freeze-drying. If so, this would lower the storage and shipping costs of animal strains considerably.

References

1. T. Wakayara and R. Yanagimachi, *Nature Biotechnol.* **16**, 639 (1998).

NET TIPS

META-LANGUAGE

Mailbox:
www.sciencemag.org/dmail.cgi?53804

XML Is Hatching

Robert Sikorski and Richard Peters

The intense promotion of Internet technologies makes it difficult for the busy scientist to separate fact from fiction. At the cutting edge, it is often pointless to try, since most new ideas fail to gain widespread acceptance in the marketplace. It is often best to watch as embryonic technologies mature in their "larval" stage, looking for signs that the technology is being incorporated into current software and tools. Use in solving practical problems and adoption by major organizations and companies are also good signs. However, many Internet "inventions" never make it to these last stages.

One technology that has now developed to the point of hatching and is well on its way to leaving the nest is the extensible markup language, or XML. XML is being developed by a large and dynamic group under the leadership of the W3 consortium, the group that brought you HTML, the language of the World Wide Web.

In the simplest sense, XML can be viewed as a way of attaching additional layers of meaning to any word or words. This is done by incorporating meta-information into the text in the form of tags. A text selection with XML tags will have the following structure: <TAG>Item</TAG>. The tags can be nested to create complex organizations. The additional meaning, or meta-data, that the tags deliver can be extracted easily by software tools. Simply viewing the tags by eye is useful. In an XML document you can extract selected pieces of data from the body of the text. If this all seems too abstract, a concrete example follows.

Let's take the field of human genetics, where gene mutation data are a staple product of this research. Unfortunately, the actual data in most genetics manuscripts (the mutation positions, the phenotypes, the assays used, and so forth) are all buried in the body of the text at publication time. For instance, to review the world's literature for all of the mutations in the gene p16 that cause melanoma, you would have to find the hundreds of papers that mention p16, read them all, extract the data manually, correct the data for the different numbering schemes used in several hundred different mutations, and so on. The task is daunting, and it underscores the need for a different way to handle this type of scientific data. Mutation databases are being formed. Such databases are easy to start, but they are very difficult to maintain in the long run. Given the numbers of genes studied, it is hard to see how high-quality databases can be kept for each one. Also, without standards, it is difficult to compare or collate data from different databases. Instead, what if we just assigned a few XML tags to the published data as it resides on the publisher's Web site? Add tags such as <Gene>p16</Gene> or <Mutation>E30A-</Mutation> and it would now be easy for software tools to extract the actual data automatically, collect it into one data structure, and allow you to visualize it in many different ways. You could feed many papers into an XML grinder to get data from multiple sources. All you need for this to work is a common set of tags to describe mutations. This set does not yet exist for genetics, but it is easy to imagine by just extrapolating from work done in other fields.

In the field of chemistry, an advanced set of XML tags has been created and deployed already. Termed Chemical Markup Language (CML), the tags provide a way of describing the structure of any molecule—the atoms, the bonds, the isotopic constitution. Think of CML as a networked version of standard chemical nomenclature.

XML is now emerging as the standard method for passing structured data over the Internet. All major Internet-based companies have supported XML in this regard. In fact, Microsoft has embedded XML into the next version of its popular Office product. With this new release, Microsoft Word documents will be automatically translated into XML. This alone means that the general use of XML documents will rise considerably.

The software development community as a whole is now focusing on incorporating XML into everything from browsers to databases. Currently, the 4.0 version of Microsoft Internet Explorer can process and display XML documents. The 5.0 versions from both Microsoft and Netscape will handle XML as well. This will make XML the de facto standard for distributed, tagged data. Additionally, many tools for creating XML and interacting with existing databases are in early releases today. Over the next few months, there will be many to choose from.

Given the large support for XML and its incorporation into the infrastructure of the next wave of Internet computing, the field of science will realize many benefits from this technology.

SITEFINDER

Robert Sikorski and Richard Peters

XML is poised to revolutionize how scientific data is exchanged and used. Below is a collection of Web sites on XML useful to both experts and beginners. Links to these sites, as well as additional XML resources, can be found at: www.medsitenavigator.com/tips

Introduction
www.byte.com/art/9801/sec5/art18.htm

www.byte.com/art/9803/sec5/art5.htm

www.forbes.com/asp/redir.asp?/tool/html/98/apr/0413/side1.htm

Background
www.w3.org/XML

XML FAQs
http://builder.cnet.com/Authoring/Xml20/index.html

www.ucc.ie/xml

www.textuality.com/xml/faq.html

www.isgmlug.org/n3-1/n3-1-18.htm

CML FAQs
http://sun01.iigb.na.cnr.it/omf/cml-1.0/doc/faq/index.html

News
www.arbortext.com/xmlnews.html

Case Studies
www.microsoft.com/xml/scenario/chrystal.asp

XML/SGML Comparisons
www.sil.org/sgml/sgml.html

Company White Papers
Microsoft
www.microsoft.com/xml/default.asp

Sun
http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm

IBM
www.infoworld.com/cgi-bin/displayStory.pl?/features/980706xml.htm

XML Specifications
www.w3.org/TR/WD-xml-lang-970331.html

Tech.Sight is published in the third issue of each month. Contributing editors: Robert Sikorski and Richard Peters, Medsite Communications, Boston, MA 01945, USA. Send your comments by e-mail to techsight@aaas.org or via the Web to the mailbox URL with each item.