CO-057

ADVANCES TOWARD EXPANDING GAZETTEER SEMANTICS

VARANKA D.
U.S. Geological Survey, ROLLA, UNITED STATES

BACKGROUND AND OBJECTIVES

Geographic queries commonly involve place names in the minds of map users. Gazetteers can facilitate geographic queries by forming the basis of interfaces to extract geographic information. Traditional gazetteers are formatted as strings of a feature name, coordinate location, classification type, and unique identifier. Before the digital transition of cartographic data, coordinates would be sought on an analogue map, and the feature would be located. In geographic information systems (GIS), names continue to function as labels for represented features, but potential functions of digital gazetteers are expanding. As digital interfaces, gazetteers can translate a name- or feature-based query to the feature identification and footprint in GIS, presenting the information as text or graphics, but such software is difficult for non-experts to use. GIS access the capabilities of the relational data model for geographic information extraction, but information models of different approaches, such as graph-based semantic technology, are rapidly spreading in use with geographic information. These models have implications pertaining to the semantics of information for users' experience.

The objective of this paper is to discuss advances in digital gazetteer capabilities of different information models from a user's point of view. Developments in gazetteer technology are reviewed from geographical literature. The motivation for this assessment is to design a gazetteer interface for *The National Map* of the U.S. Geological Survey (USGS); a collaborative effort built on partnerships and standards to improve and deliver U.S. topographic information at multiple scales and resolutions (USGS 2011). A summary of the literature assessment is followed by a discussion of gazetteer design implications for topographical data.

RECENT RESEARCH LITERATURE ABOUT GAZETTEERS

Data translation functions in traditional gazetteers, such as name-to-location (in answer to the question, where is?) or name-to-feature type (in answer to the question, what is?) are relatively easy because gazetteers simplify the representation of elements that are semantically complex (National Geospatial Intelligence Agency 2007). Recent research literature on gazetteers has presented ideas on the complexity of gazetteer components and the interoperability between different gazetteers (Goodchild and Hill 2008).

Complex geospatial features, such as those with multiple place names for a single location or whose feature type could be variably classified, require the semantic clarification of those elements. Relational databases may contain these various elements, but when decomposed from the table format, the data lose their context or meaning. For this reason, relational data or information is not sufficiently flexible to openly serve diverse users without manually translating the metadata. Gazetteer research has turned to semantic technology for potential solutions. The Resource Description Framework (RDF) standard specified by the World Wide Web Consortium (W3) decomposes data and information so that feature elements are machine-readable and can be flexibly interchanged across servers, platforms, and models. GIS/Relational DataBase Management System (RDBMS) models build on metric topological relations with location coordinates as attributes. RDF models build on networks, with triples whose location coordinates can be represented as Geography Markup Language (GML), a grammar defined to express geographical features and which was originally modeled on RDF (Cox and others 2004). The expected findings are that triples will more easily enhance queries than RDBMs do.

Existing approaches to gazetteer development seek to enhance diversity and completeness. The semantic specification of gazetteer components is seen to function best at local areas of interest, though integrated at an upper-level gazetteer system (Janowicz and Keßler, 2008). Tools for collecting vernacular place names, their feature types, and location can be on-line interfaces to databases or summarized from social networking media such as flickr.com tags (The English Project 2011; Rattenbury et al. 2007). Data interoperability between gazetteers benefit from a semantic framework. For example, varying feature type classifications can be resolved using feature type ontology.

Fully automated approaches are possible, though natural language itself is still beyond the common capabilities of automation. Standards of completeness and accuracy of gazetteers can be automatically derived by manipulating the Internet (Goldberg et al. 2009).

Spatial location has emerged as a powerful concept for linking information in libraries and record offices (Wilson et al. 2004; Hill and Zheng 1999). Locational aspects of gazetteers provide an approach to the study of historical and geographical change (Janowicz 2006).

In these approaches, spatial relations have not been leveraged, though an extensive body of research was developed in the 1990s for spatial feature queries in databases. Later research on complex feature databases provides further potential to applying topological queries to complex spatial objects (Schneider and Behr 2006).

A GAZETTEER DESIGN PLAN FOR TOPOGRAPHIC FEATURES

The National Map of the U.S. Geological Survey (USGS) is intended to serve a broad range of specialist and non-specialist users for an array of purposes by offering geospatial data organized by 7 geographic themes. The vector data model themes are hydrography, transportation, structures, boundaries; the raster data model themes are land cover, elevation, and ortho-imagery. The source for feature names is the Geographic Names Information System devised and maintained by the U.S. Board on Geographic Names (USBGN 2011). The names are embedded in thematic data layers of *The National Map* for each feature. Embedding names and features imposes constraints on the potential uses of names that text-formatted gazetteers can more flexibly accommodate, because feature data collection and representation are expensive operations to complete. Embedded names offer functional capabilities, demonstrated in this project by using query paths for topological spatial relations between geospatial features.

Eight standard spatial operators based on topological relation models from the 9-intersection model are recognized by the Open Geospatial Consortium (OGC) (Egenhofer and Herring 1991). The OGC relation terms are Equals, Disjoint, Intersects, Touches, Crosses, Within, Contains, and Overlaps, and can be used to query components of complex feature or features and their topographical context (Herring 2006). The exact terms can vary; research has shown that users mis-identify the topological relation term assigned to a particular mathematical relation (Riedemann 2005). In this research, queries are expressed using the GeoSPARQL proposed query language standard of the OGC that links SPARQL standard semantic query language and the RDF data model standard (W3C 2008). The reasoning software governing rules for the behavior of feature and their spatial relations are available in some commercially available packages.

The gazetteer queries to be enabled are to find topographical features or concepts modeled on topological spatial relations in *The National Map*. The queries resembling plain English should be capable of using objects related to the subjects of the triples. For example, a query such as ‘What river does the bridge cross?’ would identify the name of the line segment that crosses the transportation segment identified as a named bridge. The search is applied to the GIS data whose coordinates are associated with those of neighboring features through the use of the OGC standard topological relations. The matrix of relations that form the 9-intersection model, which forms the basis of the OGC standard, provides the range of relations that can exist between two features.

Landscape components for topographic concepts were specified in the semantics of vector data model for *The National Map*, called the Best Practices data model. The semantics of the Best Practices data model are based first on layer name and secondly on feature classes. For example, the Transportation thematic layer has feature classes within it of Road, Air, Rail, and Water. The Class Domains are treated as attributes for the general class and FCode coded value domain is considered a feature type. When integrated with the general topographic feature taxonomy, the resulting Transportation taxonomy has classes, subclasses, and feature types that reflect transportation processes more strongly than features found in the topographic feature taxonomy, creating a semantic shift between objects and processes.

Semantic infrastructure development for topographic data involves a wide array of new and established feature terms and a more recent need for a vocabulary of spatial relations for the subject – predicate – object format of triple data in Resource Description Format (RDF). Logical axioms specified by the ontology reasoning software govern the behavior of linkages of the nodes of the graph network in RDF. A new Open Geospatial Consortium (OGC) standard is in review for linking RDF and GML for topological spatial relation queries.

CONCLUSIONS

The application of spatial relations between topographic features or between aspects of complex features in semantic technology introduces new possibilities for enhanced gazetteer functions. A possible design leveraging spatial relations to enable the flexible geo-referencing between features and adaptation to differing contexts or circumstances is not presented in research literature. This is the approach to develop as a gazetteer interface for *The National Map* data.

REFERENCE LIST

Cox, S., P. Daisey, R. Lake, C. Portele, and A. Whiteside, (eds). 2004. OpenGIS® Geography Markup Language (GML) Implementation Specification. OGC 03-105r1 v. 3.2.1. Wayland, Mass.: Open Geospatial Consortium, Inc.

Egenhofer, M., and J. Herring. 1991. Categorizing binary topological relationships between regions, lines and points in geographic databases. In M. Egenhofer and J. Herring (eds), A Framework for the Definition of Topological Relationships and an Approach to Spatial Reasoning within this Framework, Santa Barbara, CA. 1-28.

Goldberg, D.W., J.P. Wilson, and C.A Knoblock. 2009. Extracting geographic features from the Internet to automatically build detailed regional gazetteers. International Journal of Geographic Information Science 23(1): 93-128.

Goodchild, M. and L.L. Hill. 2008. Introduction to digital gazetteer research. International Journal of Geographical Information Science 22(10): 1039 – 1044.

Herring, J.R. (ed), 2006. Open GIS implementation specification for geographic information – Simple feature access – Part 1: Common architecture. OGC 06-103r3. Wayland, Mass: Open Geospatial Consortium Inc.

Hill, L. L. 2000. Core elements of digital gazetteers: Placenames, categories, and footprints. In J.L. Borbinha, and T. Baker (eds), Proceedings of the Fourth European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000), Berlin, Springer. 280-290.

Hill, L.L. and Q. Zheng. 1999. Indirect geospatial referencing through place names in the digital library. : Alexandria digital library experience with developing and implementing gazetteers. In: Proceedings of the Sixty-Second Annual Meeting of the American Society for Information Science, Washington D.C. pp. 57-69.

Hill, L.L. 2006. Georeferencing: The Geographic Associations of Information (Digital Libraries and Electronic Publishing). Cambridge MA: MIT Press.

Janowicz, K. 2006. Towards a similarity-based identity assumption service for historical places. In M. Raubal, H.J. Miller, A.U. Frank and M.F. Goodchild (eds), Proceedings of the Geographic Information Science, Fourth International Conference GIScience 2006, Lecture Notes in Computer Science 4197, 20-23 September 2006, Munster, Germany. Berlin: Springer. 199-216.

Janowicz, K., and C. Keßler. 2008. The role of ontology in improving gazetteer interaction. International Journal of Geographical Information Science 22(10): 1129 – 1157.

The English Project. Location Lingo. The English Project @ Winchester Ltd. [http://www.englishproject.org]

National Geospatial-Intelligence Agency, 2007. Complicated Features Workshop Conference Report. Warrenton, Virginia.

Rattenbury, T., N. Good, and M. Naaman. 2007. Towards automatic extraction of event and place semantics from flickr tags. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY: ACM, pp. 103-110.

Riedemann, C. 2005. Matching names and definitions of topological operators. In: A.G. Cohn and D.M. Mark (eds) Proceedings of the Spatial Information Theory. Foundations of Geographic Information Science, International Conference, COSIT 2005, Lecture Notes in Computer Science 3693, Ellicottville, NY, USA. Berlin: Springer. pp. 165-181.

Schneider, M., and T. Behr. 2006. Topological Relationships Between Complex Spatial Objects. ACM Transactions on Database Systems 31(1): 39-81.

U.S. Board on Geographic Names. 2011. Geographic Names Information System (GNIS). U.S. Geological Survey. [http://geonames.usgs.gov/domestic/index.html]

U.S. Geological Survey. 2011. The National Map. U.S. Geological Survey. [http://nationalmap.gov/]

W3C. 2008. Semantic Web. World Wide Web Consortium. [http://www.w3.org/TR/rdf-sparql-query/]

Wilson, J.P., C.S. Lam, and D.A. Holmes-Wong. 2004. A New Method for the Specification of Geographic Footprints in Digital Gazetteers. Cartography and Geographic Information Science 31(4): 195-207.