Querying the Web of Interlinked Datasets using VOID Descriptions

Ziya Akar
Tayfun Gökmen Halaç Erdem Eser Ekinci Oguz Dikenelli

SEAGENT Laboratory
Department of Computer Engineering
Ege University, İzmir, Türkiye

April 16, 2012

Contents

Contents

The Problem

- Web of Data is growing day by day.
  - 45 datasets in 2008, 295 datasets in 2011[1]
- To benefit from Linked Data, effective query engines are needed.
- Only all relevant datasets for a query must be queried to gather complete results in a reasonable time.

Solution

Metadata of all datasets and links between them and relationships of triple patterns in queries are considered to discover all relevant datasets for a query.

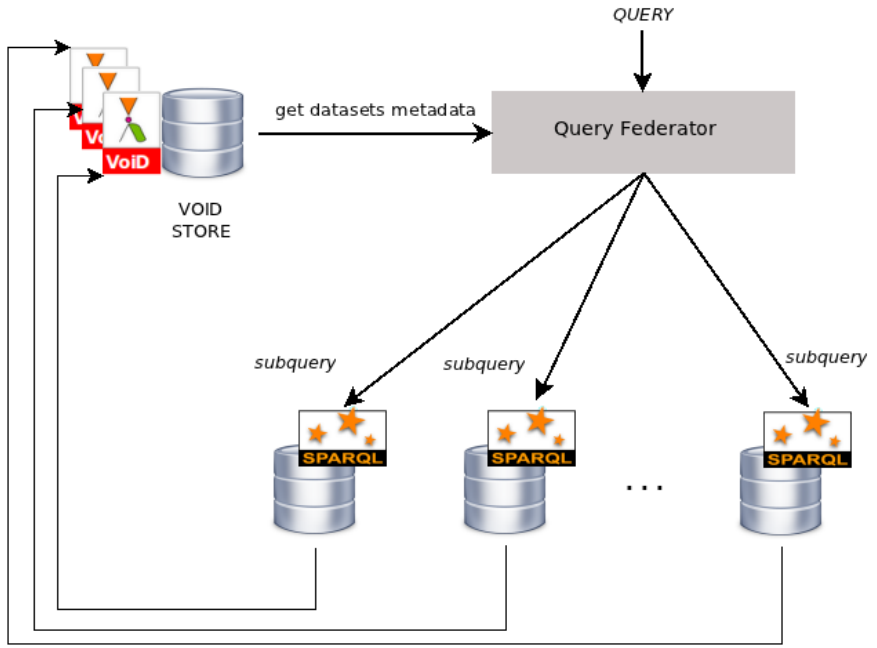
[1] http://richard.cyganiak.de/2007/10/lod/

Contents

Query Federation

- Query is divided into sub-queries respecting to dataset metadata files.
- Sub-queries are executed on selected relevant datasets.
- Requires sparql endpoints to access datasets.
- DARQ[1], FedX[2], SPLENDID[3]

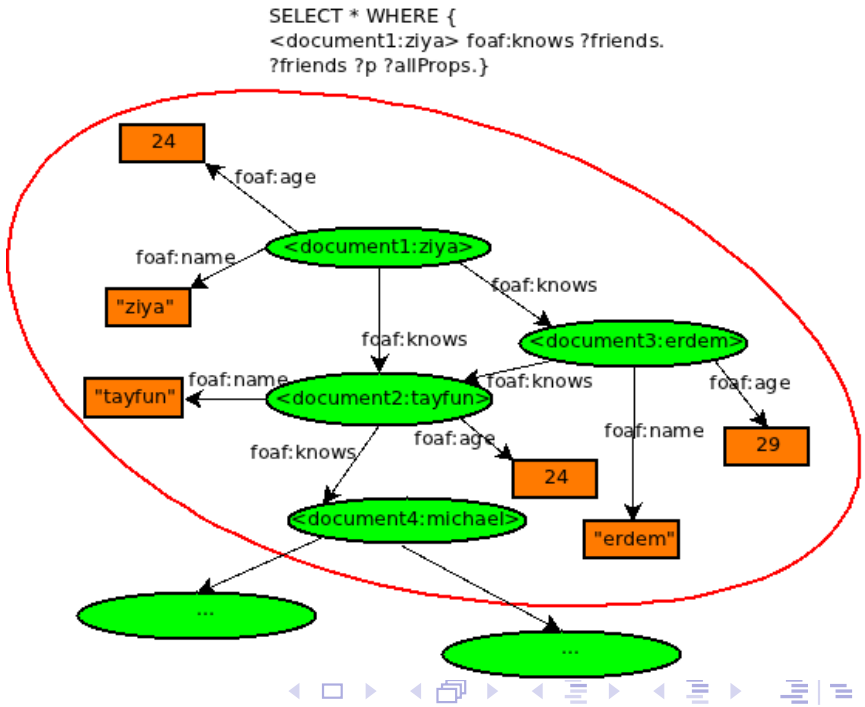
- Predicate and type indexes are not sufficient to select datasets effectively.
- Links are not taken into account.
- Using ASK queries to decide datasets increases cost.



Follow-Your-Nose

- No priori knowledge is required such as metadata of dataset.
- URIs in query are dereferenced and RDF documents are retrieved.
- Links are followed between resources in retrieved documents.
- SQUIN[4]

- Main disadvantages
  - Needs initial URIs
  - Infinite link discovery
  - Trying to retrieve large RDF graphs
  - Strictly depends on link evalution order



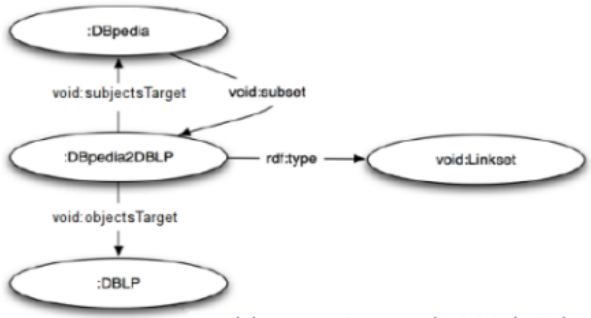
Contribution

- Our approach bases on query federation, but it incorporates follow-your-nose approach by means of VOID linkset descriptions.
- During dataset selection phase, analysis of triple pattern relationships is incorporated.
- Complete results are retrieved by querying minimum possible number of datasets.

Contents

VOID

- VOID[5] provides powerful way of description datasets. It describe links between datasets by void:Linkset concept.
- A VOID document primarily includes
  - Sparql endpoint
  - Urispace of included resources
  - Used vocabularies
  - Linkset which depicts relatives of dataset
  - Statistics about included triples.




```
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix void: <http://rdfs.org/ns/void#> .
@prefix dbp: <http://dbpedia.org/resource/> .

:DBpedia a void:Dataset ;
   foaf:homepage <http://dbpedia.org/> ;
   void:subset :DBpedia2DBLP .

:DBLP a void:Dataset ;
   foaf:homepage <http://dblp.l3s.de/d2r/> ;
   dc:subject dbp:Computer_science ;
   dc:subject dbp:Journal ;
   dc:subject dbp:Proceedings .

:DBpedia2DBLP a void:Linkset ;
   void:subjectsTarget :DBpedia ;
   void:objectsTarget :DBLP ;
   void:linkPredicate owl:sameAs .
:DBpedia void:sparqlEndpoint
     <http://dbpedia.org/sparql> ;
          void:feature [ dcterms:format
               "application/rdf+xml" ; ] ;
          void:vocabulary
               <http://xmlns.com/foaf/0.1/> .
```

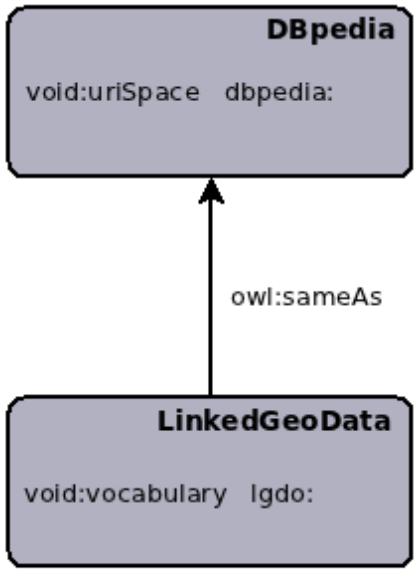
Contents

Case Study

AFTER ELIMINATION	QUERIED DATASETS
All	All
All	All
All	All
All	All
All	All
All	All
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```

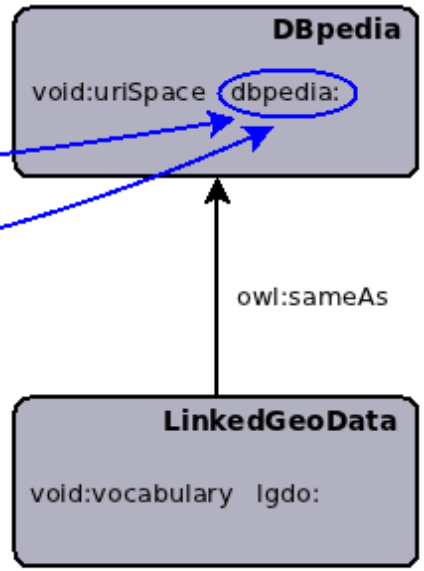


Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	~~All~~
DBpedia	~~All~~
All	All
All	All
All	All
All	All
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```

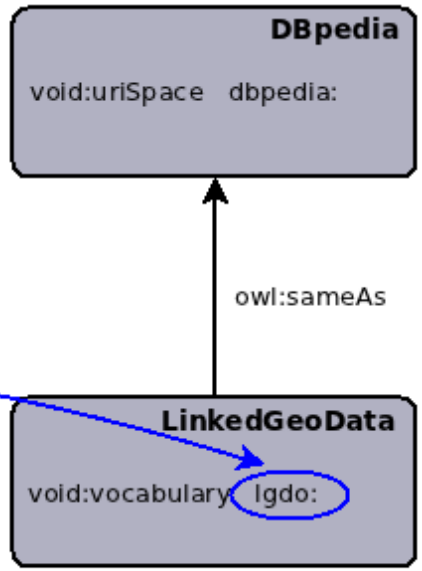


Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	~~All~~
All	All
All	All
All	All
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```

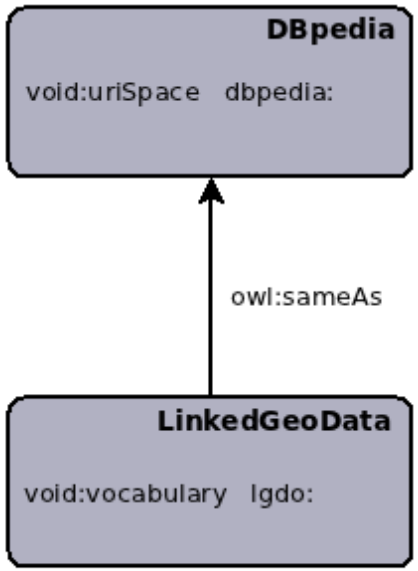


Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	LinkedGeoData
LinkedGeoData	~~All~~
All	All
All	All
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```



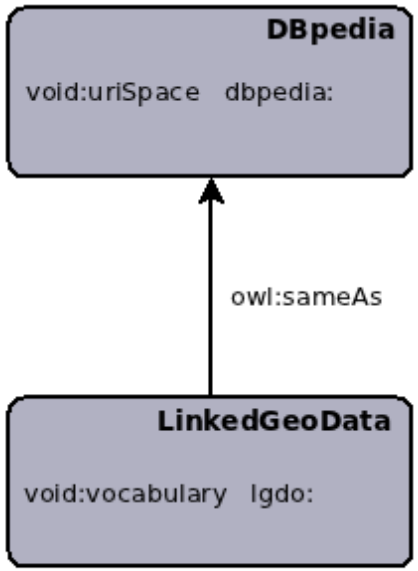
Subjects are the same.
They have to be queried from same datasets.
Thus, fourth triple pattern's datasets are eliminated

Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	~~All~~
All	All
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```



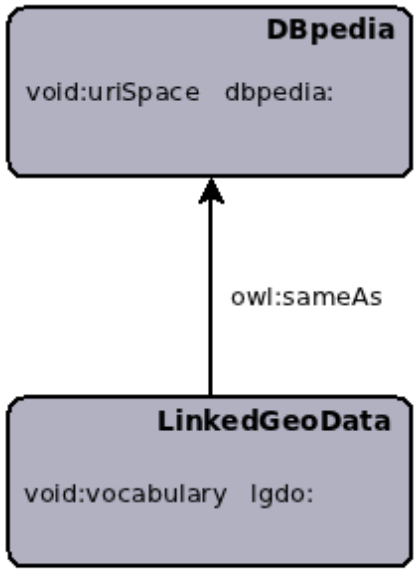
Subjects are the same.
They have to be queried from same datasets.
Thus, fifth triple pattern's datasets are eliminated

Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	~~All~~
All	All

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```



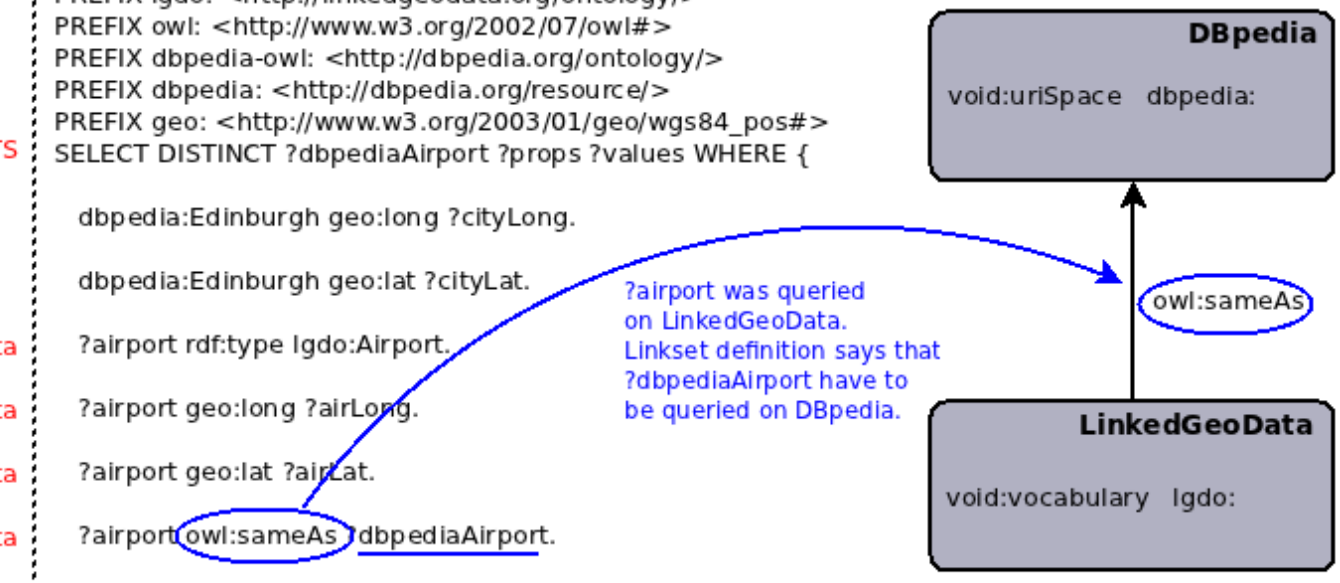
Subjects are the same.
They have to be queried from same datasets.
Thus, sixth triple pattern's datasets are eliminated

Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
DBpedia	~~All~~

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```

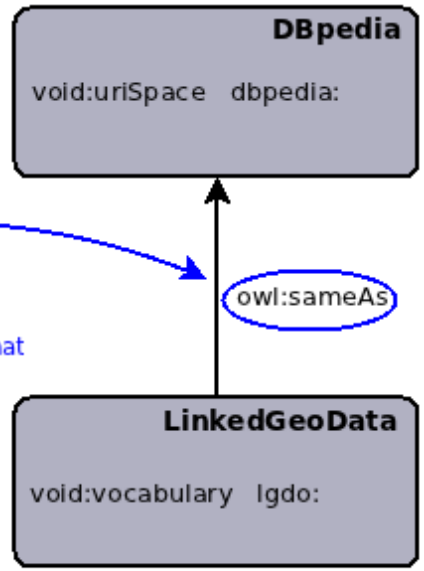


Case Study

AFTER ELIMINATION	QUERIED DATASETS
DBpedia	DBpedia
DBpedia	DBpedia
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
LinkedGeoData	LinkedGeoData
DBpedia	~~All~~

NEAREST AIRPORTS TO EDINBURGH

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX lgdo: <http://linkedgeodata.org/ontology/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia-owl: <http://dbpedia.org/ontology/>
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT DISTINCT ?dbpediaAirport ?props ?values WHERE {

  dbpedia:Edinburgh geo:long ?cityLong.

  dbpedia:Edinburgh geo:lat ?cityLat.

  ?airport rdf:type lgdo:Airport.

  ?airport geo:long ?airLong.

  ?airport geo:lat ?airLat.

  ?airport owl:sameAs ?dbpediaAirport.

  ?dbpediaAirport ?props ?values.

  FILTER(?cityLat-?airLat<1.5 &&
         ?cityLat-?airLat>-1.5 &&
         ?cityLong-?airLong>-1.5 &&
         ?cityLong-?airLong<1.5)
}
```



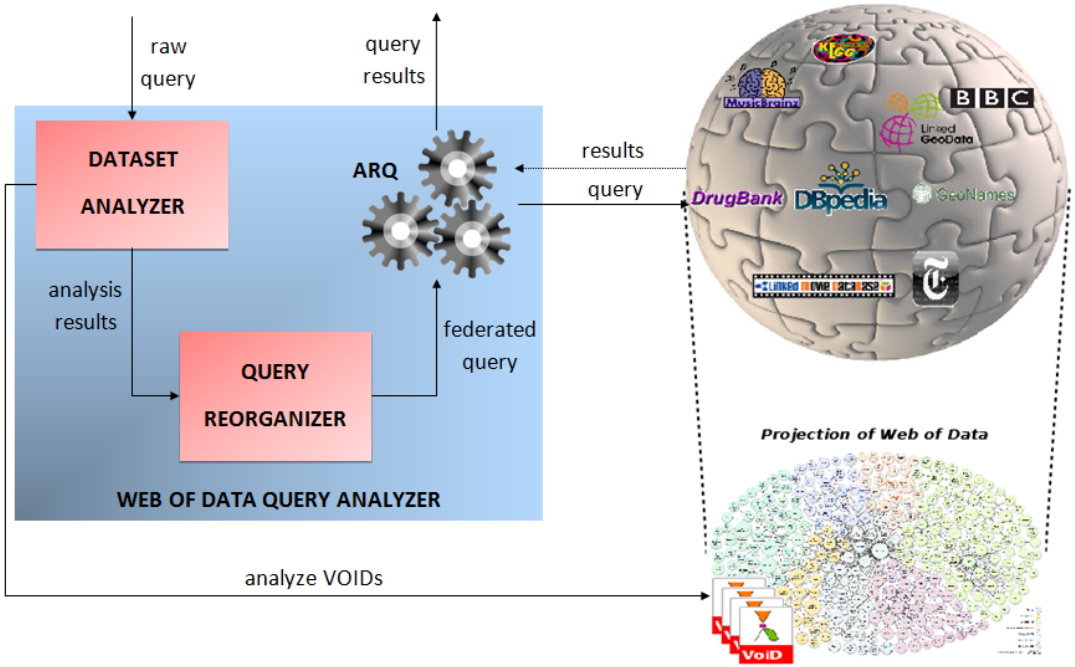
?airport was queried on LinkedGeoData. Linkset definition says that ?dbpediaAirport have to be queried on DBpedia.

Thus, last triple pattern must be queried from DBpedia.

Contents

Internal Architecture

- WoDQA uses VOID documents and triple patterns in query to discover relevant datasets.
- Main modules
  - Dataset Analyzer
  - Query Reorganizer
  - Jena ARQ

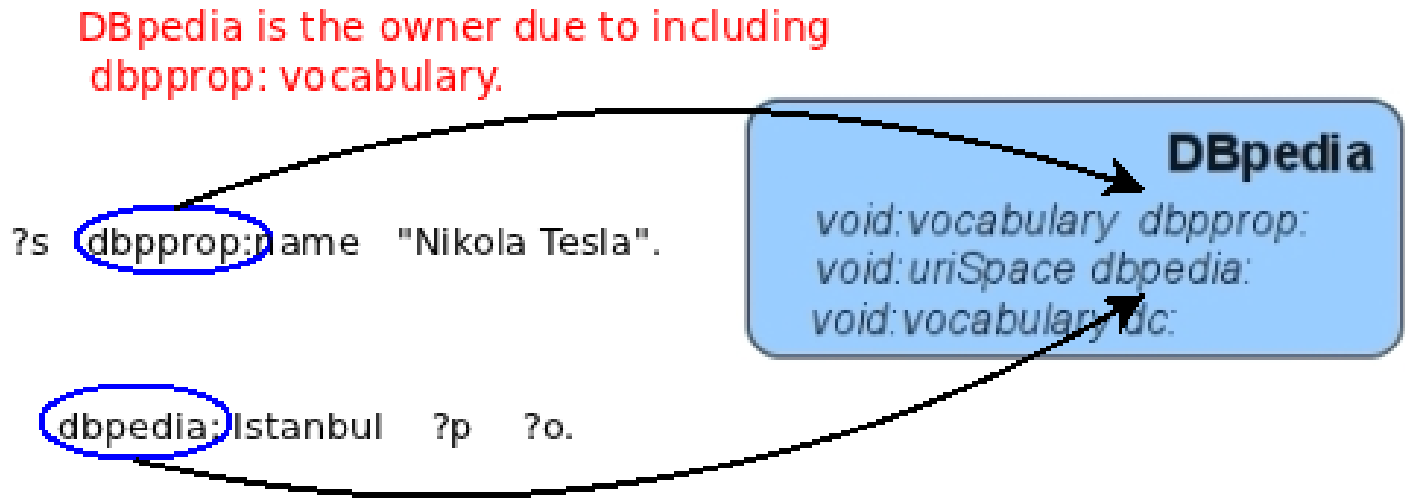


Dataset Analyzer

- WoDQA tries to find all possible related datasets for a triple pattern.
- It accepts all datasets are related for a triple pattern at first.
- Relevant datasets are found with some **discovery rules** and irrelevant ones are eliminated.
- There are 12 discovery rules under three **analysis perspectives**.
  - IRI-based Analysis
  - Linking Analysis
  - Shared Variable Analysis

IRI-based Analysis

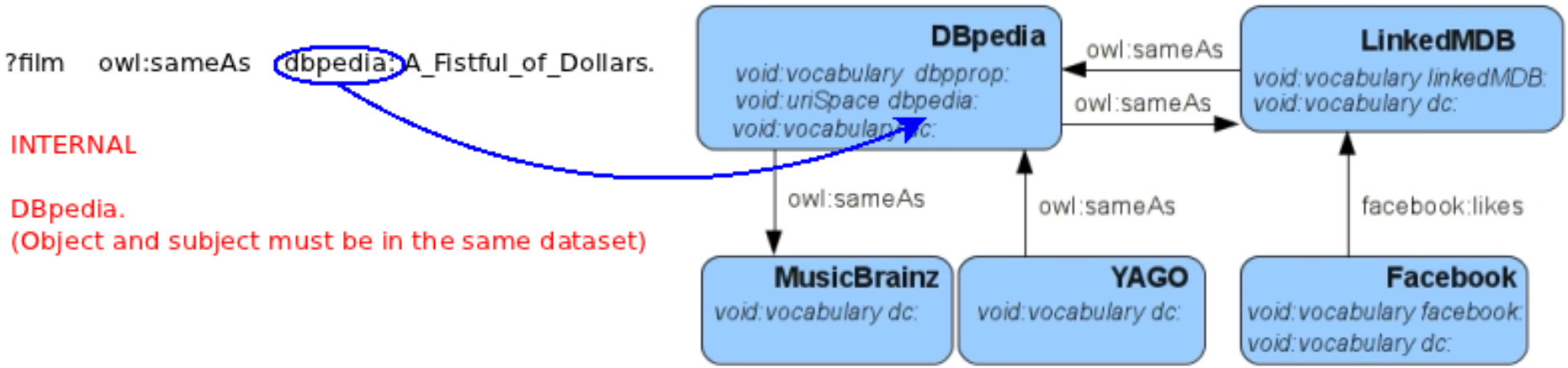
IRIs in triple patterns, void:uriSpace and void:vocabulary properties of VOIDs are considered.



Linking Analysis

Triples can link two resources which are in the same dataset (internal linking) or different datasets (external linking).

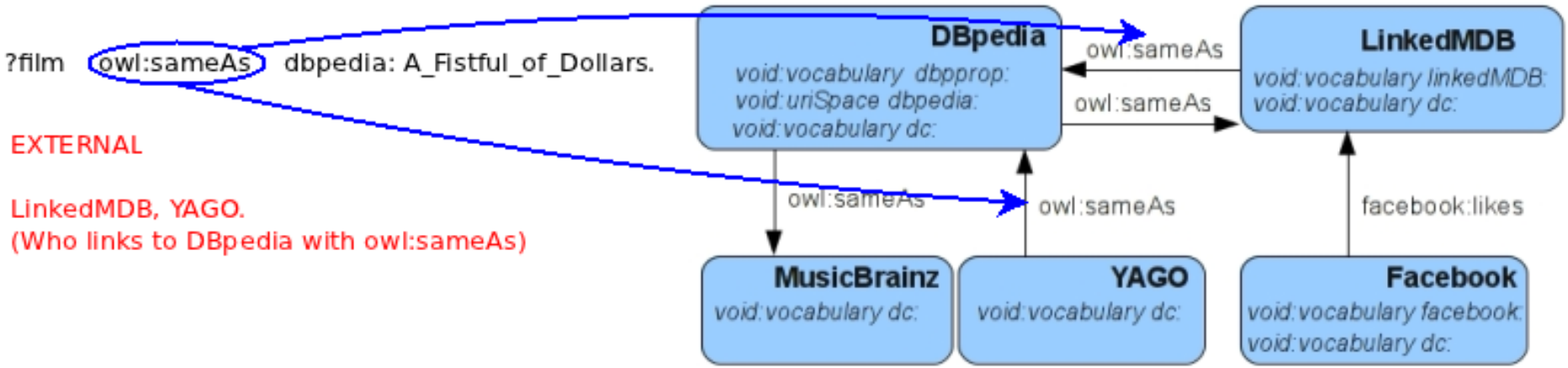
- **From the internal point of view,** subject and object must be in the same dataset. Thus, owner dataset is same and internal relevant.
- **From the external point of view,** void:Linkset descriptions of VOIDs helps us to find which dataset links to which.



Linking Analysis

Triples can link two resources which are in the same dataset (internal linking) or different datasets (external linking).

- **From the internal point of view,** subject and object must be in the same dataset. Thus, owner dataset is same and internal relevant.
- **From the external point of view,** void:Linkset descriptions of VOIDs helps us to find which dataset links to which.



Shared Variable Analysis

- Triples can contain common variables and causes to bound queried datasets of each other.
- They are analyzed together in this perspective.
- Three types of shared variable analysis patterns
  - Object of one can be subject of another one.
  - They can share same subject.
  - They can share same object.

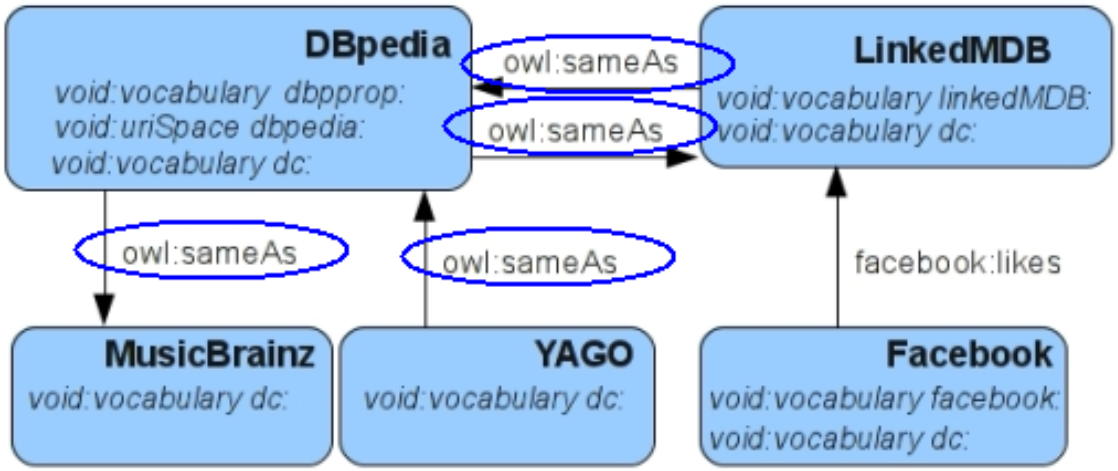
Shared Variable Analysis

Object in subject position

tp1: ?s owl:sameAs ?film.
tp2: ?film linkedMDB:producer_name "Sergio Leone".

EXTERNAL DATASETS for tp1 with LINKING ANALYSIS

DBpedia, LinkedMDB, YAGO



Shared Variable Analysis

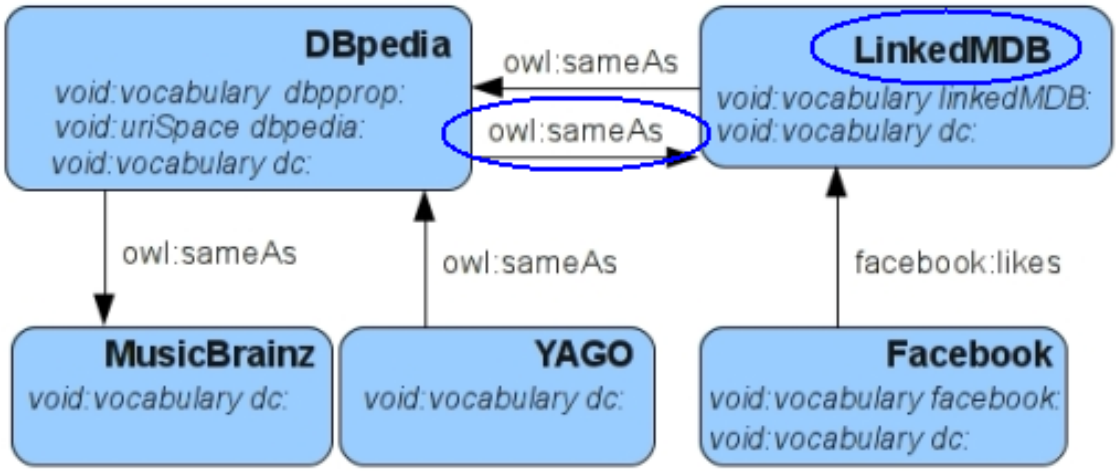
Object in subject position

tp1: ?s owl:sameAs ?film.
tp2: ?film linkedMDB:producer_name "Sergio Leone".

EXTERNAL DATASETS for tp1 after SHARED VARIABLE ANALYSIS

DBpedia, ~~LinkedMDB~~, ~~YAGO~~

(?film must be in LinkedMDB due to IRI-based Analysis
Just DBpedia links with owl:sameAs to LinkedMDB)



Query Reorganzier

According to the analysis results of Dataset Analyzer module, Query Reorganizer module rewrites queries into a federated form executable by Jena ARQ.

Initial Query

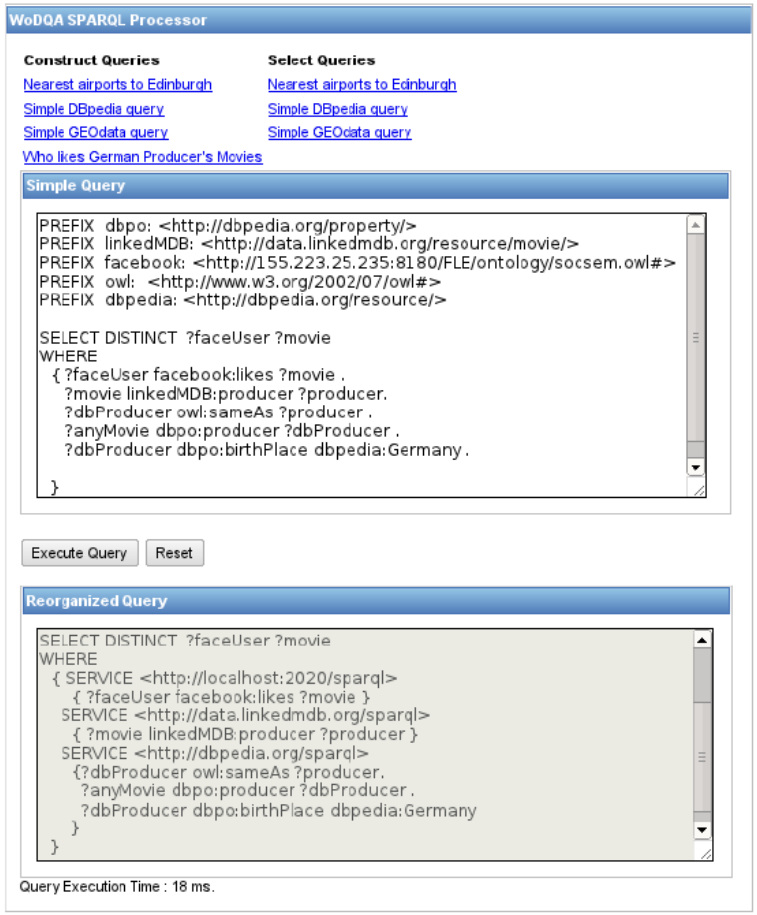
```
PREFIX dbpo: <http://dbpedia.org/property/>
PREFIX linkedMDB: <http://data.linkedmdb.org/resource/movie/>
PREFIX facebook: <http://155.223.25.235:8180/FLE/ontology/socsem.owl#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dbpedia: <http://dbpedia.org/resource/>

SELECT DISTINCT ?faceUser ?movie
WHERE
 { ?faceUser facebook:likes ?movie .
   ?movie linkedMDB:producer ?producer.
   ?dbProducer owl:sameAs ?producer .
   ?anyMovie dbpo:producer ?dbProducer .
   ?dbProducer dbpo:birthPlace dbpedia:Germany .

 }
```

Reorganized Query

```
SELECT DISTINCT ?faceUser ?movie
WHERE
 { SERVICE <http://localhost:2020/sparql>
    { ?faceUser facebook:likes ?movie }
  SERVICE <http://data.linkedmdb.org/sparql>
    { ?movie linkedMDB:producer ?producer }
  SERVICE <http://dbpedia.org/sparql>
    {?dbProducer owl:sameAs ?producer.
     ?anyMovie dbpo:producer ?dbProducer .
     ?dbProducer dbpo:birthPlace dbpedia:Germany
    }
 }
```

WoDQA Web Form

- http://etmen.ege.edu.tr/etmen/wodqa.html



Summary

- Exhaustive dataset selection on the Web of Data
  - VOID provides a powerful metadata of the Web of Data
  - Relationships between triple patterns give tips about relevant datasets
- Future
  - Automated extraction and management of VOID descriptions
  - Evaluation of our approach on the LOD cloud
  - Incorporating statistics in VOID to optimize performance
  - Discovering sameAs relationships automatically
  - Dealing with heterogenous vocabularies
- Discussion
  - Dataset availability
  - Considering queries about vocabularies

Thank you!

References I


Bastian Quilitz and Ulf Leser.
Querying Distributed RDF Data Sources with SPARQL.
The Semantic Web: Research and Applications, volume 5021 of Lecture Notes in Computer Science, chapter 39, pages 524-538.


Andreas Schwarte, Peter Haase, Katja Hose, Ralf Schenkel, and Michael Schmidt.
Fedx: A federation layer for distributed query processing on linked open data.
The Semanic Web: Research and Applications, volume 6644 of Lecture Notes in Computer Science, pages 481–486. Springer Berlin / Heidelberg, 2011.

References II


Olaf Görlitz and Steffen Staab.
SPLENDID: SPARQL Endpoint Federation Exploiting VOID Descriptions.
In Proceedings of the 2nd International Workshop on Consuming Linked Data, Bonn, Germany, 2011.


Olaf Hartig, Christian Bizer, and Johann Christoph Freytag.
Executing sparql queries over the web of linked data.
In International Semantic Web Conference, pages 293–309, 2009.


K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao.
Describing Linked Datasets - On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'.
In WWW 2009 Workshop: Linked Data on the Web (LDOW2009), Madrid, Spain, 2009.