Accommodations in Tuscany as Linked Data

Clara Bacciu, Angelica Lo Duca, Andrea Marchetti, Maurizio Tesconi

Institute of Informatics and Telematics,
National Research Council (CNR)
Via Moruzzi, 1 - Pisa, Italy
name.surname@iit.cnr.it

Abstract

The OpeNER Linked Dataset (OLD) contains 19.140 entries about accommodations in Tuscany (Italy). For each accommodation, it describes the type, e.g. hotel, bed and breakfast, hostel, camping etc., and other useful information, such as a short description, the Web address, its location and the features it provides. OLD is the linked data version of the open dataset provided by Fondazione Sistema Toscana, the representative system for tourism in Tuscany. In addition, to the original dataset, OLD provides also the link of each accommodation to the most common social media (Facebook, Foursquare, Google Places and Booking). OLD exploits three common ontologies of the accommodation domain: Acco, Hontology and GoodRelations. The idea is to provide a flexible dataset, which speaks more than one ontology. OLD is available as a SPARQL node and is released under the Creative Commons release. Finally, OLD is developed within the OpeNER European project, which aims at building a set of ready to use tools to recognize and disambiguate entity mentions and perform sentiment analysis and opinion detection on texts. Within the project, OLD provides a named entity repository for entity disambiguation.

Keywords: linked data, tourism, ontology

1. Introduction

In 2001 Tim Berners Lee introduced the concept of Semantic Web, whose main idea was to migrate from the Web of documents to the Web of data (Berners-Lee et al., 2001). In practice, instead of having many separated documents containing concepts and entities, the purpose of the Web of data is to connect concepts and contents to each other, going over documents. Thus the Web of data has led to the conversion of existing documents to linked data (Heath and Bizer, 2011), and to the creation of new datasets[1]. In addition, new data models have been implemented, to represent data in a common standard way (Allemang and Hendler, 2008). The most famous data models are the Resource Description Framework (RDF) (Klyne and Carroll, 2004) for the description of entities and the Web Ontology Model (OWL) (W3C, 2012) for the description of concepts.

The principles of Semantic Web have been applied to different fields of knowledge, spanning from cultural heritage[2] to health[3]. In this paper, we focus on the field of tourism, by building a linked dataset, which could be useful to tourists and travellers to get information about accommodation. To the best of our knowledge, only few linked datasets have been implemented in the field of tourism. They are described in Section 2..

We describe the OpeNER Linked Dataset (OLD), which has been implemented within the OpeNER project[4]. The main goal of the project is to provide a set of ready to use tools to perform some natural language processing tasks, such as named entity recognition and disambiguation, and sentiment analysis. In the context of OpeNER, OLD acts as a named entity repository for the tourism domain and it is exploited by the tool for named entity disambiguation.

OLD contains 19.140 accommodations in Tuscany (Italy). By accommodation we mean a place where people can sleep. It can be one of the following: hotel, bed and breakfast, apartment, camping, house, suite etc.

The original dataset was provided as an open dataset by Fondazione Sistema Toscana[5], the representative system of the Tuscany region. We enriched it with other useful information, such as the link of the accommodation to Facebook[6], Foursquare[7], Google Places[8] and Booking[9] whenever possible. In addition we converted the original dataset in linked data and exposed it as a SPARQL node, which is accessible at the following url: http://wafi.iit.cnr.it/opener_dataset/snorql/.

OLD exploits three common ontologies for the accommodation domain: Acco (Hepp, 2013), Hontology (Chaves et al., 2012) and GoodRelations (Hepp, 2011). In order to fulfill the principles of linked data (Bizer et al., 2009), it provides external links to DBpedia[10].

The OpeNER Linked Dataset could be useful for all people who would like to perform specific searches on accommodations, e.g. search only those which provide some specific features. It could also be used as a training set for a tool of named entity recognition in the tourism domain.

The remainder of the paper is organized as follows: in Section 2. we describe some other datasets relevant for tourism, while in Secion 3. we give an overview of the OpeNER project. In Section 4. we illustrate the data model, while in Sections 5. and 6. we focus on OLD and its enrichment with data of social media, respectively. Finally, in Section 7. we

[1] For a list of shared datasets, please look at: http://datahub.io/it/.

[2] http://www.europeana.eu

[3] http://www.lhdfound.org

[4] http://www.opener-project.org/

[5] http://www.fondazionesistematoscana.it

[6] http://www.facebook.com

[7] https://foursquare.com

[8] https://plus.google.com/u/0/local

[9] http://www.booking.com

[10] http://dbpedia.org

discuss it and we give our conclusions and future work.

2. Related Datasets

As far as we know, a comprehensive list of datasets is maintained by the CKAN repository[11]. Among all the available datasets, there are three containing accommodations: a) the Santillana Guide, b) the list of accommodations in Piedmont, Italy and c) the list of accommodations in Tuscany. Table 1 shows some information about the existing datasets.
The Santillana Guide dataset represents the content of the Santillana guide (owned by Prisa Digital) as Linked Data. The guide contains information about more than 1.500 Spanish restaurants and more than 1.500 Spanish hotels. The project exploits an ad-hoc ontology that has been developed for the tourism domain, and partially reuses the Infutur ontology[12]. The dataset constitutes a completion of El Viajero's tourism dataset[13]. It also integrates some restaurants from the Open Data Euskadi initiative[14].
The dataset of accommodations in Piedmont (Italy) uses GoodRelations and VCARD (Iannella and McKinney, 2013) ontologies. Furthermore, it includes addresses, contact information (where available) and geo-reference.
The dataset of accommodations in Tuscany (Italy) uses GoodRelations and VCARD and includes addresses, contact information (where available) and georeference.
Another useful service, which provides a browsable dataset of hotels is Hotelsbase[15]. However, it does not expose its content as Linked Data.

3. The OpeNER project

OpeNER (Open Polarity Enhanced Name Entity Recognition) is a project funded under the 7th Framework Program of the European Commission. Its main objective is to provide a set of ready-to-use modules for the processing of natural language.
In details, OpeNER focuses on building a linguistic pipeline in six European languages: English, Spanish, German, French, Italian and Dutch, in order to enable the identification and disambiguation of named entities and the analysis of sentiment in opinionated texts. As a result, the project produces a framework for the extraction of the attitude of customers regarding given topics (such as hotels and accommodations) in online reviews. This means both the analysis of large quantities of data and the aggregation of such data for the benefit of professionals within the tourism industry.
The named entities that are found in reviews can belong both to the general domain (e.g. people, locations, dates) and to the tourism domain (e.g. accommodations, attractions, point of interests). A generic entity can be recognized and disambiguated using generic linked data nodes such as DBpedia. A specific entity, instead, should be recognized through a specific linked data node. For this reason, we implemented the OpeNER Linked Dataset.

[11] http://datahub.io
[12] http://www.infutur.es/infutur/ns
[13] http://webenemasuno.linkeddata.es/index_en.html
[14] http://opendata.euskadi.net/
[15] http://hotelsbase.org

4. Data Model

All the datasets described in Section 2. exploit generic ontologies for common classes and properties, such as people and locations. However, since a standard ontology for the domain of accommodation does not exist yet, they often define ad-hoc ontologies for particular properties of such a domain. As a result, they are often mutually incompatible. This is true even for our dataset, but we tried to refer as much as possible to largely used and standardized vocabularies. This way, our dataset is more compatible with and understandable by a greater number of humans and machines.
An accommodation is characterized by some generic properties, such as its name and description, and some specific properties belonging to its domain, such as the number of rooms or the specific features it provides.
We propose to use different ontologies, depending on the properties we want to describe. Within the context of generic properties, we propose to use largely employed vocabularies such as DBpedia and VCARD. For the accommodation domain we propose to use the following ontologies: GoodRelations, Acco and Hontology. In the reminder of the section we separately describe the ontologies for accommodations.

4.1. GoodRelations

GoodRelations is a standardized vocabulary for e-commerce. It represents e-commerce scenarios through four entities: a) agent (person or organization), b) object (the product to sell or a service), c) offer (a promise made by the agent on the object) d) location (where the offer is made).
Within the context of accommodation, a hotel could be represented as a service (`gr:ProductOrService`), associated to a given location (`gr:Location`), and associated to a given agent (`gr:BusinessEntity`). GoodRelations provides also many datatype properties, which can be used for accommodation, such as `gr:description` or `gr:name` (a short textual description of the resource).

4.2. Acco

The Accommodation Ontology (Acco) is a Web vocabulary for hotels and other accommodation offers. It is designed to be used in combination with GoodRelations.
The Acco Ontology provides 20 classes, which can be classified into three categories: a) services, such as hotels, apartment, house etc., b) meals, such as breakfast, lunch, dinner etc. and c) features, such as meeting rooms, bed details, etc. The Acco Ontology represents a hotel through the class `acco:Hotel` with two properties: `acco:feature` and `acco:optionalFeature`, which specify the included and optional services, respectively, provided by the hotel. The Acco Ontology does not allow a user to specify in details which features are provided.

4.3. Hontology

Hontology is a vocabulary for the accommodation sector. The `Accommodation` class provides many subclasses, such as `Hotel`, `Apartment`, `Botel`,

Name	Source	Author	Records
Santillana Guide	http://webenemasuno.linkeddata.es	Vicomtech-IK4	64.748
Accommodations in Piedmont, Italy	http://www.linkedopendata.it/datasets/grrp	Linkedopendata.it	153.935
Accommodations in Tuscany, Italy	http://www.linkedopendata.it/datasets/grrt	Linkedopendata.it	434.714

Table 1: Comparison of the existing datasets for accommodations.

`GuestRoom` and `Hostel`. Hontology also defines the specific features provided by an accommodation through the class `Facility`, which is divided into the following subclasses: `InternalFacilities`, `ExternalFacilities` and `RoomFacilities`. Currently Hontology does not own any domain name and has no defined namespace so you have to download it locally and provide a namespace.

5. The OpeNER Linked Dataset

The OpeNER Linked Dataset contains information about accommodations in Tuscany. In particular, it contains 19.140 entries. Accommodations are classified according to the category they belong to.

Categories, which are extracted from the Acco (prefix `acco`) and Hontology (prefix `h`) ontologies are the following: a) apartment (which corresponds to `acco:Apartment` and `h:Apartment`), b) bed and breakfast (`h:BedAndBreakfast`), c) camping (`acco:CampingPitch`), d) hostel (`h:Hostel`), e) hotel (`acco:Hotel` and `h:Hotel`), f) house (`acco:House`), g) suite (`acco:Suite`). Note that not all the categories are defined in both the ontologies.

For each accommodation, information about its name and address is given. Furthermore, the features it provides are described, both through Hontology and Acco: for example, if the accommodation is equipped with a luggage room, it is provided as feature using both the ontologies. The following Turtle code is its representation through Acco:

```
acco:includedFeature [
a acco:AccommodationFeature ;
gr:name "Luggage room"@en ;
acco:value "yes"@en ].
```

while the following one is its representation through Hontology:

```
h:InternalFeature h:LuggageRoom .
```

The prefix we mean to use for our dataset is `opener:`. As for the other vocabularies, we used `h:` for Hontology, `acco:` for Acco, `gr:` for GoodRelations, `v:` for VCARD and, finally, `dbpedia-owl:` for DBpedia.

Figure 1 shows a small part of the RDF graph surrounding the resource `opener:acco-204042`, which corresponds to the Hotel Bologna in Pisa. We show only the feature regarding `Lift`, using both Acco and Hontology vocabularies. Predicates and classes are shown with their properties in italics. Circles are classes, while rectangles are literal values. The name of the accommodation is given by both `gr:name` and `v:fn`; the address is represented using both `dbpedia:owl-address` and the VCARD specific properties.

One of the best practices for the construction of a good linked dataset is to link to external, well known datasets (Heath and Bizer, 2011). This is achieved by establishing links between DBpedia and our accommodation datasets. In particular, the location of each accommodation is linked to its respective entry in DBpedia.

The OpeNER Linked Dataset is available at the following url: http://wafi.iit.cnr.it/opener_dataset while the SPARQL endpoint is available here: http://wafi.iit.cnr.it/opener/snorql/. It has been realized through a D2RQ server[16], which transforms a SQL database into a RDF one.

For example, the Hotel Bologna in Pisa can be accessed through the following direct link: http://wafi.iit.cnr.it/opener_dataset/page/acco-204042.

The OpeNER Linked Data is released under the Creative Commons license CC-BY-SA1.

6. Dataset enrichment

The original version of the dataset provided by the Fondazione Sistema Toscana contains only some specific information about accommodation, such as name, address and web site. However, nowadays many accommodations have also a page on a social media, where they advertise themselves and their offers. For this reason, we enriched the original dataset with the URLs to four of the most famous social media: Facebook, Foursquare, Google Places and Booking.

We implemented a crawler for each social media, which extracts the URL for each accommodation. In practice, the crawler searches of all the accommodations in Tuscany having a page on a social media and stores their URLs. Then, a merger algorithm compares every accommodation in the original dataset with those in the new dataset. If the matching is found, the accommodation is enriched with its URL on that social media.

The merger algorithm applies the similar text algorithm (Oliver, 1994) to the names of the accommodations to be compared. The matching is done by counting the number of matching characters in the two names. Two names are considered equal if both have the same length, which also corresponds to the number of matching characters.

7. Discussion and Conclusion

The OpeNER Linked Dataset is a data source for tourism in Tuscany. Its main purpose is its use within the OpeNER project. In particular, it is exploited by the named entity recognition process to recognize entities in the accommodation domain.

[16] http://d2rq.org

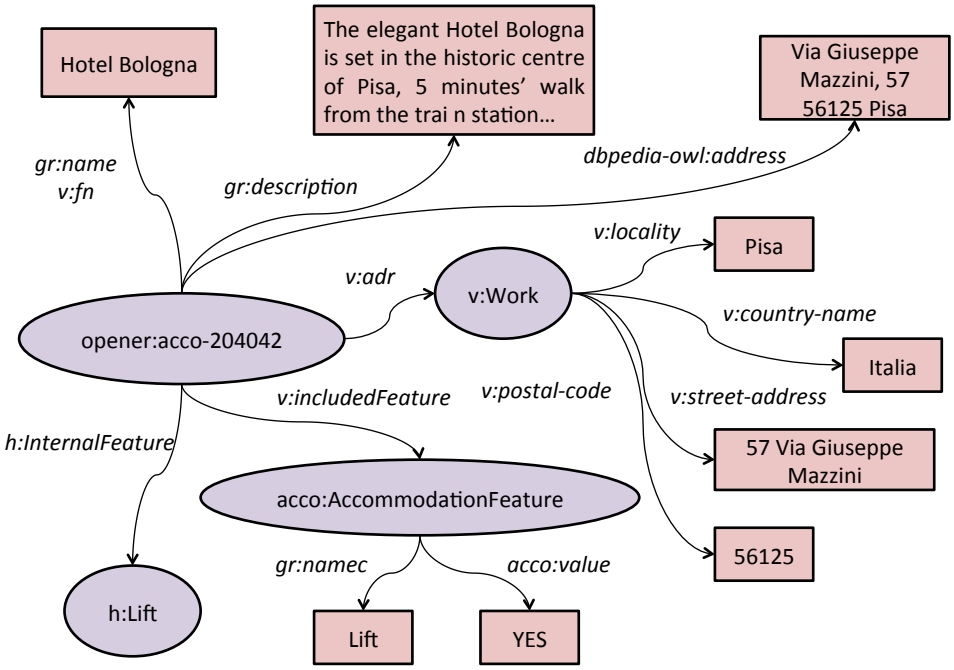

Figure 1: A small part of the RDF graph referred to resource opener:acco-204042

However the OpeNER Linked Dataset could be useful also outside this project. With respect to other Web services, which also provide a list of accommodations such as Hotelsbase, the OpeNER Linked Dataset specifies also if some given feature of an accommodation is present, such as the luggage room or the snack bar.

The presence of a great number of details could be used by people and search engines to search for an accommodation having particular features, such as the proximity to a point of interest or the fact it provides a certain service.

Future work on the dataset includes efforts to make it multilingual, i.e. provide it in the six languages of the OpeNER project: Italian (which is already available), French, German, Dutch, Spanish and English.

The purpose of OLD would be to become the Wikipedia of tourism in time, although at the moment it contains only accommodations in Tuscany (Italy).

Acknowledgements

This work has been carried out within OpeNER project, cofunded by the European Commission under the FP7 (7th Framework Programs Grant Agreement n. 296451).

8. References

Dean Allemang and Jim Hendler. 2008. *Semantic web for the working ontologist : effective modeling in RDF, RDFS and OWL.* Morgan Kaufmann Publishers/Elsevier, Amsterdam ; Boston.

Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*, 284(5):34–43, May.

C. Bizer, T. Heath, and T. Berners-Lee. 2009. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):122.

Marcirio Silveira Chaves, Larissa A. de Freitas, and Renata Vieira. 2012. Hontology: A multilingual ontology for the accommodation sector in the tourism industry. In Joaquim Filipe and Jan L. G. Dietz, editors, *KEOD*, pages 149–154. SciTePress.

Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition.

Martin Hepp. 2011. Goodrelations language reference. Technical report, Hepp Research GmbH, Innsbruck.

Martin Hepp. 2013. Accommodation ontology language reference. Technical report, Hepp Research GmbH, Innsbruck.

R. Iannella and J. McKinney. 2013. VCARD ontology. Available at: http://www.w3.org/TR/vcard-rdf/. Technical report.

Graham Klyne and Jeremy J. Carroll. 2004. Resource description framework (RDF): Concepts and abstract syntax. World Wide Web Consortium, Recommendation REC-rdf-concepts-20040210, February.

Ian Oliver. 1994. *Programming classics - implementing the world's best algorithms*. Prentice Hall.

OWL Working Group W3C. 2012. OOWL 2 Web Ontology Language Document Overview (Second Edition). World Wide Web Consortium, Recommendation, December.