Practical session - Data integration course 3

First part - Linked Open Data: Exploration of the dataset DBpedia.

DBpedia represents the factual knowledge contained in Wikipedia infoboxes. The 2016-04 release contains 28,658,449 instances (http://wiki.dbpedia.org/dbpedia-2016-04-statistics): 299,371 cities, 515,480 organisations, 3,218,716 persons ....
The RDF descriptions are extracted from Wikipedia editions written in 127 languages.

1. Look at one DBpedia page ....

URL http://dbpedia.org/page/Paris is the RDF description of Paris.

- This instance belongs to which classes of the DBPEDIA ontology? What is the most specific class?

- Are there some ontology mappings declared with another ontology?

- What is the property dbpedia-owl:abstract in Wikipedia?

2. Some Graphical tools to explore the LOD the example of relfinder

RelFinder extracts and visualizes relationships between given objects in RDF data and makes these relationships interactively explorable: http://www.visualdataweb.org/relfinder.php (try demo)

Search for links between the resource *France* and the resource *Francois Hollande*. What is the only path of length 0? Replace France by Paris and add the discovered links.

Related publication: *Philipp Heim, Steffen Lohmann and Timo Stegemann. Interactive Relationship Discovery via the Semantic Web, Proceedings of the 7th Extended Semantic Web Conference (ESWC 2010), volume 6088, series LNCS, pages 303-317. Springer, Berlin/Heidelberg, 2010.*

3. A *SPARQL endpoint* is a SPARQL protocol service, which enables users (human or application) to query a knowledge base via the SPARQL language. Results are returned in one or more machine-processable formats. Differents *SPARQL endpoint* can be used to query DBpedia: Virtuoso, OpenLink, Disco.

At this address, you can find a list of available SPARQL endpoints for various LOD datasets:

https://www.w3.org/wiki/SparqlEndpoints

A Sparql tutorial is available at:

http://www.cambridgesemantics.com/2008/09/sparql-by-example/

In the following, you will use the SPARQL endpoint DBpedia that exploits Virtuoso:
http://dbpedia.org/sparql

The predefined following prefixes can be used to simplify the query :

```
PREFIX dbp: <http://dbpedia.org/property/>
PREFIX dbr: <http://dbpedia.org/resource/>
```

3.1 How many different formats can be chosen to represent the answers?

3.2 Express and test the following queries:

a. Search for URI of persons born in France (property <http://dbpedia.org/property/birthPlace>), using the HTML format.

URI of *France* in DBpedia is the following:
http://dbpedia.org/resource/France

URI of property *birth place* is:
<http://dbpedia.org/property/birthPlace>)

b. Number of persons born in France:

```
SELECT count(DISTINCT ?personne)
{
     {?personne <http://dbpedia.org/property/birthPlace>
<http://dbpedia.org/resource/France>.}
}
```

Same query with the property *placeOfBirth:*
http://dbpedia.org/property/birthPlace.
What does it shows about the DBpedia dataset?

c. Define a simple SPARQL query that looks for URI of instances linked by a sameAs link with another URI (property: owl :sameAs). Use the example in Paris in domain then in range position for this property. Do you think that this property is symmetric? is DBpedia a saturated dataset ?

d. Count the number of entities in DBPedia for which the foaf:name is Paris (i.e 'Paris '@en). Is the property foaf:name an inverse functional property? (a key).

e. Define a query that allows knowing if the river *Amazon* is longer than the *Nile* (in Wikipedia).

```
ASK
{
  <http://dbpedia.org/resource/Amazon_River> dbprop:length
?amazon .
  <http://dbpedia.org/resource/Nile> dbpprop:length ?nile .
  FILTER(?amazon > ?nile) .
}
```

4. Publication of RDF data on the LOD. **Datalift** is a platform that can be used to publish and interlink datasets on the web of data. It provides a set of tools to facilitate the process of publishing linked datasets. The input data are raw data coming from multiple heterogeneous formats (databases, CSV, XML, RDF, RDFa, Shapefile, …). The output data produced are RDF « Linked Data ».

http://datalift.org

Related publication: *François Scharffe, Ghislain Atemezing, Raphaël Troncy, Fabien Gandon, Serena Villata, Bénédicte Bucher, Fayçal Hamdi, Laurent Bihanic, Gabriel Képéklian, Franck Cotton, Jérôme Euzenat, Zhengjie Fan, Pierre-Yves Vandenbussche, Bernard Vatant, Enabling linked data publication with the Datalift platform, in: Proc. AAAI workshop on semantic cities, Toronto (ONT CA), 2012.*

D2R server: If the original data is available in a relational database, a user can exploit **D2R Server** to publish the data along with its schema as Linked Data. This tool is part of the D2RQ Platform, a system for accessing relational databases as virtual, read-only RDF graphs. It offers RDF-based access to the content of relational databases without having to replicate it into an RDF store (http://d2rq.org).

5. Semantic annotation of unstructured (textual) data

Example of **DBpediaSpotlight** (Free University, Berlin). DBpedia Spotlight is a tool for automatically annotating mentions of DBpedia resources in text. DBpedia Spotlight looks for ~3.5M things of unknown or ~320 known types in text and tries to link them to their identifiers (URI) in DBpedia. DBpedia Spotlight is publicly available as a web service for testing purposes or a Java/Scala API licensed via the Apache License.

http://dbpedia-spotlight.github.io/demo/

5.1 Using this demo, annotate the first paragraph of the english wikipage that describe the city of Paris(France). Choose the option n-best candidate and put the lower confidence to see all the candidates with their associated confidence. Note that the annotated types can be chosen using the *select types* option.

5.2 Annotate the first paragraph of the english wikipage that describe Paris (mythology), the son of priam. Do you think that this tool exploits the context in which the word appears to disambiguate?

Related publication: *Joachim Daiber, Max Jakob, Chris Hokamp and Pablo N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction, Proceedings of the 9th International Conference on Semantic Systems (I-Semantics) 2013.*

6. Data modelling-Ontology reuse: ontology search engines.

The ontology search engine main objective is to help users of linked data and vocabularies to assess what is available for their needs, to reuse it as far as possible.

LOV (http://lov.okfn.org/dataset/lov/) has been developped in the setting of the Datalift project.

Look for the geonames ontology using LOV (use vocabs), Are ontology elements more reused than foaf (look a the incoming and outcoming links)? How many ontologies have

defined the property *birthDate* (use terms).
Other ontology (or semantic document) search engines: Falcons, Watson, Swoogle.