An Open Source System Architecture for Digital Geolinguistic Linked Open Data

Emanuele Di Buccio, Giorgio Maria Di Nunzio, and Gianmaria Silvello

Dept. of Information Engineering, University of Padua
{dibuccio,dinunzio,silvello}@dei.unipd.it

Abstract. Digital Geolinguistic systems encourages collaboration between linguists, historians, archaeologists, ethnographers, as they explore the relationship between language and cultural adaptation and change. These systems can be used as instructional tools, presenting complex data and relationships in a way accessible to all educational levels. In this poster, we present a system architecture based on a Linked Open Data (LOD) approach the aim of which is to increase the level of interoperability of geolinguistic applications and the reuse of the data.

1 Introduction

In the last two decades, several large-scale databases of linguistic material of various types have been developed worldwide. The World Atlas of Languages Structures (WALS) [1] is one of the largest projects, with 160 maps showing the geographical distribution of structural linguistic features,[1] and it is the first linguistic feature atlas on a world-wide scale. The Common Language Resources and Technology Infrastructure project (CLARIN, [2]) is a European initiative the aim of which is to create an infrastructure which makes language resources (annotated recordings, texts, lexica, ontologies) and technology (speech recognisers, lemmatisers, parsers, summarisers, information extractors) available and readily usable to scholars of all disciplines. Language resources that have been made publicly available can vary in the richness of the information they contain. Bird et al. [3] discuss three important points about the design and distribution of language resources: (i) How do we design a new language resource and ensure that its coverage, balance and documentation support a wide range of uses? (ii) When existing data is in the wrong format for some analysis tool, how can we convert it into a suitable format? (iii) What is a good way to document the existence of a resource we have created so that others can easily find it?

In this poster, we present a system architecture based on a current project named Atlante Sintattico d'Italia, Syntactic Atlas of Italy (ASIt) based on the LOD paradigm with the aim of enabling interoperability at a data-level. The LOD paradigm refers to a set of best practices for publishing data on the Web[2]

[1] http://www.wals.info/
[2] http://www.w3.org/DesignIssues/LinkedData.html

T. Aalberg et al. (Eds.): TPDL 2013, LNCS 8092, pp. 438–441, 2013.

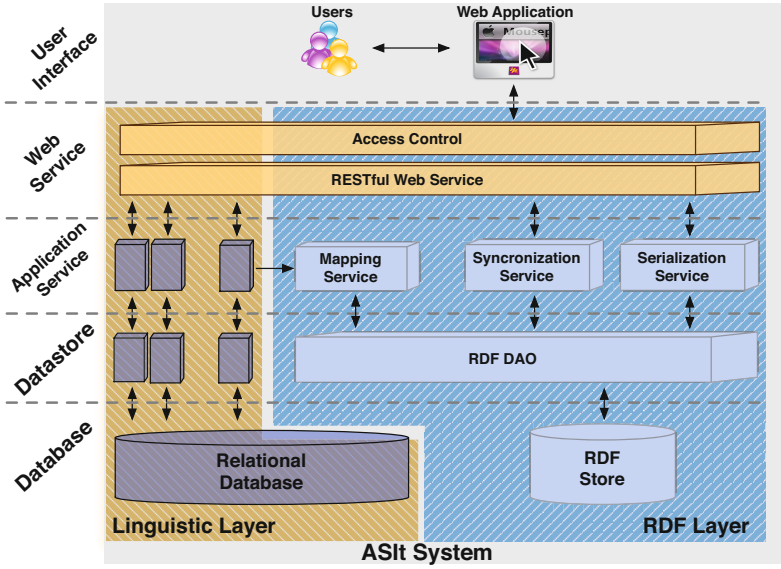

Fig. 1. The architecture of the ASIt System in which we highlight the diverse constituting levels and the RDF Layer

and it is based on a standardized data model, the Resource Description Framework (RDF). RDF is designed to represent information in a minimally constraining way and it is based on the following building blocks: graph data model, URI-based vocabulary, data types, literals, and several serialization syntaxes.

2 System Architecture

We present the architecture of the ASIt system composed by the *linguistic layer* and the *RDF layer*, as shown in Figure 1.

2.1 Linguistic Layer

The *linguistic layer* [4] has been designed to be modular, thus reducing the dependency on a particular implementation of its constituting modules. It can be framed in four different levels: *database*, *datastore*, *application service*, and *Web service*. The *database* level is constituted by a relational database, the schema of which is based on the ASIt conceptual model; the currently adopted DataBase Management System (DBMS) is PostgreSQL.[3] The *datastore* is responsible for the persistence of the linguistic resources and provides an interface to store and access linguistic data. The *application service* is responsible for the interaction with the linguistic resources; it provides an Application Program Interface (API)

[3] `http://www.postgresql.org/`

to perform operations on the resources – e.g. list sentences in a document, or list words in a sentence, and add tags to sentences and words. When linguistic resources are created or modified, the application service exploits the datastore API for the persistence of data. The *Web service* provides functionalities to create, modify and delete resources, and gather their descriptions through appropriate HTTP requests based on a RESTful Web service [5]. This level is also responsible for access control which is necessary to preserve the quality of the data maintained in the ASIt database. Indeed, only allowed users can create or modify resources, whereas there is no restriction to access resource descriptions.

2.2 RDF Layer

The *RDF layer* is responsible for persistence and access to RDF triples of linguistic data instantiated on the basis of the ontology. The RDF layer has been developed by exploiting the functionalities of the open source library Apache Jena.[4] Jena was adopted because of the variety of solutions for persistence of the RDF/S instantiation, the support of a number of RDF output formats, and the functionalities for reasoning with RDF and Ontology Web Language (OWL) data sources. A *mapping service* has been developed to instantiate the ontology starting from the data stored in the ASIt relational database. A request for creation, deletion or modification of a resource is processed by the linguistic layer that, through the proper module of the *application service*, allows the interaction with the resource and stores its new state. In parallel, by means of the *syncronization service*, the RDF layer processes the request and updates the RDF triples instantiating the ASIt ontology. This service allows for the interaction with the *RDF datastore* which is responsible for the persistence of the RDF triples in the RDF store. Therefore, the operations required by resource creation, deletion or modification are performed in parallel for each request to guarantee the synchronzation between the relational database and the RDF store. When a request for accessing a resource is submitted to the system, the RDF *serialization service* retrieves information on the requested resource from the RDF store and it returns the result in the requested output format.

3 Web Application and Final Remarks

The architecture presented in this poster allows us to expose the linguistic resources as a Linked Open Dataset. By exploiting the synchronisation services, the ASIt Geolinguistic Linked Open Dataset size grows proportionally to the size of the database.[5] Table 1 reports the statistics about the evolution of the data in ASIt in the last two and a half years. This dataset has been exposed following the guidelines in [6]. The linguistic dataset created with this architecture can be easily linked to other open datasets. As an example, the ASIt dataset is linked to DBpedia; indeed, the instances of the geographical classes `Region`,

[4] `http://incubator.apache.org/jena`

[5] The details of this dataset can be found here: `http://purl.org/asit/alld`

Table 1. Statistics about the growth of main entities/relationships of the ASIt curated database

	Jan '11	Jul '11	Jan '12	Jul '12	Jan '13
tags	524	530	532	532	532
documents	462	468	512	540	510
sentences	47,973	48,575	51,256	54,091	54,195
tags/sentences	10,364	16,731	18,080	18,369	18,371
tags/words	0	5,411	18,509	27,046	27,271

`Province`, and `Town` are linked to the corresponding instances of the dbpedia.org class `Place` through the property `owl:sameAs`.

A GUI to submit queries to the ASIt Linguistic Linked Open Dataset[6] and to dynamically produce maps on the basis of the user requests[7] are already available and publicly accessible.

One of the key points of this approach is the decoupling between the system which manages the data and the one which provides services over those data. In fact, this system has been adopted to develop a geolinguistic application build upon the presented dataset providing linguists with a tool for investigating variations among closely related languages. We also developed a graphical user interface on top of this application that dynamically produces maps on the basis of the user requests. Finally, we imagine the use of the Geolinguistic Linked Open Dataset by third-party linguistic projects in order to enrich the data and build-up new services over them.

References

1. Haspelmath, M., Dryer, M.S., Gil, D., Comrie, B.: The World Atlas of Language Structures. Oxford University Press, United Kingdom (2005)
2. Odijk, J.: The CLARIN-NL Project. In: LREC, European Language Resources Association (2010)
3. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
4. Di Buccio, E., Di Nunzio, G.M., Silvello, G.: A system for exposing linguistic linked open data. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 173–178. Springer, Heidelberg (2012)
5. Fielding, R.T., Taylor, R.N.: Principled design of the modern web architecture. ACM TOIT 2, 115–150 (2002)
6. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. In: Synthesis Lectures on the Semantic Web. Morgan & Claypool Publishers (2011)

[6] `http://purl.org/asit/rdf/sparqlGui`
[7] `http://purl.org/asit/rdf/search`