

Use XML to Structure and Manage Content

Rita E. Knox

Extensible Markup Language makes content accessible to multiple applications, including content creation. Applications are moving XML to mainstream use, increasing content integrity and reducing front and back-end costs.

WHAT YOU NEED TO KNOW

The major challenge for enterprises when using Extensible Markup Language will be to know which tools are appropriate for the level of metadata creation and management they require for the processes they want to support. "XML inside" is not enough. XML data models and sophisticated XML-aware processes will be needed.

STRATEGIC PLANNING ASSUMPTION(S)

By year-end 2009, 25 percent of enterprises will use XML-based document processes (0.6 probability).

By year-end 2007, 40 percent of knowledge workers will use XML-aware content-creation tools (0.7 probability).

Through 2006, XML-enabled e-forms will at least double in use as a standard enterprise document format (0.7 probability).

By year-end 2007, XML component management will be an explicit offering of most content management vendors (0.6 probability).

ANALYSIS

Extensible Markup Language (XML) is becoming the tool that knowledge workers use to structure content and ensure semantic consistency throughout the enterprise. By year-end 2009, 25 percent of enterprises will use XML-based document processes (0.6 probability).

"Structuring" Content: Structuring is not a matter of slotting each and every piece of information into identical structures, but of defining and labeling content components and their structures. By identifying the components, various applications can process them in different ways. For example, components could be title, summary, main points, footnotes and references. The structure of those components might be:

- *Title:* A brief (10 or fewer words) description labeling the enclosed content
- *Summary:* A three-line paragraph providing a precis of the content
- *Main point:* A section label and the content for the section
- *Footnote:* A mnemonic label for an in-content (in text or in graphic, in table) reference along with the explanation of the reference associated with the mnemonic label
- *Reference:* A pointer to another piece of content using some standard labeling convention (for example, keyword, author/date)

Application Processing: Different applications can interact with this structured, identified content in different ways. For example:

- *Content creation:* A structured editing environment can guide an author to enter different pieces of content ("components") and indicate what's required or optional. A component might have multiple parts that have particular relationships to each other — for example, a "sidebar" (that is, a short article) on a company might include its industry code, address, phone number and primary products.

- *Content presentation:* A formatting application can format content for different channels (for example, print, Web, audio) based on the content identifier. For example, a "title" might be centered, bolded in print and used as a link to the content in a Web environment.
- *Data interchange:* A "reference" can be the link between the content and back-end data resources — a relational database management system (RDBMS), the Web, or another document that causes content to be pulled into a content stream, or written to a database or third-party information resource.

History: Identifying content components for processing has evolved during the past 25 years through numerous related technology efforts. Structuring and identifying content began with Standard Generalized Markup Language (SGML; see Note 1) in the late 1980s and evolved into XML in the late 1990s (see Note 2).

Industry Pull: XML is poised to support knowledge workers, enabling process automation and significantly affecting traditional workplace applications, such as content creation and management. The IT standards associated with technical aspects of XML need to be tied to industry specifics for content and context. Knowing the technical standards behind an XML-based document or message supports application integration but does not assure or make adoption happen. What often encourages adoption is deployment of such standards by large, dominant enterprises in value chains or by industry-related communities of interest (see Note 3).

Why XML? Computers must be explicitly told what a text stream means to do something with it. Components must be identified and labeled so that applications can process them. XML does this and is the reason it's used. Using plain text (as opposed to complex data structures) is a "great leap forward" in enterprise computing. XML is a simpler specification than SGML. It is stable (that is, has not changed since it was originally issued) and has a growing set of applications that can process it. Because XML uses plain text, it is malleable in a way that processing code is not.

Uses of XML in Content Processing Applications

XML-Aware Content Creation in 2004: Features of XML-enabled content creation tools are to:

- Create or modify document models — templates in the form of schema or document type definitions (DTDs) — to control content structure
- Provide an editorial environment to create and validate the content in a model-conforming mode
- Integrate with a content management environment to access the content at some level of the granularity reflected by the document model (see Note 4)

Numerous vendors and products provide such XML-based functionality.

1. Vendors with an SGML heritage: Arbortext's Epic Editor, Adobe FrameMaker and Blast Radius' XMetaL (formerly owned by Corel). These tools have supported the aerospace, automotive, pharmaceutical and publishing industries for complex publishing applications.

2. XML-enabled versions of Microsoft Word include HyperVision's WorX, I4I's x4o and In.vision's Xpress Author for Microsoft Word.

These first two application categories support modification of the XML model, as well as the content itself.

3. Word processors have XML-aware content-editing capability, although they don't support DTD or schema editing.

- WordPerfect 12 supports XML in two modes. In standard WordPerfect, users can use style tags to mark up content to create well-formed XML. In XML mode, users enter data in templates to create valid XML that conforms to a DTD. WordPerfect can export XML conforming to the DocBook XML 4.2 DTD.
- Word 2003 supports editing in two modes: without XML validation (that is, using a standard version of Microsoft Word) and with validation (creating content in an XML-aware mode). XML content creation uses Microsoft-created schema, as well as user-selected schema, and exports XML data to conform to the user-designated schema.
- The open-source OpenOffice productivity suite can exchange data in XML but is not an XML-aware content-creation environment.

4. Ektron eWebEditPro+XML is a Web-based word processor with data validation and forms-creation capability. Altova's Authentic 2005 is a simple (and free of charge) XML form editor; content is validated in an immutable data model.

5. Document composition (aka database publishing) vendors that use XML-based approaches include Document Sciences (xPression), Exstream and Thunderhead.

By year-end 2007, 40 percent of knowledge workers will use XML-aware content-creation tools (0.7 probability).

Component-Based Content Management: XML-aware content management has been available but struggled when XML was much less in demand in the enterprise. As XML has become more mainstream, and the quantity of XML-identified data has grown, the need for XML-aware content management has also grown.

Document-centric database management approaches emerged in the late 1990s. Software AG is the only vendor of that era that is still active; however, it is doing more partnering with other vendors than selling Tamino directly as a specialized data management tool. XyEnterprise (Content@) emerged from the SGML world, as did Astoria Software, which has emerged with XML content management. These vendors offer specialized tools that store and manipulate data at the level of element tags and attributes. Some of the tools (for example, Content@) use an RDBMS to which they provide an intermediate layer interface for XML interchange. In.Vision's Xpress for Microsoft Office SharePoint is a new entrant with a specific content management integration. The ebXML Registry Information Model is the component data model often used as the metadata layer that interfaces to RDBMS repositories.

Content management/document management vendors are extending their offerings to support XML granularity. Some are doing this through partnerships. Documentum and Open Text have offerings but lack the level of functionality of the XML-aware repositories. Stellent offers integration with Software AG's Tamino. By year-end 2007, XML component management will be an explicit offering of most content management vendors (0.6 probability).

The RDBMS vendors have extended their functionality to handle XML data in numerous ways, "structured" being the closest to XML-aware repositories. Oracle Database 10g is the strongest in this group, also supporting the W3C's XML Query (XQuery) specification. IBM DB2 v.8.1, Sybase ASE and Microsoft SQL Server 2000 are other offerings.

Structured and Unstructured Data Are Merging: The evidence that structured data is fitting with unstructured data can be seen most dramatically in the growing use of electronic forms (e-forms). The textual identifiers of information and paragraphs of information are created for the benefit of human users. However, the inclusion of data models in the data fields enables data interchange between front-end forms and back-end systems that can process them. The vendors that provide e-forms technology to commercial and government organizations create these

computer-processable data descriptors using XML data models. (Document composition applications are using XML technology but generally do not use the level of XML processing seen with e-forms.)

Vendors are making investments to deploy e-forms technology in desktop-accessible applications. Through 2006, XML-enabled e-forms will at least double in use as a standard enterprise document format (0.7 probability). Examples are:

- Adobe Systems' deployment of Portable Document Format (PDF) e-forms, created with the Adobe (e-form) Designer and fillable with the free-of-charge Adobe Reader — used worldwide
- Microsoft's InfoPath as part of Office 2003
- PureEdge 8x
- Verity Liquid Office
- FileNet Forms Manager

XML e-forms integrate data-centric information with a front-end presentation to, and data input by, users. This is used in insurance claims processing, e-procurement, financial and e-commerce applications.

Key Issues

What role does XML play in managing information and data?

What are the best practices for integrating structured and unstructured data, and to develop and maintain "information architectures"?

Note 1

SGML

The approach to structuring content began in the defense, aerospace and automotive industries (where complex publishing models are a known challenge with enormous financial implications) and the publishing industry (where management of the content for reuse, re-purposing and multiple channels has been key to that industry's profitability).

SGML supported publishing initiatives in the late 1980s/early 1990s with the Computer-Aided Acquisition and Logistics Support (renamed Continuous Acquisition and Life-Cycle Support) program from the U.S. Department of Defense. SGML then moved into the mainstream for publishing applications, primarily in manufacturing and professional publishing. SGML could have been used as a data description language with applications broader than publishing, but it never developed beyond its original publishing uses.

Note 2

Enter XML

XML was issued as a World Wide Web Consortium (W3C) recommendation (the equivalent of a standard) in February 1998. Unlike SGML, which had been used primarily for publishing applications, XML was used more broadly (for example, to define the Mathematical Markup Language for mathematical expressions, and by Open Financial Exchange to describe consumer-to-bank transactions). Uses quickly expanded to other areas of financial services, manufacturing, HR, real estate and bioinformatics — an unending parade of "MLs" (markup languages) in many

domains. The relative simplicity of creating models to identify data elements had immediate and obvious appeal to many enterprises and vendors as an application-independent way of sharing data. XML is pervasive in application integration and serves as the foundation for Web services.

Note 3

XML Vertical Standards

Examples are the Voluntary Interindustry Commerce Standards for Collaborative, Planning, Forecasting and Replenishment, an industry standard for collaboration between retailers and their suppliers. RosettaNet serves a similar purpose. The XML data models deployed for property and casualty and life insurance from the Association for Cooperative Operations Research and Development (ACORD). The proliferation of industry-specific standards is a problem (see "Here's What's Wrong With XML-Defined Standards").

Note 4

XML Tools

A variety of XML-aware tools for middleware and application integration functions are not commonly used for the content-creation tasks discussed in this research. Vendors and the applications include BEA Systems (WebLogic Integration), Cape Clear (Studio, Data Interchange), IBM (WebSphere Business Integration family), Sybase (Adapter for XML, e-Biz Integrator), Tibco (Turbo XML, XMLValidate, XMLTransform, XMLCanon/Developer) and Vitria (BusinessWare family).

Acronym Key

ACORD	Association for Cooperative Operations Research and Development
DTD	document type definition
ML	markup language
PDF	Portable Document Format
RDBMS	relational database management system
SGML	Standard Generalized Markup Language
W3C	World Wide Web Consortium
XML	Extensible Markup Language

REGIONAL HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
U.S.A.
+1 203 964 0096

European Headquarters
Tamesis
The Glanty
Egham
Surrey, TW20 9AW
UNITED KINGDOM
+44 1784 431611

Asia/Pacific Headquarters
Level 7, 40 Miller Street
North Sydney
New South Wales 2060
AUSTRALIA
+61 2 9459 4600

Latin America Headquarters
Av. das Nações Unidas 12.551
9 andar—WTC
04578-903 São Paulo SP
BRAZIL
+55 11 3443 1509