HISTORICAL PLACE NAMES: FROM ARCHIVES TO LINKED OPEN DATA

Fresta Giuseppe[1], Martinelli Massimo[1], Signore Oreste[1]

[1]CNR-ISTI, Pisa, Italy, oreste.signore@isti.cnr.it

Keywords: Historical place names, Linked Open Data, Semantic Web, Ontology, Historical geography, Shared Knowledge

ABSTRACT

Place names and administrative boundaries are changing over time. The importance of historical place names and administrative/religious boundaries is widely recognized by scholars. In implementing a geographic names repository, several issues emerge, especially if the considered time range spans several centuries. Historical data get value if they can be put in their context, and this feature requires a solid data infrastructure. The pilot study leading to TGN relied on a database structure. The "ontological" approach and the LOD paradigm are offering even bigger advantages: interoperability and openness are the most relevant, because any information modeled using Semantic Web standards (like RDF and OWL) can be freely accessed and referenced by any web application. In addition, information is not bounded to be hosted on a single site/repository, but can be distributed everywhere on the Web.

The project currently under way aims to make historical place names available according to the LOD paradigm. The first data sample has been the one used in the previously recalled pilot study. Even if the ontology conforms to the "golden rules" for Linked Open Data, it is not fully satisfactory, as there is little reference to shared ontologies. Therefore, a new version of the ontology, with greater emphasis on events as the cause of changes, is currently under development. Combining several available ontologies (including CIDOC CRM) data will be represented as a set of triples, as required by the underlying RDF model, and made publicly available as LOD, while a HTML5 application will support navigation, querying and rendering of data. The appropriate framework to support map interaction as well as the possibility of supporting some kind of "social" contribution (comments and gathering additional information) are currently under investigation.

INTRODUCTION

Place names and administrative boundaries are changing over time, even nowadays. Just as an example, think about the recent cases of Germany (East Germany and West Germany born in 1949, and merged in 1991), or USSR, established in 1922 and split in several independent countries in 1991, with some boundary changes during its existence.

Historical geography raises several issues: first of all, *place names* change, as well as *administrative status* and *boundaries* (administrative, political, and religious) vary over time. Interestingly, just to add some complexity, these changes are sometimes related, but often occur *independently*.

At first glance, one could suppose that there are some "*invariants*", especially at the physical layer. Unfortunately, it is well known that even shoreline, lakes, rivers and marshes can vary.

Finally, in implementing a geographic names repository, another issue emerges, especially if the considered time range spans several centuries. For existing places the temporal duration of a name or administrative/religious boundaries is almost always imprecise (e.g. around 5th century BCE), with both the lower and upper bound given as approximate time intervals. Some present places could not have existed in the past, or there is no information available. When we can link a present place to an old one, with its name variants, we can have conflicting or missing information. Finally, boundaries can only be drafted in a very approximate fashion, if not unknown at all.

For example, you can find this kind of information about the present locality **Sezze**:

> According to a legend, the city was founded by the mythical hero Hercules … The historical *Setia* appeared around the *5th century BCE* as the *Volscan settlement* member of the Latin League. It became a *Roman colony* in *382 BC*, and flourished because of its

strategic and commercial position near the "pedemontana" way and the Appian Way, the road that connected Rome to southern Italy

The importance of historical place names and administrative/religious boundaries is widely recognized by scholars [1], [2]. A pilot study [3] supported by Getty AHIP under the auspices of CIHA considered 111 municipalities, 738 localities, described by 2676 items of information (historical names, ecclesiastical jurisdictions and historical and administrative status) accompanied by 3543 bibliographical references and resulted in a data model (the "TAU model"). The pilot study led to the well-known TGN (Thesaurus of Geographic Names) often referred as geographic names ontology [4]. TGN presently contains nearly one million place names, representing approximately 900,000 places. Data can be freely accessed and downloaded in XML format, but can't be directly referenced as Linked Open Data (LOD), even if plans in this direction have been announced.

Historical data get value if it is possible to reproduce the full context at any given point in time, with appropriate reference to place names, administrative boundaries, events, culture, and persons. Equally important is to have the possibility of showing the evolution in time of this kind of information, even better if supported by an interactive graphical representation. An example of such facility is the possibility of displaying a map showing names and boundaries at a user selected time. The disciplines involved are so many, and the amount of needed information is so large that it is chimerical to imagine that this complex knowledge can be made available at a single data source. Much more probably, some specialized knowledge repositories, which should be linked in some way to have a semantically rich picture of the context, will exist. This is not conceptually new, as any scholar is following this process, just combining her/his knowledge in different fields. The matter is simply that scholars' knowledge remains often implicit or unexpressed, while it should be made explicit and available to everyone. In addition, the present sources of information are so dispersed on the Web and in the archives that the process risks to be impossible or at least lead to unsatisfactory or incomplete results. We will see in the following how to support the scholars' needs. It is evident, however, that a solid data infrastructure is a pre-requisite.

In passing, the previously quoted issue of imprecise time intervals brings our attention towards the need of an appropriate representation of dates. We will conform to the formalism presented in [5] where we have a simple and human readable external format (with an obvious meaning of the terms):

$$\mathbf{D_p \ | \ (D_{min} - D_{max}) \ | \ (D_{min} - D_{max}) \ D_p}$$

where $\mathbf{D_x}$ can be followed by "(?)" if it is just a trial. This formalism supports a wide range of *date granularity*, can handle *multicultural calendars* (an important issue in the Web) and has been implemented and tested for several of them (Gregorian, Hebrew, French Revolution, and Muslin). Dates can be parsed according to a formal grammar and are stored in an internal suitable format. We must recall that ordering such imprecise dates is not obvious, and a *temporal algebra* is needed to select and sort time related data.

The pilot study that led to TGN relied on a database structure, as it was easily recognized that a conventional thesaurus would have been inadequate to represent all the possible combinations of different names and boundaries, and terms relationships defined in the ISO standard for thesauri are unsuitable to model the relationships needed to represent the evolution of names and boundaries. The "*ontological*" approach and the LOD paradigm are offering even bigger advantages: first of all, persistence of data across technological evolution, assuring even better independence from the underlying technology. But the greatest advantages are *interoperability* and *openness*. Any information modeled using Semantic Web standards (like RDF and OWL) can be freely accessed and referenced by any web application. In addition, information is not bounded to be hosted on a single site/repository, but can be distributed everywhere on the Web.

In short, the LOD paradigm is the way of moving from the "*Web of Documents*" to the "*Web of Data*". The *Web of Documents* is the Web we are accustomed to, therefore a Web that can be seen as a *global filesystem*, where (fairly structured) *documents* are the primary objects, connected by untyped links, and semantic of content and links is implicit. As a consequence, the

Web of Documents is mainly designed for human consumption and has a great level of simplicity. On the other hand, data are disconnected, and additional knowledge can be extracted only by human action and reasoning.
The *Web of Data*, instead, can be seen as a *global database*, where primary objects are *Things* (or their description). We have *typed links* between things (including documents), high degree of structure in (description of) things, and *semantics* of content and links is *explicit*. As a consequence, the Web of Data is designed for machines first, and humans later. Thanks to the explicit semantics, machines can perform some reasoning and deduce additional knowledge.
The LOD rules are defined in [6] and are quite simple. In essence they are:

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things.

Researchers and implementers make reference to the following "five stars" model:

1. *On the Web*: available on the Web (whatever format) but with an open license, to be Open Data.
2. *Machine-readable data*: available as machine-readable structured data (e.g. excel instead of image scan of a table).
3. *Non-proprietary format*: as above, plus non-proprietary format (e.g. CSV instead of excel).
4. *RDF standards*: all the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
5. *Linked RDF*: all the above, plus: Link your data to other people's data to provide context.

Interested readers are referred to [7].
A fully operational and effective LOD approach requires support from *ontologies*. There are many definitions of the term ontology. One of the most frequently quoted is given in [8]:

> An ontology is a formal, explicit specification of a shared conceptualisation. A '*conceptualisation*' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. '*Explicit*' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. '*Formal*' refers to the fact that the ontology should be machine readable, which excludes natural language. '*Shared*' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.

In spite of the many existing definitions of ontology, there is a general agreement about its meaning. As a matter of fact, ontology captures *shared knowledge* and is the "glue" that allows machines to understand data, make them really interoperable, and support linking among different concepts [9].
The project currently under way aims to make historical place names available according to the LOD paradigm. The first data sample has been the one used in previously recalled pilot study. In addition, the possibility of representing hierarchy of administrative bodies has been inserted in the model.

RESULTS

As a first step, we proceeded to a "flat" translation of the original TAU model. It was quite an easy task, but a closer examination showed that actual support of *semantic interoperability* was quite poor. The main reason was that it was not referring to any shared model. Therefore, we defined a bit more sophisticated ontology (**hgo**: **h**istorical **g**eographical **o**ntology), whose main characteristics are the ability to manage independent information about place names and administrative status, and modelling of administrative hierarchies. Both are improvements of the

original TAU model, where place names and administrative status were always represented as place name, and administrative hierarchies were ignored, recording just one (not necessarily the lowest) level of administrative belonging. Both these aspects can have a relevant impact when rendering data on a map, showing non necessary data and affecting the granularity of rendered boundaries.

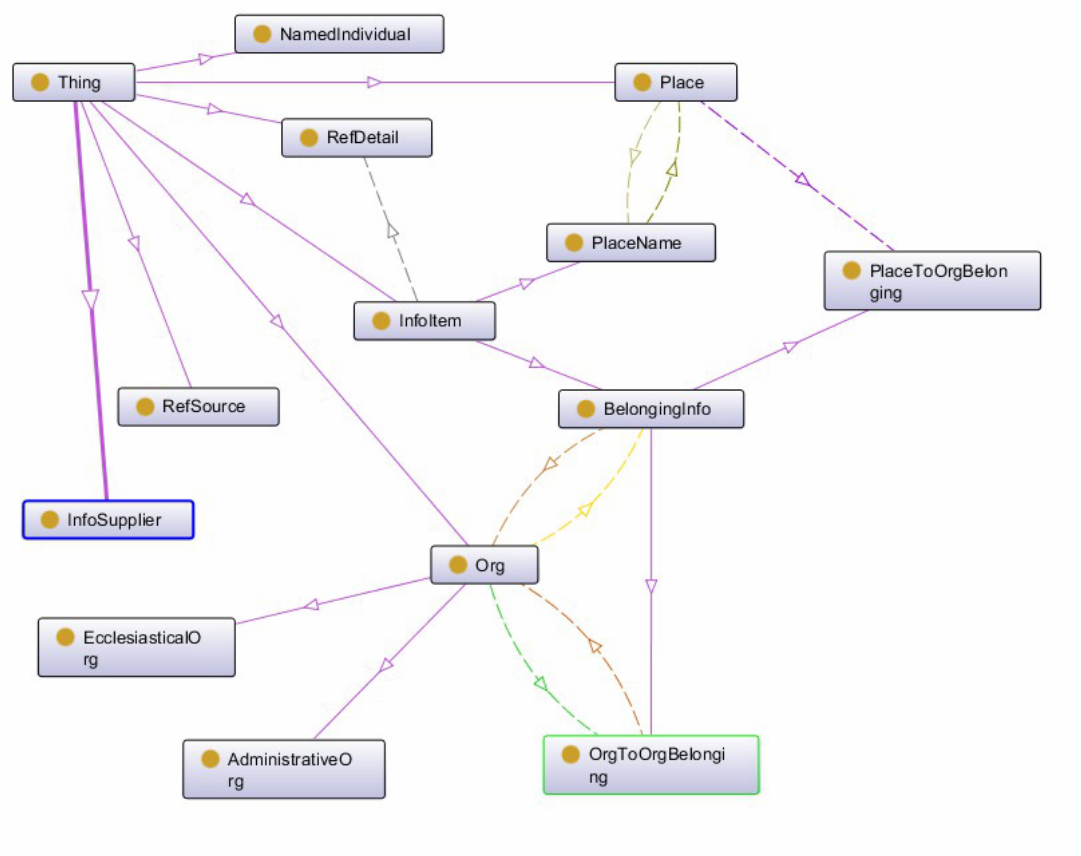

Figure 1 - A graphical representation of hgo ontology (via Protégé)

The **hgo** ontology (see Figure 1) can be considered satisfactory, as in its *intension* models name and administrative changes, supports multilingualism and overriding information, and is published on the Web, so that can be referenced by everybody. In its *extension* (data) is available as RDF, can be queried via SPARQL and can be linked by everyone from everywhere. As a result, data are no more forced to be stored in a single site. Everyone can build a repository of historical place names (and related administrative or religious info) just making reference to the **hgo** ontology, and can share data across the Web.

For example, querying *past names* of the locality presently known as Sezze can be performed with the following SPARQL query (here and in the following example we omit the complete syntax with reference to the appropriate namespaces):

```
SELECT  DISTINCT  ?PresentName  ?PlaceName  ?Since  ?Until
 WHERE { ?PlaceNameId     hgo:isPlaceNameOf   ?PresentNameId .
         ?PresentNameId   rdfs:label          ?PresentName .
         ?PlaceNameId     rdfs:label          ?pl  .
         ?PlaceNameId     hgo:startDate       ?snc .
         ?PlaceNameId     hgo:endDate         ?untl .
       FILTER regex(?PresentName, "Sezze", "i" )
       BIND(REPLACE(str(?pl), "\"", "") AS ?PlaceName) .
       BIND(str(?snc)  AS ?Since).
       BIND(str(?untl) AS ?Until).}
```

```
    ORDER BY ?PlaceName
```
returning as result:

PresentName	PlaceName	Since	Until
“Sezze”@it	“CastrumSitiense”	“sec. XIII”	“sec. XVI (?) – sec XVII (?))”
“Sezze”@it	“CastrumSitinum”	“sec. XIII”	“sec. XVI (?) – sec XVII (?))”
“Sezze”@it	“Secia”	“anno 1478”	“sec XVII (?)”
“Sezze”@it	“Setia”	“circa 383 a.C.”	“sec. XI”

clearly showing the existence of conflicting information, or better the co-existence of two different place names in the same time period
In a similar way, the SPARQL query:
```
SELECT  DISTINCT  ?PresentName  ?AdmStatus  ?Since  ?Until
 WHERE { ?PresentNameId  rdfs:label  ?PresentName .
         ?PlaceNameId  hgo:hasPlaceToOrgBelonging  ?BelongingId .
         ?BelongingId..hgo:hasUpperLevelOrg  ?org .
         ?org  rdfs:label  ?adm .
         ?BelongingId  hgo:startDate  ?snc .
         ?BelongingId  hgo:endDate  ?untl .
       FILTER regex(?PresentName, "Sezze", "i" )
       BIND(str(?snc)  AS ?Since).
       BIND(str(?untl) AS ?Until).
       BIND(str(?adm)  AS ?AdmStatus).}
```
will return as result:

PresentName	AdmStatus	Since	Until
“Sezze”@it	“Regno d’Italia”	“anno 1870”	“anno 1946”
“Sezze”@it	“Stato della Chiesa”	“anno 1414”	“anno 1870”
“Sezze”@it	“Ducato Romano”	“sec. VIII”	“sec. IX”

clearly showing that there is a lack of information about a time period.
User interaction must support the possibility of navigating among results, query formulating and refining, extending the horizon towards other information sources (e.g. historical events, persons). Interactive maps can play a significant role, but the appropriate framework to support map interaction is currently under investigation, as it is just one of the possible interaction paradigms.
At first glance, we could argue that the implementation fulfils all the requirements to be classified as a five stars LOD. However, there are some points to consider.
First of all, the ontology can’t be considered a shared ontology. Even if it can be referenced and used, we have to consider that some concepts are yet modeled in others well established ontologies, like CIDOC CRM, an International Standard, which is a core ontology with high abstraction level, extensible and suitable for automated spatial and temporal reasoning [10].
Just as an example, CIDOC CRM has its own way of representing places and place names:

- *E53_Place* may be identified by one or more instances of *E44_Place Appellation*. Places can be structured in a hierarchy;
- *E44_Place_Appellation* is a class which comprises any sort of identifier characteristically used to refer to an *E53 Place*;
- *E48_Place_Name*: “Place Names may change their application over time: the name of an E53 Place may change, and a name may be reused for a different *E53 Place*.”.

Formally:
```
      E53_Place                 P88_consists_of   E53_Place
      E44_Place_Appellation     P87i_identifies   E53_Place
      E48_Place_Name subclass_of  E44_Place_Appellation
```
In the CIDOC CRM model, Athens and Greece are both instances of E48_Place_Name.

Therefore, an appropriate modelling should reuse classes and properties defined in CIDOC CRM, adding extensions if needed. To share the knowledge, we have to take into account some other existing ontologies, like geonames [11].
Even more relevant is the issue of appropriate documentation of changes of names, administrative status, administrative belonging, etc. In fact, these changes are presently just reported with reference to the time interval and references. However, this approach ignores that *changes happen* because of some *events happen*. The event can just be a resolution, like a governmental decree, or the result of a war or something else. Modelling changes as effect of events leads to a much richer model were events and their actors can be stored, as LOD, on other specialized sites, opening the doors to real knowledge enrichment via reasoning upon available data.
Finally, even without discussing the issue of the so called Web 2.0, but returning to the roots of the Web, we must consider supporting some kind of "social" contribution (comments and gathering additional information). User interaction with rewarding and uplifting is the target.
Combining several available ontologies (including CIDOC CRM and geonames) data will be represented as a set of triples, as required by the underlying RDF model, and made publicly available as LOD, while a HTML5 application will support navigation, querying and rendering of data.

CONCLUSIONS

Historical geography is a concern both for late past times as well as the present. In 1988 a pilot study leaded to the TAU model, and subsequently to TGN.
The intrinsic nature of cultural heritage data requires a multi and interdisciplinary approach, to exploit the richness of the many existing semantic relationships. The huge amount of available data, often published on the Web, as well as the cultural differences among the scholars make unrealistic to concentrate data on a single site or force them to conform to a unified schema. A decentralized approach, based upon Semantic Web technologies, and adoption of the LOD principles can offer a viable solution and constitute a solid framework to share and enhance knowledge.
The project presented in this paper uses as test bed the TAU data, with some semantic enrichment, and makes them available on the Web according to the Linked Open Data paradigm, with a reference ontology. However, just adopting the technologies, defining ontology and publishing data in RDF doesn't implement Linked Open Data. It is important to make appropriate reference to well established and shared ontologies, which can eventually be extended to achieve a more accurate model tailored to specific needs. Even if this process is more demanding than just defining an ad hoc ontology, the achievements are much more satisfactory and in line with the Semantic Web and LOD principles.

ACKNOWLEDGEMENTS

Our warm thanks are deserved to Enrico Rendina from Centro per la ricerca e lo sviluppo di Metodologie e Applicazioni di Archivi Storici (MAAS) for raising attention on several issues about historical geography and for the useful discussions.

REFERENCES

[1] K. Janowicz, The role of place for the spatial referencing of heritage data, The Cultural Heritage of Historic European Cities and Public Participatory GIS Workshop. 17-18 September 2009, The University of York, UK, 2009
[2] Kauppinen, T., Väätänen, J., Hyvönen, E.: Creating and using geospatial ontology time series in a semantic cultural heritage portal. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 110–123. Springer, Heidelberg (2008)
[3] Papaldo Serenita, Signore Oreste, Un approccio metodologico per la realizzazione di una banca dati storico-geografica (A methodological approach to producing a historical/geographical databank - Multigrafica Editrice, Roma (1989), pp. 573, ISBN 88-7597-105-6
[4] Getty Thesaurus of Geographic Names® Online, http://www.getty.edu/research/tools/vocabularies/tgn/about.html

[5] Signore Oreste, Bartoli Rigoletto, Fresta Giuseppe, Marchetti Andrea, Issues on historical geography, Proceedings of ICHIM'97 - Fourth International Conference on Hypermedia and InterActivity in Museums - Paris, France, 3-5 September, 1997 p.252-257 (Archives & Museum Informatics, 1997)
http://www.archimuse.com/publishing/ichim97/bartoli.pdf
[6] Tim Berners-Lee: http://www.w3.org/DesignIssues/LinkedData.html (2007)
[7] Tom Heath, Christian Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. (2011) http://linkeddatabook.com/book
[8] Rudi Studer, V. Richard Benjamins, Dieter Fensel, Knowledge Engineering: Principles and Methods, Data, Knowl. Eng. 25(1-2): 161-197 (1998)
[9] Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, Ontological Engineering, Springer-Verlag (2004), ISBN 1-85233-551-3
[10] CIDOC CRM, http://www.cidoc-crm.org/
[11] http://www.geonames.org/ , http://www.geonames.org/ontology/ontology_v3.1.rdf