"I Know What You Mean": New Semantics For the New Web

Sam Berner

11 September 2002

1. A Shallow Introduction to a Deep Subject

According to the Columbia Electronic Encyclopaedia, semantics [formed from the Greek verb sAmainx: "to mean" or "to signify"] in general, the study of the relationship between words and meanings. The empirical study of word meanings and sentence meanings in existing languages is a branch of linguistics; the abstract study of meaning in relation to language or symbolic logic systems is a branch of philosophy. Both are called semantics. The field of semantics has three basic concerns: the relations of words to the objects denoted by them, the relations of words to the interpreters of them, and, in symbolic logic, the formal relations of signs to one another (syntax). In linguistics, semantics has its beginnings in France and Germany in the 1820s when the meanings of words as significant features in the growth of language was recognized. Among the foremost linguistic semanticists of the 20th cent. are Gustaf Stern, Jost Trier, B. L. Whorf, Uriel Weinreich, Stephen Ullmann, Thomas Sebeok, Noam Chomsky, Jerrold Katz, and Charles Osgood. In the linguistics of recent years an offshoot of transformational grammar theory has reemphasized the role of meaning in linguistic analysis. This new theory, developed largely by George Lakoff and James McCawley, is termed generative semantics. In anthropology a new theoretical orientation related to linguistic semantics has been developed. Its leading proponents include W. H. Goodenough, F. G. Lounsbury, and Claude Lévi-Strauss.

The study of semantics is not limited to human-based languages, and crosses over to computing science, especially in the fields of natural language understanding systems, artificial intelligence and data mining. Computational semantics is the part of natural language semantics which uses computational methods to carry out this task. In his paper, Nerbonne (1995) stresses that there is a natural division of theoretical labor between the disciplines of linguistics and computational linguistics, namely that linguistics is responsible for the description of language and computational linguistics for the algorithms and architectures needed to compute with these. On this view the theoretical fields are related by their common focus on language, and moreover, computational linguistics is dependent on linguistics for the characterization of the relations it computes. Blackburn and Bos (2001) define computational semantics as "the business of using a computer to actually build semantic construction and reason with the result". It's a messy field, and quite new: Richard Montague's (1974) pioneering work on logical approaches to natural language semantics is barely thirty years old, much relevant work on knowledge representation and computational linguistics only dates back about twenty years, and it's only over the last decade that we've seen automated reasoning and constraint systems starting to achieve the levels of performance necessary for semantic applications. Monz and De Rijke (2000) state that "Computational semantics investigates the computational properties that formal semantic theories need to enjoy to be applicable to real-world problems".

Most of the work done in the fields of artificial intelligence and intelligent systems has been aimed at making machines more intelligent. Researchers such as David Heckerman of Stanford University spent the last two decades trying to make computers emulate the human brain (Microsoft, 1999). Andrew Broad from Manchester University uses Case-based Reasoning to try to make computers "smarter than man" (Broad, 1997). Douglas B. Lenat (2002) has set upon the task of making computers more intelligent. In his view, computers lack a fundamental world-knowledge basis upon which to draw. For several years now, Lenat with his *Cycorp* company, has been building such a knowledge basis, with amazing results. These are but a few examples of a vast amount of work being done.

Enter Tim Berners-Lee with a totally opposite view of how machines and people should interact. In his major work on the "semantic web", Berners-Lee (1998) writes:

> *"Leaving aside the artificial intelligence problem of training machines to behave like people, the Semantic Web approach instead develops languages for expressing information in a machine processable form."*

The ball is once more in human's yard. Where AI gurus try to make machines outsmart people, Berners-Lee and his W3C are trying to make us use data in a smarter way [Berners-Lee and Hendler, (2001)], just like primate anthropologists have tried over decades to teach chimpanzees and orangutans sign language to communicate. The end result – apart from hundreds of thousands of dollars spent on research – is clear: people are people, chimps are still chimps, and we don't socialize over an inter-species coffee and discuss ecological politics together. At any rate, not yet.

2. The Understanding Golem

So what exactly is this new "Semantic Web"? Much of the discussion is highly technical and difficult to digest by someone outside of computational science field. However, a few simplified resources about this new creation, among them a few pages at W3C website, writings by Berners-Lee himself and others in the area of business, knowledge management and information science.

An early mention of what the Semantic Web entails is Chislenko's (1997), in which he describes his vision of an intelligent web developed from simple metadata processing to globally distributed artificial intelligence services. Knowledge representation schemes and metadata standards have been discussed since 1995 [e.g. see: Holowczak and Li (1995), Horrocks (1995) and Sheldon et al (1996)]. Chislenko provides a short summary of the then available proposals for semantic standards and a list of value-added services made possible by the semantic web. Still in the AI mould, he still subscribes to the arduous tasks of knowledge bases and representations, hoping for a "*wide variety of networked knowledge processing servers collecting and generalizing data in their own areas and cooperating with each other for "interdisciplinary" problem solving, first with direct human involvement, and then increasingly on their own*".

According to Bonner (2002), the Semantic Web is about transforming the WWW into a sort of system in which content will be accessible not only to people but to machines as well; where it could be sifted, grouped, sliced and diced and served to us in ways we would not be able to achieve ourselves (does that imply "in a more intelligent way than we can"?). Bonner is a "techie", a software architect, and the AI bias is evident. Since his specialization is business process analysis and usability issues, it isn't surprising that the main reason why he applauds the Semantic Web idea is that it comes from Berners-Lee himself and therefore has potential for making money out of writing Semantic Web-based applications.

Like WWW, the Semantic Web uses the Internet backbone and HTTP to transfer resources located through URIs. The difference lies in the fact that whereas the WWW is designed to be people readable, thus rendering many "intelligent agents" as smart as the FBI ones, the Semantic Web is intended to be optimally accessible and comprehensible to machines and software. It attempts to guide software agents through the use of coded metadata (data about data). The aim, apparently, is to render the Internet into one vast, decentralized, machine-readable database. The owners of information will be able to determine access permissions, while users will be able to indicate which sources are trustworthy. But even with various limits imposed by both sides, any functional Semantic Web application will have access to a vast range of information sources.

Berners-Lee is emphatic about keeping AI out of the Semantic Web's bonnet. "*The concept of machine-understandable documents does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined* *problem by performing well-defined operations on existing well-defined data. Instead of asking machines to understand people's language, it involves asking people to make the extra effort,"* he stated (1998b). And although writers such as Weinberger (2002) decidedly point at the fact that Berners-Lee is falling into the AI morass, the creator of the Semantic Web asserts that his is not a rerun of failed knowledge representation experiments such as Stanford's KIF (knowledge interchange format) or Cycorp's Representation Language.

3. Tools of the Trade:

In his paper "Semantic Web Roadmap", Berners-Lee describes his new creation as an effort to "*develop languages for expressing information in a machine processible form*" [Berners-Lee, 1998]. Two starting points are XML and XHTML. XML (E**X**tensible **M**arkup **L**anguage), a metalanguage written in SGML that allows one to design a markup language, is an open standard for describing data from the W3C. It is used for defining data elements on a Web page and business-to-business documents. It uses a similar tag structure as HTML; however, whereas HTML defines how elements are displayed, XML defines what those elements contain. HTML uses predefined tags, but XML allows tags to be defined by the developer of the page. Thus, virtually any data items, such as product, sales rep and amount due, can be identified, allowing Web pages to function like database records. By providing a common method for identifying data, XML supports business-to-business transactions and is expected to become the dominant format for electronic data interchange. Since its introduction, XML has been hyped tremendously as the panacea to e-commerce, but it's only the first step. The human-readable XML tags provide a simple data format, but the intelligent defining of these tags and common adherance to their usage will determine their value. Unlike HTML, which uses a rather loose coding style and which is tolerant of coding errors, XML pages have to be "well formed," which means they must comply with rigid rules. XHTML (E**X**tensible **HTML**) is the combination of HTML 4.0 and XML 1.0 into a single format for the Web. XHTML enables HTML to be eXtended (the X in XHTML) with proprietary tags. XHTML is also coded more rigorously than HTML and must conform to the rules of structure more than HTML (Dumbill, 2000).

Although compliance with an XML DTD (**D**ocument **T**ype **D**efinition) can confirm the document's internal consistency and meets the standards, it does not convey anything about the meaning of the document's content. It merely renders the document *machine-readable*, not *machine-understandable*. W3C's solution to this is the use of RDF scheme. RDF (**R**esource **D**escription **F**ormat) is about metadata for Web resources, by resources we mean any object that can be found on the Web. Essentially, RDF is a means for developing tools and applications using a common syntax for describing Web resources. It is distinctive in that it is designed specifically with the Web in mind, so it takes account of the features of Web resources. It is a syntax based on a data model, and this model influences the way properties are described in that it makes the structure of descriptions explicit. This means RDF has a good fit for describing Web resources, but, on the downside, it might cause problems within environments where there is a need to re-use or interoperate with 'legacy metadata' which may well contain logical inconsistencies. The scheme is syntax independent but can be expressed in XML, and the specification uses XML as its syntax for encoding metadata [Heery, 1998]. The basic unit of data in RDF is a triple, consisting of a subject (a resource identifier), a predicate (a characteristic, attribute, or relation), and an object (either another resource or literal data). Each element in a triple is represented as a URI, pointing to a unique Web address. These triples are rendered within an HTML document as XML metadata. The factual connections between data items that RDF establishes help make documents more comprehensible to automated readers. But there is still a lack of context in some of the data and fuzziness about how it relates to other data. To formalize meaning, ontologies and taxonomies are used [see Section 4 below]

Allen (2001) discusses various ways in which semantic metadata can be created: text parsing, embedded text, user published files, as a service provided by the data owners, crawling, specialized tools and generic servers among others.

4. Organising Meaning - Ontologies and Taxonomies

All the work on databases assumes that the data is nearly perfect. Few (if any) database systems are ready for the messiness of the Web. Any system that is "hard-coded" to understand certain terms will likely go out of date, or at least have limited usefulness, as new terms are invented and defined. Worse, there's no way for a computer or human to figure out what a specific term means, or how it should be used. The use of all these URIs is useless if we never describe what they mean. This is where schemas and ontologies come in.

It doesn't seem anyone is really sure what an ontology is, and many definitions contradict one another. It has a background in metaphysics, but has since been taken over by the Artificial Intelligence and Knowledge Representation people. According to Gruber (1993) an ontology is "a specification of a conceptualization", i.e. an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. Ontologies and taxonomies are, in functional terms, often used as synonyms. Computer scientists call hierarchies of structured vocabularies "ontologies" and librarians deploy the term "taxonomy." According to Noy and McGuiness (2001) "*ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)). An ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins.*"

The Semantic Web entails adding an extra layer of infrastructure to the current HTML Web. Metadata and structured vocabularies make it easier for databases to communicate with each other. A major problem with the Internet today is data fragmentation. With the Semantic Web, computers understand the meaning of a Web page by following hypertext links from Web documents to topic-specific ontologies. For instance, ontologies offer cross-references so a computer understands that "movie," "film," "flick," and "motion picture" are different expressions of the same concept. While intelligent agents do the visible labor of the Semantic Web, taxonomies will be facilitating communication among machines behind the scenes. For computers flung around the world to work together, a common set of terms--vocabularies--is needed and then rules that lay out how those terms work together. Taxonomies are an important part of what makes the Semantic Web "intelligent." Vocabularies and the relationships that exist between selected terms help machines to understand conceptual relationships as humans do.

Another useful definition of an ontology is a document where "*definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms.*" [Gruber (1997)]. An ontology is a document, that is, a processable resource consisting of assertions and definitions. This accords with the ISO SGML definition that a document is "a collection of information that is processed as a unit" [ISO (1986)], which puts the definition of a document's message or meaningfulness squarely in terms of the process which interprets the message. By this definition, that which is meaningful is that which is chosen and acted upon where inaction is an action. One observes and measures the behavior of the agent to determine its appropriateness. One can consider that the best measure of meaning is its usefulness to the owner. Ontologies are defined by communities whose members are presumably experts in some knowledge domain. The ontology contains assertions made by the community members about that domain. Given that the agent owner may not be a member of such community or may be encountering it for the first time, the definitions of the ontology must be tested prior to granting the source an authoritative status over the domain [Bullard, 2001]. The common use of XML/RDF enables the transport and merging of ontologies and the interoperation of the software or agents that consume the contents and behave accordingly. The agent uses these descriptions to choose among resources by identifying which resources match some set of search criteria. It may then use these resources as rules or values for rules for executing a process such as negotiation. Thus, it becomes the role of the ontologist to define a classification language, the role of the owner of a resource to choose the classification for the resource, and the role of the agent to find and potentially negotiate the use of resources based on the goals and instructions of the agent's owner.

5. The Semantic Web's 2Ps:

5.1. The Promise

The possibilities for Semantic Web-based software are nearly boundless: complex B2B or B2C transactions with no human intervention; aggregation, amalgamation, and mining of research, census, and historical data in ways beyond human capability; transparent, on-the-fly assembly, instantiation, and linking of distributed virtual applications; or just about anything people can dream up to use the power of intelligently sifting through data from millions of connected computers.

Allen (2001) calls it the "reinvigoration of the Web". And it looks promising from the information management and knowledge management perspective, too. Some of the issues involved are context-aware links, site information, collaborative filtering, collaborative categorization, annotations, related links, and corrections.

Berners-Lee has already whetted the appetite of the academic environment, promising that tools for publishing papers on the web will automatically help users to include more of this machine-readable markup in the papers they produce. Papers that include this new markup language will be found by new and better search engines, and users will thus be able to issue significantly more precise queries. More importantly, experimental results will themselves be published on the web, outside of the context of a research paper; while work will be able to be modified as a result of interaction with peers, with less need to wait for formal publication [Berners-Lee and Hendler, (2001)].

Grimes (2002) predicts the use of the Semantic web in business intelligence and knowledge management, seeing the aim in enabling profile-based software agents to search and then act on the results; thus shifting the burden of information processing to a world of automated, P2P computing. He visualized a closed loop networked system that classifies and categorises via KM and data mining techniques, applies logarithms to score and rank the results using inference rules, and derives and executes business rules.

5.2. The Problems

There are a few pre-requisites the Semantic Web must fulfill before it becomes useful. Allen (2001) mentions the fact that it has to be **globally inclusive**, **collaborative** and **interoperable**. This last one will be a major commercial problem to overcome, because in the potential market for tools that allow users to benefit from shared metadata, there will be a temptation to create customized metadata formats for each tool in order to lock in customers or meet specific short-term business needs.

The Semantic Web, as I mentioned above, aims at rendering the Internet into a searchable and decentralized database. Governments, both democratic and otherwise, do not feel comfortable with decentralized information structures, especially since the latest terrorist attacks. In fact, metadata is intended to be indexed, searched, and processed by tools, while web pages are designed for humans to view one page at a time. The best way to index something to allow fast searching and processing is to centralize it. If you consider Napster, you can see that the metadata was centralized for fast querying, and the data was decentralized for fast retrieval. This design rule, "decentralized data, centralized metadata", can be seen on the web today. At the same time as HTML pages are being decentralized further and further from the core through services like akamai and inktomi edge caching, searching functions on the Internet are being centralized into ever more powerful engines like Google. Note that it is possible to have actual metadata stored in a decentralized way, but the indexing and querying of that metadata will be most efficient if centralized. What this means is that users will demand systems that perform well, and as we push the limits of what we do with metadata (multiple levels of inference, for example), the performance strain of querying metadata will become increasingly greater. Because the performance requirements will increase as user expectations of features increase, people will tend to be drawn more to metadata services that store metadata centrally. At the same time, the net's libertarian ethic will mistrust anything that is too centralized. The social desire for decentralization will probably be no match for the consumer demand for features and performance, however, so it is likely that most collaborative metadata functions will end up collecting in few centralized services. Developers of shared metadata services will need to design with the expectation that the inexorable forces of nature will tend toward centralized metadata, and build in safeguards to make sure that their systems cannot become *too* centralized (and thus become exposed to Byzantine threats, jeopardize critical mass of user willing to use the service, etc.) This fundamental characteristic of shared metadata to flow to the center is also an opportunity for the software politicians. As market forces drive metadata into centralized repositories, these centralized services will become wonderful targets for conspiracy theories and other such sensationalized portrayals that help the software politicians maintain flock size.

Besides threats to privacy, there is another problem with metadata being centrally stored or indexed. Tools depend on the ability to query metadata (for example, a tool might ask for the top five recommended links associated with a certain page). Having a standard way in which tools can query metadata servers allows for tools to be independent of servers, and encourages tool development **and** service development. However, if there are only a few central indexing services, there will be a strong incentive for these services to compete by differentiating their services, adding features, and "improving" the query interfaces. It is likely that reducing the number of metadata services would lead to tools that are tightly-coupled to one service or another.

Another problem with the Semantic Web is standarisation. The best thing about standards is that there are so many to choose from. Such standards as the Dublin Core are very slow on the uptake. Even XML, wonderful as it is for document consistency, is not being used as often as one would have expected. The Semantic Web standards proposed by W3C is currently the main contender, but W3C is an elitist organization which has its own interests to maintain.

6. An Attempt at a Meaningful Conclusion

The Semantic Web is not here yet, and it may take some time before it is. W3C described it as "a vision for the future of the Web" [Heflin, 2002]. When the future will eventuate is still too early to predict, but it will definitely do so. Meanwhile, the technologies being developed in its anticipation will likely provide intermediate benefits to the business and academic sectors. By learning to make knowledge more accessible to machines, we are developing the means to make content more accessible and more manageable to humans as well. Mozilla, for example, uses RDF for its configuration files and contains a template-building technology, XUL, that can construct user interfaces populated with data from RDF data sources. A Javascript API for manipulation of those data sources is also available. External data sources can be integrated into Mozilla either by exposing them in the RDF syntax or by adding code to Mozilla conforming to its data source API.

The Semantic Web has already been the subject of much bluster among the XML developer community and will doubtless continue to be so. Arguments rage over the usefulness of the technology, the difficulty of using RDF, and so on. However, the Semantic Web vision of a machine-readable web has possibilities for application in most web technology -- while some complain about its lack of definition, its broad scope properly reflects the quietly radical effect it will have on the Web. The bluster is not catching everyone, though: Sean Palmer (2001) half-heartedly worries that there may be too much inference from this vastly centralized metadata, and that simple folks trying to process their shopping lists may end up in Oslo or Kyoto with some world peace plan instead. He also decries the academia and corporate bias lingering around the Semantic Web, renaming the construct as the "Pedantic Web".

Berners-Lee (1998) himself said that the problem with search engines indexing HTML and digging out thousands of irrelevant links will not go away, and that machines which guarantee answers to arbitrary questions won't come into existence. His promise is that the new "understanding" machines will have the remarkable power to answer real questions which are the stuff of our daily lives and especially of commerce. A bit like Hans the Horse, only on a larger scale. Humanity, with its tendency to err, explore and be arbitrary, will be encourage to take on an equestrian attitude to knowledge.

7. References:

Allen, J. 2001. Making a Semantic Web. [Online] Available WWW: www.netcrucible.com/semantic.html

Berners-Lee, T. 1998. Semantic Web Road Map. [Online] Available WWW: http://www.w3.org/DesignIssues/Semantic.html

Berners-Lee, T. 1998b. What the Semantic Web can represent. [Online] Available WWW: http://www.w3.org/DesignIssues/RDFnot.html

Berners-Lee, T. and Hendler, J. 2001. Scientific publishing on the 'semantic web'. In **Nature** [Online] Available WWW: http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm

Blackburn, P. and Bos, J. Representation and Inference for Natural Language. [Online] Available WWW: http://www.comsem.org/

Bonner, P. 2002. The Semantic Web. In **PC Magazine**, 21(13): IP01-03

Broad, A. 1997. Case-based Reasoning. [Online] Available WWW: http://www.cs.man.ac.uk/~broada/cs/cbr/cs3411.html

Bullard, L. 2001. Building a Better Golem. In **Markup Languages: Theory and Practice**, 2(4):337-351

Chislenko, A. 1997. Semantic Web Vision Paper. [Online] Available WWW: http://www.lucifer.com/~sasha/articles/SemanticWeb.html

Dumbill, E. 2000. The Semantic Web: A Primer. [Online] Available WWW: http://www.xml.com/pub/a/2000/11/01/semanticweb/index.html

Grimes, S. 2002. The Semantic Web. In **Intelligent Enterprise Magazine**. [Online] Available WWW: http://www.intelligententerprise.com/020328/506decision1_1.shtml

Gruber, T. 1993. A translation approach to portable ontologies. In **Knowledge Acquisition**, 5(2):199-220

Gruber, T. 1997. What is an Ontology? [Online] Available WWW: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Heery, R. 1998. What is RDF? In **Ariadne**, (14). [Online] Available WWW: http://www.ariadne.ac.uk/issue14/what-is/

Heflin, J. 2002. Requirements for a Web Ontology Language. [Online] Available WWW: http://www.w3c.org/TR/webont-reg/

Holowczak, R. and Li, W. 1995. A Survey on Attribute Correspondence and Heterogeneity Metadata Representation. [Online] Available WWW: http://www.computer.org/conferences/meta96/li/paper.html

Horrocks, I. 1995. Comparison Of Two Terminological Knowledge Representation Systems. A thesis submitted to the University of Manchester for the degree of Master of Science (Computer Science). [Online] Available WWW:

http://citeseer.nj.nec.com/cache/papers/cs/3912/ftp:zSzzSzmighp0.cs.man.ac.ukzSzpubzSzthes eszSzhorrockszSzthesis.pdf/horrocks95comparison.pdf

International Organization for Standardization (ISO). 1986. International Standard ISO 8879: 1986(E) information processing — text and office systems — standard generalized markup language (SGML). Geneva: International Organization for Standardization.

Lenat, D. 2002. From *2001* to 2001: Common Sense and the Mind of HAL. In David Stork (ed.) **HAL's Legacy: 2001's Computer As Dream and Reality**. Cambridge (MA): The MIT Press

Microsoft. 1999. Microsoft Researcher David Heckerman Translates His Passion for Data Into a Series of Technological Breakthroughs. [Online] Available WWW: http://www.microsoft.com/presspass/features/1999/03-22heckerman.asp

Montague, R.1974. Formal Philosophy: Selected Papers of Richard Montague. NJ: Yale University Press.

Monz, C. and De Rijke, M. (2000) Inference in Computational Semantics. In **Language and Computation** 1(2):151-156 [Online] Available WWW: http://staff.science.uva.nl/~christof/Papers/jlac.pdf

Nerbonne, J. 1995. Computational Semantics--Linguistics and Processing. In Shalom Lappin (ed.) **Handbook of Contemporary Semantic Theory**. London: Blackwell Publishers. pp. 461-484

Noy, N. and McGuiness, D. 2001. Ontology Development 101: A Guide to Creating Your First Ontology. [Online] Available WWW: http://ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf

Palmer, S. 2001. The Semantic Web: An Introduction. [Online] Available WWW: http://infomesh.net/2001/swintro/

Sheldon, M. et al. 1996. Services and Metadata Representation for Distributed Information Discovery. Proceedings of the W3C Distributed Indexing/Searching Workshop, MIT, Cambridge (MA): 28-29 May.

Weinberger, D. 2002. The Semantic Argument Web. In **Darwin Magazine**: (14.06.2002). [Online] Available WWW: http://www.darwinmag.com/read/swiftkick/column.html?ArticleID=421