Integrating ontologies and schema for biographic and geographic databases

Renée Sieber[1], Christopher Wellen[2], Yuan Jin[3]

[1] Geography, Environment, McGill University, renee.sieber@mcgill.ca
[2] Geography, University of Toronto, christopher.wellen@utoronto.ca
[3] Computer Science, McGill University, chengdujin@gmail.com

Abstract

This paper describes a key set of ingredients to sharing biographical and geographic information that is stored in separate databases. These ingredients include the concepts of geospatial ontologies as well as database schema. A proof-of-concept system was developed with three databases of Chinese history and geography.

Background and Relevance

Ontologies have been a dominant theme in GIScience for 10 years. The goal is to work with a higher level of semantic abstraction than the individual variables or database schema of databases. Ontologies can serve to create semantic interoperability among heterogeneous databases, that is databases of various data structures and formats and in various physical locations. We argue that application of ontologies should maintain the distinctive way of modeling concepts in each database and not enforce a single way of modeling a concept, for example, the variables comprising where a historical person lived. This is especially important for the social sciences and humanities, where there should be support for multiple, competing categorizations of concepts (Bowers and Ludascher 2004, Friedlander et al 2009). Kintigh (2006) emphasized that archeologists used numerous systems to classify, for instance, stone tools. There may be contradictory indications of where and how a person lived. Because historical geographies are often contested and researchers view their data through diverse theoretical frames, these contestations need to be preserved.

Questions remain on how the geospatial ontologies are realized. For instance, how high a level of abstraction before that abstraction becomes meaningless? How many levels of abstraction are necessary or useful? How can we operationalize what's been up till now in GIScience (at least) conceptual and logical ontologies?

Even though geospatial ontologies enjoy broad usage, at least in the academy, they are still poorly understood. This is probably because ontologies are often confused with database schema and relations among data tables. Instead, ontologies are defined as an abstraction of concepts not relations among data tables or field names (e.g., entity relation model). Ontologies comprise a set of entities, attributes, axioms and relations (Gruber, 1993). These relations in an ontology can be more descriptive than join or "is a type of". Ontologies provide the concepts behind the data. Developing ontologies also provides the opportunity for a shared conversation about what the semantics should be.

Methods and Data

We sought to integrate three databases. They are the Chinese Historical GIS, hosted at Harvard University, the Chinese Biographical Database (CBDB), hosted at Academia Sinica, Taiwan, and the Ming-Qing Women's Writers database, hosted at McGill University. The first two constitute the most significant (i.e., in number of records and completeness) databases of Chinese history and geography in the world. The third, although physically smaller compared to the others, is also the largest database of its kind.

Our approach was to build a multi tiered ontology. We follow Bowers and Ludascher (2004), who argue that ontologies should occur in multiple tiers, or levels of specificity. We chose three tiers – application ontologies (AOs), domain ontology (DO), and upper level ontology (ULO) (Figure 1). The level closest to the database is the AO, which is a mapping of each database schema; in a relational database, schemas are basically the structure of tables. A DO, shared by all databases, is a mapping between the logical objects in the dataset (specific fields in a table for instance) and objects in the conceptual schema (classes and properties). It should represent concepts as they are explicated by the domain experts, in this case Chinese history and geography. The ULO is the highest level of abstraction and should be independent of the knowledge domain. It represented categories such as place and person and was used generally to guide a standard creation of concepts in the DO.

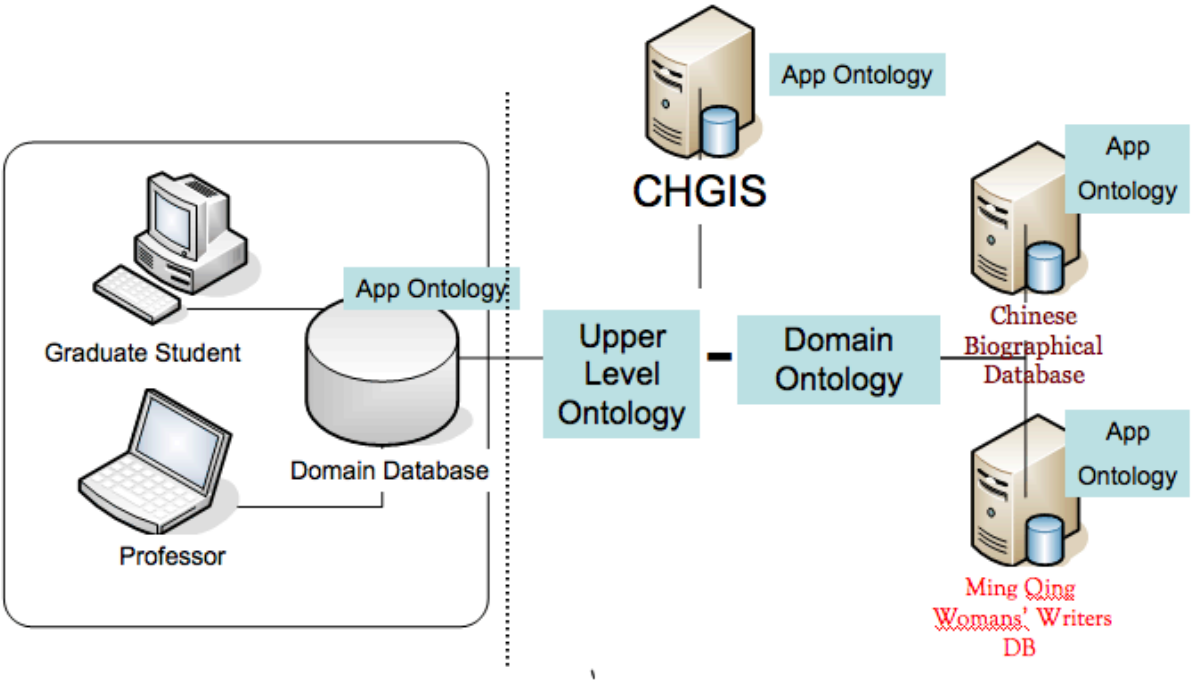

Figure 1. System Diagram, showing databases and their ontologies.

Geography present its own problems to ontologies (Egenhofer 2002) because many geographic features are determined not by single records in a geographic database but by topological relations among records as well as multiple and nested ways to represent

place. We were guided by our previous work (Wellen 2008) in development of ethnographic geospatial ontologies.

We used primarily open source or freely available software because we operated in a knowledge domain where developers did not have extensive computing resources. Much open source software is supported by large software user communities so it is conceivable that humanities resources could be extended by this support. These open source/free components included D2R and Protégé. A companion piece of software, D2RQ, converted the database schema into the application ontologies. We were guided by Zhao et al. (2008), who pioneered the work in the use of D2R and D2RQ for geospatial ontologies. Zhao et al. (*ibid*.) worked with only geospatial data; ours combined geospatial and nonspatial data. SPARQL and RDF were used to represent the semantics. D2R utilizes SPARQL and RDF.

Results

Figures 2 and 3 show the actual integration of the ontologies/schema. The figures illustrate the ULO and how it connected to the more specific ontologies and then to the database schema. The ULO was designed to operate at a high level of abstraction, in that it represents place as a feature, with properties such as feature type. To develop the ULO, we relied on existing ontology standards such as GeoOWL (for space) and Friend of a Friend (FOAF, for people).

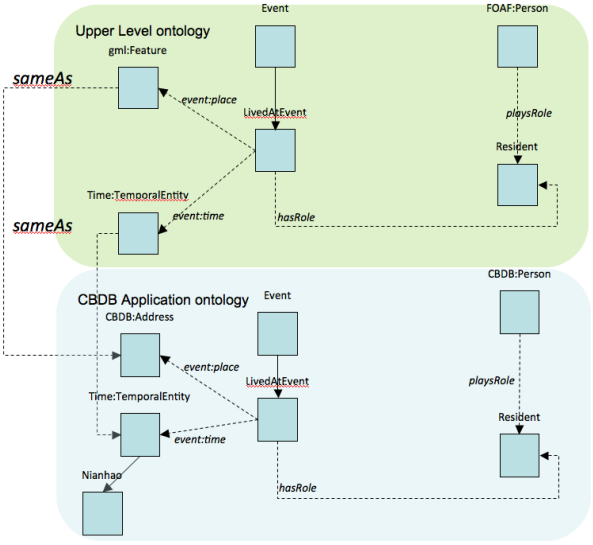

Figure 2. Upper level ontology/application ontology integration

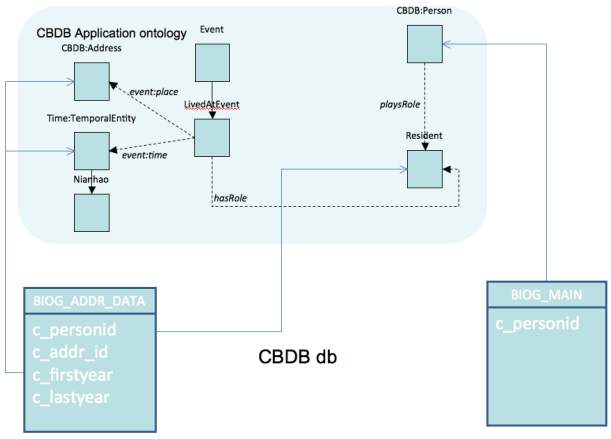

Figure 3. Application ontology/database schema integration

Figures 2 and 3 demonstrate that proposed methods do not necessarily translate into actuality. This was due to our choice of open source software. D2R was useful because it has a strong user community that could support developers from the social science and humanities. More importantly D2R handled the query brokerage (e.g., "where does famous woman poet x live?" requires a division of the questions into subquestions appropriate to each databases and a prioritization of the order in which the subquestions are asked of the databases). Query brokerage undergirds the functioning of the system architecture. Instead of being separate, the DO was collapsed into the ULO. It functions as a subset of the ULO. Like the ULO, the DO is shared across all databases. The two figures trace the integration of the abstract and generalized concept of geographic feature to greater specificity about how location are represented in the CBDB. What is not shown is the same abstraction of feature is linked to geographic locations in the CHGIS.

Using D2R required us to make use of D2RQ, which automates the translation of database schema into AOs. D2RQ generated separate files where D2R demanded only one database. The AO in Figure 3 represents the CBDB concept of "lived at". But each database had a different way to represent "lived at". So there was considerable manual modification to get the AOs to link to the ULO/DO.

Some final points on geography: We modeled geospatial data mostly as attributional—names and point features instead of features composed of multiple x, y's. This was dictated by the way the geospatial data was modeled in the three databases. Our relations were mostly mereotopoligical and places were disambiguated by dates, so we created relationships like preceeded_by. The geography in our case study is simple relative to other geospatial ontologies but did provide us with interesting challenges, nonetheless.

Conclusions

We were able to develop a proof-of-concept tiered ontology for our domain, which is Chinese historical biography and geography. We are left with numerous research questions for using ontologies for integrating databases, many of which are non-technical. How do we handle varying institutional issues, such as who maintains the ontology? How do we manage open/closed source systems? Lastly, how do we attend to the uneven institutional resources for computing in the social sciences and humanities? These questions and others provide a substantial domain for geospatial ontology research.

References

Bowers S, Lin K, and Ludascher B (2004) On Integrating Scientific Resources through Semantic Registration. Paper read at 16th International Conference on Scientific and Statistical Database Management (SSDBM'04).

Egenhofer M J (2002) Toward the geospatial semantic web. Proceedings of the 10th ACM international symposium on Advances in geographic information systems. McLean, Virginia, USA, 1-4.

Friedlander A et al. (2009) Working Together or Apart: Promoting the Next Generation of Digital Scholarship. Report of a Workshop Cosponsored by the Council on Library and Information Resources and The National Endowment for the Humanities. www.clir.org/pubs/reports/pub145/pub145.pdf

Gruber, T R (1993). A Translation Approach to Portable Ontologies Specifications. *Knowledge Acquisition* 5(2): 199-220.

Guarino N, Giaretta P (1995) in *Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing*, ed Mars N (IOS Press, Amsterdam), pp 25-32.

Wellen, C C (2008). Ontologies of Cree Hydrography: Formalization and Realization. Masters of Science Thesis. Department of Geography, McGill University, Montreal, Canada.

Kintigh K W (2006) The promise and challenge of archaeological data integration. *Am Antiq* 71(3): 567-578.

Zhao, T, Zhang C, Wei M and Peng Z R (2008) Ontology-based geospatial data query and integration. *Lecture Notes in Computer Science* 5266: 370-392.