# SPRING: Ranking the results of SPARQL queries on Linked Data

Kunal Mulay and P Sreenivasa Kumar


17th International Conference on Management of Data
COMAD 2011, Bangalore, India

# Outline

- Introduction
- Architecture
- Ranking scheme
  - Ranking datasets
  - Ranking resources
  - Ranking Triples
- Experimental setup
- Results
- Conclusion
- Future work

# Introduction

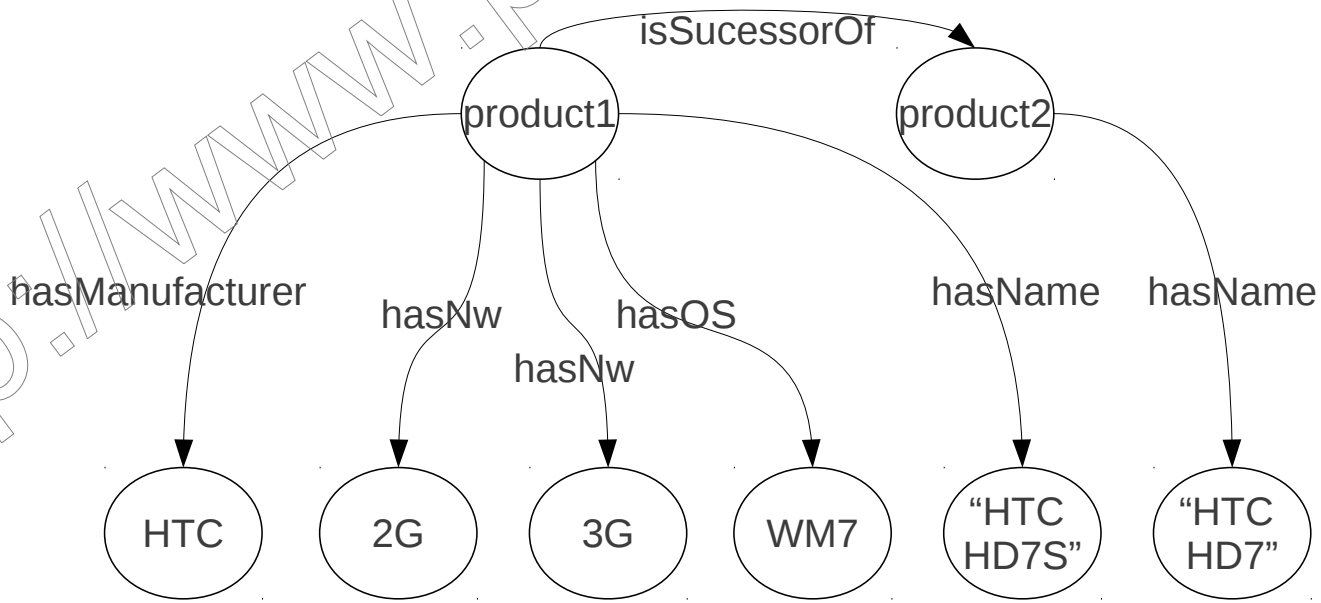
- The amount of semantic data present on the web has been increasing in the recent past
  - Research organizations and government data
  - Linked Open Data
- The web of data community uses Resource Description Framework(RDF) for data and metadata representation
- The data represented can be queried using the SPARQL query language
- Current search engines use shared vocabulary for annotating the text data (schema.org)

# Resource Description Framework

- Each data element is represented by a triple
  - (Subject - Predicate - Object)
  - Each of them is represented by a URI, but Objects can be character strings also
- The collection of such data elements forms a huge graph called RDF graph
  - Subject and object are represented by nodes
  - Predicate is represented as an edge between these two nodes

# Example(RDF data)

"HTC HD7S" could include the following triples:
(product1, hasName, "HTC HD7S")
(product1, hasManufacturer, HTC )
(product1, hasNetwork, 2G )
(product1, hasNetwork, 3G)
(product1, hasOS, Windows7Mobile)
(product1, isSuccessorOf, product2)
(product2, hasName, "HTC HD7")

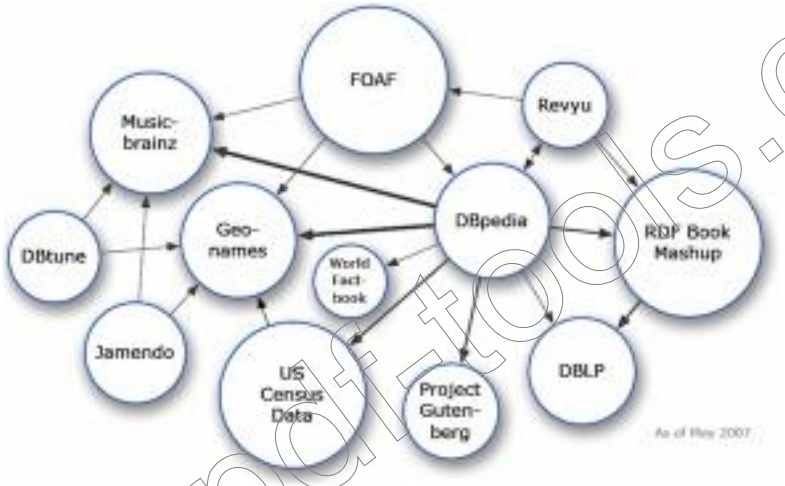


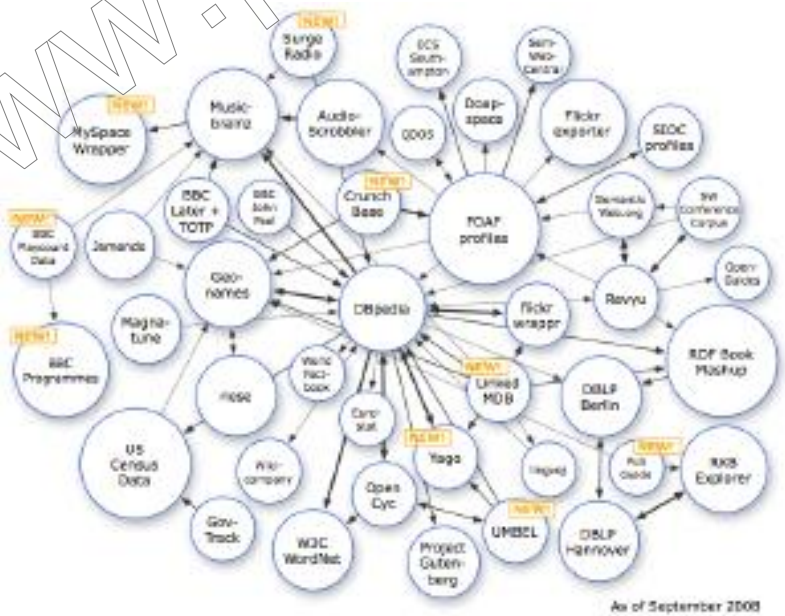
# Linked Open Data

- Linked Open Data(LOD) is an initiative to connect the data that wasn't previously connected
- It recommends the best practice for connecting and sharing pieces of data using URIs and RDF
- Principles of publishing Linked Data
  - Use URIs as names for things
  - Use HTTP URIs, so that people can look up those names
  - When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
  - Include links to other URIs, so that they can discover more things

# LOD Growth

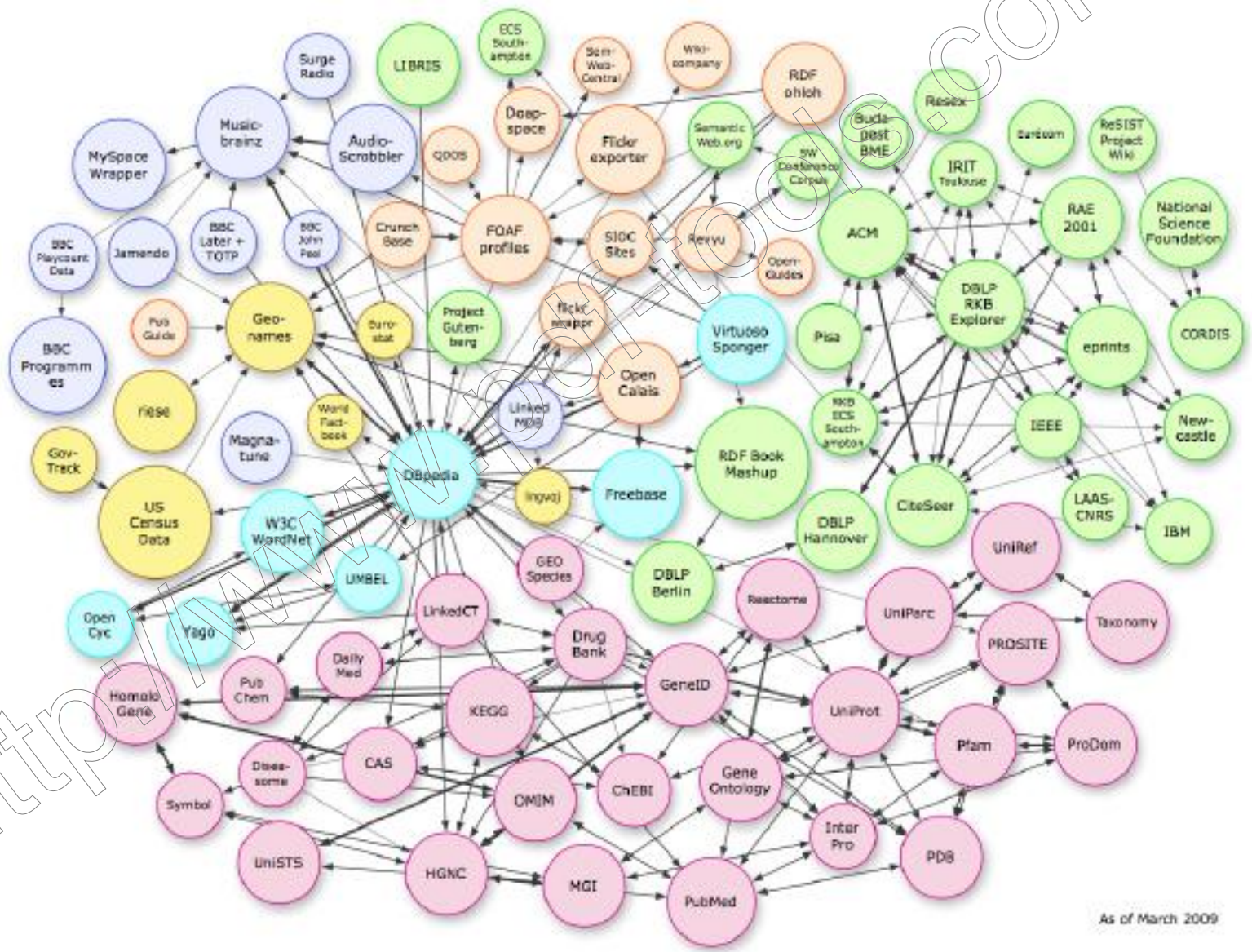
Number of datasets (2007) : 18, Number of datasets (2008) : 45



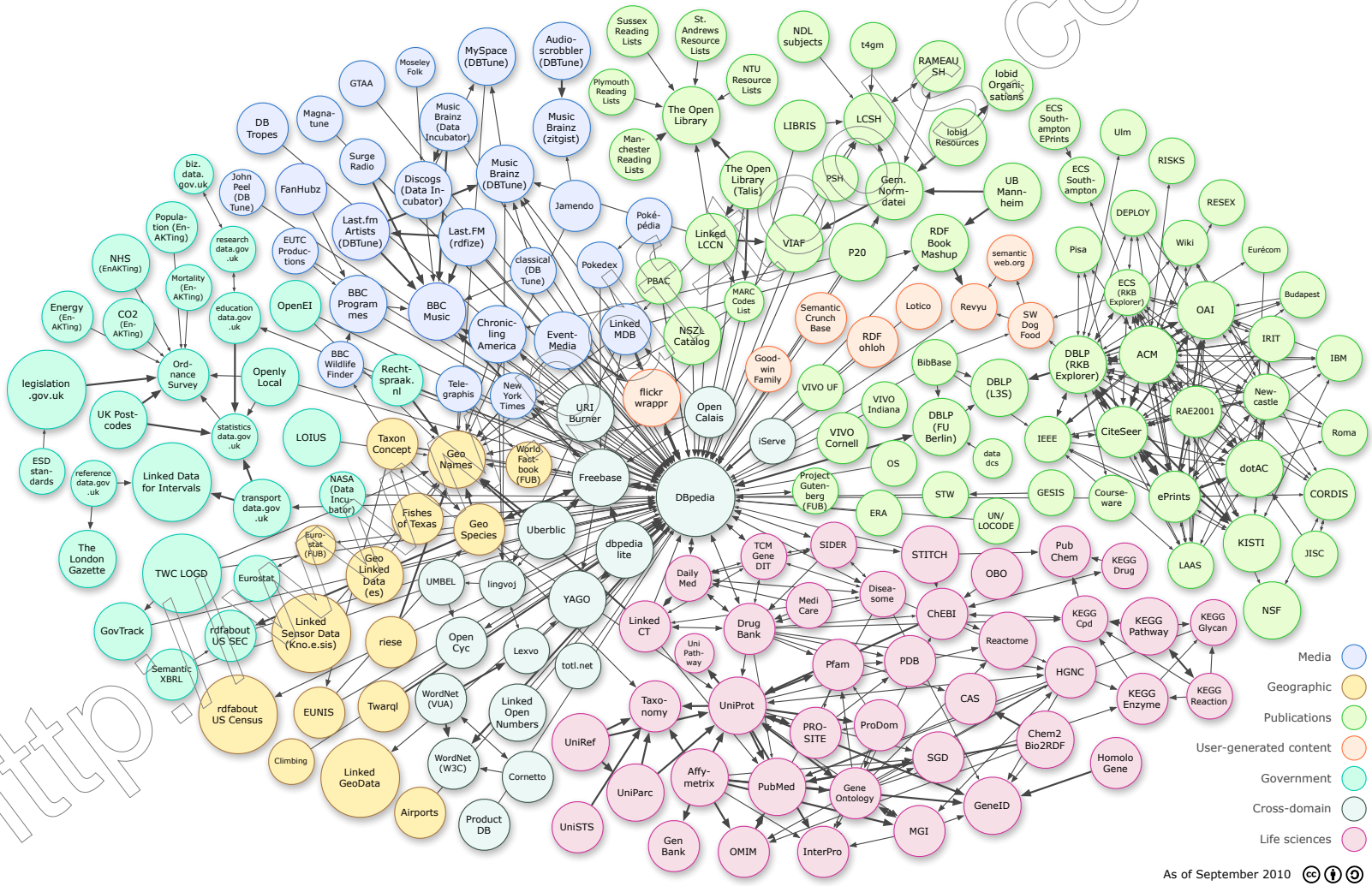
# LOD Growth

Number of datasets (2009) : 95



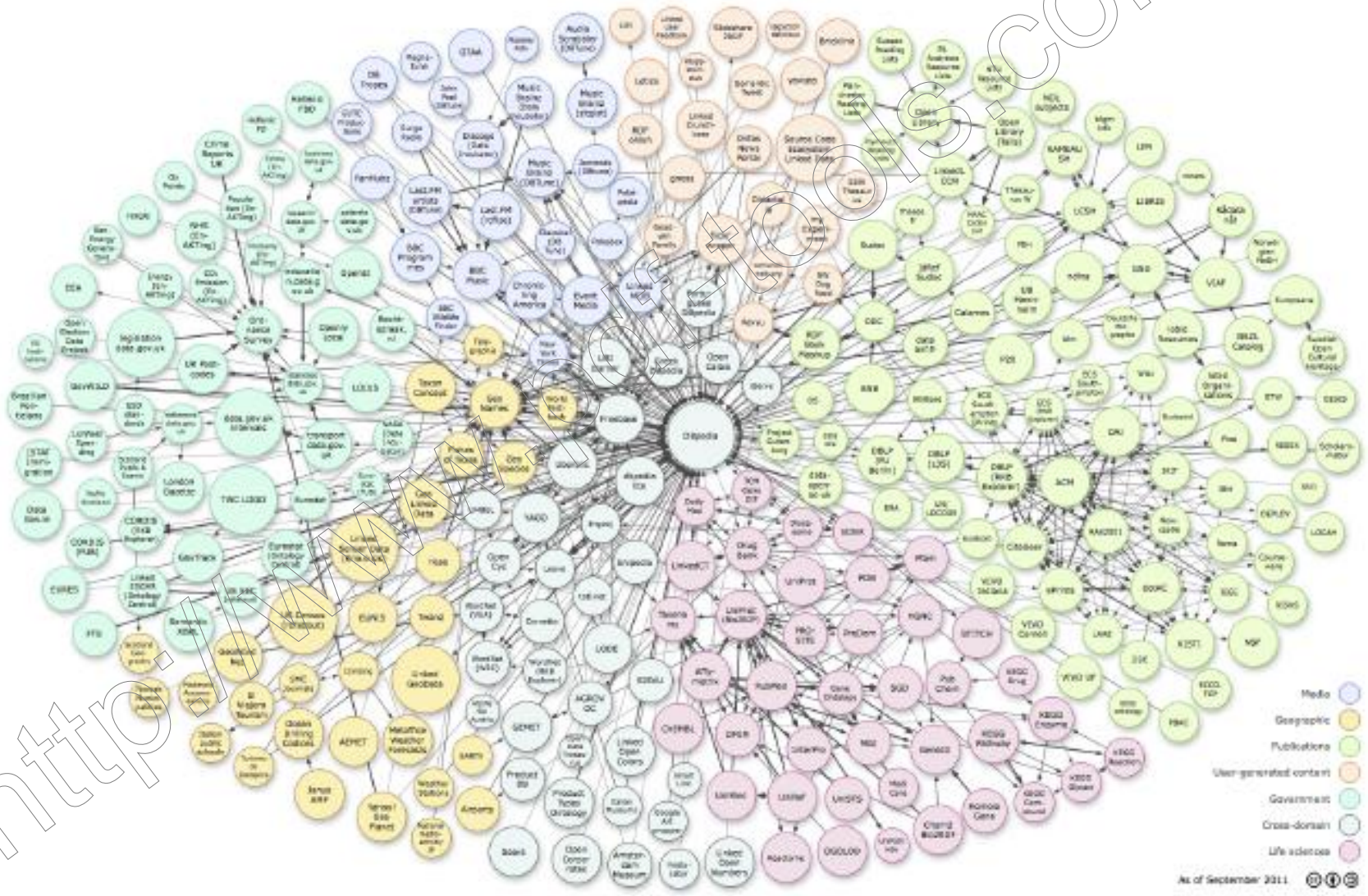
# LOD Growth

## Number of datasets (2010) : 203



# LOD Growth

Number of datasets (2011) : 295



# SPARQL

- A SPARQL query returns the nodes that satisfy the specified conditions
- The conditions are specified using
  - triple patterns or basic graph patterns
- Example SPARQL query on the RDF data:
  - Get names of products having manufacturer as HTC and support 3G networks.

```
SELECT ?name WHERE {
  ?product <hasManufacturer> <HTC> .
  ?product <hasNetwork> <3G> .
  ?product <hasName> ?name . }
```

- Get the full name of scientists born in any city of Switzerland and having the doctoral advisor who is born in Germany.
  - SELECT ?GivenName ?FamilyName WHERE {
    ?GivenName <givenNameOf> ?p .
    ?FamilyName <familyNameOf> ?p .
    ?p <type> <scientist> .
    ?p <bornInLocation> ?city .
    ?city <locatedIn> “Switzerland” .
    ?p <hasDoctoralAdvisor> ?a .
    ?a <bornInLocation> ?city2 .
    ?city2 <locatedIn> “Germany” . }

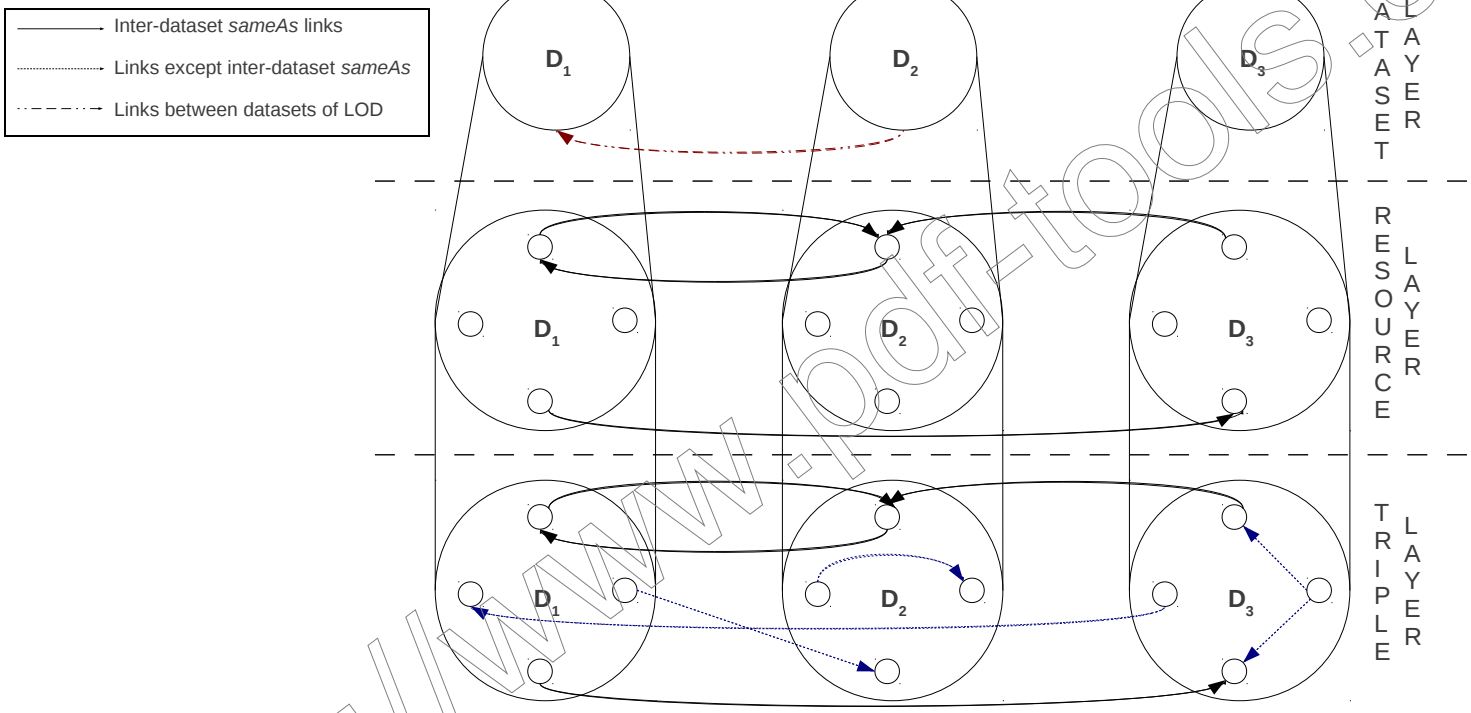
# Motivation

- Even though SPARQL queries are fairly specific, their results may have a large number of triples
  - Ex., Get names of all student - advisor pairs who authored a research paper together.
- One physical entity can be represented in many different datasets
  - The entity Tim-Lee is present in DBpedia, Yago, DBLP, Freebase, uriburner, semantic web dog food, etc
- We use mutual consensus between datasets to rank order the entities
- We call this system as SPRING( **SP**arql **R**esult rank**ING**)

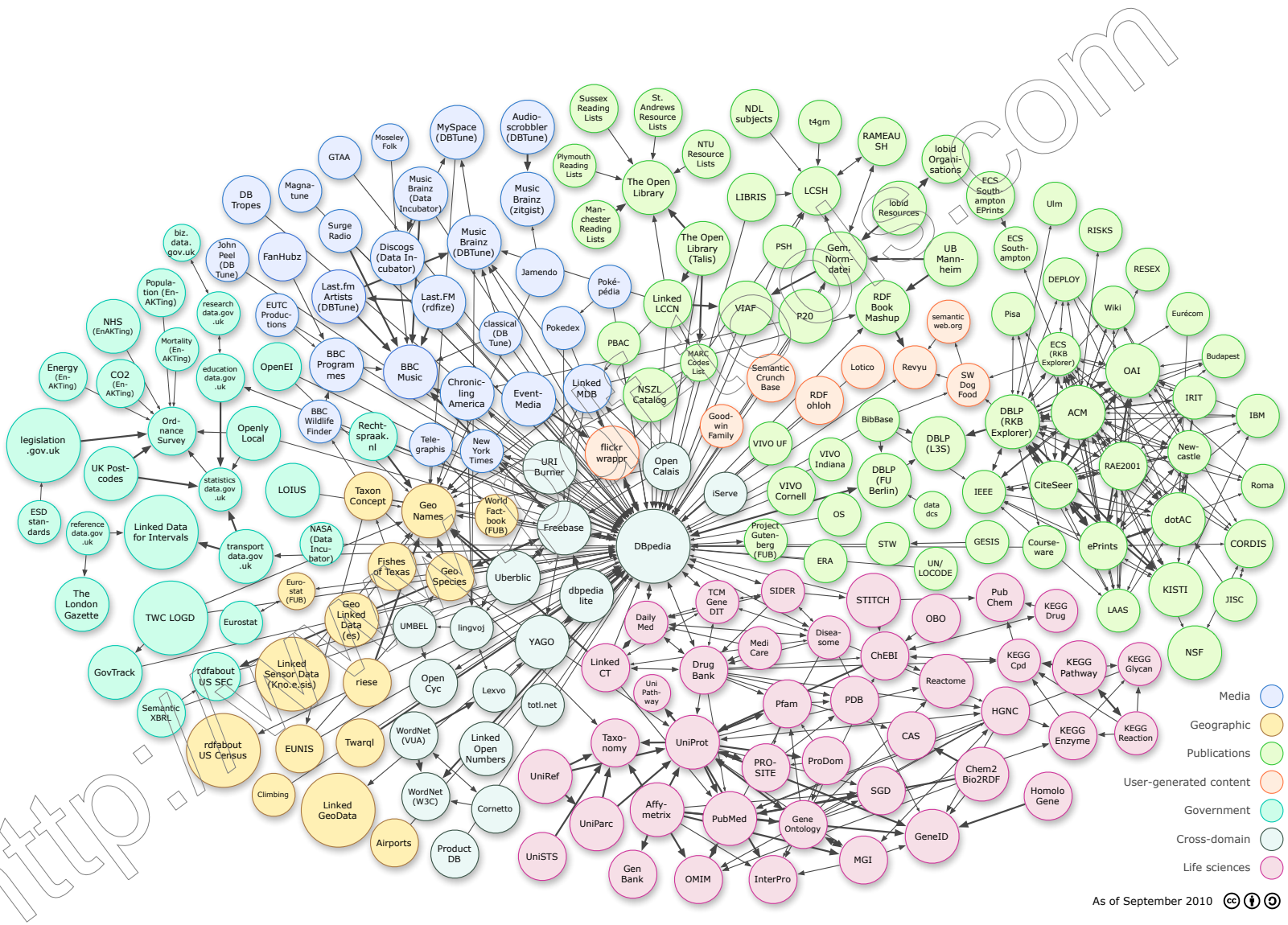
# Ranking Scheme

- Makes use of a three level ranking scheme
- Rank datasets on the basis of links existing between them
  - Make use of all links existing between resources of two datasets
- Rank entities on the basis or inter-dataset *owl:sameAs* links
  - Make use of only *owl:sameAs* links
- Rank triples on the basis of the entities present in the triple

# Architecture



# Ranking Datasets



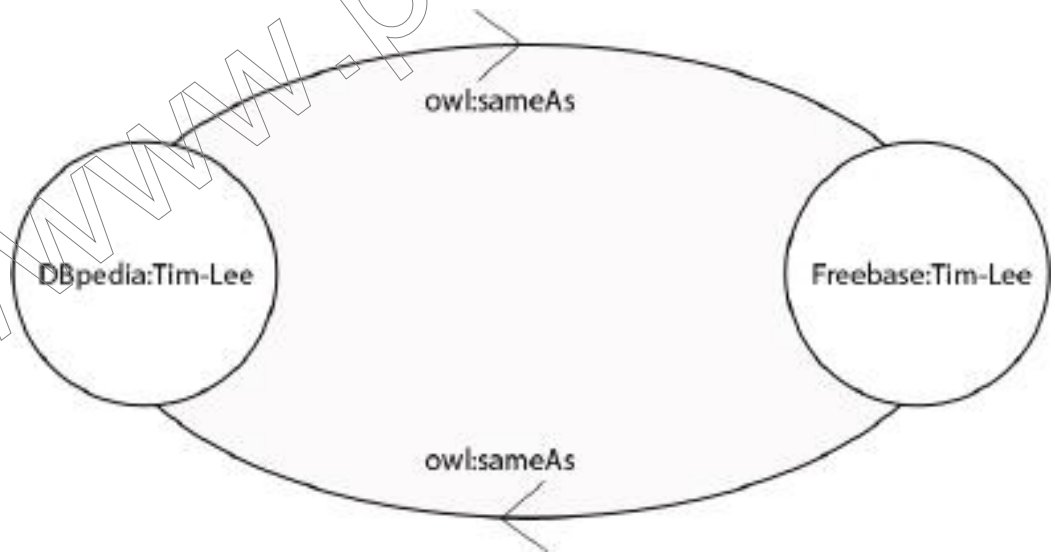
# Datasets Ranking Score

- Make use of the links existing between two datasets in LOD cloud
- A directional link exists from $d_i$ to $d_j$
  - If there are at least 50 links from resources of $d_i$ to $d_j$.
- The ranking score for a dataset $d$, denoted as $R_{ds}(d)$, is defined as:

$$R_{ds}(d) = \frac{\text{Total number of incoming links to dataset } d}{\text{Total number of datasets in LOD cloud}}$$

# Ranking Entities

- Make use of the consensus existing between two datasets
- The consensus is captured by *owl:sameAs* links between resources of different datasets
- *owl:sameAs* means that
  - The two URI's connected by *owl:sameAs* actually refer to same object

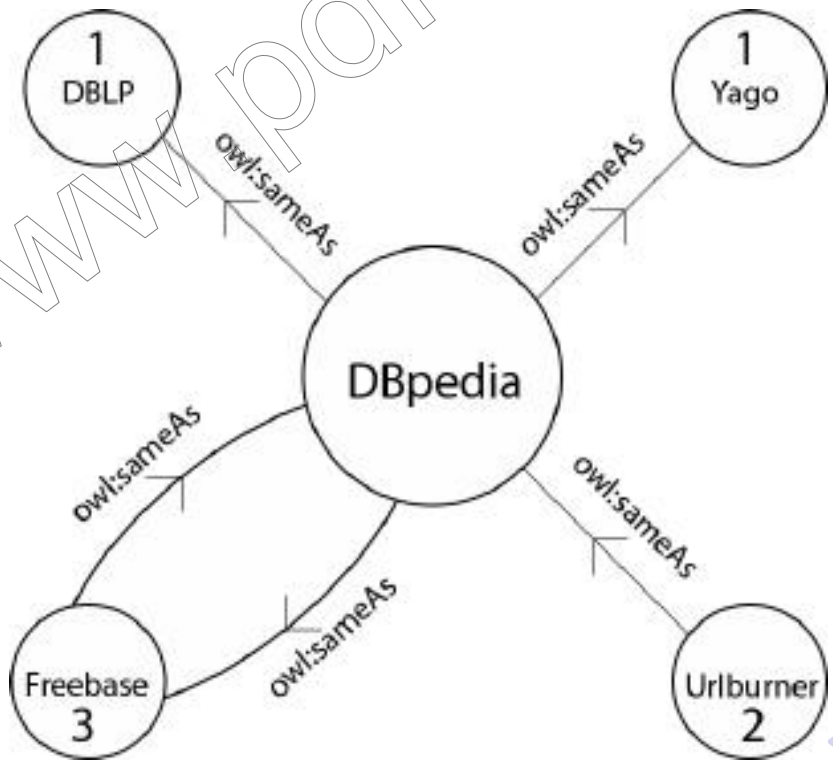


# Practical Use of *owl:sameAs*

- *sameAs* predicate is defined as *symmetric* and *transitive*, but in practice these are some times violated
- There are three possible types of links between two resources
  - *sameAs* link exists but resources are different
    - The link between two "football" resources, where one is rugby and other is soccer
  - *sameAs* link exists and resources are approximately same
    - In OpenCyc Sodium is defined in its pure form, while in DBpedia it is defined to include isotopes also
  - No *sameAs* link exists, but resources are same

# Types of Links

- The *owl:sameAs* links can be classified into three types:(with respect to a particular node)
  - Type 1: Outgoing link to a resource in a different dataset
  - Type 2: Incoming link from a resource from a different dataset
  - Type 3: Bidirectional links between a pair of resources, that are not in the same dataset
- Only type 2 and type 3 can be used for ranking entities



# Ranking Score for Resources

- Mutual score: Uses type 3 links

$$R_{mutual}(r) = \sum_{i=1}^{n} R_{ds}(\text{dataSetOf}(r_i))$$

$dataSetOf(r_i)$ denotes the dataset, say $d_i$, that contains resource

$r_1$, $r_2$,...., $r_n$ are the resources from different datasets having Type 3 links to resource $r$

- Partial score: Uses type 2 links

$$R_{pa}(r) = \sum_{k=1}^{m} \frac{R_{ds}(\text{dataSetOf}(r_k))}{p_k}$$

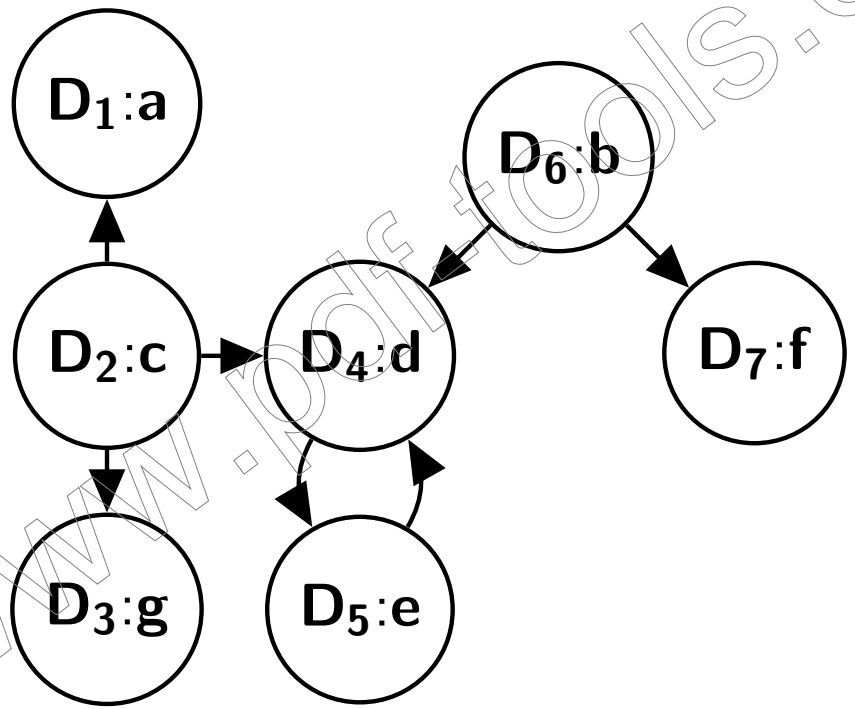
m = Number of resources, each from different datasets having type 2 links to r
$r_1$, $r_2$,...., $r_m$ are the resources, each from different datasets having Type 2 links to resource $r$
$p_k$ = Number of outgoing and non-bidirectional links from $r_k$ to other resources

- Total score:

$$T(r) = R_{mutual}(r) + R_{pa}(r)$$

# Example

- Example:




$$R_{mutual}(d) = R_{ds}(\text{dataSetOf}(e))$$
$$R_{pa}(d) = \frac{R_{ds}(\text{dataSetOf}(b))}{2} + \frac{R_{ds}(\text{dataSetOf}(c))}{3}$$
$$T(d) = R_{mutual}(d) + R_{pa}(d)$$

# Ranking Triples

- Triple ranking score
  - Triple is defined as subject - predicate - object

$$R_{triple} = (T(r_{subject}) + T(r_{predicate}) + T(r_{object}))/3$$

- Predicate score is not calculated using this scheme
  - Predicates do not use *owl:sameAs*
  - For experimental purpose we assigned ranking score 1 to well known predicates and 0 to all others

# Experimental Setup

- Billion Triple Challenge(BTC) dataset – 3.2 billion triples
- Chose BTC because most of the datasets of LOD were available in it
- Found 55 datasets present in both LOD and BTC dataset
- Out of these 55, only 31 datasets have links between datasets in LOD
- The experiment is performed on these 31 datasets, called BTC-cloud dataset (10 million triples)
- Used Allegrograph for storing *sameAs* network

# Preprocessing Data

- Convert N-Quads to N-Triples format
- Removed unwanted triples and triples containing literals
- Mapped resources to their domain names
- Found the common datasets between LOD cloud and BTC
- Divided BTC-cloud dataset into 31 sub-datasets for calculating score

# Results

| Dataset | DS-score | Dataset | DS-score | Dataset | DS-score | Dataset | DS-score |
|---|---|---|---|---|---|---|---|
| DBpedia | 0.5806 | Linkedmdb | 0.0967 | Jamendo | 0.0645 | Openguides | 0.0322 |
| Geonames | 0.2580 | DBLP Hannover | 0.0645 | Opencyc | 0.0645 | telegraphis (capitals) | 0.0322 |
| Musicbrainz | 0.1290 | Freebase | 0.0645 | Eurostat | 0.0322 | telegraphis (countries) | 0.0322 |
| Drugbank | 0.1290 | Lingvoj | 0.0645 | Project Gutenberg | 0.0322 | Crunch Base | 0 |
| Yago | 0.1290 | SW Conference Corpus | 0.0645 | Myspace | 0.0322 | DBLP (rkb) | 0 |
| Dailymed | 0.0967 | Factbook | 0.0645 | Umbel | 0.0322 | OS (rkb) | 0 |
| Linkedct | 0.0967 | DBLP Berlin | 0.0645 | Surgeradio | 0.0322 | Opencalais | 0 |
| Diseasome | 0.0967 | Revyu | 0.0645 | Wikicompany | 0.0322 | | |

- The above table shows the dataset score of each dataset used in the experiment
- DBpedia is the highly connected dataset, hence it gets the highest rank

# Results

| Resource(Dihydrofolate Reductase) | Ranking score |
|---|---|
| http://www4.wiwiss.fu-berlin.de/drugbank/resource/targets/365 | 0.2903225805 |
| http://dbpedia.org/resource/Dihydrofolate_reductase | 0.129032258 |
| http://mpii.de/yago/resource/Dihydrofolate_reductase | 0.0 |

| Resource(Apple Island) | Ranking score |
|---|---|
| http://sws.geonames.org/4984314 | 0.580645161 |
| http://dbpedia.org/resource/Apple_Island | 0.258064516 |

- The above tables shows the scores of resources identified by two different URI's
- Here we find that the entity from domain specific dataset is ranked higher than cross-domain datasets

# Results

select ?s where {
?s <rdf:type> <dbpediac:LandscapeArtist>}

| Subject (?s) | Predicate | Object | Triple score |
|---|---|---|---|
| dbpediar:Charles_Leickert | rdf:type | dbpediac:LandscapeArtists | 0.397849462 |
| dbpediar:Bernard_Hailstone | rdf:type | dbpediac:LandscapeArtists | 0.376344086 |
| dbpediar:John_Marin | rdf:type | dbpediac:LandscapeArtists | 0.376344086 |
| dbpediar:Lucius_Richard_OBrien | rdf:type | dbpediac:LandscapeArtists | 0.333333333 |

- The above table shows the result of the SPARQL query ordered by its triple score

# Conclusion

- Proposed a framework for ranking SPARQL query results
- Consensus play an important role in ranking semantic web data
- Traditional ranking methods cannot be applied to rank semantic web data
- The ranking scheme makes use of the statistics provided by LOD providers
- The ranking scheme can only be applied to objects and different method is needed to rank relationships

# Future Work

- An algorithm to update the ranking score locally, when a link is updated
- Extend the framework to rank predicates
- Use link discovery methods to discover new links and rank the resulting data
- The transitive nature of *owl:sameAs* is not used for any calculation, it may produce better results

**Thank You**

# References


- linkeddata.org
- Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net
- http://www.w3.org/TR/rdf-sparql-query
- http://www.w3.org/TR/rdf-concept
- http://www.franz.com/agraph/allegrograph/
- http://km.aifb.kit.edu/projects/btc-2010
- T. Neumann and G. Weikum RDF-3X: a RISC-style Engine for RDF. In PVLDB, 2008.
- L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Stanford Digital Library Technologies Project, 1998.
- J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, September 1999.