# Tourism in Apulia: Driving Open Data towards an Apulian Tourism Ontology

Pierpaolo Basile*, Leo Iaquinta†, Rita Schiralli*, Lucia Siciliani* and Giovanni Semeraro*
*Department of Computer Science, University of Bari Aldo Moro, Italy
Email: {pierpaolo.basile,giovanni.semeraro}@uniba.it, r.schiralli2@studenti.uniba.it, siciliani.lu@gmail.com
†So.Co.In. System, Noci, Italy
Email: leo.iaquinta@gruppopace.com

According to the current Italian regulation the open data have become a "legal obligation" and public administrations have published on-line databases by making their open data. Generally, these datasets are not linked to the Linked Open Data (LOD) cloud [1] on the web, but they are provided as non-proprietary open format (e.g., CSV instead of Excel). Unfortunately, the data openly available is not very useful on its own. Data value increases when data is linked because a certain amount of semantics is made explicit and hence it can be exploited in advanced data-driven applications.

In 2007, the Linked Open Data project was launched to stimulate researchers and organizations publishing their data in RDF format and adopting shared vocabularies, in order to express an agreed semantics and interlink the data to each other. Nine years later, 150 billions of RDF triples and almost 10,000 linked datasets are available on the Web, thus representing a rapidly growing piece of the big data puzzle. These interconnected RDF statements form a huge decentralized knowledge base, called Linked Open Data (LOD) cloud. The LOD cloud covers many topical domains, ranging from government and geographical data to structured information about media (movies, books, etc.) and life sciences. The typical entry point to these data is DBpedia [2], the RDF mapping of Wikipedia which is commonly considered as the nucleus of the emerging Web of Data.

Data that contain link to other ones are the best format (five stars) in the open data classification proposed by Tim Barnser-Lee, however public administrations hold data in structured/relational format that need some processing for obtaining five stars open data. The transformation of data in linked data is a time consuming task and requires some background knowledge about semantic web technologies. In this paper, we propose a case study in which a set of Apulia open data about tourism are transformed in five stars open data and they are linked to other datasets in the LOD. The output of the transformation process is an OWL ontology containing 40 classes and about 36,300 instances linked to some datasets in the LOD, such as: DBpedia, Geonames, Km4city [3], Proton and FOAF.

The transformation process requires three preliminary steps:

1) the selection of exiting datasets already published as open data;
2) the assessment of selected datasets;
3) the ontology designing.

For the first step, we analyze the open data website[1] of the Apulia region. The site contains several datasets in CSV format, we analyzed each dataset in order to assess its relevance with respect to the tourism domain. The results of this analysis is summarized in Table I.

TABLE I
NUMBER OF DATASETS CLASSIFIED ACCORDING TO THE LEVEL OF RELEVANCE.

| | low | medium | high |
|---|---|---|---|
| dataset.puglia.it | 91 | 5 | 6 |

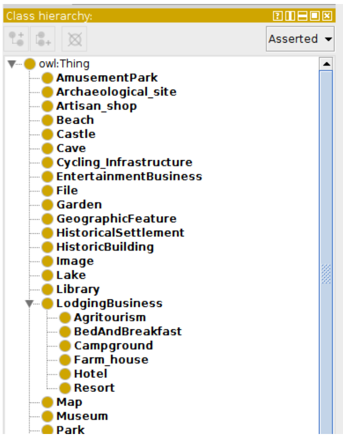
LOD datasets are based on the use of shared vocabularies and ontologies that allow to structure and describe the data using a common set of resources. Ontologies and vocabularies are the building blocks for creating formalized and reusable data. Linking and reusing them contributes to the growth of Linked Open Vocabularies (LOV). Following this idea for the design of the ontology, we select from the LOV a set of ontologies and vocabularies that are suitable for both the tourism domain and the datasets available on the regional website. After a deep analysis, we select the following ontologies and vocabularies:

- DBpedia[2] is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web. DBpedia provides us commons sense concepts such as *Beach*, *Castle* and *Cave*;
- Geonames[3] is a geographical database available under a creative commons license that contains over 10 million geographical names and consists of over 9 million unique features whereof 2.8 million populated places and 5.5 million alternate names. Geonames is used to geo-localize point of interest (POI) and other geographical instances;
- Geo[4] is the RDF vocabulary that provides the Semantic Web community with a namespace for representing lat(itude), long(itude) and other information about spatially-located things, using WGS84 as a reference datum;

[1] http://www.dataset.puglia.it
[2] http://wiki.dbpedia.org
[3] http://www.geonames.org
[4] https://www.w3.org/2003/01/geo

- Shema.org[5] is a general purpose vocabulary that covers entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. Over 10 million sites use Schema.org to markup their web pages and email messages;
- Km4city[6] is a knowledge base developed by the DISIT lab of the University of Florance containing information about the open data of Florence Municipality regarding weather, POI, digital location, cultural activities, schools, commercial activities, etc.;
- PROTON (PROTo ONtology)[7] is a lightweight upper-level ontology, serving as a modeling basis for a number of tasks in different domains;
- FOAF[8] is an ontology that describes persons, their activities and their relations to other people and objects. It is widely used on the Web.

We model the ontology "Tourism in Apulia" by partially reutilizing these ontologies and introducing new concepts: not always it is possible to find a match in LOVs, as in the case of *cycleways*. Classes and proprieties have been manually created by using the Protégé software[9]. In order to transform CSV files into ontological data, we need to map the information in CSV to classes and proprieties in the ontology. Starting from the dataset about POI in Apulia region[10] we devise a set of procedures to implement the mapping process of tabular data into ontology-based knowledge in order to finally achieve the linked version of selected datasets. Figure 1 shows a snapshot of the class hierarchy, we model several kind of POI for example *Castle*, *Historic Building*, *Beach*. We add some second level classes, for example the class *Lodging Business* has several sub classes, such as *Agriturism*, *Bed&Breakfast*, *Hotel*, etc.



Class hierarchy:

Asserted

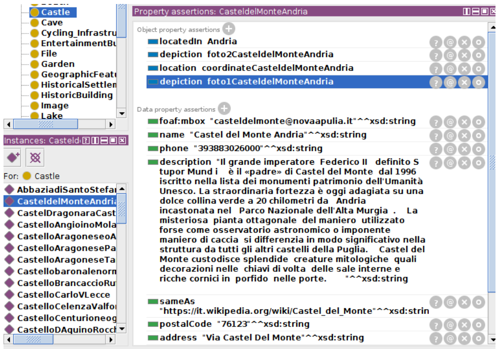
- owl:Thing
  - AmusementPark
  - Archaeological_site
  - Artisan_shop
  - Beach
  - Castle
  - Cave
  - Cycling_Infrastructure
  - EntertainmentBusiness
  - File
  - Garden
  - GeographicFeature
  - HistoricalSettlement
  - HistoricBuilding
  - Image
  - Lake
  - Library
  - LodgingBusiness
    - Agritourism
    - BedAndBreakfast
    - Campground
    - Farm_house
    - Hotel
    - Resort
  - Map
  - Museum
  - Park

Fig. 1. A snapshot of the class hierarchy.

Then, we define some object proprieties:

- *locatedIn*: used to link a POI to a geographic feature taken from Geonames;
- *location*: used for geo-localize objects according to Geo;
- *depiction*: allows to link an image to each object.

Figure 2 reports the description of the POI instance *Castel del Monte*, object properties are reported in blue, while data properties are in green. We can note that the *foaf:mbox* property is used to model the e-mail and the *sameAs* property links to the Wikipedia page.



Castle
Cave
Cycling_Infrastru
EntertainmentBu
File
Garden
GeographicFeatu
HistoricalSettlen
HistoricBuilding
Image
Lake

Instances: Casteld

For: Castle

AbbaziadiSantoStefa
CasteldelMonteAndria
CastelDragonaraCast
CastelloAngioinoMola
CastelloAragoneseoA
CastelloAragonesePa
CastelloAragoneseTa
Castellobaronalenorn
CastelloBrancaccioRu
CastelloCarloVLecce
CastelloCelenzaValfo
CastelloCenturioneog
CastelloDAquinoRocch

Property assertions: CasteldelMonteAndria

Object property assertions

locatedIn Andria
depiction foto2CasteldelMonteAndria
location coordinateCasteldelMonteAndria
depiction foto1CasteldelMonteAndria

Data property assertions

foaf:mbox "casteldelmonte@novaapulia.it"^^xsd:string
name "Castel del Monte Andria"^^xsd:string
phone "393883026000"^^xsd:string
description "Il grande imperatore Federico II definito S tupor Mund i è il «padre» di Castel del Monte dal 1996 iscritto nella lista dei monumenti patrimonio dell'Umanità Unesco. La straordinaria fortezza è oggi adagiata su una dolce collina verde a 20 chilometri da Andria incastonata nel Parco Nazionale dell'Alta Murgia . La misteriosa pianta ottagonale del maniero utilizzato forse come osservatorio astronomico o imponente maniero di caccia si differenzia in modo significativo nella struttura da tutti gli altri castelli della Puglia. Castel del Monte custodisce splendide creature mitologiche quali decorazioni nelle chiavi di volta delle sale interne e ricche cornici in porfido nelle porte. "^^xsd:string
sameAs "https://it.wikipedia.org/wiki/Castel_del_Monte"^^xsd:strin
postalCode "76123"^^xsd:string
address "Via Castel Del Monte"^^xsd:string

Fig. 2. The snapshot of the instance *Castel del Monte*.

After the design of the "Tourism in Apulia" ontology we obtain 30 first level classes and 10 second level classes, 5 object properties and 12 data properties. Classes and properties are taken from existing ontologies and vocabularies, only in one case we define a new class (*Cycling_Infrastructure*). Table II summarizes the number of classes and properties taken from each LOV. Finally, we supply the "Tourism in Apulia" ontology with instances derived from selected datasets by computational procedures.

TABLE II
NUMBER OF CLASSES AND PROPERTIES TAKEN FROM EACH LOV.

| LOV | Classes | Object Properties | Data Properties |
|---|---|---|---|
| DBpedia | 20 | 1 | 6 |
| Geonames | 2 | 2 | 1 |
| Geo | 1 | 1 | 0 |
| Shema.org | 9 | 0 | 2 |
| Km4city | 6 | 0 | 0 |
| Proton | 1 | 0 | 1 |
| FOAF | 0 | 1 | 2 |
| *NA* | 1 | 0 | 0 |

## REFERENCES

[1] C. Bizer, "The emerging web of linked data," *IEEE intelligent systems*, vol. 24, no. 5, 2009.

[2] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia," *Semantic Web Journal*, 2014.

[3] P. Bellini, I. Bruno, A. Cavaliere, D. Cenni, M. DiClaudio, G. Martelli, S. Menabeni, P. Nesi, G. Pantaleo, and N. Rauch, "Km4city: Smart city ontology and tools for city knowledge exploitation," in *European Data Forum*, 2015.

[5] http://schema.org/

[6] http://www.disit.org/km4city/schema

[7] http://www.ontotext.com/proton

[8] http://xmlns.com/foaf/spec/

[9] http://protege.stanford.edu/

[10] http://www.dataset.puglia.it/dataset/luoghi-di-interesse-turistico-culturale-naturalistico