



Robert J. Boeri &  
Martin Hensel

## Manage Your Metadata

**m**ost searches on popular Internet search engines yield hundreds or thousands of results, few of which correspond to what the query was intended to find. Why can't the engines do a better job of giving us what we want? Because HTML was defined as a presentation language, not as a way to structure document information. Although structure was added to support such objects as tables, the few structures available only make us want more, such as forms support. And no standard can satisfy every need. Thus XML, a standard for creating markup languages (like SGML, but Web-enhanced), was a response both to the explosion of Internet content and the need to create tags as needed.

Yet even with the potential for greater flexibility and specificity XML offers, a basic problem remains: How can I identify the attributes of an object (such as a text string), to say whether that string represents a word, a number, or a date? Without such information, XML-coded sites will be structured, linked, and styled without providing the necessary "metadata" tags to communicate what their information means.

### ENTER THE META MANAGERS

ICE, or Information Content and Exchange, is a protocol aiming to develop a consistent vocabulary for describing and managing the exchange of content and electronic assets between Web businesses. Web sites based on ICE will facilitate electronic commerce, including Web superstores. For example, a Web travel site could license other Web-based restaurant and hotel guides, vacation club materials, and similar travel-related content. Companies to watch that will facilitate this new generation of business Web commerce include Vignette Corporation with its Story Server product and Firefly Network, a Cambridge, Massachusetts-based private company facilitating personalized and secure Web business relationships.

ICE will also facilitate syndicating content, helping publishers increase sales by making content licensing easier. Moreover, content buyers could add their own value and redistribute the content. For example, restaurant reviews could be distributed to newspapers as well as CD/Web media and even to cable outlets. Each could add its own value to that content via formatting, enriching links, integrating with other content, and so on.

Even more abstract is the Resource Description Framework, or RDF. RDF is a metadata framework providing interoperability between Web applications exchanging machine-understandable information. In fact, RDF can be applied to any resource that can be named by a Uniform Resource Identifier (URI—a more generalized way of pointing to Web resources than a URL). RDF creates a framework for creating, manipulating, and searching information. Over time, RDF could transform the Web into a coherent digital library. However, RDF itself is only a framework, and specific implementations (called vocabularies) must be developed to fulfill its promise. RDF is championed by Microsoft (RDF builds on Microsoft's X&IL-Data proposal), Netscape (building on Netscape's Meta Content Framework), content-providers like CNN and ABC News, and search systems suppliers like Alta Vista and Yahoo.

One leading group building on RDF is called the Dublin Core, an international effort to make much of the Web into a modern-day equivalent of the ancient, grand Library of Alexandria. Dublin Core ([http://purl.org/metadata/dublin\\_core](http://purl.org/metadata/dublin_core)) is a set of **15** metadata elements designed to facilitate discovery of electronic resources. Although this clearly emphasizes Web sources, it is not intended only for Web content. Moreover, the type of resources being described are not restricted to text; they may include multimedia types such as pictures and sound. This effort has attracted the attention and talent of international resource description groups, such as museums and libraries.

Dublin Core's 15 metadata elements, each optional and repeatable, are grouped into three categories: content, intellectual property, and instantiation. The seven content tags include the familiar Title, Subject, Description, and the like. The four intellectual property tags are Creator, Publisher, Contributor, and Rights. Lastly, instantiation includes Date, Type, Format, and Identifier.

The list of international projects using Dublin Core is long and growing. Samples include Europe's Euler (integrated bibliographic databases, academic journals, and mathematical Internet sources), Germany's Subject Area Information for Earth (Earth Sciences information on Internet servers, CD-ROMs, and reference books), and Florida International University Digital Library (cataloging images, sound, and video for all subjects in the university's teaching and research portfolios).

### HOW META CAN GET BETTER

What impact will metadata initiatives such as RDF have on content developers and search systems? By guaranteeing metadata to be interoperable and able to be processed by software automatically, these metadata standards will provide a uniform way to search for information. Clearly this will enhance information searches by providing more precise results than even the best queries to a Web search system can now provide. And if you provide content, for the Web or elsewhere, you will want to be sure your content is accessible to as many users looking for it as possible.

**Robert J. Boeri** ([bboeri@world.std.com](mailto:bboeri@world.std.com)) and **Martin Hensel** ([mhensel@hensel.com](mailto:mhensel@hensel.com)) are co-columnists for *INFORMATION INSIDER*. Boeri is an Information Systems Publishing Consultant at a Boston-area loss prevention and control service company. Hensel is founder of Martin Hensel Corporation, a Newton, Massachusetts-based consulting firm that builds SGML-based editorial and production systems for publishers, corporations, interactive services, and compositors.

Comments? Email us at [letters@onlineinc.com](mailto:letters@onlineinc.com), or check the masthead for other ways to contact us. ■