Proposal to participate in the NCGIA specialist meeting on Digital Gazetteer Research and Practice

Krzysztof Janowicz
Institute for Geoinformatics, University of Muenster
Robert-Koch-Str. 26-28, 48149 Muenster, Germany
Email: janowicz@uni-muenster.de
Webpage: http://ifgi.uni-muenster.de/~janowicz

Similarity-Based Identity Assumptions for Historical Places

The domain of cultural heritage is very heterogeneous; the themes or exhibits that museums and related institutions are concerned with range from history of science to various kinds of art and historical documents, and biodiversity. Accordingly, the number and type of preserved exhibits range from millions of collected organisms to a small number of valuable paintings. Creating and maintaining metadata about exhibits and historical facts in general gets increasingly important for scholars and curators in order to structure, manage, and query their own data. As long as metadata is used for internal workflows only (such as the preparation of an exhibition), each institution may develop and maintain their own schema and representation format; however, to refine and enrich their own knowledge base or to answer complex scientific questions, interchange with external sources becomes necessary. Cleaning up the local knowledge base is especially important because one needs to keep in mind that historical knowledge may be vague, incomplete, or even misleading. To support these tasks the Committee on Documentation (CIDOC) provides a well established and standardized core ontology (called CIDOC CRM; ISO 21127) [1] intended to annotate heterogeneous cultural heritage information to make it available in a machine-readable format (RDF) and reasonable way for knowledge integration, mediation and interchange. The long-term vision is publishing all annotated datasets through web services and therefore create a shared network of interlinked historical information enabling automatic metadata harvesting. The CIDOC conceptual reference model can be regarded as the underlying semantic level that provides meaning within the intended cultural heritage data infrastructure (which can bee seen analogously to a Spatial Data Infrastructure) by delivering a common metadata schema. Instead of trying to reach a community wide agreement on definitions for concrete entity classes (such as types of exhibits) the strength of CIDOC CRM lies in defining an abstract but interrelated vocabulary describing the fundaments of historical facts, namely established links (relations) between places, actors, objects, and events.

To make use of external data sources, however, a common language is not sufficient. It must be guaranteed that the collected metadata refers to the same real world phenomenon (which could be a historical place, person, event, or object) as the local datasets. Global authorities (such as the Alexandria Digital Library Gazetteer Server) provide unique identifiers and annotated datasets for some common kinds of real world phenomena. Scholars can refer to these global identifiers in addition to (or instead of) their local identifiers and therefore reduce maintenance effort and redundancy on the one hand and enable data interchange on the other. If compared datasets refer to the same global identifier and the scholar decides to trust the global authority as well as the external party that linked their dataset to the specific identifier, it can be assumed that the same real world phenomenon is meant.

Nevertheless, so far most datasets do not refer to global authorities and scholars must decide as the case arises whether the harvested information is relevant for their own knowledge base. There are several reasons for this:

- Knowledge about historical places is often vague and incomplete.
- Non-unique place names (even within the same area)
- Place names refer to cities, rivers, valleys, mountains, etc.
- Misinterpreted place names (e.g. 'Al Wahat' → oasis)
- Names also refer to varying geopolitical units (e.g. nomads) or prominent (artificial) landmarks (e.g. telegraph stations)
- Out-dated place or even country names (e.g. UDSSR)

Finally, the most significant reason why global identifiers provided by Gazetteers can only partially solve the problem of identity is that using Gazetteers to determine whether two datasets refer to the same real world place, presumes that all involved institutions have manually annotated millions of local datasets beforehand, which is not the case until now. Therefore an identity assumption assistant should support scholars in analysing the harvested metadata and returning promising datasets - in a way that the external datasets *probably* refer to the same real world place addressed by the local data. The identity assumption theory used by such an assistant should be non-rigid in a way that it returns a ranked list of estimations instead of trying to automatically conclude safe predictions from vague historical data.
If, in practice disambiguation via gazetteers and other global authorities (such as for historical figures) is often difficult, expensive and error-prone (especially for subordinate geopolitical units, events, actors, etc.) an identity assumption service should use the links established via the CIDOC CRM annotation between places, actors, objects, and events as additional *reference* points. In other words, taking Goodchild's geographic reality (geoinformation as a spatiotemporal location vector and an attribute/thematic vector [2]) and Kuhn's notion of semantic reference systems [3] into account, the underlying idea is to use thematic information as support for spatiotemporal reference. The same way as the spatiotemporal location vector is interpreted by a spatiotemporal reference system, thematic information is interpreted by a semantic reference system defined by CICOC CRM as a formal ontology, and similarity and classical (spatiotemporal & subsumption) reasoning as functions over this defined terminology. Proposing similarity as part of the puzzle of identity assumptions is drawing the metaphor from our geographical notion of location to the location within a network of historical facts and the spatial 'next-to' relation to a thematic one based on similarity assessments (see [4, 5] for further details on the identity assumption theory itself and similarity measurement between conceptualisations described in formal languages).

References

[1] Crofts, N., et al.: *Definition of the CIDOC Conceptual Reference Model (version 4.2)*. 2005.
[2] Goodchild, M.: Geographical data modeling. Computers and Geosciences (1992). 18(4) p. 401- 408.
[3] Kuhn, W.: Geospatial Semantics: Why, of What, and How Journal on Data Semantics III. Springer Verlag LNCS 3534 (2005) p.1-24
[4] Janowicz, K.: Towards a Similarity-Based Identity Assumption Service for Historical Places. In M. Raubal, H. Miller, A. Frank, and M. Goodchild, Eds. Geographic Information Science - Fourth International Conference. Springer Verlag LNCS 4197 (2006) p.199-216.
[5] Janowicz, K: Sim-DL: Towards a Semantic Similarity Measurement Theory for the Description Logic ALCNR in Geographic Information Retrieval. R. Meersman, Z. Tari, P. Herrero et al. (Eds.): SeBGIS 2006, OTM Workshops 2006, LNCS 4278 (2006) p. 1681 - 1692