CS561

Web Data Management

Spring 2013

Assignment 2

Professor: Vassilis Christophides

Deadline: **22/04/2013**

Introduction

The most central data sources in the linked data cloud is DBpedia, a big knowledge base which is essentially a "translation" of parts of Wikipedia into RDF.

The DBpedia data set uses a *large multi-domain ontology*, which has been derived from Wikipedia. The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases.

Each thing in the DBpedia data set is *denoted* by a de-referrenceable IRI- or URI-based reference of the form http://dbpedia.org/resource/Name, where *Name* is *derived* from the URL of the source Wikipedia article, which has the form http://en.wikipedia.org/wiki/Name.

Every DBpedia entity name resolves to a description-oriented Web document (or Web resource). Each DBpedia entity is described by various properties.

DBpedia provides three different classification schemata for things:

- Wikipedia categories: represented using the SKOS vocabulary and DCMI terms.
- The YAGO Classification: derived from Wikipedia category system using Word Net [details]
- Word Net Synset Links: generated by manually relating Wikipedia infobox templates and Word Net synsets, and adding a corresponding link to each thing that uses a specific template.

Using these classifications within SPARQL queries allows you to select things of a certain type.

You can find the current ontology classes of DBpedia in http://mappings.dbpedia.org/server/ontology/classes/.

The DBpedia data set is interlinked with various other data sources. Figure 1 gives an overview of some of these data sources *(as of September 2011)*.

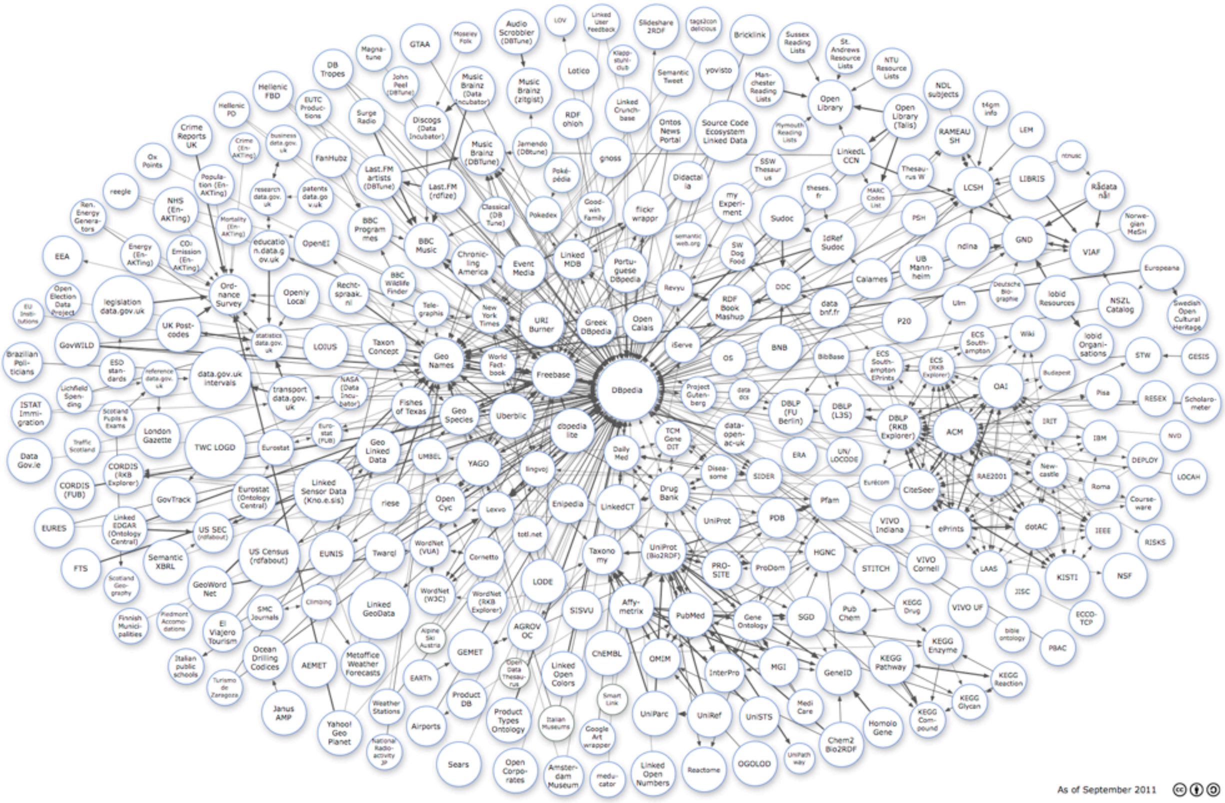

Figure 1 - Interlinking DBpedia

You can find more information about the structure of DBpedia on http://wiki.dbpedia.org/Datasets.

In this assignment, you are asked to display your understanding on the concepts of RDF/RDF(S) and querying RDF databases using SPARQL. You will do that by querying the DBpedia knowledge base.

Tools

For this assignment, you will have to use the Virtuoso SPARQL Query Editor*, available online at the following webpage:

http://dbpedia.org/sparql

* In case of downtime of the Virtuoso SPARQL Query editor, you can find alternative endpoints at: http://dbpedia.org/OnlineAccess#h28-3

This query editor provides a public SPARQL endpoint over the DBpedia data set. With this tool, you will be able to test your queries over the knowledge base and export the results in various formats.

You will not need to download any aditional software.

For the needs of this assignment, you will probably have to use the following namespace prefixes:

category, dbpedia, dbpedia-owl, dbpprop, foaf, freebase, owl, rdf, skos, xsd, yago

You can find the above prefixes and their mappings, along with many other, in http://dbpedia.org/sparql?nsdecl.

Querying DBpedia using SPARQL

The goal of this assignment is to express the following queries in SPARQL format and run them using the Virtuoso SPARQL endpoint over the DBpedia knowledge base.

Browse the knowledge captured by DBpedia by starting from a resource that you are familiar with (e.g. the actor Jack Nicholson) and follow the links to other DBpedia resources. To familiarize yourselves with querying DBpedia using SPARQL and a SPARQL endpoint, run the example queries found at http://wiki.dbpedia.org/OnlineAccess?v=53r.

Afterwards, express the following queries in SPARQL:

- **Query 1**
  Find all movies starring this year's winner of the Oscar for Best Actress. Output the movie names and the names of the co-starring actors and actresses in each movie.

  (you can use *dbpedia-owl, dbpedia, foaf, dbpprop* prefixes and the properties *starring* and *name*)

- **Query 2**
  Find the books of the Harry Potter series that have more than 800 pages. Output the book titles and their page count.

  (you can use *dbpedia, dbpprop* prefixes and the properties *books, pageCount* and *name*)

- **Query 3**
  Find the undisputed boxing champions. Output the number of undisputed boxing champions grouped by nationality.

  (you can use *rdf, dbpedia-owl, dbpedia, dbpprop* prefixes and the properties *type, Boxer, title, nationality* and *name*)

- **Query 4**
  Find all the current players of Barcelona FC who where born after 01/01/1990. Output the names of the players and their birth dates. Also output the given name of the players, where it is available.

  (you can use *dbpedia-owl, dbpedia, xsd, dbpprop* prefixes and the properties *currentClub, birthDate, fullname, givenName* and *date*)

- **Query 5**
  Find all movies that last longer than two hours. Output the movie names, their runtime and their director.

  (you can use *rdf, dbpedia-owl, dbpprop* prefixes and the properties *type, runtime, Film, director* and *name*)

- **Query 6**
  Find the capitals of Greece, Italy, Spain and France. Output the names of the capitals and their postal codes.

  (you can use *rdf, dbpedia-owl, dbpedia, yago, dbpprop* prefixes and the properties *type, CapitalsInEurope, country, postalCode* and *name*)

- **Query 7**
  Find all software developed by Microsoft that is available for Mac OS X. Output the names of the software.

  (you can use *dbpedia-owl, dbpedia, dbpprop* prefixes and the properties *developer* and *operatingSystem*)

- **Query 8**
  Find all persons born in Greece. Filter out those for whom a title is not provided by DBpedia. Output the names and the birth dates of the remaining persons.

  (you can use *dbpedia-owl, dbpedia, dbpprop* prefixes and the properties *birthPlace, birthDate, title* and *name*)

- **Query 9**
  Find the three oldest Beatles members. Output their names.

  (you can use *dbpedia-owl, dbpedia, dbpprop* prefixes and the properties *bandMember, birthDate* and *birthName*)

- **Query 10**
  Find the member countries of the European Union. Remove from the query result the countries that their names do not contain the string "land". Output the names of the remaining countries.

  (you can use *category, foaf, dcterms, dbpprop* prefixes and the properties *subject, European_countries, homepage* and *name*)

Files to submit

The files you should submit to the course's email address (hy561@csd.uoc.gr) are the following:

1. A text file containing the above queries expressed in SPARQL. In each query you will have to include the prefixes you used.
2. 10 XML files, each one containing the results for the corresponding query.

Further reading

1. DBpedia: A Nucleus for a Web of Open Data
2. DBpedia - A Crystallization Point for the Web of Data
3. DBpedia overview (presentation)
4. DBpedia presentation