

information insider



XML: The New Document Standard Robert J. Boeri Martin Hensel

EMedia Professional, June 1998

[Copyright © Online Inc.](#)

Everybody within a few degrees of hypertext circles knows that eXtensible Markup Language (XML) is hot. But what is it, how does it compare with HTML and SGML, and how will it affect your electronic publishing efforts?

Definitions first: XML is both the name of a specific standard and an umbrella term for three distinct, complementary standards: XML, The eXtensible Linking Language (XLL), and the eXtensible Style Language (XSL). And true to its SGML lineage, XML really isn't a language at all, but rather a standard for creating markup languages.

XML will change everything from the way you code your Web pages to the way you manage digital document collections. Its first point of impact will be Web applications, but soon XML will encompass CD/Web hybrids and structured off-Web data as well. As the relentless move toward making Web browsers the universal interface to information, it will matter less where that information is found. And as XML becomes integrated with Web browsers, XML's impact on data will grow far beyond the Web.

MARKUP LANGUAGES: A BRIEF HISTORY

Standard Generalized Markup Language (SGML), an International Standards Organization standard approved in 1986, defined the rules for creating textual markup languages. Although HTML has SGML roots (and in its pure form, HTML is truly an application of SGML with a Document Type Definition), HTML as practiced would fail any SGML stress test. Until recently, HTML was largely presentation-oriented (e.g., `` for "bold") instead of concerned with structure (e.g., `<list_item>`) as is SGML. But like Henry Ford's Tin Lizzy, HTML gave birth to a vigorous new industry, was simple and affordable, and came without options. At the other extreme, SGML was designed to be quintessentially general yet was neither simple nor affordable. The SGML specification alone is about 300 pages long.

In 1997, the Web "market" was ripe for something in between these extremes; enter XML. Unlike its SGML parent, XML is not an ISO standard but instead was developed by a consortium of nearly 300 companies under the aegis of the World Wide Web consortium. And its specification is about a tenth the size of its parent's. XML was conceived by the scent of commercial profit, the enormous popularity of the Web, the limitations of HTML, and the realization that file systems from local hard drives to the Web needed integration and access via a Web browser. Accepted as a recommended standard in February 1998, XML now comes with an industry guarantee that it is stable and will deliver the interoperability that the Web demands.

WEB-ENABLED AND STRUCTURED

XML by itself can be used to model and deliver structured data without any reference to documents, and that may prove one of its earliest uses. However, to be useful as a document delivery standard, its two companions—the style and linking standards—must also be defined. XSL defines how to display each markup tag: color, size, font attribute, and the like. Think of XSL as a downsized version of SGML's Document Style Semantics Specification Language, which attempted to define rules for displaying SGML-tagged data.

XLL, whose SGML parent is HyTime, is designed to enhance the hypertext links that make the Web work. Instead of providing one-to-one paths between documents, XLL will take advantage of XML structure. Today, clicking on a link such as "Drug Family" could transfer you to a specific anchor point in an aspirin medical document; clicking on a link such as "Drug Interactions" might take you to a different anchor point in that same aspirin document. In XLL, clicking on the link could let you jump to a pop-up list of sections in the aspirin document, such as drug interactions, drug family, warnings, or other sections.

Beyond allowing publishers to create their own custom tag sets, XML has been designed specifically to allow real-time use. By trimming the generality of SGML, XML will allow browsers to interpret and display a stream of tags and data. Further, XML's structure will enable focused searching (e.g., find all "toxic" within "Warning" tags).

True to its SGML roots, XML will require publishers to have a clear idea of what their documents' structures can be. Follow the rules of XML, and your parser can infer the structure of your documents from clues in the use of the tags. But publishers beware: unlike HTML, which you could abuse with nonstandard extensions or even misspelled tags (which browsers would gracefully ignore), XML is strict. In fact, the XML standard says in effect that XML browsers must refuse to process an XML document that isn't at least implicitly structured, or "well-formed."

LINKS TO SGML ROOTS: COMMON THREADS IN HYPERTEXT'S FUTURE

Like XML, HTML 4.0 is a recommended specification. You don't have to look very hard to see how these two standards, although fundamentally different, reveal their common SGML ancestry. HTML 4.0 places tags into three categories of suggested use, each corresponding to an SGML Document Type Definition: Strict (including all tags and attributes that have not been deprecated—discouraged from use and possibly soon to be removed from the spec), Transitional (all strict plus all deprecated tags and attributes), and Frames (all Transitional plus tags supporting frames).

Note that most deprecated tags have been demoted to that status because they dealt with visual presentation, and SGML and XML both separate form from content. Instead, HTML 4.0 encourages the use of style sheets to apply form to your content much like XML. As HTML evolves along a path that draws it even closer to its once-dissimilar cousin, XML's impact becomes as ubiquitous as the Web itself.

Robert J. Boeri (bboeri@world.std.com) and *Martin Hensel* (mhensel@hensel.com) are columnists for INFORMATION INSIDER. Boeri is Information Systems Publishing Consultant at Factory Mutual Engineering of Norwood, Massachusetts. Hensel is president of Texterity Inc., a Newton, Massachusetts-based consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and composers.

Comments? Email us at letters@onlineinc.com.

[\[EMedia Home\]](#) [\[News\]](#) [\[Magazine Issues\]](#) [\[Subscriptions\]](#) [\[Web Resources\]](#)

Copyright © 1998, Online Inc. All rights reserved.

info@onlineinc.com

[This site created for best results under Netscape.]