Infinitely extensible markup language?

XML looks set to usurp HTML as the standard for web publishing. Its in-built addressing of semantics means that it may become a knowledge management standard too. David Green looks forward to its widespread implementation throughout the business information industry.

HTML is dead. There have been no new standards developments to HTML for some time now. It is being replaced by XML. While easy to learn, HTML was limiting in that it only addressed the design and layout of information, and not its meaning. Given that most internet users and systems are primarily concerned with information retrieval and exchange, this limitation was quite a handicap.

The migration from HTML to XML as the *de facto* web publishing mechanism will have far reaching implications for information professionals and publishers alike.

Much work to date has been on the development and agreement of open standards. In recent months the World Wide Web Consortium (W3C), an international industry consortium that sets open standards for the web, has finally rubber-stamped the remaining related publishing and linking standards that complement XML.

Now XML will move into a new phase – wide-scale implementation. It is here that information professionals' skills in information management, classification schema and indexing, search skills and records management will be called upon.

For the business information industry, which has witnessed consolidation into three mega-players – the Thomson Group, Reed Elsevier (owner of Lexis-Nexis) and Factiva – distribution has become a key factor in gaining competitive edge. We no longer think about internal and external sources. Instead, the goal is to seamlessly aggregate both into unified information. The enterprise information portal – such as Factiva Select – is an XML content feed that allows corporate customers to host and integrate news into their intranet environment.

In many ways MAID's LiveIntranet product pre-dated this – however that was fundamentally flawed in that it was based on their proprietary InfoSort indexing technology and not on open standard technology such as XML. Nobody wants to be locked into a single supplier if it can be helped.

Newscast

Information management can be said to follow a cycle of discovery, acquisition, cataloguing and dissemination. XML content management systems will allow information managers to centrally manage independent content stores

Information management can be said to follow a cycle of discovery, acquisition, cataloguing and dissemination. XML content management systems (such as Interwoven) will allow information managers to centrally manage independent content stores. Data can be pulled from several sources, aggregated, and documents (web page or other format) generated 'on the fly'. Agent-based indexing and retrieval tools such as Autonomy can also add value by identifying related terms within and between documents and data sets, and can automatically generate XML-based hyperlinks.

Just as XML is a technology standard, there is much scope for it also to become a knowledge management standard. For example, a taxonomy would be integral to supplying the rules for automatically XML-tagging internal data.

Although XML tags content that scripts can then manipulate in complex ways, until recently the system interrogating the data needed to know what each tag was used for. In other words XML allowed users to add arbitrary structure to documents without saying what that structure meant. This has been resolved with the W3C issue of XML schema. These will define shared mark-up vocabularies and provide hooks to associate semantics with them.

To reiterate, the central tenet of XML is that it addresses semantics. Tim Berners-Lee, a director of W3C and often referred to as 'the godfather' of the internet, has been working on 'the semantic web', which he describes as an extension of the internet as it is today. The semantic web will allow programs to browse around and exchange data without human intervention, in effect turning the internet into a single giant computer.

What is eXtensible Markup Language (XML)?

XML is a semantically focused open technology that allows far greater possibilities than mere metadata. Not only does it enable explicit description of the content, but through related technology standards it allows manipulation of how data should be formatted and output. This takes the web page beyond a flat display of data and allows the user to manipulate the data. Every major player in the technology industry is touting this XML-driven interoperable future. Indeed, in the database arena, XML has already become the standard approach for distributing data from one application to another.

Microsoft is also placing a multi-million dollar bet on this vision of the near-future interoperable internet with its .NET project. This will allow for the automatic exchange of content and messages between software programs, applications and databases and, where appropriate, towards people.

Clearly this raises the requirement for verification and authentication of information sources in order to address data security and personal privacy concerns. XML schema will allow for better validation and assurance in information exchange (for example, ecommerce transactions) through digital signatures and other verification tools.

Again, another recently W3C-issued standard has resolved the other outstanding impediment to XML's generic adoption. Extensible Stylesheet Language (XSL) makes complex formatting of documents possible. This allows authors to write once and publish many times and to many platforms, for example different content formatted for print, web and mobile channels. In the future documents will be nebulous entities generated 'on the fly'. Automated personalised editions could be created for each customer.

While this may allow for the optimal storage of data, virtual data repositories used to generate multiple documents raises a records management issue. Like any other electronic document management system, there will be a need to save transactional documents for legal, regulatory or business purposes, as opposed to saving the base data elements. These documents must be as accessible as today's hardcopy documents. Again this is another area of XML implementation that information professionals are best placed to address for their organisations.

David Green
is a member of the Institute of Information Scientists.