An Introduction to XML

By Stephen Turbek
22 January 2001 Razorfish Report 047

XML, the eXtensible Markup Language, is an open and flexible standard to format and exchange information. Though generally invisible to users, it is affecting computing, including Web servers, wireless devices, and voice recognition.

Mark Up Languages

The HyperText Markup Language (HTML) is the code used to write Web pages. It was created by Tim Berners-Lee as a simple way to add formatting to a text file. 'Tags' indicate where layout changes should 'mark up' the text. For example, `<b>**makes text bold**</b>` on the browser screen. XML and HTML both grew from SGML, a mark up language used in the publishing industry.

HTML was designed to make text more readable for humans, but it has precisely the opposite effect on readability for computers. The flexibility and inconsistency with which pages are built makes it impossible for a program to identify them in a page and extract that information out.

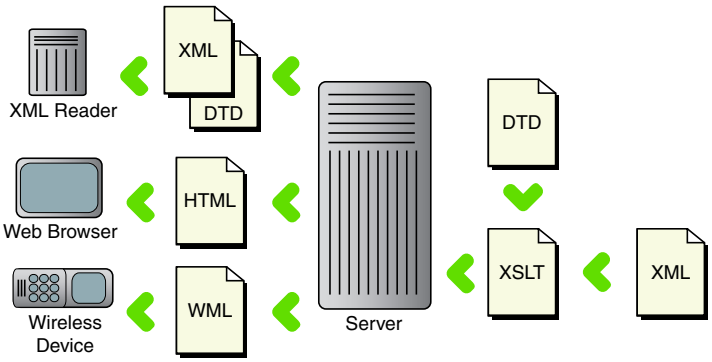
Stephen Turbek (stephen@razorfish.com) is in the razorfish science department.

> XML organizes data for sharing in a simple, open format.

XML Organizes Data

XML is not a visual mark up language, but a data mark up system. One uses it to identify and logically group the information, for example:

```
<full-name>
<title>Mr</title>
<first>Stephen</first>
<last>Turbek</last>
</full-name>
```



XML can be sent directly, or transformed into other data types.

Another computer could read this text, identify each data element and use them. XML can be later integrated with style information to create HTML pages, but its primary use is exchanging data between software. Separating data from presentation results in greater flexibility for the new class of sophisticated sites and less work in the long run.

In the Old Days…

Exchanging information between computers, particularly between organizations, has always been a troublesome task. Applications stored data in proprietary formats, so translation software would have to be written.

Structured Query Language (SQL) based databases were an improvement, as the common data format, structure layout, and standard extraction language simplified exchange. However, it was very brittle, any change in the layout of the data and it would be inaccessible.

Exchanging data in an XML-based language allows use of a common data format, a common structure layout, and standard extraction software. In addition, it's more resilient adding fields or making minor layout changes which does not cause well-written software to break. As long as the other organization supports the agreed-upon DTD (discussed later) the source data can change without effecting the receiver.

XML is Extensible

This aspect is so important, it is in the name. XML is actually a meta-language, a language used to describe other languages. HTML tags are defined by the browsers,

other tags are ignored. With XML, the author or industry group creates their 'XML Namespace", defining the tags, the structure, and values. XML itself only defines how tags can be written and organized.

The system is so flexible it is used for thousands of specialized services, such as the speech recognition language VoiceXML, the multimedia coordinating language SMIL, and the vector graphics format SVG.

XML is an Open Standard

It is important to recognize the impact of an open standard. The XML meta-language has been used to create thousands of specialized data formats based on the individual needs of industry groups. Two companies wishing to share data could simply adopt one of these formats, without having to licence someone's proprietary format.

XML's Family

XML files store and organize data, but several other types of documents in the XML family increase its dependability and utility:

Defining Types

The Document Type Definition (DTD) is like a check-sheet to compare an XML document against. It defines the tags in the XML document, their organizational structure and sometimes what data the tags should hold. While not necessary, it is used to make sure the XML is "well formed" -organized correctly and without error. XML Schemas are another approach, defining the document structure in XML itself.

Transforming XML

Another benefit of XML's tag-based format is that it can be read by humans, something not possible with binary data formats. To present it attractively for human understanding requires a bit more work. eXtensible Stylesheet Language Transformations (XSLT) are instructions for transforming XML documents into other XML documents or into HTML files. Sophisticated manipulation can be done to XML documents including math functions, combining data elements, and renaming them on the fly.

> Thousands of applications and services are now built using XML.

XML with Style

While XML is primarily for data exchange between computers, it will often need to also be viewed by people, typically on a Web browser. Cascading Style Sheets (CSS) are documents that define how to visually format the data on the screen. Using the above example, one rule could define that the last name element should always be printed in bold. Generally one CSS document is used by a large number of XML documents for a consistent formatting. While CSS and XSLT overlap somewhat, XSLT is typically more powerful, while CSS can be simpler.

The Document Object Model

The "DOM" is a specification for programming software to create, modify and control documents. This eases integration of XML functionality into new software.

XML is changing the way we interact with information on the Internet. Separating information and presentation enables computers to exchange the information much easier, and enables the data to be formatted for multiple platforms, such as wireless devices. XML does have limitations. Like a ASCII text file using tags to identify data, it is not as efficient as direct database access and may be more difficult to make massively scalable.

New ways of accessing information, such as with wireless devices, will encourage the use of XML. Separating data from presentation also enables sites to be updated visually without recreating thousands of pages. While XML is in some ways an invisible technology, its impact is being seen in new services that are built faster and work better.

The author thanks Oz Lubling, Scott Martin, and Jeff Milton for reviewing the report.

for more information

The official XML standards
http://www.w3.org/XML

An XML portal
http://www.xml.com

A good overview
http://hotwired.lycos.com/webmonkey/authoring/xml

Razorfish Reports are published for our colleagues and the interested public.