An Overview of Electronic Publishing and Extensible Markup Language (XML)

John Kane
Joseph R. Makuch

Abstract

Many knowledge-based organizations are expanding their publishing efforts to include electronic publishing. This article gives evidence of this move and discusses the factors that have been instrumental in promoting electronic publishing. The importance of information stucture and adherence to open standards are emphasized as critical components of digital document management systems. The development and use of Standard Generalized Markup Language (SGML) and Hypertext Markup Language (HTML) are discussed along with their strengths and weaknesses as tools of electronic publishing. An emerging alternative, Extensible Markup Language (XML), is described as having features that may reduce some of the impediments to producing and managing documents digitally.

Introduction

Electronic publishing involves the computer-assisted preparation, presentation, transmittal, storage, and retrieval of digital documents (Vanoirbeek & Coray, 1992). Some digital documents may exist solely in electronic form, while others also may be published on paper. Digital documents need not be limited to text, graphics, and pictures: they can include sound and video and provide hyperlinks to related resources.

Electronic publishing is a growing field: The seventh edition of *The Directory of Electronic Journals, Newsletters and Academic*

The authors are with the U. S. Department of Agriculture, Agricultural Research Service, National Agricultural Library in Beltsville, Maryland. John Kane is Coordinator of Electronic Publishing and Archiving , the Information Systems Division. Joseph R. Makuch, an ACE member, is Coordinator of the Water Quality Information Center.

Discussion Lists (1997) contains more than three thousand four hundred titles of electronic serials, which is twice the number included in the previous (1996) edition (ARL Publications, 1998).

Examples of electronic publishing initiatives include those by the Association for Computing Machinery (Rous, 1993), the Electronic Text Center (Ream, 1993), Johns Hopkins University Press (Lewis & Kelley, 1995), the Networked Digital Library of Theses and Dissertations (1998), the Scholarly Communications Project (McMillan, 1995), and the University of Chicago Press (Owens, 1993).

Why Publish Electronically and Why Now?

What is driving knowledge-based organizations to publish electronically? A major reason is that many interrelated social and technological changes have occurred, and are occurring, that foster the move to electronic publishing. We have more documents to publish and manage, we need to publish faster and less expensively, and we need to make documents universally available. Fortunately, the state of technology allows substantial progress on these issues.

Much of the world is now in the postindustrial or information age, characterized by economies dominated less by manufacturing and resource extraction and more by providing services and information. The amount of information is increasing at rates never before experienced, and the significance of information is increasing substantially as well. Human knowledge is now doubling every ten years and more scientific knowledge has been created in the past decade than in all of human history (Kaku, 1997). And as the general pace of society has become faster, it has become necessary to access and process information rapidly in order to remain competitive. The timeliness of information often determines its value.

Coupled with these changes is the rapid evolution of computers and software. Better than any device to date, computers efficiently store, manipulate and distribute large amounts of information (von Hagen, 1992). Computers have steadily become more powerful, faster, less expensive, and easier to use. Such changes provide an increasingly available tool that is well designed for handling the massive amounts of information contained in various types of documents (Van Houweling, 1994).

The emergence of the Internet as a widely-available means to move documents electronically has also fostered the growth of electronic publishing. One estimate indicates there are fifty-seven million users of computers that can access information by interactive Trans-

mission Control Protocol/Internet Protocol (TCP/IP) services such as the World Wide Web (the Web) or File Transfer Protocol (FTP). The number of users almost doubles every year and is expected to reach seven hundred and seven million by January 2001 (Quarterman, 1997).

Another reason for the interest in electronic publishing is cost containment by publishers. Earl (1996) reports that commercial publishers and other observers estimate that savings from publishing electronically could be about thirty percent. This is because costs to publishers associated with paper documents—printing, transportation and storage—can be eliminated with electronic publishing. Of course, there are new expenses: hardware, software and human and organizational costs associated with the change (Lamberton, 1992). And the administrative and editorial costs to produce the "first copy" are present for both electronic and paper publications. Holmes (1995) estimates that for the National Research Council (NRC) of Canada these first-copy costs account for eighty percent of the total cost of producing a NRC journal. Only twenty percent is for the marginal printing cost and distribution.

Continuity and Change

The agricultural research and education system in the United States is a knowledge-based enterprise that creates and transmits knowledge. Universities are a major component of this system. Commenting on universities as knowledge-based organizations, Van Houweling (1994, p. 9) states:

> Since the knowledge world changes so continuously and so rapidly, there are always new challenges, new information to be sought, new processes to be understood, and students with new needs to understand. As a result, our enterprises are centered on challenge and opportunity, not organization and process. Our focus in not on routine, but on change.

One of the "new processes to be understood" is electronic publishing. The agricultural research and education system must continue serving its clients. But client needs are changing, so the system too must change. New tools are needed for new times. Expanding the media mix to include digital documents provides additional possibilities for enhancing agricultural knowledge management.

Digital Documents

Online digital documents, accessible through a digital library, have a number of characteristics that make them appealing to users.

Table 1 summarizes these characteristics.

Table 1 *Desirable Characteristics of Online Digital Documents from a User's Perspective*

Characteristic	Comment
Available 7 days/wk., 24 hrs./day	The extension office or library never closes
Timeliness and immediacy	Documents can be "published" faster and once published, are immediately available, never out of stock and quickly updated
Location is irrelevant	Distance between document and user doesn't impede access
Comprehensive coverage	User isn't limited to the holdings of a particular location
Simultaneous usage	One document can be independently used by many people at the same time
Secure preservation	Documents aren't missing pages or otherwise damaged
Multimedia	In addition to text and visuals, documents can contain sound and video
Interactivity	Documents can provide hyperlinks to related resources; the user can customize the document's appearance and content; search capabilities allow specific information to be located rapidly within and among documents

Note: These characteristics assume user access to a digital library.
Adapted from: Drabenstott and Burman (1994) and Van Houweling (1994).

Electronic publishing should go beyond the paper-publishing paradigm and take advantage of technology to offer new and better

ways of producing, using, and managing information. According to Boyce, Pilachowski and Dalterio (1993), "the whole point of making information available electronically is to take advantage of the host of new uses for the information that would not otherwise be possible with a simple printed page" (p. 133). But these possibilities cannot be realized without standards.

Standards and Information

Information without structural standards is chaotic. If the ink markings on this page were randomly distributed, they would not be able to convey the authors' ideas to the reader. But since the ink markings have a structure (letters, words, sentences, etc.), meaning can be communicated to readers who share these standards. For clarity and manageability, the structure of information must be explicit.

In the electronic realm, proprietary products that deviate from standards leave users at the mercy of the marketplace and commercial interests. Commercial word-processing and desktop-publishing software packages that were popular in the past can be very difficult to read now. An organization with responsibility to provide access to documents encoded with proprietary products would have to archive the software and maintain the hardware necessary to view the information. Something more generic and open is needed for the long-term use of digital information.

Markup: SGML and HTML

Structural, nonproprietary (i.e., open) standards have grown out of initiatives in the publishing sector. In traditional publishing, structure is elaborated in a process called "mark up." Any document being prepared for publication goes through some form of mark up (the term "markup" without a separation between the words generally refers to electronic markup) which defines the layout of the material. An editor and a typographical designer make manual notations in a document telling a typesetter how to arrange the elements of a document with attributes like typeface, size, pagination and margin size. Standardized editorial notations used in this mark up allow different typesetters to turn out a predictable and commonly recognized product. Without this well-defined layout or structure, a document may be misleading or even incomprehensible.

With regard to electronic documents, markup can be described in two different ways: procedural or descriptive. Procedural markup is prescriptive. Codes in the software describe what should be done

with the marked text: make it bold or italic or indent it. The problem is that when the specific software that reads these codes is not available, the information is indecipherable. Procedural markup is also limited in that it describes what an element of the document looks like and not what it represents. That means if you searched a database of documents but wanted just to search titles or wanted to separate out other elements of a document that were meaningful, you could not do it because those elements have not been explicitly identified. By contrast, descriptive markup describes what the structural elements are: title, paragraph, list, citation, etc.

In the late 1960s IBM and a few other special-interest publishers began to look at problems unique to digital documents. IBM researchers developed a way to replicate manual "mark up" as machine or general "markup" in Generalized Markup Language (GML). It became clear that a standard dealing with document structure could not cover every potential structural document type, so GML was reworked as a metalanguage—a language about a language, or a set of rules for building an application to describe any type of document.

Under the auspices of the International Standards Organization (ISO), the concepts of "generalized markup" were elaborated and formalized as Standard Generalized Markup Language (SGML) in ISO Standard 8879 in 1986. The development of SGML evolved from desires to automate the editorial process and recognized the limitations of relying on document appearance as a mechanism for management. Hockey (1993), Owens (1993), Ream (1993), and Rous (1993) describe electronic publishing efforts using SGML. Cover (1998a, 1998b, 1998c) provides many examples of SGML usage in business, academia, government, and the military.

SGML is a robust metalanguage for describing a document and therein lies its beauty. PC Magazine has said SGML is not just a publishing tool, but a "...new paradigm for working with information" (Karney, 1995, p.144). It is a metalanguage that defines markup codes embedded in documents. But these markup codes are descriptive, not prescriptive (Hockey, 1993).

SGML Structure

Any particular SGML document is configured in four parts, all of which are archived together. These are (a) a declaration which defines how specific options in the standard are being implemented for a particular application or document, (b) a document type definition (DTD) which describes the relationship of structural elements

(tags) in a document type, (c) the instance or text marked up with the tags, and (d) a style sheet describing how the structural elements will appear when output is produced. (Since 1997, style sheets are covered by the Document Style Semantic and Specification Language (DSSSL) Standard; however, appearance is still largely handled by proprietary applications.)

The way the standard is applied—rules (declaration), content (instance), structure (DTD) and appearance (style sheets)—are all handled and archived as unique entities. It is the combination of all four elements that make up a particular document. The same document could be generated with a different appearance given another style sheet or even in a different medium with a modification of the DTD and declaration. For example, Braille is built into most commonly used DTDs today, so with the proper software and hardware the same document can be accessed as a two dimensional document or in Braille.

Using SGML

Broad use of SGML has been limited by its complexity. For many information professionals the intellectual, technological, and financial investment necessary to put a SGML-based system into place, and master the skills to use it, is too high. In addition, SGML's application to the Web—note that SGML predates the Web—has been confounded by the number of optional features it allows. For example, options such as tag lengths and character sets must be synchronized in both sending and receiving systems for a particular document to be read. So while SGML is a sound archival and publishing standard, it has proven difficult for SGML to achieve wide application on the Web.

In 1990, Tim Burnes-Lee at the European Laboratory for Particle Physics (also known as CERN) chose a sampling of tags described in a DTD used at CERN and came up with Hypertext Markup Language (HTML) to take advantage of the "linking" potential of both the Web and SGML. With the release of the freely-available Mosaic browser, the use of HTML became widespread by 1993. HTML's advantage has been that its DTD relies on a limited set of tags and on tags that connote appearance (like bold, italics, and headings). Because of this, authors use HTML editing software much like they use word processors: composing documents based on how they look without considering document structure. In addition, the tags (DTD) and rules of application (declaration) are hard-coded into editor and browser software. Different vendors, however, can hard-code

different tags into their proprietary browsers which has caused some confusion.

Since there are relatively few tags, and construction of a document with HTML is much like authoring a document with word-processing software, HTML as a particular application of SGML has been widely implemented. However, while providing ease of use, HTML has not addressed the problems that the originators of SGML intended to solve. HTML does not describe structure with any degree of sophistication or reliability. That makes long-term archiving of HTML documents extremely dubious. The authors of SGML realized that no single tag set could describe every document type. And a simple tag set, while it may be easy to use, cannot possibly cover the variety of documents that need to be archived. The great weakness of HTML is that it is limited to a single tag set or DTD and cannot accommodate any additional tags. It is not extensible.

Not only is the clarity of structure in an HTML document an issue over time, but also its management. With vast amounts of information available online, systems have to go beyond full text retrieval for the management of information. SGML grew out of the recognition that structure can convey meaning just as content does. A flexible and robust format to describe information allows meaning to be conveyed. To paraphrase a common SGML concept: "If you want to search it, tag it." One of the developers of SGML has stated that "markup should be rigorous so that the techniques available for processing rigorously-defined objects like programs and databases can be used for processing documents as well" (Goldfarb, 1990, p. 8). By contrast, HTML provides neither the stability for clear retrieval and access to information over time nor the sensitivity necessary for effectively retrieving documents from large data collections. Documents coded in Portable Document Format (PDF) have similar problems. PDF facilitates page-based screen display and printing, but does not provide the robust underlying structure necessary for context-sensitive information searching (Milligan, 1997).

XML: Extensible Markup Language

To date, SGML has proved to be too complex for widespread use on the Web, while HTML is too limited for adequate archiving and management of large document collections. An alternative that combines the rigor of one with the simplicity of the other is needed.

In 1996, the World Wide Web Consortium (W3C) formed a committee to address this and other issues. The committee identified three aspects of HTML that needed attention:

(1) Extensibility—permitting authors to define their own tags as they needed them;

(2) Structure—allowing elements to be nested inside other elements (like title inside article, chapter or citation) so that database schemes and object-oriented hierarchies could be described, and;

(3) Validation—letting authors automatically check a document's structure against a specific standard structure (or DTD) (Bosak, 1997).

Using experience with SGML and the Web, the W3C committee developed the Extensible Markup Language (XML). Connolly, Khare and Rifkin (1997) note that "XML is not a collection of new ideas: it is a selection of tried-and-true ideas" (online). XML was specifically designed to rectify the shortcomings of HTML while making SGML easily implemented on the Web. XML is a simplified subset of the SGML standard options and by defining that subset it makes the creation of a document easier and predictable. That XML is less complicated is dramatically pointed out by comparing the XML and SGML specifications: XML specifications are covered in twenty-six pages while SGML specifications require five hundred pages (Khare & Rifkin, 1997). XML is not an application of SGML or a set of tags or a specific DTD as HTML is.

What the specification states (it can be viewed at http://www.w3.org/TR/REC-xml) is that there are essentially two types of XML: "well-formed" XML and "valid" XML.

The only requirements for a "well-formed" document are that it be in plain text (i.e., ASCII) and include:

- one or more elements,
- a root element that cannot appear inside itself (tags that bracket the document),
- opening and closing tags that are nested sequentially (like boxes inside boxes),
- open tags with a "/" at the close bracket (e.g., <graphic file="cow.gif" id="ab4321"/>), and
- all attributes quoted (note "cow.gif" and "ab4321" in previous example).

A well-formed XML document does not use a DTD, making it extremely easy to write and deliver to the Web. "Valid" XML must have a DTD, but a script could be written to extract the tags from a well-formed document, configure them into a DTD, and make the document valid by including the DTD. The purpose of a DTD, however, is to encourage some degree of consistency and predictability by using a recognized DTD so that tags are semantically clear. Without that, there is limited ability to compare the structure of different documents and search across them with anything other than a full-text search.

Already there has been a great deal of interest in, and work done with, XML. From the 1997 Seybold Seminar, Waldt (1997, p. 4) reports that "in addition to Gates, every keynote speaker, including John Warnock of Adobe, John Gage of Sun, and Mike Hoimer of Netscape spoke praise of XML and made clear statements about its potential benefit to distributing information on the Web." That potential is approaching fruition in a series of proposed XML-based formats that address specific HTML related problems. Examples of emerging XML-based formats include Web Interface Definition Language (Allen, 1997) and several works in progress, such as Synchronized Multimedia Integration Language and Resource Description Framework from the W3C (World Wide Web Consortium, 1997, online).

The commercial sector is reacting quickly to XML: Microsoft announced at the 1997 Seybold Conference that its new browser is compliant with XML. And DataChannel Inc. and other companies have formed a sixteen-member council to "promote development of real-world applications based on XML" (Taft, 1998, online).

Summary

Electronic publishing holds a great deal of potential, some of which has already been realized. A problem that has been fostered by the popularity of the Web, however, is that a standard, HTML, with very limited value for long-term document management, has come into common use. SGML, which predates the Web, is a good standard for publishing and archiving, but because of its complexity, its widespread application to the Web has been limited. One solution may be XML. In February 1998, the *Extensible Markup Language (XML)* 1.0 Specification became a W3C recommendation, "signifying that the [specification is] stable, contribute[s] to Web interoperability, and [is] supported for industry-wide adoption by the W3C membership" (W3C, 1998,online). More information about XML can be found at http://www.w3.org/XML.

References

Allen, C. (1997). *WIDL: Application integration with XML* [Online]. Available: http://www.webmethods.com/technology/widl.html [1998, March 31].

ARL Publications. (1998). *ARL announces seventh edition of the directory of electronic journals, newsletters and academic discussion lists* [Online]. Available: http://arl.cni.org/scomm/edir/pr97.html [1998, April 1].

Bosak, J. (1997). *XML, java, and the future of the web* [Online]. Available: http://sunsite.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm [1998, March 31].

Boyce, P. B., Pilachowski, C. & Dalterio, H. (1993). Editorial-electronic publishing in astronomy; projects and plans of the AAS. Reprinted in Okerson, A. (Ed.), *Scholarly publishing on the electronic networks, the new generation: Visions and opportunities in not-for-profit publishing* (pp. 133-135). Washington, D. C.: Association of Research Libraries.

Connolly, D., Khare, R. & Rifkin, A. (1997, Autumn). The evolution of Web documents: The ascent of XML. *World Wide Web Journal* [Online], *2(4)*. Available: http://www.cs.caltech.edu/~adam/papers/xml/ascent-of-xml.html [1998, March 31].

Cover, R. (1998a). *General SGML/XML applications* [Online]. Available: http://www.sil.org/sgml/gen-apps.html [1998, April 2].

Cover, R. (1998b). *SGML: Academic projects* [Online]. Available: http://www.sil.org/sgml/acadapps.html [1998, April 1].

Cover, R. (1998c). *SGML: Government, military, and heavy industry* [Online]. Available: http://www.sil.org/sgml/gov-apps.html [1998, April 2].

Drabenstott, K. M. & Burman, C. M. (1994). *Analytical review of the library of the future.* Washington, D. C.: Council on Library Resources.

Earl, L. (1996). Whither the electronic journal? *SLS UK user group, Yvonne Fuller memorial bursary 1996 winning paper* [Online]. Available: http://www.lib.ic.ac.uk:8081/leah.htm [1998, March 31].

Goldfarb, C. F. (1990). *The SGML handbook.* Oxford: Oxford University Press.

Hockey, S. (1993). Encoding standards; SGML and the text encoding initiative: What and why? In Okerson, A. (Ed.), *Scholarly publishing on the electronic networks, the new generation: Visions and opportunities in not-for-profit publishing* (pp. 59-64). Washington, D. C.: Association of Research Libraries.

Holmes, A. (1995). A publisher's view of the changing paradigm of scholarly publishing. In Fairley, C. (Ed.), *Electronic journals and the paradigm shift* (pp. 26-31). Proceedings of the 1995 spring forum of the Canadian serials industry systems advisory committee, Toronto, Ontario. April 25, 1995. Canadian Serials Industry Systems Advisory Committee.

Karney, J. (1995, February 7). Tag masters. *PC Magazine*, 14(3), 144-162.

Khare, R. & Rifkin, A. (1997, July/August). X marks the spot: Using XML to automate the Web. *IEEE internet computing* [Online], 1(4), 78-87. Available: http://www.cs.caltech.edu/~adam/papers/xml/x-marks-the-spot.html [1998, March 31].

Kaku, M. (1997). *Visions: How science will revolutionize the 21st century* (p. 4). New York: Anchor Books, Doubleday.

Lamberton, D. (1992). Cyberspace economics. In Cook, B. (Ed.), *The electronic journal: The future of serials-based information* (pp. 89-99). Binghamton, NY: Haworth Press.

Lewis, S. & Kelley, T. (1995, November). Project muse: Tackling 40 journals. In Okerson, A. (Ed.), *Scholarly publishing on the electronic networks: Filling the pipeline and paying the piper* (pp. 103-112). Proceedings of the fourth symposium. Washington, D. C.: Association of Research Libraries.

McMillan, G. (1995). Scholarly communications project: Publishers and libraries. In Okerson, A. (Ed.), *Scholarly publishing on the electronic networks: Filling the pipeline and paying the piper* (pp. 135-145). Proceedings of the fourth symposium. Washington, D. C.: Association of Research Libraries.

Milligan, J. (1997, November 10). XML - *New hypermedia/web's future* [Online]. Available: http://128.230.1.252/archives/recmgmt.html [1998, March 31].

Networked digital library of theses and dissertations [Online]. (1998). Available: http://www.ndltd.org [1998, April 1].

Owens, E. (1993). Electronic text and scholarly publishers: How and why? In Okerson, A. (Ed.), *Scholarly publishing on the electronic networks, the new generation: Visions and opportunities in not-for-profit publishing* (pp. 65-72). Washington, D. C.: Association of Research Libraries.

Quarterman, J. (1997). *1997 users and hosts of the internet and the matrix* [Online]. Available: http://www.mids.org/press/pr9701.html [1998, March 31].

Ream, D. (1993). *The University of Virginia's electronic text center: An interview with David Seaman. Virginia Librarian, 39(2)*, 6-10 [Online]. Available: http://etext.lib.virginia.edu/articles/VirgLib/virglib.html [1998, March 31].

Rous, B. (1993). Electronic publishing: A five-year plan. In Okerson, A. (Ed.), *Scholarly publishing on the electronic networks, the new generation: Visions and opportunities in not-for-profit publishing* (pp. 29-42). Washington, D. C.: Association of Research Libraries.

Taft, D. K. (1998). Developers rally to promote XML—Sixteen-member council seeks new era of active-content apps. *TechWeb News,* March 23, 1998 [Online]. Available: http://www.techweb.com/se/directlink.cgi?CRN19980323S0106 [1998, April 2].

Van Houweling, D. E. (1994). Knowledge services in the digitized world: Possibilities and strategies. In Chiang, W. S. & Elkington, N. E. (Eds.), *Electronic access to information: A new service paradigm* (pp. 5-16). Proceedings from a symposium held July 23 through 24, 1993. Mountain View, CA: The Research Libraries Group, Inc.

Vanoirbeek, C. & Coray, G. (1992). *EP92:Proceedings of electronic publishing, 1992* (p. i). Cambridge, UK: Cambridge University Press.

von Hagen, J. L. (1992). The electronic journal: Is the future with us? In Cook, B. (Ed.), The electronic journal: *The future of serials-based information* (pp. 3-16). Binghamton, NY: Haworth Press.

World Wide Web Consortium. (1998, April 4). *Technical reports and publications* [Online]. Available: http://www.w3.org/TR [1998, April 4].

Waldt, D. (1997). SGML at Seybold? Can it be true? <TAG>: *The SGML Newsletter*, 10(10).