Metadata's many meanings and uses

Conrad Taylor is Chair of the Electronic Publishing Specialist Group of the British Computer Society, and is working for the BCS 'KIDMM' study project (Knowledge, Information, Data and Metadata Management.) He teaches electronic publishing courses to industry clients.

'Metadata' is a term used in connection with the management of data and information, especially in digital form. Librarians talk about it; also publishers, especially in scientific and scholarly fields. Metadata has become important for government, the health service and education; and it has long been an important subject for data managers.

However, when you get representatives of these different communities together, you often find they cannot agree on what metadata means.

Here, I attempt a framework of understanding which may unite these disparate understandings – also casting light on a range of practices not typically labelled as 'metadata', but which do the same sort of job.

Etymology of meta+data

The term 'data' comes from the Latin *datum*, meaning '[that which is] given'. In science and in computing, data are the 'givens', the simple facts or records. Examples of data could be 'George' as someone's given name, '35° C' as the temperature of a liquid, or '5,323 metres' as the height of a mountain.

The prefix 'meta' is Greek in origin – *μετά* – a preposition that, according to context of use, can mean 'with/beside' or 'after' (compare German *mit* and Swedish *med*), and is a root of such words as 'metaphor' and 'metaphysics'.

In more recent uses, 'meta-' often signifies a concept that is (a) abstracted from another concept, and (b) providing a viewpoint from which to analyze the latter. For example the American philosopher WV Quine in 1937 coined the term 'metatheorem' to mean 'a theorem about theorems'; and the word 'metacognition' is used to describe the ability of sapient beings to think about their own thought processes.

On this model, Metadata might be defined as meaning 'data about data' – and often is. However, because the term has been adopted in an ad-hoc way by different groups of data and information practitioners, it has come to have more than one meaning. As I shall explain, it was in the context of database management that the term was first adopted, about 35 years ago; but in the last 15 years or so, people managing document-like objects and media assets (e.g. librarians, publishers) have adopted it to mean something substantially different. It is this latter use which has become the dominant one today.

μετά +DATA

"The word is half Latin and half Greek. No good can come of it!"

— C P Scott, on 'television'

'Metadata': what it means to the Data Management community

Some credit the coining of the term 'metadata' to Jack E. Myers, who used it conversationally around 1969, founded The Metadata® Company and in 1986 registered Metadata as a trademark, yet seems not to have defined it clearly.

Professor Bo Sundgren of the Stockholm School of Economics referred to 'meta-data' and 'meta-information' in his 1973 PhD thesis, *An Infological Approach to Data Bases*, in which he distinguished (a) objects in the real world (b) information about those objects and (c) information that describes the nature of that information – the 'meta-data'. In the 1970s, 'metadata' became increasingly used in the data management community as a common generic term to indicate various kinds of formal definitions that describe and control how data is managed and used within a computer database system.

This meaning of 'metadata', and issues that lie behind it, may be illustrated by some simple examples. To start with, let us take a record of sales from a small bookshop, as they might be written by hand into a ledger or notebook:

Date	Book title	Author	Price
Oct 8th, 2005	Neuromancer	William Gibson	£4.99
8 Oct 2005	The Amber Spyglass	Philip Pullman	£6.99
9 Oct, 2005	African Eldorado	John Carmichael	£12.99
9th Oct 2005	The Book Before Printing	David Diringer	£7.45

Fig. 1
A tabular record of sales from a small bookshop. Note the inconsistent way in which the dates have been recorded: this is a barrier to computer manipulation.

As a visual record for simple stocktaking and accounts, this table is satisfactory. But let us imagine this data transferred to a computer, not only as a means of storage, but also so that operations can be performed upon it – say, adding up sales totals on a week-by-week basis. This will require data to be structured in accordance with a set of rules, and recorded in a consistent way.

Delimited files, interpretive rules

We must distinguish between (a) the visual means of presenting a database, such as the neatly-ruled table above, and (b) the formal means by which the data is stored in the machine. Our example database – after a bit of cleaning up of the 'date' field of course – may be stored as a text file, in which certain characters (which I've coloured red below) are made to play a special role:

```
08-10-2005|Neuromancer|William Gibson|04.99¬
08-10-2005|The Amber Spyglass|Philip Pullman|06.99¬
08-10-2005|African Eldorado|John Carmichael|12.99¬
08-10-2005|The Book Before Printing|David Diringer|06.99¬
```

Individual *records* of sales are separated by carriage returns, represented by the ASCII[1] codepoint 13 – here made visible as ¬. Fields within each record are separated by the vertical pipe figure | (ASCII codepoint 124). Data files organised in this way may be described as 'delimiter-separated'.

1. ASCII – the American Standard Code for Information Interchange, is a standard introduced in 1967 for the representation of English-language characters, numerals, punctuation and symbols by numerical codes, expressed as sequences of binary digits. Thirty-three of the ASCII codes are 'control characters' such as tab, carriage return, line-feed.

Various ASCII characters may act as delimiters. The comma is often used thus, and data files structured this way are known as CSV (comma-separated value) files, often used for exchanging spreadsheet data. But book titles often contain commas in them, which is why I used a different delimiter here.[2]

Any software that works with database records will have to 'know' the rules by which they are structured, and how to deal with each component. We may express the rules for our example datadase records in English, thus:

- The records in this database table are separated by carriage returns, and fields within records by pipe characters…
- The first field in each record is the **Date of Sale**; it must be recorded in the form DD-MM-YYYY to avoid ambiguity. In the case of single-digit values for the day or month, they must be padded with a leading zero…

Rules of relation

More sophisticated database systems are usually **relational**: they contain a number of tables, with look-up relationships between them.

Let us make our bookshop example more sophisticated. We shall maintain a distinct 'book-reference' table of book titles, authors, prices, ISBN codes &c. When a customer presents a book at the till, the barcode scanner reads the ISBN. The book being thus identified, its title and price are picked up from the book-reference table, and imported to the table used to print the receipt and record the sale; at the same time, the stock records tables can be updated.

Relations make our 'rules' more complex, because they have to define all the tables which together constitute the relational database, and the look-up relationships whereby some fields get their values by referring to other tables. (Indeed, some fields may derive contents through calculations performed on the values of *several* fields in *several* other tables.)

The value of explicit metadata

For the successful operation of a database, it is not necessary that these rules be expressed *explicitly* – it is merely sufficient that software which processes the data is programmed to behave consistently and accurately as it reads data from the file, performs operations on it, and writes values into the files.

However, data managers have increasingly come to recognise the value of expressing the structural and processing rules of databases in a more explicit fashion. Just consider the **Y2K problem**, or **Millenium Bug**, which resulted from a widespread practice in the 1960s and onwards of using two digits to record the year, rather than four, creating the potential for computers to confuse, say, the years 2001 with 1901. (So, people who rented videos from Blockbuster and returned them after 1 January 2001 were being charged late-return fines of $91,250, as if they were 100 years overdue!)

2. To be fair, CSV files can deal with the existence of commas within fields, and they do so by the use of double-quote characters to 'protect' commas that do not play the role of delimiters.

Part of the reason why worried businesses and governments had to spend an estimated global total of $300 billion in 'fixing' the Y2K problem was that a large number of data-processing applications and the databases they worked with had been developed in an undocumented fashion, as US Federal Reserve Chairman Alan Greenspan testified before the Senate Banking Committee:

> I'm one of the culprits who created this problem. I used to write those programs back in the 1960s and 1970s, and was proud of the fact that I was able to squeeze a few elements of space out of my program by not having to put a 19 before the year. Back then, it was very important… It never entered our minds that those programs would have lasted for more than a few years. As a consequence, they are very poorly documented. If I were to go back and look at some of the programs I wrote 30 years ago, I would have one terribly difficult time working my way through step-by-step.[3]

Had these applications been explicitly documented, locating problem fields and figuring out what depended on them would have been less painful.

Data management uses of metadata

Data managers define 'metadata' as meaning: explicitly recorded definitions of what data objects stand for, what values they are allowed to have, how they are recorded physically and what the relationships are between data objects.

It's been said that 'metadata turns data into information'. By itself '1334' is meaningless, but it becomes meaningful if you know (a) it is an elevation above sea-level; (b) it is expressed in metres; (c) is an attribute of a location 56˚47′51·49″N, 5˚0′9·98″W which (d) is called 'Ben Nevis'.

Metadata concepts may be used early in the process of devising a database system, in the process of **data modeling**, in which an organisation decides on the kinds of data it needs to store, and what it needs to be able to do with it.

In early planning stages, the result is a *logical data model* describing the entities required, their attributes, and the relationships between them, in a manner largely independent of the database management system (DBMS) in which it will be implemented – thus, 'we'll need to specify a date and time'. Later this process moves towards a *physical data model* which gets down to the nitty-gritty of implementation in the DBMS, including such details as how entries will be encoded – 'dates will be recorded as DD-MM-YYYY'.

Unfortunately, as the Millenium Bug case illustrates, it is often necessary to dig into existing data management systems and figure out the metadata in a *retrospective* manner. **Data profiling** projects are undertaken to do this, almost always for hard business reasons. The organisation may need to know if existing data can be used for other purposes; they may plan to merge new kinds of data into the DBMS and want to know the risks of doing so. Business mergers can also create the need to reconcile database metadata sets, as data assets developed initially by the separate companies are merged.

3. Alan Greenspan, Chairman of the US Federal Reserve, before the Senate Banking Committee, 25 Feb 1998. ISBN 0-16-057997-X. Cited in Wikipedia.

Often there are issues of **data quality**, which is only in part about whether the values recorded are true or not; the quality issue is also about whether the values recorded are within permitted ranges, or in the correct format.

Understanding database metadata becomes essential in **data warehousing** projects, which are undertaken to create a 'corporate memory' of past records and transactions that can be exploited to provide management information. Maximum benefit is obtained by extracting information from the company's various operational databases and loading it into a comprehensive system within which these records can be inter-related and analysed. Inevitably one finds that, for example, one database refers to males and females as *M* and *F*, whereas from the German side of the operation comes a database using *H* and *D* as labels. The total process within which this metadata is reconciled is known as **ETL**, standing for **Edit**–**Transform**–**Load**.

The GIS community and its metadata concerns

David Haynes[4] notes that one set of data managers that showed an acute and early interest in metadata were those working with Geographical Information Systems (GIS). A variety of organisations, government agencies and systems vendors found themselves collaborating 'on the ground', making it necessary to ensure interoperability between how GIS systems define location.

Haynes dates this to the late 1980s; but even by 1980, work towards standards for GIS location data had progressed to the point where the US Geological Survey took on the role of lead agency in the development of what emerged in 1992 as SDTS, the Spatial Data Transfer Standard. (Work on developing international standards for interoperability continues today through OGC, the open Geospatial Consortium, Inc.).

Z39.50 – a library application of database metadata

During the 1970s, work began on trying to solve the problem of making a distributed search across multiple databases from a single user interface or query form, especially in the field of library and museum catalogue databases. Essentially, the problem was caused by a lack of consistency between the field definitions of the different databases – an example of a metadata mismatch.

To harmonise metadata across such a diverse collection of databases in different languages would be impossible, so the solution, a network protocol standard called Z39.50, translates a query into a coded form, which in turn is mapped to the categories used by the target databases. The University of Glasgow Library, for example, has declared that its database category 'Title' is mapped to the Z39.50 user code 4; as long as the University of Bologna uses the same code 4 for its equivalent field, a pan-European search of university libraries should be able to return relevant results regardless of the languages in which the databases are constructed.

4. David Haynes, 2004. *Metadata for Information Management and Retrieval*, Facet Publishing ISBN: 1-85604-489-0. David Haynes is Head of Consultancy Services at CILIP: the Chartered Institute of Library and Information Professionals.

'Metadata': what it means in the world of online information

In the 1990s, a new set of meanings began to emerge for the term 'metadata', emerging from a collaboration between librarians, scholars and computer scientists. The context for this collaboration was as follows:

- computer hard disk storage had become sufficiently economical and capacious to store extended electronic texts online;
- links were being set up between NFSNet, Usenet, Bitnet, JANET and other networks to create an increasingly global Internet;
- Tim Berners-Lee and colleagues at the CERN particle physics laboratory had invented the Hypertext Markup Language and the World Wide Web, which went live in August 1991; and
- a free Web browser, *Mosaic*, and a free Web server program, NCSA *HTTPd*, had been developed at the National Center for Supercomputer Applications in Illinois (1993).

There rapidly followed an explosive growth in online publishing on the Web, pioneered by universities and research institutions. But people also noted the chaotic nature of Web-based information, and the difficulty of finding useful information just by following hyperlinks between sites ('surfing the Web').

This problem inspired two kinds of approach towards making the Web manageable. The first was the development of Web search engines and their free-text indexes, two early examples being Webcrawler and Lycos (1994) – now eclipsed by Google, Yahoo! and others.

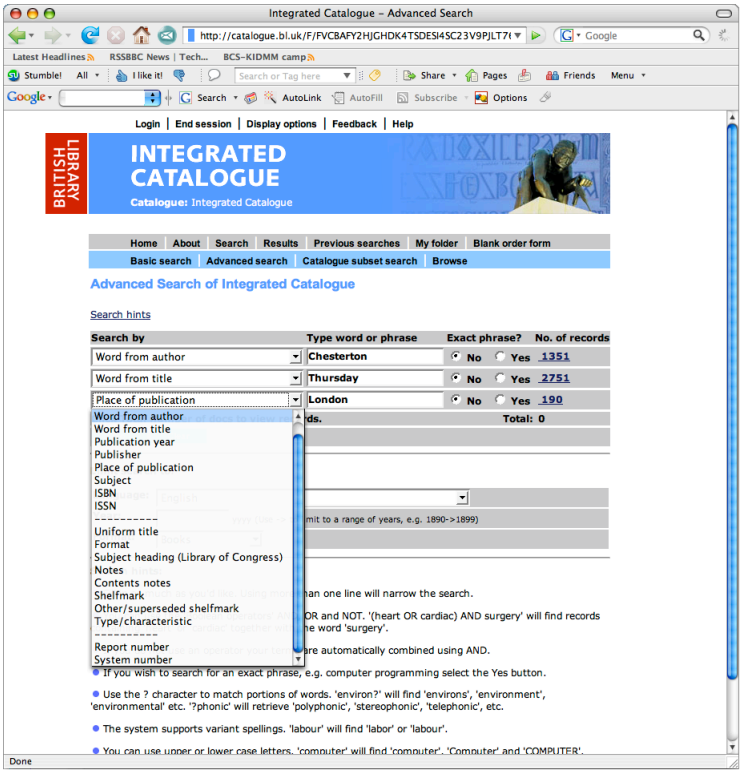
The second approach was based on a very old idea: the library catalogue. It was within the community of diverse experts promoting this approach that the term 'metadata' gained a new currency, gradually slipping over to mean something rather different from its meaning in data management circles.

A basis in the history of cataloguing

Making catalogues of collections of information resources and organising them into categories has a long history within librarianship. Archaeological evidence demonstrates that 4,000 years ago, custodians of clay-tablet libraries in Sumeria practiced document classification and catalogue compilation.* Through the centuries, librarians developed the useful practice of working with **surrogates** for the information sources they manage – it is much more convenient to search a catalogue than to search every shelf of a library. They have also developed **classification schemes**, so that books on similar subjects can be grouped together on the shelves and in the catalogue.

The last 200 years has seen a great deal of progress in standardising how bibliographic data should be recorded in catalogues, in what order it should be presented, and even what punctuation should be used to separate the parts of the record. This is not the place to review that history, but special mention should be made of MARC, the standard for Machine Readable Cataloguing, which from the 1970s onwards has made it possible to exchange bibliographic

* I wish to record my thanks to Aida Slavic of the UCL School of Library, Archive and Information Studies for these historical insights, and for many helpful comments on drafts of this paper.



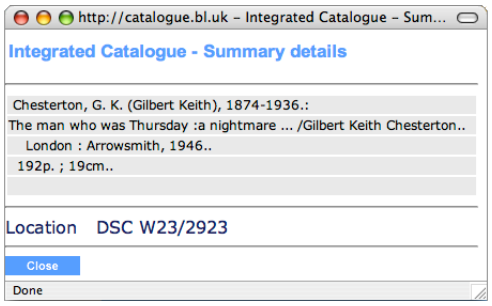


Fig. 2

British Library Integrated Catalogue

Quite sophisticated queries can be made against the catalogue, by defining up to three database fields against which you wish to search at the same time, and a text search string for each.

To the librarian, the contents of these fields constitutes 'metadata', though a data manager would regard the contents as 'data' – a mismatch of meanings between these two communities, as discussed below.

data electronically between libraries. This also laid the foundation for later development of searchable electronic catalogues. (A British Library example of a searchable electronic catalogue is shown above in Fig. 2.)

Resource description metadata – a new twist to the term

As librarians began to examine the problem of organising and cataloguing information resources on the World Wide Web, they began to use the term '**metadata**' to refer to summaries of the attributes of such resources which could usefully be collected to build electronic catalogues of them. As Lorcan Dempsey of the UK Office for Library and Information Networking (UKOLN) wrote in a paper for the British Library's R&D department in 1994:

> Metadata is information about resources, and is of various types, and levels of fullness. In this article it is used inclusively to refer to names, locations and descriptive data which facilitate access or selection. In some cases, the metadata may be no more than a file name and location; in others, in library systems, for example, structured descriptive data may be manually created. Resources are the actual information objects of interest. This article will not say much about the resources themselves, but will focus on their discovery…[5]

Seven months after the publication of Dempsey's paper, an important workshop held in the USA in the town of Dublin, Ohio marked a further step in the development of this community's use of the word 'metadata'. The host

5. Lorcan Dempsey, 1994. *Network Resource Discovery: a European Library Perspective*. In *Libraries, networks and Europe: a European networking study*. Neil Smith (ed). London: British Library Research & Development Department, 1994. Also online at http://www.lub.lu.se/UB2proj/LIS_collection/lorcan.html

agencies for the March 1995 **OCLC/NCSA Metadata Workshop** were the Online Computer Library Center,[6] based in Dublin, OH, and the National Center for Supercomputer Applications, already mentioned for its role in developing tools for the World Wide Web. The 52 invited participants were a diverse group of librarians, information scientists, computer scientists and experts in text encoding and the application of mark-up languages.

The event had as its goal: 'reaching consensus on a core set of metadata elements to describe networked resources'. The result of the workshop was the setting up of the **Dublin Core Metadata Initiative**, of which more below (see p. 14 and following).

The report of the March 1995 workshop explains that while participants considered free-text indexation to be inadequate, they also realised that trying to compile MARC data for Web pages would be too onerous. In any case, the range of information resources on the Web was more diverse than MARC had been designed to deal with; and an attempt in 1992 by OCLC to catalogue the Web, *NetFirst*, had shown the problems of this approach.

Instead, the workshop participants sought to define a small core set of data elements, similar to the metadata model for a catalogue database, which could be used to compile simplified records about 'document-like objects' (DLOs), as they dubbed them. The report explains:

> [A] reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves, without having to undergo the extensive training required to create records conforming to established standards.[7]

Fashionable versus unfashionable definitions?

This is a good place to note the difference between this definition of 'metadata' and that employed by data managers, as explained on page 2 and following. A data manager would insist that what the librarians are calling 'metadata' in the passages quoted above can't be metadata *because they are data to be contained in a catalogue database*.[8] That database would then have its own metadata, as defined by Lundgren and others in the early 1970s.

Reading the Dempsey 1994 paper and the OCLC/NCSA 1995 workshop report, one can almost catch the meaning of metadata changing before one's eyes. It is not so surprising that in the context of a meeting between diverse disciplines (with their diverse vocabularies), jointly attempting to address the explosive new phenomenon of online information, a stylish-sounding piece of jargon got repurposed this way. (Note also that version 2.0 of the HTML specification was published in 1995, defining a new **meta** element for adding

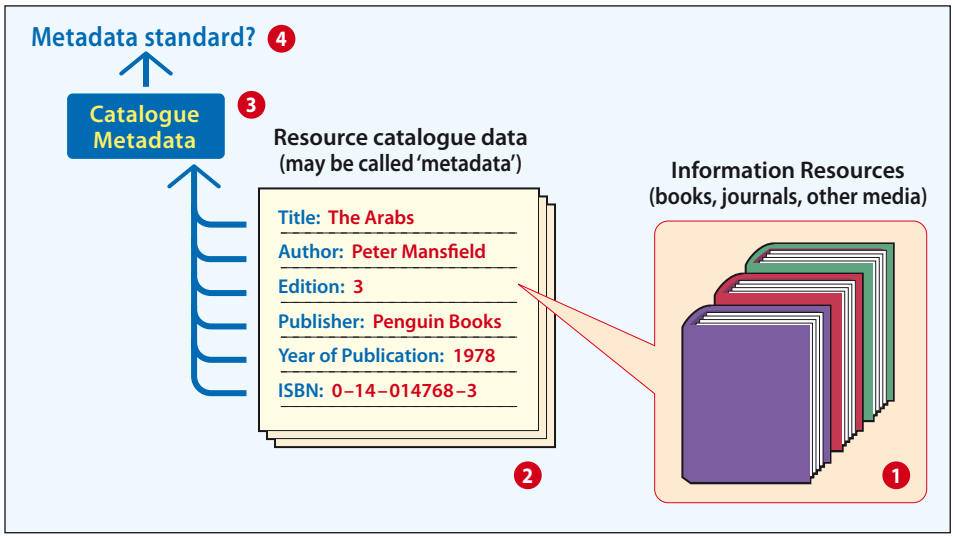
6. Since the late 1960s, the Online Computer Library Center, formerly the Ohio College Library Centre, has played a pioneering role in the computerisation of bibliographic data. OCLC maintains WorldCat, a worldwide database of records in MARC format that is contributed to and shared by over 10,000 libraries worldwide.
7. Report – http://dublincore.org/workshops/dc1/report.shtml
8. See for example the debate captured in *Account of the KIDMM discussion meeting, 6 March 2006*, available from http://www.epsg.org.uk/KIDMM/workshop.html. (KIDMM is the Knowledge, Information, Data and Metadata Management project within the British Computer Society.)

Fig. 3

One man's meat is another man's meta…

The diagram, discussed in the text, illustrates that what a librarian would call 'metadata' is, for the data manager, simply 'data' in a database system that has its own 'metadata'.

But there are many situations in computing in which information makes sense only within a higher framework of definitions, which could be thought of as a general manifestation of 'meta-ness'.



such embedded information as authorship, expiry date, keywords etc. This may also be seen as an indicator of the 'fashionableness' of the term.)

From one standpoint, the data managers are right – they got there first with their definition. However, we must note:

- History has moved on – and due, to the great deal of attention being paid to the problems of managing online information resources, it is the revised resource-description meaning of metadata that has become the one in majority use.
- In any case, it is not strictly true that the new-style metadata is something which necessarily lives in the fields of a database – it is often embedded within information resources themselves, and part of the intention of the Dublin Core Metadata Initiative has been to make this possible.

My general advice to anyone wanting to speak of 'metadata' is to be aware of these different traditions of meaning, and quality what kind of metadata you are talking about if there is any risk of confusion.

A hierarchy of meta-ness

I tend to unify these diverse uses by thinking in general about 'meta-ness' – by which I mean *any* situation in which one set of descriptive information sets the framework whereby another set of information is to be interpreted. This is more common than you may think.

Fig. 3 at the top of this page represents a hierarchy of 'meta-ness' in which a collection of information resources – in this case, books in a library – lies at the bottom [1]. A librarian extracts various kinds of information from the books to add to a catalogue database [2]; this, marked red in my diagram, is what she is likely to refer to as 'metadata', because for her it stands in a 'meta' (*above, abstracted from*) relationship to what she considers the *important* thing, the book or other information resource.

To a data manager, the metadata in this application is that which describes the purpose and function of the fields in the library database – those fields

the names of which are indicated in Fig. 3 by the blue type in the catalogue sample [2]. One could abstract these definitions from the actual database instance, and formally define them as a metadata model [3].

Yet the hierarchy of metadata need not stop here. What if the metadata for the library catalogue is derived from a standard [4], facilitating exchange of catalogue data between institutions – MARC, for example? Standards are not generally described as 'metadata', but they do seem to stand in a relation of 'meta-ness' to the database models which are implementations of them. And so it goes on, for these standards rest upon the acceptance of other standards – right down to the ASCII convention that certain binary numbers will mean certain characters, numerals and punctuation; and this in turn rests on the standard acceptance of there being eight bits in a byte.

In computing, chains of 'meta-ness' are absolutely commonplace. Whether it is Internet protocols, file formats, standards for the formatting of magnetic media, text encodings, mark-up languages, programming languages or whatever – practically everything in computing makes sense to a computer only in reference to a framework of prior definitions.

Metadata on the move

In a world of distributed information objects, there is a very real advantage to having resource description metadata firmly attached to the resource itself. That is why the front pages of a modern book of non-fiction – especially if it is a scholarly one – contain such items as the ISBN code and perhaps Library of Congress 'Cataloging-in Publication Data'. If I need to quote you the ISBN for a book I am recommending to you, I don't need to search a database for it, I can just read it off the back or inside front page of my copy.

There are lots of good reasons why electronic information resources and media should also be able to carry descriptive metadata around with them; here are two:

- In a workflow situation, such as exists in the news publishing industry, there are great benefits to having byline, caption and copyright data inseparable from stories and pictures.
- Where electronic information resources are exposed to the world on the global Internet, embedded descriptive metadata can be interrogated by software agents to compile catalogues.

This goal, however, poses challenges that are not trivial. Not only should the embedded metadata be structured in a standardised way to be of any real use, but the means of embedding that metadata in electronic files also has to be standardised, so that those who write software to access it are dealing with something predictable.

To illustrate how resource description metadata may be embedded in electronic resources, we could do with a case study or two; and for various reasons, I shall take my first example from the world of photography.

Metadata use in commercial photography

Metadata has become central to the management of digital image collections, for three reasons. Firstly, in the commercial use of photographs in publishing there are some essential packets of information in text form that need to be transferred between businesses – copyright, credit, location, date and caption information. Secondly, computers' ability to analyse the visual content of images is still in its infancy, making textual labelling the most efficient way to assist search and retrieval. Thirdly, for many decades picture libraries and agencies have had established practices of categorising and indexing images, using textual entries in databases.

It has long been common for picture libraries to sort their holdings into categories, and attach keywords to images for purposes of retrieval – usually in accordance with home-made categorisation schemes.

There is nothing wrong with having 'private' categorisation schemes, but in certain sectors and for certain industries it is helpful if everybody uses the same scheme. In 1979, the International Press Telecommunications Council (IPTC) – a worldwide consortium of news agencies and news industry vendors – defined a standard series of metadata elements for documenting images, and unambiguous transfer of this data between photographers, agencies and publishers. These elements were revised in 1991 in collaboration with the Newspaper Association of America.

- The **Contact** fields are: Creator, Creator's Job Title, Address, City, State/Province, Country, Phone(s), Email(s), Website(s).
- The **Content** fields are: Headline, Description, Keywords, Subject Codes and identification of the Caption Writer.
- The **Image** fields comprise Date Created, Intellectual Genre, IPTC Scene Code, and location data.
- The **Status** fields record such things as copyright, usage terms and special instructions.
- Although most of these fields can be completed in any way that the participating agency wants, the **Keywords**, **Subject Code** and **Scene Code** entries are special – they must be chosen from a list maintained at *www.newscodes.org*. (See examples, Fig. 4 below)

The IPTC codes serve as an excellent example of a standardised metadata scheme in the service of a business community, not least in its adoption of a 'controlled vocabulary' of keywords and codes for subject and scene type.

Fig. 4: Some sample IPTC Subject Codes, and their explanatory labels in English, Spanish and German

Code	TopicType	English	Spanish	German
01000000	Subject	arts, culture and entertainment	arte, cultura y espectáculos	Kultur, Kunst, Unterhaltung
01001000	SubjectMatter	archaeology	arqueología	Archäologie
01003000	SubjectMatter	bullfighting	toros	Stierkampf
01007001	SubjectDetail	jewellery	joyas	Schmuck

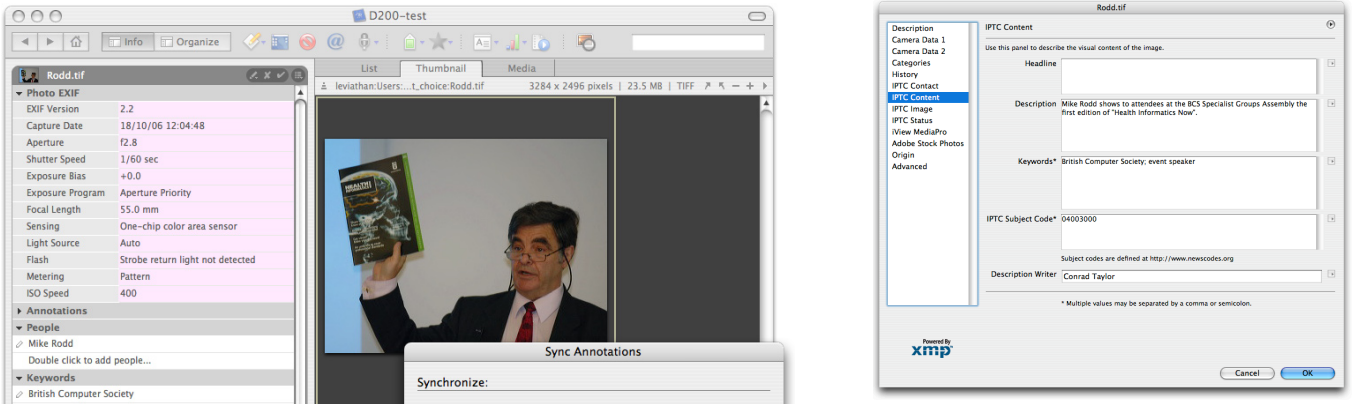


Fig. 5

Exif, IPTC and XMP metadata

Left: iView MediaPro imports Exif data from a digital camera image, acts as an editing tool for adding IPTC and other annotations, and can embed this metadata back into the image.

Above: The edited metadata viewed in Photoshop.

Embedding metadata into digital images

In 1994, Adobe Systems created a specification for embedding IPTC data into images in the JPEG or TIFF formats;[9] at the same time, Adobe enabled their Photoshop application to add this information through dialogues accessed with the *File Info* command. This boosted support of the 'IPTC Headers', as they became known in the publishing industry: when an image is transferred along the production chain, the metadata travels with it.

Following the adoption of digital photography, JEIDA, the Japan Electronic Industry Development Association, developed a standard defining how digital images should be stored – **Exif**.[10] In 2002, version 2.2 of Exif added a wide range of metadata fields for recording shooting data such as shutter speed, focal length, aperture, camera model, metering mode, and of course the exact date and time at which the picture was taken.

Photographers use specialist database programs such as Extensis Portfolio or iView MediaPro to manage image collections. When files are imported to these from a digital camera, the Exif data gets imported too – in Fig. 5, Exif data shows in the top left information panel (pink background).

Within the software, additional metadata such as keywords, IPTC subject codes or names of people can be added. Some of this metadata (such as IPTC Header content) can be embedded back into the files themselves, so it will travel on to the next user of the image. The top right panel in Fig. 5 shows the edited-then-embedded metadata being viewed afterwards in Photoshop.

9. The **Tag Image File Format** (TIFF) was developed in the mid 1980s by Aldus Corporation with Microsoft, and custodianship of the standard is currently in the hands of Adobe Systems: the last major revision is version 6.0 (1992). Between 1988 and 1992, the **Joint Photographic Experts Group** (JPEG) also agreed their standard for image compression, and a file formats was constructed on the basis of this. The significance of these two standards for a discussion of metadata is that they both allow for the addition of extra data fields, and this has been exploited for metadata information interchange.
10. **Exif** is not written in full capitals, not being a true acronym, but is sometimes described as the Extensible Image Format.

Colour management metadata

Within the publishing industry, especially the print-oriented sector, there is tremendous concern with the quality reproduction of digital images. This concern is shared, of course, by such print custoimers as retailers who sell through mail-order catalogues, and advertisers.

The problem is that a photograph – say, a studio fashion shot of a woman's dress – is captured by one digital colour device (the camera), and edited on another (the computer monitor); a page proof may then be printed from a studio printer for the customer's approval. Electronic artwork is then sent to a print works, where printing plates are made, and tens of thousands of copies run off on a press. How do we ensure that colour appears consistently on all these devices? That judgements about quality and any adjustments needed are made with reference to a fair representation of the image? That the the person who orders that dress because of how it looks in the catalogue is not disappointed when it arrives?

Colour management is a special kind of metadata problem. Digital colour is defined by numbers (e.g. red 206, green 102, blue 53) – but these numbers are meaningless unless we know how a physical device like a scanner, camera, monitor or printer represents them. The numbers may stay the same, but the colours come out differently. It would be better if we had a systematic way of tweaking the numbers, so the *colours* come out the same.

The solution recommended by the **International Color Consortium** (ICC)[11] is to measure the colour behaviours of each device – for example, the colours produced by a computer monitor can be measured with a colorimeter. This generates a **colour profile**, which expresses the colour characteristics of the device in relation to a neutral and standard 'connection space' such as CIE–Lab. If the generated profile follows the standardised format promoted by the ICC, it can be called an **ICC profile**. There are published *generic* profiles that can be used instead, e.g. for CRT monitors using EBU phosphors – though such profiles won't be as accurate as one measured from the device itself.

This paper is not about colour management, so I shall sidestep the messy business of how it is implemented! – but for our purposes, what is worthy of note is that by associating a particular digital image to an ICC profile, the numbers that define the colours in it get a meaning.

Colour management is the application of those meaningful sets of numbers, by transforming the expression of them within the colour spaces of different devices as the image moves from one step to another in the publishing chain. It is fair enough to regard ICC profiles as a form of image metadata; and the ICC standard itself as a form of *meta*-metadata that defines how those profiles are written so that machines can read and work on them.

Colour-profile metadata is stored at a system level in the computers used in the publishing process – my Macintosh computer has dozens of **.icc** files, which are colour profiles, most of them generic. When I retrieve digital

11. http://www.color.org/

camera images from my Nikon D200 camera, they include a metadata label that tells my computer whether the image-numbers should be interpreted in relation to the sRBG colour space or the Adobe RGB (1998) one (the user of this model of camera can choose to use one or the other.)

Finally, the image-editing and publishing software I use gives me the ability to embed this ICC profile metadata into the image itself, so the exact meaning I have given to the colours will be understood further down the line.

A little more on Dublin Core

We have already mentioned the Dublin Core Metadata Initiative, a project which started at the OCLC/NCSA 1995 conference in Dublin, Ohio. In the years since, many other workshops, conferences and working parties have carried forward DCMI's work of:

- developing metadata standards for resource search and retrieval across different subject areas;
- defining frameworks for the interoperability of metadata sets;
- facilitating the development of community- or subject-specific metadata sets that work within these frameworks.

The fifteen core elements

The Dublin Core Metadata Element Set has expanded from from a original 13 elements to 15 in version 1.1. Their names and a short description is given below, but it is worth examining the reference document to see the kind of *best practice* DCMI recommends for how to populate these categories.[12]

DC Element	Definition
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content of the resource.
Subject	A topic of the content of the resource.
Description	An account of the content of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the content of the resource.
Date	A date of an event in the lifecycle of the resource.
Type	The nature or genre of the content of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A Reference to a resource from which the present resource is derived.
Language	A language of the intellectual content of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content of the resource.
Rights	Information about rights held in and over the resource.

12. The latest version can always be accessed at http://dublincore.org/documents/dces/

Universality and specificity

The Dublin Core project is in many ways much more ambitious than any particular industry-oriented metadata project, such as the news industry's IPTC metadata. Dublin Core does aim to be truly universal in its application to just about any kind of electronic information resource – though as Stuart Weibel and Eric Miller of OCLA explain, it is only a starting-point.

> The Dublin Core Metadata Element Set [DCMES]… can be viewed as the common semantic building block of Web metadata. Its 15 broad categories (elements) are useful for creating simple, easy-to-understand descriptions for most information resources. Most communities need additional semantics to fully describe their resources, however. …
> The DCMES is the basic block, but other chunks of metadata can be combined with it to form richer descriptions.[13]

In some cases, there will be an argument for augmenting the DCMES with additional metadata elements suitable to a particular domain. In many cases, however, effort will go instead into creating and using reference 'schemes' that give a much higher degree of specificity to how the content of a DCMES element is defined, to suit the needs of particular communities of interest.

An example described in the March 1995 OCLC/NCSA workshop report illustrated this, populating the *Subject* field with six keyphrases drawn from the Library of Congress Subject Headings scheme (`scheme=LCSH`), and then defining the *Object Type* of the resource as a 'monograph', applying the Anglo-American Cataloging Rules (`scheme=AACR2`).

This is where much of the current work in applying Dublin Core metadata is going on, in various DCMI Communities and Task Forces that are looking into fields of application within education, government, digital preservation and archiving, enterprises and corporations and others.

Semantic labelling

Huge investments of effort are being made, in particular, in the devising of **controlled vocabularies and classification schemes** which could be used within metadata element content definitions. The aim here is to progress beyond arbitrary description, in favour of rigorous classification-coding wherever possible, because then it will be possible to use computers as our helpers in searching for the information resources we need.

Who knows if computers will ever be able to make sense of the actual contents of electronic information resources such as documents and images and diagrams? No matter. If we can achieve a greater degree of accuracy in metadata coding – if we can at least make the *labels* attached to information resources machine-processable in some significant way – we'll have achieved much. I shall give a government example of controlled vocabulary labelling on page 19.

13. Stuart Weibel and Eric Miller, 2000. *Building Blocks for the Semantic Web.* Online at http://dublincore.org/2000/09/28-xmlcomarticle.html

Technologies for attaching metadata to electronic information resources

Embedding metadata as structured text

There is never any technical problem with collecting metadata and holding it within a repository such as a database – sometimes known as 'standalone metadata' – but there certainly are challenges involved in embedding it.

We have already seen that certain image file formats – TIFF and JPEG – have been specified in such a way as to lend themselves to having resource-descriptive metadata embedded in them: this is the mechanism by which Exif and IPTC metadata is embedded. Equivalents are now being sought for other kinds of information resource, particularly textual ones.

How can one structure data that is embedded in text files? There is a well established precedent, with its roots in a typesetting control language, GML, developed at IBM in the early 1980s. This work led to an ISO standard, the Standard Generalized Markup Language (SGML), which provided a means for defining entities and embedding them within text files. Although SGML still exists and is in use, it has been superceded by a vastly more popular simplified version – the Extensible Markup Language, or XML.

There is no fixed vocabulary of elements or 'tags' provided by XML, which is often described as a 'metalanguage' – a set of rules for anone who wants to create a custom element set, and a markup language to denote those elements. Once again, as with Dublin Core, XML provides a happy blend of universality and specificity. To receive embedded metadata written in XML, a file format does not have to be re-engineered with an intricate fixed system of data fields; it is sufficient to provide for a single metadata block, of variable size, into which the structured resource description text file is written. This method also allows for metadata definitions to evolve and elaborate over time without 'breaking' the file format.

Below, I shall discuss two XML-based methods now being used to attach metadata to information resources:

- the Resource Description Framework, or RDF, developed by the World Wide Web Consortium (W3C);
- the Extensible Metadata Platform (XMP), an initiative of Adobe Systems, which is based on RDF.

RDF as a metadata framework

According to Rael Dornfest, chief technology officer at O'Reilly Media, the ideas for RDF emerged from reflections about an earlier metadata project, PICS – the Platform for Internet Content Selection. This provided a simple metadata format for classifying and rating Web pages, primarily to protect children on the Internet and filter pornographic content. The PICS project gained experience in deploying metadata vocabularies, digital signatures

and the like, but lacked a 'namespaces' mechanism to allow all the various independently-managed metadata vocabularies to play together.

The purpose of RDF is to provide a standardised framework or method for describing metadata and interchanging it. To explain how it works, we'll need to review some definitions.

- **Resource** — in RDF-speak, a Resource is anything that can have a URI (Uniform Resource Identifier). This includes every Web page, each of which has a Uniform Resource Locator (URL), a limited type of URI that refers to a network location. One can create URIs to define component parts of documents. Indeed, one could create URIs to refer to people, to cars, to books in a library, or to abstract concepts. Anything with identity can receive a URI; then, you can make statements about it.
- **Property** — a Property is a particular type of Resource, one that can be used as a property of other resources. Examples of RDF Properties for publications could include Title or CreationDate. We treat them as Resources in their own right (a) because they have identity and (b) so that they in turn can be assigned Properties.
- **Statement** — an RDF Statement is a combination of three things: a Resource, a Property, and a Value. Two examples: {*Hamlet*} {is a type of *literary work*} {value = *play*}; also, {*Hamlet*} {has a *creator*} {value = William Shakespeare}.[14]

The Value of a Property may just be a string of text or numbers, or may be treated as a Resource in its own right. Certainly in a library context we would expect William Shakespeare to be so treated.

Statements can themselves be treated as Resources, and so they too can have Properties. For example, the statement that {Lucien French} {has access rights} with the {value = 'may upload files to the server'} is likely to provoke the response 'Where's the proof of that?' – and that statement would therefore have to have a verification property, which might be the URI of an encrypted digital certificate.

The RDF standards specify a straightforward way of expressing Statements in the syntax of XML. In fact, RDF defines a specific XML mark-up language, RDF/XML, for representing RDF information and exchanging it between systems.

Readers wanting to delve further may read the 'RDF Primer' put together by Frank Manola and Eric Miller.[15] Here, I shall limit myself to some general observations about the implications of RDF for knowledge and information management, retrieval and publishing.

14. The notation I am using here, with {braces} to indicate Resource, Property and Value, is purely my own creation for the convenience of explanation.
15. http://www.w3.org/TR/rdf-primer/

The key thing to realise is that RDF allows communities of interest the freedom to create their own customised metadata vocabularies: no-one is forced to shoehorn their particular type of information into an inappropriate categorisation scheme. However, because RDF standardises the way in which statements about resources are expressed, it enables machines to parse them easily. In a networked environment, machine-readable explanations of Properties can be made available on-line, so that machines can learn about vocabularies they haven't encountered before.

The RDF Primer mentioned above includes a number of case studies of RDF vocabularies already in use, and one of these is Dublin Core. Dublin Core metadata expressed in RDF can be placed directly into XML and HTML files, and XMP (see below) is emerging as a mechanism to embed it in other kinds of file such as PDF.

For the magazine publishing industry, another interesting emergent RDF application is **PRISM** – the *Publishing Requirements for Industry Standard Metadata*.[16] The emphasis of this project has been to provide extensive means for categorising subjects, using multiple subject description taxonomies; to perform 'rights tracking' for resources such as photos, the use of which has been licensed from others, and for the onward re-use of content in other editions, media etc; and to ensure that metadata is not discarded as content moves from one stage in the publishing process to another.

An RDF application that is having a great deal of impact on the Web is **RSS** – the *RDF Site Summary* standard, though may interpret the acronym as *Really Simple Syndication.* RDF statements allow items placed on one Web site to be syndicated across to others automatically, and summaries of the latest contents of news sites and blogs to be broadcast to Web browser menus and newsreader software.

XMP – Extensible Metadata Protocol, an RDF implementation from Adobe Systems

Adobe Systems first introduced XMP in 2001, as part of the Acrobat 5.0 suite of applications for processing PDF files. Essentially, it is a standardised way for organising document metadata, and embedding it in a variety of file types. The syntax that is used is that of RDF, and the markup by means of which it is structured is an application of XML.

Eleven Adobe applications currently support XMP. These are the image editing programs *Photoshop* and *Illustrator*; the page make-up application *InDesign* and its associated copy-editing program *InCopy*; the technical document composition program *FrameMaker*; the Web page editor *GoLive*; the workflow applications *Bridge* and *Version Cue* and the server products *Adobe Document Server* and *Adobe Graphics Server*. The forthcoming Adobe image management tool *Lightroom* will also support XMP.

16. See http://www.prismstandard.org/

What kinds of metadata does XMP support?

The glib answer is 'any – that's what Extensible means'. In practice, XMP supports for a metadata scheme depends on the availability of templates called *XML Schema* (i.e. the metadata for the metadata, if you like). Custom schema can be added by organisations that need them, but among the ones supported by Adobe software 'out of the box' you will find:

- Dublin Core Schema
- XMP Rights Management Schema
- XMP Dynamic Media Schema
  (to describe e.g. video and music)
- Camera Raw Schema[17]
  (an extensive range of parameters that describe technical metadata embedded in camera raw image files)
- the Exif Schemas for digital photography metadata.

XMP take-up and use

Most use of XMP to date has been made by users of Adobe's own applications, collaborating within commercial publishing workflows. The Adobe Creative Suite 2 (CS2) software bundle includes a lightwork workflow management system for studio workgroups called 'Adobe Bridge', within which the version-control and rights management metadata aspects of XMP are quite useful.

The photography management program iView MediaPro supports XMP, as we have seen, as do Extensis Portfolio and Apple Aperture.

Another group of applications that support XMP are concerned with document management and workflow within corporate and organisational environments, e.g. IBM's DM2 Content Manager and EMC | Documentum.

However, there is not much evidence to date of XMP being taken up in library and archive management applications.

Metadata in public affairs

Thus far I have drawn most of my examples from the worlds of publishing and librarianship, largely because it was with reference to those practices that the newer concept of metadata arose and were first applied. However, in closing I must stress that metadata is becoming a 'hot topic' in a wider variety of fields, such as in government and health service management.

The reason, quite simply, is because formerly closed physical filing systems are being replaced with networks of electronic resources, some of which must remain confidential, some of which are meant for public access, and all of which need to be accessed efficiently online within enormous and proliferating collections. Metadata offers to provide the key both to access management, and to more efficient retrieval of what is appropriate.

17. For an extensive description of Camera Raw files and the role of exposure metadata within them, see my paper *Using Raw Files from Digital Cameras* in this series.

As an example, we may consider e-GMS, which is the British Government's **e-Government Metadata Standard**, at version 3.1 at the time of writing.[18] The government is committed to the use of metadata to improve access to electronic resources within national and local government agencies.

Metadata ways and means

There are various mechanisms by which such metadata might be attached to electronic documents. They could be collected as 'standalone metadata', in database catalogues. They could be embedded in proprietary document information fields, such as those in Microsoft Word documents. They could be embedded as RTF, or XMP. In the case of Web pages in HTML format, the metadata may be embedded as a series of `<meta>` tags.

Each of those means of attaching descriptive metadata has its own kind of 'meta-ness', in the sense of depending on various industry standards. The fact that these various committees have been hammering out such standards and how to implement them in software relieves government of any need to devise its own – and of course also makes the process cheaper, as metadata authored to a broadly-adopted standard can be processed with off-the-shelf or customised software.

The metadata of metadata

The hard part of document description metadata almost always boils down to two main factors:

- **ensuring the quality** of metadata, given that much of it can only be generated by diligent humans; and
- **constraining the descriptions** that are possible for each metadata element, through formal definition of more specific attributes, and the establishment of controlled vocabularies.

The e-GMS is based on Dublin Core – version one of the standard consisted of little more than basic Dublin Core. As e-GMS has become more elaborate, it has added extra attributes for each element – for example, for each kind of metadata declaration there is now a defined attribute of *obligation* – is this declaration mandatory in all circumstances, or only in some, or not at all?

As a further example, consider how e-GMS defines the recommended content of a **Coverage** Dublin Core element. This is intended to indicate either a geographical limitation (e.g. London Borough of Hackney) or a spatial one (e.g. July 2005). Two alternative forms of HTML syntax are suggested by the standard, and a preference indicated for one of them. Here is an example of a Coverage Temporal statement in recommended HTML format:

```
<meta name="DCTERMS.temporal" scheme="DCTERMS.W3CDTF"
content="2006-04-20"/>
```

18. http://www.govtalk.gov.uk/schemasstandards/metadata_document.asp?docnum=1017

It is vital here to understand the function of the **scheme** statement. Several alternative ways of defining time are supported by e-GMS, and there is an oblication to declare which one is using, and then use the correct encoding for that scheme. The three temporal schemes permitted are those of the Government Data Standards Catalogue, the World Wide Web Consortium's Date & Time Format (which is the one used in this example), and the Dublin Core Metadata Initiative's 'Period' definitions.

Controlling vocabulary in public

It is when it comes to filling out the **Subject** field that there is perhaps the greatest risk of chaos. If people were allowed to invent their own categories and keywords, the same categorisation term might be used for dissimilar subjects, or a single subject could be labelled a dozen ways.

In e-GMS, and other e-Government applications, this is solved by referring users to the **IPSV**, the Integrated Public Sector Vocabulary scheme.[19] Thus if you have a document relating to newspaper and magazine publishing, the IPSV preferred term for metadata labelling that would be *Communications Industries*, which has an ID code of 490.

The IPSV also usefully establishes a hierarchy between these terms: thus *Communications Industries* can be a sub-term of either *Business Sectors* (685) or *Information and Communication* (758). This could be exploited in a search system by letting the use of a broader search term disclose the results labelled with one of its subordinate terms, or a synonym.

Metadata for humans and for machines

This whole business of resource description metadata is one in which there are complementary roles for humans and machines. Sometimes the use of a machine to create the resource in the first place makes it easier to scoop up the appropriate metadata – in digital photography, for example, the time is automatically collected (and, with some GPS-capable cameras, even where the photographer was on the earth's surface!). But for categorisation, as in the example just given from IPSV, you want a human to give the final say-so.

Tim Berners-Lee, who invented the World Wide Web, has a vision that he calls 'The Semantic Web'. In this he imagines a Web in which machines understand, at some useful level, the meaning of resources on the Web and are able to act as our intelligent agents. The Semantic Web is beyond the scope of this paper, but it is clear that metadata will play a big part in helping machines to help us deal with those resources in an automated way.

19. Integrated Public Sector Vocabulary – see http://www.esd.org.uk/standards/ipsv/