Scalable Extraction of Implicit and Explicit Schema Information in Linked Open Data

Ansgar Scherp

Data and Web Science, U Mannheim

Matthias Konrath, Thomas Gottron

Web Science and Technologies, U Koblenz

UNIVERSITÄT MANNHEIM

Linked Data

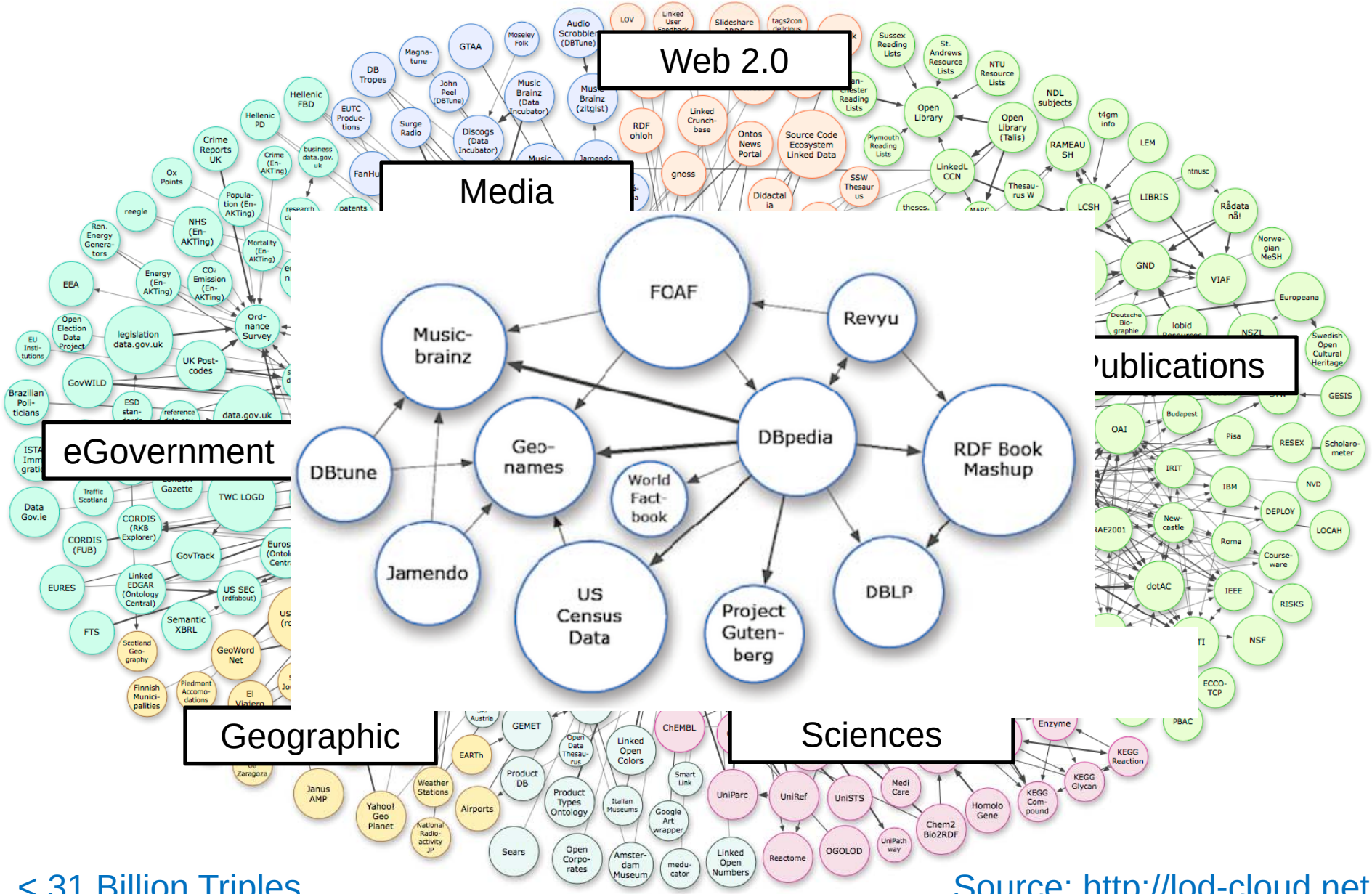
- We witness a major movement in the Web …
- Publishing and interlinking of data of different quality, purpose and source on the Web
- Technology + Social Phenomenon

World Wide Web	Linked Data
Documents	Data
Hyperlinks	Typed Links
HTML	RDF
Addresses (URIs)	Addresses (URIs)

Relevance of Linked Data?

schema.org

Linked Data: May ‘07 → Sept. ‘11



< 31 Billion Triples

Source: http://lod-cloud.net

Linked Data: Based on Four Principles

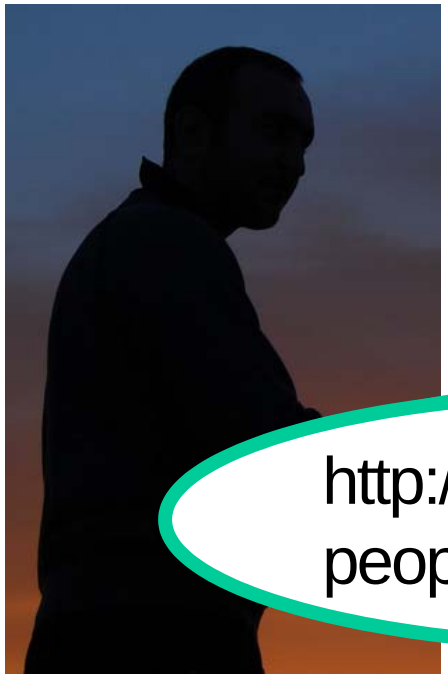
1. Identification
2. Interlinkage
3. Dereferencing
4. Description

1. Use URIs for Identification

2. Interlinking of Ressources

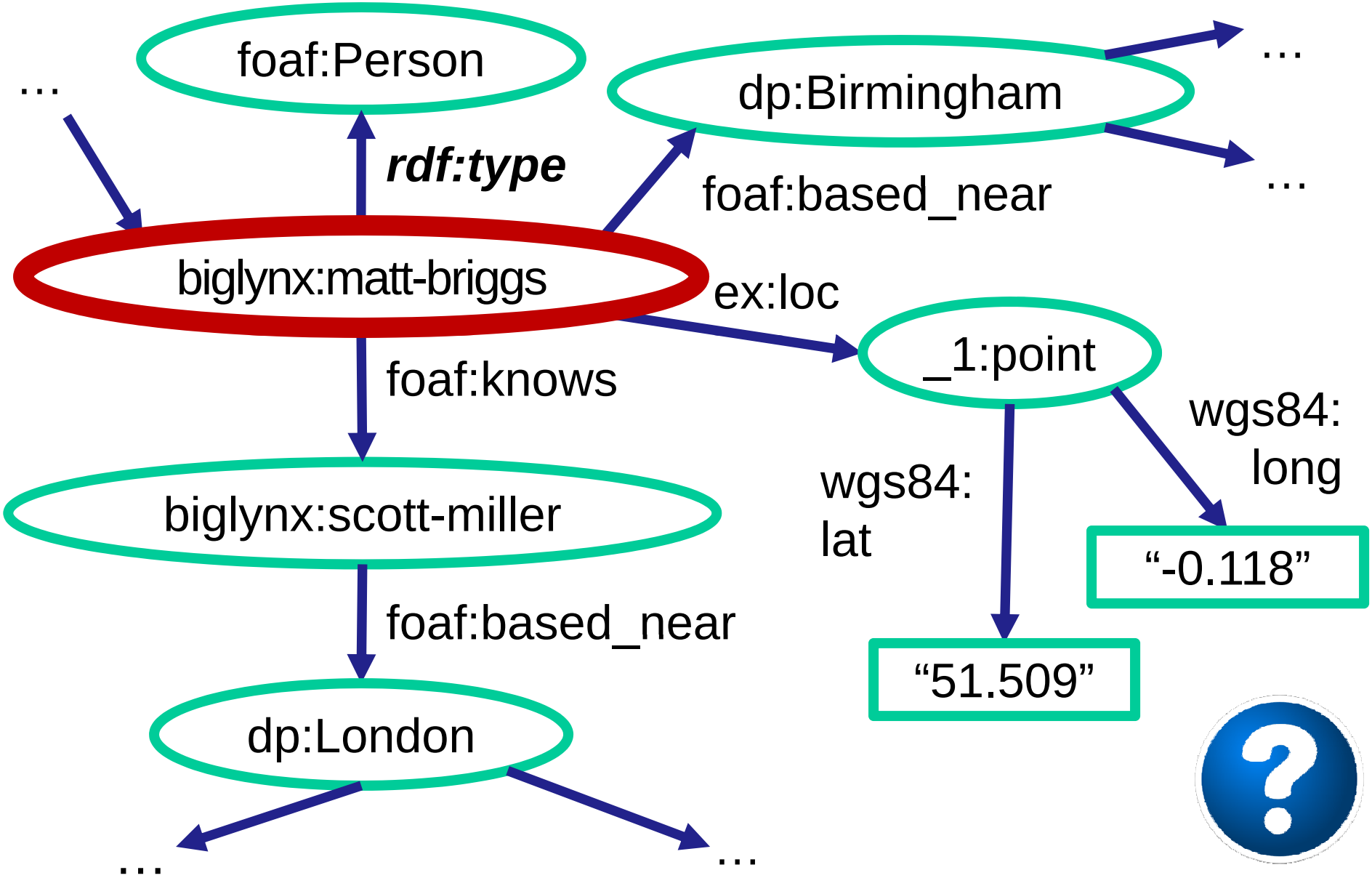
3. Dereferencing of URIs

- Pretty easy to look up web documents
- How do we “look up” things of the real world?

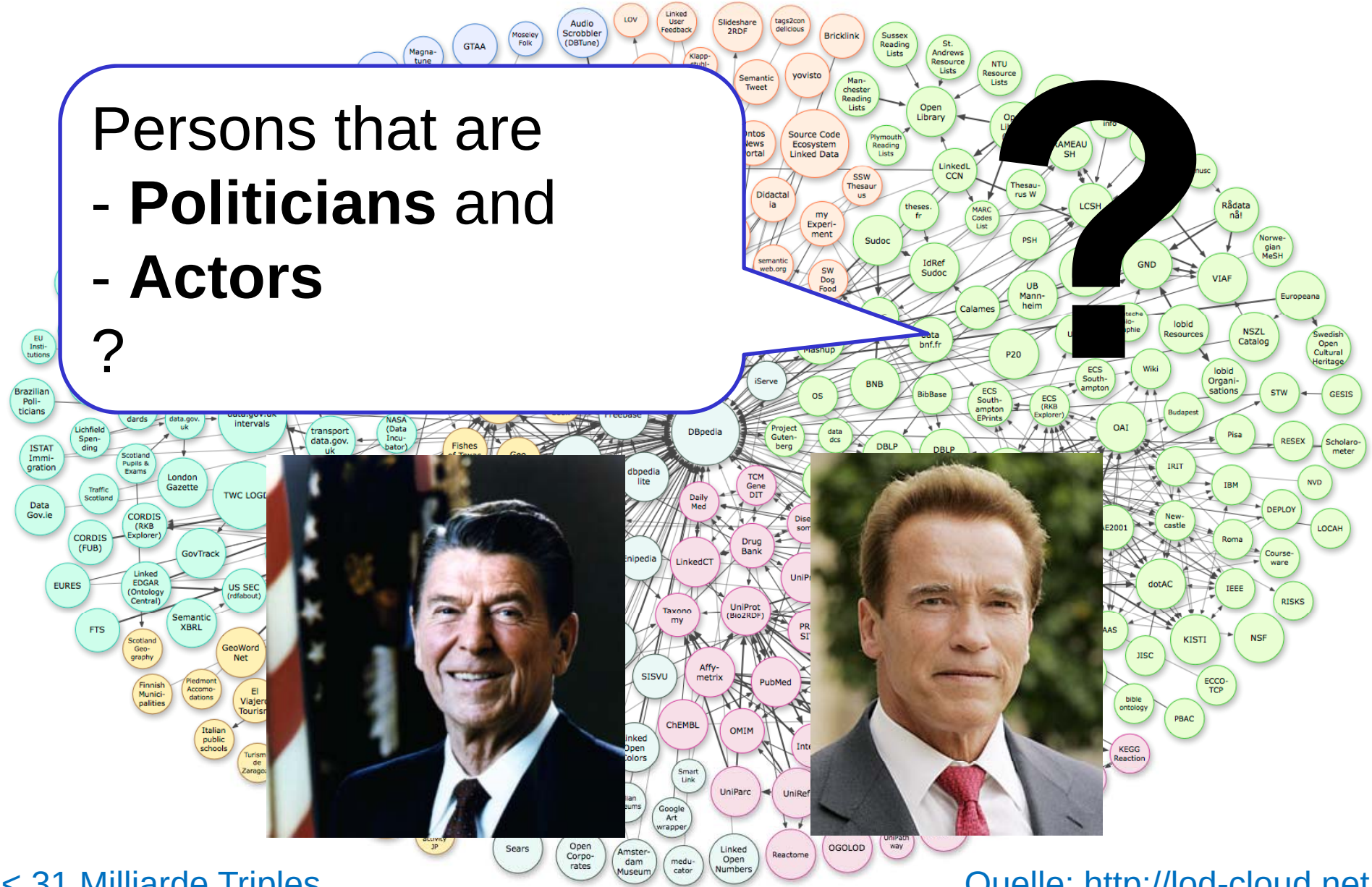


http://biglynx.co.uk/
people/matt-briggs.rdf

4. Description of URIs



Searching for Data Sources

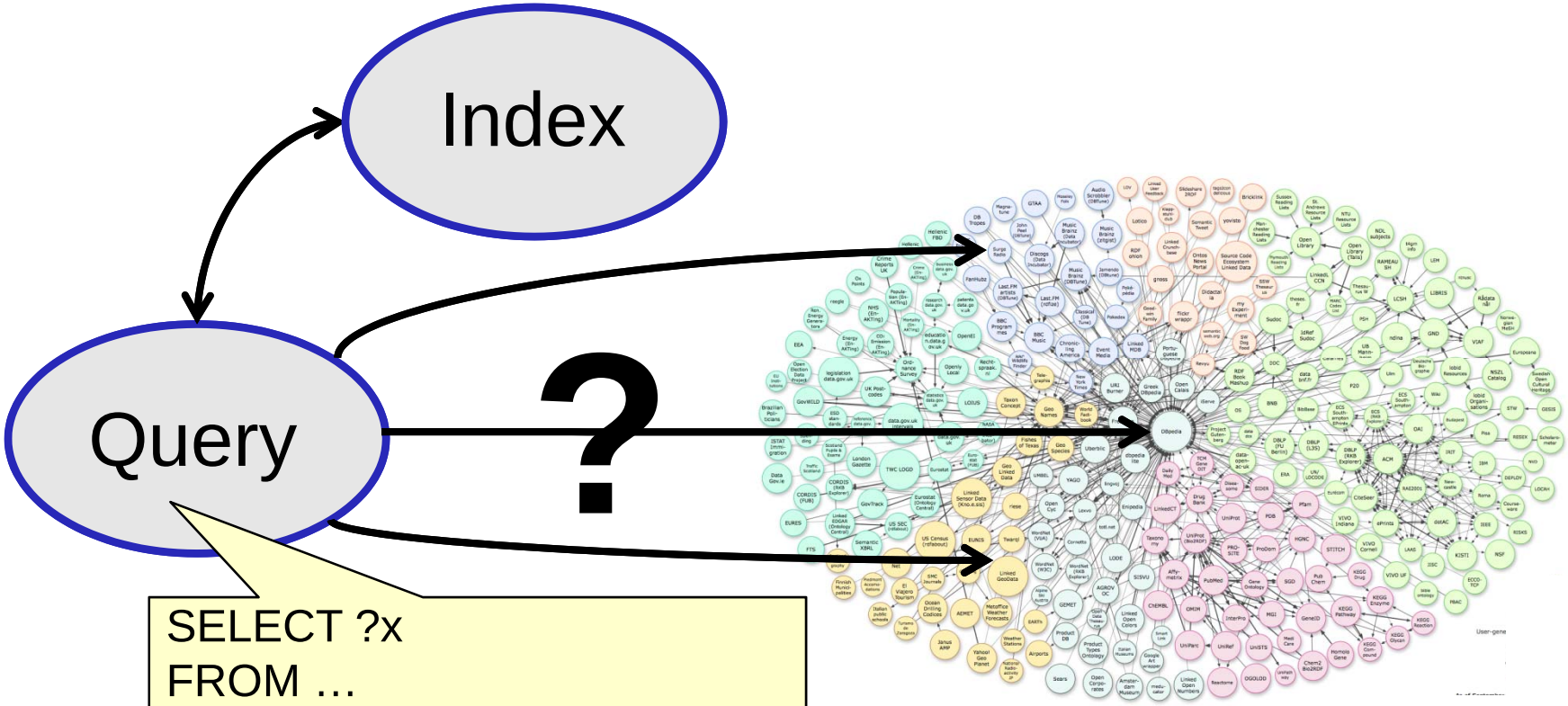


< 31 Milliarde Triples

Quelle: http://lod-cloud.net

Searching for Data Sources (2)

- No single federated query interface provided
- Schema-based index to find data sources



Idea

- Schema-based index
  - Define families of graph patterns
  - Assign instances to graph patterns
  - Store the source information (context URI)

- Construction
  - Stream-based for scalability
  - Little loss of accuracy

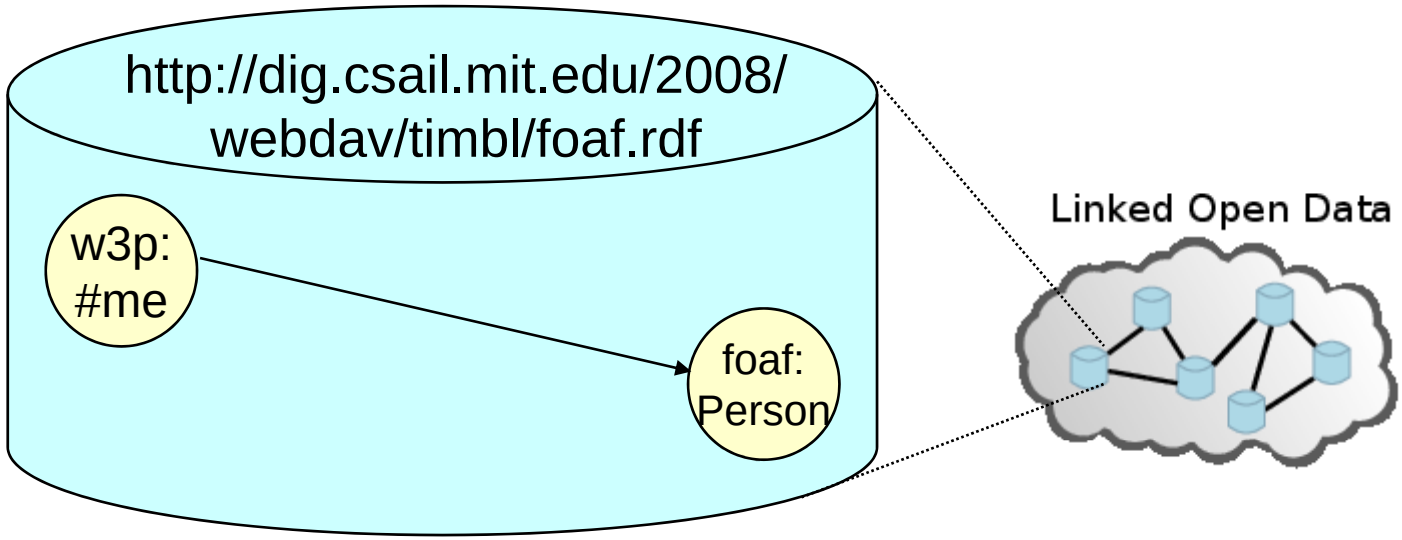
Input Data

- n-Quads

  <subject> <predicate> <object> <context>

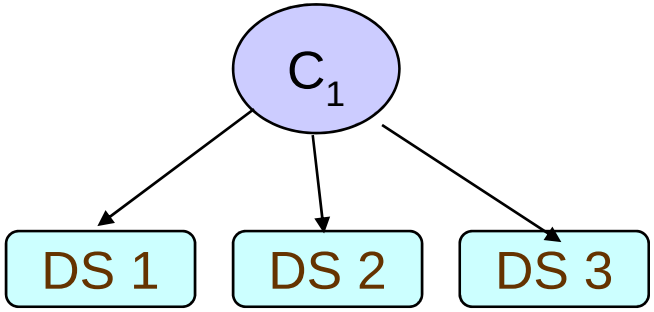
- Example:

```
<http://www.w3.org/People/Connolly/#me>
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
<http://xmlns.com/foaf/0.1/Person>
<http://dig.csail.mit.edu/2008/webdav/timbl/foaf.rdf>
```

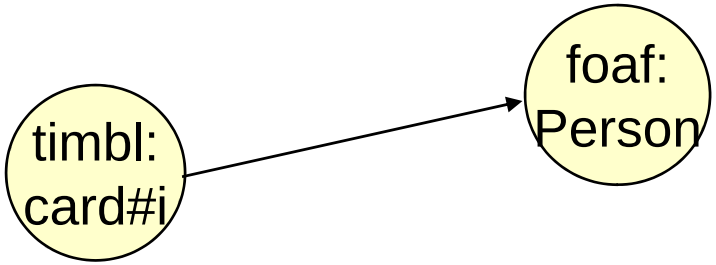


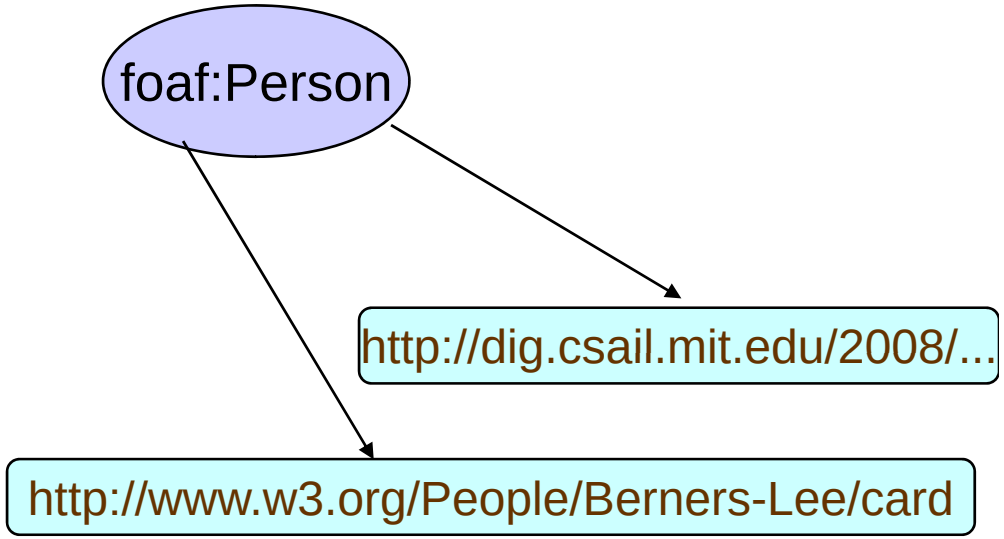
Layer 1: RDF Classes

- All instances of a particular type



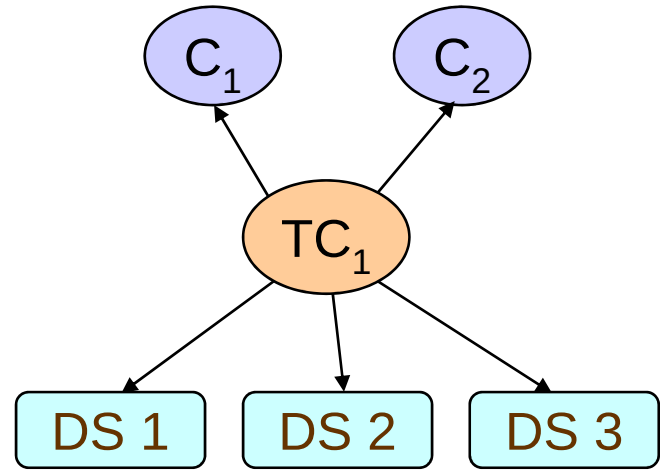

```
SELECT ?x
FROM …
WHERE {
   ?x rdfs:type foaf:Person .
}
```



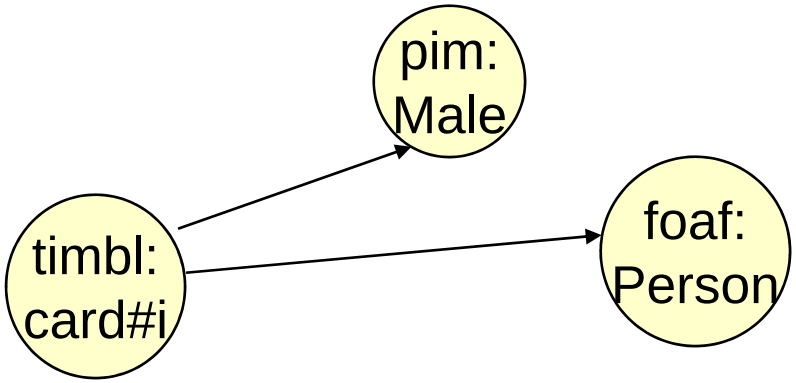


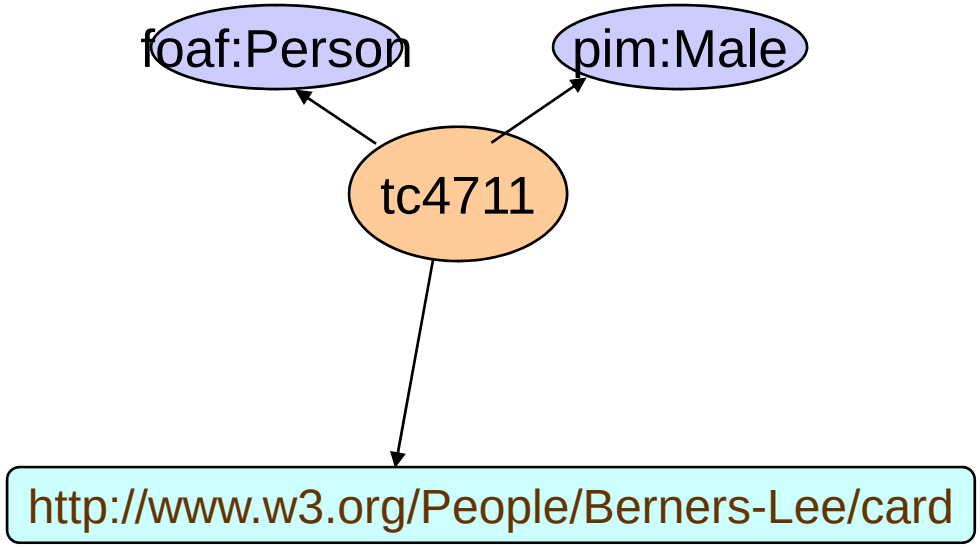
Layer 2: Type Clusters

- All instances belonging to exactly the same set of types



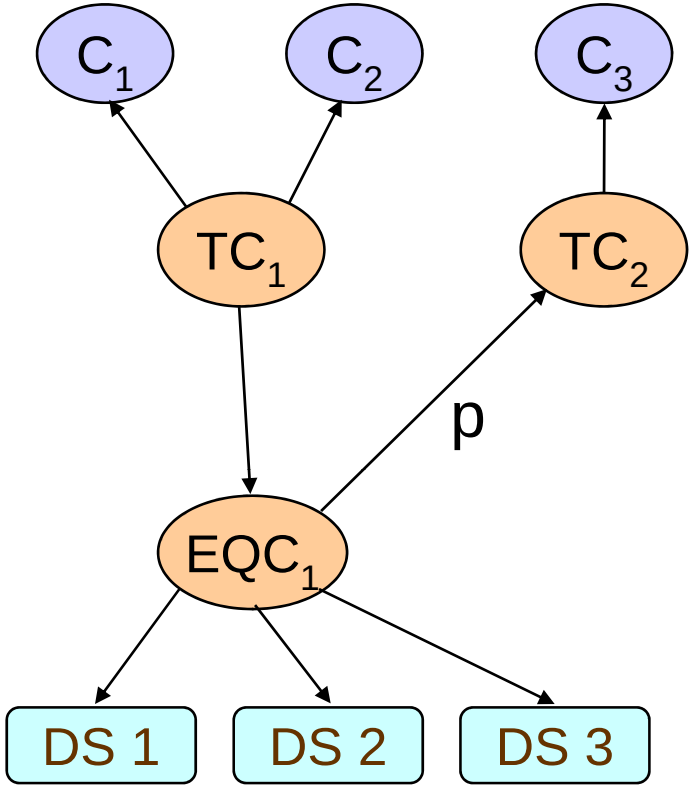

```
SELECT ?x
FROM …
WHERE {
   ?x rdfs:type foaf:Person .
   ?x rdfs:type pim:Male .
}
```





Layer 3: Equivalence Classes

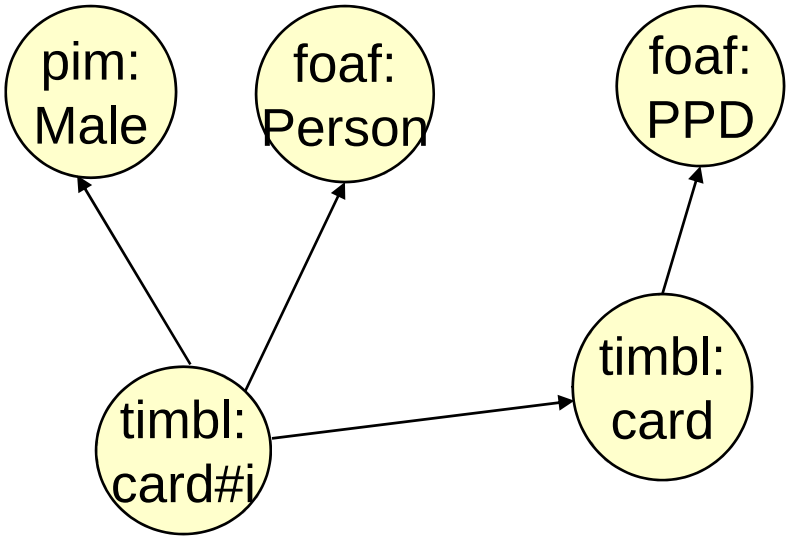
- Two instances are equivalent iff:
  - They are in the same TC
  - They have the same properties
  - The property targets are in the same TC

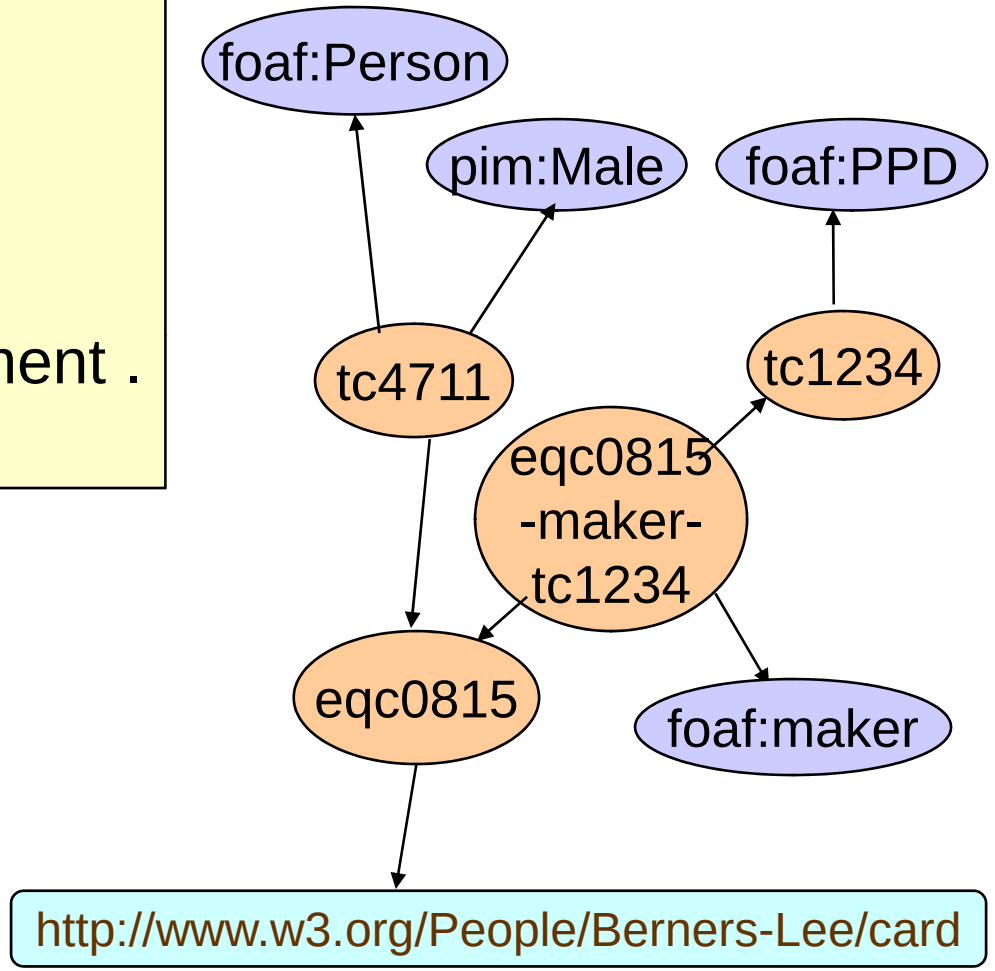


- Similar to 1-bisimulation

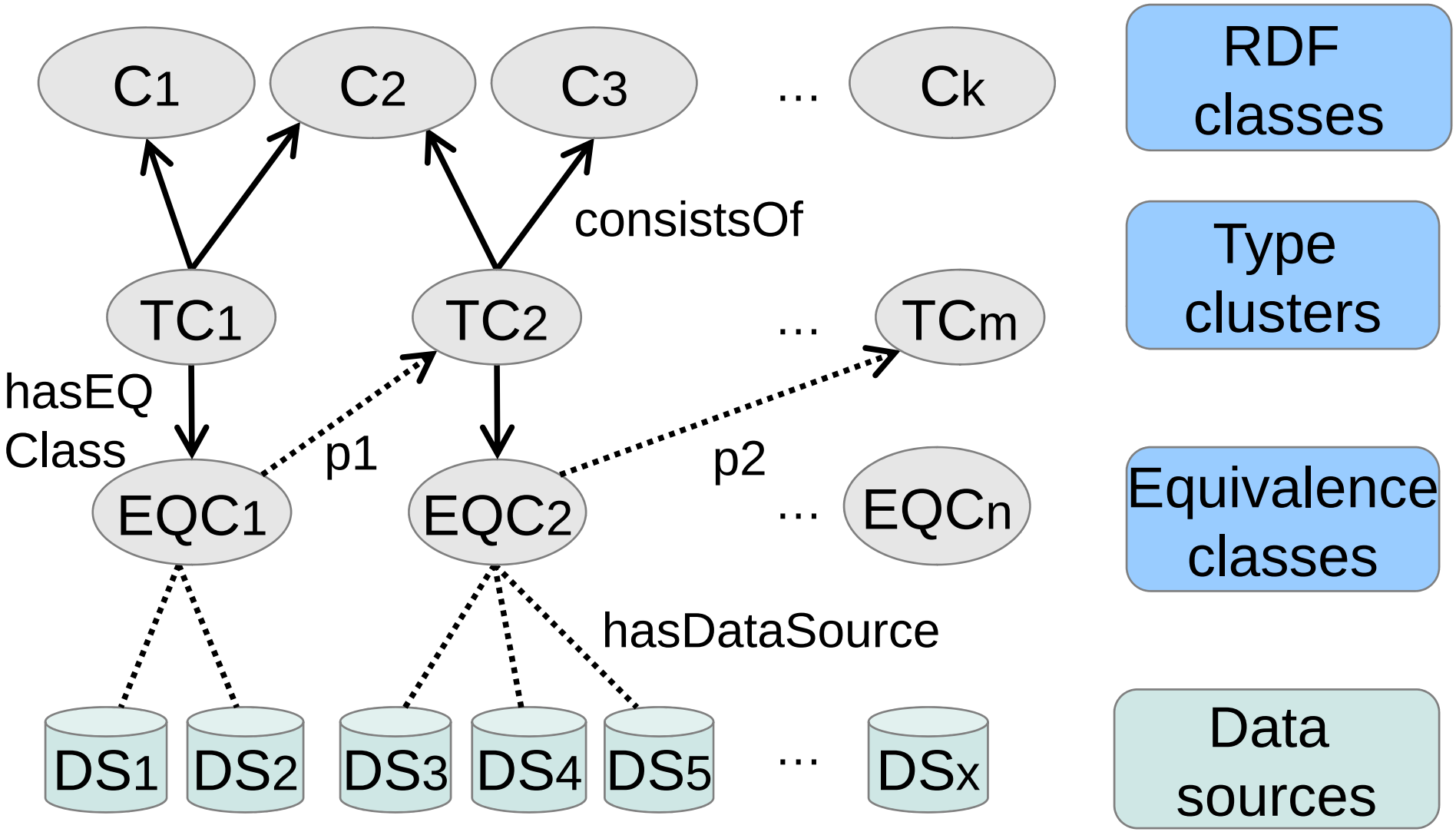
Layer 3: Equivalence Classes

```
SELECT ?x
WHERE {
   ?x rdfs:type foaf:Person .
   ?x rdfs:type pim:Male .
   ?x foaf:maker ?y .
   ?y rdfs:type
      foaf:PersonalProfileDocument .
}
```



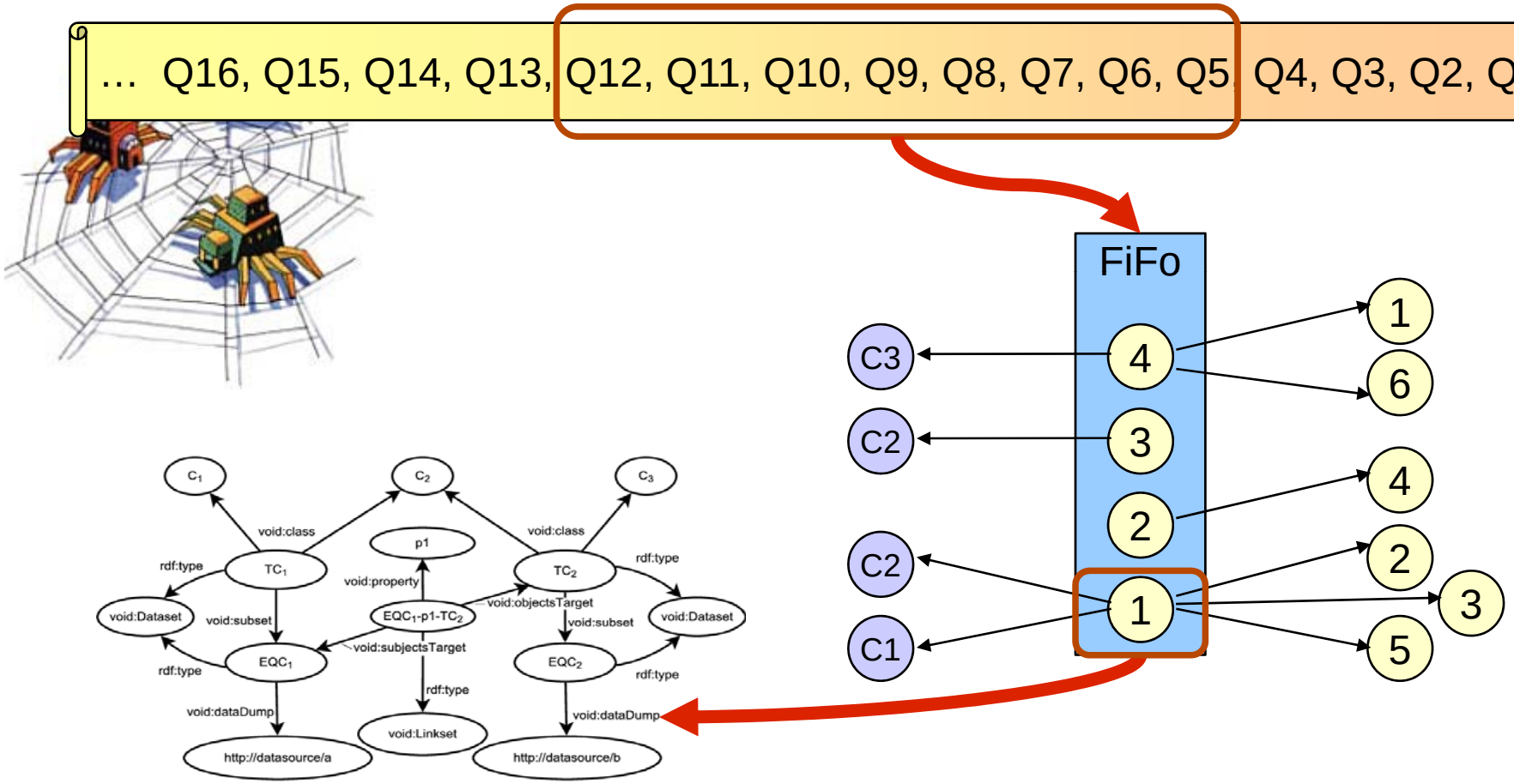


Building the Schema and Index



Building the Index from a Stream

- Stream of n-quads (coming from a LD crawler)



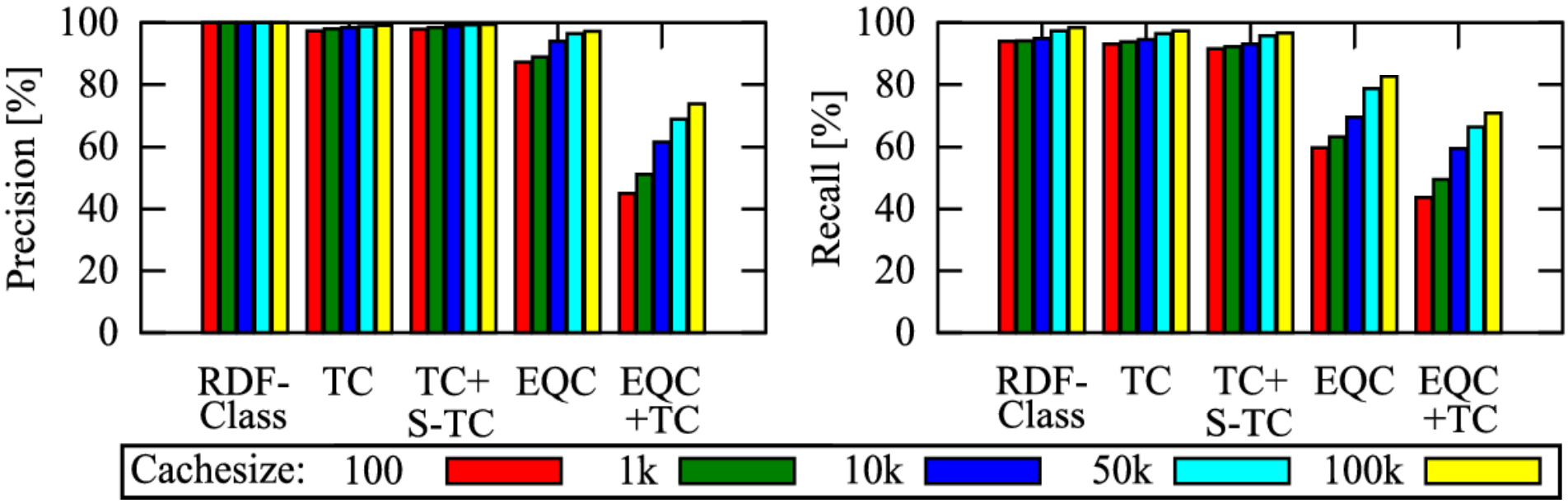
- Linear runtime complexity wrt # of input triples

Computing SchemEX: TimBL Data Set

- Analysis of a smaller data set
- 11 M triples, TimBL's FOAF profile
- LDspider with ~ 2k triples / sec

- Different cache sizes: 100, 1k, 10k, 50k, 100k
- Compared SchemEX with reference schema
- Index queries on all Types, TCs, EQCs
- Good precision/recall ratio at 50k+

Quality of Stream-based Index Construction



- Runtime increases hardly with window size
- Memory consumption scales with window size
- Commodity hardware (4GB RAM, single CPU)

Computing SchemEX: Full BTC 2011 Data

	1st billion	2nd billion	full dataset
#triples	1 billion	1 billion	2.17 billion
#instances	187.7M	222.6M	450.0M
#data sources	13.5M	9.5M	24.1M
#type clusters	208.5k	248.5k	448.6k
#equivalence classes	0.97M	1.14M	2.12M
#triples index	29.1M	24.8M	54.7M
Compression ratio	2.91%	2.48%	2.52%
runtime (hh:mm)	6:51	6:05	15:16
average runtime per 10M chunk	247 s	219 s	252 s
standard deviation	80 s	12 s	57 s
#triples/sec.	40.5k	45.6k	39.5k

Cache size: 50 k

Billion Triple Challenge 2011

challenge.semanticweb.org

A new application award

Semantic Web Challenge

http://challenge.semanticweb.org

Home | Criteria | Upcoming Challenge | Registration | Sponsors/Contact | Former Challenges

News

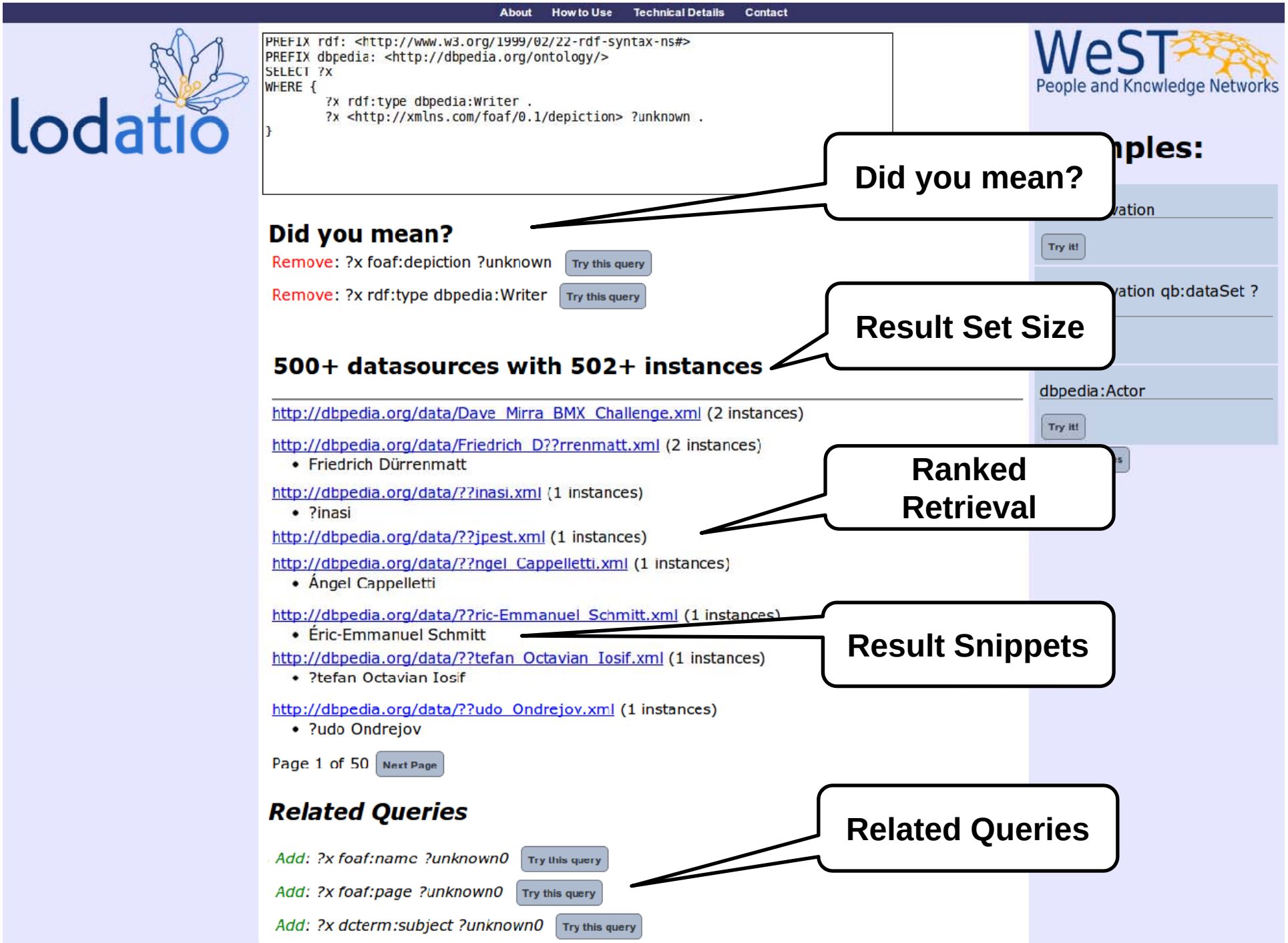
Winners of the Semantic Web Challenge

Billion Triples Track

Winner **SchemEX -- Web-Scale Indexed Schema Extraction of Linked Open Data**

Mathias Konrath, Thomas Gottron, and Ansgar Scherp

[JWS'12]


About
How to Use
Technical Details
Contact
lodatio
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dbpedia: <http://dbpedia.org/ontology/>
SELECT ?x
WHERE {
?x rdf:type dbpedia:Writer .
?x <http://xmlns.com/foaf/0.1/depiction> ?unknown .
}
WeST
People and Knowledge Networks
Did you mean?
Did you mean?
Remove: ?x foaf:depiction ?unknown
Try this query
Remove: ?x rdf:type dbpedia:Writer
Try this query
Result Set Size
500+ datasources with 502+ instances
http://dbpedia.org/data/Dave_Mirra_BMX_Challenge.xml (2 instances)
http://dbpedia.org/data/Friedrich_D??rrenmatt.xml (2 instances)
Friedrich Dürrenmatt
http://dbpedia.org/data/??inasi.xml (1 instances)
?inasi
Ranked Retrieval
http://dbpedia.org/data/??jpest.xml (1 instances)
http://dbpedia.org/data/??ngel_Cappelletti.xml (1 instances)
Ángel Cappelletti
http://dbpedia.org/data/??ric-Emmanuel_Schmitt.xml (1 instances)
Éric-Emmanuel Schmitt
Result Snippets
http://dbpedia.org/data/??tefan_Octavian_Iosif.xml (1 instances)
?tefan Octavian Iosif
http://dbpedia.org/data/??udo_Ondrejov.xml (1 instances)
?udo Ondrejov
Page 1 of 50
Next Page
Related Queries
Related Queries
Add: ?x foaf:namc ?unknown0
Try this query
Add: ?x foaf:page ?unknown0
Try this query
Add: ?x dcterm:subject ?unknown0
Try this query
ples:
vation
Try it!
vation qb:dataSet ?
dbpedia:Actor
Try it!