# Is XML in Your Future?

Art Rhyno

*Presenter*

**SUMMARY.** What is the significance of XML for library services? This article looks at the potential impact of some current XML technologies on libraries and identifies some key XML applications for moving library information between systems. The importance of XML for future directions in content management, metadata, and library standards like MARC are examined. *[Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <getinfo@haworthpressinc.com> Website: <http://www.HaworthPress.com>]*

In the Pulitzer Prize winning book, Guns, *Germs,* and Steel, author Jared Diamond explores the factors that have impacted the fates of human societies, including the use of the written word in supporting the flow of information. Diamond notes that the Sumerians' introduction of phonetic representation, in which a symbol represents a sound rather than an object, may have been the most important single step in the whole history of writing. The Sumerian record keepers had started with pictographs, which required a different picture for each word, but this led to the need for a huge number of consistent symbols for anything but the simplest records. By using phonetic signs, the Sumerians created building blocks that would be used by many others for developing writing systems for thousands of years, and would help societies to effi-

[Haworth co-indexing entry note]: "Is XML in Your Future?" Rhyno, Art. Co-published simultaneously in The Serials Librarian (The Haworth Information Press, an imprint of The Haworth Press, Inc.) Vol. 42, No. 1/2, 2002, pp. 143-153; and: *NASIG 2001: A Serials Odyssey* (ed: Susan L. Scheiberg, and Shelley Neville) The Haworth Information Press, an imprint of The Haworth Press, Inc., 2002, pp. 143-153. Single or multiple copies of this article are available for a fee from The Haworth Document Delivery Service [1-800-HAWORTH, 9:00 a.m. - 5:00 p.m. (EST). E-mail address: getinfo@haworthpressinc.com].

ciently transmit technologies, discoveries, and the "commands of the monarchs and merchants who organized colonizing fleets."[1]

## FROM SUMERIAN WRITING TO THE WEB

Like the ancient Sumerian record keepers, Web developers are often confronted with the difficulty of assigning consistent meaning to pieces of data. HTML (Hypertext Markup Language),[2] which underpins almost every Web page in your browser, is concerned with presentation and not content. With HTML, a fixed number of elements or "tags," such as <h1></h1> and <b></b>, control how information is presented on a Web browser. For example, the title of an archival finding aid on the Web might be presented as "<h1>Bruce J. S. Macdonald Papers, 1896-1986</h1>" or "<strong>Bruce J.S. Macdonald Papers, 1896- 1986</strong>." While most Web users may be able to discern the title of a resource with HTML, it is much trickier for a computer that is trying to index or manipulate information based on how the resource is structured.

XML (extensible Markup Language)[3] solves this problem by allowing meaningful tags to be defined. For example, in the text "<titleproper>Bruce J. S. Macdonald Papers, 1896-1986</ titleproper," the tags "<titleproper>" and "</titleproper>" delimit or "mark" the proper title of the finding aid. Like HTML, XML has its roots in SGML (Standard Generalized Markup Language)> a standard that was approved by the International Organization for Standardization (ISO) in 1986 for creating new markup languages. SGML has built up a somewhat deserved reputation for complexity and size, despite having a solid track record for handling content as diverse as scribbled notes, helicopter manuals, and stone tablets. In order to bring SGML forward into a Web-enabled universe, the decision was made by the World Wide Web Consortium (W3C) to simplify and apply some well-chosen rules to SGML while retaining SGML's ability to create and specify tag sets. In February 1998, the W3C published the XML 1.0 specification. It turned out to be a watershed event for the Web as almost all W3C standards are now based on XML. From here on, any substantive Web technology is seemingly expected to define its relationship with XML.

## THE TAG'S THE THING?

XML can be used to define new tags to describe the structure of a resource but it is important to understand that this alone is not what makes

XML an important Web technology. If additional tags are arbitrarily defined for the title of a finding aid, such as <thisisthepropertitle> and <propertitleishere>, and are used instead of <titleproper>, any program or technology trying to work with the aid may miss a critical component of the document's structure. To avoid inconsistencies between tag sets, XML retains SGML's concept of a DTD (Document Type Definition), a mechanism that defines the tag sets or "vocabularies" for an application and their proper use. DTDs have long been a powerful component of SGML, allowing programs to "validate" content, both in the process of creating the content and in ensuring it can be processed by other programs.

While DTDs are the most common validation tool to be used with XML documents, the W3C has recently standardized a successor to DTDs called XML Schema[5] that actually uses XML to define vocabularies instead of the cryptic syntax required by DTDs. Schemas make it easier to share vocabularies, so that an existing schema can be used but overridden at the point a new feature is needed. Schemas are also a good example of the kind of tools that the W3C is continually introducing to support XML activities. These technologies are license-free and are backed by industry heavyweights as well as a worldwide community of developers.

Although it is far beyond the scope of this paper to describe all the XML-related efforts emerging from the W3C, it is important to note that XML represents a litany of powerful technologies that are constantly under active development. Some of the applications fueled by these technologies may dramatically impact how libraries interact with information and the Web, and a few broad categories deserve special mention.

## *CONTENT PUBLICATION AND MANAGEMENT*

XML is often described as the successor to SGML, and it is not surprising that many SGML applications are being recast in XML. For example, the Text Encoding Initiative (TEI)[6] standard for textual materials and the Encoded Archival Description (EAD)[7] standard for finding aids are well-established SGML applications that provide the backbone for many digital library collections. By moving these standards to XML, TEI and EAD materials can now be maintained in XML with a much greater variety of mainstream tools and can be published on the Web and in other formats with a number of different presentation technologies, especially style sheets. Style sheets come in two powerful flavors: CSS (Cascading Style Sheets)[8] and XSL (Extensible Stylesheet Language).[9] When combined with a transformation tool called XSLT (Extensible Stylesheet Language Transformations),[10] a style sheets allow XML content to be dynamically transformed to HTML and virtually any other format. Style sheets are to presentation what XML is to content, and make for a clean separation as well as a greater degree of sanity when trying to support both content management and flexible publication options for the same collection.

XML-aware content publication and management frameworks can also assist organizations dealing with large amounts of XML content. The Cocoon[11] publishing framework from Apache[12] provides a free server-based implementation for serving XML documents to non-XML sources and can even produce Portable Document Format (PDF)[13] files on request. The popular open source application server Zope[14] has extensive XML capabilities, including a powerful search engine, and Endeavor Information Systems has recently introduced a product called ENCompass[15] that can manage XML documents in addition to other digital library resources.

TEI and EAD are just two of many standards for dealing with structured documents and almost any content that needs to be managed for network access can benefit from XML. HTML itself can be marked up in XML using a standard called XHTML.[16] The result is cleaner and easier-to-manage HTML content. XML editors, such as Altova's XML Spy and SoftQuad's Xmetal, can perform validation of XML documents and provide a sophisticated XML editing environment, and XML capabilities are currently available for both Word and WordPerfect. On the database side, Oracle's XQL and Microsoft's SQL Server 2000 can process and format XML based on existing data, and numerous approaches are being taken to "XML-enable" legacy systems. And this is just the tip of the iceberg. If you have ever dealt with moving documents between different word processors and in and out of the Web, you will appreciate the eagerness on the part of application developers to include XML support in their products.

## *INTEGRATION*

Libraries frequently need to integrate disparate systems. In a university setting, for example, the library may want to load records from the registrar's office. Libraries of all types import and export MARC records and many libraries are starting to use EDI (Electronic Data Interchange) in technical processing. EDI has been one of the first "business-to-business" integration activities to embrace XML because it allows application developers to leverage mainstream XML tools. As EDI becomes cheaper and easier to implement, libraries may benefit from the move towards retooling EDI to work as XML-based "interactive" transactions enabled by the Web rather than the cumbersome "system" or "batch" loading that is often done now.

In addition to energizing existing standards, XML has the potential to help libraries offer new services by tapping into systems that are usually out of the library's reach. For example, if a patron wants to load due dates for library materials into a PDA (Personal Digital Assistant) or import the information into a desktop calendar system, an XML standard for defining calendar information makes this integration much more possible. The library system can conceivably output the due date information in an XML format which can then be imported into the calendar application.

XML is a great way to bring data together because virtually any type of information that humans have ever attempted to manage seems to be the subject of an XML standardization initiative, from recipe sharing[17] to mind reading.[18] Table 1 lists just a few XML-based applications for moving data with XML.

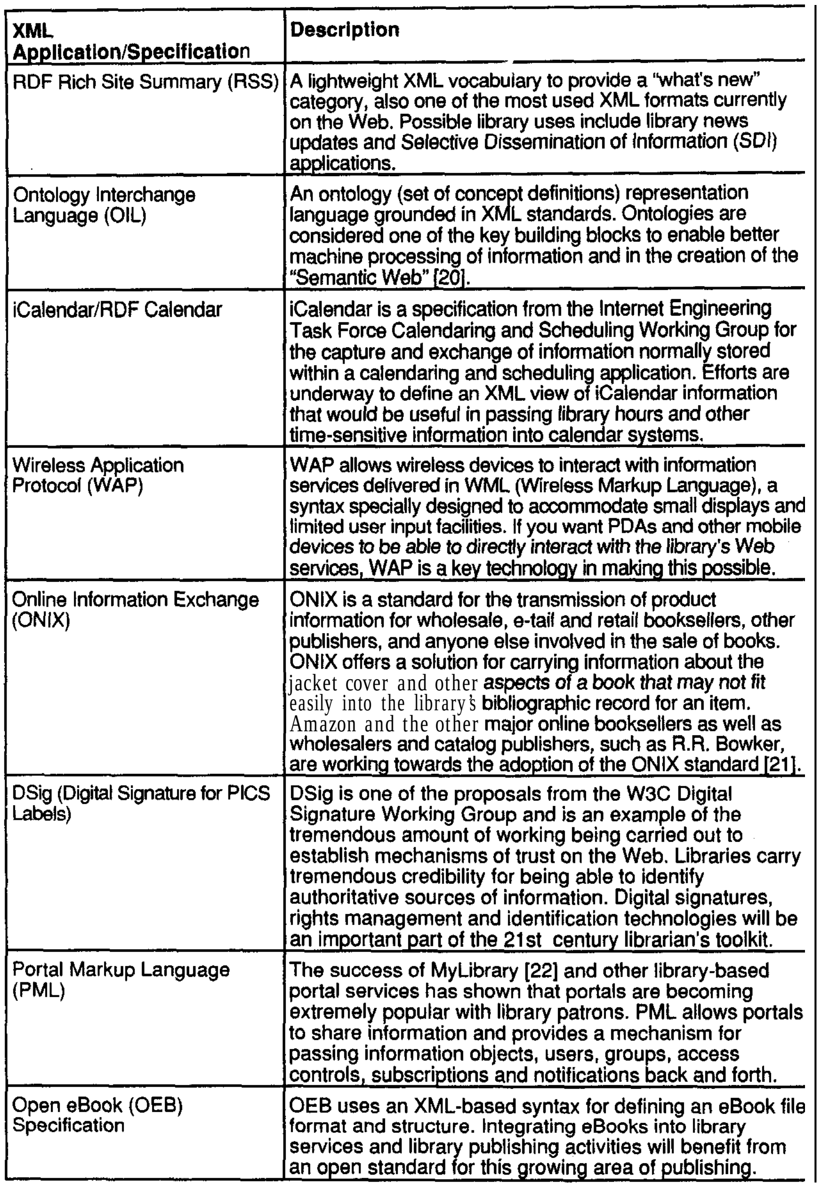
To fully appreciate all the options, one of the best starting points is Robin Cover's *The XML Cover Pages,* [19] an amazingly up-to-date and comprehensive site for tracking emerging XML initiatives, An active electronic discussion of the uses of XML in libraries can also be found in the XML4Lib list hosted by the Berkeley Digital Library SunSITE. [Notes 20, 21, and 22 appear in Table 1. *Ed.* ]

## *METADATA*

Metadata is typically defined as "data about data" and deals with problems related to the description of resources and resource discovery. Although the term "metadata" is often associated with Internet resources and has become fashionable in many popular technical magazines, libraries have created metadata about bibliographic materials for most of their existence, conveying rules for its creation through several generations of cataloging codes.

In 1997, the W3C announced the first draft of the Resource Description Framework (RDF),[23] the central component of its metadata activi-

TABLE 1



| XML Application/Specification | Description |
|---|---|
| RDF Rich Site Summary (RSS) | A lightweight XML vocabulary to provide a "what's new" category, also one of the most used XML formats currently on the Web. Possible library uses include library news updates and Selective Dissemination of Information (SDI) applications. |
| Ontology Interchange Language (OIL) | An ontology (set of concept definitions) representation language grounded in XML standards. Ontologies are considered one of the key building blocks to enable better machine processing of information and in the creation of the "Semantic Web" [20]. |
| iCalendar/RDF Calendar | iCalendar is a specification from the Internet Engineering Task Force Calendaring and Scheduling Working Group for the capture and exchange of information normally stored within a calendaring and scheduling application. Efforts are underway to define an XML view of iCalendar information that would be useful in passing library hours and other time-sensitive information into calendar systems. |
| Wireless Application Protocol (WAP) | WAP allows wireless devices to interact with information services delivered in WML (Wireless Markup Language), a syntax specially designed to accommodate small displays and limited user input facilities. If you want PDAs and other mobile devices to be able to directly interact with the library's Web services, WAP is a key technology in making this possible. |
| Online Information Exchange (ONIX) | ONIX is a standard for the transmission of product information for wholesale, e-tail and retail booksellers, other publishers, and anyone else involved in the sale of books. ONIX offers a solution for carrying information about the jacket cover and other aspects of a book that may not fit easily into the library's bibliographic record for an item. Amazon and the other major online booksellers as well as wholesalers and catalog publishers, such as R.R. Bowker, are working towards the adoption of the ONIX standard [21]. |
| DSig (Digital Signature for PICS Labels) | DSig is one of the proposals from the W3C Digital Signature Working Group and is an example of the tremendous amount of working being carried out to establish mechanisms of trust on the Web. Libraries carry tremendous credibility for being able to identify authoritative sources of information. Digital signatures, rights management and identification technologies will be an important part of the 21st century librarian's toolkit. |
| Portal Markup Language (PML) | The success of MyLibrary [22] and other library-based portal services has shown that portals are becoming extremely popular with library patrons. PML allows portals to share information and provides a mechanism for passing information objects, users, groups, access controls, subscriptions and notifications back and forth. |
| Open eBook (OEB) Specification | OEB uses an XML-based syntax for defining an eBook file format and structure. Integrating eBooks into library services and library publishing activities will benefit from an open standard for this growing area of publishing. |

ties. Originally an extension on W3C's PICS (Platform for Internet Content Selection)[24] description technology, RDF would draw on several metadata-related proposals, including an infrastructure called the Warwick Framework[25] designed to broaden the scope of the Dublin Core Metadata Setz[26] by representatives from industry and the library, research and academic communities. The Warwick Framework was designed to support any metadata vocabulary, and RDF allows multiple metadata vocabularies to be used together. Libraries have had great success in providing access to Internet resources that meet the library's selection criteria and applying rich descriptive data to such resources. Packaging these records into RDF and sharing the results with search engine providers and Internet directory initiatives like the Open Directory Project[27] may be important contributions that libraries can make to the Web community's efforts to provide better access to Web content.

The widespread interest in XML-based metadata technologies has also resulted in many different Web applications for information navigation, and it is possible that libraries will be able to take advantage of some of these developments to augment access to the physical and digital objects in the library's collection. For example, Topic Maps create a virtual organization and navigation layer above diverse electronic resources.[28] There is also a project in connection with the Harmony Project[29] called MetaNet[30] that seeks to enable semantic mappings between synonymous metadata terms from different vocabularies. Sharing metadata is also at the heart of the Open Archives Initiative (OAI),[31] a protocol that enhances access to e-print archives as a means of increasing the availability of scholarly communication.[32] With OAI, data providers can expose metadata about available content, and e-print archives can be accessed and queried in one step.

## *MARC*

Perhaps no other library standard may be impacted as dramatically by XML as MARC. Lane Medical Library at Stanford has initiated an exciting experiment by releasing a DTD, software, and a host of related resources for moving MARC records to what has been christened XMLMARC.[33] The European Union project ONE-2 has also published a discussion paper about using XSLT for MARC XML Record Conversion.[34] Recent work on combining XML and RDF Schemas[35] holds great promise for bringing together the syntactic side of MARC creation with the semantics represented by the metadata assigned to a resource. XML Schema holds the key for a flexible syntactic validation while RDF Schema exposes the metadata to other applications. MARC records could move from being largely library-specific entities requiring quirky editing programs to mainstream resources that can be pulled into and updated by third party applications.

In looking at MARC, it is worth examining what lessons EDI may provide to the library community about the process of taking an existing computer standard and re-architecting it for XML. Some early and existing EDI implementations in XML are literal mappings between EDI's cryptic syntax and a set of only slightly less cryptic XML tags. More recent efforts like SIMPL-EDI[36] concentrate on core EDI requirements and attempt to leverage previous EDI experience against what needs to be in the XML version.[37] Similarly, Lane has constructed a DTD based on what MARC fields are actively used. While some MARC constructs may be trimmed in the move to XML, others are ripe for expansion, particularly the 856 field that records electronic location and access information.

## THE WEB AS A GLOBAL DESKTOP

The Web browser started out as an application that sat on the edge of the desktop. Now it sits at the center and is, in some cases, so close to the operating system that the distinctions between one and the other are starting to blur. Microsoft has tightly integrated Internet Explorer into the desktop, and as part of its .Net[38] initiative, uses XML and an XML-based protocol called SOAP (Simple Object Access Protocol)[39] to pass information between network services so it can follow a user around on the network. Mozilla,[40] the open source browser development started by Netscape, uses a technology called XUL (XML-based User Interface Language)[41] that can be used to customize the browser and expand the options for delivering a rich interface for Web interaction. The W3C has also reworked Web forms with a standard called Xforms[42] that will make using forms far less clunky than they are now. In combination with other XML and non-XML Web technologies, the Web browser is becoming a viable focal point for mainstream applications.

Imagine linking the Oxford English Dictionary Web service to your word processor while seamlessly drawing on bookmarks, annotations, and other Web resources, and having the ability to interact with the majority of desktop applications from any Web-enabled device. XML has a key role in making the Web the premier entry point for a mobile desktop.

## *ANCIENT INNOVATIONS AND MYSTERIOUS DISKS*

While the introduction of phonetic representation by the Sumerians would be passed forward to modem times, Jared Diamond uses the famous and still mysterious Phaistos Disk as an example of an innovation that was lost to future generations despite having obvious value to its creators.[43] The Phaistos Disk, discovered in 1908 in an excavation of the ancient Minoan palace at Phaistos on the island of Crete, is a series of signs applied to clay that represent one of humanity's most ingenious attempts to build a printing system before the ideas of ink and the printing press were formed a thousand years later.[44] With XML, libraries have an opportunity to avoid the fate of the Phaistos Disk and plug our fascinating but not always appreciated technologies into a phenomenon that may be as profound as the creation of the Web itself. Is XML in your future? You bet it is!

## NOTES

1. Jared M. Diamond. Guns, *Germs, and Steel.* New York: W. W. Norton & Co., 1997, 218-222.

2. World Wide Web Consortium (W3C), *HyperText Markup Language Home Page.* http://www.wc3.org/MarkUp/.

3. W3C, *Extensible Markup Language (XML),* http://www.w3c.org/XML/.

4. W3C, *Overview of SGML Resources,* http://www.w3c.org/MarkUp/SGML/.

5 W3C, *XML Schema, http://www.w3c.org/XML/Schema/.*

6. W3C, *Text Encoding Initiative Home Page. http://www.uic.edu/orgs/tei/,*

7. W3C, *Encoded Archival Description (EAD) Official Web Site.* http://www.loc.gov/ead/.

8. W3C, *Cascading Style Sheets Home Page.* http://www.w3org/Style/CSS/.

9. W3C, *Extensible Stylesheet Language (XSL).* http://www.w3.org/Style/XSL/.

10. W3C, *XSL Transformations (XSLT) Version 1.0.* http://www.w3.org/TR/xslt/.

11. Apache XML Project, *Cocoon.* http://xml.apache.org/cocoon/.

12. *The Apache Software Foundation.* http://www.apache.org/.

13. Adobe Systems Incorporated, 2001; see reference specifications and patent notes at: http://partners.adobe.com/asn/developer/technotes/acrobatpdf.html.

14. *Welcome to Zope.org:* http:/www.zope.org; Zope has several XML tools, but the latest is ParsedXML: http://www.zope.org/Members/karl/ParsedXML/Parsed XML.

15. Endeavor Information Systems, 2001; see product description at.http://www.endinfosys.com/prods/encompass.htm.

16. *XHTML* TM *1.1 -Module-based XHTML.* http://www.w3.org/TR/xhtm111/.

17. *RecipeML-The Recipe Markup Language.* http://www.formatdata.com/recipeml/index.html.

18. *Mind Reading Markup Language.* http://www.owlnet.rice.edu/~shiwala/mrml.html.

19. *The XML Cover Pages-Home Page.* http://xml.coverpages.org/sgml-xml.html.

20. For an introduction to the Semantic Web see Berners-Lee, Tim, James Hendler, Ora Lassila, "The Semantic Web," *Scientific American*, May 2001. http://www.sciam.com/2001/0501issue/0501berners-lee.html; one ongoing debate on the possibilities of the Semantic Web can be found on *Slashdot:* http://slashdot.org/ askslashdot/01/03/21/0739222.shtml. I thank Kevin S. Clark (ksclarke@stanford.edu) for this reference.

21. *Content Guard Home Page,* http://www.contentguard.com.

22. A good starting point for learning more about MyLibrary is Morgan, Eric Lease, "MyLibrary@NCState: The Implementation of a User-centered, Customizable Interface to a Library's Collection of Information Resources" http://my.lib.ncsu.edu/about/sigir-99; see also the development page: http://hegel.lib.ncsu.edu/development/mylibrary/.

23. W3C, *Semantic Web Activity: Resource Description Framework (RDF).* http://www.w3.org/RDF/.

24. W3C, *Platform for Internet Content Selection (PICS).* http://www.w3.org/PIGS/.

25. A good treatment of the development of the Warwick Framework can be found in: Lagoze, Carl, "The Warwick Framework: A Container Architecture for Diverse Sets of Metadata," *D-Lib Magazine,* July/August 1996, http://www.dlib.org/dlib/njuly96/lagoze/07lagoze.html.

26. *Dublin Core Metadata Initiative,* http://dublincore.org/.

27. *ODP-Open Directory Project,* http://dmoz.org/.

28. Eric van der Vlist has written an article that nicely ties together RSS, RDF and Topic Maps, see: van der Vlist, Eric, "Building a Semantic Web Site," *XML.com,* http://www.xml.com/lpt/a/2001/05/02/semanticWebsite.html.

29. *Harmony Main Page,* http://www.ilrt.bris.ac.uk/discovery/harmony/.

30. *MetaNet Query,* http://sunspot.dstc.edu.au:8888/Metanet/Top.html.

31. *OAI Home Page,* http://oaisrv.nsdl.cornell.edu/.

32. The role of OAI in scholarly communication is described in Luce, Richard E., "The Open Archives Initiative: Interoperable, Interdisciplinary Author Self-arching Comes of Age," *NASIG Conference Proceedings,* San Diego, June 22-25 2000, http://lib-www.lanl.gov/lww/articles/oai_nasig2000.htm; see also: Tennant, Roy, "Open Archives: A Key Convergence," Feb. 15, 2000, http://www.libraryjournal.com/articles/infotech/digitallibraries/2000215_13666.asp.

33. *Medlane Experiment-MARC to XML,* http://xmlmarc.stanford.edu/.

34. *Resource Description Framework (RDF) Schema Specification 1.0,* http;//www.w3.org/TR/rdf-schema/.

35. Hunter, Jane and Carl Lagoze, "Combining RDF and XML Schemas to Enhance Interoperability Between Metadata Application Profiles," Tenth International World Wide Web Conference May 1-5, 2001, Hong Kong, http://archive.dstc.edu.au/RDU/staff/jane-hunter/www10/paper.html.

36. *What is SIMPL-EDI?* http://www.ecommerce.ac.uk/esc/simpl-edi.html.

37. Arofan T. Gregory examines this approach for EDI in "XML schema design for business-to-business e-commerce," XML Europe 2000, June 12-16, 2000, France, htp://www.gca.org/papers/xmleurope2000/papers/s2l-01html.

38. *Microsoft.NET,* http://www.microsoft.com/net/.

39. *Simple Object Access Protocol (SOAP) 1.1, http://www.w3.orglTR/SOAP/.*

40. *mozilla.org,* http://www.mozilla.org/.

41. *Iction to XUL,* http://www.mozilla.org/xpfe/xptoolkit/xulintro.htrnl.

42. W3C, *Xforms-The Next Generation of Web Forms,* http://www.w3.org/MarkUp/Forms/.

43. Diamond, p. 239-241.

44. There are many resources on the Phaistos Disk, but one of the most enjoyable is *The Phaistos Disk-The Oldest Hard Disk Ever,* http://www.fsai.fh-trier.de/~mourisj/diskos/diskos.html; a somewhat more reverent description can be found at *The Phaistos Disk,* http://www.millennia.org/phaistos.html.

## CONTRIBUTOR'S NOTE

Art Rhyno is Head of Systems, University of Windsor, Ontario, Canada.