COMPUTER SCIENCE

New Language Could Meld the Web Into a Seamless Database

The World Wide Web, circa 1998: A chemist wants to know how to perform a certain reaction. She does a literature search on the Web and copies down by hand the chemicals she needs. She does another search to find the vendors who sell those chemicals and jots down their names. She sends a purchase order to Company A, which pays a clerk to enter the order by hand.

The Web, circa 2000: The chemist enters some words describing the reaction. An intelligent Web agent finds it and asks her if she wants to order the reagents. She clicks "Yes," and the Web agent takes care of the rest. A few days later, she receives her chemicals.

The key to eliminating the human intervention between the chemist and her chemicals is a new Web invention called extensible markup language. XML, developed by a group of technology companies and universities called the World Wide Web Consortium (W3C), goes beyond merely displaying data—the strength of hypertext markup language (HTML), the language that currently dominates the Web—to making it meaningful for other computers.

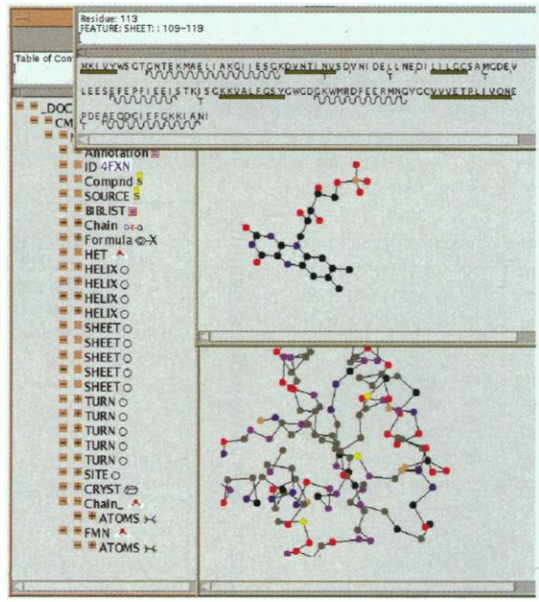
In HTML, for example, $CaCO_3$ is nothing more than a set of letters on a screen. But in XML and related languages, a "tag" can identify it as the chemical symbol for a compound. On another Web page, the same tag might go with the words "calcium carbonate," and on a third, with "Catalog Item No. 1311." An intelligent search engine can tell from the tags that they represent the same compound. Then it can download the accompanying data on, say, chemical properties or bulk prices, in a form ready for the user's software to manipulate.

Earlier this year, the W3C approved the first version of XML, and software companies are working on XML-capable browsers. Ultimately, say Web developers, the language could open the way to "a new Internet" that would be easier to search and exploit, offer more flexible formatting of documents, and might even usher in the ballyhooed age of electronic commerce. As Tim Berners-Lee, the creator of the Web, told the *Los Angeles Times*, "Whereas phase one of the Web puts all the accessible information into one huge book, if you like, in phase two we will turn all the accessible data into one huge database." For scientists, says Peter Murray-Rust of the University of Nottingham in England, a contributor to XML, "I think it will make a major difference. ... For the first time we've got something capable of managing most of the information we deal with."

The success of XML, however, will hinge on the ability of professional societies and others to agree on what kind of information they want to share and how that information is to be structured, because XML offers far more latitude to its users than HTML does. Like any markup language, HTML annotates text with certain instructions or "tags," which a computer program, such as a Web browser, can spot and act upon. Designed to serve as a simple lingua franca for the Web, HTML has a very small set of tags, all of which (at least in early versions of HTML) refer only to the display of text. The tags <H1> and <H2>, for example, tell a browser to render the ensuing text as a main heading or a subheading.

By contrast, XML allows groups of users to define their own tags. It does this by prescribing only the syntax of a file (how tags look and how they are used), and not the semantics (what they mean). The tags can, for example, contain information about the text, rather than just about how to display it. In an online bibliography, the title and author of a book could be tagged as <TITLE> and <AUTHOR>, instead of <H1> and <H2>. When a user searched for <AUTHOR> Gates, he would find Bill Gates's *The Road Ahead*, without wading through spurious links to transistor gates, garden gates, or even articles about Bill Gates.

The new language is actually an adaptation of SGML (for standard generalized markup language), which was developed in the 1980s for technical communication but was vastly too complicated for wide use on the Web. In 1996, Jon Bosak of Sun Microsystems formed a working group in the W3C to simplify SGML for Web use. Although the committee originally envisioned the language as an "SGML Lite," the members eventually realized it was different enough to deserve its own name. Companies that develop Web browsers are now implementing XML in their products. Microsoft's Internet Explorer, for example, already contains an XML parser—a program that separates tags from text—as does the publicly available source code for Netscape Communicator. Parsers make it possible to transfer information from one database to another, but they can't display an XML file on screen. The W3C is now drawing up a format for the additional applications needed to display XML files: "style sheets," which turn tags into formatting instructions.



SOURCE: P. MURRAY-RUST

Information on demand. Chemical markup language assembles data from different sources into a single interactive document describing a protein molecule.

Of course, no style sheet would be able to handle the Babel of information that would result if users all made up their own tags. For XML to work, common-interest groups will have to agree on a shared vocabulary. "The technology is the small part," says Dan Connolly of the W3C, one of the designers of the new language. "The large part is getting people together to agree."

That has already happened in a few cases. Mathematicians, who have long been hamstrung by HTML's inability to handle complex equations, can now post their pages in MathML, a rather straightforward application developed by the W3C in which the XML tags represent mathematical formatting instructions. A more ambitious XML offshoot called chemical markup language (CML) is the brainchild of Murray-Rust, who is a crystallographer as well as a Web expert. CML enables the viewer's computer to meld information stored in separate databases into a seamlessly linked, interactive document, made to order. By clicking on different parts of a CML page on a certain protein, for example, a user can call up windows showing its molecular structure, its sequence, and the structure of a ligand it attaches to. A click on the sequence in one window then lights up the corresponding part of the molecular model in another.

Meanwhile, computer scientists are working on the smart browsers and souped-up search engines that will make the fullest use of XML. Such "intelligent agents" would be able to answer queries current search engines can't touch: "Is there a university in a state bordering Virginia with an ROTC program, Japanese classes, and a Computational Biology major?" The answer (University of Maryland) happens to be where one of the first such search engines resides. The experimental browser, developed by computer scientist James Hendler, can make the necessary connections because it works with an advanced markup language in which the XML tags indicate not only meanings but relationships between entities (universities are located in states, and majors are found at universities).

Even XML aficionados don't expect to see these kinds of tags popping up on every Web site. For displaying ordinary text documents, HTML is likely to remain the standard, and XML-capable browsers will still be able to read pages written in HTML. But they believe that for specialized Web applications—in science, for example—XML will quickly make converts. "People predicted the Web would fail because no one would want to learn HTML," says Tim Finin, a computer scientist at the University of Maryland, Baltimore County. The pessimists were wrong, he notes, and "the same thing will happen with XML."

–Dana Mackenzie

Dana Mackenzie is a science and math writer in Santa Cruz, California.

GENOMICS RESEARCH

Year of the Cat—in More Ways Than One

Far from the limelight shining on the human and mouse genome projects, researchers have also been laboring on the genomes of a half-dozen other mammals. One of these efforts, the Feline Genome Project, is about to hit a major milestone. Researchers at the National Cancer Institute's (NCI's) Laboratory of Genomic Diversity in Frederick, Maryland, expect to complete a genetic map this year—appropriately, the Chinese Year of the Cat.

The immediate goal of the NCI effort, which will cost about $3 million, sounds modest: a map with about 950 markers spread across the roughly 3 billion bases in the cat genome. That's far less detailed than the human and mouse maps, each with more than 20,000 markers on genomes that also contain about 3 billion bases. And there are no plans to undertake the massive job of determining the complete cat genome sequence. But the cat map could nevertheless turn out to be a useful guide to human genetic diseases.

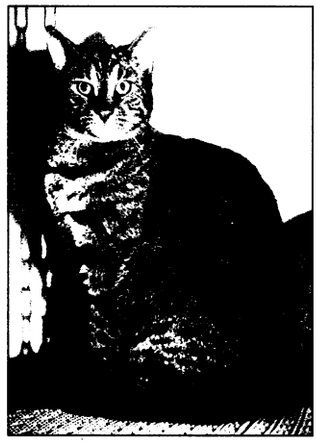
Cats and humans share almost 60 inherited diseases, including polycystic kidney disease, diabetes, heart muscle disorders, and certain common immune cell cancers. Once the cat map is in hand, the NCI group plans to use it to track down the cat disease genes and then mine the comparable regions of the human and mouse genomes for candidates for human disease genes.

If the same genes turn out to be at fault in both species, then cats would also provide good models for the human diseases. Animals like cats and dogs offer a "tremendous, rich resource of genetic diseases that can't be studied in mice," says Don Patterson, director of the Center for Comparative Medical Genetics at the University of Pennsylvania School of Veterinary Medicine.

Stephen O'Brien, chief of the NCI team, began mapping the cat genome 20 years ago because he thought that it might help him find cat genes that regulate the effects of a cancer-causing feline virus. When his first map attempt (*Science*, 16 April 1982, p. 257) revealed that gene arrangements on human and cat chromosomes are very similar—much more so than those on human and mouse chromosomes—O'Brien quickly grasped the potential for "comparative mapping" between humans and cats. He has been working to complete the feline map ever since.

For the bulk of the mapping effort, the O'Brien group is using standard crossbreeding experiments—an effort aided by the availability of Asian leopard cats that the NCI team borrowed from the National Zoo in Washington, D.C., and bred with domestic cats to produce a handsome cross known as the Bengal. The leopard cats "were in the right place at the right time," notes NCI geneticist Leslie Lyons.

By following the inheritance pattern of cat genes in blood samples taken from crosses through three generations, the researchers can establish their relative positions on the map, because genes that are located near one another on chromosomes tend to be inherited together. Breeding domestic cats with Asian leopard cats helped, because the genes of the two species are sufficiently different that tracking cat genes to the third-generation crosses can be readily accomplished.



LESLIE LYONS

Cat tracks. Tracing genes in hybrids, such as this Asian leopard–domestic cat cross, is helping to produce a cat genome map.

To fill some holes and add detail, the NCI team also applied a newer technique—radiation hybrid mapping—used for human genome mapping. NCI geneticist Bill Murphy irradiated cat cells, breaking apart their chromosomes, and then fused the cells with hamster cells. The cat chromosome fragments integrate into the hamster chromosomes in these hybrid cells, which are then tested for the presence of cat DNA markers. The nearer two markers are on a cat chromosome, the greater the likelihood that they will end up on the same fragment—and in the same cell.

Together, the crossbreeding and radiation mapping techniques will enable the NCI team to create a map showing the relative positions of about 350 genes. But although genes are needed to match to the corresponding genes in other species, they don't vary enough from one individual to another to be used as markers for tracking down unknown genes. So again using blood samples from the crossbred cats, geneticist Marilyn Menotti-Raymond is mapping cat microsatellite markers, short segments of repeating DNA that are variable enough for tracking down genes for diseases and other traits. When it's ready, the cat map will include about 600 microsatellite markers.

The next step will be to use the map to pin down genes that cause cat—and presumably also human—diseases. Lyons and Menotti-Raymond have already collected cat families with cancer, retinal atrophy, and polycystic kidney disease that they will use to track down human gene counterparts. Patterson expects the cat to be a better mirror of complex human diseases than lab mice, which he says are "essentially manmade organisms" that have lost many recessive traits and diseases through inbreeding. But the usefulness of cats as models of human disease remains to be shown.

O'Brien also expects benefits from comparing the genetics of the 37 cat species, which range from the tiny sand cat to the majestic lion. "Less than a tenth of 1% of all mammalian species that have existed survive," he notes. O'Brien believes the survivors' genes contain disease prevention secrets preserved by natural selection that could be identified by these studies. "We want to find out the things they have conjured up," he says, "because we might not be as clever ourselves."

–Ken Garber

Ken Garber is a science and health writer in Ann Arbor, Michigan.