Welcome to the Semantic Web

Tim Berners-Lee

from *The Economist* "The World in 2007"

Data integration will be the web's next leap forward, predicts **Tim Berners-Lee**, inventor of the World Wide Web and director of the World Wide Web Consortium.

Today, digital information about nearly every aspect of our lives is being created at an astonishing rate. Hidden amid all of these data is the key to knowledge about how to cure diseases, make more money and govern our world more effectively. Yet the technical tools and social practices that shape the way we manage, share, integrate and analyze this under-used treasure trove are sorely out of date. The good news is that a number of technical innovations (with names like AJAX, XML, RDF and OWL) along with new social arrangements regarding data are advancing the World Wide Web towards what we call the Semantic Web.

Progress towards better data integration will happen using the same basic technology that has made the World Wide Web so successful: the link. The power of the web today, including the ability to find information quickly, derives from the fact that people publish documents in standard formats, and then link them together. The value of the web increases in more than linear fashion with the number of links; this has been called the "network effect". The Semantic Web will derive its power in a similar way, but through the linking of data rather than documents.

To appreciate the need for better data integration, compare the enormous volume of experimental data produced in commercial and academic pharmaceutical laboratories around the world with the frustratingly slow pace of drug discovery. Life-science researchers are coming to the conclusion that in many cases no single lab, library or genomic data repository contains the information necessary to discover new drugs. Rather, the information needed to understand the complex interactions between diseases, biological processes and the vast array of chemical agents is spread across disparate databases, spreadsheets and documents.

As a result, progress towards better drug discovery depends on technologies that enable sharing and integration of data, as well as on changes in institutional practices in order to allow exploration of the links in these data. This is not to suggest that all pharmaceutical companies simply free their data, but rather that they explore more flexible licensing models that allow greater value to be created through the combination of their own intellectual property and that of others.

For this sort of integration to happen, an essential technical step is to publish the data using Semantic Web standards (RDF, OWL, SPARQL), and to link them together with definitions of the terms used to express the data. For example, when publishing experimental results about the behavior of a particular chemical in a larger biological process, one must indicate which vocabularies are being used to describe the biological pathway and the chemical. Then, when someone else wishes to integrate those data-for example, with other experiments documented in the research literature-that person can use that same vocabulary to match article keywords to chemical names.

There's money in it too

Scientists are not the only ones who need and will benefit from better data integration. Consider the financial-services industry. Successful investment strategies are based on finding patterns and trends in an increasingly diverse set of information sources (news, market data, historical trends, commodity prices). Leading-edge providers of financial information are now developing services that a1low users easily to integrate the data they have themselves-about their own portfolios or from their in-house market models-with the data delivered by the information service. The unique value creation lies in the integration service itself, not in the raw data on its own or even in the software tools, most of which will be built on open-source components.

The key to this integration is to use common data formats that link the information with identifiable vocabularies. The Semantic Web does not require that everyone use the same vocabulary, any more than we can expect the entire world to speak a single language. Instead, it provides tools for the translation from one set of terms to another. These translations will integrate larger and larger collections of information across the web.

Semantic Web technology can provide significant benefits within particular user communities. But the most exciting discoveries will come from the serendipitous combination and integration of data drawn from diverse sources. The investment house that has taken the Semantic Web plunge, for example, will find that if it is interested in analysing the market value of drug companies based on their drug-discovery success, it will be possible to combine market-research data held by the investment company with evidence of the contribution that different companies have made to ongoing research.

As with the first phase of the World Wide Web, innovation and new value creation will depend as much on changing institutional practices as on the simple development of these new technologies. The technology is necessary, of course, but not sufficient to reap the potential benefits of data integration.

So readers should take the following homework assignment for 2007 from this article: make an inventory of the data you are responsible for and think about which parts would be most likely to be re-used if you were to share them on a corporate intranet or on the internet using Semantic Web standards. Just as in the early days of the web ten years ago, the advances in data integration will benefit all those who contribute, often in unexpected ways.