Here come the X Files

Finally, the Web's future looks like it might be functional

David Green

The Web is a wonderful place, but one of its problems is caused by the fact that Web pages are constructed using HTML. HTML (or Hyper Text Markup Language) is a subset of a more robust publishing technology called SGML (Standard Generalised Markup Language). Unlike SGML, however, HTML addresses the design and layout aspects of the information but ignores its meaning. Given that a large percentage of Internet users go to the Web primarily for information retrieval, why aren't Web pages constructed in the semantically focused SGML, rather than the aesthetically concerned HTML? The answer is, simply, because SGML is too complicated for the Web.

Many large Web publishers are circumventing some of the limitations of HTML by integrating databases 'behind the scenes' of a publicly accessible Web page. However, such information resides in a plethora of different types of database, many with bespoke, tailored data structures and, as a result, it is not readily transferable or accessible to other applications, including search engines.

Which brings us to XML (Extensible Markup Language). This is a simplified subset of SGML specifically designed for the Web. XML is primarily concerned with the semantics of the information to which it is attached. As an information professional you're probably asking 'so what?'. After all, you're already familiar with metadata, whether it is Reuters Textline, Infosort or Predicasts index codes (to name a few), or simply the meta tags hidden in the head section of a Web page.

XML is different in that it is an open technology (ie it is platform independent) and offers far greater possibilities than straightforward metadata. Unlike simple indexing, which describes the information contained in the entire document, XML allows information to be segmented into portions, which are referred to as 'classes'. Each portion is uniquely described, thus allowing direct access to the particular segment of the information in which the user is interested.

As XML describes data either as nested hierarchies or as links, this semantic markup enables more detailed searching akin to 'traditional' online search languages. For example: publication equals *IWR*, author equals David Green, and so on. In other words, XML not only enables explicit description of Web page content, but it also describes the rules for manipulating each data set contained within the information.

Imagine – the power of 'traditional' online search systems applied to every document on the Web. XML will not only revolutionise both the 'visible' and 'invisible' Web but is also good news for publishers, and therefore for users of their products, as it allows reuse of the information for different outputs. Rather than downloading financial data in comma delimited format for import to an Excel spreadsheet, XML will enable the data to be simply slotted in. Want to import data from a Web document to a database? Simple. With XML you can just drag and drop.

XML will also make it possible for documents to be generated automatically from databases and read in a variety of programs from Web browsers to word processing packages: automated personalised editions could be created for each customer, reflecting their different needs. XML is also flexible in the way that the information is presented: the raw data which has been instantly pulled from a database can be slotted into any range of style sheets. So – downloading the table of contents of a market research report from a Web-based server? With HTML it is necessary to request a new page from the Web server each time you want to click into a particular section of the table of contents. By shifting the processing power away from the server and onto the user's computer, XML frees the user to zip in and out of a dynamic table of contents without requesting a new Web page each time. Of course, this kind of functionality greatly reduces the demands placed on bandwidth which is of obvious benefit to everyone.

Some online users are sceptical about the ability, or otherwise, of XML to deliver the benefits it promises, citing the example of the much-hyped Java cross-platform programming language. But Java's uptake was limited by the existence of two non-compatible versions, one distributed by Microsoft, the other by Sun and almost everyone else. Fortunately, XML has avoided this fate and a single standardised version has been agreed by the W3 Consortium.

So where do we go from here? In order to be able to read data that has been generated using XML, applications must be 'XML enabled'. Until this happens XML will not have the effect it might have on search engines, electronic commerce, EDI (electronic data interchange), interactive electronic publishing and so on.

But the time when all this will be possible is not far off. Internet Explorer 5 and the next version of Netscape Communicator will support XML, as will Microsoft Office and products from database vendors such as Oracle and IBM and many other software vendors. XML is not backwards-compatible with HTML, but any Web page created in HTML 3.2 or higher can easily be converted.

In conclusion, I believe that XML will take a little time to become widespread – but not that long! In addition to all the benefits I've mentioned, XML complements developments in other areas of the Web: search engines that analyse the link relationships on the Web rather than the content on Web pages; intelligent agents that are making more and more requests for structured data from databases over the Web ... XML won't change the way the Web looks, but it will certainly radically change the way it is produced and used.

David Green is the former Head of Publishing at Informed Business Services and can be contacted via his Web site at www.clickmedia.freeserve.co.uk.