Toward the Semantic Geospatial Web*

Max J. Egenhofer
National Center for Geographic Information and Analysis
Department of Spatial Information Science and Engineering
Department of Computer Science
University of Maine
Orono, ME 04469-5711, USA
http://www.spatial.maine.edu/~max
max@spatial.maine.edu

ABSTRACT
With the growth of the World Wide Web has come the insight that currently available methods for finding and using information on the web are often insufficient. In order to move the Web from a data repository to an information resource, a totally new way of organizing information is needed. The advent of the Semantic Web promises better retrieval methods by incorporating the data's semantics and exploiting the semantics during the search process. Such a development needs special attention from the geospatial perspective so that the particularities of geospatial meaning are captured appropriately. The creation the *Semantic Geospatial Web* needs the development multiple spatial and terminological ontologies, each with a formal semantics; the representation of those semantics such that they are available both to machines for processing and to people for understanding; and the processing of geospatial queries against these ontologies and the evaluation of the retrieval results based on the match between the semantics of the expressed information need and the available semantics of the information resources and search systems. This will lead to a new framework for geospatial information retrieval based on the semantics of spatial and terminological ontologies. By explicitly representing the role of semantics in different components of the information retrieval process (people, interfaces, search systems, and information resources), the Semantic Geospatial Web will enable users to retrieve more precisely the data they need, based on the semantics associated with these data.

* This work was partially supported by a grant from the National Imagery and Mapping Agency under grant number NMA401-02-1-2009. Max Egenhofer's work is further supported by the National Science Foundation under grants IIS-9970123, EPS-9983432, and IIS-9613646; the National Institute of Environmental Health Sciences, NIH, under grant number 1 R 01 ES09816-01; and the National Imagery and Mapping Agency under grant numbers NMA201-01-1-2003, NMA201-00-1-2009, and NMA202-97-1-1023.

GIS'02, November 8-9, 2002, McLean, Virginia, USA.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *abstracting methods, dictionaries, linguistic processing*. Information Search and Retrieval – *information filtering, query formulation, retrieval models, search process*.

General Terms
Design, Languages.

Keywords
Semantics, spatial information geographic information systems, semantic geospatial web, semantic spatial web.

1. INTRODUCTION
With the growth of the World Wide Web has come the insight that currently available methods for finding and using information on the Web are often insufficient. Today's retrieval methods are typically limited to keyword searches or matches of sub-strings, offering no support for any deeper structures that might lie hidden in the data or that people typically use to reason; therefore, users may often miss critical information when searching the Web. At the same time, the structure of the posted data is flat, which increases the difficulty of interpreting the data consistently. Higher-level computational operations that need to compare, query, analyze, combine, or integrate data cannot be carried out due to the lack of methods that make compatible information available.

There would exist a much higher potential for exploiting the Web if tools were available that better match human reasoning. In this vein, the research community has begun an effort to investigate foundations for the next stage of the Web, called the *Semantic Web* [1, 6]. Current efforts include the Extensible Markup Language XML, the Resource Description Framework RDF, Topic Maps, and the DARPA Agent Markup Language DAML+OIL. Much like in the late 1980s during the early days of object orientation [3], currently the different research approaches to the Semantic Web are ground breaking and experimental; often, however, they are incompatible and incoherent.

A rich domain that requires special attention is the semantics of geospatial information. The enormous variety of encodings of geospatial semantics makes it particularly challenging to process requests for geospatial information. Work in the area of GIS interoperability [8, 15] and the work led by the Open GIS Consortium addressed some basic issues, primarily related to the geometry of geospatial features. The Geography Markup

Language GML provides a syntactic approach to encoding geospatial information (opengis.net/gml/01-029/GML2.html) through a language in which symbols need to be interpreted by users, because associated behavior is not accounted for. While the Semantic Web approaches provide generic frameworks to describe ontologies and capture semantics, they do not relate explicitly to some of the most basic geospatial properties Web users require, such as the kinds of entities and relationships that are most useful for a particular information-processing task. Current approaches do not address the issue of how any particular semantics can actually be exploited by people who are searching the Web to accomplish a particular task.

2. *LAKES IN MAINE*

The lack of appropriate methods for geospatial information retrieval and processing reflects the way operators of large information resources currently operate. A Web user, for instance, who wants to find on the World Wide Web an appropriate data set for a specific task cannot use one of the current search engines to retrieve the desired data sets by simply expressing the request as:

*Find data sets that allow me to analyze **lakes in Maine**.*

Current search engines only examine Web content for relevant keywords (i.e., ***lakes***, ***Maine—in*** typically gets dropped because it is a frequently occurring term) or a fixed character sequence (i.e., ***lakes in Maine***), but would not be able to address the semantics of this request.

- What would happen if a data set references lakes by counties, and one needs another data set to link counties to Maine?
- Or if there is available a layer with the geometry of lakes, and another layer with the geometry of the States in the US, what is the semantics of ***in*** to perform the spatial join?
- Must a lake be completely inside Maine in order to qualify, or could it extend into the neighboring state, or country?
- Would an inventory with lakes ***inside*** Maine be appropriate?

This non-exhaustive list of possibilities demonstrates the variety of semantic issues that may need to be involved in finding the right data set for a Web user. The burden of performing a successful search is put almost entirely on the user, which may mean that important information may be missed in a query. We envision that the ***Semantic Geospatial Web*** will be capable of processing such requests with a varying degree of geospatial content such that users obtain automatically results that match their tasks. A key issue for the Semantic Geospatial Web is that it captures, analyzes, and tailors geospatial information, much beyond the purely lexical and syntactic level.

3. PRESENTATIONS OF GEOSPATIAL SEMANTICS

To capture meaning of information, we consider four presentations of geospatial semantics for the Web [5]:

- *Natural language with minimum markup* on the Web is perhaps still best exemplified by basic HTML or XHTML. The chief bearers of this presentation of semantics are people and documents. Interfaces and search systems do bear some natural-language semantics, but typically interfaces bear only small subsets of natural language (such as basic search instructions), and systems typically exploit only certain full-text indexing techniques on the natural language.
- *Simple metadata* on the Web most often occurs as specially designated tags describing access points of documents [14]. Such tags are commonly found in XML-based languages. Although users do need to interact with metadata semantics during information retrieval, metadata semantics is chiefly borne by the interface, the documents, and the search system.
- *Data models* endow documents or other Web resources with an identifiable conceptual structure. One currently popular data model used on the Web is the Resource Description Framework, RDF (www.w3.org/TR/1999/REC-rdf-syntax-19990222). The conceptual structure supplied by RDF is given in terms of entities, relationships, and attributes. The semantics of data models is chiefly borne by the interface, documents, and the search system. Users bear data model semantics only insofar as they need to express their query in terms acceptable to the data model. The advent of documents bearing data-model semantics is a relatively recent advance on the Web.
- *Logical (model-theoretic) semantics* provides a correspondence among terms and real-world entities, which allows for automated reasoning. As with metadata and data models, logical semantics is borne principally by interfaces, documents, and search systems. DAML+OIL (www.daml.org/2001/03/reference.html) is one currently popular way of making logical semantics available to search engines and users. The semantics of DAML+OIL are expressed by tags that can travel with the documents themselves. The challenge remains of how best to make this semantics available to the user via the search interface.

These presentations of geospatial semantics on the Web all recognize that people and machines will need to interact cooperatively in order to exploit the different meanings available to them; however, meanings are of limited use by themselves. For example, in the "lakes in Maine" request we expect that whatever meanings that do exist in the context of a Web search be exploited by the person who makes the request in order to carry out a particular information-processing task. Thus, in any consideration of geospatial semantics on the Web, there needs to be an explicit taking into account of the way in which particular geospatial tasks can be done. This is relatively easy to state, but difficult do.

Fortunately, we see a convergence between a fundamental principle of semantics on the Web and the task-based nature of spatial data querying. All the previously mentioned approaches to semantics on the Web attempt to represent meanings by certain primitive objects and the operations that can be performed on these objects to create more complex entities and relationships. There is a strongly parallel situation in geospatial data querying (or task-based geospatial information retrieval). Specifically, many tasks that web users want to perform with geospatial data consist of combining different kinds of primitive objects in meaningful ways.

4. THE SEMANTIC GEOSPATIAL WEB

The accomplishment of the Semantic Geospatial Web will require a much broader research agenda, but initially two research issues are apparent: (1) we need a plausible canonical form in which to pose geospatial data queries and (2) we need methods to assess the semantics of available data sources to see whether their semantic structure can be exploited for any particular geospatial query task. Our starting point for a meaningful retrieval of information on the Semantic Geospatial Web is a *geospatial request* that has clearly defined semantics associated with each element. A rough sketch of a geospatial request describes it as *geospatial constraints* of the form[1]:

<geospatial request> ::= <geospatial constraint>
[<logical connective> <geospatial request>]

where multiple geospatial constraints can be part of conjunctions or disjunctions using the logical connectives "and" or "or" (<logical connective> ::= "and" | "or"). A geospatial request resembles the WHERE part of a spatial SQL query [4], but will provide a much richer and more meaningful interpretation of the geospatial constraints than today's strict database queries do. Each geospatial constraint is made up of three parts—two geospatial terms that are linked by a geospatial comparator:

<geospatial constraint> ::= <geospatial term>
<geospatial comparator>
<geospatial term>

The semantics of the geospatial comparators is captured in a geospatial-relation ontology. Geospatial terms refer to either a geospatial class (such as the term "lake" in the initial example) or a geospatial label (such as the term "Maine" in the same example).

<geospatial comparator> ::= ! based on the geospatial-relation ontology used

<geospatial term> ::= <geospatial class> | <geospatial label>

<geospatial class> ::= ! based on a geospatial feature ontology

<geospatial label> ::= ! based on a geospatial gazetteer

A geospatial-relations ontology will include the treatment of synonyms (e.g., in, inside, within), the algebraic properties of the spatial relations (e.g., in (a, b) and in (b, c) $\rightarrow$ in (a, c)), and the mapping of spatial terms onto corresponding geometries (e.g., in (a, b) $\leftrightarrow$ interior (a) $\supset$ (boundary (b) $\cup$ interior (b)). In a similar way, one needs to develop methods that describe the semantics of geospatial feature classes and geospatial labels [7, 13]. The sources of these terms come from two different ontologies—geospatial feature ontology that captures the feature type definition [11, 12] and a geospatial gazetteer such as USGS's Geographic Names Information System, NIMA's GeoNet Names Server, or the Alexandria Digital Library [9].

[1] This is a very generic construct that generalizes to other kinds of information-retrieval requests (i.e., <request> ::= <constraint> [<logical connective> <request>]). The focus of this proposal, however, is on the geospatial domain.

5. BEYOND THE SEMANTIC GEOSPATIAL WEB

The Semantic Geospatial Web will be a significant advancement in the meaningful use of spatial information. With the flexible incorporation of geospatial semantics into the web, geospatial information retrieval will become precise to the level that the results of user queries will be immediately useful, without weeding out irrelevant hits. At that level, transactions over meaningful geospatial content will become possible. At the same time, however, the much broader relevance of spatial concepts and terms for general information retrieval and processing will become more apparent [2]. In particular, the spatial terms and their related spatial concepts are widely used metaphorically across different application domains. Therefore, the Semantic Geospatial Web will be only a stepping-stone toward the *Semantic Spatial Web*,

6. ACKNOWLEDGMENTS

Discussions with Jim Farrugia, David Mark, and Barry Smith helped me develop some of the concepts of the Semantic Geospatial Web. Participants at the NRC Workshop on the Intersection of Geospatial Information and Information Technology (http://www7.nationalacademies.org/cstb/project_geospatial.html) and the NSF Workshop on Database and Information Systems Research for Semantic Web and Enterprise (http://lsdis.cs.uga.edu/SemNSF/) provided useful feedback.

7. REFERENCES

[1] J. Berners-Lee, J. Hendler, and O. Lassila, The Semantic Web, *Scientific American*, vol. 184, no. 5, pp. 34-43, 2001.

[2] G. Cai, GeoVSM: An Integrated Retrieval Model for Geographic Information, in: M. Egenhofer and D. Mark, Eds., *Geographic Information Science—Second International Conference, GIScience 2002*, Boulder, CO, 2002, vol. 2489, *Lecture Notes in Computer Science*, Springer, pp. 70-85.

[3] K. Dittrich, Object-Oriented Database Systems: The Notation and The Issues, in: K. Dittrich and U. Dayal, Eds., *International Workshop in Object-Oriented Database Systems, Pacific Grove, CA*, Washington, D.C., 1986. IEEE Computer Society Press, pp. 2-4.

[4] M. Egenhofer, Spatial SQL: A Query and Presentation Language, *IEEE Transactions on Knowledge and Data Engineering*, vol. 6, no. 1, pp. 86-95, 1994.

[5] J. Farrugia and M. Egenhofer, Presentations and Bearers of Semantics on the Web, in: *Special Track on Semantic Web at FLAIRS 2002*, Pensacola, FL, 2002, pp. 408-412.

[6] D. Fensel and M. Musen, The Semantic Web: A New Brain for Humanity, *IEEE Intelligent Systems*, vol. 16, no. 1, pp. 24-25, 2001.

[7] F. Fonseca, M. Egenhofer, P. Agouris, and G. Câmara, Using Ontologies for Integrated Geographic Information Systems, *Transactions in GIS*, vol. 6, no. 3, pp. 231-257, 2002.

[8] M. Goodchild, M. Egenhofer, R. Fegeas, and C. Kottman Eds. *Interoperating Geographic Information Systems*. Kluwer, 1999.

[9] J. Hastings and L. Hill, Treeatment of "Duplicates" in the Alexandria Digital Library Gazetteer, in: M. Egenhofer and D. Mark, Eds., *GIScience 2002—Collection of Extended Abstracts and Poster Summaries*, Boulder, CO, 2002, pp. 64-66.

[10] J. Hendler, Agents and the Semantic Web, *Intelligent Systems and their Applications*, vol. 16, no. 2, pp. 30-37, 2001.

[11] A. Rodríguez, M. Egenhofer, and R. Rugg, Assessing Semantic Similarities Among Geospatial Feature Class Definitions, in: A. Vckovski, K. Brassel, and H.-J. Schek, Eds., *Interoperating Geographic Information Systems—Second International Conference, INTEROP'99, Zurich, Switzerland*, vol. 1580, *Lecture Notes in Computer Science*, Springer-Verlag, 1999, pp. 189-202.

[12] R. Rugg, M. Egenhofer, and W. Kuhn, Formalizing Behavior of Geographic Feature Types, *Geographical Systems*, vol. 4, no. 2, pp. 159-179, 1997.

[13] B. Smith and D. Mark, Geographical Categories: An Ontological Investigation, *International Journal of Geographical Information Science*, vol. 15, no. 7, pp. 591-612, 2001.

[14] A. Taylor, *The Organization of Information*. Englewood, CO: Libraries Unlimited, Inc., 1999.

[15] A. Vckovski, K. Brassel, and H.-J. Schek Eds. *Interoperating Geographic Information Systems—Second International Conference, INTEROP'99, Zurich, Switzerland*, *Lecture Notes in Computer Science* vol. 1580. Berlin: Springer-Verlag, 1999.