Journal of Documentation, Vol. 54, No. 2, March 1998

THE

Journal of Documentation

VOLUME 54 NUMBER 2 MARCH 1998

METADATA: A CURRENT VIEW OF PRACTICE AND ISSUES

LORCAN DEMPSEY *and* RACHEL HEERY
{L.Dempsey, R.M.Heery}@ukoln.ac.uk

UK Office for Library and Information Networking, University of Bath
Bath BA2 7AY

This paper describes emerging metadata practice and standards. It gives an overview of the environments in which metadata is used, before focusing on metadata for information resources. It outlines an approximate typology of approaches and explores different strands of metadata activity. It discusses trends in format development, metadata management, and use of search and retrieve protocols. It concludes by discussing some features of future deployment of metadata in support of network resource discovery.

1. INTRODUCTION

However, our feeling is that at this point 'metadata' as a descriptive term has become so debased by overuse (and means so many different things in different communities and contexts) that it is now virtually meaningless without extensive qualification; unfortunately, it has also become a very fashionable term. The very vagueness of the term metadata today makes it easy to offer sophisticated-sounding proposals about using metadata in various ways which seem to be almost impossible to reduce to practice, or which are extremely pedestrian when actually implemented.

Clifford Lynch, Avra Michelson, Craig Summerhill, Cecilia Preston [1]

What a supreme irony that those who proclaim and pursue vision are the least likely to attain it. … And – often – those who are later considered visionary were earlier considered nerds.

Robert Venturi [2]

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

THIS PAPER AIMS to outline some of the issues surrounding the design and deployment of metadata, with special reference to current UK and international developments. An embracing vision sees metadata as pervasively disseminated throughout the network to characterise attributes of people, services, software components and data, in support of self-describing, dynamically reconfigurable distributed systems and services (later examples should make this clearer).

To attempt such a vision, though, would be to fall foul of the charges laid in the first quote above and to submit to the hubris implied in the second. For example, to have promoted a future view of developments three years ago might have ignored the transforming influence of the Web. We are still in a technical construction phase in which the visionary nerd can have major unanticipated influence, in which many theoretical issues remain unresolved, and in which organisational and business issues are yet to be addressed. At the same time, significant commercial and research interest is now focused on issues in this area.

Accordingly, we have a more limited ambition in this article. An opening section approaches a definition by way of example. In subsequent sections, the focus is narrowed to a particular 'type' of metadata, that which describes 'information and document-like resources', and largely to one function of metadata, that of resource discovery. This choice is determined by its presumed readers' interests, but, more importantly, by its authors' competences.

Within this scope, the paper notes some directions and provides significant background material so that there is enough context for the reader to relate these issues to wider developments and trends and to have some sense of some of the environments in which these discussions are taking place. The focus is broadly descriptive rather than analytical.

2. WHAT IS METADATA?

'Metadata is data about data': this is a routine definition, though it is too terse to take us very far. This section provides some examples of metadata and its use before proposing a fuller definition.

2.1 An approach by example

What does metadata look like?

- Many resources include self-descriptive data. Documents typically carry descriptive data: title, author, maybe an abstract, and so on. In a digital version these may be marked by some form of tagging. There may be provision for including some form of metadata in the file. For example an HTML (Hypertext Mark-up Language) document allows data to be placed in the Head. A title can be put in and conventions exist or are being defined for fuller data. AltaVista has some recommendations, which if followed, will enable it to recognise data for harvesting into its search engine. In an SGML (Standard Generalised Mark-up Language) context, there are agreements for the encoding of article headers or, within the TEI (Text Encoding Initiative), of quite rich metadata for electronic texts. Some image files, PNG (Portable Network Graphics)

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

or TIFF (Tagged Image File Format), may also contain some form of structured, descriptive data.

- An interesting document type is an Internet email message which contains simple metadata in attribute value pairs in the mail header. Programs can use this data to provide threaded discussion archives, searches for names or topics, and so on.
- The Electronic Libraries Programme (eLib) has funded several so-called subject-based information gateways. These provide databases of resource descriptions which facilitate discovery of resources in particular subject areas. Typically, the records provide enough descriptive information about network services (often web-sites, but also newsgroups, individual documents, and so on) to allow some precision in searching and some (human) judgement of relevance before committing to retrieving a resource. Their records are rather like those in abstracting and indexing services, with some additional attributes for technical and service characteristics.
- Archivists, records offices, data archives, businesses and government organisations are interested in 'records'. Records are not merely data; they need to satisfy requirements for evidence, what 'it means for written testimony about an act in the past to be considered trustworthy in the future'. Such data may be critical for items such as patient records, data sets, business transactions and legal documents as well as scholarly materials. 'Records require associated metadata which allows a user to audit that the records are comprehensive, identifiable, complete, and authorised' [3]. Data about how to extract information content, about provenance, use history, ownership, terms and conditions of use established by the creator, and so on, needs to be considered in the light of indefinite future use. Preservation raises major metadata issues.
- Statistical data sets need to include supporting documentation to make them useful and usable, which covers such aspects as survey design, processing and analysis, and the data set itself.
- Geographical Information Systems will need to weigh data as it applies to a situation. Metadata may be incorporated into the structure of the data set being operated on to support its proper application [4].
- A database management system contains a data dictionary which preserves the integrity of the database by constraining the operations that can take place on data. Metadata is stored in a data dictionary.
- Z39.50 is an information retrieval protocol. Z39.50 servers can provide data about the databases they make accessible and the facilities they support through an 'explain' database. This might include something about terms of availability or more technical data about supported searches, and so on to allow a client to make sensible decisions.
- It is widely expected that future network systems will be based on distributed objects. Monolithic applications will be disaggregated into components. New applications will be built from multiple autonomous components specialised for particular tasks, put together as and when required. Ideally, one would want an application to discover a

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

component when required, and not to have to have advance hardwired knowledge of all possible requirements. CORBA (Common Object Request Broker Architecture) is one proposed framework for managing interactions between distributed objects. Metadata is an important feature of the CORBA architecture. All objects in a CORBA compliant system must be self describing using a C++-based language: the Interface Definition Language. These descriptions are stored in the Interface Repository, and describe the operations and data types supported by the object. Entries in the repository store enough information about an object to allow other objects to interact with it [5].

What types of things do users, programs or people, need to know about resources?

- A user needs to discover the existence of resources. This is the primary current focus of discussion: resource discovery. However, much of the discussion assumes a flat space of unrelated resources.
- A user may need to know rather more about a resource than some basic description if it is to be useful within a business or research context. Variously, its provenance, archival history, and various forms of intellectual responsibility; its integrity and authenticity; its relationship to other resources; particular domain-specific features; and so on.
- A user as provider or potential provider needs to know about intellectual property rights attaching to a resource, about levels of use made of it.
- A user may need to know whether a resource is fit for use at various levels: whether it is possible to extract its information content with available tools; whether it can be rendered with available equipment; whether a document is a textbook or a scholarly monograph. Additionally, for example, a potential user of an engineering model might need to know what functional goals a device model had, what design assumptions underlay it, and so on.
- A user as customer needs to know under what terms and conditions a resource may be made available. As the Internet becomes a more mature environment, this becomes more important: many of the resources of interest will need to be paid for. However, rather more than a mere listing of price may be needed. Terms and conditions may be situational: they may depend on particular characteristics of a user (frequent user; member of staff of a particular institution …) or on particular characteristics of the use (a discount for night-time use to achieve global load balancing …) and so on. Terms and conditions metadata are seen as crucial, although poorly developed.
- A user as client or agent needs to know about technical interfaces, access protocols, searches and formats supported, etc.
- A user as parent may wish to know what ‘rating’ a particular server has received within some scheme: whether it contains material which is likely to be unsuitable in some way for his or her child to see, and so on.
- A future user … who knows what a future user might expect his or her forebears to provide?

This range supports the point made in the opening quote from Lynch and colleagues. Metadata includes the descriptive or subject data traditionally

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

included in library catalogues, but a great variety of other data also. It should be clear from the foregoing that it does not make much sense talking about metadata divorced from actual uses. Metadata supports particular processes (discovery, preservation, use etc.); any discussion needs to take account of the resources in question and the operations that need to be supported. Metadata for resource discovery is well developed and people feel that they have some understanding of it; there is much more work to be done on appropriate metadata for other operations.

2.2 A definition

These examples confirm that we are looking at a diffuse environment of use. Many others could be offered. However, at this stage we can offer a preliminary fuller definition of metadata:

> metadata is data associated with objects which relieves their potential users of having to have full advance knowledge of their existence or characteristics. It supports a variety of operations.

A user might be a program or a person.

As the proportion of the intellectual record which appears on the network grows, appropriate metadata is seen as a central part of a mature information, business and technical environment. In an indefinitely large resource space, users need to have advance knowledge which allows them to discover resources, know what terms they are available under, assess their potential usefulness, be assured of their authenticity, and so on. Metadata needs to be directed at human users, but increasingly it needs to be addressed at programmatic users. The ability to store searches and user profiles, to consolidate retrieved results from several resources, to filter and summarise, to pass off some of the drudgery of information seeking to programs will be increasingly necessary. These services may be a prelude to more capable agents, autonomous programs which act on behalf of users in distributed, heterogeneous environments. Metadata will assist effective human use of resources; it will be essential for effective programmatic use of resources. Metadata is knowledge which allows human and automated users to behave intelligently.

3. WHAT IS A RESOURCE?

Much of the earlier work on resource discovery in an Internet context had a very simple view of a resource: a resource might be a server or site, or a file on a server. Recently, within the Web community an intermediate level of granularity has been identified: that of a ‘collection’, some explicit aggregation of Web pages. We still use ‘resource’ quite casually though it can cover a wide spectrum of possibilities. A resource might be a file, or a database, or a record in a database, or the metadata about a database. Increasingly, however, resources will be a complex of data and services which may be opaque to a human or robot user. A current awareness service, for example, may operate on a range of resources which remain hidden to the user. A high level interface may be available which hides the internal arrangement or supply of a resource. A resource may be opportunistic or fugitive, existing only in response to a particular conjunction of events or a

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

particular query. Resources will be mutable and dynamic. And again, when we look beyond the 'information' realm, the diversity is significant.

This immaturity in our view of what a 'resource' is, is very clear in current approaches to and discussions about metadata, particularly in those discussions about 'generic' metadata, where one of the design criteria is to be hospitable to a wide range of resources. It is apparent for example in the discussions surrounding the UK subject based services [6] and in parallel discussions about the Dublin Core [7, 8], where there is some tension between a requirement for simplicity and a requirement to recognise the diversity of resource types and the descriptive demands they raise.

Of increasing importance will be the description of relationships: resources are multiply related to others, in ways in which it would be useful to know.

In this context, one can recognise some desirable developments. One would be a typology of resource types which oriented some of the discussion and development. A more fundamental requirement is the ability to name a resource. Clearly, our view of what is a resource is closely related to our ability to name it and issues of naming and identification are now central to current research in networked information. For further information about approaches to naming see Powell [9].

4. REFERENCE MODELS

4.1 Introduction

It would be useful to have a reference model which outlined objects and concepts of interest and the relationships between them: an ontology, to use the borrowed philosophical term. However, partly for reasons suggested in the last section, there is no general view which has guided design and development of metadata formats across domains, where domain is defined by some combination of professional, sectoral or technical characteristics. Even when we narrow our focus to the 'information domain' as suggested above, a variety of organisational and discipline-specific initiatives are in place between which there are different levels of mutual knowledge or influence. In some cases, frameworks are only now being established and organising principles may not be explicit. Clearly, certain application areas require different approaches: geospatial data has different characteristics to bibliographic data, for example. Nevertheless, we suspect that there is significant redundancy in several areas. See Burnard and Light [10], for example, for a comparison of CIMI (Consortium for the Computer Interchange of Museum Information), EAD (Encoded Archival Description), and TEI, SGML-based approaches in museums, archives and electronic texts communities respectively. This may not be an issue now, but will be when it comes to interworking across domains.

A shared ontology, which conceptualised the objects and relationships that needed to be represented in particular metadata formalisms would clarify understanding and facilitate future mapping between domains. Some domain specific metadata approaches are listed below, but first it might be useful to outline a rough sketch of a model generalising the approach of one particular domain, the library community.

Libraries have evolved very full theoretical, technical and organisational apparatuses for resource description and discovery, particularly for books, and there is

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

a full body of experience which can be prospected. However, the aim here is not to suggest that libraries have built on an explicit ontological base for their work (this is not the case), but to give us a handle on some terms and concepts for later discussion and a comparative perspective.

4.2 A look at books

4.2.1 The traditional view Underpinning library cataloguing practice has been a three-fold conception of the 'book':

- *The copy (or item)* This is the actual physical item which is handled and read. Traditionally, cataloguers have not been very interested in the copy *per se* unless special circumstances – if it is very old or rare, or has had an interesting owner – mean it is accorded some special attention. Some copy-level data is assigned to assist in its effective management as part of a collection: a shelfmark or other locating information and a control number. The technology of print typically means that it is one of multiple identical copies of a publication. The book in hand was equal to all other copies as a representative of the publication. There is a copy of *At swim- two-birds* in our office.
- *The publication (or manifestation)* For the reason suggested above, the main library interest focuses on the 'publication' or 'title' of which the copy is an example. Library practice is to start with the published object, exemplified by the copy in hand. The core of library metadata is a physical description derived from this copy – author, title, publisher, place and date of publication, dimensions, format, and so on. In theory, the book is self-describing. Author and title are transcribed as on the copy. The physical object is central as an instance of a publication. Our copy of *At swim-two-birds* is the 'first Four Square edition 1962'.
- *The work* However a publication is only one possible manifestation of a work. There are many other manifestations of *At swim-two-birds*. The work is the intellectual content which may be embedded in a variety of publications. A library cataloguer adds headings or access points which allow the publication description to be retrieved. These headings relate to the work: typically author, title and maybe subject are added. Individual publications may represent author or title names slightly differently; by applying special rules, the aim is to regularise the form in which these are noted in access points. In practice there is often a further stage. To ensure consistency among the access points, an authority list may be used which records preferred regular forms of headings and relationships between them. For example, Flann O'Brien is the author of *At swim-two-birds*. But this is only one of the names under which he is known; he is also Myles na gCopaleen, Brian O'Nolan and other variants. An authority list would establish one of these as authoritative and link the others to it. In some cases, a 'uniform title' may be assigned to an item where the work has some variety in title (the Bible is the obvious case).

So (in theory) the library constructed metadata consists of descriptive data derived from the publication, some copy-specific data, and headings. Headings

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

are supposed to operate at the 'work' level. They are metadata which aims to relieve the user of having to know in advance the individual characteristics of all the manifestations of a work, the different versions of an author name, or all the works on a particular subject.

Cataloguing also provides for relationships to be expressed between these three 'objects' (work, publication, copy). Barbara Tillett identifies a variety of relationships following a review of cataloguing codes: equivalence (copy, reproduction), derivative (editions, translations, ...), descriptive (commentary, criticism, …), whole-part, accompanying (e.g. parts of a kit), sequential, shared characteristic (same author, publisher, etc.)[11]. A variety of linking devices have been developed to express these relationships, influenced, she notes, by the technology used to create the catalogue.

So, provision is made for description of individual published objects and for integrating them into the collections of which they are a part by means of collocation of headings and references. These integrating mechanisms work more or less well depending on the intellectual effort applied to their creation. However, there are some problems.

Emphasis is given to the published object, starting from its format and physical description. As we move into an environment where, increasingly, 'content' may be manifest in several different formats this is a disadvantage and is one of the issues facing those using MARC for cataloguing of electronic resources. The means for bringing together works and indicating relationships do not always work very well and this will create an issue for merging results across many databases. Duplicates will have to be identified. But there is also the 'Humphrey Clinker problem', a term coined by OCLC for the fact that a search on a large database like OCLC's retrieves a large variety of manifestations of the same work [12]. The user is presented with a long listing, based on publication, which has to be read through. The user really wants access at the work level, and an indication of the relationships between publications.

4.2.2 Generalising the book model One of the characteristics of the networked environment is that previously distinct areas may converge. For libraries, 'bibliographic' data and commercial data were kept apart: the booktrade and the library have occupied different worlds. Evaluative data (e.g. book reviews) followed a different channel again. Increasingly, these and other types of data may have to be available together, to support selection decisions and so on. The range of data of interest in these contexts has been discussed elsewhere [13]. If we apply the 'book model' to some unelaborated 'network resource', we could identify the following types of data at each level:

- intellectual content data: describing the characteristics of the work itself, without reference to any particular delivery channel or format: author/other responsibility, title, subject descriptor, narrative description, genre/category/level, review/rating/evaluation, date of creation, ….
- publication or source data: describing the formal and business characteristics of a particular manifestation of the work: identifier, publisher, edition, distributor, terms and conditions, intellectual property rights, format/structure, interface/interchange;

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

- copy-specific: supporting the use of specific manifestations of publications – who owns them, who uses them, where are they, and where there is some special interest (where a copy has some historical or other interest), aspects of provenance, use or marking, integrity/authenticity (in relation to the source): identifier, date of receipt/creation, transaction use, provenance/history, location, owner, integrity, physical characteristics.

A copy inherits publisher/source and content data. To describe a particular resource one wants to say something about its content, the particular business and technical characteristics of the publisher/source, and something about its status within the use chain, to assure a potential user of what it is. Outwith the library world, current practice suggests that no distinction would be made between 'headings' and 'description' as we have outlined it above. One issue which will have to be addressed as users begin to search across domains is the variety of controlled vocabularies and authority lists in use.

4.2.3 Developments in cataloguing theory Some recent treatments have addressed bibliographic 'ontology' issues. In an important article, Michael Heaney models the 'cataloguing world' in object-oriented terms. He argues that users will be best served by an approach which allows users to search for 'works', the intellectual objects of interest to them, rather than by the current approach which puts the 'publication' or 'manifestation' at the centre of attention. He models the bibliographic world in terms of works/texts, publications/manifestation, copies/items, and agents. Agents might be people or corporate bodies. The precise nature of the relationship between agents (authors, translators, publishers, printers, and so on) and other objects is to be modelled as link attributes. He sketches how such an approach can be supported by current techniques and technologies [14].

The IFLA (International Federation of Library Associations and organisations) Study Group on the Functional Requirements for Bibliographic Records has proposed a draft report for review [15]. This report develops an entity-relationship model of the bibliographic world. It introduces a fourth entity, an 'expression', which is a realisation of a work in a particular form. So, for example, a work may be expressed as recorded sound, or in a score. This document also discusses a range of relationships.

4.2.4 Cross domain issues This is a very preliminary sketch. It is presented as a gesture towards what would be a useful cross-domain exercise: the identification of the objects of interest and the relationships between them prior to any representation in particular metadata models. It has the merit of disentangling some of the levels at which objects of interest exist: much current discussion assumes that objects live in a very redundant flat space.

However, other domains may take different starting points. We have suggested that library practice has been to start with the 'publication'/'manifestation', though it might be useful to be more influenced by the 'work'. Interestingly, it has not focused on the 'collection', a notable lack as we move towards a distributed environment in which there is currently little support for database or catalogue

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

selection. In other domains, some notion of 'collection' may be more central. One such is the archives community, which, in the terms presented above, also has more focus on the copy/item level, as the objects of interest to them tend to be unique. A related but different set of functional requirements, including those of evidentiality and establishing context, are in operation. Any 'ontology' will have a different starting point with different objects of interest. Bearman outlines a view based on a study of the archival literature [3]. Similarly, geospatial data, for example, may impose different requirements. The Federal Geographic Data Committee (FGDC) has identified four functional requirements which have guided metadata development: discovery or location (to find what is available), fitness for use (does a data set meet a particular need), access (data needed to acquire a set of data), and transfer (data needed to process and use a set of data).

This is an area which will have to be further developed in coming years. Different domains have developed specific intellectual, professional and technical apparatuses. Domain specific user requirements are met by these systems. However, users also have cross-domain needs. A cultural historian may be interested in the image of the city in modernist literature. A preliminary investigation might wish to consider the Dublin of Joyce, the Prague of Kafka and discussions of Paris by Walter Benjamin. The user would like to look for books and serial articles, for demographic and other social data sets, for images. A child doing a school project on butterflies and natural selection provides another example. He or she may wish to discover the existence of some museum objects, some textbook or encyclopedia discussions and some images. One might expect much of the material for these exercises to be available on the network, in whatever organisational or business framework. However, if current patterns of access are continued, effective use of this variety of resources would be time-consuming and tedious.

Users would thus be well served by some framework which allows domain cross searching. At the same time, many institutions contain within their own collections several domains. For example, a large museum or research library is likely to contain some selection of the following: books, serials, archives, museum objects, specimens, image collections, maps and other geospatial data. They might also wish to provide access to external resources on the network. Even within an institution each domain may have different technical and professional practices associated with it. There are good management and organisational reasons for continuing this specialisation but these institutions are faced with the challenge of providing unified access to their collections. For example a user researching a certain specimen will want to discover relevant exhibits as well as documents.

In summary, libraries and related organisations want to facilitate useful access to the intellectual record, irrespective of what domain it is in. One approach will be to facilitate common 'shallow' access; a common reference model would also be helpful.

5. AN APPROXIMATE TYPOLOGY OF METADATA FOR NETWORK RESOURCE DESCRIPTION

Initial interest in new forms of metadata to manage electronic resources has led to increased awareness of the diversity of formats available for resource description.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

Various subject communities and market sectors are strongly attached to their own metadata formats; considerable effort has been expended on developing specialist formats to ensure fitness for purpose; there has been investment in training and documentation to spread knowledge of the format; and, not least, systems have been developed to manipulate and provide services based on these formats. For these reasons it is inevitable that many of the diverse approaches will continue to exist, and new formats will be created to respond to new user communities and market opportunities.

Specialised formats are optimised for use in particular contexts and by particular communities of user. Inevitably, there is a tension between this drive for specialism and the ambition (whether it derives from the individual searcher or central funding agencies) for a level of interoperability which will allow searching across domains.

We will analyse metadata approaches using approximate groupings based on shared characteristics. The level of complexity of the format can be used as a defining characteristic and this allows a typology of metadata to be constructed along a continuum of successively richer formats. A pattern of association with other factors such as method of creation, search and retrieve protocols, and status as international standards can be shown to follow the placement within this typology. (This analysis is based largely on Dempsey and Heery [16] and Heery [17]).

The following table groups formats into bands along a continuum of growing structural and semantic richness. This allows us to identify shared characteristics of each grouping. This is of necessity a 'loose' typology, any one format may not possess all characteristics of the grouping in which it is placed, but overall the model helps us to compare and contrast characteristics.

In practice some formats may be used in particular circumstances in an a-typical way (e.g. a simplistic 'pseudo-MARC' record can be created with one or two fields), and in such instances a format could be placed in different bands depending on usage. We should note here that discussion focuses on broad approaches or formats: each has a different approach to the relationship between syntax, semantics and the construction of content, which we do not explore in detail.

Band one (see Figure 1) is data derived from full-text indexes of the resources themselves, which might suggest not using the term 'metadata' at all. It is likely to improve in quality as summarisation and extraction technologies improve. Included in Band One is the data generated by the systems in use by the Internet indexing services. Band Two formats are simply structured and generic in scope. It includes several formats which have emerged in the computer science world to support search and directory services. An exception to this is the Dublin Core, a simple metadata element set emerging from a broadly consensual initiative. Band Three is broadly inclusive and is characterised by domain-specific initiatives. Here we find fuller, more structured formats, usually developed to meet specific functional requirements and within particular disciplinary or curatorial traditions. Newer formats in Band Three are often SGML-based or are moving in that direction. A characteristic of some formats in Band Three is that they are typically part of a wider framework which includes markup of 'content' (TEI for example). There are of course many other formats than are mentioned here.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

Band One	*Band Two*	*Band Three*	
(*full text indexes*)	(*simple structured generic formats*)	(*more complex structure, domain specific*	(*part of a larger semantic framework*)
Proprietary formats	Proprietary formats		
		FGDC	TEI headers
	Dublin Core	MARC	ICPSR
	IAFA/WHOIS++ templates	GILS	EAD
	RFC 1807	…	CIMI
…	…		…

Figure 1. *Typology of metadata formats*

It is possible to extend this model to associate other factors with the position of the format on the continuum. Records can be associated with more or less 'rich' retrieval and analysis processes (Z39.50, emerging query routing, text analysis). The bands of records typically have common characteristics in other aspects, for example:

- environment of use: the type of service using the metadata format;
- creation method: the level of manual involvement. This in turn will affect level and type of resource required for record creation and the cost;
- function of record: the variety of functional requirement. Records may be used for location, selection, evaluation, analysis and so on;
- complexity of designation: simpler records do not permit the complex designation of sub-fields, qualifiers etc;
- associated search protocols: more complex formats are associated with the more featured search and retrieve protocol (Z39.50) whereas the simpler formats tend to be associated with directory service protocols or with proprietary approaches.

This pattern of association is summarised in Figure 2.

As noted above, one can argue about this classification. GILS (Government Information Locator Service) is designed to provide descriptive information about government materials, but also people, organisations, events and so on [18], but its origin and emphasis put it in Band Three. One could also internally differentiate Band Three in various ways. For example, GILS, MARC, the Federal Geographic Data Committee's 'Content Standard for Digital Geospatial Metadata' [19] point to independent resources. Metadata in the Text Encoding Initiative and Encoded Archival Description [20] may sit in the same file as the resource it describes. This distinction is hinted at in the organisation of Figure 1. One could also note that several of the formats one might include in Band Two (LDIF – light-weight directory interchange format, for example) lack agreed schema for 'content' description.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

	Band One	*Band Two*	*Band Three*
Enviroment of use	Global Internet search services; Web indexing services	Selective Internet search services; directory services	Descriptions of scholarly collections; other important repositories
Function	Location	Discovery; location; selection	Location; selection; evaluation; analysis; documentation
Creation	Robot generated	Robot plus manual input	Intellectual expertise required, often involving dedicated 'information' staff
Designation	unstructured	Attribute value pairs; limited structure	Subfields; qualifiers; structured mark up
Associated search protocols	http with CGI form interface	http with CGI form interface; directory service protocols (WHOIS++, LDAP) with query routing (Common Indexing Protocol)	Z39.50 SGML browsers and querying
Status	Proprietary	Emerging Internet standards	Domain specific standardisation

Figure 2. *Characteristics of banded metadata formats*

5.1 Band One

5.1.1 Environment of use This is the arena of web indexers and includes services which attempt to offer global indexing of the web and more circumscribed services. They may use simple internal record structures or be based on indexes. A typology of such services is outlined in Koch [21].

5.1.2 Creation method Data is created automatically by a gathering process involving web crawlers (also known as spiders or indexing robots). Such software extracts data automatically from web pages often using the content of the HTML title tag and the first section of the body of the text. Inconsistency of authors' use of HTML may lead to incomplete and unhelpful content. To improve matters, some

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

services seek to enhance the metadata on which they operate: for example different services have different conventions to allow authors of web pages to designate the title and description of a particular web resource.

5.1.3 Function of record Because they typically operate on resources themselves and are often very wide in coverage these services are more effective for location rather than discovery. If a user is looking for a known item, they can be reasonably useful. But for more general subject searching a user may find many resources which have to be sifted and relevant resources may be missed because they are not indexed with appropriate terms. Nor, in many cases, is the description full enough to allow the user make relevance judgements in advance of actually retrieving the resource.

These are automated services and rely on the documents themselves for indexing. They do not provide data about the status of a resource: whether it is fit for the purpose or whether it is what it purports to be and so on. In addition they only cover publicly available web pages; less visible resources (whether because they are commercial, or because they are hidden behind CGI interfaces, or others) are less well covered.

Because of the way they are designed, crawlers parallel the disorganised nature of the web itself – they operate exclusively at the 'copy' level described above. They are not equipped to recognise duplicated files or relationships between resources. It should be noted that the providers of these services are actively working to improve them in various ways.

5.1.4 Complexity of designation Typically there is little by way of structure in Band One: they may consist of full-text indexes. The limited semantic content of what they collect and index – HTML tagging and URLs – may be exploited.

5.1.5 Associated search protocols These services are typically searched using the basic web protocol (HTTP) with CGI (Common Gateway Interface) scripts. There is very limited fielded searching. Harvest, mentioned again below, is a suite of tools for distributed web indexing and searching.

5.2 Band Two

5.2.1 Environment of use Band Two formats tend to be used for services incorporating simple resource descriptions. Often these are selective in some way. These services may have more sophisticated extraction algorithms, or may manually create metadata referring to selected resources. These Band Two formats include the simpler sorts of metadata built by hand, and the more complex of the automatically generated records (often manually amended and enhanced).

Services using these types of formats include OCLC's NetFirst, based on its own internal format, and the UK Electronic Libraries Programme (eLib) subject-based information gateways. Some of the eLib services use their own internal format; some use ROADS templates based on IAFA/WHOIS++ templates [22]. (IAFA stands for the Internet Anonymous FTP Archive Working Group of the Internet

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

Engineering Task Force which originally developed formats to describe network resources. These have been adapted for use by the WHOIS++ protocol.) Such services often focus on descriptions at the server level, creating records for use in repositories or collections of resources. Often, these services involve some selectivity in what they describe and may have more or less explicit criteria for selection. For these reasons, the metadata may be expensive to create, again driving an interest in author- or publisher- generated description and automatic extraction techniques such as those piloted by Essence as part of the Harvest suite of tools [23].

This is an area where there is likely to be significant consensus and development work in the near future, some of it under the auspices of the World Wide Web Consortium.

5.2.2 Creation method Typically formats are simple enough to be created by non-specialist users, or not to require significant discipline-specific knowledge. Descriptions may be manually created, or may be manual enhancements of automatically extracted descriptions. They may be created to be loaded directly into the discovery service database or to be harvested automatically. Web site administrators and authors may create such records locally on their sites, or the records may be created by a central agency on their own databases.

5.2.3 Function of record Band Two includes data which contains sufficiently full description to allow a user to assess the potential utility or interest of a resource without having to retrieve it or connect to it. They support ‘directory’ type services.

5.2.4 Complexity of designation The Band Two metadata tends to be based on simple record structures influenced by RFC-822 style attribute-value pairs. They include a variety of descriptive and other attributes. Formats here do not contain elaborate internal structure but contain sufficient level of designation to allow for some fielded searching. Descriptions tend to be of discrete objects and do not capture the variety of relationships which might exist between objects. The formats do not easily represent hierarchical or other aggregated objects. This is usually by design: there is a necessary trade-off between simplicity and expressiveness. IAFA/WHOIS++ templates are perhaps the most detailed with different template types for different types of object (document, user, service etc.), and there has been some consideration given to ‘clusters’ of data which are likely to be repeated across records and to variants within records.

Their purpose is to be hospitable to the non-specialist description of information objects of different types and from different domains and so are not concerned with the very specific requirements of any one domain. Because of some similarity of construction and content across formats in this band, conversion between them, though inevitably lossy, is feasible. For this reason (as well as the relatively low cost of record creation) these are candidate formats for services that interoperate between domains and media types.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

There has been much recent interest in the Dublin Core, which has been developed to act as a simple description format [8]. Interesting recent discussion over the Dublin Core and the eLib subject based services has exposed the tension between simplicity and structure – which is mentioned below in discussion of the Dublin Core format.

5.2.5 Associated search protocols Selective search services are being delivered through emerging distributed searching and directory approaches on the Internet, notably Harvest, WHOIS++, LDAP (Lightweight Directory Access Protocol) and Dienst. New proposals are likely to emerge as data standards become more mature.

An unknown factor is the influence of Netscape's ongoing work based on Harvest technologies. It is working with Resource Description Messages, as a framework for search and retrieval of metadata [24].

5.3 Band Three

5.3.1 Environment of use There are now metadata initiatives across major scholarly disciplines and within curatorial traditions as they prepare for effective digital use of their materials. Developments include the Inter-university Consortium for Political and Social Research (ICPSR) SGML codebook initiative to describe social science data sets, the Encoded Archival Description (EAD), FGDC's Content Standards for Digital Geospatial Metadata, and the approach developed within the Consortium for Computer Interchange of Museum Information (CIMI).

As might be expected with large international communities of use, there may be more than one initiative in a domain. For example, CIMI proposes a framework for the creation, search and retrieve of metadata, but there are other approaches within the museums community. The Geospatial area is quite well developed, and the Federal Geographic Data Committee has been steering the Content Standards for mapping and geospatial data. Included in this band are TEI independent headers [25]. A range of other initiatives could be mentioned.

This band includes MARC, the format widely used within the library community as the basis for library catalogues, and the most mature of the formats discussed. Within the MARC community there has been some exploration of extending present systems to include metadata referring to electronic resources which complement existing collections. MARC has particular value in enabling integration of metadata into existing library systems. OCLC's Intercat project has provided a pilot in which the USMARC community has explored the creation of electronic resource descriptions, their retrieval by end users, and the requirements for changes to current library systems to manipulate the records.

5.3.2 Creation method They require specialist knowledge to create and maintain. Records will typically be created manually, probably requiring knowledge of the format and some familiarity with cataloguing rules in order to achieve high quality consistent results. Within the library community elaborate business models have been developed for sharing the effort of record creation [26]. There has been considerable investment in training and documentation to spread

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

knowledge of the formats and generations of library management systems have been developed to manipulate and provide services based on MARC.

5.3.3 Function of record This band includes fuller descriptive formats which may be used for location and discovery, but also has a role in documenting objects and, very often, collections of objects. They allow for some level of analysis of content and navigation around aggregations of objects. They are expressive enough to capture a variety of relationships at different levels. Typically, they are associated with research or scholarly activity and cater for specialist domain-specific requirements. We discussed some of the variety of functional requirement across domains above.

5.3.4 Complexity of designation The documentation band contains some very full frameworks for the description of multiple aspects of objects and collections of objects. In some cases, the frameworks describe metadata objects as one type only of information object: they are concerned with 'information content' also. Typically, work is proceeding within an SGML context and the example of the Text Encoding Initiative has been quite influential. Within the social sciences, museums, archives and geospatial data communities work is progressing on establishing Document Type Definitions (DTDs). These may relate to collection level description, item level description, and allow various levels of aggregation and linkage appropriate to the domain. They cater for a very full range of attributes appropriate to documenting data sets or other resources. These can be distinguished from the range in Band Two by fullness (they go into more detail), structure (they contain richer structuring devices), and specialism (they may be specific to the relevant domain).

It seems likely that specialist users will want to search such data directly, but that to make data more visible to more general 'discovery' tools, there may be export of data in some of the simpler formats used in Band Two. Indeed, the Dublin Core has been explicitly positioned as a basis for semantic interoperability across richer formats, although it has not yet been widely used in this context.

5.3.5 Associated search protocols Z39.50 is the most widely used protocol in this area [27]. This was developed in a library cataloguing environment and the MARC community, particularly the USMARC community, was influential in its development. Z39.50 is now used by GILS and CIMI and is being investigated in the geospatial community. In particular, there has been some interest in the Z39.50 profile for access to digital collections [28].

5.4 Publishers' metadata formats

There is at present little integration between formats used for printed material by publishers and libraries but there are a number of parallels worth investigation.

Within the book world records are created at various stages and by different parties in the process of supplying printed publications to the reader. These records are created to fulfil different requirements but have overlapping functions

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

(see Figure 3). The provision and exchange of such trade and bibliographic data has been part of the book world for a considerable time, and various trade, commercial and co-operative apparatuses co-exist within the bookseller, publisher and library worlds. Although there has been some consideration of a more organised evolution of the bibliographic record, the issue of a 'single record system' remains outstanding. In 1987 a seminar attended by publishers and librarians in Newbury considered whether the most appropriate record content and format could be sustained throughout the process to meet the various users' requirements (publishers, suppliers, libraries). It was noted that: 'One important strand to emerge was the idea of an evolving bibliographic record where through more organised articulation of current record supply, or less likely, through the development of an all through single record system, current requirements might be more efficiently met' [13].

The growth of electronic publishing has highlighted the importance of metadata and has encouraged renewed attention to the sharing of record creation responsibility between publishers and information managers. There are parallels with the banded categories used above for analysing the characteristics of metadata in the publishing world. Typically the Cataloguing in Publication CIP) record is now used only for printed material; in the EU BIBLINK project several national libraries are looking at the use of a simple metadata format such as Dublin Core to replicate the CIP process for electronic publications.

The more complex metadata created when publishing electronically in SGML formats offers opportunity for advancing the 'one set of metadata elements' throughout the use chain. One might envisage mapping of SGML to other formats which might be transported using a container architecture such as Warwick Framework.

The library community is only beginning to address how to bring 'book world' metadata and 'network' metadata into the same context of use.

6. DUBLIN CORE AND WARWICK FRAMEWORK

The Dublin Core is a simple resource description format. It has attracted considerable attention recently, partly because of the eloquence and consensus-building activity of its principal proponent, Stu Weibel, but importantly because it has situated itself as a potential solution for three pressing requirements [7]. The first is to have a generally acceptable simple resource description format which is hospitable to the description of a wide range of resources. Following recent

Simple			*Rich*
Publishers CIP forms	CIP MARC	'Item' descriptions EDI messages	SGML article headers SGML book headers

Figure 3. *Booktrade metadata formats*

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

discussions, the Dublin Core (a list of fifteen elements with some qualifying structure) is being adapted to take on board some of the concerns of those with an interest in image metadata and to address some structural and content issues. The second target use is to provide a semantic base for metadata embedded or attached to HTML (and subsequently other) documents. The third target use is to provide a base for semantic interoperability between richer metadata domains. Richer record formats might map a core set of data onto Dublin Core to provide a common set of elements for discovery purposes – this might be implemented in various service and technical environments; for example, at the time of writing, there is discussion about creating a Dublin Core-based attribute net for Z39.50.

Dublin Core looks at one aspect of metadata – simple description – but there is an evident tension to extend the element set to enable more complex description for particular specialist domains, as well as to extend the types of resource described e.g. printed material. In addition, achievement of a concrete syntax in the first target area (HTML) is impacted by innovative proposals within that community such as PICS (Platform for Internet Content Selection), XML and Web Collections. At the time of writing, following the Fourth Dublin Core workshop in Canberra March 1997, there is a recommended syntax that will allow implementation to proceed.

Satisfying the need for competing, overlapping, and complementary metadata models requires an architecture that will accommodate a wide variety of separately maintained metadata models. UKOLN and OCLC jointly organised a conference in Spring 1996 to examine various general metadata issues and the Dublin Core in particular. The venue was Warwick and a new requirement was identified and scoped [8], which resulted in the Warwick Framework proposal [29]. It was concluded that an architecture for the interchange of metadata packages was required. A package is conceived as a metadata object specialised for a particular purpose. A Dublin Core-based record might be one package, a MARC record another, a terms and conditions record another, and so on. Such discrete packages might be numerous and varied in content and even source. Users or software agents would need the ability to aggregate these discrete metadata packages, hence the notion of a container-package architecture.

This architecture should be modular, to allow for differently typed metadata objects; extensible, to allow for new metadata types; distributed, to allow external metadata objects to be referenced; recursive, to allow metadata objects to be treated as ‘information content’ and, in turn, to have metadata objects associated with them.

Although there is wide agreement that this is a sensible direction, the Warwick Framework has not been widely implemented at the time of writing, and certain issues remain outstanding. There is still considerable intra-domain discussion; the need for inter-domain exchange is recognised but the applications framework for this is not in place. Because of the variety of metadata approaches that have been discussed here, and the variety of metadata requirements which may exceed the current provision of any one format, the Warwick Framework has attracted a lot of interest as a simple but potentially very powerful architectural component.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

7. SOME TRENDS

7.1 Blurring of the edges

Of course there is movement across our suggested Bands and this will increase. Author or site produced metadata will become more important for many purposes. This may be harvested unselectively, or only from selected sites. An important motivation for this is to overcome some of the deficiencies of current crawlers without a provider incurring the cost of record creation. In some respects, the crawlers will assume some of the characteristics of Band Two. (AltaVista, for example encourages web authors to embed metadata. Within the Desire project, the Nordic Web Index, a Nordic 'search engine', is being enhanced to be 'metadata aware': where it is available it will harvest discovered embedded metadata.)

At the same time, communities using the richer 'documentation' formats will wish to disclose information about their resources to a wider audience. How best to achieve this will have to be worked out: perhaps 'discovery' records will be exported into other systems. These trends suggest that the Band Two will become more important as a general-purpose access route, maybe with links to richer domain-specific records in some cases.

Various obvious contrasts between the bands are clear. The web crawlers currently operate at a very fine-grained level: they see a world of pages. The services in the middle band face an interesting development challenge: to reconcile the economic and service goals of simplicity and a generic approach with the desire to make descriptive practices responsive to the relatedness of the information world at various levels. The domain specific approaches tend to focus on the description of particular 'collections' and to capture some of this relatedness, but are currently in various stages of development and resources may not be yet visible through general purpose tools.

7.2 Implementation

Standards-based resource discovery services are also in early stages. Examination of the descriptions collected in Dempsey and Heery [16] shows that many formats are still under development or are not widely implemented.

In Band Three (Figure 2), the 'documentation category', in particular, communities of users are working towards consensus and in some cases robust interoperating implementations are some time away.

The 'discovery category' (Band Two, Figure 2), IAFA/WHOIS++ templates are in use in several projects, and are deployed in WHOIS++ directory services. Dublin Core is being piloted in several projects, but an agreed syntax is only now being defined. RFC-1807 describes the format used within the NCSTRL project (Networked Computer Science Technical Report Library URL: http://www.ncstrl.org). SOIF (Summary Object Interchange Format) is widely used as the internal format for Harvest, but there is no agreed 'content' definitions. LDIF (the record used in LDAP, the Lightweight Directory Access Protocol) is in a similar position, lacking an agreed set of schema for resource description. LDIF and SOIF have attracted much interest as a result of Netscape's decision to base its directory server and catalog server products on LDAP and Harvest respectively.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

MARC is the obvious exception here: it is long established in use, although is not widely used for description of network information resources. An interesting development in the US is Federal encouragement for some approaches which make government information more readily available or which improve the operations of federal agencies. This is the case with GILS and the FGDC recommendations [19].

7.3 Maturing environments of use

The discipline or control exercised over the production of collections of resources will improve as the Web becomes a more mature publishing environment. There will be managed repositories of information objects. Such repositories may be managed by information producing organisations themselves, universities for example, by traditional and 'new' commercial publishers, or by other organisations (the Arts and Humanities Data Service in the UK, for example, or industrial and other research organisations, archives, image libraries, and so on). This is not to suggest that the existing permissive electronic publishing environment will not continue to exist in parallel. One concern of a managed repository will be that its contents are consistently disclosed and that descriptions are promulgated in such a way that potential users, whoever they might be, are alerted to potentially relevant resources in that repository.

In parallel, existing and emerging professional sectors within different curatorial traditions are exploring the impact of the electronic environment on their practice. This will result in more robust technical, service and professional frameworks for the creation, propagation and use of metadata.

7.4 A variety of creators and use chains

There will be a variety of metadata creators. These fall into three broad categories: 'authors', repository managers, and third party creators. As its importance becomes more apparent, 'authors' are likely to create descriptive metadata: a major incentive for this will be the emergence of editing and harvesting tools based on Dublin Core or other simple metadata sets. Descriptive data will be similarly associated in other objects by those responsible for their creation. 'Embedding' metadata may not be a sustainable model however and other frameworks are also being investigated. Repository managers, who have some responsibility for a resource and the data that describes it, will also create metadata. Third party creators (including, for example, the eLib information gateways) create metadata for resources that they themselves may not manage or store.

Metadata may sit separately from the resources it describes; in some cases, it may be included as part of the resource. Embedded HTML tags are probably the simplest example of the latter case, but it is common in some of the domain-specific SGML frameworks mentioned above. For example, a TEI (Text Encoding Initiative) header needs to accompany conformant TEI documents. However, independent TEI headers may also exist, which describe documents that may be physically remote.

Metadata, once created may be shared with others. Take for example, author-created metadata embedded in HTML documents. This may be collected by robot

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

or other means. Value will be added to this data at various stages along whatever use chain it traverses: by a local repository manager, by subject-based services, by crawler-based indexing services, by various other intermediary services. Value may be added through automatic or manual means: automatic classification or the addition of subject headings respectively, for example. These intermediary services might include librarians and others who now invest in current awareness and SDI (selective dissemination of information) services, as well, maybe, as current abstracting and indexing services. Many authors may only provide basic information: typically they will not be conversant with controlled subject descriptor schemes, record all intellectual or formal relationships with other resources, and so on.

A different use chain might be traversed by fuller metadata associated with the scholarly edition of an electronic text, for example. Full documentary metadata would be available to assist in the analysis and use of the text, but a subset might be output to a general-purpose discovery service. There might be a link back to the fuller metadata from the shorter record. A part of the motivating rationale for the Warwick Framework, described more fully above, was that resources may have different sets of associated metadata reflecting particular requirements.

A number of factors, including the perceived value of a resource, will determine the relative balance between author-produced, added value and third-party original descriptions in different scenarios. The metadata ecology and economy is still in development.

7.5 Cross domain development

Different repositories will have different requirements and priorities. Examples are a social science data archive, a university web site, a commercial publisher's collection of electronic journals, an archival finding list, and so on. Objects on a university web site may be briefly and simply described. A data archive may need extensive documentation. We have outlined a variety of the approaches through which such requirements are being met.

One can suggest, however, that for many users, their needs will be met best by working across these metadata sources. This was suggested above as one of the motivating uses for the Dublin Core, as a basis for interoperability. Other approaches are also being considered. This will be an active area of research and technical development over the next few years, and some of the issues have been sketched earlier in this article.

7.6 Resource discovery

Metadata currently occupies a number of resource spaces, where a resource space is defined as the sum of resources which are accessible through a particular protocol. There is a large http resource space, as the Web has become the *de facto* user interface to network resources. However, http does not provide a search interface or an explicit framework for managing metadata. Other protocols and approaches will be used. These were discussed above and some are repeated in Figure 4, divided in line with the three bands.

A resource provider 'exports' (push or pull) some representation of resources. These resources may exist at different levels of granularity. For example, there

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

Band One	*Band Two*	*Band Three*
CGI scripts	Harvest, Metadata Content Framework, RDM, LDAP WHOIS++, Dienst	Z39.50, Z39.50 profile for access to digital collections, SGML browsers

Figure 4. *Technologies associated with banded metadata formats*

may be a server level description; there may be desciptions of the individual objects on the server. There may be descriptions of various intermediate 'collections' or 'logical archives'. A user can export a representation of their interests, typically at the moment a simple query, though the use of profiles will increase. In a 'middle layer' a number of services might be available based on aggregation, distribution and matching of representations as well as on other services such as name resolution and so on. These 'middleware' services are currently underdeveloped, but, again, are crucial to the development of mature information services. Different intermediaries may act at several stages: to create metadata, to create user profiles, to use middleware services to present a user with a managed information 'landscape', and so on.

Take a simple case: the use of one of the Internet search engines. Information providers 'export' the whole resource in the form of Web pages. Robots create a representation based on some text analysis. Users represent their interests in the form of single, simple keyword searches. The middleware provides a limited discovery service based on a match between limited represention of user interest and resources.

Consider the subject services supported by the Electronic Libraries Programme. Several of these use the ROADS system to provide services (Resource Organisation and Discovery in Subject-based services). ROADS is developing a distributed systems framework based on WHOIS++, an Internet search and retrieve protocol initially developed for directory applications. The subject services are intermediaries who 'catalogue' resources within particular subject areas. They make those descriptions available as web-accessible databases but also as WHOIS++ servers. WHOIS++ will provide a consistent search and retrieve interface and a technique for routing queries to relevant servers, based on the Common Indexing Protocol which defines 'centroids', inverted index style representations of database content [30]. The aim is to provide the eLib subject gateways with the ability to act autonomously with individual web interfaces, while also being able to be accessed as a collective unit within a WHOIS++ framework. Currently, the resource providers make only their resources available; in due course they may provide metadata, maybe using Dublin Core, maybe using other agreements. ROADS is being equipped with a robot which will harvest these descriptions to give the 'cataloguers' a head-start as they approach the task of resource description. ROADS provides middleware which will help users discover resources based on richer representations than the web indexes, which will route queries to appropriate servers, and which will gradually incorporate other metadata management techniques [6, 31].

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

A third example is provided by the recent initiative in UK higher education to set up 'clumps' [32]. A clump is an aggregation of catalogues. A clump may be 'physical' where it has a continual physically aggregated existence. A clump may also be 'virtual', where the participating catalogues are not physically brought together. How closely coupled the members of a virtual clump might be is seen as a discretionary matter, depending on the particular service scenario involved. The relationship might be entirely dynamic or user-defined, or it might be determined by long-standing service agreements among a group of service providers. The 'glue' that creates a clump was seen to be Z39.50. Other UK initiatives are looking at Z39.50 to 'clump' access to metadata resources. There is some experimentation in the archives world, and at the time of writing the Joint Information Systems Committee is about to fund a Z39.50 cross-archive demonstrator. The Arts and Humanities Data Service will be exploring its use in the cross-domain environment in which it operates [33], and this will provide unified access to metadata stores associated with time-based media, electronic texts, visual data, archaeology data, and some others.

However, these are still 'islands' of interworking. A common set of protocols and data formats will improve interworking, supplemented by a federating layer of middleware which hides underlying protocol and format differences, guides the user to appropriate metadata sources, and provides other services.

At the same time, programs will collect and manipulate data in a variety of ways, passing it off to other applications in the process. Data may be collected and served up in various environments, converted on the fly to appropriate formats: it may be collected by robot, added to a local 'catalogue', or pulled into a subject-based service. The metadata we have been talking about refers to network information resources. This will need to be integrated at some level with the large, albeit highly fragmented, metadata resource for the print literature. There may also be metadata about people, about courses, about research departments and about other objects. Programs might periodically look for resources that match a particular user profile, might search for people with a particular research interest, and so on.

8. CONCLUSION

Metadata will be pervasive of viable digital information environments, to the extent that, as our opening quote suggests, it may be difficult to sustain a general conversation about it. Sensible discussion will involve metadata for particular purposes or within particular communities. Here, our focus has been on resource discovery within 'information' communities. We conclude with some comments about market and policy issues which have largely been outside the scope of this article.

Curatorial professions – libraries or archives, for example – are examining theoretical, service and technical issues. What, for example is a national bibliography when a large part of the national intellectual record exists on the network? Different national libraries are responding in different ways in the context of legal deposit and national bibliography frameworks (see Rugaas for example [34]). The importance of metadata is also highlighted in the 'post-custodial' debate in

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

the archives world [35]. Similarly, it becomes an issue within particular funding or service contexts. The UK Joint Information Systems Committee is making a significant investment in metadata activity within higher education, either as separately funded work, or in association with other funded services, and commands a significant collective intellectual and service resource. A mechanism which supported effective leverage of that resource in concerted ways would be of benefit. The interest of US Federal Agencies in metadata has already been noted. These and other activities share a common interest, effective use of the intellectual record, which is seen to justify investment in rich metadata apparatus. This interest is driving significant intra-domain development, and, as sketched above, will begin to drive a cross-domain research and development agenda. Such an agenda includes, amongst other things, investigation of resource ontologies, semantic interoperability, protocol frameworks, collection and database representation.

As part of a wider pattern of activity, interest in metadata is becoming evident in the product development of Netscape and Microsoft (and in Apple's Meta Content Framework [36]), in the refinement of the internet search engines, in Intranet solutions, and in a variety of other products and services. We are likely to see the emergence of routine tools for local metadata management and creation which assist in more organised disclosure of local resources, and of a greater variety of commercial and other directory services. Metadata will become integral to the web and desktop applications as an organising component as we see a general shift to support for information management and navigation. It is for these reasons that metadata has become of strategic interest to the World Wide Web Consortium, the organisation which oversees technical Web development, and significant work is going into the generalisation of PICS to provide a framework for communication of various different types of resource description.

This intense focus will help create a richer information landscape, one which, as yet, we only dimly discern.

ACKNOWLEDGEMENTS

This work has been supported by the ROADS, Desire and Biblink projects. The first of these is funded under the Electronic Libraries Programme of the Joint Information Systems Committee of the UK Higher Education Funding Councils. The latter two are funded under the EU Telematics Application Programme. UKOLN is funded by the JISC, the British Library Research and Innovation Centre and the EU, with support from the University of Bath. The authors are responsible for any views expressed in this article.

REFERENCES

Note: this is an emerging area which is not yet well covered in the literature. Most of the best sources of information are on the Web. Some URLs are given below. UKOLN maintains a set of pages with links to many other sources of information. These are at http://www.ukoln.ac.uk/metadata/.

1. Lynch, C., Michelson, A., Summerhill, C. and Preston, C. *The nature of the NIDR challenge*. Draft of April 10, 1995.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

2. Venturi, R. *Iconography and electronics upon a generic architecture: a view from the drafting room*. MIT Press, 1966.
3. Bearman, D. and Sochats, K. *Metadata requirements for evidence*. N.d. http://www.lis.pitt.edu/~nhprc/BACartic.html.
4. Danko, D. *Perspectives in the development of ISO metadata standards.* Paper presented at the Earth Observation (EO)/GEO World Wide Web Workshop '97, February 4-6, 1997. http://www.fgdc.gov/Communications/Metadata/nimapaper.html.
5. Orfali, R., Harkey, D. and Edwards, J. *The essential distributed objects survival guide*. John Wiley, 1996.
6. Dempsey, L. ROADS to Desire: some UK and other European metadata and resource discovery projects. *D-Lib Magazine*, July/August 1996. http://www.dlib.org/dlib/july96/07dempsey.html.
7. Weibel, S. Metadata: the foundations of resource description. *D-Lib Magazin*e, July 1995. http://www.dlib.org/dlib/July95/07weibel.html.
8. Dempsey, L. and Weibel, S. The Warwick Metadata Workshop: a framework for the deployment of resource description. *D-Lib Magazine*, July/August 1996. http://www.dlib.org/dlib/july96/07weibel.html.
9. Powell, A. *Identification – related resources.* (Web page created as part of EU project BIBLINK) http://www.ukoln.ac.uk/metadata/BIBLINK/wp2/links.html.
10. Burnard, L. and Light, R. *Three SGML metadata formats: TEI, EAD, and CIMI*. A Study for BIBLINK Work Package 1.1. December 1996. http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/sgml/.
11. Tillett, B. Bibliographic relationships in library catalogues. *International Cataloguing and Bibliographic Control*, 17(1), 1988, 3-6.
12. Svenonius, E. Clustering equivalent bibliographic records. *Annual Review of OCLC Research July 1987-June 1988*. 1988, 6-8.
13. Dempsey, L. *Bibliographic records: use of data elements in the book world*. Centre for Bibliographic Management, Bath University Library, 1989. (BNB Research Fund Report No. 40)
14. Heaney, M. Object-oriented cataloging. *Information Technology and Libraries*, September 1995, 135-153.
15. IFLA Study Group on the *Functional Requirements for Bibliographic Records*. Functional requirements for bibliographic records. (Draft report for world-wide review) IFLA Universal Bibliographic Control and International MARC Programme, Deutsche Bibliothek, May 1996.
16. Dempsey, L. and Heery, R. *Metadata: an overview of current resource description formats.* Version 1, 1997. (with contributions from Hamilton, M., Hiom, D., Knight, J., Koch, T., Peereboom, M. and Powell, A.) http://www.ukoln.ac.uk/metadata/DESIRE/overview/. (Work Package 3 of Telematics for Research project DESIRE (RE 1004))
17. Heery, R. with contributions from Clayphan, R., Day, M., Dempsey, L. and Martin, D. *Metadata formats*. 1996. http://www.ukoln.ac.uk/metadata/BIBLINK/wp1/d1.1/. (Work Package 1 of Telematics for Libraries project BIBLINK (LB 4034))

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

18. Christian, E. GILS: what is it? Where is it going? *D-Lib Magazine*, December 1996. http://www.dlib.org/dlib/december96/12christian.html.
19. Nebert, D. Supporting search for spatial data on the Internet: what it means to be a Clearinghouse Node. (Revised 10/96) In: *Proceedings of the Sixteenth Annual ESRI User Conference, 1996*. http://www.esri.com/base/common/userconf/proc96/TO100/PAP096/P96.HTM.
20. Pitti, D. The Encoded Archival Description (EAD) DTD. In: *Proceedings of the ASIS Annual Meeting*, 33, 1996, 287.
21. Koch, T., Ardö, A., Brümmer, A. and Lundberg, S. *The building and maintenance of robot based internet search services: a review of current indexing and data collection methods*. (Prepared to meet the requirements of Work Package 3 of EU Telematics for Research, project DESIRE. Version D3.11v0.3). 1996. http://www.ub2.lu.se/desire/radar/reports/D3.11/.
22. Heery, R. ROADS: Resource Organisation and Discovery in Subject-based Services. *Ariadne*, 3, 1996. http://www.ukoln.ac.uk/ariadne/issue3/ roads/.
23. Schwartz, M.F. Internet resource description at the University of Colorado. *Computer*, 26(9), September, 1993, 25-34.
24. Hardy, D. *Resource Description Messages (RDM): Technical Specification*. Draft 1.0b3. http://www.nlc-bnc.ca/documents/libraries/cataloging/metadata/rdm.htm.
25. Giordano, R. The documentation of electronic texts using Text Encoding Initiative Headers – an introduction. *Library Resources and Technical Services*, 38(4), 1994, 389–401.
26. Dempsey, L. Bibliographic access: patterns and developments. In: Dempsey, L. ed. *Bibliographic access in Europe: first international conference*. Gower, 1990, 1–29.
27. Dempsey, L., Russell, R. and Kirriemuir, J. Distributed library systems: Z39.50 in Europe. *Program*, 30(1), 1996, 1–22.
28. *Z39.50 profile for access to digital collections*. Draft seven (final draft for review). May 3 1996. http://lcweb.loc.gov/Z3950/agency/profiles/collections.html.
29. Lagoze, C. The Warwick Framework: a container architecture for diverse sets of metadata. *D-Lib Magazine*, July/August 1996. http://www.dlib.org/dlib/july96/lagoze/07lagoze.html.
30. Allen, J. and Mealling, M. *The architecture of the Common Indexing Protocol (CIP)*, Internet Draft (Work in Progress), IETF, 1997. ftp://ftp.ietf.org/internet-drafts/draft-ietf-find-cip-arch-00.txt.
31. Heery, R. Review of metadata formats. *Program*, 30(4), October 1996, 345-373.
32. Dempsey, L. and Russell, R. Clumps or … organised access to printed scholarly material: outcomes from the third MODELS workshop. *Program*, 31(3), July 1997, 239–249.
33. Greenstein, D. Operational requirement for the AHDS's HTTP/Z39.50 gateway and selected Z39.50 servers. 1997. Available from http://ahds.ac.uk/.
34. Rugaas, B. How legal is your deposit? In: Dempsey, L., Law. D. and Mowat, I., eds. *Networking and the future of libraries* 2. London: Library Association Publishing, 1995, 174-178.

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib

Journal of Documentation, Vol. 54, No. 2, March 1998

35. Cunningham, A. Commentary: journey to the end of night: custody and the dawning of a new era on the archival threshold. *Archives and Manuscripts*, 24(2), 1996, 312-321.
36. Guha, R.V. *Meta Content Framework: a whitepaper*. N.d. http://mcf.research.apple.com/wp.html.

(Revised version received 30 July 1997)

Aslib, The Association for Information Management
Staple Hall, Stone House Court, London EC3A 7PB
Tel: +44 (0) 171 903 0000, Fax: +44 (0) 171 903 0011
Email: pubs@aslib.co.uk, WWW: http://www.aslib.co.uk/aslib