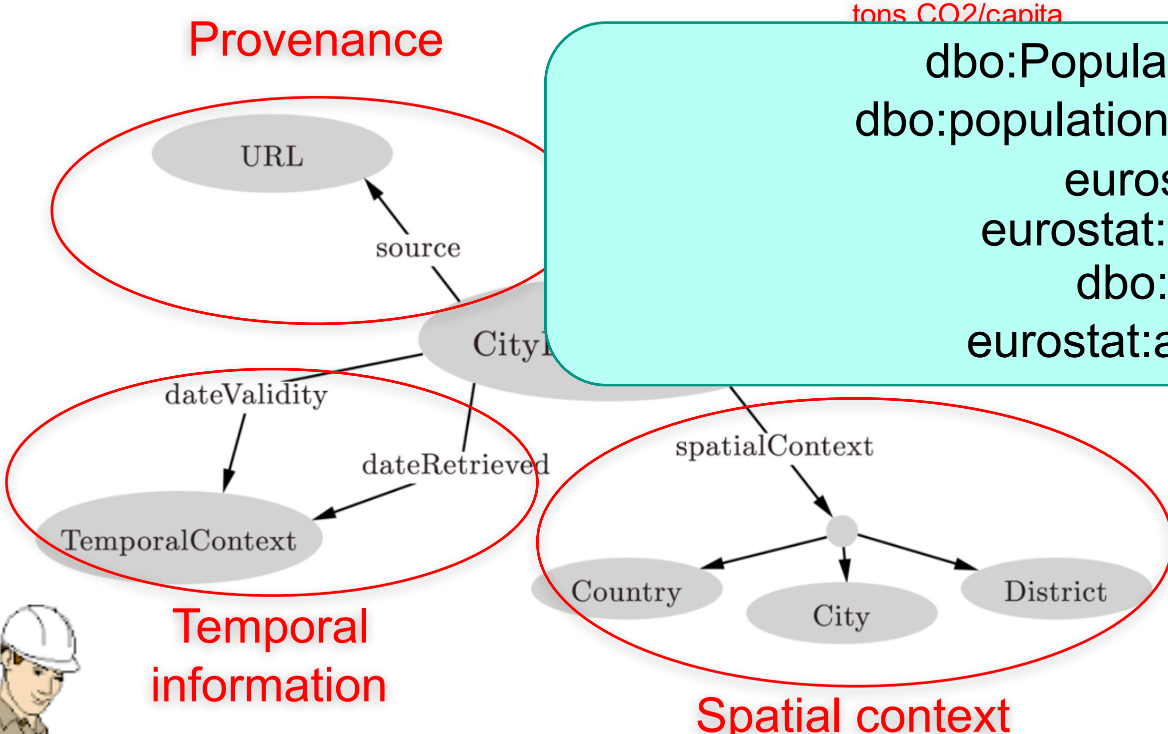
dbo:PopulatedPlace rdfs:subClassOf :Place.
dbo:populationDensity rdfs:subPropertyOf :populationDensity.
eurotstat:City rdfs:subClassOf :Place.
eurotstat:popDens rdfs:subPropertyOf :populationDensity.
dbpedia:areakm rdfs:subPropertyOf :area
eurostat:area rdfs:subPropertyOf :area

A concrete use case: The "City Data Pipeline"

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

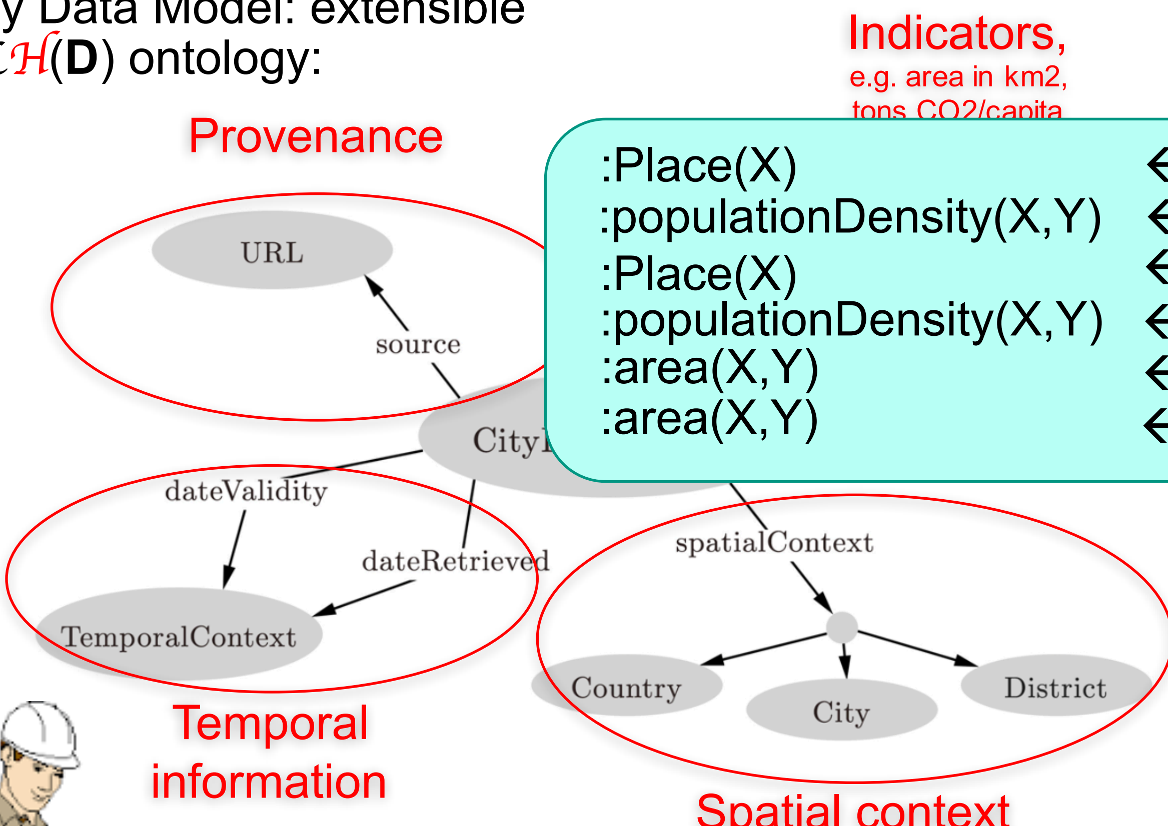
Indicators,
e.g. area in km2, tons CO2/capita

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



A concrete use case:
The "City Data Pipeline"

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



: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)

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 [Krötzsch, 2012, 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) }
```

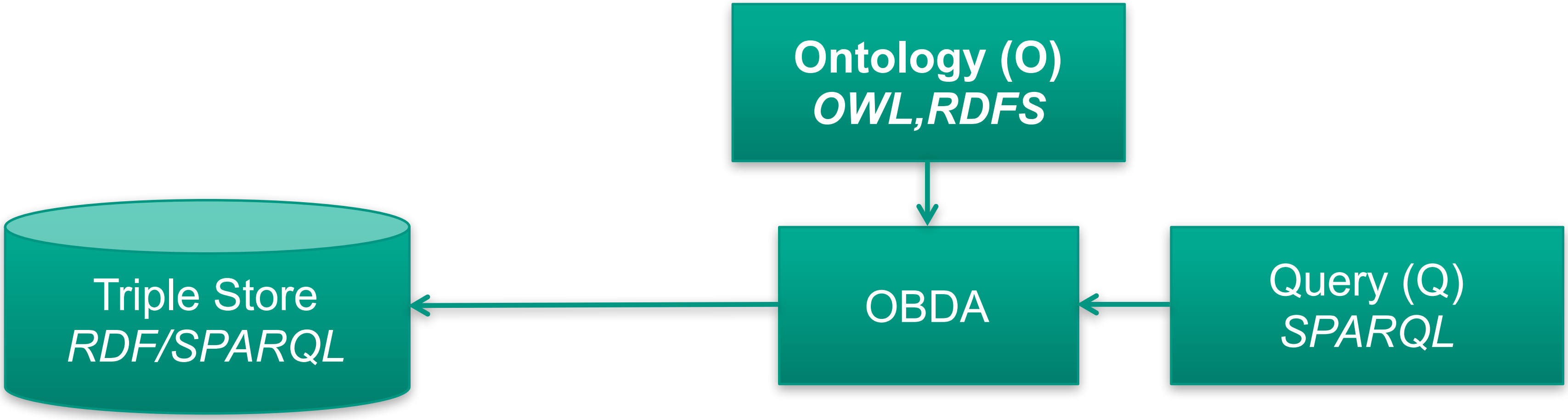
Approach 2: Query rewriting (input: conjunctive query (CQ) + Ontology output: UCQ)

```
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!

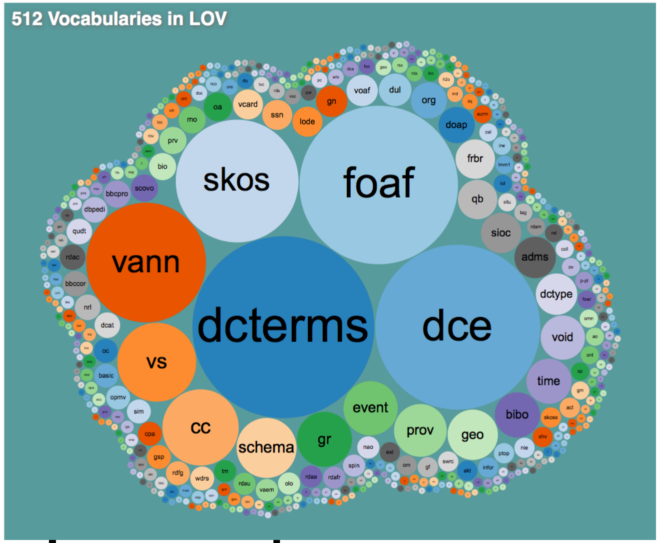
Where to find suitable ontologies?



Ontologies and mapping between Linked Data Vocabularies

- Good Starting points: 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]

Specific Steps (non-exhaustive, overlapping!)

- Extraction
- Inconsistency handling
- **Incompleteness handling** (sometimes called "Enrichment“, sometimes imputation of missing values…)
- Data Integration (alignment, source reconciliation)
- Aggregation
- Cleansing (removing outliers)
- Deduplication/Interlinking (could involve
- Analytics
- Enrichment
- Change dedection (Maintainance/Evolution)
- Validation (quality anaysis)
- Efficient, sometimes distributed (query) processing
- Visualization

Recall that slide from the beginning? What did we actually cover and where could Semantic Web techniques help?

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

Incompleteness Handling: Are RDFS and OWL 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}
```

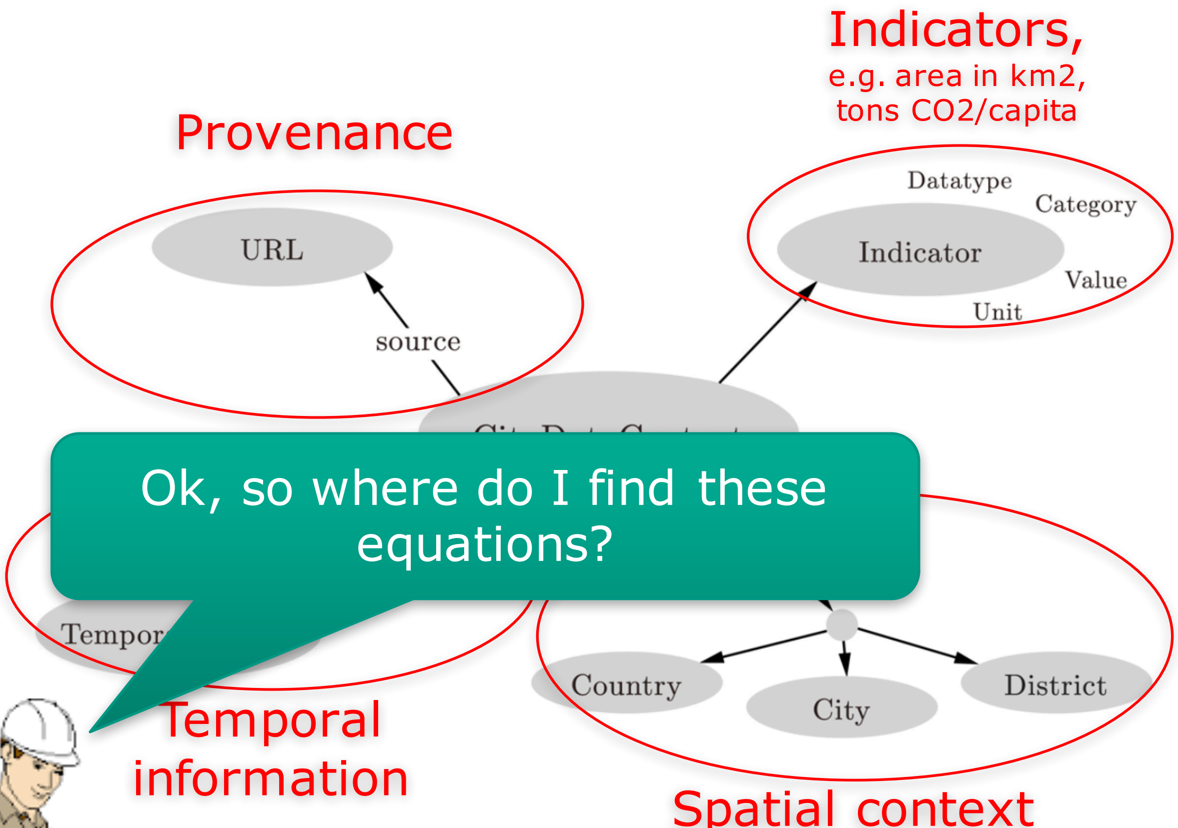
$$q(PD) \leftarrow (S, \text{popDensity}, PD)$$
$$q(PD) \leftarrow (S, \text{population}, P), (S, \text{area}, A), PD := P/A$$
~~$q(PD) \leftarrow (S, \text{popDensity}, PD'), (S, \text{area}, A'), (S, \text{area}, A), PD := P/A, P := PD' * A'$~~

.. 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"



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/

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.

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

Ontology of units of Measure (OM)

search concepts in this ontology

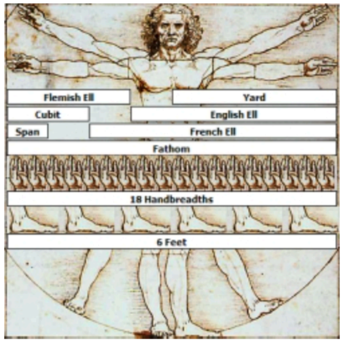
OK

download this ontology

RDF/XML OK

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

Incompleteness Handling: Are RDFS and OWL and equations enough?

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

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

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]

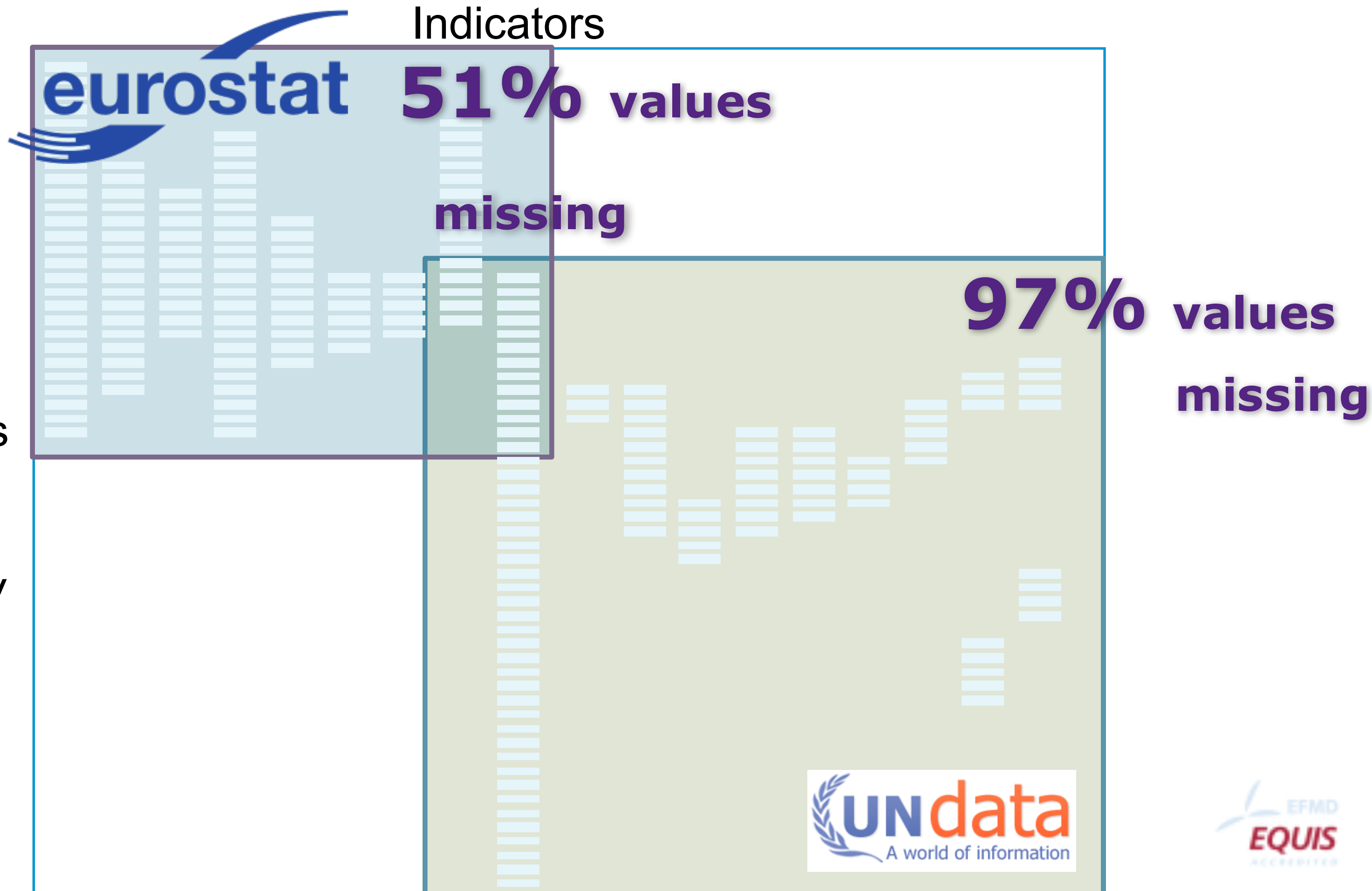
But Eurostat data is incomplete... I don't have the avg. income for all provinces or countries in the EU!

Hmmm...we actually need Claudia!

Integrated Open Data is (too?)sparse

Challenges – Missing values [Bischof et al. 2015]

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

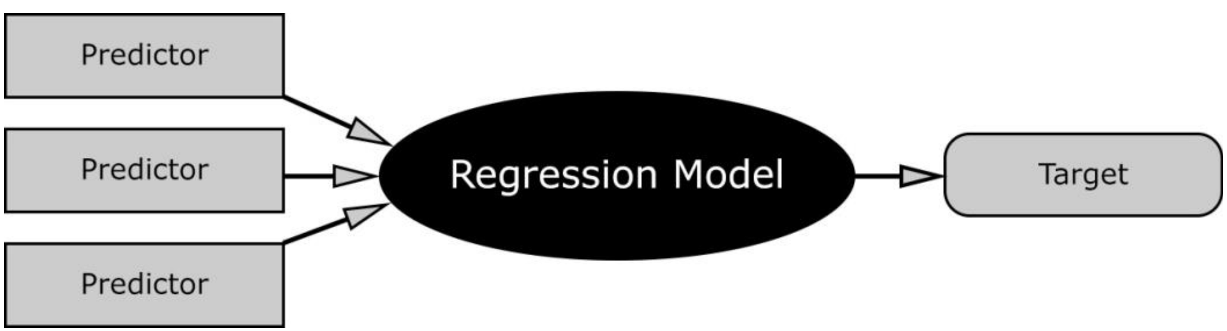


We don't get very far here with equations… Let's try Data Mining/ML!

Missing Values – Hybrid approach choose best imputation method per indicator [Bischof et al. 2015]

- 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)
- Let’s pick the “best method per indicator:

Validation: 10-fold cross validation



However: many/most machine learning methods need more or less complete training data! More trickery needed, cf. e.g. [Bischof et al. 2015] … or ask Claudia ☺

Specific Steps (non-exhaustive, overlapping!)

- Extraction
- Inconsistency handling
- Incompleteness handling (sometimes called "Enrichment“, sometimes imputation of missing values…)
- Data Integration (alignment, source reconciliation)
- Aggregation
- Cleansing (removing outliers)
- **Deduplication**/Interlinking (could invol
- Analytics
- Enrichment
- Change dedection (Maintainance/Evolution)
- Validation (quality anaysis)
- Efficient, sometimes distributed (query) processing
- Visualization

Last but not least…**Really** Don’t forget the basic steps, e.g.

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

Duplicates/Ambiguities:

EDITION US

THE HUFFINGTON POST

INFORM • INSPIRE • ENTERTAIN • EMPOWER

NEWS POLITICS ENTERTAINMENT WELLNESS WHAT'S WORKING VOICES VIDEO ALL SECTIONS

- http://www.huffingtonpost.com/2013/10/29/12-places-with-the-same-n_n_4170470.html

TRAVEL

22 Places That Have The Same Names But Are Actually Absurdly Different

10/29/2013 07:25 am ET | **Updated** Oct 29, 2013

180

Suzy Strutner
Associate Lifestyle Editor, The Huffington Post

GETTY

We were heartbroken this weekend when we heard of a woman who had used her late husband's airline miles to book a dream trip to the Spanish city of Granada, only to realize mid-flight that her plane was instead bound for the tropical island nation of Grenada, waaay out in the Caribbean Sea.

The incident got us thinking about other destinations that sound the same but are actually totally different and put travelers at risk for similar mix-ups. Consider this a public service announcement, people, and take notes.

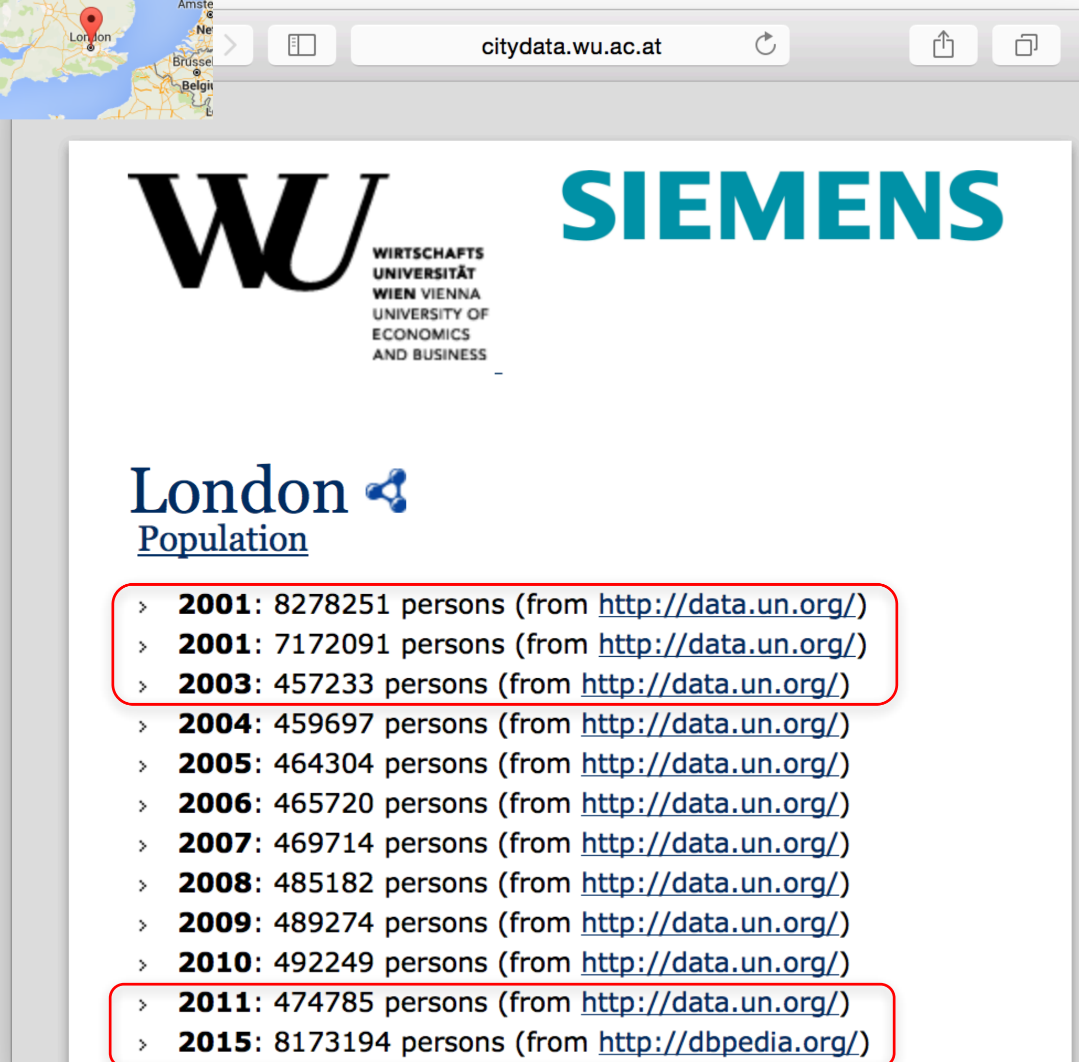
Granada, Spain vs. the nation of Grenada Ok, for the record: Granada is the city in Spain that is home to the historic Alhambra. Grenada is an island grouping in the Caribbean Sea.

Ambiguities/Inconsistencies affected also some older versions of our City Data Pipeline:

- This example on the right was due to naïve object consolidation/deduplication, **BUT:**

- Open Data is often incomparable/inconsistent in itself (e.g. across years the method of data collection might change)

→ inconsistencies **across** and **within** datasets are common



A concrete use case:
The "City Data Pipeline"

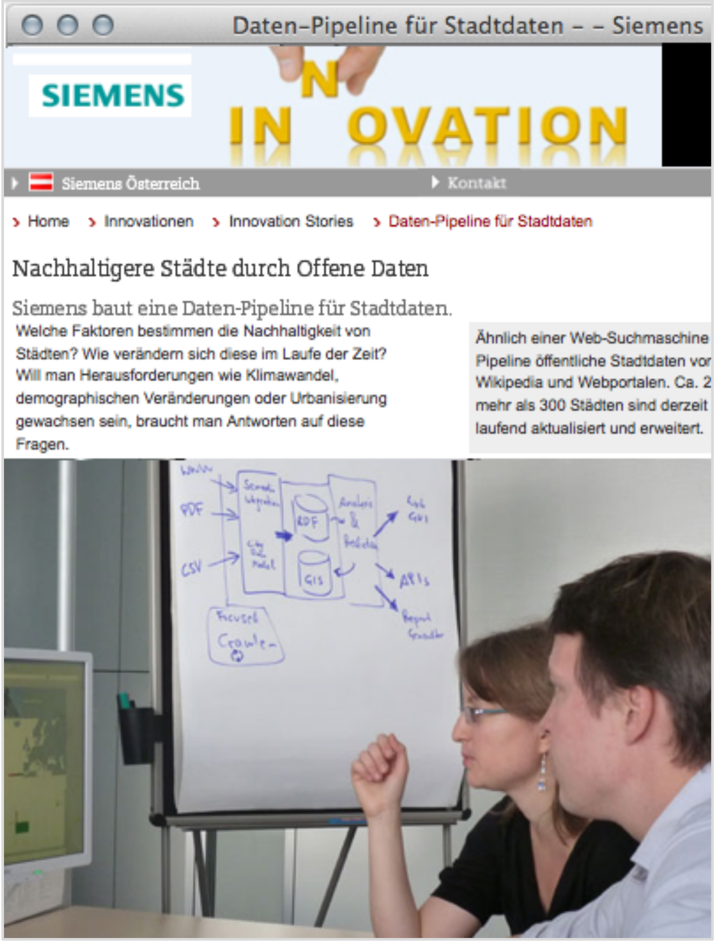
Idea – a "classic" Semantic **Web** use case!

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

citydata.wu.ac.at

(How) can ontologies help me?

- Are ontology languages expressive enough?
- Which ontologies could I (re-)use?
- Is there enough data at all?
- **Where to find the right data?**
- Where to find the right ontologies?
- How to tackle inconsistencies?

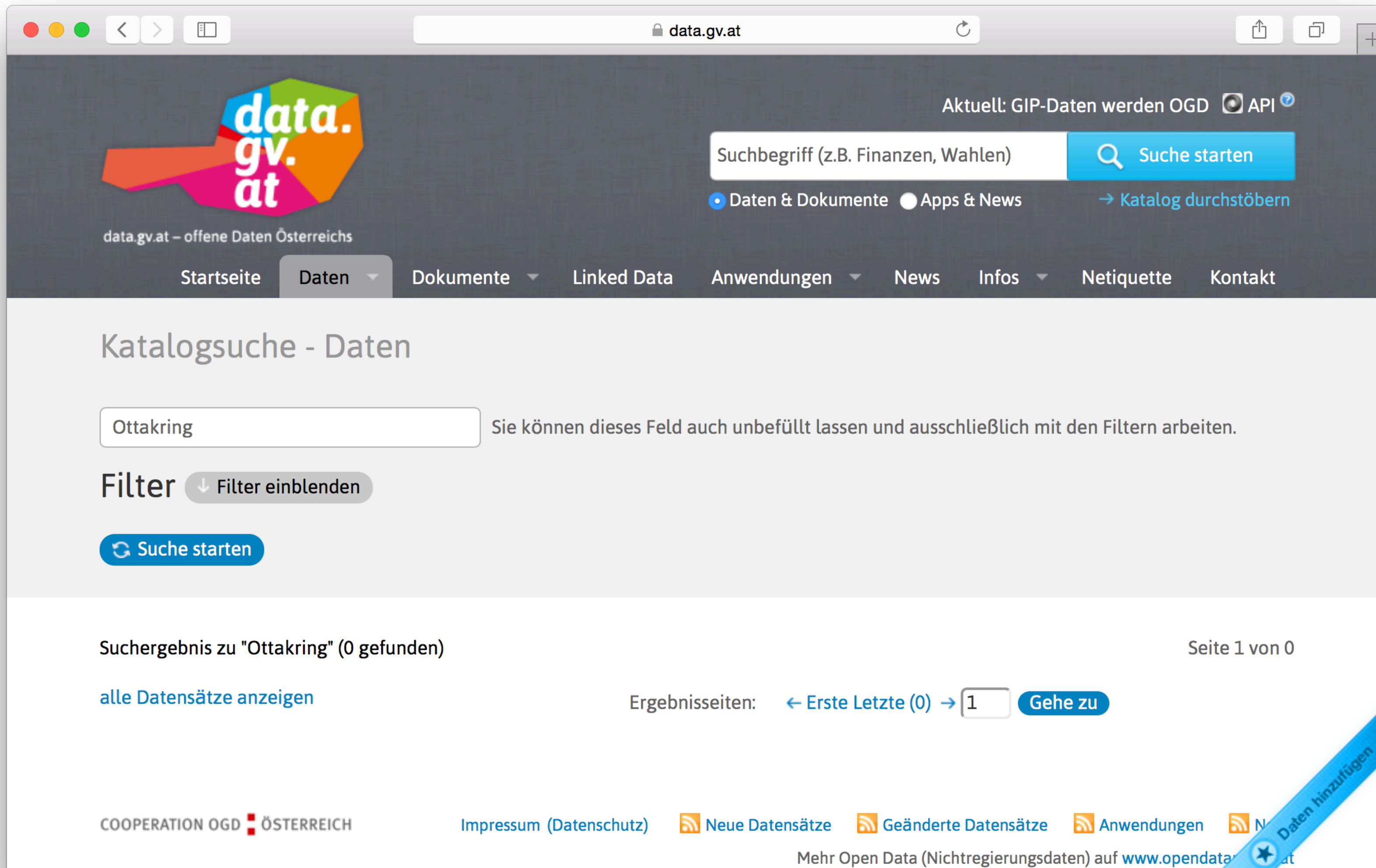


Where to find the <u>data</u>?

- Bad news:
  - Finding suitable **ontologies** to map data sources to is not the only challenge:
  - Foremost… even before a Data workflow starts, a main challenge is to find the right Datasets/Resources
  - Semantic Web Search engines... Failed? ☹
    - https://www.w3.org/wiki/Search_engines
- ... The obvious entry point:
  - **Open Data portals**
    - Still quite messy cf. http://data.wu.ac.at/portalwatch/
    - Different formats, encodings, metdata of varying quality
    - No proper Search!
- … but again: Semantic Web Technologies **could** help here!

No reason not to try again and succeed this time! ☺

Open Data Portal search is a big problem... Why?



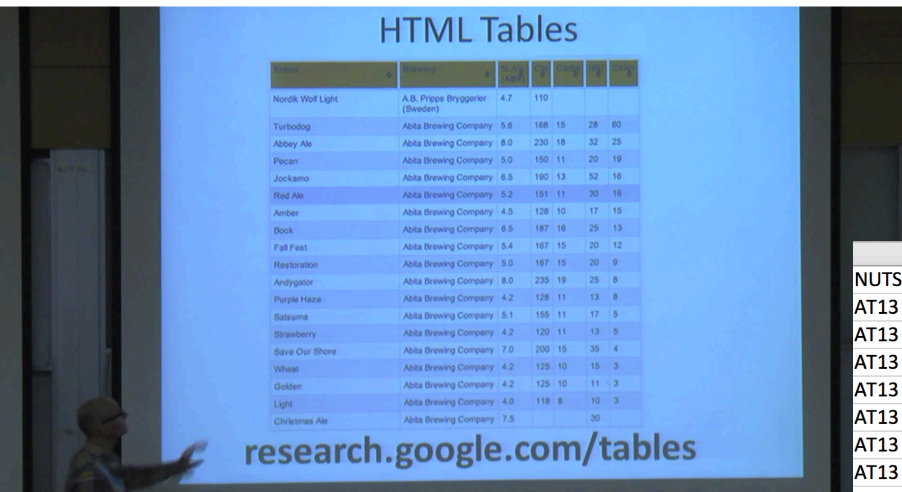
How to search in/for Open Data?

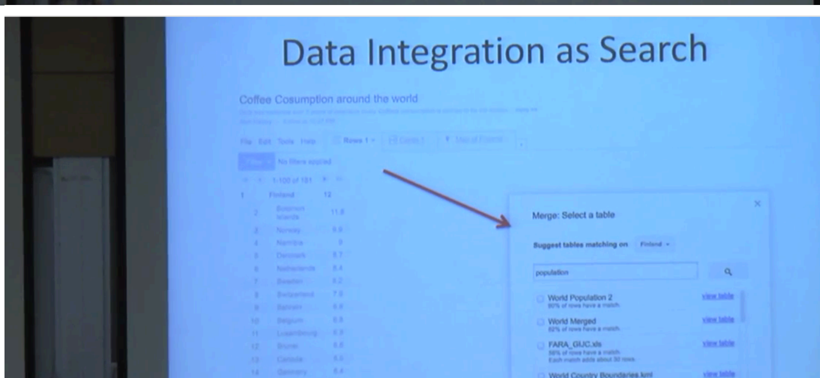
https://www.youtube.com/watch?v=kCAymmbYIvc

Cf. Work on structured Data in Web Search by Alon Halevy

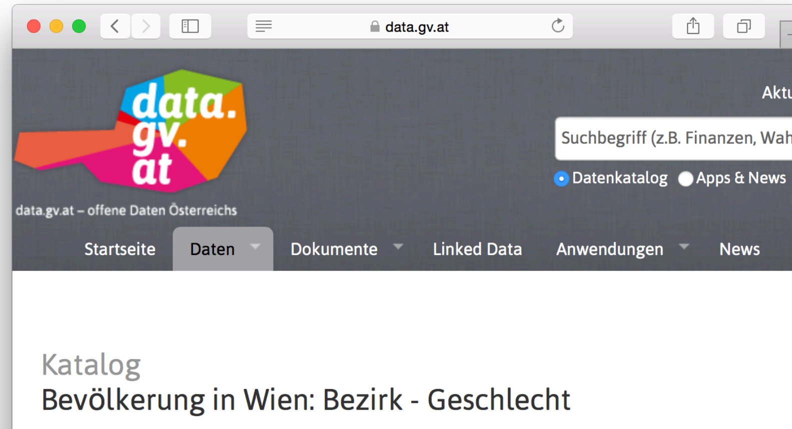
... BTW: google has partially given it up on it it seems.

→ *Some more recent work in a SW & Open Data context: [Neumaier et al., 2015+2016] [Ramnandan et al. 2015] cf. also mini-projects!*





VS.



B	C	D	E	F	G	H	I
NUTS2	NUTS3	DISTRICT_CODE	SUB_DISTRICT_CODE	POP_TOTAL	POP_MEN	POP_WOMEN	REF_DATE
AT13	AT130	90101	0	16131	7726	8405	01.01.2014
AT13	AT130	90201	0	99597	48650	50947	01.01.2014
AT13	AT130	90301	0	86454	41085	45369	01.01.2014
AT13	AT130	90401	0	31452	14903	16549	01.01.2014
AT13	AT130	90501	0	53610	26299	27311	01.01.2014
AT13	AT130	90601	0	30613	14833	15780	01.01.2014
AT13	AT130	90701	0	30792	14703	16089	01.01.2014
AT13	AT130	90801	0	24279	11855	12424	01.01.2014
AT13	AT130	90901	0	40528	19286	21242	01.01.2014
AT13	AT130	91001	0	186450	91638	94812	01.01.2014
AT13	AT130	91101	0	93440	45541	47899	01.01.2014
A						7122	01.01.2014
A						7770	01.01.2014
A						5870	01.01.2014
A						7571	01.01.2014
A						0643	01.01.2014
A						7894	01.01.2014
A						6157	01.01.2014
AT13	AT130	91901	0	69242	31849	37393	01.01.2014
AT13	AT130	92001	0	84305	41200	43105	01.01.2014
AT13	AT130	92101	0	148947	71633	77314	01.01.2014

Compared to Web (Table) search…

a) This looks like a slightly different problem…

b) Can linking to "Open" knowledge graphs help? (wikidata, dbpedia?) … Probably.

CONCLUSIONS

Conclusions

Heterogeneous Web Sources

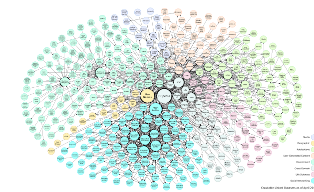
EUROPEAN DATA PORTAL

eurostat

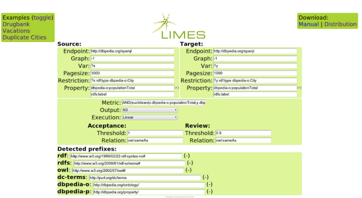
IDB

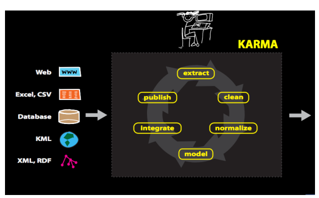
Inter-American Development Bank

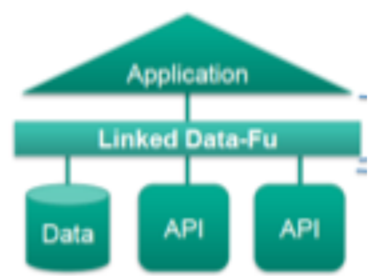
The World Bank

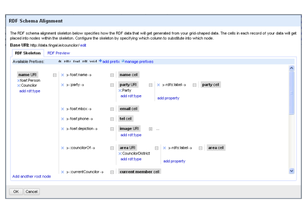


Tools & Pipelines to Access/Integrate Web Sources

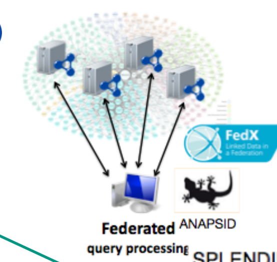








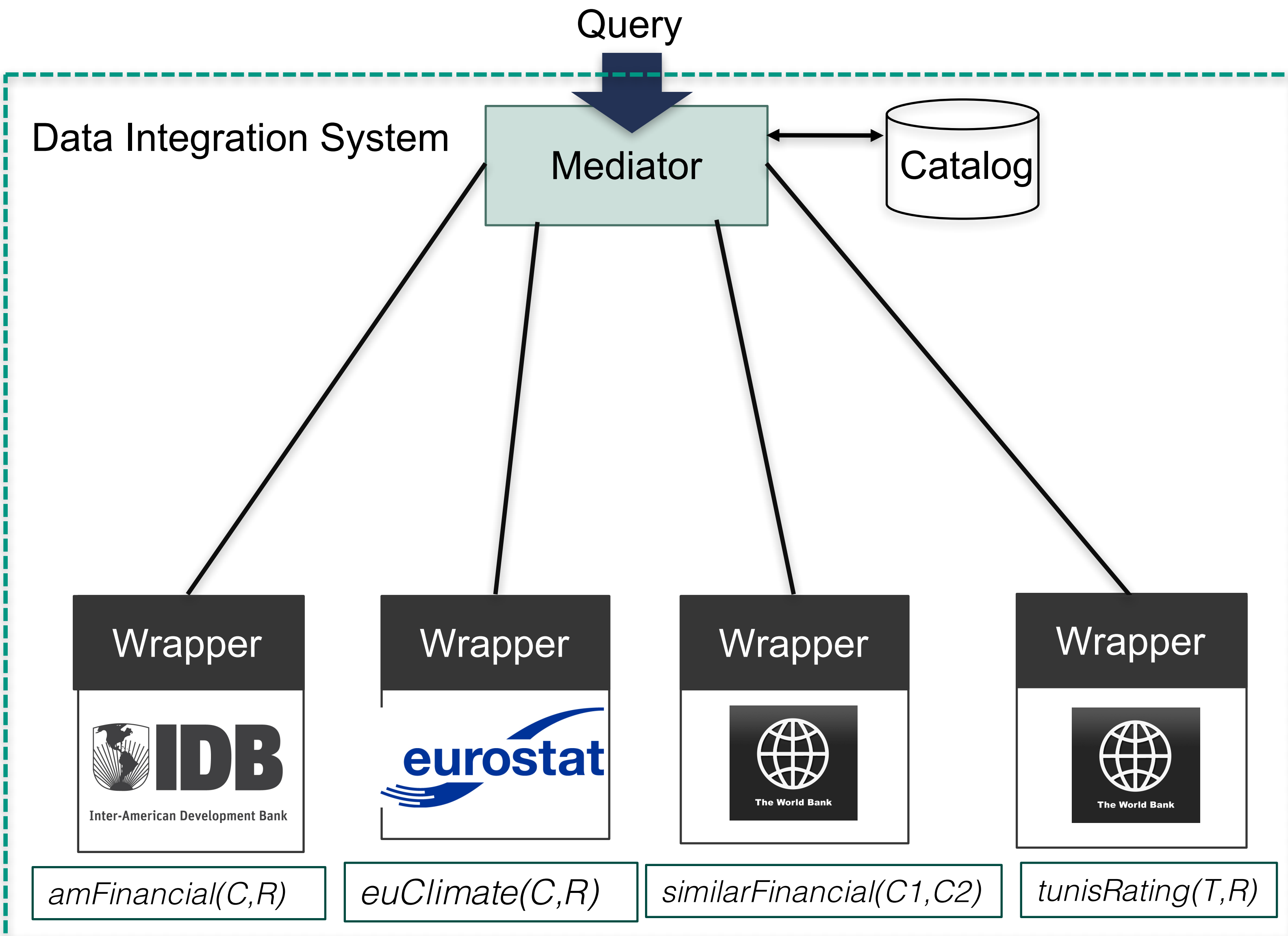
RDB2RDF Systems



CSV2RDF Systems

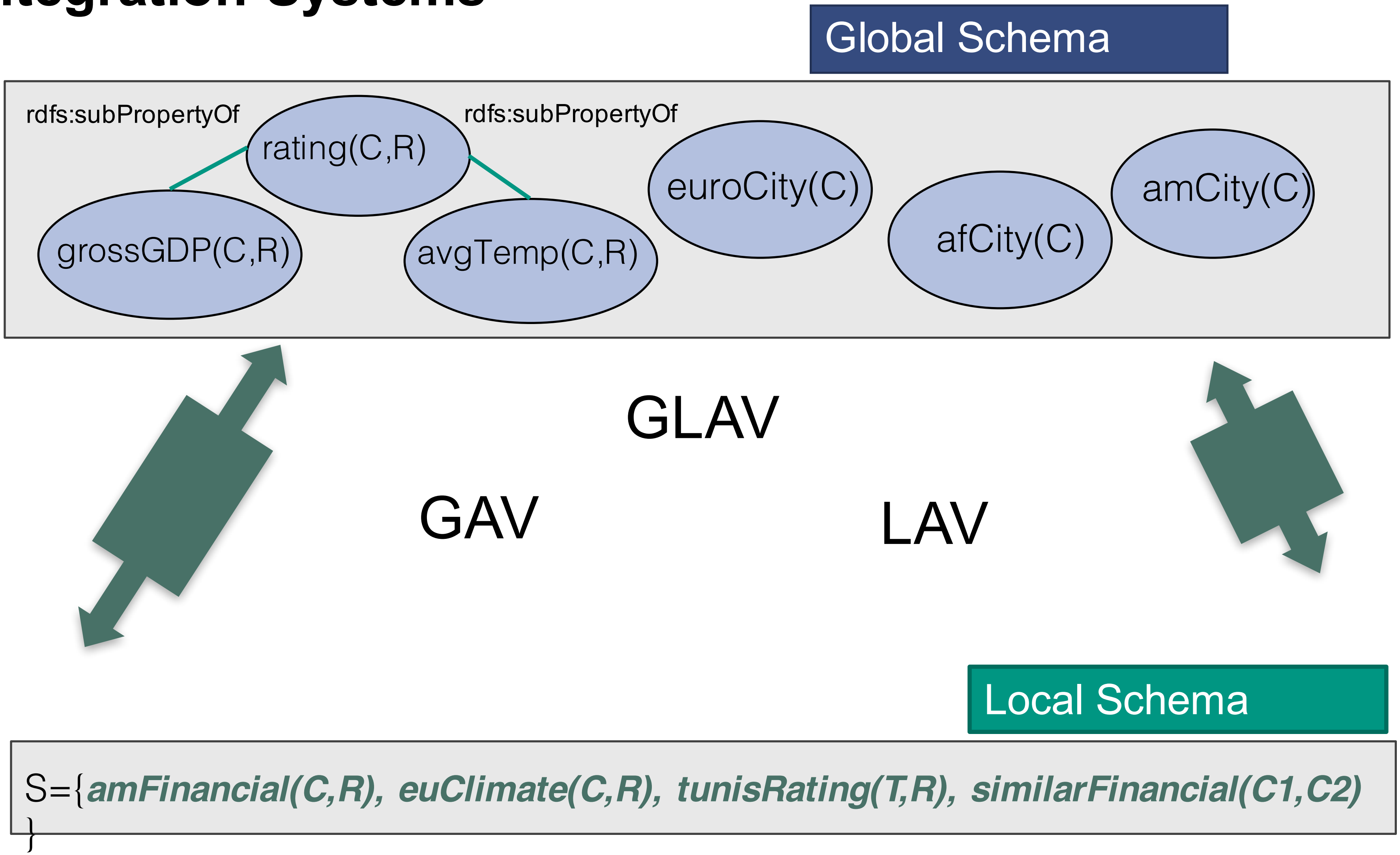
W3C®

Conclusions



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

Integration Systems



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 AND more recently work on OBDA
- Non-Clean Data requires: Statistics & machine learning (outlier detection, imputing missing values, resolving inconsistencies, etc.)

- Despite 15 years into Semantic Web:“Finding the right data“ remains a major challenge!

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
- [Gentile, et al. 2016] Anna Lisa Gentile, Sabrina Kirstein, Heiko Paulheim and Christian Bizer. Extending RapidMiner with Data Search and Integration Capabilities
- [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)
- [Corby et al. 2015] Olivier Corby, Catherine Faron-Zucker, Fabien Gandon: A Generic RDF Transformation Software and Its Application to an Online Translation Service for Common Languages of Linked Data. International Semantic Web Conference (2) 2015: 150-165
- [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, Patrik Schneider: Collecting, Integrating, Enriching and Republishing Open City Data as Linked Data. International Semantic Web Conference (2) 2015: 57-75

References 2

- [Doan et al. 2012] AnHai Doan, Alon Y. Halevy, Zachary G. Ives: Principles of Data Integration. Morgan Kaufmann 2012, ISBN 978-0-12-416044-6, pp. I-XVIII, 1-497
- [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
- [Montoya et al. 2014] 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).

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 on the Web with High Availability. ISWC2014

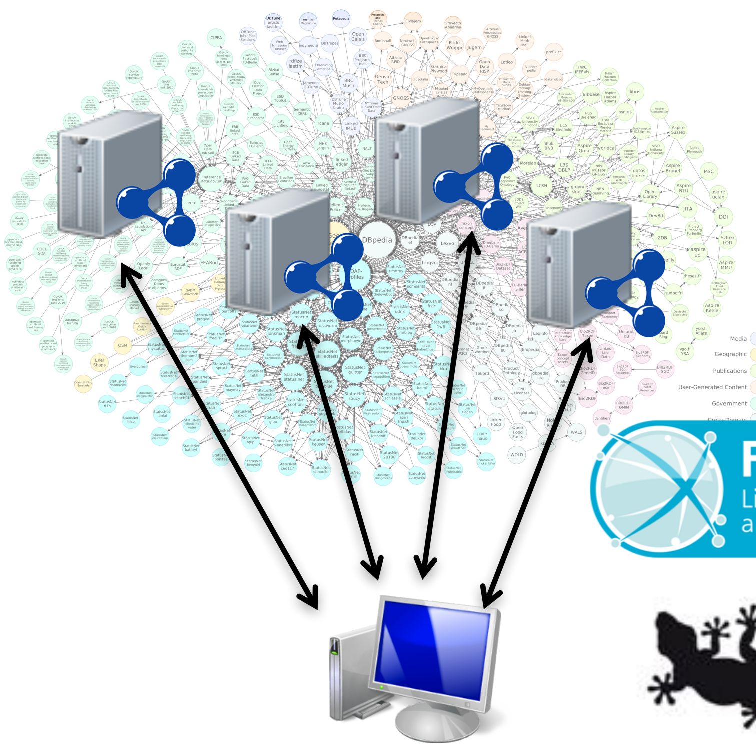
References 5

- [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)
- [Sarma et al. 2012] Das Sarma, A., Fang, L., Gupta, N., Halevy, A., Lee, H., Wu, F., Xin, R., Yu, C.: Finding related tables. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. pp. 817–828. ACM (2012)
- [Venetis et al. 2011] Venetis, P., Halevy, A.Y., Madhavan, J., Pasca, M., Shen, W., Wu, F., Miao, G., Wu, C.: Recovering semantics of tables on the web. PVLDB 4(9), 528–538 (2011)
- [Ramnandan et al. 2015] Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P.A.: Assigning semantic labels to data sources. In: ESWC 2015. pp. 403–417
- [Neumaier et al. 2015] Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment & evolution of open data portals. In IEEE International Conference on Open and Big Data, Rome, Italy, August 2015.
- [Neumaier et al. 2016] S. Neumaier, J. Umbrich, J. Parreira, A. Polleres. Multi-level semantic labelling of numerical values, ISWC2016, to appear.

TRENDS & OPEN RESEARCH QUESTIONS (SOME)

Federations of SPARQL Endpoints

Publicly available SPARQL endpoints



FedX
Linked Data in a Federation

ANAPSID

Federated
query processing

SPLENDID

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;
….

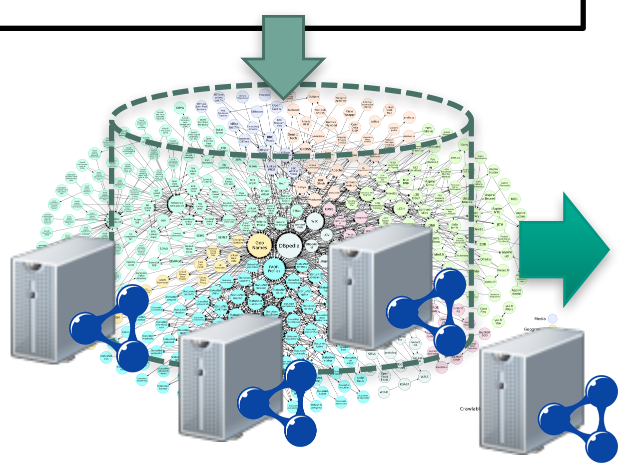
http://dbpedia.org/sparql := http://xmlns.com/foaf/0.1/name;
http://dbpedia.org/ontology/award;
http://dbpedia.org/ontology/almaMater;
http://www.geonames.org/ontology#name;
http://www.geonames.org/ontology#parentFeatures;
….

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;
…. .

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' .}
```

Federated Query Engine



{'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/'}

References

- Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, Michael Schmidt: FedX: Optimization Techniques for Federated Query Processing on Linked Data. International Semantic Web Conference (1) 2011: 601-616

  http://www2.informatik.uni-freiburg.de/~mschmidt/docs/iswc11_fedx.pdf
- 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

  http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Research_Paper/03/70310017.pdf

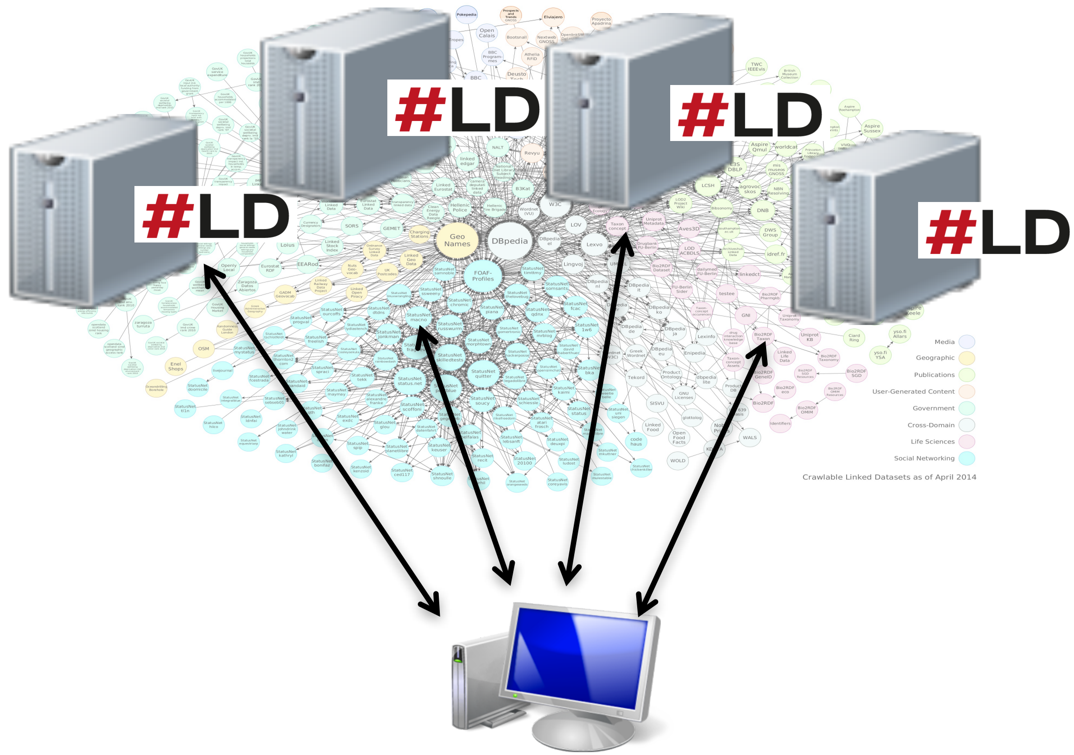
<u>RQ1:</u> Can a federation of SPARQL Endpoints be seen as a Data Integration System?

- Describe the problem presented in the related papers as a Data Integration System.
- Select the most suitable mapping approach to describe the Data Integration System.
- Use the mediator and wrapper architecture to describe the Data Integration System.
- Illustrate with an example the Data Integration System, and show the features implemented by the mediator and wrappers of the Data Integration System

SPARQL Query Execution using LAV views

Publicly available Linked Data Fragments (LAV views)

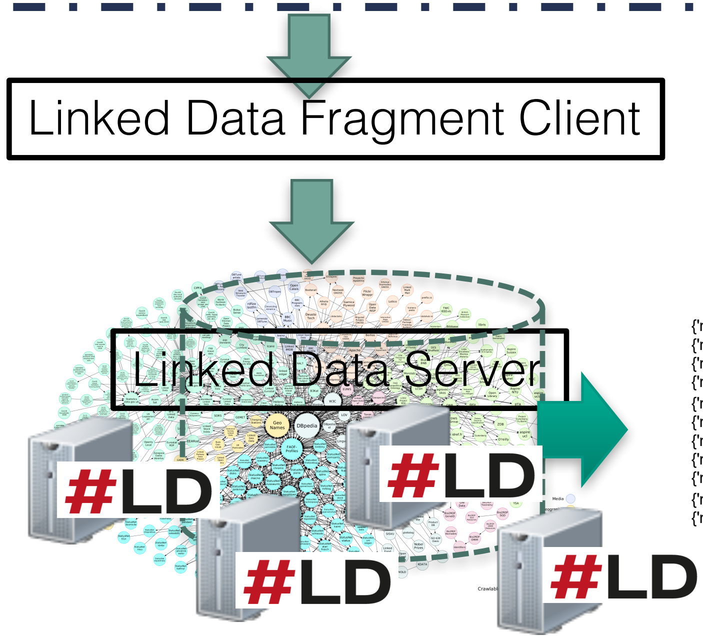
Linked Data Fragment Server



Linked Data Fragment Client

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/'}
```

References

SPARQL Query Execution using Linked Data Fragments

- Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, Pieter Colpaert:

  Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. J. Web Sem. 37: 184-206 (2016)

  http://www.sciencedirect.com/science/article/pii/S1570826816000214

- Maribel Acosta, Maria-Esther Vidal: Networks of Linked Data Eddies: An Adaptive Web Query Processing Engine for RDF Data. International Semantic Web Conference (1) 2015: 111-127

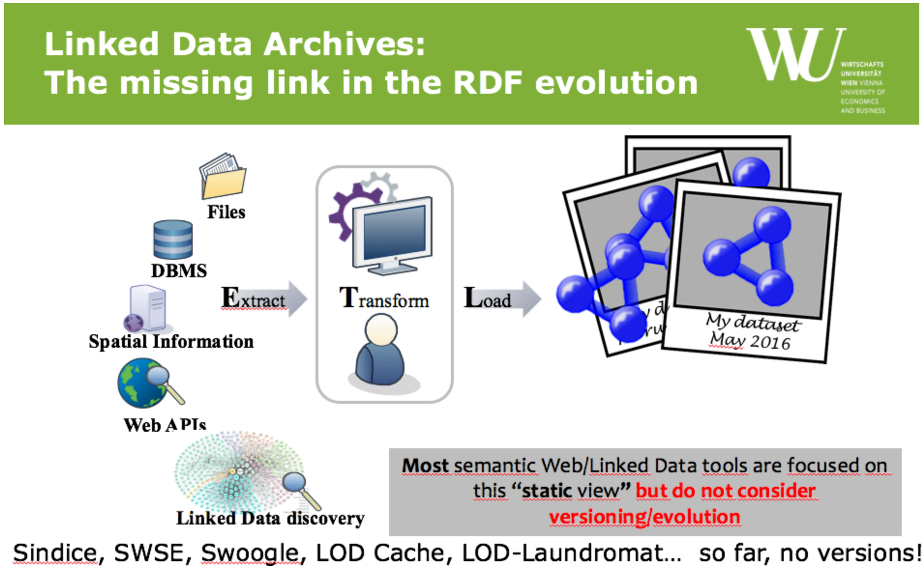
  http://www.aifb.kit.edu/images/f/f0/Acosta_vidal_iswc2015.pdf

<u>RQ2:</u> Can a federation of Linked Data Fragments be seen as a Data Integration System?

- Describe the problem presented in the related papers as a Data Integration System.
- Select the most suitable mapping approach to describe the Data Integration System.
- Use the mediator and wrapper architecture to describe the Data Integration System.
- Illustrate with an example the Data Integration System, and show the features implemented by the mediator and wrappers of the Data Integration System

RQ3: What challenges does archiving of RDF and Open Data involve?

- If Open Data is Big Data, **archiving** Open Data and RDF Data is even one order of magnitude more!
- Challenges on creating (crawling), maintaining, storing and querying such archives:
- cf. slides
- "On Archiving Linked and Open Data" at the 2nd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2016),



http://polleres.net/presentations/20160530Keynote-MEPDaW2016.pptx

<u>RQ4:</u> How to publish and use Linked Open Data alongside Closed Data?

- Which policies need to be supported?
- How to describe these policies?
- How to enforce them, how to protect and securely store closed linked data?
- Surprisingly few starting points in **our** community on access control/encryption for RDF/Linked Data, cf. e.g.
  - **S. Kirrane. Linked data with access control. PhD thesis, 2015. NUI Galway https://aran.library.nuigalway.ie/handle/10379/4903**
  - Mark Giereth: On Partial Encryption of RDF-Graphs. International Semantic Web Conference 2005: 308-322
- Lots of work on policy languages, e.g. ODRL:
  - Simon Steyskal, Axel Polleres:
    Towards Formal Semantics for ODRL Policies. RuleML 2015: 360-375
  - Simon Steyskal, Axel Polleres: Defining expressive access policies for linked data using the ODRL ontology 2.0. SEMANTICS 2014: 20-23

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

- Work on **one** of the overall Research Questions (**too generic on purpose!!!!**) RQ1-RQ6 from the slides before **in your mini-project groups!**
- **4 questions/11 groups → 1 RQ can be chosen by at most 3 groups!**
  - **RQ1-2 → Maria Esther**
  - **RQ3-4 → Axel**

For each problem you work on:

1) **Problems**: Why is it difficult? Find obstacles. Define concrete open (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** presentation per group addressing these 4 questions and findings.

Tips:

- Think about how much time you dedicate to which of these four questions.
- **Don't** start with 3)
- Prepare some answers or discussions for afinal plenary session which can be presented in a **2-3 min pitch SUMMARIZING your discussion**
  - **no more than 2 slides**
  - **focus on take-home messages**

→ *Please email your* ***notes*** *and (link to) slides to axel[at]polleres.net ...*
We will review them and provide feedback during tmrw morning!