digital humanities
Lausanne – Switzerland '14
UNIL | Université de Lausanne
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ANNOTATING TEXTS WITH ONTOLOGIES, FROM GEOGRAPHY TO PERSONS AND EVENTS

Lana, Maurizio
Università del Piemonte Orientale
Ciotti, Fabio
Università di Roma Tor Vergata
Magro, Diego
Università di Torino
Peroni, Silvio
Università di Bologna
Tomasi, Francesca
Università di Bologna
Vitali, Fabio
Università di Bologna

Category: Poster
Session: 1
Date: 2014-07-10
Time: 14:00:00
Room: Amphipôle Common Area

1. Introduction

1.1. Overview

Geolat - geography for latin literature aim is to annotate every placename in the latin literature using a geographical ontology for the ancient world which must be built from scratch. This will allow scholars but also students and citizens to start reading texts choosing a specific area or place they are interested to. The starting point is the Latin literature because of its founding meaning for the European culture [1].

But this same model can be applied to every other literature because nothing in it is language-dependent. Should a scholar be able to browse a map of Europe (or of the world) to choose a specific place and start examining which authors in which works wrote about it is a completely new way or conceiving the study of literature.

But this model can be expanded from place names to persons and events, allowing to browse texts by the names of people they contain or by types and subtypes of events. That is texts are meant as collections of information. Not always and not necessarily factual information because literary texts can speak of mental, fictional representation of places or people. Nevertheless what texts say often constitutes the canvas of what real places and people and events actually are or were.

Concurrent 'sayings' about the same place, person, or event, can be managed strictly connecting the annotation to the textual object containing the statement(s). So an ontology of textual objects is needed.

1.2. Methodology

The main methodology is that of starting from a prototype where a geographical ontology is used to annotate place names in latin texts. At this level a startup funding is sufficient to build a prototype showing and example of what could be obtained.

Then the prototype can be expanded:

adding more literatures;
and/or adding more ontologies.
In both cases a great effort is needed to accomplish the task and adequate funding is indispensable.

As the ontology is the core of the project something must be said about it. Conceptually, the annotation mechanism is based on EARMARK [2]. EARMARK is an OWL 2 DL ontology that defines document meta-markup (elements, attributes, comments and text nodes), and it can be used to express facts about the inherent semantics of the markup elements and of the content of a text [3] according to a

precise semiotic model [4]. Thus, using EARMARK it is possible to create annotations on (plain or marked up) documents and document fragments according to a multilayer architecture (that can be aligned, in principle, with the Open Annotation Data Model [5]) and allows one not only to specify by an annotation the referent of a geographic place name, but also to provide a set of data describing such referent. All of the entities involved in an annotation are identified by an IRI and both the annotation, its metadata and its content are specified according to the RDF data model. In this way, all these pieces of information can be published on the Web of Data following the Linked Data principles [6]. The semantics of these data are explicitly expressed by means of formal ontologies, thus allowing the Geolat system itself as well as external applications to capture (at least partially) the data, to draw inferences on them and facilitating data integration. Given the specificity of the domain and of the task, a Geolat ontology has been built: We will discuss it and its relationship with other geographical ontologies widely used in the Linked Data world, such as Geonames [7].

2. Getting Started

What is here described is a running project - geolat - with some possible extensions.

2.1 The Geolat Project

Geolat project is intended to build a complete digital library of classical and late latin texts, to annotate them in every place name and then to offer a double interface to browse the texts: a cartographic one, and a mixed one (textual + cartographic). The first one allow users to choose a region on a map and to obtain a list of authors, and works, and place names pertaining to that region; the second one allows to search for a specific place whose name is written in a text field; the search will show a list of passages, a display of places onto a map, histograms of presence of place names in various sections of the related works, list of other cooccurent place names.

2.2 The Geolat Prototype

A prototype of geolat system will be built for the first months of year 2014 and will be avaliable online at the address http://www.geolat.eu. It will offer a small digital library with some ten latin texts. The placenames contained in these texts will be manually annotated using a geographical ontology specifically conceived for the classical (latin) world and literature. A cartographic interface will allow to browse the texts clicking regions or places showed onto a map.

2.3 Building the Digital Library

As it was said, the first step is having a digital library of texts. The texts in the library will be annotated using a mixed approach: the in-line annotation is based on the TEI standard, and links to an external ontology which describes geographical entities how the ancient Romans did: a forest is the home of a nymph, a river is sacred to a goddess, a city can have more than one founder, and so on.

2.5 Browsing the Texts through a Map-based Interface

The map-based interface is the more complex part of this project because of its novelty and because of some technical problems which must be solved (e.g. finding geographical places in a given radius from another one).

3. Going further

Project geolat can go further in two non mutually exclusive ways:

adding literatures
adding ontologies (that is, types of entities).

3.1 More Literatures

Adding more literatures to the initial latin one would allow to build distinct layers of literary interpretation and to extend the scope of analysis and queries. For instance we could ask in which texts (and inside them in which passages) of Latin and French literature we find contextual references to the city of Appida and to the person of Caesar.

From the technical point of view different literatures mean different digital libraries which must define a formal protocol for interoperability. There are some basic conditions to assure this interoperability (use of UTF-8 characters encoding; adoption of at least a basic markup - preferably in XML - describing and identifying its structure: title, sections, subsections, but also paragraphs, sentences, words; adoption of open software systems).

But the only way to assure the level of interoperability we envisage to heterogeneous collections is agian based on sharing conceptual and ontological assumption.

3.2 More Ontologies

Adding more ontologies, that is more types of categories, would offer a more complex access to the texts and the knowledge they contain.

3.2.1 Persons

Recognizing and semantically annotating geographical entities in only a first step to the goal of providing an innovative access to literary works. As witnessed by the work of Francesca Tomasi [8], persons and characters play a central role in many literary productions. In particular, we will describe how the system support the identification and the formal representation of semantic relationships between texts and persons (or characters), among persons themselves and, in general, between persons and other resources and how the EAC-CPF Ontology [9] can be exploited to this aim. The analysis of the context, as conceived in the archival domain, is a fundamental approach to connect people to documents on the basis of the role or function covered. Roles are the key tool to manage self-explanatory relationships. For this reason the Pro Ontology [10] will be also considered. We will also describe how the integration of the geographical and the personal perspectives in a homogeneous framework enhances the benefits provided by both.

3.2.3 Events

Since people participate to events and many relevant events take place in the real world (meetings, battles, travels, etc.), we believe that the notion of event is a sort of conceptual glue between geographical places and people that should be captured within our system. We will discuss this issue.

Events are of different types and are bound to time, to a specific point or duration in time.

Good examples of ontologies for events are the "Simple Event Model":

lov.okfn.org/dataset/lov/details/vocabulary_sem.html[11] or the "Linking Open Descriptions of Events": linkedevents.org/ontology

3.2.4 Textual Objects

Moreover, the work of Peroni and Vitali on semantic publishing [12] [13] proved how the semantic technologies can be exploited in order to enrich the meaning of documents published on the Web by specify meaningful relationships between them [14]. We will discuss how this approach can be applied in to the digital library provided within the discussed system, in order to complement it with a rich network of semantically linked resources connecting both the productions by themselves and their contents (places, persons, characters and events) in a coherent semantic graph.

References

1. **M. Lana, Geolat**: *Geography for Latin Literature*, in (forthcoming) ISAW papers 7, Current Practice in Linked Open Data for the Ancient World Editors: Thomas Elliott, Sebastian Heath, John Muccigrosso, sfsheath.github.io/lawdi-publication/isaw-papers-7.xhtml

2. **Di Iorio, A., Peroni, S., Vitali, F.** (2011). *Using Semantic Web technologies for analysis and validation of structural markup*. In International Journal of Web Engineering and Technologies, 6 (4): 375-398. Olney, Buckinghamshire, UK: Inderscience Publisher. DOI: 10.1504/IJWET.2011.043439

3. **Peroni, S., Gangemi, A., & Vitali, F.** (2011). *Dealing with markup semantics*. In Proceedings the 7th International Conference on Semantic Systems (I-SEMANTICS 2011): 111–118. New York, New York, US: ACM Press. DOI: 10.1145/2063518.2063533

4. **Picca, D., Gliozzo, A. M., & Gangemi, A. (2008). *LMM: an OWL-DL MetaModel to Represent Heterogeneous Lexical Knowledge*. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 08). Marrakech, Morocco: European Language Resources Association (ELRA). ISBN: 2-9517408-4-0**

5. *Open Annotation Data Model*, W3C Open Annotation Community Group 2013, www.openannotation.org/spec/core

6. **Tom Heath and Christian Bizer**, *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool 2011

7. *Geonames Ontology*, www.geonames.org/ontology/documentation.html

8. **] Tomasi, Francesca**. (2013). *Digital editions as a new model of conceptual authority data*, JLIS.it 4.2 21-44

9. **] Mazzini, Silvia, and Francesca Ricci**. (2011). *EAC-CPF Ontology and Linked Archival Data*. In Semantic Digital Archives (SDA) Proceedings of the 1st International Workshop on Semantic Digital Archives. ceur-ws.org/Vol-801/

10. **Shotton David, Peroni Silvio**. (2010). *Pro - The Publishing Roles Ontology*purl.org/spar/pro/ (last modified 2013-05-15)

11. **Willem Robert van Hage, Véronique Malaisé, Roxane Segers, Laura Hollink, Guus**

Schreiber.*Design and use of the Simple Event Model (SEM)- Web Semantics: Science, Services and Agents on the World Wide Web*, Volume 9, Issue 2, July 2011, Pages 128–136 dx.doi.org/10.1016/j.websem.2011.03.003 Postprint at: www.cs.vu.nl/~guus/papers/Hage11b.pdf

12. **Peroni, S., Shotton, D., Vitali, F.** (2012). *Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents*. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502

13. **Peroni, S., Shotton, D., Vitali, F.** (2012). *Scholarly publishing and the Linked Data: describing roles, statuses, temporal and contextual extents*. In Presutti, V., Pinto, H. S. (Eds.), Proceedings of the 8th International Conference on Semantic Systems (i-Semantics 2012): 9-16. DOI: 10.1145/2362499.2362502

14. *Semantic Publishing and Referencing Ontologies*: purl.org/spar