Open Data Project Report
Martinello Pierfrancesco

Introduction

The purpose of the project was to clean open data about cultural points of interest located in the regions Sardinia and Puglia and in the city of Palermo, to create an ontology for these data and linking them to other knowlegde bases, consenquentially creating 5 stars data.

About the data

The data were taken from the Open Data Portal of Palermo, for the data redarding the over mentioned city and from the Italian Open Data Portal for the two regions taken in exam . The data downloaded from these sources were heterogeneous among them and a cleaning operation was considered necessary in order to delete unnecessary columns and errors due to different encoding. The original files were kept intact and shared with the proejct.

Licenses

The data taken in exam were distrubuted with the following licenses:

- CC BY 4.0 for the data about the city of Palermo;
- IODL 2.0 for the remaining data.

These two liceces are equivalent and both allows complete freedom on sharing and modification of the data and the creation of derivative work, at the only condition of recognize the paternity of the data to their respective owners.

Data Cleaning

As it was mentioned before an operation of data cleaning is been necessary. To do that is been used the cleaning software OpenRefine, that allows to simplify the opereation of data cleaning with its instruments for text recognition. In fact it was thus possible to standardize the fields who did not have any information, replacing them with blank strings. This instrument was preferred to automatic cleaning programs made using the Python language because some of the data did contain the symbol ¿ due to an encoding error. It represented both the symbol – and ‘, thus an automatical program could have wrongly prefer one to the other. More details on the file called *modifications_log.txt* .

Four stars data creation

The cleaned date were then processed by a Python program, using the library RDFLib in order to create a knowlegde graph, insert data as nodes of the said graph and then serialize the result into a *Turtle* file. Two versions of the program are purposed, one where the created ontology was explicitly inserted into the knowledge graph before the data were processed and one where the ontology is implicit (and described in the file *implicit_ontology.txt*).

URIs

The fake domain used is http://poifinder.it . There’s a section dedicated to the ontology at this URI http://poifinder.it/ontology/ and a section for the resources at the URI http://poifinder.it/resource/ . When a series of nodes is added, it is available at the URI http://poifinder.it/resource/{city_name}/{poi_name}

Link to other knowlegde bases and five stars data creation

There are links to a extern knowledge base and to an RDF vocabulary. Said vocabulary is Geo and it is been introduced in order to have a better manipulation over geographical coordinates. The external knowledge base used is DBPedia.org and it is been used in order to create links between the ontology and other knowledge sources. For each point of interest, the locality where is settles is directly linked to its corresponding city in DBPedia (e.g. for the city of Palermo, the URI associated is http://dbpedia.org/resource/Palermo).

Results

The knowledge base obtained can be easily used in order to show information about the Points of interest using SPARQL queries and it could be linked to online services such as https://www.openstreetmap.org/ in order to obtain a route between the position given by the final user to the coordinates of the point of interest.

This documentation, the original files, the programs and the resulting file are shared with licence CC BY 4.0 which is compatible to the licences of the original files.