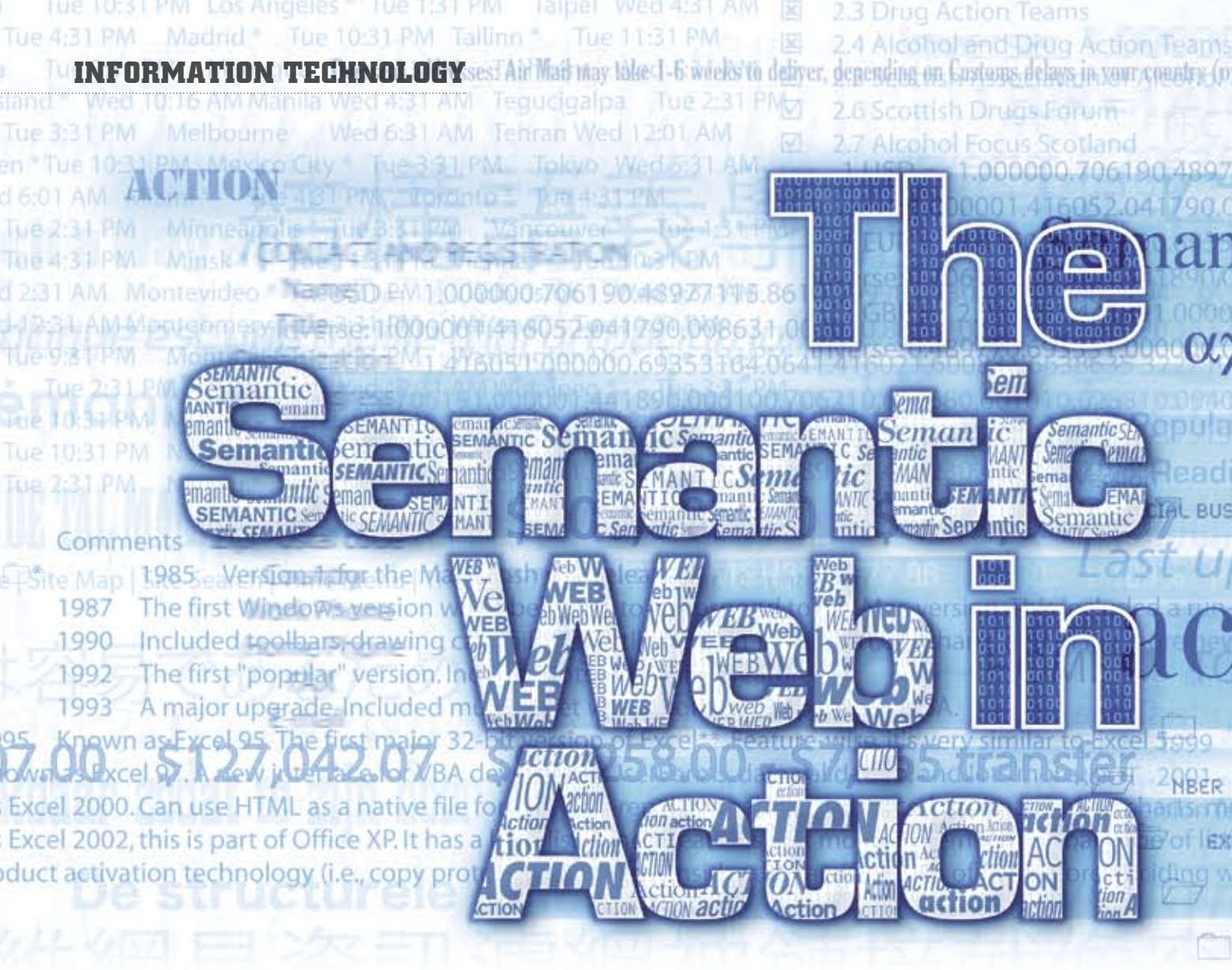
The Semantic Web in Action



KEY CONCEPTS

- A wide variety of online Semantic Web applications are emerging, from Vodafone Live!'s mobile phone service to Boeing's system for coordinating the work of vendors.
- Scientific researchers are developing some of the most advanced applications, including a system that pinpoints genetic causes of heart disease and another system that reveals the early stages of influenza outbreaks.
- Companies and universities, working through the World Wide Web Consortium, are developing standards that are making the Semantic Web more accessible and easy to use. *—The Editors*

Corporate applications are well under way, and consumer uses are emerging

Six years ago in this magazine, Tim Berners-Lee, James Hendler and Ora Lassila unveiled a nascent vision of the Semantic Web: a highly interconnected network of data that could be easily accessed and understood by any desktop or handheld machine. They painted a future of intelligent software agents that would head out on the World Wide Web and automatically book flights and hotels for our trips, update our medical records and give us a single, customized answer to a particular question without our having to search for information or pore through results.

They also presented the young technologies that would make this vision come true: a common language for representing data that could be understood by all kinds of software agents; ontologies—sets of statements—that translate information from disparate databases into common terms; and rules that allow software agents to reason about the information described in those terms. The data format, ontologies and reasoning software would operate like one big application on the World Wide Web, analyzing all the raw data stored in online databases as well as all the data about the text, images, video and communications the Web contained. Like the Web itself, the Semantic

SLIM FILMS

BY :: Lee Feigenbaum, Ivan Herman, Tonya Hongsermeier, Eric Neumann and Susie Stephens

Web would grow in a grassroots fashion, only this time aided by working groups within the World Wide Web Consortium, which helps to advance the global medium.

Since then skeptics have said the Semantic Web would be too difficult for people to understand or exploit. Not so. The enabling technologies have come of age. A vibrant community of early adopters has agreed on standards that have steadily made the Semantic Web practical to use. Large companies have major projects under way that will greatly improve the efficiencies of in-house operations and of scientific research. Other firms are using the Semantic Web to enhance business-to-business interactions and to build the hidden data-processing structures, or back ends, behind new consumer services. And like an iceberg, the tip of this large body of work is emerging in direct consumer applications, too.

Just below the Surface

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

se•man•tic web
[si-'man-tik 'wĕb]
—noun

A set of formats and languages that find and analyze data on the World Wide Web, allowing consumers and businesses to understand all kinds of useful online information.

Combining Concepts

Search engines on the World Wide Web cannot provide a single answer to a broad-ranging question such as "Which television sitcoms are set in New York City?" But a new Semantic Web engine called pediax can, by analyzing different concepts (*top, in approximated form*) found on Wikipedia's seven million online pages. Pediax, which grew from the DBpedia project to extract information from Wikipedia, provides a clean result (*bottom*) that merges text and images.

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.

Perhaps the most visible examples, though limited in scope, are the tagging systems that have flourished on the Web. These systems include del.icio.us, Digg and the DOI system used by publishers, as well as the sets of custom tags available on social sites such as MySpace and Flickr. In these schemes, people select common terms to describe information they find or post on certain Web sites. Those efforts, in turn, enable Web programs and browsers to find and crudely understand the tagged information—such as finding all Flickr photographs of sunrises and sunsets taken along the coast of the Pacific Ocean. Yet the tags within one system do not work on the other, even when the same term, such as "expensive," is used. As a result, these systems cannot scale up to analyze all the information on the Web.

The World Wide Web Consortium—an ad hoc organization of more than 400 companies and universities co-hosted by the Massachusetts Institute of Technology in the U.S., the European Consortium for Informatics and Mathematics in France, and Keio University in Japan—has already released the Semantic Web languages and technologies needed to cross such boundaries, and large companies are exploiting them. For example, British Telecom has built a prototype online service to help its many vendors more effectively develop new products together. Boeing is exploring the technologies to more efficiently integrate the work of partners involved in airplane design. Chevron is experimenting with ways to manage the life cycle of power plants and oil refineries. MITRE Corporation is applying Semantic Web tool kits to help the U.S. military interpret rules of engagement for convoy movements. The U.K.'s national mapping agency, Ordnance Survey, uses the Semantic Web internally to more accurately and inexpensively generate geographic maps.

Other companies are improving the back-end operations of consumer services. Vodafone Live!, a multimedia portal for accessing ring tones, games and mobile applications, is built on Semantic Web formats that enable subscribers to download content to their phones much faster than before. *Harper's Magazine* has harnessed semantic

LUCY READING-IKKANDA (*illustration*); AGILE KNOWLEDGE ENGINEERING AND SEMANTIC WEB (*screen shot*); COURTESY OF APPLE, INC. (*laptop*)

ontologies on its Web site to present annotated timelines of current events that are automatically linked to articles about concepts related to those events. Joost, which is putting television on the Web for free, is using Semantic Web software to manage the schedules and program guides that viewers use online.

Consumers are also beginning to use the data language and ontologies directly. One example is the Friend of a Friend (FOAF) project, a decentralized social-networking system that is growing in a purely grassroots way. Enthusiasts have created a Semantic Web vocabulary for describing people's names, ages, locations, jobs and relationships to one another and for finding common interests among them. FOAF users can post information and imagery in any format they like and still seamlessly connect it all, which MySpace and Facebook cannot do because their fields are incompatible and not open to translation. More than one million individuals have already interlinked their FOAF files, including users of LiveJournal and TypePad, two popular Weblog services.

As these examples show, people are moving toward building a Semantic Web where relations can be established among any online pieces of information, whether an item is a document, photograph, tag, financial transaction, experiment result or abstract concept. The data language, called Resource Description Framework (RDF), names each item, and the relations among the items, in a way that allows computers and software to automatically interchange the information. Additional power comes from ontologies and other technologies that create, query, classify and reason about those relations [*see box on page 95*].

The Semantic Web thus permits workers in different organizations to use their own data labels instead of trying to agree industry-wide on one rigid set; it understands that term "X" in database 1 is the same as term "Y" in database 2. What is more, if any term in database 1 changes, the other databases and the data-integration process itself will still understand the new information and update themselves automatically. Finally, the Semantic Web enables the deployment of "reasoners"—software programs that can discover relations among data sources.

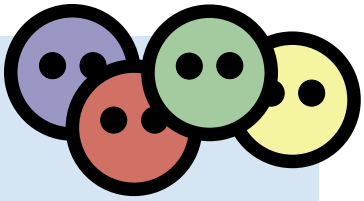
Just as the HTML and XML languages have made the original Web robust, the RDF language and the various ontologies based on it are maturing, and vendors are building applications based on them. IBM, Hewlett-Packard and Nokia are promoting open-source Semantic Web frameworks—common tools for crafting polished programs. Oracle's flagship commercial database, 10g, used by thousands of corporations worldwide, already supports RDF, and the upgrade, 11g, adds further Semantic Web technology. The latest versions of Adobe's popular graphics programs such as Photoshop use the same technologies to manage photographs and illustrations. Smaller vendors—among them Aduna Software, Altova, @semantics, Talis, OpenLink Software, TopQuadrant and Software AG—offer Semantic Web database programs and ontology editors that are akin to the HTML browsers and editors that facilitated the Web's vibrant growth. Semantic Web sites can now be built with virtually all of today's major computer programming languages, including Java, Perl and C++.

We are still finding our way toward the grand vision of agents automating the mundane tasks of our daily lives. But some of the most advanced progress is taking place in the life sciences and health care fields. Researchers in these disciplines face tremendous data-integration challenges at almost every stage of their work. Case studies of real systems built by these pioneers show how powerful the Semantic Web can be.

Case Study 1: Drug Discovery

The traditional model for medicinal drugs is that one size fits all. Have high blood pressure? Take atenolol. Have anxiety? Take Valium. But because each person has a unique set of genes and lives in a particular physical and emotional environment, certain individuals will respond better than others. Today, however, a greater understanding of biology and drug activity is beginning to be combined with tools that could predict which drugs—and what doses—will work for a given individual. Such predictions should make custom-tailored, or personalized, medical treatments increasingly possible.

FRIEND OF A FRIEND



Users of a grassroots, semantic social network system—Friend of a Friend—have created a vocabulary that describes the personal information they want to post and finds common interests. The network (*logo shown*) can also integrate information from isolated, commercial systems such as MySpace and Facebook. See www.foaf-project.org

[THE AUTHORS]

All five authors have participated in various projects to develop Semantic Web technologies. **Lee Feigenbaum,** formerly at IBM, is vice president of technology and standards at Cambridge Semantics, Inc. **Ivan Herman** leads the Semantic Web Activity initiative at the World Wide Web Consortium. **Tonya Hongsermeier** is corporate manager of clinical knowledge management and decision support at Partners Healthcare System. **Eric Neumann** is executive director of Clinical Semantics Group Consulting. **Susie Stephens** was principal product manager at Oracle Corporation and has recently become principal research scientist at Eli Lilly and Company.

WWW.FOAF-PROJECT.ORG

Personalized medicine will become possible only when semantics makes medical databases smarter and easier to use.

The challenge, of course, is to somehow meld a bewildering array of data sets: all sorts of historic and current medical records about each person and all sorts of scientific reports on a number of drugs, drug tests, potential side effects and outcomes for other patients. Traditional database tools cannot handle the complexity, and manual attempts to combine the databases would be prohibitively expensive. Just maintaining the data is difficult: each time new scientific knowledge is incorporated into one data source, others linked to it must be reintegrated, one by one.

A research team at Cincinnati Children's Hospital Medical Center is leveraging semantic capabilities to find the underlying genetic causes of cardiovascular diseases. Traditionally, researchers would search for genes that behave differently in normal and diseased tissues, assuming that these genes could somehow be involved in causing the pathology. This exercise could yield tens or hundreds of suspect genes. Researchers would then have to pore through four or five databases for each one, trying to discern which genes (or the proteins they encode) have features most likely to affect the biology of the disorder—a painstaking task. In the end, investigators often cannot afford the hours, and the work falters.

The Cincinnati team, which includes a Semantic Web consultant, began by downloading into a workstation the databases that held relevant information but from different origins and in incompatible formats. These databases included Gene Ontology (containing data on genes and gene products), MeSH (focused on diseases and symptoms), Entrez Gene (gene-centered information) and OMIM (human genes and genetic disorders). The investigators translated the formats into RDF and stored the information in a Semantic Web database. They then used Protégé and Jena, freely available Semantic Web software from Stanford University and HP Labs, respectively, to integrate the knowledge.

The researchers then prioritized the hundreds of genes that might be involved with cardiac function by applying a ranking algorithm somewhat similar to the one Google uses to rank Web pages of search results. They found candidate genes that could potentially play a causative role in dilated cardiomyopathy, a weakening of the heart's pumping ability. The team instructed the software to evaluate the ranking information, as well as the genes' relations to the characteristics and symptoms of the condition and similar diseases. The software identified four genes with a strong connection to a chromosomal region implicated in dilated cardiomyopathy. The researchers are now investigating the effects of these genes' mutations as possible targets for new therapeutic treatments. They are also applying the semantic system to other cardiovascular diseases and expect to realize the same dramatic improvement in efficiency. The system could also be readily applied to other disease families.

Similarly, senior scientists at Eli Lilly are applying Semantic Web technologies to devise a complete picture of the most likely drug targets for a given disease. Semantic tools are allowing them to compile numerous incompatible biological descriptions into one unified file, greatly expediting the search for the next breakthrough drug. Pfizer is using Semantic Web technologies to mesh data sets about protein-protein interaction to reveal obscure correlations that could help identify promising medications. Researchers there are convinced that these technologies

[ANALYZING DATABASES]

Which Genes Cause Heart Disease?

Hundreds of genes could potentially contribute to heart disease. Researchers at Cincinnati Children's Hospital Medical Center are using Semantic Web tools to find the most likely culprits by analyzing numerous online databases and scientific references (*left, on screen*), revealing possible causative connections (*right, on screen*). For example, they have pinpointed suspect genes related to a chromosomal region linked to dilated cardiomyopathy, a weakening of the heart's pumping ability.

CHANDRA GUDIVADA, ANIL JEGGA, ERIC BARDES, SCOTT TABAR AND BRUCE ARONOW *Cincinnati Children's Hospital Medical Center (screen shot)*; COURTESY OF APPLE, INC. (*monitor*)

will increase the chance for serendipitous discoveries, accelerate the speed of delivering new drugs to market and advance the industry as a whole toward personalized medicine. "This is where the Semantic Web could help us," says Giles Day, head of Pfizer's Research Technology Center informatics group in Cambridge, Mass.

In each of these cases, the Semantic Web enhances drug discovery by bringing together vast and varied data from different places. New consumer services are being built in similar fashion. For example, the British firm Garlik uses Semantic Web software to compare previously incompatible databases to alert subscribers that they might be the target of an identity thief. Garlik culls disparate personal identity information from across the Web, integrates it using common vocabularies and rules, and presents subscribers with a clear (and sometimes surprising) view of their online identity.

Case Study 2: Health Care

The health care industry confronts an equally dense thicket of information. One initiative that has been deployed since 2004 was developed at the University of Texas Health Science Center at Houston to better detect, analyze and respond to emerging public health problems. The system, called SAPPHIRE (for *s*ituational *a*wareness and *p*reparedness for *p*ublic *h*ealth *i*ncidences using *r*easoning *e*ngines), integrates a wide range of data from local health care providers, hospitals, environmental protection agencies and scientific literature. It allows health officials to assess the information through different lenses, such as tracking the spread of influenza or the treatment of HIV cases.

Every 10 minutes in the greater Houston area, SAPPHIRE receives reports on emergency room cases, descriptions of patients' self-reported symptoms, updated electronic health records, and clinicians' notes from eight hospitals that account for more than 30 percent of the region's emergency room visits. Semantic technologies integrate this information into a single view of current health conditions across the area. A key feature is an ontology that classifies unexplained illnesses that present flulike symptoms (fevers, coughs and sore throats) as potential influenza cases and automatically reports them to the Centers for Disease Control and Prevention. By automatically generating reports, SAPPHIRE has relieved nine nurses from doing such work manually, so they are available for active nursing. And it delivers reports two to

[HOW IT WORKS]

Making the Semantic Web Tick

Several formats and languages form the building blocks of the Semantic Web. They extend similar software technologies that underlie the World Wide Web itself and have been published as standards by the World Wide Web Consortium's Semantic Web Activity initiative.

:: RDF FORMAT. The most fundamental building block is Resource Description Framework (RDF), a format for defining information on the Web. Each piece of data, and any link that connects two pieces of data, is identified by a unique name called a Universal Resource Identifier, or URI. (URLs—the common Web addresses that we all use, are special forms of URIs.) In the RDF scheme, two pieces of information, and any notation indicating how they are connected, are grouped together into what is called a triple. For example, an online reference to the famous television animal "Flipper," a reference to the relationship "is a," and a reference to the concept of "dolphin" could be joined in the triple shown below.

< uri for **Flipper** > < uri for **Is A** > < uri for **Dolphin** >

URIs can be agreed on by standards organizations or communities or assigned by individuals. The relation "is a" is so generally useful, for example, that the consortium has published a standard URI to represent it. The URI "http://en.wikipedia.org/wiki/Dolphin" could be used by anyone working in RDF to represent the concept of dolphin. In this way, different people working with different sets of information can nonetheless share their data about dolphins and television animals. And people everywhere can merge knowledge bases on large scales.

:: ONTOLOGY LANGUAGES. Individuals or groups may want to define terms and data they frequently use, as well as the relations among those items. This set of definitions is called an ontology. Ontologies can be very complex (with thousands of terms) or very simple. Web Ontology Language (known as OWL) is one standard that can be used to define ontologies so that they are compatible with and can be understood by RDF.

:: INFERENCE ENGINES. Ontologies can be imagined as operating one level above RDF. Inference engines operate one level above the ontologies. These software programs examine different ontologies to find new relations among terms and data in them. For example, an inference engine would examine the three RDF triples below and deduce that Flipper is a mammal. Finding relations among different sources is an important step toward revealing the "meaning" of information.

< uri for **Flipper** > < uri for **Is A** > < uri for **Dolphin** >

< uri for **Dolphin** > < uri for **Subclass Of** > < uri for **Mammal** >

< uri for **Flipper** > < uri for **Is A** > < uri for **Mammal** >

:: OTHER TECHNOLOGIES. The Web consortium is crafting inference engines as well as many other technologies. One is SPARQL, a query language that allows applications to search for specific information within RDF data. Another is GRDDL, which allows people to publish data in their traditional formats, such as HTML or XML, and specifies how these data can be translated into RDF. For more, see www.w3.org/2001/sw

If two databases joined by the Semantic Web have different privacy criteria, the software will have to honor both sets of rules.

three days faster than before. The CDC is now helping local health departments nationwide to implement similar systems, replacing tedious, inconsistent and decades-old paper schemes.

The nimbleness of Semantic Web technologies allows SAPPHIRE to operate effectively in other contexts as well. When Hurricane Katrina evacuees poured into Houston's shelters, public health officials quickly became concerned about the possible spread of disease. Within eight hours after the shelters were opened, personnel at the University of Texas Health Science Center configured SAPPHIRE to help. They armed public health officials with small handheld computers loaded with health questionnaires. The responses from evacuees were then uploaded to the system, which integrated them with data from the shelters' emergency clinics and surveillance reports from Houston Department of Health and Human Services epidemiologists in the field. SAPPHIRE succeeded in identifying gastrointestinal, respiratory and conjunctivitis outbreaks in survivors of the disaster much sooner than would have been possible before.

SAPPHIRE's flexibility showcases an important lesson about Semantic Web systems: once they are configured for a general problem—in this case, public health reporting—they can quickly be adapted to a variety of situations within that field. Indeed, the CDC would like to roll out a single, integrated, SAPPHIRE-style illness alert system nationwide.

SAPPHIRE succeeds because it can unify information from many places, which can then be used for different goals. This same attribute is fueling FOAF's grassroots growth. By using an agreed-on Semantic Web vocabulary, the FOAF system finds common interests among friends and acquaintances, even if they do not belong to the same social-networking sites such as MySpace or Facebook. FOAF enthusiasts are also now developing semantic trust networks—white lists of trusted senders—as a way to fight e-mail spam.

Crossing Boundaries

The success of SAPPHIRE and other applications has prompted calls for more Semantic Web integration in health care. The Food and Drug Administration and the National Institutes of Health have both recently declared that a shift toward research into translating data across boundaries is necessary for improving the drug development and delivery process.

The same work will enhance the traditional computerized clinical decision support (CDS) systems that medical professionals use—knowledge bases that contain the latest wisdom on therapeutic treatments. Each hospital, physicians' network and insurance company has had to custom-design its own system, and all of them are struggling mightily to stay current. Every time an advance is made about diagnoses, clinical procedures or drug safety—which is often—administrators must rework their systems. The personnel time required is usually far greater than most of these organizations can afford. Furthermore, because the custom systems are frequently incompatible, making industry-wide insights or deciphering best practices is slow and cumbersome. What is more, "we are investigating Semantic Web technologies because traditional approaches for data integration, knowledge management and decision support

[UNIFYING INFORMATION]

Is a Flu Outbreak Under Way?

Public health officials take longer than they would like to recognize new disease outbreaks, because they must manually compare disparate reports in incompatible formats from many hospitals and doctors' offices. Researchers at the University of Texas Health Science Center have built a Semantic Web system that quickly and automatically tracks and analyzes these online data across the Houston area. It presents officials with clear trends, such as the incidence of flu symptoms across different age groups over time (*center, on screen*); a sharp rise would indicate early signs of outbreaks.

COURTESY OF THE CENTER FOR BIOSECURITY AND PUBLIC HEALTH INFORMATICS RESEARCH, UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON (*screen shot*); COURTESY OF DELL, INC. (*laptop*)

will not scale to what is needed for personalized medicine," says John Glaser, chief information officer at Partners HealthCare System in Boston.

To remedy this situation, Agfa HealthCare has constructed a prototype CDS system based on Semantic Web technologies. When a person inputs a change into one part of a system, records that should be altered in other parts of the system or in the systems of another institution are automatically updated. For example, Agfa's prototype transforms standard radiology protocols into Semantic Web notation and integrates them with other common knowledge bases, such as clinical guidelines from medical societies. Institutions can maintain their own internally standardized content, yet end users such as hospitals can readily integrate new content, greatly reducing the labor hours required.

As systems such as Agfa's are implemented across the health care network, medical knowledge bases will become smarter, easier and less expensive to use. Imagine a patient who is prone to blood clots and has a genetic mutation that, according to current medical literature, will respond well to a new anticlotting medication. Over the ensuing months, however, new studies show that particular variants of this mutation actually cause that same drug to increase clotting. This patient's clinician must be notified to change the therapy for anyone with this variant. How could notifications such as this be effectively carried out given that thousands of genes are involved in hundreds of diseases across millions of patients? Meeting this challenge will not be possible without robust semantic approaches.

Daily Life, Too

The same Semantic Web technologies that are transforming drug discovery and health care are being applied to more general situations. One example is Science Commons, which helps researchers openly post data on the Web. The nonprofit organization provides Semantic Web tools for attaching legally binding copyright and licensing information to those data. This capability allows a scientist, for example, to instruct a software applet to go find information about a particular gene—but only information that comes with a free license.

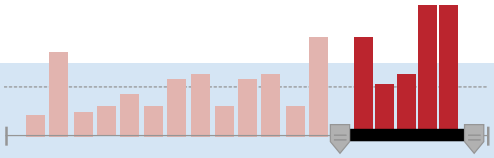
DBpedia is an effort to smartly link information within Wikipedia's seven million articles. This project will allow Web surfers to perform detailed searches of Wikipedia's content that are impossible today, such as, "Find me all the films nominated for a Best Picture Academy Award before 1990 that ran longer than three hours."

As applications develop, they will dovetail with research at the Web consortium and elsewhere aimed at fulfilling the Semantic Web vision. Reaching agreement on standards can be slow, and some skeptics wonder if a big company could overtake this work by promoting a set of proprietary semantic protocols and browsers. Perhaps. But note that numerous companies and universities are involved in the consortium's semantic working groups. They realize that if these groups can devise a few well-designed protocols that support the broadest Semantic Web possible, there will be more room in the future for any company to make money from it.

Some observers also worry that people's privacy could become compromised as more data about them from disparate sources is interlinked. But Semantic Web advocates argue that the protections are the same as those used in the nonlinked world. If two databases joined by the Semantic Web have different privacy criteria, then the software will have to honor both sets of rules or create a set that covers both. When SAPPHIRE joins patient databases, it adheres to the privacy requirements of both or it won't proceed; the nurses who had formerly performed the same mergers manually imposed the same practice.

The Semantic Web will probably operate more behind the scenes than the World Wide Web does. We won't see how it helps Eli Lilly create personalized drugs; we'll just buy them. We won't know how Vodafone makes cool ring tones so readily available, but we'll appreciate how easy they are to download. And yet, soon enough the Semantic Web will give more direct power to us, too, allowing us to go on eBay and not just say "find me the Toyota Priuses for sale" but "find me only used, red Priuses for sale for less than $14,000 by people who are within 80 miles of my house and make them an offer." Grand visions rarely progress exactly as planned, but the Semantic Web is indeed emerging and is making online information more useful than ever. ■

BLOG ANALYZER



Oracle Technology Network has demonstrated a Semantic Web site that can analyze blogs, podcasts and discussion groups to find related commentary about specific topics. It also can produce visualizations of its findings, such as tag clouds (*below*) that show whose blogs are drawing the most traffic (*larger names*) and bar charts that identify the most concentrated discussions. Project details are available at http://otnsemanticweb.oracle.com

MORE TO EXPLORE

The Semantic Web. Tim Berners-Lee, James Hendler and Ora Lassila in *Scientific American,* Vol. 284, No. 5, pages 34–43; May 2001.

Books about the Semantic Web are described at **http://esw.w3.org/topic/SwBooks**

Case studies of how companies and research groups are applying the Semantic Web can be found at **www.w3.org/2001/sw/sweo/public/UseCases**

Guides to RDF are indexed at **http://planetrdf.com/guide**, and tools to develop Semantic Web pages are available at **http://esw.w3.org/topic/SemanticWebTools**

Related blogs and RSS feeds can be accessed at **http://planetrdf.com**

COURTESY OF ORACLE