FORTH INSTITUTE OF COMPUTER SCIENCE
isl information systems laboratory
Semantic Access and Retrieval group

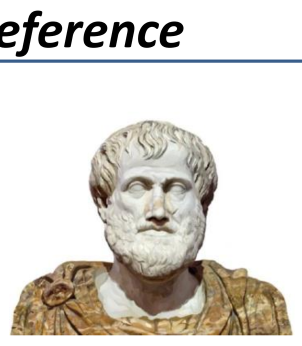
Services for Large Scale Semantic Integration of Data

1.Motivation

- The **ultimate objective** of **Linked Data** is **linking** and **integration** for **enabling discovery** and **integrated query answering** and a **big number of RDF datasets** has already been **published** and this **number keeps increasing.**
- However, it is **not currently evident** how **connected** the **LOD cloud** is, only **measurements** between **pairs of datasets** are available. It is **not possible** to **find** the **number of common URIs** between **3 or more datasets**.
- **Measurements** and **indexes** involving **more than 2 datasets** are important for:

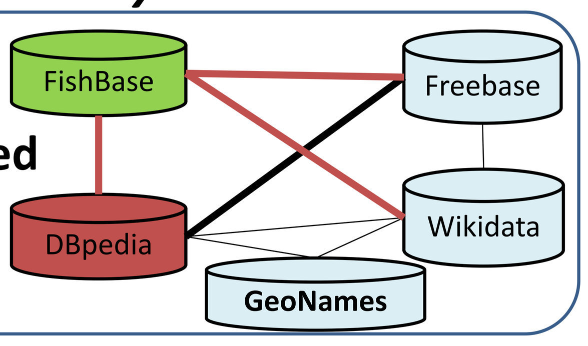
Object Coreference

Give me all the URIs of Aristotle

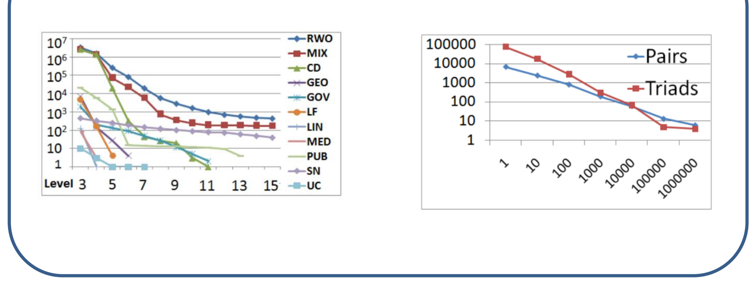


Dataset Discovery & Selection

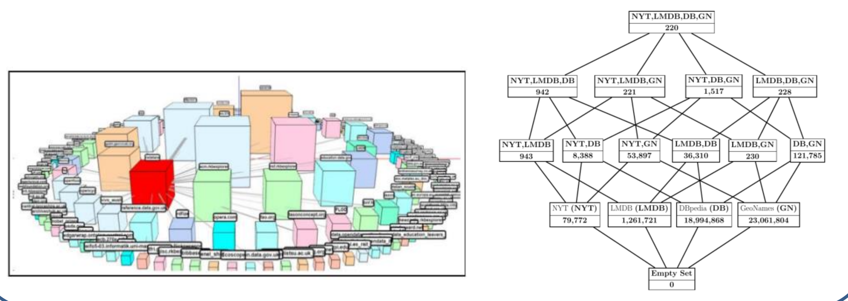
Give me the K most connected Datasets to my Dataset



Connectivity Assessment & Monitoring



Informative Visualizations



2. Problem - Running Example

<u>**We focus on how to compute efficiently**</u>

- ❑ <u>The datasets containing a particular (or equivalent) URI</u>
- ❑ <u>the number of common or equivalent URIs (i.e. same real world objects) in any subset B</u>

URIs of Datasets

1. NYT	2. DBpedia	3. Yago	4. Geonames
nyt:jordan_michael	dbp:Michael_Jordan	yg:Michael_Jordan	geo:Texas
en_wiki:san_diego	en_wiki:Canada	yg:Aristotle	geo:Dallas
dbp:Texas	dbp:Texas	en_wiki:san_diego	en_wiki:Portland
de_wiki:USA	dbp:Aristotle	yg:Socrates	geo:Las_Vegas
en_wiki:Portland	dbp:Las_Vegas	dbp:Las_Vegas	de_wiki:USA
en_wiki:Las_Vegas	dbp:Houston	geo:Texas	en_wiki:san_diego

SameAs Relationships

dbp:Michael_Jordan `owl:sameAs` yg:Michael_Jordan
dbp:Michael_Jordan `owl:sameAs` nyt:jordan_michael
dbp:Aristotle `owl:sameAs` yq:Aristotle
dbp:Texas `owl:sameAs` geo:Texas
dbp:Las_Vegas `owl:sameAs` en_wiki:Las_Vegas
en_wiki:Las_Vegas `owl:sameAs` geo:Las_Vegas

1. Prefix Index

Prefix	dataset ID
http://data.nytimes.com/ (**nyt**)	1
http://en.wikipedia.org/ (**en_wiki**)	2,3,4,1
http://dbpedia.org/ (**dbp**)	1,3,2
http://de.wikipedia.org/ (**de_wiki**)	1,4
http://yago-knowledge.org/ (**yg**)	3
http://www.geonames.org/ (**geo**)	3,4

2. SameAs Catalog

URI	ID
nyt:jordan_michael	1
dbp:Michael_Jordan	1
yg:Michael_Jordan	1
dbp:Aristotle	2
yq:Aristotle	2
dbp:Texas	3
geo:Texas	3
dbp:Las_Vegas	4
en_wiki:Las_Vegas	4
geo:Las_Vegas	4

3. Creation of Element Index

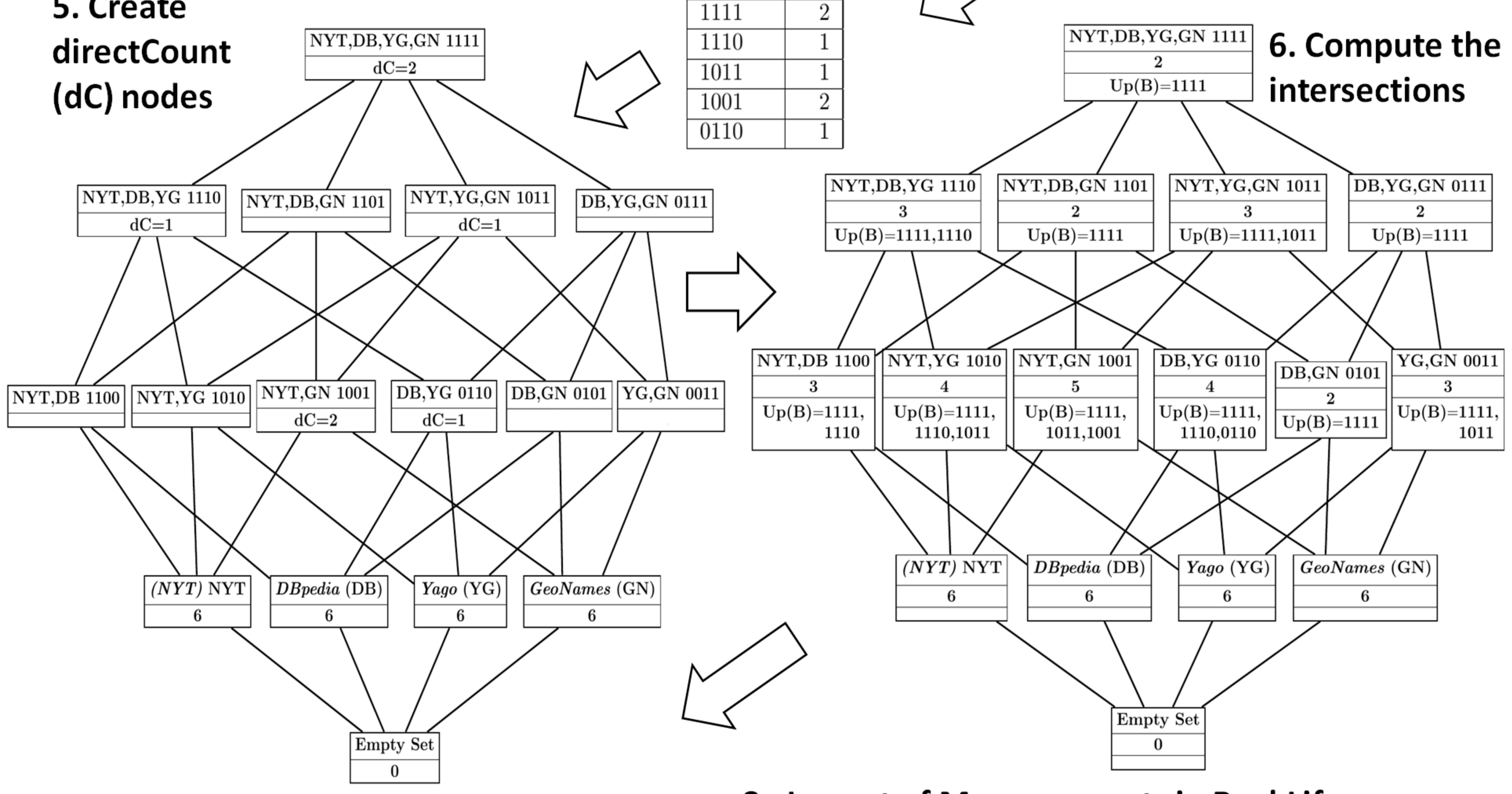
URI or SameAs ID	ID Number	Bit Array
1 (SameAs ID)	1,2,3	1110
2 (SameAs ID)	2,3	0110
3 (SameAs ID)	1,2,3,4	1111
4 (SameAs ID)	1,2,3,4	1111
de_wiki:USA	1,4	1001
en_wiki:Portland	1,4	1001
en_wiki:san_diego	1,3,4	1011

4. Direct Counts List

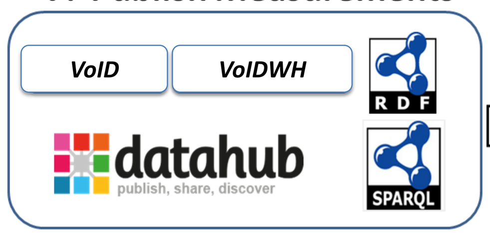
Subset	dC
1111	2
1110	1
1011	1
1001	2
0110	1

5. Create directCount (dC) nodes

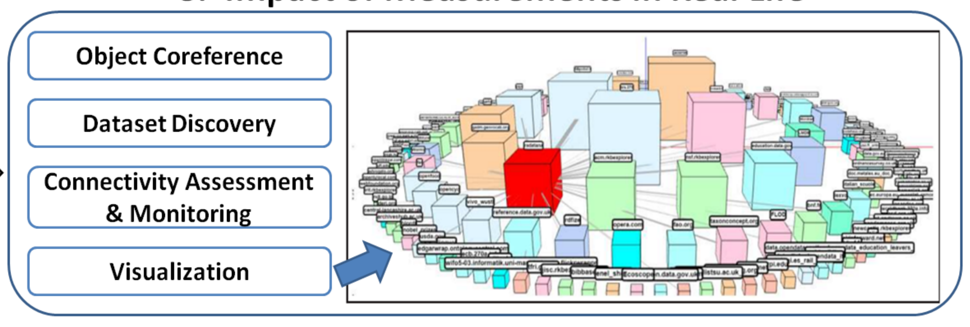
6. Compute the intersections



7. Publish Measurements



8. Impact of Measurements in Real Life

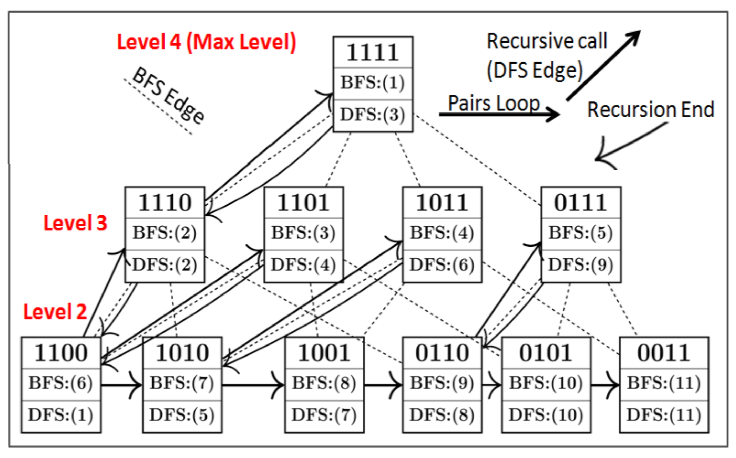


3. The Proposed Indexes

- ❑ **1. Prefix Index:** An index which lists all namespaces and for each one what datasets contain them, e.g., **see step 1 of Running Example**.
- ❑ **2. SameAs Catalog:** A catalog that computes the symmetric & transitive closure of **owl:sameAs** relationships. All the URIs that belong to the same class of equivalence (i.e., referring to the same entity) are getting the same signature, e.g., **see step 2 of Running Example.**
- ❑ **3. Element Index:** For each real world object (i.e., URI or signature) appearing in two or more datasets, this index stores the datasets where it occurs (e.g., **see step 3 of Running Example),** by exploiting
  - ✓ **SameAs Catalog** for replacing a URI with its signature
  - ✓ **Prefix Index** for identifying the possible datasets where a URI occurs
  - ✓ **ASK queries** for checking if a URI exists at least in two datasets

4. The Lattice of Measurements

- ❑ A **lattice** is a partially ordered set which can be represented as a Directed Acyclic Graph (DAG) where the edges points towards the direct supersets.
- ❑ We compute the intersection of any set of datasets by **making the measurements of the lattice <u>incrementally</u>:**
  - ***directCount(B):*** the frequency of subset B in the element index. (e.g., **see steps 4 & 5 of Running Example**)
  - ***Up(B):*** the supersets of B that can be found in directCount List (e.g., **see step 6 of Running Example)**
  - <u>The sum of the directCount of Up(B) gives the number of common real world objects in B.</u>
- ❑ We propose **two incremental** algorithms that require only one index scan for computing the lattice (or a part of it) and exploit lattice and set theory properties.
- ❑ **Top-Down approach using Breadth-First Search (BFS)** starts from the maximum level (i.e., quad in our example). Then, it continues with the computation of the intersection of triads and finally of the pairs.
- ❑ **Bottom-Up approach using Depth-First Search (DFS)** starts by computing the intersection of a pair and continues upwards following a "Height First Search".



Lattice Traversal (BFS and DFS)

5. Experimental Evaluation

New connections thanks to the closure:

- ✓ **19 millions** of newly discovered **owl:sameAs** pairs!
- ✓ **2,393** of newly discovered connected pairs of datasets!

Measuring the current status of LOD:

- ✓ DBpedia, Freebase and Yago share **2.7 millions** of real world objects
- ✓ Only **2.3 %** of real world objects exist in **3 or more datasets**

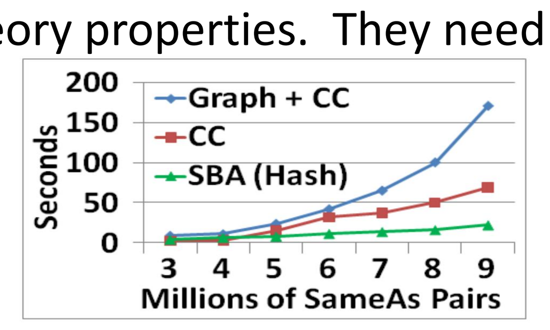
Category	Value
Prefix Index Size	63,803
Unique Real World Objects	141,269,960
Element Index Size (*rwo*)	6,242,344
Element Index Size (URIs)	17,840,499
Asks Number	6,684,242
rwo in 3 or more D_i	3,293,248
URIs corresponding to *rwo* in 3 or more D_i	12,296,650
Num. of Lattice Nodes (threshold $\geq$ 30)	130,525,631
Num. of Lattice Nodes (threshold $\geq$ 20)	1,541,968,012

Category	Value
SameAs Triples	13,158,621
SameAs Catalog Size	18,789,593
SameAs Triples Inferred	19,450,107
Pairs sharing at least 1 real world object	6,708
New Pairs discovered due to SameAs Alg.	2,393
Triads sharing at least 1 real world object	74,432
New Triads discovered due to SameAs Alg.	48,658
SameAs Unique IDs	6,218,958

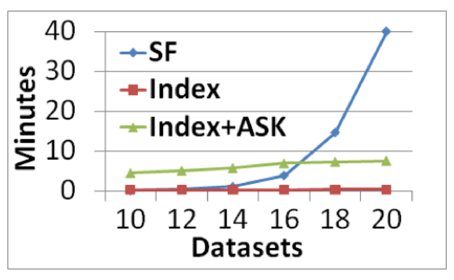
Datasets of subset B	$co_{\sim}(B)$
1: {DBpedia,Freebase,Yago}	2,709,171
2: {DBpedia,Freebase,Wikidata}	1,950,319
3: {DBpedia,Yago,Wikidata}	1,435,713
4: {Yago,Freebase,Wikidata}	1,434,407
5: {DBpedia,Yago,Freebase,Wikidata}	1,434,404
6: {DBpedia,GADM,Freebase}	107,968
7: {DBpedia,GeoNames,Freebase}	98,985
8: {DBpedia,GADM,Wikidata}	96,968
9: {GADM,Freebase,Wikidata}	96,968
10: {DBpedia,GADM,Freebase,Wikidata}	96,968

Time Efficiency

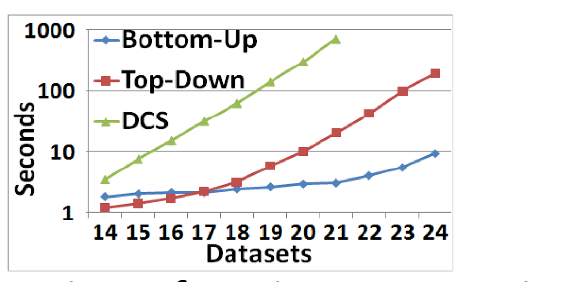
- ✓ Signature-based algorithm (**SBA**) needs **45 seconds** to **compute** the **closure** of **13 millions of owl:sameAs pairs** and is faster than a common Connected Components algorithm (**CC**)!
- ✓ Index approach is faster than a straightforward (**SF**) method that performs binary search.
- ✓ **1.5 billion** of **subsets intersections** computed in **35 minutes** with **the bottom-up algorithm**.
- ✓ Incremental approaches are faster than methods (e.g., **DCS**) which do not exploit lattice & set theory properties. They need **less than 10 seconds** for **over one million subsets.**



Exec. Time of computing closure



Exec. Time for varying D



Exec. Time of Lattice computation

6. Publishing and Exchanging Measurements

<u>**TRY LODsyndesis :**</u> *<u>www.ics.forth.gr/isl/LODsyndesis/</u>*

& FIND links to: datahub, a 3D visualization page, an active SPARQL Endpoint & a list of answerable queries.

Contact

Michalis Mountantonakis
Yannis Tzitzikas (coordinator)

http://www.ics.forth.gr/isl/LODsyndesis

BlueBRIDGE
Very Large Data Bases