

It's all Semantics: The Semantic Web and PR

James L. Horton

Wouldn't it be nice to have a media web site that could do the following? A reporter goes to the site, clicks on one link and every release, reference to a product and allied service is instantly lined up for the reporter from as many databases as the company has. This can be done today using a strangely named extension of the World Wide Web called the "Semantic Web."

The idea behind the Semantic web came in the late 1990s from the inventor of the World Wide Web – Sir Tim Berners-Lee. Berners-Lee was bothered by the way that data on web pages was unable to relate to data on other web pages. Data was trapped and static in a system that related documents to documents but not data in the documents to other data. Yes, search engines crawl web pages and capture data but search engines can be overly focused or wildly imprecise. Ask for "printer cartridges" on Google and you will get hundreds of stores selling printer ink cartridges, but you won't get a result that shows printers and their cartridges lined up by type and serial number. That data is buried on respective web pages.

Berners-Lee believes the next step in the evolution of the World Wide Web needs to be relating data to data. He chose the term "Semantic" to describe this insight. Berners-Lee today calls the word choice unfortunate. He believes he should have called his insight the "Data Web." No matter what name one uses, the Semantic Web is a system of relating data in web pages that may change the way PR works.

Tagging data

The Semantic Web is similar to tagging systems individuals use in blogs to help search engines like Technorati (<http://www.technorati.com/>) find topics and pull them up. For example, if I blogged about baking an apple pie, I might give two tags to my blog posting – "recipe," "apple pie." Technorati in turn would look for these tags when a user searches for recipes or apple pie.

Similar meta-tagging systems are used to identify web pages to search engines. Meta-tags are invisible to web page users (unless they look at source code), but they tell search engine software what a web page is about. Here for example are the meta-tags for www.online-pr.com -- "PR, public relations, online PR, Online public relations, marketing PR, promotion, publicity, public relations online."

When someone searches on any of these terms, Google knows to call up online-pr.com.

Other tagging systems are based on extensible mark-up language (XML), from which web formatting coding (HTML) derived. Today, millions of documents are formatted in XML because of its tagging system. XML tagging allows users to define concept and formatting elements in documents and then use these elements to tag data for formatting and search. To do this, XML provides a place for users to create a data-and-formatting dictionaries for information to follow. For example, a client some years ago loaded 100,000 pages of repair manuals for an airplane into an XML file then tagged each reference to tens of thousands of parts in the plane. Technicians could search on the part number and get every instance of repair related to it including an illustration of the part for reference. This system worked without breaking down text into fields for coding into a relational database – a prohibitively expensive task.

Why another tagging system? Because other tagging systems are inherently limited. The Semantic web strives to unite data across multiple locations and databases. It wants to free data. Blog tagging is determined by individual taggers. One blogger's concept may have little or no relationship to another's. For example, another blogger might use the tag, "good eating" to identify the post about baking an apple pie. A search engine cannot distinguish among "recipe," "apple pie" and "good eating." As far as a search engine is concerned, these are three separate topics. Blog tagging systems include del.icio.us (<http://del.icio.us/>), the social bookmarking software, and Digg (<http://digg.com>) the content sharing site.

Meta-tags for web pages are descriptors of the entire page or even of the entire site. Tags may not describe the discrete data on the page. It is like having the tag "food store" to describe a supermarket's web site, but not the 20,000 stock keeping units inside the store's site.

XML tagging systems define at the beginning of a document what tags will mean within the document. One can tag words, phrases, data for formatting and data search purposes. One tag might be an instruction to italicize a word. Another tag might be a classification of the word under a defined heading such as "canned goods." The problem with XML tagging is that it is pertinent to each document, but once one leaves the document for another, there may be a different set of definitions that make tags incompatible.

There are groups that have developed standardized XML tags for industries, but there is no guarantee everyone in an industry will use the system. Consider, for example, XBRL a classification tagging system for business. (<http://www.xml.com/pub/a/2004/03/10/xbrl.html>). The idea is that every business would use the same classifications so data in XML documents can be

searched across industry. However, once one leaves the XBRL framework, precise data searching ceases to work.

The Semantic Web

The Semantic web uses something called the Resource Description Framework (RDF) to define data. It starts with self-description of data that other tagging systems use, but RDF names each item and relationships between items in way that allows computers and software to exchange them automatically. It works best when groups agree to common schemes of data tagging as they do in XML. However, groups can be of any size, which means tagging is decentralized and doesn't require agreement across industries. RDF uses a triple descriptor— a noun, verb and object. The triple descriptor is called a Universal Resource Identifier (URI). A typical web page address such as www.online-pr.com is a special form of a URI. The descriptor for the concept apple pie might be “Apple pie is a baked good.” Or the descriptor might refer to a common recipe for apple pie such as “Apple pie is <http://allrecipes.com/Recipe/Mile-High-Apple-Pie/Detail.aspx>,” an available online recipe for apple pie to which many different people could link. This methodology is also called Data Linking. URI codes also tell where they came from, and they can be geo-coded to relate to maps.

Why bother with Data Linking? Isn't Google good enough? Yes and no. Google finds data on web pages but not necessarily data in databases, and it has a tendency to return a huge number of hits for terms, most of which aren't useful. For example, for the concept, “apple pie,” Google had 6.39 million hits on the day the term was searched. It is probably higher now. Hundreds of thousands of those references would have no usefulness to the searcher, especially if the concept the searcher is looking for is not physical apple pie. There is a colloquial use of apple pie – “apple pie of my eye” – that doesn't refer to actual pie but to the beauty of an individual or thing. In fact, Google had 15 discrete references to “apple pie of my eye,” but one had to know to search precisely on the term. As with all searching, results are limited by underlying data and how it occurs in documents. For example, one could write a whole recipe for apple pie without using the term “apple pie.” Google would not see that data if one uses the search term, “apple pie.”

The Semantic web creates common formats for data integration and permits combination of data from multiple sources. It is used mostly now for databases of information and less for coding data on web pages, but the potential for coding web pages is there. In a Dec. 2007 article in *Scientific American*, the authors summarized the Semantic Web as follows:

“The Semantic Web is not different from the World Wide Web. It is an enhancement that gives the Web far greater utility. It comes to life when people immersed in a certain field or vocation, whether it be genetic research or hip-hop music, agree on common schemes for representing

information they care about. As more groups develop these taxonomies, Semantic Web tools allow them to link their schemes and translate their terms, gradually expanding the number of people and communities whose Web software can understand one another automatically.”

In sum, the Semantic Web is used for data integration, for resource discovery and for cataloging.

The Semantic Web is not perfect. No system can be. Critics of the Semantic Web point out that abstractions are incomplete by nature. They cannot describe the complete reality of anything. So, even though the Semantic Web links data, it won't handle every possible conceptual relationship in every instance that a user might need. It will do better than the Web does today in finding and relating data.

Ontology

What happens when one has tagged data in a document, so it can be found? That brings up the next step of the Semantic web. Tags for concepts must be shared across documents and data or somehow coordinated if they are not identical. Thus, the next step was to collect URIs and to place them in computerized lists called ontologies. This is done through a formal Ontology Web Language called OWL. OWL establishes the relationships of concepts within the data set, which allows these relationships to be translated across other databases. Ontologies are like a library classification. They begin with the individual object (the book), then classes or sets (Novels, History) with descriptions of attributes and relationships among them. Ontologies have to be structured so they are recognizable in the Resource Description Framework but once they are, it becomes easier to solve problems such as the following:

- **Mailing list.** On one mailing list, data is filed under Zip code. On another list, it is filed under Postal code. With coordinated ontologies, the computer understands that postal and zip code are the same.
- **Finding the right service.** A city offers hundreds of online services on multiple websites. Using ontologies and a search engine, the city created a way for citizens to make natural language requests that found the right service on the right web site. (This actual case refers to the City of Zaragoza in Spain. The case is located at www.w3.org/2001/sw/sweo/public/UseCases/Zaragoza/)
- **Creation of a database that allows natural language inquiries of Wikipedia.org.** Dbpedia (<http://dbpedia.org/About>) is described as a “community effort to extract structured information from Wikipedia and to make this information available on the Web. DBpedia allows you to ask sophisticated queries against Wikipedia and to link other datasets on the Web to Wikipedia data.”

- **Creation of a music database.** MusicBrainz (<http://musicbrainz.org/>) that tracks nearly 390,000 artists and 6.8 million tracks using the semantic web and a music ontology.
- **Building a social networking system.** The Friend of a Friend project (FOAF) (<http://www.foaf-project.org>) is an open-source system based on the Semantic web.

Ontologies allow creation of data webs that span multiple databases, web pages and the internet itself depending on how they are deployed. Ontologies can be short or thousands of terms long. There is no limit on size.

Inference engines

One level above ontology is a computerized inference engine that examines multiple ontologies in order to find new relationships among terms and data in them. With such engines one can do the following:

- **Creating an integrated clinical database from multiple discrete databases.** This is what the Cleveland Clinic accomplished with 195,000 patient records and 54.2 million RDF statements. The database allows the clinic to track patients longitudinally across pharmacy records, local and group practices and primary, secondary and tertiary hospitals. (See www.w3.org/2001/sw/sweo/public/UseCases/Cleveland)
- **Created an internal index for a magazine that relates discrete concepts to the same concepts from multiple issues.** See: <http://www.harper's.org/archive/2003/10/0079762>. The magazine, *Harper's*, uses its inference engine to deliver references from back issues in a coordinated and easy-to-understand fashion. It is less of a data dump and more of an organized presentation.

Inference engines use computerized reasoning within the rules of the Semantic web and ontologies. The result is greater data inclusion and precision.

Semantic web and PR.

PR is about information and transparency. Often there is too much information spread in too many places on the web and in organizational databases that is not transparent. The Semantic web provides tools to gather data more easily into one place.

Think, for example of what could be done with a company's multi-year backlog news releases, product fact sheets and specifications, speeches, white papers and other materials deposited on a web site for media use. From personal experience, I can tell you that reporters, editors and producers are defeated

regularly by complex web sites because data is buried throughout the site under different names and concepts, and search engines are inadequate. The benefit of providing this data in an integrated and coordinated fashion is that it allows outsiders to understand more quickly an organization's message. On the other hand, if the organization has changed its message over time, it would reveal that as well. There will be instances in which the transparency of the Semantic web may prove to be too open for proprietary reasons, but there is an answer for that. Don't code the data, and it will remain largely invisible.

The Semantic Web is a development that PR practitioners should know about, but they needn't learn how to apply about underlying machinery. That is for technologists. The important point is that as web pages and databases grow more complex, there is a way to make them transparent. The Semantic web is growing rapidly within the technology community. It will reach the general web community soon enough. Now is the time to prepare for it.

#