Journal Publishing Technologies: XML

By Anne Hoekman, Managing Editor,
Journal of International Business Studies

As managing editors, many of us simply ship off our manuscripts to a publisher after the review process is complete and the acceptance letter has been issued. But do we know what happens to the manuscript when it gets there? Being aware of the processes and technologies used by journal publishers and printers might not have direct impact on the daily tasks of a managing editor, but it is helpful for understanding the production, distribution, and consumption of your journal – without which your journal (and your job!) wouldn't exist. One such important technology is XML.

XML 101

Extensible Markup Language (XML) is a technology used for structuring and organizing data. It's called a "markup language" because it uses tags to mark and delineate pieces of data. The "extensible" part means that the tags are not pre-determined; users can define them based on the type of content they are working with. According to the World Wide Web Consortium (W3C), XML's goal is to "structure, store, and transport information," focusing on "what data is," rather than focusing on "how data looks." So the tags are based on the *content* rather than on the *appearance* (this would be HTML) of a document. For example, a reference such as:

> Shah NC. Viewpoint: Consultation time—time for a change? Still the "perfunctory work of perfunctory men!" *Br J Gen Pract* 1999;49:

could be marked up as the following:

```
<ref id="B1">
  <label>1</label>
  <nlm-citation citation-type="journal">
    <person-group person-group-type="author">
      <name>
        <surname>Shah</surname>
        <given-names>NC</given-names>
      </name>
    </person-group>
    <article-title>Viewpoint: Consultation time—
      time for a change? Still the "perfunctory
      work of perfunctory men!"
    </article-title>
    <source>Br J Gen Pract</source>
    <year>1999</year>
    <volume>49</volume>
    <fpage>497</fpage>
  </nlm-citation>
</ref>
```

(From a sample article at http://dtd.nlm.nih.gov/publishing/tag-library/2.3/index.html.) This reference sample is just one piece of the XML document; the whole article, plus each reference, would be tagged in such a way.

Standard Generalized Markup Language (SGML) is a similar technology; it's actually a markup language that is the predecessor to XML. SGML was developed in the 1980s and was implemented widely in the publishing industry, in part to simplify the publishing process (i.e., to save time and money) and provide a way to structure text that was "system independent" (Cave, p. 144). Some publishers still use SGML; in this article I discuss just XML as it is and can be used in journal publishing.

As one of XML's main purposes is to transport data, it is a great tool for exchanging electronic information, such as between an editorial office and a publisher. This is one area where a managing editor might come briefly into contact with this, though you may not know it. For example, the *Journal of International Business Studies* (*JIBS*), which I work for, uses the online peer-review system Manuscript Central. When a manuscript has been accepted, we send it through Manuscript Central

to the publisher's FTP site. The system sends all of the manuscript's metadata (title, authors, keywords, acceptance date, etc.) in XML format.

And this leads into the final key point about XML: as W3C explains, "XML does not DO anything.… Someone must write a piece of software to send, receive or display it." XML contains only the content of a text or publication; it requires a software application to send the information, and it requires a Document Type Definition (DTD) or schema to present the text (e.g., in the form of a website or a printed article). And, going back to the transfer of manuscript metadata to the publisher, a DTD is used to format that information too.

The DTD or schema defines the types of elements that are required in the XML document, and it defines the relationship of those elements in the document. So, for example, a DTD can define the content's output as a PDF for sending to the printer with one set of rules, and a second set of rules can transform the content into HTML for web display. As will be illustrated below, this ability to create multiple outputs from a single source is one of the main draws for using XML. This is beneficial in many industries but is especially so for the publishing industry in an era when online presentation and consumption of content is ever increasing.

XML and journal publishing

Publishers have been using XML since its inception in the late 1990s; according to Judith Wusteman, "journal publishing was one of the first industries to employ markup technologies successfully" (p. 125). What follows in this section are some XML uses and processes in the industry.

According to Boismenu and Beaudry in their book *Scholarly Journals in the New Digital World* (p. 108), when implemented fully, XML can accomplish all of the following for publishers: receive manuscripts, prepare copy, prepare proofs, and produce metadata. In order to implement, the publisher must transform the article, which is probably most often received as a Word document, into XML.

Boismenu and Beaudry outline the two processes used by publishers to convert manuscripts to XML (pp. 97–102). The first is to lay out the text with page-layout software, such as QuarkXPress or PageMaker. After layout, the article is sent to the printer for the print publication, and then for web publication, the publisher use a conversion program designed to convert the page-layout document to XML. An XML editor can also be used in place of a conversion tool. This method is more print-based as the main focus is laying out the print version, then secondarily converting to XML for web use.

The second model involves first applying a style sheet to the Word version upon receipt (identifying elements like title and headers), and then converting the article to XML. From there, the XML document is used to create a page-layout version for print and an HTML version for the web. This method is more flexible for producing multiple formats; it can "service the greatest number of current and future dissemination media" (Boismenu & Beaudry, p. 102), an important distinction in this digital era. According to Alex Brown, this process of converting a manuscript to XML upon receipt has become part of the "new orthodoxy" of journal production (p. 153).

XML is also used by journal publishers to create metadata, which can automatically be pulled from the text when it's formatted with XML. Metadata are important because they "describe the attributes and content of articles and are used for information

retrieval, management, description, access, and preservation" (Boismenu & Beaudry, p.105). Good metadata allow for more effective searches, enabling researchers to better locate and cite your journal's content.

Why XML?

Large journal publishers are all using technology like XML, but there is not yet an industry standard for its use. Agreement has not yet been reached on the level of "granularity" (how big of chunks of text) to use in tagging articles (see Wusteman, p. 126) and which DTD to use. Most scholarly publishers use their own DTDs for marking up article headers and full article text (Cave, 147).

So there are perhaps some drawbacks to XML and problems with it in the journal publishing industry; however, it has made its mark on publishing in many positive ways. The use of XML has many features for improving web presence, distribution, and budget.

1. Ease of producing multiple formats

XML allows publishers to easily create different formats of the same article because XML tags only the content, not the presentation, of the articles. Articles published in *JIBS*, for example, are available in three formats: the standard print version, an online PDF version (which follows the same formatting as the print version, but with slight differences, such as in pagination), and an online HTML version. By using XML, these three different formats can be created from a single source. Online presentation and presence is important in the digital age, and it is a cost and time saver to create both the print and online versions from a single source.

2. Rapid dissemination of information

When publishers can more easily and quickly produce multiple formats of an article by using XML, this can also lead to more rapid distribution of their journals' content. One example of what XML technology can facilitate is the Advance Online Publication (AOP) system from Palgrave MacMillan (publisher of journals like *Nature* and *JIBS*).

AOP articles are published online ahead of their print publication date, and these web versions are final versions – they are unchanged when they appear in print. Advance web publication allows for faster production of time-sensitive data, which means it can be read and cited sooner. This can be especially important for science journals.

According to industry expert Bob Hecht, XML is an "enabling technology"; an XML production process is a more efficient production process, and thus enables features like AOP.

3. Wider dissemination of information

As mentioned earlier, metadata makes journal content more effectively searchable. And with online researching and distribution of journal content now the norm, most journals can be accessed through various online indexes and databases, such as ProQuest or Science Direct. Libraries then subscribe to online journal access through these vendors.

Making journal articles available through online vendors most likely makes it searchable for a much bigger audience than the journal's circulation, and vendors often require submission of XML metadata for indexed articles.

A major example of how XML and metadata can be used to more widely share content is CrossRef, which was established by scholarly publishers as "an infrastructure for linking citations across publishers" (see www.crossref.org). In this system, a researcher

reading an electronic article can directly access a paper that article cites by simply clicking on the citation in the reference section.

CrossRef works by assigning each article a unique Digital Object Identifier (DOI). Publishers submit article metadata to the CrossRef DOI database in XML format, and then CrossRef registers each DOI and the URL where the article resides. The publisher also submits the citations listed in each article to the Reference Resolver, which then links the citations to their data in the DOI directory (if the citations have been previously registered with CrossRef). Thus, CrossRef can create a link in an article's reference section to the actual article cited.

With its mission "to serve as the complete citation linking backbone for all scholarly literature online, as a means of lowering barriers to content discovery and access for the researcher," CrossRef uses XML metadata to widen journal article dissemination and increase ease of locating articles, allowing your journal greater visibility.

Conclusion

There are many other ways XML technology can influence production of your journal and how it is consumed, from using it to create RSS feeds alerting subscribers of new issues to using it as a quality assurance tool.

Boismenu and Beaudry leave us with the overarching importance of "enabling" technologies: "The value of a collection of journal articles lies in the quality of the articles, but also in the number and accessibility of journals and articles on a given subject" (p. 108). You as the managing editor contribute to the process of publishing high quality scholarship, and your publisher, using technologies like XML, contributes to the efficient and wide-reaching dissemination of that research to scholars in the field.

As XML becomes more prolific, and as methods of distribution continue to grow and change, this is an important technology for anyone in journal publishing to learn and understand.

References

Cave, F. (2003). Article metadata standards: An historical review. *OCLC Systems & Services*, 19, 144–148.

Brown, A. (2003). XML in serial publishing: Past, present and future. *OCLC Systems & Services*, 19, 149–154.

Boismenu, G., & Beaudry, G. (2004). *Scholarly Journals in the New Digital World*. Translated by Maureen Ranson. Alberta: University of Calgary Press.

Hecht, B. Who in their right mind wouldn't want XML? (Interview with Data Conversion Laboratory). Retrieved April 22, 2008, from http://www.dclab.com/bobhecht.asp.

World Wide Web Consortium. Introduction to XML. Retrieved April 20, 2008, from http://www.w3schools.com/XML/xml_whatis.asp.

Wusteman, J. (2003). XML and e-journals. *OCLC Systems & Services*, 19, 125–127.

Further resources

- Journal Publishing Tag Set Library (http://dtd.nlm.nih.gov/publishing/tag-library/2.3/index.html): examples of articles tagged in XML along with what the final (PDF) version looks like
- DCL's Technical Library (http://www.dclab.com/techlibrary1.asp?GRP=21): information on XML in scientific, technical and medical journals, including interviews with industry experts
- *OCLC Systems & Services* 19 (2003): special issue on XML and e-journals