A Geospatial Semantic Web

Jonathan W. Lowe

A semantically intelligent World Wide Web is quickly becoming reality. And the potential it holds for the geospatial practitioner may be this industry's "next big thing."

For the past decade, a growing group of academic and industry technologists has been laying the groundwork for an artificially intelligent World Wide Web. Inspired by Tim Berners-Lee, inventor of the Web and current director of the World Wide Web Consortium (W3C), this group of individuals has established methods for representing not just words, but the meanings of words, in a format that machines can understand. Berners-Lee and his colleagues in numerous industries envision a future in which reasoning engines and Web-crawling agents inductively respond to questions such as "Is Route 66 in Arizona?" rather than simply returning Web pages that contain text matches for "Route 66" and "Arizona." Although today's Web may hold the raw data to answer such a question, only a human being is able to cobble together the answer by knowing that "Route 66" is a road, "Arizona" is a state, "in" is a geographic relationship, and all three mean something when strung together in a sentence. That same human must also know how to find and read a map. Without interpretive guidance from a human brain, how could a computer program possibly make these leaps alone?

Enabling a machine to understand human language requires additions to the way we represent words (or more generically, data) on the Web. The task of assigning machine-recognizable meaning to words includes identifying classes (for instance, a county class), taxonomic hierarchies (for example, counties subdivide states, which subdivide nations), and relationships (a county is a political boundary, or counties contain cities). Formalizing these categorizations into standard structures establishes ontologies. Standards have emerged for both simple ontologies — the Resource Description Framework (RDF) — and for more complex ontologies — the Ontology Web Language (OWL). Combining these ontologies with existing hypertext markup language (HTML) and extensible markup language (XML) transforms the Web into what Berners-Lee calls "The Semantic Web," forming the basis for artificial intelligence, machine interpretation of "natural language," more relevant query results, and several other advances. On the Semantic Web, for instance, computer programs called agents can tell the difference between similar words such as "in," "inside," and "intersects," or can deduce whether a land parcel is owned by a person or a company, and process each case differently.

Though the Semantic Web may initially sound like an over-ambitious pipe dream, most major industries already have efforts underway to establish their own ontological standards. Noteworthy examples are RosettaNet (electronics), the Open Travel Alliance (travel), STAR (automotive), and OpenCyc (the world!), though there are dozens more. Represented by the Open Geospatial Consortium (OGC), the geospatial industry has recently launched an interoperability experiment based on the Semantic Web concept (see sidebar titled "OGC Specifies Spatial Semantics" on page 35), and some academic groups have already published geospatial ontologies. This column explains the vision that is guiding the Semantic Web, the implementation that is empowering it, and its implications for geospatial practitioners.

The Vision

The vision behind the Semantic Web is far-reaching and has broad application potential, but begins, by design, within many tightly focused, conceptually specific study areas, one of which is geospatial. Interest in semantics for geospatial practitioners seems to focus on improving the relevance of Web-search results and the interoperability of spatial data, as demonstrated by the University Consortium for Geographic Information Science (UCGIS) "challenge" for further research of the geospatial Semantic Web, with authors Frederico Fonseca and Amit Sheth driving the initiative. Longtime geospatial academic Max J. Egenhofer is also leading investigations into the use of semantics to "enable users to retrieve more precisely the data they need, based on the semantics associated with these data." In a paper titled "Toward the Semantic Geospatial Web," Egenhofer illustrates the need for semantics when

Net Results columnist **Jonathan W. Lowe** covers the role of emerging technologies in the exchange of spatial information. Lowe is the owner of Local Knowledge Consulting (Berkeley, California), where he designs and implements spatial Web sites. He can be contacted at info@giswebsite.com.

searching the Web for datasets supporting the analysis of "lakes in Maine" (see the sidebar titled "Egenhofer's 'Toward the Semantic Geospatial Web.'")

A more semantic Web would contain a variety of metadata about all of a site's important words and their meanings, giving a semantically aware search engine a better chance of returning relevant results to Egenhofer's query.

Of obvious benefit to the geospatial industry, a Semantic Web offers something not previously available on a grand scale: interoperability that joins not only tightly structured professional spatial data such as that in online spatial databases, but also the unstructured, informal geographic information sprinkled throughout many Web pages but not originally intended by its authors for geospatial processing. For instance, a user might ask of the Semantic Web, "Find other Web sites with places close to the places mentioned in this Web site." In a semantic implementation, place names would be identifiable by a machine as geographic references, alerting a search engine to the difference between, for instance, "Whitman" the public high school and "Whitman" the poet. Whitman High School has a place, whereas Walt Whitman does not (other than for an application involving cemetery mapping, perhaps).

The Semantic Web is one of those ideas that, similar to the Web Services concept, is much bigger than just the geospatial industry. Some geospatial pundits describe a future in which geospatial information permeates our lives so completely that it becomes unremarkable. When Berners-Lee introduced the Semantic Web concept at the XML 2000 conference, he opened with an imaginary example of an automated "agent" searching through semantic material to book an appointment with the most conveniently located doctor within the user's medical plan. The agent used geospatial semantic information to limit the list of possible doctors to those with offices near the user's home.

Egenhofer's "Toward the Semantic Geospatial Web"

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

- What would happen if a dataset references lakes by counties, and one needs another dataset 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 U.S., 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 dataset 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.

Because semantics work at the level of natural language rather than information-technology system and database-schema jargon, the geospatial element of the example was not unusual or out of context. Semantics offer at least the theoretical opportunity for GIS to melt into a larger information technology framework.

The Implementation

How does such an ambitious agenda as Berners-Lee's take shape at the technical level? As it turns out, the mentally challenging part is in grasping the meaning of "meaning," not implementing it. After all, what constitutes meaning?

Most parents remember vividly their child's first words, often involving a person or thing ("mommy," "kitty," "ball") or an event ("eat," "bye"). Though not typical, it's conceivable that an infant's first spoken word is pure verbal imitation — he's just copying the sound "ball" without realizing that it names a bouncy, round toy. An infant who only mimics word sounds is the human equivalent of today's Web search engines. A search engine scans the Web for any text that users provide and returns links to pages with matches, but it cannot know that the text, "ball," refers to a beach toy on one page, a formal dance on another, and a portion of the human foot on a third. (Nearby text on the same page may provide a flimsy source of context, but hardly a reliable or fully predictable one.) Since it doesn't know the meaning of "ball," it can't process that word differently depending on context, so a search for "ball" brings back Web pages with all three kinds of balls, leaving the interpretation to the human searcher.

> Semantics offer at least the theoretical opportunity for GIS to melt into a larger information technology framework.

Fortunately, babies learn fast. Even if an infant's first words are only repeated sounds, he soon connects those sounds to the objects or events they name, screaming "Kitty!" and pointing excitedly when the cat prowls into the nursery. When a child can repeatedly identify the

```
@prefix m:   <http://www.example.org/meeting_organization#> .
@prefix g:   <http://www.another.example.org/geographical#> .

<http://meetings.example.com/cal#m1>
     m:Location [ g:zip "02139"; g:lat "14.124425"; g:long "14.245" ];
     m:chair    <http://www.example.org/people#fred> .
```

Figure 1. A Notation3 description of the location and chairperson of a meeting, also referencing a calendar URI (from "Examples — Getting into RDF & Semantic Web using N3," Tim Berners-Lee, www.w3.org/2000/10/swap/Examples.html).

same object, such as a cat, by the same name "kitty," he proves that he knows what the word means. Though he may not even be able to walk, he has already surpassed the intelligence inherent in today's Web and unwittingly begun a lifelong relationship with semantics — the study of meaning in language.

The first step in the development of the Semantic Web takes a cue from babies by linking every word to one or more unique uniform resource identifiers (URIs) and tagging words on Web pages accordingly. For instance, semantic developers might establish three different URI links for the word "ball." Semantic Web authors referring to toy balls on their pages would then tag the text "ball," with the toy ball URI. So, the first step toward a smarter Web is to link individual elements of our language to the Web's core identifier, a URI, which makes it accessible to all Web users as a standard point of reference.

Perhaps less dramatic than her first word but still worthy of family news headlines is the moment when a baby strings together related words to utter her first real sentence ("Kitty eat mommy?"), proving she recognizes a relationship between objects in the world as well as a corresponding relationship between the words representing those objects. Objects don't exist in isolation; they interact as members within a world of events, rules, and hierarchies of meaning.

Of course, babies may make taxonomic mistakes — kitties don't eat mommies, for example — but, right or wrong, her first sentence proves that our baby is beginning to build a mental model of how the world works and a basis for layer upon layer of logical reasoning. To enable automated machine reasoning, Semantic Web developers again seem to have taken their cues from infants; the next step in enabling an artificially intelligent Web is to build relationship "triples" with the standard URIs described earlier. These triples capture hierarchies and relationships, such as the assertions "cats are mammals" or "mammals are warm-blooded." Imagine one semantic developer providing these first two triples, while on a different site, another provides "dogs are mammals" and tags the same "mammals" URI. Though distributed across the Web, these triples now provide a knowledgebase of relationships between words such that an inference engine can infer that "if a dog is a mammal, and mammals are warm-blooded, then dogs are warm-blooded," even though the last triple may not exist explicitly on the Web.

There are several languages or syntaxes for annotating triples and posting them to the Web, RDF (an XML serialization) and Notation3 (N3, a plain text approach) being two of the most common options. The allure of RDF and N3 lies in their simplicity, or in what Berners-Lee calls the "principle of least power." Just as HTML's loose-but-simple specification presented a very low entry barrier to curious developers (and so allowed for rapid expansion of the World Wide Web), RDF and N3 are also designed simply for relatively easy learning and adoption (see Figure 1).

Today's Web browsers still read HTML, though more sophisticated markup approaches such as extensible hypertext markup language (XHTML) in combination with cascading style sheets (CSS) continue to gain prominence. Likewise, with the use of RDF and N3, their simplicity benefits rapid expansion but is also a limitation to more sophisticated reasoning. For instance, RDF and N3 do not support the logical property of negation. In other words, there is no way to represent the fact that "kitties don't eat mommies" using RDF or N3. Consequently, other semantic languages such as OWL and DAML,

OGC Specifies Spatial Semantics

In April, the Open Geospatial Consortium (OGC) Review Board approved the launch of a Geospatial Semantic Web (GSW) Interoperability Experiment to move toward the development of a Geospatial Semantic Web that would enable the discovery, query, and consumption of geospatial content based on formal semantic specification.

The GSW experiment aims to augment existing Web-focused mapping specifications (such as OGC's Web Feature Service and Filter Encoding specifications) with a semantic query capability. This would be accomplished by defining an ontology for the geospatial community. The experiment will also explore an appropriate distributed architecture to support specific use scenarios. For information on participating in this interoperability experiment, contact George Percivall at gpercivall@opengeospatial.org.

capable of denoting sophisticated logical structures (inverses, unambiguous and unique properties, restrictions, cardinalities, disjoint lists, and so on) provide a more sophisticated alternative to the simpler RDF or N3 approach. For instance, unlike RDF or N3, OWL allows a semantic developer to say that one property is the inverse of another. "Water surrounds islands" is the inverse of "islands are surrounded by water." This enables an inference engine to make more accurate deductions about other triples, resulting in closer matches to searches involving islands and water than would be possible with RDF or N3 ontologies.

Initially confusing, this alphabet soup of ontologic syntaxes really boils down to the same essential concept: that by establishing a URI for any term, the Web acquires mutually comprehensible references for individual elements of human language — people, places, things, actions. And stringing those URIs together as subject-verb-object triplets captures our human understanding of how the world works, whether expressed in N3 or RDF or DAML (DARPA [Defense Advanced Research Projects Agency] Agent Markup Language) or OWL or any other formal syntax. Deployed on the Web, these triples allow computer programs to manipulate the content of Web pages sensibly, with "intelligence." Any collection of URI triplets comprehensive enough to represent either a broad or narrow domain of experience, such as medicine or wines or geometry, is called an ontology. For a glimpse of existing ontologies, visit any of the links at Stanford University's Ontology Library (http://protege.stanford.edu/ontologies/ontologies.html), the DAML Ontology Library (www.daml.org/ontologies/), or SchemaWeb (www.schemaweb.info), all of which include geographic information ontologies.

What you find may leave you wondering, "What next?" After all, ontologies don't do much on their own. To derive any benefit from the Semantic Web, individuals must create agents and inference engines — programs that ingest ontologies in order to solve problems such as Egenhofer's "Lakes in Maine" search, or the UCGIS "naïve" spatial data search. One semantics enthusiast, Geoff Chappell, has converted the free TIGER/line census and streets dataset from shapefile format into RDF syntax and designed queries to interrogate the RDF (http://labs.intellidimension.com/tiger/).

For a non-geospatial example of a user interface against an ontology, visit Stanford University's Knowledge Systems Lab "Wine Agent," which recommends wines based on the composition of the food in the meal (or vice-versa) based on wine and food OWL ontologies (http://onto.stanford.edu:8080/wino/index.jsp).

> The real experts in meaning within any given industry are not the technologists, but the practitioners. A hydrologist surely knows better about the meaning of hydrological jargon than does a geospatial hydrology application designer.

Business Implications

In a March 2003 Harvard Business School Press article titled "Semantic Applications, or Revenge of the Librarians," David Moschella shares his thoughts about the implications of the Semantic Web. "The real future of technology," he says, "rests on the concept of standardizing human language, not computer programs. And that puts the customer in total control." He sees the Semantic Web shifting the information technology industry away from its current supplier-driven model to a customer-driven model. After all, the real experts in meaning within any given industry are not the technologists, but the practitioners. A hydrologist surely knows better about the meaning of hydrological jargon than does a geospatial hydrology application designer. Moschella also points out that standard customer-created metadata (ontologies) will not be a major competitive battleground, but "the effective use of these standards can become important areas of competitive advantage."

Worth Following

There's not enough room in one column for more than a quick dip into the ocean of Semantic Web concepts, and the territory has only begun to be navigated by our industry. Of interest to all industries, researchers and developers are also investigating inference engines (how do users query the Semantic Web?), issues of trust (is this ontology reliable?), equivalence (is the English "in" equivalent to the French "en"?), metadata proofs (what inferences lead to this answer, and are the supporting ontologies logically sound?), and connections between semantics and the more familiar structured data of files and databases (can an ontology be generated automatically from this spatial database?). Clearly, some high-profile players are actively involved, but interestingly, access to the work is in some cases by membership only. With proven staying power, a leader, and broad support — but not yet a household word — the Semantic Web has the earmarks of a "next big thing." ⊕