

METADATA OVERLOAD

Ian Beeson

University of the West of England, Bristol
Faculty of Computing, Engineering & Mathematical Sciences
Coldharbour Lane, Bristol BS16 1QY
ian.beeson@uwe.ac.uk

ABSTRACT

The spread and proliferation of metadata threatens to overwhelm the informational content that metadata was originally invented to supplement. The emergence of new kinds of metadata and markup are traced, and reasons for their spread are explored. The concept of metadata needs to be disaggregated and refined so that we can understand more precisely how documents are produced and managed now that information increasingly takes digital and hypertextual forms. In this reworking of the idea and purpose of metadata, more attention needs to be paid to document creation and composition, and to where authority lies over these processes.

Keywords

Metadata; markup; information overload; hypertext; Semantic Web; information architecture; controlled vocabulary.

1. INTRODUCTION: KINDS OF METADATA

Until recently, the concept of *metadata* has been quite clear and quite circumscribed. In Library and Information Science (LIS), it has referred in particular to two kinds of 'data about information' or information about documents:

- (1) the description of the physical aspects of a document, as included in a catalogue – what the information object *is*;
- (2) the description of the subject content of a document, as included in indexes or classification schemes – what the information object *is about*.

These two forms of metadata have been at the core

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

© 2006 Higher Education Academy

Subject Centre for Information and Computer Sciences

of library work for many years. Gilchrist and Mahon [1] observe, in connection with the expansion of digitized information, the emergence of a third class of metadata:

- (3) the description of various aspects of the management or administration of a document, including ownership, version, status, security, and rights.

In the newer field of Information Architecture, which has developed out of the spread of web technologies in organizational and inter-organizational contexts, and which, at least for some of its authors, has explicit roots in LIS [2], metadata originally had a fairly limited connotation, referring to information about a document rather than to document content.

In HTML, (the HyperText Markup Language, used for writing web pages), metadata is specified in the *head* section of the document (which is not displayed by browsers), and is restricted to a small number of types, including document title, base URL (for resolving links), links to external documents, styles to be applied, scripts to be run, and a catch-all *meta* type. These types are specified as HTML elements, using *tags*. In the HTML specification, the meta element is quite generally defined, consisting of a declaration of a *property* and of a *value* for that property. Properties come in name/content pairs and are frequently used to identify authors, general document content, or keywords, for search engines to find.

The meaning of a property and its set of legal values can be defined in a *profile* defined for the head element. For example, Dublin Core (DC), a set of recommended properties for electronic bibliographic descriptions, could be attached to a document as a profile, to ensure conformance to the DC recommendations.

In her discussion of metadata, in the context of information architecture, Wodtke [3] lists a set of types similar to those presented earlier from the LIS perspective:

- (1) *intrinsic* metadata – what kind of thing the information object is (what size, whether compressed, what software produced it);
- (2) *descriptive* metadata – what kind or genre of information it is, what it's about;
- (3) *administrative* metadata – how the object is to be handled, its status, who owns it.

The American National Information Standards Organization (NISO) [4] also lists three main types of metadata:

- (1) *structural* metadata: specifies how compound objects are put together (eg, how pages are ordered into chapters);
- (2) *descriptive* metadata: used for document discovery and identification (elements such as title, author, abstracts, keywords);
- (3) *administrative* metadata: information to help manage a resource, including creation and access information, file type and other technical information, and, as major subtypes, rights management and preservation metadata.

Comparing these three categorizations of metadata, we may note that:

- *descriptive* metadata, as identified by Wodtke and NISO, is usefully divided in the general LIS categorization into catalogue-type and classification-type information;
- all include an *administrative* category, though Wodtke's *intrinsic* metadata would be included under NISO's administrative metadata;
- NISO's *structural* category is missing in the other two.

Putting the three categorizations together, four types of metadata can be distinguished:

- (1) metadata for resource description or identification (what the information object is)
- (2) metadata for resource discovery (what the object is about)
- (3) structural metadata (how the object is composed)
- (4) administrative metadata (how the object is to be handled).

The first two kinds of metadata are the basic and original forms, which have to do with keeping an inventory of documents and then with making the documents accessible to people who might be interested in their content. These two forms, in the shape of catalogue or index, exist outside the base documents and point to them, while at the same time some of the metadata would typically also be included inside the documents (eg title, author, abstract, subject class). From the perspective of the document as a published item, structural metadata relates to its pre-publication, and

administrative metadata to its post-publication, existence. These two become important general forms of metadata only after the large-scale migration of information to the web.

2. THE GROWTH OF METADATA

If you look at the HTML source of a web page, you see a mixture of original information, intended to be rendered visible to viewers by a browser, and HTML tags which surround and thread through it. A web page, as an HTML document, is a text file made of HTML elements, where these elements are defined by their tags, which usually come in opening and closing pairs, and are written in angle brackets. Some of these tags, especially in the head of the document, specify metadata, as described above, but there are many others which can be used in the body of the document. These serve a number of purposes:

- to show the structure of the document: its headings, paragraphs, line breaks, and the inclusion of lists and tables;
- to show how the content is to be presented: use of fonts, colours, and different forms of emphasis;
- to link to other documents: to permit navigation among documents, to incorporate images or other media into the document, or to put the document under the control of a template or profile of some kind.

Do all these tags constitute metadata? None of them are *data*, in the sense of the original content of the document, but rather show how the content is to be displayed, expanded (by inclusion of non-textual data), or navigated. Increasing sophistication in the design of web pages means that the content is hard to pick out among the welter of tags. If one learns HTML, it is quite easy to interpret the first two kinds of tags, but the third create links to other documents which are not immediately visible.

(One might counter at this point that HTML is not primarily intended for human viewing, but for browser interpretation. That may be true; but the HTML code, not the rendered page, is the information object, and information scientists at least should have an interest in it.)

The usual general term for these tags in the body has been *markup*. The first kind may be called structural markup, and is close to the NISO idea of structural metadata. The second kind, presentational markup, is deprecated in modern versions of HTML, the preference being to put this kind of markup in an external document called a *style sheet*, and link to it. That reduces the clutter on the screen, but puts the markup out of sight. The third kind may be called associational markup, and

situates the document currently being viewed as a node in a wider web of documents.

All these kinds of markup are increasingly seen as kinds of metadata. Boiko, for example, from the perspective of content management, asserts: 'Markup languages (such as XML) are the major way that you apply metadata to content.' [5].

In the original design of HTML, a primary purpose, alongside support for *links*, was to establish a simple and consistent syntax for document *structure* which would be sufficiently straightforward for browsers to handle. This early push to interoperability lies at the heart of the Web's success, but we should note it involves some surrendering of authorial control over how a document looks.

The evolution of HTML in the direction of the more powerful metalanguage XML (eXtensible Markup Language) is bringing with it a further kind of markup, called *semantic* markup. Semantic tags indicate what a portion of the document *means*. XML provides the general syntax for semantic markup, but the names and scopes of the tags have to be supplied separately, in the form of a *controlled vocabulary*, agreed for some domain of interest.

Here is a fragment of this latter kind of markup, borrowed from Glushko and McGrath [6]:

```
<OrderLine>
  <LineItem>
    <BookItem>
      <ISBN>0262072610</ISBN>
      <BasePrice>99.95</BasePrice>
    </BookItem>
    <Quantity>300</Quantity>
  </LineItem>
</OrderLine>
```

This encodes an order line for 300 copies of the book with the given ISBN at a unit cost of 99.95 (presumed US dollars). The content here is just the three values 0262072610, 99.95, and 300. The rest is markup. Imagine semantic markup combined with structural and associational markup, and you can see that the content, from the perspective of the human reader at any rate, becomes completely overwhelmed by the metadata.

To the recognised problem of *information overload*, suffered by humans, we can now identify a further challenge for the Information Age, *metadata overload*, suffered by information.

3. CAUSES OF THE GROWTH OF METADATA

Three major causes may be suggested for the explosive spread of metadata:

- (1) ease of publishing;
- (2) dissolution of the document;
- (3) the drive to machine processing of documents.

3.1 Ease of publishing

The availability of powerful word processing and other content creation software, coupled with the advent and rapid expansion of the World Wide Web, has made it unprecedentedly easy for people to create and disseminate documents. Belew comments:

'The number of content-producers (writers) is rapidly approaching the number of content-consumers (readers)! Never before has the machinery of producing and distributing media been as widely available as it is today. Our collective expectations as to just what documents are "out there," not to mention the care and authority with which they have been authored, are in terrific flux.' [7]

Leaving aside for the moment the uncertainties over the volume and provenance of documents, another consequence of the ease of publishing on to the Web is the erosion of previous professional and temporal separations between writing, editing, publishing, and printing. In earlier times, libraries received copies of finished works - the writing, editing, publishing and printing had all been done. All libraries had to do was catalogue, classify and index the works, to keep track of what they had, and to make the material accessible. The important metadata therefore, related to the keeping and finding of finished works.

When works can be written, edited, and published on the web, *markup*, which is pre-publication activity, comes more strongly into the picture. Structural and presentational markup, in particular, become critical metadata for content creators, editors, and publishers. Although publishing houses and content management regimes may try to keep these activities separate, this looks like a hangover from previous ways of working, which may not be sustained in the longer term.

3.2 Dissolution of the document

In earlier times, the information objects stored and made available by libraries were books, journals and other physical materials. Their physical presence and unity as objects were undeniable, they could only be stored in one place, and the metadata relating to them was, in comparison to their content, minuscule and clearly supplemental.

As Landow has persuasively pointed out [8], the transition from text to hypertext, particularly on the Web, spells the end of the formerly taken for granted unity and fixity of texts. He identifies three aspects of what might be called the dissolution of the document:

- (1) fragmentation or atomization of documents into component parts, or *lexias* (reading units), or chunks of (reusable) content;
- (2) blurred beginnings and endings: as text becomes non-linear, and furthermore as textual elements in the electronic medium become endlessly revisable and recombinable, it becomes hard to say where a document begins or ends, or when it is started or finished;
- (3) loss of borders or boundaries: the ease and ubiquity of *linking* on the Web precipitate a transformation of originally solid and bounded documents into navigable associative networks of pieces of text, within which readers move freely between one piece to another, and in and out of original texts, and annotations, commentaries and interpretations.

Thus the document, on the Web at least, ceases to be a coherent finished work which offers a clear locus for the attachment of metadata, and dissolves instead into a shifting and interpenetrating network of pieces or *lexias*. Where is the classifier or cataloguer to pin the metadata on such slippery concoctions? To preserve what used to be possible, and to retain authority and control over the corpus of material, it is necessary to keep track of versioning, variation and ownership of an increasing mass of increasingly rich and complex, but also increasingly fluid, documents. As text turns into hypertext, the document dissolves into its subatomic particles, only for these particles to return, again and again, in new constellations. When documents are never finished, there is no transition from a publication point to a post-publication phase. Administration or management of documents now becomes much more problematic, and the explosion of administrative metadata is triggered.

3.3 The drive to machine processing of documents

'Metadata is machine understandable information for the web'. So states the World Wide Web Consortium (W3C) [9]. Even more than the other two developments, which relate to the changing boundaries of publishing activity and the delinearization of text, this is the principal engine behind the explosion of metadata – the drive towards interoperability, and to the Semantic Web. The reasoning is that, in order to solve the problem of information overload, some of the interpretation of data must be moved away from humans into software. Documents must therefore become

meaningful to machines. Since machines lack consciousness and intentionality, they cannot interpret the data directly, but must have meanings spelled out explicitly, using markup and rules. And for machines in different organizational contexts to extract meaning in the same way, the markup and rules must be established globally (for some domain of interest). This is why documents will be increasingly freighted with semantic markup, and why *controlled vocabularies* are being developed apace, to fix the markup and rules for different domains interested in information exchange – for instance in e-commerce, or medicine, or public administration.

Take for example the Integrated Public Sector Vocabulary (IPSV) being developed in the UK, described as 'an encoding scheme for populating the e-GMS Subject element' [10]. e-GMS is the e-Government Data Standard, which is based on the Dublin Core (DC) metadata standard, a standard for cross-domain information resource description. Dublin Core specifies 15 elements which can be included in the metadata for a resource: *title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights*. Among these, Description is an account of the content of a resource, and may take the form of an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content; while Subject is described as 'a topic of the content of the resource' [11], with the following comment: 'Typically, Subject will be expressed as keywords, key phrases or classification codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme'.

Dublin Core thus recommends that the Subject(s) of a resource be specified, but does not specify them itself, leaving that task to some body constituted as an authority in the domain of interest.

IPSV is then an example of a controlled vocabulary for the DC Subject element. The authority behind it is the Office of the Deputy Prime Minister and the Cabinet Office e-Government Unit.

IPSV is a controlled vocabulary with 16 top-level terms:

- business and industry
- economics and finance
- education and skills
- employment, jobs and resources
- environment
- government, politics and public administration
- health, well-being and care
- housing
- information and communication

- international affairs and defence
- leisure and culture
- life in the community
- people and organizations
- public order, justice and rights
- science, technology and innovation
- transport and infrastructure

Multiple hierarchies of preferred terms hang below these top-level terms – 2700 in all for the full IPSV, and about 500 in an abridged version. One can see an entire theory of government and public administration encapsulated in the vocabulary.

A guide to the use of IPSV [12], which assumes the user has access to metatagging software linked to a content management system, states the overriding principle of use: ‘...by some means or another, one or more relevant IPSV preferred terms must get added to the Subject metadata’.

The syntax in the HTML/XML head section would look like this (for example):

```
<meta name="dc.subject" scheme="IPSV"
content="Housing benefit" />
```

Here, the preferred term ‘Housing benefit’, from the IPSV, has been identified as a subject for the document; and the term is declared to be part of a controlled vocabulary (IPSV), itself designed to supply terms for the DC subject element. The document thus opens on to, through one tag in its header, a whole world of government, described in conformity with the IPSV, which in itself is an artefact of the Dublin Core standard. The individual document sits under a huge carapace of meaning.

And this is only the metadata in the heading. It will help retrieval systems find the document when appropriate terms are supplied. For the document to be interpreted in its detail by software, semantic markup must be interlaced into the body of the document. The obvious source for the XML tags will be the controlled vocabulary which it has taken such strenuous efforts to invent. Eventually, one can assume, it will be appealing to content creators (or their managers) to use words from this lexicon in the creation of the content itself.

4. IS YOUR METADATA REALLY NECESSARY?

Haynes [13], building on earlier discussions by Day and Gilliland-Sweetland, proposes a ‘five-point model’ of metadata according to its different purposes:

- (1) resource description
- (2) information retrieval
- (3) resource management
- (4) ownership and authenticity

(5) interoperability

In terms of previous categorizations and discussion in this paper, we can see assembled here the two ‘original’ forms of metadata (for description and discovery; or description and subject in DC terms), a division of administrative metadata into two kinds, and a final category, interoperability, which sweeps together structural and semantic markup under a single purpose of information exchange between software systems.

The extent of this spectrum of purposes is so broad as to suggest a concept that has become too encompassing to be very useful. Metadata, which once took the form of supplementary lists enabling access to physical documents in a library, has invaded and consumed documents at the very time that the documents themselves have begun to dissolve into virtuality. Metadata is no longer supplementary, but dominant. When a secondary concept becomes so extended that it overwhelms its primary referent, the time has come to re-examine the processes and interests which lie behind it, to see whether a more precise conceptual framework might replace it. The separation of types of metadata by purpose in Haynes’s model is useful, and points a way forward towards replacing the single concept by a set of concepts geared to different purposes or interests.

We can understand the rapid recent growth of metadata as a consequence of the historic shift from text to hypertext, of the dematerialization of documents, and of the erosion of the boundaries between authoring, editing, publishing and distribution.

4.1 Document and metadata in focus

As we try to develop a clearer understanding of the nature and role of documents in the midst of these developments, the following lines of approach could be helpful:-

- *Restore the primacy of content.* Document content has been quartered and colonized by markup and templates. It is time to pay more attention to the quality and coherence of content (even while acknowledging the loss of unity in the move to hypertext), and to consider where authority for content creation lies in different circumstances.
- *Restore the conceptual division between metadata and markup.* It might be worthwhile, in order to preserve metadata as a useful term, to restrict it to refer to information about a document in a header or an external profile, while keeping markup to refer to structural, presentational or semantic tagging in the body of a document.

- *Pay more attention to the composition of documents.* Under the twin assaults of hypertext and markup, the document has been pulverized almost to vanishing point. Yet, in however new a guise it may appear, the document, or some such body of written content, must remain as the fundamental focus for information or content management. We need to understand more thoroughly the processes by which a document is created, perhaps not until the point of reading, at once as an assembly of lexias or components and as a node in a network of documents
- *Question the expansion of semantic encoding.* We should consider the feasibility, desirability, and consequences of semantic encoding of documents in order to facilitate interoperability and machine processing of documents. The invention and imposition of controlled vocabularies involves a continual heavy workload (as meanings shift). There are not only deeper questions here about the extent to which machine understanding is achievable, or the social and political consequences of attempting to control vocabulary, but more immediate pragmatic questions about the costs of encoding and about how control of what can be said is in fact exerted. The benefits of discovery and extraction of meaning by machine, which may be considerable in particular applications, need to be argued case by case.

4.2 Levels of metadata and markup

The areas of interest we have covered in this paper might be separated finally into the following layers, following a general semiotic model of communication:

- (1) Document *content*. Where does content creation end and markup begin? Is document layout a legitimate concern for authors as well as browsers? Punctuation, now firmly embedded in content, would once have been considered 'markup'. Included at this level might be what is now considered to be presentational markup (or presentational metadata, in the form of style sheets), as well as the 'resource description' aspect of metadata, since it comprises a summary of the content.
- (2) Document *composition*. How is content structured and arranged in a document, how are components organized into a document, and how are documents organized into a network (a 'hypertext')? This area would cover structural and associational markup, as well as the whole connecting apparatus of links, templates, and profiles.
- (3) Document *semantics*. What do documents mean, and how are they to be understood, by

humans and by machines? The focus here is what documents are *about*. This level of analysis brings together the original 'resource discovery' (information retrieval, indexing and classification) aspect of metadata, as well as the new burgeoning area of semantic markup and controlled vocabulary.

- (4) Document *pragmatics*. What are documents to be used for, and how are they to be managed? This is the (also burgeoning) area of administrative metadata.

Instead of viewing metadata as a single continuum, defined in contrast to data or information, we can in this way disaggregate the concept into four levels of analysis which focus on the content, composition, meaning or use of documents, and couple these four with metadata and markup, separately at each level, as two different ways of organizing information (metadata by external rules or templates and markup by embedded indications).

5. CONCLUSIONS

As originally conceived, metadata fulfilled the useful role of enabling a library to catalogue and index its materials, so that users could access them. On the Web, metadata in the header was also conceived to help people find documents, and markup in the body was introduced to simplify and standardize the rendering of documents to screen by browser software. Both these developments can be appreciated as ways of making information more available to people.

Later developments in metadata and markup, and the convergence of the two, cannot be seen unequivocally in the same light. It is at least arguable that the recent expansion of administrative metadata reflects institutions' desire to retain control of published materials to a degree which may be unsustainable, given the increasing complexity and fluidity of documents, and a fundamental shift in the nature of publishing. It is equally plausible that the race towards semantic encoding of documents springs from a desire to support, not freer or more extensive exchange of information among people, but increasing levels of machine processing of information, and increasing control over what people say and do.

The present analysis is intended to help focus a necessary debate over the role of metadata and markup in the production, use and control of information in our institutions.

6. REFERENCES

- [1] Gilchrist, A. and Mahon, B. (eds.), *Information Architecture: Designing Information Environments for Purpose*. London: Facet Publishing (2004). Preface to Part 3.

- [2] Rosenfeld, L. and Morville, P., *Information Architecture for the World Wide Web*. Sebastopol, CA: O'Reilly (2nd ed., 2002).
- [3] Wodtke, C., *Information Architecture: Blueprints for the Web*. Indianapolis: New Riders (2002).
- [4] National Information Standards Organization, *Understanding Metadata*. Bethesda, MD: NISO Press (2004). <http://www.niso.org/standards/resources/UnderstandingMetadata.pdf> (accessed 13 April 2006)
- [5] Boiko, B., *Content Management Bible*. Indianapolis, IN : Wiley (2nd ed., 2005). p. 498.
- [6] Glushko, R.J. and McGrath, T., *Document Engineering*. Cambridge, MA: MIT Press (2005). p. 194.
- [7] Belew, R.K., *Finding Out About: a Cognitive Perspective on Search Engine Technology and the WWW*. Cambridge (UK): Cambridge University Press (2000). p. 297, emphasis in original.
- [8] Landow, G.P., *Hypertext 2.0*. Baltimore, MD: John Hopkins University Press (1997). Ch. 3.
- [9] World Wide Web Consortium (W3C), *Metadata and Resource Description*. <http://www.w3.org/Metadata/> (accessed 13 April 2006).
- [10] Cabinet Office e-Government Unit (UK), *Abridged IPSV (Integrated Public Sector Vocabulary): Version 1:0 – Hierarchical presentation* (April, 2005). http://www.esd.org.uk/standards/ipsv%5Fabridged/IPSV_AbridgedV1.0Hierarchies.doc (accessed 13 April 2006).
- [11] Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set, Version 1.1: Reference Description* (July, 1999). <http://dublincore.org/documents/1999/07/02/dces> (accessed 13 April 2006).
- [12] Cabinet Office e-Government Unit (UK), *Guide to Meta-tagging with the IPSV*. (Version 1.2, March 2006). <http://www.esd.org.uk/documents/IPSVHowToMetatag.pdf> (accessed 13 April 2006).
- [13] Haynes, D., *Metadata for Information Management and Retrieval*. London: Facet Publishing (2004).