Language, Cultural Influences and Intelligence in Historical Gazetteers of the Great War

Robert Warren
Big Data Institute, Dalhousie University
Halifax, Canada
rhwarren@dal.ca

Bo Liu
Big Data Institute, Dalhousie University
Halifax, Canada
BoLiu@Dal.Ca

***Abstract*—Historical gazetteers trace locations that have been long forgotten while allowing for the cross-referencing of locations across different documents. In this work, we present the problem of managing a gazetteer of geometries, features and names during the Great War on the Western Front. The careful tracking of provenance information and the novel use of existing semantic web standards allows for the discovery of both the quality of the cartographic work done by both sides and the cultural influences between belligerents.**

***Keywords*-Historical Gazetteers, Linked Open Big Data**

I. Introduction

The availability of plentiful and electronically accessible data sources is driving a renewed interest in historical gazetteers as a means of locating places that have been forgotten or changed over time. Besides providing spatial contexts to toponymy, gazetteers in the Linked Open Data (LOD) contexts can be used for information discovery and integrating different data sets.

In this paper we review some issues in handling complex historical gazetteer databases that come about when binarizing (transforming scanned images of maps into digital features) maps dating from the Great War. Complexity and size are not new issues in gazetteers but linking them through Semantic Web technologies with machine generated data is creating new opportunities in integrating multiple data sources.

In the past, some of these issues would have been dealt with using workarounds or textual comment fields. Big Data, or more accurately the Big Data of online archives, means that these coping mechanisms are no longer possible. The data binarized from thousands of maps can re-position a feature over a dozen times simply because of changing survey techniques, even before change and movement is taken in account.

A number of open and LOD-driven gazetteer are in current use including Geonames[1], the UK Ordnance Survey, Ordnance Survey, the Norway historical Gazetteer, Past Places [1] and the Linked Geo Data [2] version of OpenSteetMap. Previous approaches to dealing with change have been dealt with by the Finish Seco [3] project with an event driven framework that records partOf relationship between features. The Linked Geo Data [2] ontology does not record changes in itself, but does preserve the identity of points that make up geometries. The Pelagios [4] project has made the most comprehensive effort so far to have fully specification labels, features and geometries. This paper deals with the naming of features, or toponyms, across languages and cultures under uncertainty.

A. Big Data Toponymy

Inherently a gazetteer that is to be linked across projects and data sources is expected to be multicultural and multilingual: names serve as an identification of an object. However, even in mono-cultural data-sets that span multiple eras this can become a problem: names can change (e.g. through the fusion of communities, see [3]), different spellings or transliterations are introduced, vernacular names disappear and are replaced by an officially surveyed names, etc.

The resolution of these bodies is important is that it opens access to a number of historical documents that were not intended to reference the gazetteers such as personal letters and diaries. These documents are rich in local knowledge that is not available from official sources that focus on large scale events.

The problem is compounded when the nomenclature used is one of local knowledge that is not accumulated in an official body of knowledge like a governmental gazetteer. A typical example is when the local population references a particular landmark with a nickname of the sort of “The McDonald Farm”. The McDonald family may not have held that farm or field in several generations, or even lived in the community, but the name has been retained as a reference to the geographic feature.

Direct translation is not an answer since different cultures will use different names that reference deeper cultural roots, as in the case of the English referencing “The English Channel” and the French “La Manche”. The benefit of a cross-linguistic gazetteer to record these differences is that the actual name used projects the belief and background

[1] http://www.geonames.org/

of the writer which can in turn be used for provenance recording.

This issue was a concern in creating digital maps of the Western Front during the Great War as the same trenches or farms would be referenced by maps of different belligerents with different names. Some of these features, such as trenches would change shape over time and tracking this movement was also an area of interest.

B. Feature and Geometry

There exists a distinction between the thing (the feature) and its physical shape and/or location (the geometry). The typical example is one of a river whose bed moves over time as its flow changes. The river does not change in nomenclature or identity, but its course does, hence requiring this difference between feature and geometry.

Geographic Information Systems (GIS) have taken different approaches to modelling geographic information in that 1) both feature and geometry instances can be merged, 2) the level of granularity used in describing the geometry and 3) how the change in geometry is modelled. The separation of feature and geometry is used by GeoSparql [5] while the Linked Geo Data [2] version of OpenSteetMap does not differentiate them.

The recording of the individual points that make up the geometric shape is done through encapsulation in Well Know Text (WKT) or Geographical Markup Language (GML) by GeoSparql. The NeoGeo vocabulary and Linked Geo Data make use of a series of geo:Point's whose ordering through lists forms the geometry. The encapsulation used by GeoSparql is meant to simplify implementation by programmers by recycling previously written code. In the case of historical changes and uncertain data, each change or alternate location requires re-instantiating the entire encapsulating string, incurring a large amount of redundancy when modelling movement. In these cases, NeoGeo and Linked Geo Data approaches are preferable since the underlying geometry points can be reused.

Expressing these physical location requires knowing an absolute position which is not always possible and even when coordinates are available, this is only to a certain accuracy. How to communicate imprecision and uncertainty about the location and/or the shape remains an open problem. As of the writhing of this paper, the OGC's Geography Markup Language properties such as horizontalAbsoluteAccuracy have not been ported to the GeoSPARQL [6]. Some experimentation using a complex amalgam of the CRM-CIDOC and GeoSPARQL [7] vocabulary has been demonstrated, but with a high computational cost.

This is a concern in that a cornerstone of the Scientific Method is to keep track of significant digits when recording measurements; position information should be no different. It is a pressing concern that none of the current standards for recording positions differentiate between 4.56, 4.560 and $4.5\overline{6}$. Currently, 'approximate' positions are recorded using a mixture of bounding boxes, simple point references and under-specified properties pointing to approximation polygons.

II. Proposed Solution

There currently are three basic sources of name information in use in the Semantic Web: the FOAF (Friend of a Friend) ontology, the SKOS series of vocabularies and the basic RDF label property. While FOAF has it roots in an early attempt at communicating social networking information, SKOS focuses on documenting taxonomy documents using various labels properties such as Preferred, Alternate or Hidden.

FOAF is a mature ontology that focused on the person and has a limited number of properties for naming. While currently the de facto standard for representing people on the Semantic Web, it still has a number of outstanding issues (see Brown and Simpson [8]) that were never resolved when dealing with non-western cultures.

It's property foaf:name remains popular as a means of naming an instance, while SKOS has seen widespread adoption for vocabularies extracted from other knowledge systems, such as the Heritage Vocabularies [2].

An extension to the SKOS vocabulary has been the SKOS eXtension for Labels (SKOS-XL) which closely mirrors SKOS using classes which allows us to annotate labels. We use the SKOS class skos-xl:Label related to the feature through the skos-xl:labelRelation property. This does not preclude the use of additional foaf:name or rdfs:label properties itself which can remain for convenience.

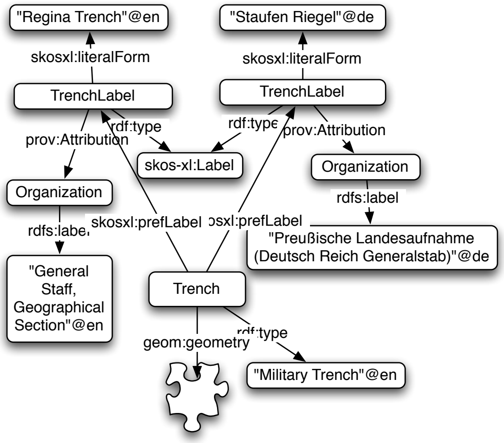

Figure 1: The trench named Regina Trench by the Canadian army was known as Staufen Riegel by the German army.

Figure 1 is an example of an German-held trench that was part of the Hindenburgh line during the Battle of the Somme in the Great War. It was named Regina Trench by the

[2] http://www.heritagedata.org/blog/vocabularies-provided/

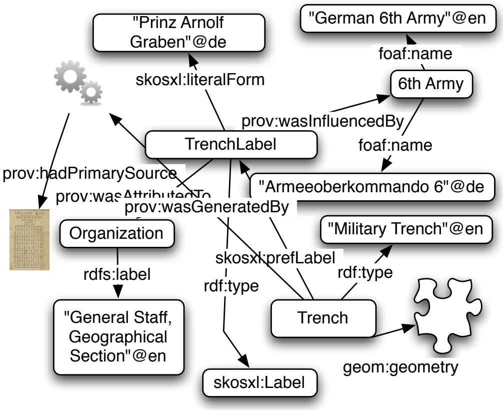

Figure 2: In this case, the Geographic Section of the General Staff (GSGS) used the German army name for its own maps.

Canadian units that were attacking it while called Staufen Riegel by the defending German units.

In this case, the data obtained from both Canadian and German documents references the same feature but with different names. At the cost of additional complexity we can trace whether the name of the location implies either the German or Canadian experience of the event. Furthermore while both views concern the same military trench (feature) within time and space, the provenance of the different nomenclature is recorded as being from two different surveying sources. This allows us to resolve the feature entity to a single one while keeping track of two different and concurrent nomenclature perspectives without selecting an "authoritative" one.

Few cultures are insular and it is common for one organization to borrow artifacts from another. In the case of maps of the Great War it was common for armies to label their maps with the feature names of their adversaries to facilitate the orientation of their troops. Figure 2 is a representation of the structure used to track not only the name assigned to the feature but what influenced the Geographic Section's choice of labels, which is in this case the German 6th Army's nomenclature.

There is no documentation available as to which units were responsible for gathering the data or whether one of the field survey units organically absorbed the data from local knowledge. Thus we do not want to instantiate a series of organizations and activities who contents are unknown, but we do wish to record some information about this process. The cultural influence can be recorded here using prov:wasInfluencedBy from the W3 Provenance ontology which while loosely specified, represents a process that is easy to detect but hard to qualify without a detailed understanding of the General Staff Geographic Section's internal processes. However, if we do wish to reference the intelligence activity that occurs within this army, without designating a specific military intelligence unit, we add that its creation was informed (prov:wasInformedBy) by military intelligence as an activity (prov:Activity).

III. Conclusion

In this paper we presented some of our current approaches to recording complex geographical changes to event data from the Great War in a manner that is not lossy. One of the challenges of both Big Data and the Semantic Web is recording detailed information without creating a new information management problem: it is difficult to convince practitioners to write very complex and detailed RDF/OWL documents about the place of death of Admiral Nelson at the Battle of Trafalgar, when simply specifying the orlop deck of the HMS Victory as a feature will achieve the same thing. Simplifying structures and best practices will remain a topic of heavy debate for some time to come.

In closing, an observation is that a part of the promise of the Semantic Web is not completely about creating "correct information" as much as recording partial information in a useful manner. Combined with the sheer volume of Big Data, this will allow researchers to infer new knowledge by building on previous work as much as their own.

Acknowledgements

This paper follows up on gazetteer and geometry problems identified during stimulating conversations with Bettina Waldvogel and Thomas Scharrenbach at the Swiss Federal Institute for Forest, Snow and Landscape Research.

References

[1] H. S. et al., "Pastplace: the historical gazetteer service from the people who brought you a vision of britain through time," in *UK Archives Discovery Forum*. National Archives, 2013.

[2] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "Linkedgeodata: A core for a web of spatial open data," *Semantic Web Journal*, vol. 3, no. 4, pp. 333–354, 2012.

[3] T. Kauppinen, J. Väätäinen, and et al., "Creating and using geospatial ontology time series in a cultural heritage portal," in *European Semantic Web Conference*, 2008, pp. 110–123.

[4] L. Isaksen, R. Simon, and et al., "Pelagios and the emerging graph of ancient world data," in *Proceedings of the 2014 ACM Conference on Web Science*, 2014, pp. 197–201.

[5] OGC, "OGC GeoSPARQL - a geographic query language for rdf data," Open Geospatial Consortium, Tech. Rep. OGC 11-052r4 OGC 11-052r4, September 2012.

[6] R. Battle and D. Kolas, "Enabling the geospatial semantic web with parliament and geosparql," *Semantic Web Journal*, 2011.

[7] M. Doerr and G. Hiebel, "Crmgeo: Linking the cidoc crm to geosparql through a spatiotemporal refinement," Institute of Computer Science, Tech. Rep. GR70013, April 2013.

[8] S. Brown and J. Simpson, "The curious identity of michael field and its implications for humanities research with the semantic web," in *IEEE Big Hum. Data*, 2013, pp. 77–85.