



WHAT'S ALL THE NOISE ABOUT XML?

Sometimes, a new technology suddenly seems to be everywhere. In 1998, XML (Extensible Markup Language) has been just such a technology. Why all the noise? The buzz over XML stems mainly from two problems: how publishers can have better control over the presentation of documents on the World Wide Web, and how computers can exchange information that is structured for further processing.

WIN TREESE

Putting it Together



Both of these problems have been serious barriers to the increased use of the Web for many kinds of applications, and XML is one way to solve them.

WHAT IS XML?

XML is one example of a “markup language,” which is a formal way of annotating documents to indicate how they should be interpreted, presented or otherwise processed. By “document,” we mean not only things like memos and reports, but almost anything that can be passed from one computer to another, including images, order forms, invoices and the like.

The markup of a report might be very simple: the title, the author, headings of various kinds, the body text of the report and some figures or tables. The markup for an order form might be very different, including a list of items to be purchased, quantities, descriptions, part numbers and prices. In both cases, the markup tells us something about the structure of the document, and we may also use that knowledge to render the document onto a screen or piece of paper. For example, the headings of a report could be typeset in larger bold type than the body text, or the items in an order form might be displayed in a table.

For most of us, the most familiar markup language is Hypertext Markup Language (HTML), commonly used on the Web. Even if we don't create HTML by hand, most Web users have seen “raw HTML” at one time or another. A simple example in the report style might be:

```
<html>
<body>
<h1>A Short Report</h1>
<I>Xavier M. Lang</I>
<p>This report introduces XML.</p>
</body>
</html>
```

When rendered by a browser, the result might look something like this:

A Short Report

Xavier M. Lang

This report introduces XML.

It's not too hard to see how the markup relates to the presentation. The elements in angle brackets, such as `<html>`, are called “tags.” The `<html>` tag indicates the beginning of an HTML document, and the corresponding tag `</html>` indicates the end. Most tags in HTML come in pairs, with a slash to indicate the closing tag of the pair. The `<h1>` tag marks a “heading level 1,” which in this case is the title of the report. HTML's `<I>` tag indicates italics, used for the author of the report, and we mark a regular paragraph with `<p>`.

Even in this short example, we can see some of the limitations of HTML. Here we have used the markup primarily to describe how the document should be presented, such as using italics for the author's name. The focus on presentation means that we have lost some useful information, because a computer program processing the document can't easily identify the title or author of the report. If we wanted to search a database of reports, having such information could be very valuable in making those searches effective.

XML gives us a way to solve this problem. It allows us to define tags for marking up the document to describe the real structure, as well as a way to tell a browser (or other program) how to render it for display. One way to write our report example in XML would be:

```
<?XML version="1.0">
<report>
<title>A Short Report</title>
<author>Xavier M. Lang</author>
<p>This report introduces XML.</p>
</report>
```

*XML allows us
to define tags
to describe a
document's real
structure, as well
as a way to tell a
browser how to
render it for
display.*



Win Treese <treese@acm.org> is Corporate Systems Architect at Open Market Inc., and co-author of *Designing Systems for Internet Commerce* (Addison-Wesley, 1998).

In this example, the markup tells us about the structure of the document. A companion standard to XML, called the Extensible Style Language (XSL), lets us describe the way that such a document should be displayed (or printed).

A BIT OF HISTORY

The history of XML really began more than 20 years ago, with the development of what was to become the Standard Generalized Markup Language (SGML). SGML provides a complex and powerful way to define document formats and use them to create documents. It has been widely used in the publishing industry for many years. When Tim Berners-Lee first created the World Wide Web, he used a very restricted subset of SGML, with the addition of hyperlinks, that he called HTML.

The invention of HTML enabled the rapid growth of the Web, and many applications have been built with it. But as the Web grew, many designers were frustrated with the lack of control it offers for presentation. In reaction, they developed many clever tricks for layout, such as putting text and images in tables with invisible edges, or using blank images to fill in white space precisely. They also began to demand ways to extend HTML.

At the same time, many users of SGML were also frustrated by the limits of HTML, even as they saw its obvious appeal on the Web. The popularity of SGML had been limited by its complexity, special cases and optional features that were poorly understood. Developers found it hard to write programs that worked well with it. But SGML also had the power that the Web needed to grow.

In January 1997, the World Wide Web Consortium (W3C) created a working group for XML to define the syntax for a new markup language, one that would be much simpler to use than SGML but would offer the extensibility needed for many different applications.

XML IN PRACTICE

XML is already being used in many applications, ranging from protocol definitions to interactive programs using XML-enabled Web browsers. Many of the early examples are projects in electronic commerce. One such example is the Open Buying Initiative (OBI), which provides a way to define the interactions between trading partners. Another is the Open Trading Protocol (OTP), which is intended to provide a framework for consumer electronic commerce that can incorporate different kinds of purchasing and payment protocols. It is being developed by the OTP Consortium (<http://www.otp.org>).

Other efforts from the W3C are using XML as well. These include MathML, which provides a standard language for describing mathematical equations, and the Platform for Privacy Preferences (P3P), which provides a way for Web sites to describe their privacy policies and for end users to control what private information is disclosed. These are just a few of the many projects already underway with XML, and we can look forward to many more as XML tools become more common. ~

PERMISSION TO MAKE DIGITAL OR
HARD COPIES OF ALL OR PART OF THIS
WORK FOR PERSONAL OR CLASS-
ROOM USE IS GRANTED WITHOUT FEE
PROVIDED THAT COPIES ARE NOT
MADE OR DISTRIBUTED FOR PROFIT OR
COMMERCIAL ADVANTAGE AND THAT
COPIES BEAR THIS NOTICE AND THE
FULL CITATION ON THE FIRST PAGE.
TO COPY OTHERWISE, TO REPUBLISH,
TO POST ON SERVERS OR TO REDIS-
TRIBUTE TO LISTS, REQUIRES PRIOR
SPECIFIC PERMISSION AND/OR A FEE.
© ACM 1091-556/98/1200
\$5.00

FOR MORE INFORMATION

The XML specification was developed by the World Wide Web Consortium; the W3C's project page about XML can be found at <http://www.w3.org/XML/>. Peter Flynn's "Frequently Asked Questions about the Extensible Markup Language" page is at <http://www.ucc.ie/xml/>. The XML.com Web site (<http://www.xml.com>) is another excellent source of information about XML, as is Robin Cover's SGML/XML Web Page at <http://www.oasis-open.org/cover/xml.html>.

There are many books about XML (as I write this, a search for "XML" at Amazon.com yields 39 entries). Some good ones include *XML: Principles, Tools, and Techniques*, edited by Dan Connolly (O'Reilly/Travelers' Tales); *Presenting XML*, by Richard Light and Tim Bray (Sams); and *XML by Example: Building E-Commerce Applications*, by Sean McGrath (Prentice Hall Computer Books).