nature webdebates

Text markup and the cost of access

Jon Bosak
Sun Microsystems
jon.bosak@sun.com

Nature has asked me to assess the possible impact of extensible markup language (XML) on the dissemination of scientific literature. I argue that the potential benefits of XML are real but can be achieved only by adding considerably to the costs of production.

XML is a method for defining special markers or �tags� that can be inserted into text to indicate its logical structure and to make explicit the meaning or rhetorical role of its component parts. In a scientific article, for example, XML tags can be used to distinguish the title of the article from the names of its authors or the cells in a table. If done properly, �marking up� text in this way makes each document into a miniature, hierarchically structured text database.

This has two important consequences. First, it enables more sophisticated retrieval in automated searches. Second, by separating the presentation of a document from its content and structure, it enables the visual form of a document to be determined at run-time, thus allowing the presentation of text to be automatically adapted to the capabilities of different publishing media.

Some scientific publishers already use XML, though so far without the standardization that would enable the benefits of XML to be realized outside the individual publishing house. The adoption of a standard XML markup for all scientific publications would allow users to locate desired material more quickly, greatly facilitate the publication of scientific literature in multiple media and substantially reduce the effort needed to integrate papers from multiple sources into a shared repository.

However, the increased utility of properly marked-up text cannot be achieved without cost. Markup is useful because it provides a way to add information that aids in processing the text, performing for the machine a function analogous to the function that careful typography performs for the human reader � differentiating the text from the abstract and the footnotes, for example. This added layer of information is the product of human thought, however, and therefore requires human labour.

In industrial contexts, it is commonly estimated that marking up technical material adds about 40% to the work of writing the text. This is a continuing cost, in addition to the costs of training and tools. Consistent markup also imposes upon authors, even professional technical writers, a mode of working that many find uncongenial � and consistency is the *sine qua non* of XML-enabled search functionality.

A body of information inconsistently provided with metadata will return dangerously misleading results to a system expecting that metadata. For example, a sophisticated query keyed to search for papers in which a certain term appears tagged as a chemical reagent will find papers in which that term has been given the standardized tag for chemical reagents but not those whose authors failed, due either to ignorance or to a lack of resources, to embed this additional information in the text. Partial adoption of the tagging scheme means that an intelligent search engine asked to aggregate papers on this basis will return a number of papers answering the query but will have no way to list the papers that actually contain the target information that has not been tagged appropriately. Such a system is much worse than a simple full-text search, because it gives a user the false impression that a precise and comprehensive report has been generated. In a field such as medical research, misleadingly incomplete query results of this kind could have disastrous consequences.

So although XML markup, properly applied, can greatly increase the utility of scientific texts, the cost of using it effectively is not trivial.

The same is true of more powerful indirect information-management techniques based on XML. Here the analogous traditional forms are the index and the bibliography. Properly constructed electronic indexes, also known as �concept maps� or �topic maps�, can be vastly more capable than traditional printed indexes. Like their traditional counterparts, they can be created without modifying the text itself, although they are more useful and easier to construct if the text they refer to has been uniformly marked up. But just as the utility conferred by a traditional index adds considerably to the cost of producing a traditional text, so the potentially much greater utility conferred by a properly constructed topic map can be achieved only through a concomitant investment of intellectual capital.

The principle here is quite simple: the more information we add to a document, the better use we can make of it. But the information has to come from us; it is not going to come from the computer. XML does nothing to change this basic fact. The potential for increased access to information made possible by the new XML-based technologies is indeed spectacular, but the notion that we can noticeably improve upon full-text searching without thorough standardization and a continuing investment of consistent, expert labour is, in my opinion, a chimaera. I believe that the big question in making scientific texts more accessible is how to provide the additional expert labour.

Anyone capable of writing an acceptable scientific paper can master a consistent metadata format and can learn the subject-matter classification of their specialty. This indicates that we should look to the authors of scientific articles for a considerable portion of the extra work needed to enable better online access. In particular, we should be able to expect authors to use a standardized XML markup for scientific articles and to expect authors or their editors to classify the work properly in one or more standard registries. But this would still leave the question of how to define and implement an online catalogue system that would allow humans and search tools to assemble variously related texts.

One cost-effective way to establish and maintain online catalogues is suggested by the Open Directory Project. This self-sustaining directory of the World Wide Web is driven by classification data entered by the creator of each web page. Management of the registry is accomplished through a system of volunteer area editors, each responsible for the organization and maintenance of some small part of the overall taxonomy. One can imagine a directory such as this for scientific literature that would be created and maintained by a distributed network of domain experts.

In my opinion, we cannot evade the real costs of supporting automated access to scientific literature, but I believe that we can take steps to see that these costs are equitably shared within the community of scientists and publishers. The considerations I�ve outlined above suggest a programme something like this:

1. Adopt a standard XML markup for scientific texts within each specialty and require all authors working in that specialty to adhere to it.
2. Adopt a standard format for bibliographic data and require all authors and publishers to provide such data in texts made available in electronic form.
3. Institute a collaborative project to catalogue scientific papers using a distributed system based on the labour of volunteer editors.

The Internet provides a technical infrastructure that can greatly improve access to the scientific literature. It remains to be seen whether the scientific community is ready to accept the conformity and provide the extra layer of collaborative work needed to reach this goal.

Jon Bosak organized and led the working group that created XML. He subsequently served for two years as chair of the XML Coordination Group of the World Wide Web Consortium. He is a founding member of OASIS, the Organization for the Advancement of Structured Information Standards, and he chaired the committee that developed the OASIS process for the definition of industry-specific XML markup standards. He also served on the Advisory Board of the Electronic Business XML initiative (ebXML), a joint project of OASIS and the United Nations body for Trade Facilitation and Electronic Business (UN/CEFACT). At Sun Microsystems, where he holds the title of Distinguished Engineer, Bosak originated the strategy used for the Web distribution of documentation about the Solaris operating system. Before joining Sun, he was responsible for developing the online delivery system for Novell NetWare manuals based on SGML, the predecessor to XML, and was a primary force behind the development of the DocBook standard for the markup of technical documentation.

Articles by Jon Bosak can be found at his home page