OpeNER Accommodations as Linked Data

Clara Bacciu, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi
Institute of Informatics and Telematics National Research Council, via Moruzzi 1, 56124 Pisa, Italy
E-mail: {name.surname}@iit.cnr.it

Abstract. The OpeNER Linked Dataset contains information about accommodation for three locations: Amsterdam, Tuscany (Italy) and Spain. For each accommodation, it provides the type (e.g. hotel, bed and breakfast, hostel etc.), and other useful information, such as a short description, the location, the number of rooms and the features it provides. The dataset has been built starting from two Web sites, which give information about accommodation: Booking.com and Google+ local. Furthermore, it exploits three common ontologies for the accommodation domain: Acco, Hontology and GoodRelations. Finally, the dataset contains 19.973 entries: 1.043 entries for Amsterdam, 15.371 for Tuscany and 3.559 for some localities of Spain.

Keywords: Tourism, Accommodation, Linked Data

1. Introduction

In this paper, we describe the Accommodation Linked Dataset implemented within the OpeNER project [8]. This dataset contains all the accommodations of Amsterdam, Tuscany (Italy) and some cities of Spain. It has been deployed for the needs of the OpeNER project, which focuses on such locations. By accommodation we mean one of the following: hotels, bed and breakfasts, apartments, campings, hostels, houses and suites.

Data have been retrieved from two sources: Booking.com [1] and Google+ Local[6]. Booking.com is online booking facility, which provides information about accommodations through a query-based system. We have retrieved interesting data through ad-hoc scrapers, which build automatic queries and examine the HTML pages of results. Google+ Local, instead, is a part of the Google+ social network that deals with places and business activities. It provides some APIs, which allow a developer to retrieve information about such activities. However, their free APIs present some restrictions, i.e. they allow a single user to perform maximum 100 text searches per day. For this reason, the retrieval of information is very slow and time-consuming.

We have integrated data coming from Booking.com with those retrieved from Google+ Local. The match between the entries of Booking.com with those of Google+ Local is done on the Google ID, which is retrieved through the Google APIs. Google+ Local receives as input the name and the address of an accommodation and returns its Google ID as a result.

Our dataset, named OpeNER Linked Dataset, is exposed as a SPARQL node, which is accessible at the following url: `http://wafi.iit.cnr.it/opener-acco`. It exploits three common ontologies for the accommodation domain: Acco [2], Hontology [7] and GoodRelations [5]. Furthermore, it is linked to the DBpedia Linked Dataset. Maintainance is done through ad-hoc scrapers, which periodically check whether there are new entries on Booking.com and Google+ Local, and update the dataset.

The OpeNER Linked Dataset can be exploited by all those people who want to perform specific searches on accommodations, e.g. search only those which provide some specific features.

The remainder of the paper is organized as follows: in Section 2 we discuss some related datasets, while in Section 3 we illustrate our data model. In Section 4 we define the OpeNER Linked Dataset and, in Section 5, finally, we discuss about it.

2. Related Datasets

A comprehensive list of datasets is maintained by the CKAN repository[1], which contains the following projects providing accommodations datasets: a) the *Santillana Guide dataset*, 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 1500 Spanish restaurants and more than 1500 Spanish hotels. The project exploits an ad-hoc ontology thet has been developed for the tourism domain, and partially reuses the Infutur ontology[2]. The dataset constitutes a completion of El Viajero's tourism dataset[3]. It also integrates some restaurants from the Open Data Euskadi initiative[4].

The dataset of accommodations in Piedmont (Italy) uses GoodRelations [5] and VCARD [9] 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 geo-reference.

All the described datasets exploit generic ontologies for common classes and properties, such as persons and locations. However, since a standard ontology for the domain of accommodation does not exist, they often define ad-hoc ontologies for such a domain. As a result, they are often 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.

Another useful service, which provides a browsable dataset of hotels is Hotelsbase[5]. However, it does not expose its content as Linked Data.

3. Data Model

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. For this reason, 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 used vocabularies such as DBpedia [4] and VCARD [9].

For the accommodation domain we propose to use the following ontologies: the GoodRelations [5], Acco [2] and Hontology [7] [10] ontologies.

In the reminder of the section we separately describe the ontologies for accommodations.

3.1. GoodRelations

GoodRelations [5] 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).

3.2. Acco

The Accommodation Ontology (Acco) [2] 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 the features in details.

[1] The Data Hub is a community-run catalogue of useful sets of data on the Internet: http://datahub.io/

[2] http://www.infutur.es/infutur/ns

[3] http://webenemasuno.linkeddata.es/

[4] http://opendata.euskadi.net/w79-home/eu/

[5] http://www.hotelsbase.org

Name	Source	Author	Triples
Santillana Guide dataset	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.

3.3. Hontology

Hontology [7,10] is a vocabulary for the accommodation sector. The `Accommodation` class provides many subclasses, such as `Hotel`, `Apartment`, `Botel`, `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.

4. OpeNER Linked Dataset

4.1. Dataset Extraction

In this section we describe the sources of data and which information we have extracted from them.

4.1.1. Google+ local

Google+ local [6] is a service provided by Google, allowing users to create a profile of their commercial activities. It is the new version of Google Places. It contains more than 50 million dynamically generated places, of which 8 million have been claimed by the businesses themselves. With respect to the locations referred by the OpeNER project, Google+ local contains about 50.000 accommodations in Spain, 15.371 accommodations in Tuscany (Italy) and 1.043 accommodations in Amsterdam. It associates each location to a static HTML page. In order to retrieve information about accommodations, we have implemented some ad-hoc scripts, which exploit the API provided by the Web site. Table 2 shows the extracted information for each accommodation.

Name	Meaning
name	the name of the accommodation
address	the place where the accommodation is located
latitude	the latitude where the accommodation is located
longitude	the longitude where the accommodation is located
type	the type of accommodation
google id	the identifier of the accommodation on Google+ local
URL	the URL of the accommodation on Google+ local
logo	the URL of the logo of the accommodation

Table 2

Extracted information from Google+ local for each accommodation.

4.1.2. Booking

Booking.com [1] is one of the leading online accommodation reservation agencies. It contains more than 41,000 destinations, organized per location. With respect to the locations referred by the OpeNER project, Booking.com contains 17,244 accommodations in Spain, 3,860 accommodations in Tuscany (Italy) and 669 accommodations in Amsterdam. Booking.com associates each accommodation to a static HTML page, accessible through a search engine, which is browsable through HTTP requests, each performed through the GET method.

We have developed some ad-hoc scrapers for the extraction of information from Booking.com. Table 3 shows the extracted information for each accommodation.

Name	Meaning
hotel name	the name of the accommodation
address	the place where the accommodation is located
summary	a short description of the accommodation
URL	the URL of the accommodation on Booking.com
language	the language of the page on Booking.com
services	the features provided by the accommodation
number of rooms	the number of rooms of the accommodation

Table 3

Extracted information from Booking for each accommodation.

4.2. Dataset Integration

In order to build a more complete resource, we have integrated data coming from Booking with those coming from Google+ local. The integration has been done through the match on the Google+ local ID. In particular, for all the Booking.com accommodations, we have retrieved their Google+ local ID, by exploiting a

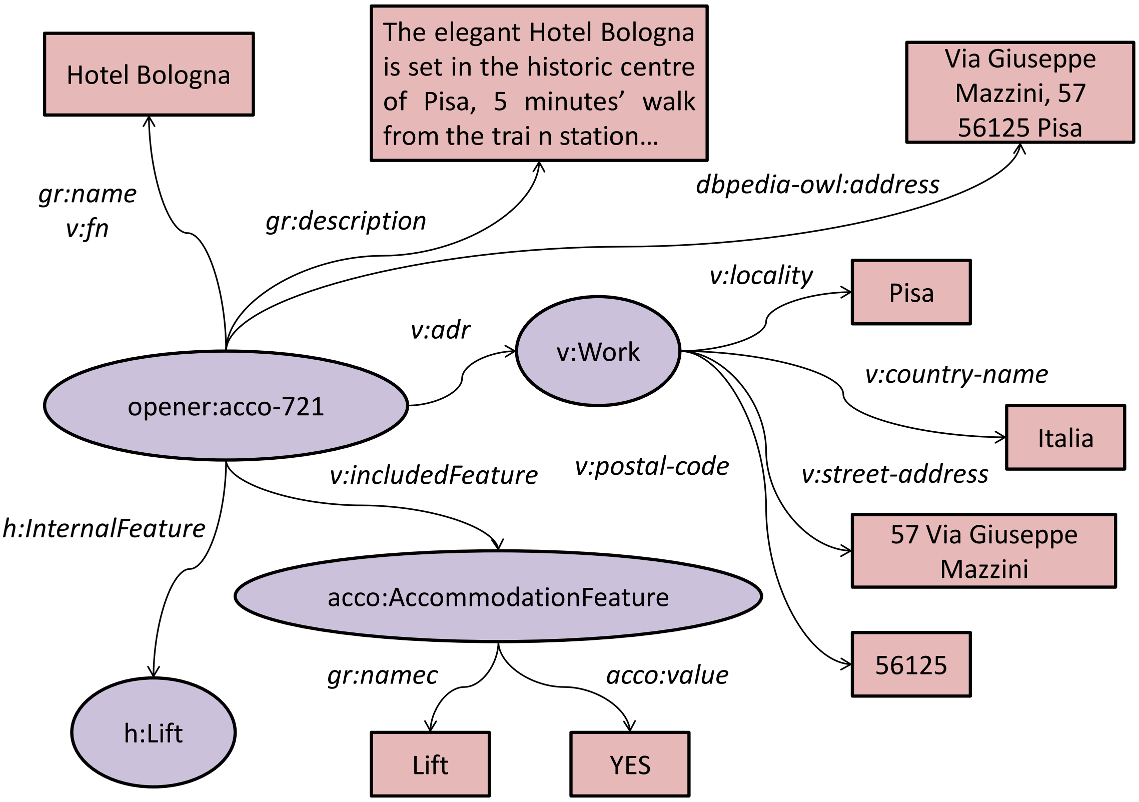

Fig. 1. A small part of the RDF graph referred to the resource opener:acco-721.

service provided by the Google Places APIs[6]. Such a service requires a string as parameter, which contains the name and the address of the accommodation, both provided by Booking. In practice we have exploited Google for the reconciliation of data. However, this service presents some limitations: it allows the same user to perform only 100 textsearch per day. For this reason, at the moment, the OpeNER Linked Dataset provides only the integration of Booking.com and Google+ local for Amsterdam. Currently we are integrating also Tuscany, but the integration is not complete yet.

4.3. Linked Data set Structure

The OpeNER Linked Dataset contains information about accommodation in Amsterdam, Tuscany and some locations in Spain. In particular, it contains 19.973 entries: 1.043 entries for Amsterdam, 15.371 for Tuscany and 3.559 for some localities of Spain. Accommodations are classified according to the category they belong to. Categories, which are extracted from the Acco (acco) and Hontology (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 provides a *luggage room* as a feature, it is provided using both the ontologies. The following Turtle codeis its representation through Acco:

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

[6] https://maps.googleapis.com/

Locality	Booking	Google+ local	Resolved Google IDs	integrated items	not integrated items
Amsterdam	669	1.043	568	474/568	195/669
Tuscany	3.860	15.371	0	0	0
Spain	3.559	about 20.000	0	0	0

Table 4

Statistics about the OpeNER Linked Dataset.

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, `v:` for VCARD, `gr:` for GoodRelations and finally `dbpedia-owl:` for DBpedia.

Figure 1 shows a small part of the RDF graph surrounding the resource `opener:acco-721`, 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 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.

At the moment the OpeNER Linked Dataset is provided only in English. Table 4 shows some statistics about the OpeNER Linked Dataset. The table shows for each locality, how many entries we have retrieved from Booking.com and Google+ local. Furthermore, it describes how many Google+ IDs we have resolved starting from Booking.com, and how many Booking.com entries we integrated with Google+ local (integrated items) and the remaing not integrated entries (not integrated items). Currently we have integrated only the data adout Amsterdam, due to the limitations given by the free Google APIs service.

One of the best practices for the construction of a good Linked Dataset is to link to external, well known datasets [11]. 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. We have obtained 2 links to DBpedia for Amsterdam, 521 links for Tuscany and 20 for Spain.

4.4. Availability

The OpeNER Linked Dataset is available at the following url: `http://wafi.iit.cnr.it/opener-acco`, while the SPARQL endpoint is available at the following url: `http://wafi.iit.cnr.it/opener/snorql/`. It has been realized through a D2RQ server [3], which transforms a SQL database into a RDF one. For example, the *Inntel Hotels* in Amsterdam can be accessed through the following direct link: `http://wafi.iit.cnr.it/opener/acco-1`.

4.5. Licensing and Maintainance

Licensing is a big issue of our dataset. Although it has been developed within the OpeNER project, which is a European Research project, we have taken data from two different contributions (Section 4.1), which provide free access to their databases, through a form-based access or an API. We are currently investigating if they allow access to their resources through mechanisms which are not web-form based or API-based. For this reason, at the moment the exposed linked dataset contains only the information which does not depend on the specific source, such as the name and the address of the accommodations. However, we are going to get a specific licence in order to expose all the information, including accommodations description and other sensitive data.

In order to maintain the OpeNER Linked Dataset updated, some ad-hoc scrapers periodically check whether new accommodations have been added to Booking.com and Google+ Local and add them to the dataset.

5. Discussion

The OpeNER Linked Dataset is an important data source for tourism in Amsterdam, Tuscany and some localities in Spain. 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. However the OpeNER Linked Dataset can be used also for other purposes. With respect to other Web services, which also provide a list of accommodation (e.g. http://www.hotelsbase.org), the OpeNER Linked Dataset specifies also if some specific feature of an accommodation is present, such as the *luggage room* or the *snack bar*. The presence of a great number of details can be used by people to search for a accommodation providing a particular service, such as the proximity to a point of interest or its features.

Future work on the dataset include efforts to make it multilingual, i.e. provide it in six languages: English (which is already available), French, German, Dutch, Spanish and Italian. Furthermore, we are going to extend the dataset also to other localities, such as the whole Italy and the whole Holland and Spain.

Acknowledgements

This work has been partially supported by the European Project OpeNER (FP7/2007-2013).

References

[1] Booking Web Site: http://www.booking.com.

[2] The Accommodation Ontology Language Reference: http://ontologies.sti-innsbruck.at/Acco/ns.html.

[3] The D2RQ project: http://d2rq.org.

[4] The DBpedia Ontology: http://wiki.dbpedia.org/Ontology.

[5] The GoodRelations Ontology: http://www.heppnetz.de/ontologies/goodrelations/v1.

[6] The Google+ Local Project: https://developers.google.com/+/api/.

[7] The Hontology Ontology: http:// metashare.tilde.com/repository/browse/hontology/ a83c9d04cb7a11e1a404080027e73ea2359e10ea62b940109aabe03684aa5ea4.

[8] The OpeNER project: http://www.opener-project.org.

[9] The VCARD Ontology: http://www.w3.org/Submission/vcard-rdf/.

[10] Marcirio Silveira Chaves, Larissa A. Freitas, and Renata Vieira. Hontology: a multilingual ontology for the accommodation sector in the tourism industry. *Proceedings of the 4th International Conference on Knowledge Engineering and Ontology Development, Barcelona, Spain*, pages 149–154, October 2012.

[11] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 1(1):1–136, 2011.