How XML Supports a Unified Content Strategy

David M. Lentz

Introduction

Extensible Markup Language (XML) is a language that organizes and describes data. XML is not specific to technical communication or any industry. An XML document could contain data that prescribes a computer's configuration, or it could describe the content stored within a health-care-specific content management system. These two documents would differ in their content and structure, but each would be based on the rules of XML.

My purpose in writing this paper is to connect what I've learned about XML as a software developer with what I've learned about content management systems in this class. Our discussion, reading, and exercises relating to DITA showed me an application of XML different from what I've seen in the past. My goal in this assignment is to synthesize professional knowledge with new information from this semester's readings and discussion.

In this paper I will introduce and explain some fundamental characteristics of XML, and describe how these characteristics make XML useful for structured writing and content management.

What is XML?

XML is a markup language that allows users to create, organize, and describe content. XML files are plain text and can be viewed in any basic text editor. They contain data that is arranged inside structures called elements. Attributes can be added to elements to describe the data they contain.

W3C is the organization that has published the XML standard. W3C maintains recommendations that describe how XML works, its syntax and structure.

Programming Languages

Though XML is a language used within the domains of computers and the Internet, it is not a programming language. Programming languages influence the behavior of a computer, determining such things has how it manipulates data and how a user interacts with it. Modern computers can be thought of as general-purpose computing devices, in contrast to earlier single-purpose machines that were designed and constrained to execute only a small set of pre-determined functions. By applying a programming language within a general-purpose computer, programmers can create logic, rules, and behaviors that make the computer useful.

Markup Languages

By contrast, markup languages do not express logic or instructions. Instead, they organize and describe data. They arrange pieces of data using a structure that helps convey what each piece of data is, and how it relates to the other pieces of data. In the case of an XML-based content management system, the data is the content stored by the system.

A key characteristic of XML is separation of format from content. XML doesn't describe how its content is to look, only what it is. Formatting instructions can be contained in a separate file to determine the appearance of the content in a published document.

Another markup language is HTML, the language used to create web pages. Originally, HTML contained formatting information as well as data, mixing presentation and content. The data is the text and images displayed to the user, and the HTML markup provides instructions to the web browser on how the data should be presented.

The current version, HTML5, has more in common with XML. HTML5 separates presentation and content better than previous versions. Where previously HTML contained instructions to show text in a certain size or weight, HTML5 instead contains information about the data, and leaves out formatting instructions. Formatting can be described in a separate document, a Cascading Style Sheet, which gives the web browser specific information on how to display each type of data in the document.

XML's Relevant Characteristics

XML is the basis for content management systems and documentation standards used by technical communicators. Three characteristics of XML that facilitate these uses are its ability to represent structured content, its extensibility, and its standardization.

Structure

The structure of every XML document is based on elements and attributes. Elements take the form of a set of container tags. One tag designates the start of the container, and is followed by the container's data, and then a separate tag designating the end of the container. These containers may be nested, so that an XML structure can represent, for example, a list element existing inside a list.

Elements may have attributes associated with them. If present, an attribute appears as part of the element's start tag, and describes the element. Using attributes, an XML document can apply metadata to an element, describing the data it contains.

This structure is useful because it allows a simple file to represent an information product model. The IPM describes content structure defined by an information architect.

Though XML is human-readable and stored in plain text files (you can use a tool like Notepad to read one), there are many software applications designed specifically for editing XML documents. Because each well-formed XML document contains a standard collection of attributes and elements, XML editors can understand what to expect. The elements and attributes beyond the base collection that defines XML will adhere to syntax defined by the standard and rules defined by an associated schema (the IPM), so each of the various XML editors can know what to expect and can enforce the rules.

Extensibility

The tags used within an XML document are not limited to the basics of the XML standard. Once an XML document includes the minimum elements necessary, more tags can be added arbitrarily as needed. XML is extensible in that it "allows you to create the tags to suit the needs of your content and your authors." (Rockley & Cooper, 2012, p. 267)

XML's extensible quality means that it can be applied in new contexts. The elements and attributes that comprise an XML document in a given domain are extensions of XML. They aren't inherent in the language, but are added as needed when XML is put into use describing that domain's content.

As a result, using XML to describe a new domain doesn't require any change to the standard. Each application of XML starts from the same basic standard, and the nature of the language allows for domain-specific elements and attributes to be added to support each new use.

XML can scale to support the information needs of different types of organizations, and can evolve to continue to support them over time.

Standardization, Well-Formedness, and Validity

A well-formed XML document adheres to the rules defined in the XML standard. Though XML is human-readable, well-formedness is typically confirmed by software specifically designed to parse a document and alert the user if it fails to meet the XML standard. As an example, W3C (the organization that maintains the XML standard) provides an online tool called the Markup Validation Service at http://validator.w3.org. A user can upload an XML file and see whether it contains structural or syntactic errors.

For an XML document to be useful within a particular application, such as a CMS, it must also be valid. A valid XML document is one that adheres to the rules defined in a schema to which the XML document is associated.

Schemas contain rules and constraints. An XML document associated with a schema is held to those rules. When an authoring tool enforces the rules of XML and any applicable schema, an author's content meets the standards of the schema. This system of authoring tool and an associated schema holds authors throughout the organization to the same set of rules, and "can ensure that all your information products are structurally consistent." (Rockley & Cooper, 2012, p. 271)

In addition to a more consistent body of content, an XML-based content management system can create documents that are more likely to be reusable. As each author contributes content, the CMS enforces validity and well-formedness. The schema, which provides the rules for validation, represents the information product models on which the CMS is based. This system assures that all content adheres to the organization's IPMs. Because the CMS contains content that is consistent with the IPMs, the organization knows more about each piece of content. The content is findable according to any metadata requirements, and is written to a standardized structure. As a result, the organization is better at finding its own content, and knowing how to reuse it (given that the structure of the content is known.)

How XML Supports a Unified Content Strategy

Rockley and Cooper define a unified content strategy as a "repeatable method of identifying all content requirements up front, creating consistently structured content for reuse, managing that content in a definitive source, and assembling content on demand to meet customer needs." (Rockley & Cooper, 2012, p. 10) In this section I will discuss characteristics of XML that facilitate a unified content strategy. Then I will discuss DITA and DocBook, two XML schemas used in content management.

Content Definition

The first step in implementing a unified content strategy is identifying content requirements. These requirements can be encoded into an information product model, which Rockley and Cooper describe as "a hierarchical ordering of components" (Rockley & Cooper, 2012, p. 135). The order and hierarchy of an IPM can be expressed in an XML schema as a series of elements and attributes.

The schema is used to "define the required structure of a document." (Rockley & Cooper, 2012, p. 270) It reflects the IPM determined in the content definition phase, and can be used to enforce the rules of the IPM against any content creation that happens later in the process.

Content Creation

Though XML is human-readable and stored in plain text files (you can use a tool like Notepad to read one), there are a number of software applications designed specifically for creating and editing XML documents. These editors are programmed to recognize the attributes and elements that make up a well-formed XML document. Users can see and edit the XML data (though not necessarily the tags that make up its raw source code). The editor can enforce the rules of XML to assure the author's content remains well-formed and valid.

Because XML editors can enforce the rules embodied in the schema that represents the organization's IPM, content creators are guided in their work and constrained to producing content that is consistent with the rest of the organization's content. Different authors are held to the same rules, and the resulting content more consistent and reusable.

Content Management

Storing, organizing, and finding content are important in a unified content strategy. XML files are plain text, and not in a proprietary or specialized format. Text files are generally smaller than many other formats. These characteristics make it easy for an organization to store the files necessary for a CMS.

The XML standard is open source. By writing an appropriate schema, any organization can arbitrarily add to XML's ability to store and organize data. Modifying existing schemas to include new types of data doesn't require a change in software. Schema changes can be made without negatively affecting the organization's existing content, and can allow the organization to maintain an IPM even as its content needs evolve.

Content Delivery

XML documents contain content, but they avoid prescribing the appearance of that content. The colors and typefaces and other elements that make up a document's presentation are not encoded along with the data. Rather, these decisions are expressed in separate documents called XSL style sheets. An organization's different channels may have different style sheets, even though the content exists in only one source. This separation of content from presentation is one way XML facilitates reuse.

XML Applications

As a standard, XML does not favor any particular usage, industry, or application. It can be extended for specialized use within a particular domain. In this section I will describe two applications of XML created specifically for use in publishing and technical communication.

Because these methods and tools are based on XML, they demonstrate these qualities:

- An information model can be expressed as XML.
- XML can serve as the basis for tools and processes that enforce the information model.
- XML can serve as the basis for a method that enforces the separation of format and content.
- As an extensible language, XML can evolve and scale to support the information needs of organizations into the future.

DITA

Darwin Information Typing Architecture (DITA) is an XML implementation whose schema defines a set of four document topics: generic topic, concept, task, and reference. XML editors that support DITA will guide the author to adhere to the format of the topic. The collection of topics is stored in a CMS and compiled into deliverables within various channels through the use of DITA maps.

DITA maps facilitate reusability of content. A map defines how a particular deliverable such as a help system or brochure is composed of DITA topics. Once a topic is written, any number of maps may include that topic in a printed or online deliverable. This way, a single topic may be published in more than one channel.

DocBook

DocBook is a schema used to create content primarily for software and hardware documentation. DocBook is more mature and more complex than DITA. It has been in use since 1991 and is actively maintained. Several editing programs support DocBook, and some of them are available free of charge. The OpenOffice.org word processor is probably the most popular software that supports DocBook. Microsoft Word does not support DocBook.

Conclusion

XML is a language that provides a standard and neutral foundation. It allows experts to create schema that define structured data useful to a particular domain. In other words, information architects can express the IPMs their organizations need in the form of XML schemas. XML editors can enforce the rules of the schemas and the XML standard. This system helps authors create content that is reusable and consistent.

XML has been customized (or extended) for use in many applications, such as publishing, content management, word processing, and software configuration. It has potential to serve many more.

XML is standardized, lightweight, and human-readable. Yet it can serve as a foundation for a unified content strategy in a variety of industries. This means that a fundamental and free technology can provide organizations increased reusability and consistency. It can increase the effectiveness of editors, who can rely more on the XML editing software to automatically enforce the organization's IPM. It can allow organizations to better serve readers without increasing the size or expertise of their authoring teams.

XML and HTML5 act as the foundation for many documentation tools. These standards allow for open documents that can be generated using one tool, read by another and maintained by yet another. The market for software such as browsers and word processors appears at least as crowded as it has been at any time in the past, and this is healthy. As these different tools converge upon standard markup languages like XML and HTML5, it becomes evident how the languages benefit the audience, the content provider, and even the companies offering products that "understand" them. These standards are actively maintained, and the opportunity for those software companies is working with the standards rather than competing against them with a proprietary offering.

Works Cited

Rockley, A., & Cooper, C. (2012). *Managing Enterprise Content: A Unified Content Strategy.* Berkeley, CA: New Riders.