

Extending Gazetteer Service in Geographic Information Retrieval

A Position Paper

Naicong Li, Jordan Henk

The Redlands Institute, University of Redlands

naicong_li@institute.redlands.edu, jordan_henk@institute.redlands.edu

Hill (2000) identifies three basic elements in a digital gazetteer: *name*, *footprint*, and *type* (or category). Gazetteer entry *types* are organized in a hierarchical classification system. The name and footprint properties of geographical features enable automatic association between non-spatial data (e.g. text documents) and the location of geographical features, whereas the feature type hierarchy provides automatic organization of the spatially indexed data. Used in a geographic information retrieval system (GIR, Larson 1996), such spatial indexing makes it possible to answer thematic (“what”) and locational (“where”) queries. The “theme” in the query may be spatial, referring to a geographic feature (e.g. the City of Redlands) or to a category of such features (cities). It can also be non-spatial, referring to some property of the feature (e.g. the city’s population). While the *name* and *footprint* properties of geographic features are formally defined in a gazetteer, the typical set of properties associated with a particular feature category often are not. Similarly, the typical set of relationships among geographic features (other than the *is-a* relationship, such as the *capital-of* relationship) are often not present in gazetteers. GIR systems could potentially benefit from explicit definitions of these properties and relations by using them to automatically index and later retrieve relevant data, especially in the case of highly structured data such as GIS datasets.

The Redlands Institute at the University of Redlands is conducting applied research in GIR using a prototype platform known as the Geospatially Referenced Information Portal (GRIP). GRIP was designed to facilitate knowledge synthesis across a disparate set of information resources (sources such as internet engines, local file systems, digital libraries, databases, etc. - all stored in heterogeneous formats, possibly with non-standardized structures). GRIP (which is being extended to support knowledge management tasks for a desert tortoise science and habitat management project) currently uses a project gazetteer with a feature type hierarchy modeled after the ADL Feature Type Thesaurus (FTT). This gazetteer has over 12,000 features extracted from recent versions of the USGS GNIS gazetteer (for the states of CA, NV, AZ, and UT); USGS Federal Land polygons augmented with boundaries for states, counties, state parks, etc.; and further supplemented with project-based GIS data. Using this gazetteer, GRIP supports basic spatial searches and basic queries (including theme and location).

Our users are often interested in the exact value or range of values of a certain property of a geographic feature (e.g. the actual tortoise population density of a critical habitat unit), or the statistics on these values, or retrieving a subset of geographic features of a certain type based on some specification of their property values. Sometimes the data value already exists in the project GIS dataset(s), sometimes GRIP needs to perform spatial and/or non-spatial processing to derive the answer. To handle such queries, we have adopted an approach similar to those described in Fonseca et al. 2002 and Lutz et al. 2006. We are prototyping using a light weight (not heavily axiomatized) project ontology for our user community. This ontology contains project-based concept categories (both spatial and non-spatial). For each concept category, we define a set of relevant properties (e.g. the ‘area’ of a ‘habitat’), including definitions of the property value type and value range. We also define its relationship (of types *is-a*, *partitive*, *causal*, *functional*, *associative*, etc.) to objects of other categories when applicable. Data in the project database are linked to appropriate categories as instances of these categories. Such links are established by a process similar to “registration mapping” as described in Owers and Ludäscher 2004, and in Lutz and Klien 2006. The categories in this ontology thus can be used as an index into the data in the project geodatabase. The categories can also be used to index document metadata, and they may be used to annotate text documents (through recognition of their corresponding natural language expressions in the documents) to produce sub-document level metadata for various levels of discourse units such as word, phrase, sentence, paragraph, etc. This ontology is designed to import pre-established upper level ontologies including an ontology of spatial relations with definitions of topological relations, distance and direction. It also incorporates definitions of frequently used spatial and non-spatial operations which can be invoked when GRIP is processing the more complex spatial search

(Larson 1996, Jones et al. 2004). Using the properties, relations and rules defined in this ontology, GRIP will be able to deduce implied information through inferencing and reasoning.

GRIP exposes this ontology to its user in one of its query formulation interfaces. One advantage of ontology-driven query formulation is that the category, instance and property in the ontology may be used as a set of controlled vocabulary terms in query formulation (thus by-passing the need of disambiguating user queries in natural language). Another advantage is that this approach facilitates user exploration (and therefore discovery) of the information stored in the project database(s). A novice user, unfamiliar with desert tortoise habitats, may browse the ontology and learn about relevant factors (such as climate conditions, vegetation cover, soil, threats, etc.) through the property definitions for the concept of “Desert Tortoise Habitat”. The user may also obtain a more comprehensive knowledge of threat factors by browsing all the sub categories of “Threat”.

Our research is intended to produce a generic platform in which a project ontology and a project gazetteer may be integrated to enhance geographic information retrieval. Because of its relatively simple structure and the use of an efficient indexing scheme, the gazetteer has a much faster performance time during a search. The gazetteer is also sufficient for answering the most common types of queries regarding theme and location. The ontology and its instance data are deployed (with slower performance) when the user drills down into a query to discover more detailed information about a category or an instance (knowledge not stored in the gazetteer).

References

Alexandria Digital Library project, <http://webclient.alexandria.ucsb.edu/>.

Bowers, S. and B. Ludäscher, 2004, An Ontology-Driven Framework for Data Transformation in Scientific Workflows. In *International Workshop on Data Integration in the Life Sciences (DILS'04)*, March 25-26, 2004, Leipzig, Germany.

Fonseca, F., M. Egenhofer, P. Agouris, and G. Camara. 2002. “Using Ontologies for Integrated Geographic Information Systems”, *Transactions in GIS*, 6(3), 2002.

Hill, L. L. 2000. Core elements of digital gazetteers: placenames, categories, and footprints. In J. Borbinha & T. Baker (Eds.), *Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000* (pp. 280-290). Berlin: Springer.

Hill, L. L. 2006. *Georeferencing*. The MIT Press. Cambridge, Massachusetts.

Jones, C., A. I. Abdelmoty, D. Finch, G. Fu, and S. Vaid. 2004. The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. In *Proceedings of the 3rd International Conference on Geographic Information Science*, pages 125–139.

Larson, Ray R. 1996. Geographic Information Retrieval and Spatial Browsing. In L. Cl, Smith and M. Gluck, eds., *Geographic Information systems and Libraries: Patrons, Maps, and Spatial Information*, 81-123. Urbana-Champaign: Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign.

Lutz, M. and E. Klien. 2006. Ontology-based Retrieval of Geographic Information, *International Journal of Geographical Information Science*.