



*Meaningful pursuit: He invented the World Wide Web. So when Tim Berners-Lee talks about the next big thing, people listen.*

# A SMARTER WEB

*How do you endow the Internet's chaotic pile of bits with a structure that makes information easier to find and use? It's all a matter of semantics.*

BY MARK FRAUENFELDER

*Photographs by Patricia McDonough*

TECHNOLOGY REVIEW November 2001

Tim Berners-Lee must feel like he's in a time warp. In the early 1990s, he spent a frustrating year trying to get people to grasp the power and beauty of his idea for a scheme known as an Internet hypertext system, to which he gave the beguiling name the World Wide Web. But since the Web didn't yet exist, most people couldn't imagine the implications of what he was talking about. Berners-Lee persevered, and with the help of the few people who shared his vision, his invention became the fastest-growing media distribution system in history. A decade later, Berners-Lee is struggling with the same problem—only this time, he's trying to articulate his dream of a Semantic Web. The idea is to weave a Web that not only links documents to each other but also recognizes the meaning of the information in those documents—a task that people can ordinarily do quite well but is a tall order for computers, which can't tell if “head” means the leader of an organization or the thing on top of a body. “The Semantic Web is really data that is processable by machine,” says Berners-Lee, who is director of the MIT-based World Wide Web Consortium. “That's what the fuss is about.” Today's World Wide Web is fundamentally a publishing medium—a place to store and share images and text. Adding semantics will radically change the nature of the Web—from a place where information is merely displayed to one where it is interpreted, exchanged and processed. Semantic-enabled search agents will be able to collect machine-readable data from diverse sources, process it and infer new facts. Programs that weren't made to be compatible with each other will share previously unmixable data. In other words, the ultimate goal of the Semantic Web is to give users near omniscience over the vast resources of the Internet, turning the millions of existing database islands into a single gigantic database Pangea. To compare the Semantic Web with today's Web,

Technology Review November 2001

[www.technologyreview.com](http://www.technologyreview.com)

pg 1 of 7

Berners- Lee—an intense person who speaks in low-volume bursts—offers the following scenario: Imagine registering for a conference online. The conference Web site lists the event time, date and location, along with information about the nearest airport and a hotel that offers attendees a discount. With today’s Web, you have to first check to make sure your schedule is clear, and if it is you have to cut and paste the time and date into your calendar program. Then you need to make flight and hotel arrangements, either by calling reservations desks, or by going to their Web sites. “There’s no way you can just say, ‘I want to go to that event,’” explains Berners- Lee, “because the semantics of which bit is the date and which bit is the time has been lost.” But on the Semantic Web, he asserts, those bits will be labeled; the software on your computer will recognize those labels and automatically book your flight to the conference and reserve a hotel room with the click of a button. The Semantic Web will also be a richer, more customizable Web. Imagine running your cursor over the name of the hotel and being informed that 15 percent of the people who’ve voted on its quality say it’s excellent. If you happen to know that the hotel is a dump, you can instruct your browser to assign those people a trust level of zero. (The polling information would be saved on a third-party “annotation server” that your Web browser accessed automatically.) By assigning high levels of trust to people who match your tastes and interests, and “bozo-filtering” the people who don’t, the Web will start looking more like your Web. It’s an enormous undertaking. The first step is to establish standards that allow users to add explicit descriptive tags, or metadata, to Web content—making it easy to pinpoint exactly what you’re looking for. Next comes developing methods that enable different programs to relate and share metadata from different Web sites. After that, people can begin crafting additional features, like applications that infer additional facts from the ones they’re given. As a result, searches will be more accurate and thorough, data entry will be streamlined and the truthfulness of information will be easier to verify. At least that’s the goal. Many feel it can’t be done. Even though things are heating up in research labs, the Semantic Web as envisioned by Berners-Lee is hampered by social and technical challenges that some critics say may never be solved. But that’s not stopping the World Wide Web Consortium and other organizations from trying. The U.S. Defense Advanced Research Projects Agency (DARPA) and commercial enterprises such as Network Inference in Manchester, England, are already developing tools for building the Semantic Web infrastructure—as well as applications for using it. And according to Berners-Lee, with growing numbers of people beginning to grasp how the Semantic Web will “allow more and more sophisticated agents to do things on their behalf,” we’ll soon see some glimmers of what could be in store.

### **Untangling the Semantic Web**

In his crowded office on the third floor of MIT’s Laboratory for Computer Science building, research scientist Eric Miller doesn’t seem bothered by the pounding and grinding noises coming from heavy equipment on the construction site next door. As the head of the Semantic Web project, the friendly and energetic Miller is too enthralled with his new job to notice. “I’m the luckiest guy alive,” he says. “I

get paid for what I'd do for free." Berners-Lee tapped Miller to head up the consortium's Semantic Web Activity because of Miller's involvement in Webbased knowledge management projects and his ability to enthusiastically articulate the concepts behind the Semantic Web. Standing next to a whiteboard covered in diagrams of metadata in action, Miller explains that the fundamental idea behind the Semantic Web is to make the Internet more useful to people by making the information floating all over the Web more easily manipulated by computers. Today, by contrast, most content is formatted for human consumption. When you read a news article online, for instance, you can easily pick out the headline, byline, dateline, photo credit and so on. But unless these things are explicitly labeled, a computer has no idea what they are. It simply sees a bunch of text. In the Semantic Web, a news story will be marked with labels that describe its various parts, making it easy, among other things, for a search engine to find articles written by Jimmy Carter and not stories written about him. That's not possible today, at least not on a global scale. The formatting tags used to create Web pages are part of the hypertext markup language (HTML), and they describe only what a Web page's information looks like (boldface, small, large, underlined, etc.). The Semantic Web would go beyond cosmetics by including tags that also describe what the information is: tags would label text as designating, for instance, subject, author, street address, price or shipping charge. These descriptive tags are the metadata—the data about the data. Metadata is not a new concept, nor one restricted to the Internet. A library's card catalogue—with its records describing a book's title, author, subject, year and location on the shelves—is metadata. The Web made it trivially easy to exchange documents between previously incompatible computers (a few of today's Web users may recall the headaches of the 1980s, when computers from different makers were electronic islands). The Semantic Web will take this a step further, making it possible for computers to exchange particular pieces of information from within documents.

*AS THE CONSORTIUM DEVELOPS TECHNOLOGIES FOR THE SEMANTIC WEB, HUNDREDS OF ORGANIZATIONS, COMPANIES AND INDIVIDUALS ARE CONTRIBUTING TO THE EFFORT.*

### **Beyond Metadata**

You can't have a Semantic Web without metadata, but metadata alone won't suffice. The metadata in Web pages will have to be linked to special documents that define metadata terms and the relationships between the terms. These sets of shared concepts and their interconnections are called "ontologies." Say, for example, that you've made a Web page listing the members of a faculty. You would tag the names of the different members with metadata terms such as "chair," "associate professor," "professor" and so on. Then you'd link the page to an ontology—one that you created yourself or one that someone else has already made—that defines educational job positions and how they relate to each other. An appropriate ontology would in this case define a chair as a person, not a thing you sit on, and it would indicate that a chair is the most senior position

in a department. By defining the relationships between terms, ontologies can then be used by applications to infer new facts. Suppose you have created a Web page that teaches schoolchildren about condors, and have added metadata to the content. You could link to an ontology (or more likely, several ontologies) that define the various terms and their relationships: “California condor is a type of condor from California.” “Condor is a member of the raptor family.” “All raptors are carnivores.” “California is a state in the United States.” “Carnivores are meat eaters.” By using both metadata and ontologies, a search engine or other software agent could find your condor site based on a search request for “carnivores in the U.S.”—even if your site made no mention of carnivores or the United States. Because ontology development is a big undertaking, it’s likely that site creators will link to third-party ontologies. Some will be free, others will be sold or licensed. One issue that will have to be confronted: just as with dictionaries and atlases, political and cultural bias will creep into ontologies. A geography-based ontology maintained by the Chinese government, for instance, would probably not define Taiwan as a “country.” But that hardly impedes the vision. As the World Wide Web Consortium continues to develop standards and technologies for the Semantic Web, hundreds of organizations, companies and individuals are contributing to the effort by creating tools, languages and ontologies. One major contributor is DARPA—the folks responsible for a great deal of the technology behind the Internet (see “DARPA’s Disruptive Technologies,” TR October 2001). These days, DARPA is contributing tens of millions of dollars to the Web consortium’s Semantic Web project and has developed a semantic language for the U.S. Department of Defense called DARPA Agent Markup Language that allows users to add metadata to Web documents and relate it to ontologies. University of Maryland computer science professor Jim Hendler—who was until August manager of the DARPA program—has been working closely with Berners-Lee and Miller to ensure consistency with the consortium’s efforts. Last December, Hendler announced the creation of a language that combines the DARPA Agent Markup Language’s capabilities with an ontology language, developed in Europe, called OIL (which stands for both Ontology Inference Layer and Ontology Interchange Language). A developer of this new language, University of Manchester lecturer Ian Horrocks, also advises the World Wide Web Consortium on the Semantic Web. In January, he cofounded a company called Network Inference to develop technology that uses ontologies and automated inference to give Semantic Web capabilities to existing relational databases and large Web sites. Recently, an Isle of Man-based data services company called PDMS began using Network Inference’s technology to add Semantic Web capabilities to corporate databases. Dozens of other companies, from Hewlett-Packard to Nokia, are contributing to Semantic Web development (see “Spinning the Semantic Web,” below).

*EVEN IF BERNERS-LEE AND HIS COHORTS MEET THE TECHNICAL CHALLENGES, THAT WON'T BE ENOUGH FOR THE SEMANTIC WEB TO CLICK INTO PLACE. THERE IS A BIG QUESTION WHETHER PEOPLE WILL THINK THE BENEFITS ARE WORTH THE EXTRA EFFORT.*

## **Too Much, Too Late?**

Miller believes the seamless flow and integration of information resulting from these moves will make it possible to process knowledge in a way “that solves problems, brings people closer and spurs on new ideas that never could happen before.” Others, though, are not so optimistic about the Semantic Web. “It’s rather ambitious,” says R.V.Guha, who led development of the Web consortium’s Resource Description Framework efforts in the late 1990s. (This framework is an essential tool for describing and sharing metadata.) “It would be nice if such things existed,” he says, “but there are some really hard research problems that need to be solved first.” One issue concerns inference. The time it takes a computer to draw new conclusions from data, metadata and ontologies on the Web increases rapidly as rules are added to a system. Inference falls into the same category as the classic “traveling-salesman problem” of planning the shortest route through a number of cities. It’s not hard to figure out the best of all possible routes when you’re dealing with just a very few locations. But when you get up to only 15 cities, there are more than 43 billion possible routes. The same kind of runaway situation exists for inference, where brute-force searches for answers could lead to time-wasting paradoxes or contradictions. And even if Berners-Lee and his cohorts meet the technical challenges, that won’t be enough for the Semantic Web to click into place. There is a big question as to whether people will think the benefits are worth the extra effort of adding metadata to their content in the first place. One reason the Web became so wildly successful, after all, was its sublime ease of creation. “The Web today is the simplest, most primitive form of hypertext,” says former Sun Microsystems Distinguished Engineer Jakob Nielsen, cofounder of the Nielsen Norman Group, a Web design firm in Fremont, CA. “And that’s why it was so easy to implement; that’s why everybody could...start putting up their own Web pages; that’s why the Web is so big.” However, while most people may be comfortable doing simplistic editing, such as marking a text as “bold,” Nielsen points out, “They cannot do semantic editing, where they say, ‘This is the author’s name,’ or ‘This is the name of people I’m quoting.’” Of course, such pessimism may be ignoring recent history. Not so long ago, the notion of millions of people learning to write HTML code seemed farfetched —yet that’s exactly what happened. Still, the hurdle of creating a Semantic Web will be higher. People can use HTML any way they want. They commonly use tables for nontabular purposes, for instance, and slap on the “subhead” tag merely to apply boldface. These kluges and shortcuts usually have only cosmetic consequences. But the same type of fudging—say, by employing “bibliography” tags to list a DVD collection— could make a page’s metadata unusable. The fact that metadata wasn’t implemented right from the Web’s start could also make it harder for the Semantic Web to gain acceptance. One particularly tough skeptic is Peter Merholz, cofounder of Adaptive Path, a San Francisco-based user experience consultancy. “This stuff has to be baked in from the beginning,” says Merholz, who calls the Semantic Web “an interesting

academic pursuit” with little bearing on society. “The Semantic Web is getting a lot of hype simply because Tim Berners-Lee—the inventor of the World Wide Web—is so interested in it,” he says. “If it were just some schmuck at some university in Indiana, nobody would care.”

### **Initial Threads**

Even Berners-Lee admits that the path to the Semantic Web may be a bit slower than that to the World Wide Web. “In a way we don’t need to move too fast,” he says, “because the theory people need to look at it to make sure we’re not too crazy, and other people need to check out the ideas in practice before they’re picked up and used too extensively.” When asked to peek into his crystal ball, the evangelist of exchangeable data predicts that some of the Semantic Web’s first commercial applications will aim to integrate the different information systems that typically coexist in large organizations. (Wouldn’t it be nice to take care of business at the motor vehicle department or hospital without having to fill out a half-dozen largely redundant forms? The Semantic Web can help here.) And even though the Semantic Web still resides chiefly on the drawing board, you can see hints of its power on some existing Web sites. Consider Moreover Technologies’ search engine that crawls thousands of news sites several times a day, making it a favorite for news junkies. Moreover’s software agents have been programmed to look at the font tags (the HTML labels that tell Web browsers how large or small to make the text appear on the screen) to determine whether or not a particular page is a news story. If a Moreover agent finds a string of six to 18 words tagged as large type near the top of a page, it will assume it is a headline and place it in a database. Of course, since the agent is only making a guess, sometimes it selects a page that isn’t news after all. So Moreover has to apply additional filtering to get rid of pages that don’t contain articles. That’s still a far cry from the ultimate goal—but it’s a good start. And even the Semantic Web champions don’t pretend to grasp exactly where such steps will lead. After all, who predicted Amazon.com or eBay back when Berners-Lee turned on the switch of the world’s first Web server in December 1990? But the point is that people want more intelligence from the Web than they’re getting—and a growing number of computer scientists share the twinkle in Berners-Lee’s eye, and the feeling that the Semantic Web holds the answer. “It’s great,” says the inventor of the World Wide Web, “to have that grass-roots enthusiasm around again.”



***Maestro of metadata: Semantic Web project leader Eric Miller wants to link the Net’s information islands into a database Pangea.***

## Spinning the Semantic Web

A sampling of companies developing tools and applications for the Semantic Web

COMPANY	FOCUS
Aidministrador Nederland (Amersfoort, the Netherlands)	Software to classify sites based on content and relationships
CognIT (Halden, Norway)	Software to share information among different applications
Hewlett-Packard (Palo Alto, CA)	Java-based tools to create and maintain metadata
Intellidimension (Windsor,VT)	Development of databases with semantic properties
Invention Machine (Boston, MA)	Semantic searching tools
Network Inference (Isle of Man)	Software to create ontologies and inference engines
Nokia (Espoo, Finland)	Markup and ontology languages
Ontoprise (Karlsruhe,Germany)	Ontology-editing and inference software