

Presentations and Bearers of Semantics on the Web

James Farrugia and Max J. Egenhofer

National Center for Geographic Information and Analysis
Department of Spatial Information Science and Engineering
University of Maine, Orono, ME 04469-5711, USA
{jim,max}@spatial.maine.edu

Abstract

We use an example from information retrieval to explore Web-based semantics, which can be presented in different ways and carried by different bearers of the information retrieval process. We describe four presentations of semantics: natural language with minimal markup, simple metadata, basic data models, and logical semantics. Each presentation of semantics can be borne by human users, search interfaces, documents, or search systems. We show that different presentations of semantics vary across a spectrum of expressiveness and arbitrariness of meanings that are shared and processed during information retrieval.

Introduction

The Semantic Web deals with meanings that humans can understand and that machines can process (Berners-Lee *et al.* 2001; Fensel and Musen 2001; Hendler 2001). We explore semantics on the Web by focusing on how people and computers can use semantics to enhance Web-based information retrieval. People conducting searches on the Web exploit semantics to the extent that meaningful clues and sensible processing are incorporated into the interfaces, documents, and search systems they use. For instance, people can choose an interface that allows them to specify meaningful access points of the documents they wish to retrieve. Computers exploit semantics, and thus help people find information on the Web, by representing certain kinds of useful relationships, such as the IS-A relationship in “a hex-head bolt is a type of machine bolt” (Berners-Lee *et al.* 2001). Once a computer represents this particular relationship, it can then draw particular conclusions about hex-head bolts based on their relationship to machine bolts.

We can see how people and computers work together in exploiting semantics for information retrieval by considering how meanings are presented and how they are carried by different parts of the information retrieval process.

We consider four presentations of semantics on the Web: (1) natural language, (2) simple metadata, (3) basic data models, and (4) logical (model-theoretic) semantics. These four presentations cover the major ways that semantics is currently presented on the Web.

In focusing on information retrieval, it is also useful to consider “what it is that has the semantics” (Uschold 2001).

Uschold notes that any of the following may have the semantics: terms or expressions referring to the subject matter of the Web content; terms or expressions in an agent communication language; and a language for representing the above information. In this paper, we focus on different bearers of semantics, which are more relevant to information retrieval. Specifically, we consider: (1) the human user, (2) the search interface, (3) the documents being searched, and (4) the system doing the searching.

The chief benefit of considering different presentations and bearers of semantics is that we can understand semantics on the Web as a spectrum that varies according to expressiveness and arbitrariness, ranging from meanings that exist only in people’s heads to meanings that are represented according to agreed-upon conventions, thus permitting these meanings to be shared. Our conception of a spectrum of semantics is related to the continuum of ontologies described in McGuinness (Forthcoming). The chief difference between the two is that in the ontology continuum, ontologies are organized according to the “spectrum of detail in their specification” (McGuinness Forthcoming), whereas in the semantic spectrum the presentation of meanings varies according to the expressiveness and arbitrariness of the presentation. We show the relevance of this semantic spectrum to the success of the information retrieval process.

Consider the following query and sample Web page.

Find one or more Web documents that describe a moving company that can move a family from Chapel Hill, North Carolina to San Francisco, California.

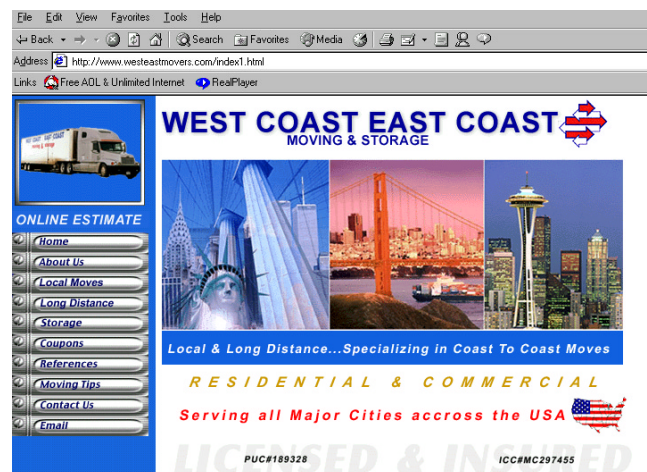


Figure 1 A sample query result.

We use this sample query and result to explain how different presentations and bearers of semantics combine to help users find Web documents that satisfy their queries.

The remainder of this paper is structured as follows. First we give an overview of the four presentations of semantics on the Web and explain their relationships to the different bearers of semantics in information retrieval. Next we discuss semantic issues for our example. Then we use our example to explain the role and limitations of the four presentations of semantics. Finally, we provide conclusions and directions for future study.

Four Presentations of Semantics on the Web

Although people generally agree that semantics is the study of meaning, different researchers approach this study from different perspectives (Partee 1999). Two of these perspectives concern us here. One of them considers meaning as part of the internal language of people's minds. A representative of this perspective is conceptual semantics (Jackendoff 1990). The second perspective, called truth-conditional semantics, considers semantics as a set of relations between language and the world (Larson 1995). The spectrum of semantics stretches between these two poles. At one end, natural language semantics seeks to provide an account of meanings inside people's heads (the *I-semantics* of Jackendoff). At the other end, logical (model-theoretic) semantics provides a formal specification of certain meanings by making explicit connections between particular symbols and the entities to which the symbols refer.

We consider four presentations of semantics widely used on the Web:

- *Natural language with minimum markup* on the Web is perhaps still best exemplified by basic HTML or XHTML (W3C HTML working group 2001). 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 exploit only certain full-text indexing techniques on the natural language.
- *Simple metadata* on the Web often occur as specially designated tags describing access points of documents (Taylor 1999). 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 (Lassila and Swick 1999). 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 (van Harmelen *et al.* 2001) is one currently popular way of making logical semantics available to search engines and users. The semantics of DAML+OIL is 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.

The preceding categorization of semantics on the Web is not the only way such a categorization can be made, but we use it here with advantage to illustrate a spectrum of semantics for information retrieval. Natural language semantics lies on one end of this spectrum, being the presentation of semantics most closely resembling the perspective of conceptual semantics. Logical (model-theoretic) semantics lies on the other end of this spectrum, representing the truth-conditional perspective of semantics.

Exploiting Semantics in Information Retrieval

Consider the retrieval result from Figure 1. In order to retrieve this result, we would like people and computers to be able to exploit meaningful information such as the following:

- Moving companies typically list themselves on the Web according to the origins or destinations they serve;
- Chapel Hill is part of the Raleigh-Durham area of North Carolina;
- The Raleigh-Durham area is on the East Coast of the US;
- San Francisco is part of the Bay Area of California;
- San Francisco is on the West Coast of the US;
- San Francisco is a major US city.

This meaningful information deals with three different domain-specific semantics. First, the information related to moving companies pertains to the semantics of what we might call the *moving domain*. Second, the information about cities, their respective metropolitan areas, and their locations on particular coasts belongs to the *specific geographic domain of the United States*. Third, the information that relates cities to their larger metropolitan areas (via *part-of* relations) belongs to a *general geographic domain* that deals with part-whole relations among geographic entities. All three domains contain meanings that can be used to improve the precision of retrieval results. We next investigate how the four

presentations of semantics can help users exploit meanings from these various semantic domains.

Roles and Limitations of the Semantic Presentations

Natural Language

Natural-language semantics is borne primarily by the human users in their efforts to match their information needs with the semantics made available through the interface, the documents, and the system. Even though users may be able to *express* their information needs in natural language, rarely will they enter a full natural-language query to a search interface. Rather, users typically reduce natural-language meanings to a list of terms that they expect will be found in the desired documents. Such a truncation from the richness of natural language to a list of terms is, in fact, a move away from semantics and towards syntax.

For the example query, a simple term search on <http://www.google.com> might be “moving companies chapel hill san francisco.” This query does not retrieve (within the first twenty results) a link to the Web page in Figure 1. The reason is that the query is a simple word list, the syntax of which carries no special meanings that the search system can exploit. Further, the search system does not store domain semantics and so cannot make the necessary mappings (e.g., from *San Francisco* to *West Coast* city) that would allow the page in Figure 1 to be retrieved.

Since most Web search interfaces and systems do not offer access to the semantics of different domains, and since documents are rarely encoded to represent domain-specific semantics, users must essentially guess about the nature of the term-matching done by the system if they are to achieve satisfactory results with natural language semantics they must. Thus, the retrieval results depend more on human ingenuity and raw syntax than on the semantic capabilities of the interfaces and systems. This problem can be alleviated somewhat by the incorporation of metadata semantics into the interface, documents, and search system.

Natural-language semantics, akin to the I-semantics of Jackendoff, is both the most expressive presentation of semantics on the Web and the most arbitrary. It is the most expressive because any thought that can be formulated can be expressed in (some) natural language. It is the most arbitrary, because it depends on individual mastery and application. It poses the greatest problems for using machines to help share mutually understood meanings.

Metadata

Since the primary bearers of metadata semantics are the interface, the documents, and the search system, users need access to this semantics if they are to exploit it for

information retrieval. This access is typically provided either directly through the interface, or indirectly through the documents and search system. So for instance, if the interface offers the ability to search a *City* access point, and if the search system tags certain US cities as *East Coast* or *West Coast*, then by searching on *Chapel Hill* or *San Francisco* the user has the possibility of retrieving documents like the one in Figure 1. Of course, to make the retrieval results more precise, the interface and system will need to include more such tags, as well as mappings that cover the moving and geographic domains.

Metadata on the Web, at least outside of controlled bibliographic environments, are still largely *ad-hoc*, not recorded in agreed-upon ways that would allow the metadata to be shared. This situation is improving somewhat by the use of shared metadata frameworks such as the Dublin Core (Dublin Core Metadata Initiative 1999), but the meanings that can be expressed within the Dublin Core are still essentially only keyword based. Metadata semantics is limited in that it typically does not supply explicit conceptual models of Web resources, nor does it offer a well-defined semantics, which is necessary for inferences to be carried out automatically. The next two presentations of semantics address these issues.

Metadata semantics, in its most common form, lies one step away from the arbitrary meanings that are expressible by natural language. Metadata semantics typically uses *ad-hoc* or agreed-upon recorded subsets of natural language as tags that offer access to the documents of interest. Users have to learn to recognize the available access points, and interfaces can help by making these access points explicitly available. The meanings expressed by metadata semantics, although written down, are not automatically shareable unless steps are taken to make them so. Further, metadata semantics by themselves do not permit searchers to exploit the meanings of a conceptual structure that may exist for a given subject domain related to the documents of interest. Finally, metadata semantics by itself does not allow automated inferences to be made.

Data Models

A data model of Web resources allows searchers to take advantage of a codified and potentially shareable conceptual structure. If users have access to the data model used to structure the data, they can employ a search interface and an appropriate query language to mine the relationships available from the data model. If users do not have access to the data model, an intelligent search engine could map user query terms to terms appropriate for querying the data model. In this case, the system would allow users to exploit a shareable data model to improve the precision of their retrieval results.

For the moving-company example, we might have the following set of RDF-like object-attribute-value triples:

- (moving company, serving_east_coast, west_coast_east_coast_moving_and_storage)

- (moving_companies, destination_city, san francisco)
- (city, lying_on_the_east-coast, Raleigh-Durham_area)
- (Chapel_Hill, is_part_of, Raleigh-Durham_area)
- (city, lying_on_the_west-coast, San Francisco)

If the search system has access to such a set of triples from a data model that deals with moving domains and geographic domains, then the system could take advantage of the semantics of these domains to retrieve documents, such as the Web page in Figure 1, that are deemed to satisfy these relationships.

Data models, by making explicit the relationships among entities, as well as the values of entity attributes, create a conceptual structure that can be exploited in information retrieval. To be exploited, this conceptual structure needs to be made available, at least to the search system. If the search system is the only bearer of this semantics, it needs to process the documents in advance according to the data model and then map the user's query to the relevant parts of the data model, retrieving for the user the documents indicated by the model. It would be better, of course, if the data model can be made explicit to the user through the interface.

Even though explicit data models allow users access to a conceptual structure (a shareable conceptual structure in the case of RDF), simple data models by themselves will not allow search systems to make inferences that are important to meaningful information retrieval.

Logical (Model-Theoretic) Semantics

In the moving-company example, logical semantics could be used to formalize the meaning of relations such as the *part_of* relation for geographic regions. One part of this formalism might include a logical axiom that relates the *part_of* relation for geographic regions with the *on* relation for geographic regions, such as "For X, Y, and Z geographic regions, if X is a *part_of* Y and Y is *on* Z, then X is *on* Z." (We use this axiom only for illustration and do not necessarily accept it.)

A searcher could take advantage of this logical semantics by entering only *Chapel Hill* into the interface. The system could then, using a sufficient fact base, infer that Chapel Hill is on the East Coast. Such a logical semantics for the geographic domain could be combined with an appropriate semantics for the moving domain to allow the user to exploit automatic inference and so retrieve the document shown in Figure 1.

Logical (model-theoretic) semantics goes one step further than data models by specifying the correspondence between language primitives and entities in a domain of discourse, as well as an interpretation that assigns truth values to legitimate expressions in the language based on the correspondence established between the primitive terms and the entities in the domain of discourse. Logical semantics restricts the expressiveness and arbitrariness found in the other presentations of semantics by specifying

precisely the relationships between terms and the world. Because any specification of logical semantics forms part of a logical system (Gabbay 1998), logical semantics allows automated inferences to be made.

Conclusions and Future Work

The role of the Semantic Web in information retrieval is to control the presentations of meanings carried by different bearers of semantics, so that meanings can be shared via agreed-upon conventions, resulting in greater precision of retrieval results. The current effective use of semantics for Web-based information retrieval still depends largely on the ingenuity of the individual human user to guess how his or her information need should be mapped to interfaces, documents, and search systems. The searcher is thus reduced to using rudimentary syntactic means to try to retrieve relevant documents. The searcher's need to guess reveals a lack of effective semantic alternatives.

With a geographic search example we demonstrated the need for different domains of semantics to retrieve a result that is acceptable to the user's intention:

- A task ontology (i.e., moving) is needed to identify the critical object classes (i.e., cities, transportation companies, origins, destinations, time, price) and their relationships.
- An upper-level geographic ontology captures fundamental large-scale spatial concepts (e.g., part-of relations and their properties such as transitivity; a path as a directed link and its associate properties). It relies on mechanisms that allow logical inferences about these properties.
- A regional geographic ontology in the form of an enhanced gazetteer that captures geographic names and places, and provides inferences about other spatial relations, such as containment and directions.

Current presentations of semantics on the Web form a spectrum that goes from full arbitrary expression of meanings that cannot be effectively computed to restricted, but tractable and well-defined sets of meanings. Each presentation of semantics plays a particular role for the four bearers of semantics (people, interfaces, documents, and systems). Future uses of semantics for Web-based information retrieval will require the integration of semantic information from various semantic domains

Acknowledgments

This work was partially supported by the National Imagery and Mapping Agency under grant number NMA202-97-1-102 and the National Science Foundation under grant numbers IIS-9613646, IIS-9970123, and EPS-9983432. Max Egenhofer's work is further supported by 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 NMA201-00-1-2009 and NMA201-01-1-2003.

References

- Berners-Lee, T., Hendler, J. and Lassila, O. (2001) The Semantic Web. *Scientific American* 184(5): 34-43.
- Dublin Core Metadata Initiative (1999) Dublin Core Metadata Element Set, Version 1.1: Reference Description. <http://www.dublincore.org/documents/dces/>
- Fensel, D. and Musen, M. (2001) The Semantic Web: A Brain for Humankind. *IEEE Intelligent Systems* March/April 2001: 24-25.
- Gabbay, D. (1998) *Elementary Logics: A Procedural Perspective*. Prentice Hall Europe, London.
- Hendler, J. (2001) Agents and the Semantic Web. *Intelligent Systems and their Applications* 16(2): 30 - 37.
- Jackendoff, R. (1990) *Semantic Structures*. MIT Press, Cambridge, Massachusetts.
- Larson, R. (1995) Semantics. in: L. R. Gleitman and M. Liberman, (Eds.), *An invitation to cognitive science*. Vol. 1. Language, pp. 361-380, MIT Press, Cambridge, Massachusetts.
- Lassila, O. and Swick, R. R. (1999) Resource Description Framework (RDF) Model and Syntax Specification. <http://www.w3.org/TR/REC-rdf-syntax/>
- McGuinness, D. L. (Forthcoming) Ontologies Come of Age. in: D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, (Eds.), *The Semantic Web: Why, What, and How*. MIT Press, Cambridge, Massachusetts.
- Partee, B. H. (1999) Semantics. in: R. A. Wilson and F. C. Keil, (Eds.), *The MIT Encyclopedia of the Cognitive Sciences*. pp. 739 - 742, MIT Press, Cambridge, Massachusetts.
- Uschold, M. (2001) Where is the Semantics in the Semantic Web? *Workshop on Ontologies in Agent Systems*. <http://cis.otago.ac.nz/OASWorkshop/Papers/WhereIsTheSemantics.pdf>
- van Harmelen, F., Patel-Schneider, P. and Horrocks, I. (eds.). (2001) Reference description of the DAML+OIL (March 2001) ontology markup language. <http://www.daml.org/2001/03/reference.html>
- W3C HTML working group (2001) XHTML™ 1.0: The Extensible HyperText Markup Language: A Reformulation of HTML 4 in XML 1.0. <http://www.w3.org/TR/xhtml1/>