# A Geolinguistic Web Application Based on Linked Open Data

Emanuele Di Buccio
Department of Information Engineering
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
dibuccio@dei.unipd.it

Giorgio Maria Di Nunzio
Department of Information Engineering
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
dinunzio@dei.unipd.it

Gianmaria Silvello
Department of Information Engineering
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
silvello@dei.unipd.it

## ABSTRACT

Digital Geolinguistic systems encourage collaboration between linguists, historians, archaeologists, ethnographers, as they explore the relationship between language and cultural adaptation and change. In this demo, we propose a Linked Open Data approach for increasing the level of interoperability of geolinguistic applications and the reuse of the data. We present a case study of a geolinguistic project named Atlante Sintattico d'Italia, Syntactic Atlas of Italy (ASIt).

## Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Database Management—*Spatial databases and GIS*; J.5 [**Computer Applications**]: Linguistics

## General Terms

Design, Management, Standardization

## Keywords

Linked Open Data, Digital Geolinguistics, RDF, Ontology

## 1. INTRODUCTION

Geolinguistics is an interdisciplinary field that incorporates language maps depicting spatial patterns of language location or the results of processes that lead to language change [3]. In this context, the linguistic atlas has proved to be a vital tool and product of geolinguistics since the earliest stages of the field, and it has provided a stage for the incorporation of modern Geographic Information System (GIS).

Several large-scale databases of linguistic material of various types have been developed worldwide. The World Atlas of Languages Structures (WALS) 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 study of dialectal heritage is the goal of many research groups in Europe as well. The aim of the EU-sponsored Common Language Resources and Technology Infrastructure project (CLARIN)[2] 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. The heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections [1]. For example, the Edisyn search engine – the aim of which was to make different dialectal databases comparable – "in practice has proven to be unfeaseable".[3]

The research direction we pursue in our work is to move the focus from the systems handling the linguistic data to the data themselves. For this purpose the Linked Open Data (LOD) paradigm is very promising, because it eases interoperability between different systems by allowing the definition of data-driven models and applications.

We present a case study of a linguistic project named ASIt the objective of which is to investigate variations among closely related languages. The project aims to implement a system that enables the management of a resource of curated dialect data and provides access to grammatical data, also through an advanced user interface specifically designed to update and annotate the primary data. The ASIt linguistic corpus is constituted of eight different questionnaires written in Italian and almost 500 translations from the original Italian questionnaires to one Italian dialect. A questionnaire is a set of sentences built to study specific linguistic phenomena. Currently there are more than 240 different dialects, for a total of more than 50,000 sentences.

## 2. A GEOLINGUISTIC WEB APPLICATION

One of the requirements of the ASIt system is to search and browse specific grammatical structures rather than searching for a specific word or combination of words; moreover, result presentation and browsing should explicitly consider the linguistic research task. We developed a graphical user interface on top of the ASIt System [2] that dynamically

[1] http://www.wals.info/
[2] http://www.clarin.eu/
[3] http://www.dialectsyntax.org/


*SIGIR'13,* July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.

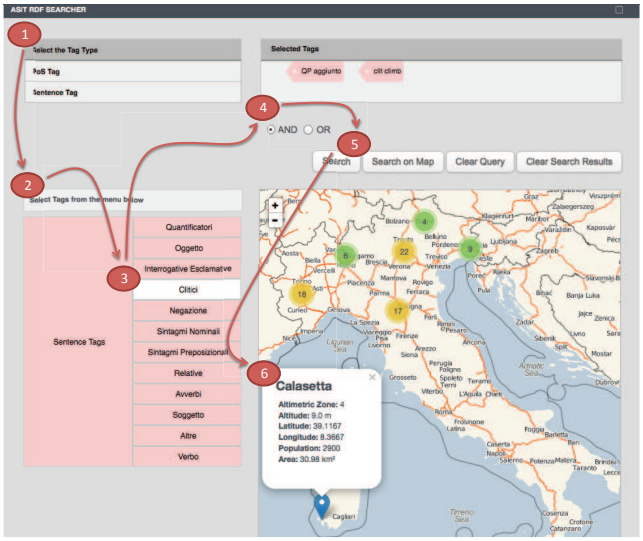


**Figure 1: ASIt RDF GeoSearch GUI.**

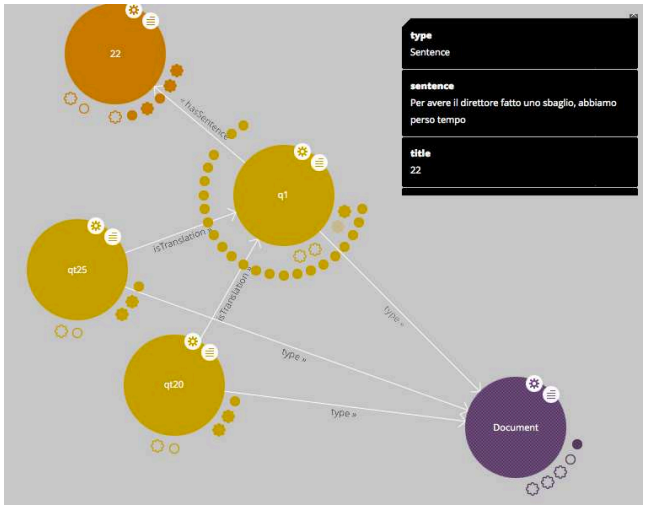


**Figure 2: ASIt RDF Browsing GUI.**

produces maps on the basis of the user query. The interface is available at the URL:[4]

`http://purl.org/asit/rdf/search`

A screenshot is reported in Figure 1; the circles highlight the steps performed by a user in a tag-based search. When the user accesses the search page, a table with the list of supported tag types is presented (step 1). Tags of a given type are hierarchically organised in a tag tree. Once the user clicks on a tag, the tag is added to the selected tag list (step 4); the user can remove a tag from the list by clicking on it. Moreover, the user can select the Boolean constraint for the tag-based search.

Two different types of result presentation are currently supported. The first one consists of the list of sentences that satisfy the query. For each of the sentences in the result set, information on the document that contains that sentence, the sentence identifier and fulltext is reported. The user can browse the RDF Graph associated to these resources as shown in Figure 2.

The second result presentation supports the users in the investigation of the geographical distribution of linguistic phenomena. When the user clicks on the "Search on Map" button, the results are shown on a dynamically generated map. When the user clicks on the marker corresponding to a specific location, information on the location is displayed. For instance, in step 6 of Figure 1 the user clicked on the marker corresponding to the location "Calasetta".

### *Technical Details*

The ASIt dataset is represented by RDF triples (subject, predicate, object) and maintained in a triple store. The triples can be accessed through HTTP requests to a SPARQL Endpoint.[5]

All the components of the interface, e.g. the tag tree and the result list, show the data generated by SPARQL queries.

The RDF Browsing GUI shown in Figure 2 is based on the open source LodLive Library.[6] The Leaflet javascript[7] library is used to visualise the results on a map; we adopted this library in order to achieve a complete open data approach since it relies on *OpenStreetMap* data.[8] In order to display the locations in the result set, we used the Leaflet marker cluster plugin. Each cluster is depicted as a circle; the number in the centre of the circle refers to the number of locations that belongs to the cluster. When the user clicks on a cluster, a zoom is performed in order to show all the locations within the cluster. Currently, location clustering is based on the default criterion implemented by the plugin; future extension of our interface could support diverse clustering strategies, e.g. based on the features of the dialects spoken in the considered locations.

## Acknowledgment

This work has been supported by the project "Un'inchiesta grammaticale sui dialetti italiani: ricerca sul campo, gestione dei dati, analisi linguistica" (Bando FIRB – Futuro in ricerca 2008) and the PROMISE network of excellence (contract n. 258191) project, as part of the 7th Framework Program of the European Commission.

## 3. REFERENCES

[1] C. Chiarcos. Interoperability of corpora and annotations. In *Linked Data in Linguistics*, pages 161–179. Springer Berlin Heidelberg, 2012.

[2] E. Di Buccio, G. M. Di Nunzio, and G. Silvello. A system for exposing linguistic linked open data. In *TPDL*, volume 7489 of *Lecture Notes in Computer Science*, pages 173–178. Springer, 2012.

[3] S. Hoch and J. J. Hayes. Geolinguistics: The Incorporation of Geographic Information Systems and Science. *The Geographical Bulletin*, 51(1):23–36, 2010.

[4]The Web application is optimised for Firefox browsers.

[5]SPARQL is a recursive acronym for *SPARQL Protocol and RDF Query Language*; a SPARQL endpoint is the URI at which an HTTP server services HTTP requests and sends back HTTP responses for SPARQL Protocol operations — see `http://www.w3.org/TR/sparql11-protocol/`

[6]`http://lodlive.it/`

[7]`http://leafletjs.com/`

[8]`http://www.openstreetdata.org/`