Toward the Semantic Web

June 22 2010, by Larry Hardesty

When the World Wide Web went live in 1991, it consisted of static pages of text connected to each other by hyperlinks, and that's pretty much what it remained for years. But from the outset, the Web's inventor, Tim Berners-Lee, had envisioned a much more sophisticated Web, a so-called Semantic Web, which wouldn't just store data but would actually know what it meant.

Now an MIT professor, Berners-Lee also directs the World Wide Web Consortium (W3C), a standards body whose industrial participants include everybody from Adobe to Yahoo, and which maintains an office at MIT's Computer Science and Artificial Intelligence Lab. The W3C has just published a new standard that should help bring the Semantic Web that much closer to fruition.

If the current Web is like a giant text file - which you can search for instances of particular words - the Semantic Web would be like a database, where every item of information is categorized, and new queries can combine categories in any imaginable way. For instance, if you were looking for the menus of a particular type of restaurant in a particular part of town, you could pull up just those pages featuring menus - not page after page of review sites that happened to use the word "menu."

But while an ordinary database has categories selected in advance by a programmer, the Semantic Web is "a database where each person controls their own data," says Sandro Hawke, systems architect at the World Wide Web Consortium (W3C). "You have your own parts of the database, so you can put whatever data out there that you want."

A giant networked database where people control their own data has obvious advantages: huge numbers of people can contribute to it, and they can ensure that their contributions aren't categorized or recorded incorrectly. But it also has an obvious disadvantage: There's no guarantee that people will organize and label their data in a uniform way.

To take a simple example, suppose that two nearby medical clinics put their staff lists online. Semantic Web technologies would allow the clinics to categorize the information in the lists. But suppose that one clinic chose to label the surnames of its doctors "surname," and the other clinic chose the label "last name." A Web search that listed local doctors by "surname" might not pick up those labeled "last name," and vice versa.

In fact, an existing Semantic Web standard, the Web Ontology Language, solves this problem. The language gives programmers a way to specify that, for instance, "last name," "surname," and maybe "family name" or just "last" indicate the same types of data.

The case for rules

But what if a third clinic, while still adopting Semantic Web technology, chooses to dump first names, last names, and middle initials into a single category, labeled "name"? A direct mapping of category to category will no longer work. Instead, unifying the data on different sites requires a rule, such as, Put everything up to the first space character in "first name," anything after the last space character in "last name," and anything else in "middle."

The newly released Semantic Web standard is called the Rule Interchange Format, or RIF, and it gives Web programmers a way to write rules for translating between data on different sites. But that's not the only purpose rules serve on the Web. For instance, Hawke points out, an online Web retailer might offer customers free shipping if their total purchases exceed some threshold in a given time period; but the retailer's Web servers might store no data about its customers other than individual invoices. The code for sifting through the invoices and determining whether to offer the discount is another example of a rule. "Part of the standards game is to have these very different use cases around the same table and then get one standard that can be used in all these different pieces of software," Hawke says.

If the RIF standard becomes widely adopted, it's likely to go unnoticed by most Internet users. The Web is already replete with pages that aggregate data from other sites: A personalized Google home page, for instance, might include headlines from several different news sources, weather reports from yet another site, and stock prices from still another. When such content aggregators are already popular online destinations, it can be hard to convey exactly what the advantage of a Semantic Web would be. But as Hawke puts it, "You can always build something to aggregate data you already know about"; what the Semantic Web offers is a way to aggregate data you *don't* already know about. A small site that

lists weekend events in a particular neighborhood, for instance, could retrieve data from sources that didn't even exist when it was built, as long as they categorized their data according to [Semantic Web](#) standards.

Although it has been nearly 20 years since Berners-Lee launched the first website, if his original idea finally comes to fruition, "it'll happen so quickly that no one will know," Hawke says. "They'll just notice the Internet doing more cool things."

More information: www.w3.org/

Provided by Massachusetts Institute of Technology