

Dissecting our information society

David Penfold, Chairman of the Electronic Publishing Specialist Group, discusses the issues surrounding the organisation and access of information and content.

We are constantly told that we are living in an 'information society'. While it is true that access to information, mainly via the World Wide Web and the Internet, is easier than ever before, how information should be organised and how it can be accessed effectively and efficiently is, even now, not as widely understood as it should be. There is still a tendency for websites to be put together without any real understanding of the information architecture that should underlie them and with more emphasis on visual effect than on the underlying structure.

Fortunately, this situation is changing. The UK government, for example, has realised that if people are to access public information this must be structured and encoded in such a way that access is straightforward and easily understandable. And it is perhaps a truism to say that the ease of access is inversely proportional to the amount of thought (and work) necessary to achieve this.

Not only has there been an increasing awareness of the importance of organising and coding content for maximum accessibility, there has also probably been an increase in awareness of the importance (and indeed, the commercial value) of content. This has been driven, on the one hand, by the fact that more and more people are able to access the Web. On the other hand, it has been made easier to implement as a result of the development of both de facto standards such as XML (Extensible Markup Language) and RDF (Resource Description Framework) and software, based both on these standards and on more traditional tools, such as databases.

XML is now almost endemic in a very wide range of applications. This is mainly

because its coding structure is so well formalised that it can be used as a method of encoding data as it is passed between applications. In addition, XML is the basis for a large number of higher-level encoding systems. These include such diverse areas as the INDECS system (see <http://www.indecs.org> or, for a basic introduction <http://www.bisg.org/wp02.pdf>),

Monsieur Jourdain, in Moliere's *Le Bourgeois Gentilhomme*, suddenly realises that "I've been talking prose for the last forty years and have never known it."

which is at the heart of several of the recently developed digital rights managements systems, and SMIL (Synchronized Multimedia Integration Language – pronounced *smile*), which provides a method of authoring interactive audiovisual presentations, integrating video, audio, text etc, (see <http://www.w3.org/AudioVideo/>), providing an open-source product like Macromedia's Flash.

In addition, there is SVG (Scalar Vector Graphics – see <http://www.adobe.com/svg/> for a good introduction), a very flexible method of handling vector graphics on the Web and many markup languages, the abbreviations for which end in ML, eg ebXML, SPML and LDML. Details of many of these, particularly the business-oriented applications can be accessed from <http://www.xml.org> or <http://xml.coverpages.org>, while the World Wide Web Consortium (W3C) site (<http://www.w3.org>) provides the basic 'standards' information.

So XML and its 'children' are being used to encode data in many application areas. However, before one can encode

information, one needs to determine what should be encoded and how. The 'what' is a huge topic and should be organised on the basis of the discipline that is known as 'information architecture'. This is a concept that is less well-known (and even less well-understood) than it should be and concerns carrying out an analysis of what information is available within an

organisation and how that information is organised (and probably could be re-organised) so that it is most easily accessible. For more explanation of the concepts involved, see, for example, <http://argus-acia.com>, <http://www.infoarch.ai.mit.edu>, or <http://hotwired.lycos.com/webmonkey/design/tutorials/tutorial1.htm>

A related area that has grown over the last year has been content management. Rather in the same way as for XML, this is not limited to conventional areas, such as document management, but is now being extended to cover whole organisations, under the term Enterprise Content Management, which includes areas that have not conventionally been considered to fall in this area (financial information, for example). The Butler Report *Enterprise Content Management* (February 2003; Butler Direct, Hull) describes this well. However, here we are more concerned with what might be called 'access to conventional information'.

An important concept in access to information is metadata. However, metadata

● Multimedia & Electronic Publishing

Not only has there been an increasing awareness of the importance of organising and coding content for maximum accessibility, there has also probably been an increase in awareness of the importance (and indeed, the commercial value) of content.

is really not new. In a parallel situation, Monsieur Jourdain in Molière's *Le Bourgeois Gentilhomme* suddenly realises that, "I've been talking prose for the last forty years and have never known it." In just the same way, people have been talking metadata for at least the last forty years and probably never known it!

So what is metadata, at least within the context of publishing and access to information? We are all familiar with Yellow Pages, library catalogues, entertainment guides, indexes, menus, parts lists and Web portals, to name but a few. These are examples of applications of metadata – data/information about things or about other data.

Metadata is often data that is not otherwise part of the intrinsic data. If we take a photograph as an example, the associated metadata could include the photographer's name, the technical data (exposure, aperture etc), the date when the photograph was taken and the subject (which itself could be divided into many other metadata fields). Of course, in some situations, for example a book, the author's name will be both part of the data forming the book and part of the metadata describing it.

If we assume that we have designed our information architecture, ie what information we have and how it is organised, there are a number of important questions that have to be asked when preparing a metadata application:

- How should metadata be classified? Many classifications are based on the Dublin Core (15 elements, including title, creator, subject etc – see <http://dublincore.org>), but different applications require different extensions to this, ie different metadata fields.
- What structural approach should be taken – a thesaurus, a taxonomy or an ontology? A well-known taxonomy is, for example, the ACM Computing Classification Scheme (www.acm.org/class/1998), while thesauri include information about the relationships between terms, and an ontology goes into even more details about the relationships (see *Tomatoes are not the only Fruit – A Guide to Controlled Vocabularies*, by Maewyn Cumming; <http://www.govtalk.gov.uk/>

schemasstandards/gcl_document.asp?docnum=681). The approach taken usually depends on how the data is stored and structured, how the software used accesses the information and what skills and resources are available for the encoding.

- What software/data structure approaches should be used – a database, a content management system, or tagged files? The Resource Description Framework (RDF – written in XML) provides a possible solution (see <http://xml.com/pub/a/2001/01/24/rdf.html>, <http://www.dlib.org/dlib/may98/miller/05miller.html>, or <http://www.ibm.com/developerworks/library/wrdf/#resources>, or <http://www.w3.org/RDF> for detailed explanations).
- How should metadata be created? Approaches include: manual coding, through on-screen data entry, although this is time-consuming and leaves unsolved problems with legacy data; conversion from existing data, although this requires appropriate software that can cope with the incompatibility of schemes and ways of handling incomplete data sets; automatic coding software. In practice, the last option, the use of automatic coding software, is a useful way of providing the initial coding. If time and personnel are available, checking this is a good idea, if only because different programmes use different approaches, for example frequency of occurrence of terms, position of terms in a document etc, so two programmes will not necessarily give the same metadata.

These are not simple problems and they all have to be solved for every application. In spite of the problems, there are many example applications available today. The UK government has established the e-Government Metadata Standard (e-GMS – see www.govtalk.gov.uk/interoperability/metadata_document.asp?docnum=524). In the educational world, the government and related bodies are developing standards (see <http://metadata.ngfl.gov.uk/>; www.curriculumonline.gov.uk/). Other organisations working on developing metadata include, the BBC for its many websites (including many local BBC sites),

museums and galleries for making their collections available in digital form, and publishers of online publications.

While metadata provides the most powerful approach to finding information on the Web, there are various other approaches that can be taken. These include:

- Free text searching (see, for example, <http://www.proc.britac.ac.uk>).
- Automatic indexing and linking (see, for example, <http://www.active-navigation.co.uk>).
- Document Object Identifiers (DOIs) – www.doi.org. This is an extension of the ideas in the ISBN (International Standard Book Number) to parts of documents. The granularity can be chosen to match the application. DOIs are based on XML coding and make it possible on the Web, via a database that is similar to a Domain Name Server, to find information even if the URL has changed. DOIs are the basis of the CrossRef system (<http://www.crossref.org>), which makes it possible to link directly from the reference list in a journal paper to the URL of the paper referenced.
- Automatic classification (data mining), as discussed briefly above.

While many of these alternatives may help to find information, what they will not provide is the next step on the Internet – the Semantic Web – put forward by Tim Berners-Lee see <http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC5-88EF21>). In the Semantic Web software uses many kinds of metadata to help us find and filter information that today requires many hours of effort.

Thus, metadata, and accurate, well structured metadata can not only help us find information more effectively today, but will be essential if the Semantic Web is ever to become a reality. ■

David Penfold is a publishing consultant, writer and editor. He is a visiting teacher at the London College of Printing, chairman of the Electronic Publishing Specialist Group and a member of the Knowledge Services Board.