In at the Shallow End: Metadata and Cross-domain Resource Discovery

Lorcan Dempsey, Rosemary Russell, and Rachel Heery[1]

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

Contents

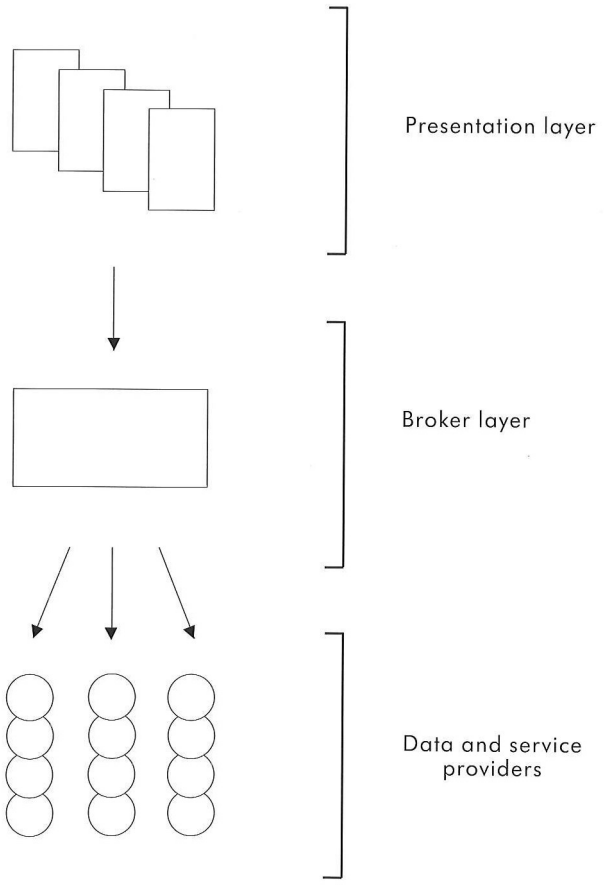
[1] First published in: *Discovering Online Resources Across the Humanities: A Practical Implementation of the Dublin Core*, ed. Paul Miller and Daniel Greenstein (Bath: UK Office for Library Networking; Arts and Humanities Data Service, 1997), pp. 63-71. Online version available at: http://www.ukoln.ac.uk/metadata/publications/ahds-ukoln/

Work reported here on the refinement and implementation of Dublin Core style metadata and on the development of Z39.50-based tools for searching and retrieve such metadata from different underlying information systems advances a wider aim: the effective integration of distributed and heterogeneous networked information resources. In this contribution, Dempsey, Russell, and Heery place the AHDS's and UKOLN's work in a wider context, documenting complementary developments, and illuminating areas where further work is required.

1. Three tiers

The three-tier model for information discovery and delivery which is being explored in a range of initiatives, of which the AHDS is one is shown in Figure 5.1 below.

Figure 5.1. A three-tier model of information discovery and delivery.



Presentation - where the user accesses resources. The Web has emerged as a unifying presentation layer.

Broker - a layer of software which mediates access between users and available resources. Depending on its functional richness, this layer may hide the heterogeneity of available resources, may provide navigation and selection support, may allow users to interact with resources in various ways, may provide authentication, management and other services. In this case, a requirement is to develop a federating solution which allows services to develop autonomously while projecting a single unified image to the user. They need to be seen as a single service, rather than as a series of individual opportunities.

Data and service providers - The specifics of the AHDS's proposed implementation are discussed in Chapters 3 and 4 of this volume, and aspects of the interaction between the 'broker' or 'gateway' or 'middleware' and the individual services are elaborated. Central to the construction of the broker will be protocol support (to manage interactions between components) and metadata (including data which supports these interactions). A central objective is to provide metadata support for discovery and navigation. The purpose of this short piece is to sketch an informal introduction to some wider metadata issues and to briefly relate them back to AHDS service scenarios.

2. Aspects of the metadata landscape

2.1 On the discrimination of metadatas (Dempsey and Heery forthcoming)

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 infinitely 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 software tools that will carry out a range of discovery, transaction, use, and other functions on resources. The ability to store searches and user profiles, to consolidate retrieved results from several resources, to filter and summarise, and to pass off some of the drudgery of information seeking to software will become 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 use of those resources by software tools. Metadata is data which allows human and automated users to behave intelligently.

It should be clear at this stage that it does not make much sense talking about metadata in the abstract, divorced from actual implementations. Metadata supports particular processes to do with resources (discovery, preservation, use, payment, etc.); any discussion needs to take account of the resources in question and the operations that need to be supported. Metadata which supports resource discovery is well-

developed and people feel that they have some understanding of it; there is much work to be done on appropriate metadata for other operations.

One definition of metadata might be:

> *Metadata is data associated with objects which relieve their potential users of having to have full advance knowledge of their existence or characteristics. A user might be a program or a person, and metadata may support a variety of uses or operations.*

2.2 A variety of sectors

Within the constrained area we have established, it is worth noting the variety of active channels of current development. Much of the activity can be seen to be driven by interests under three non-exclusive headings: curatorial traditions, web developments, directory and search services.

2.2.1 Curatorial traditions

Libraries, museums, archives, and electronic text archives. Each has a different approach to the use of metadata. They are interested in describing the resources they manage, in controlling them, in preserving them, in documenting them, in assisting use of them, in providing context for them. They have evolved full, and different, metadata apparatuses, which allow richly structured description of resources and, in some cases, rich depiction of relationships. Typically, they manage scholarly collections, where there is seen to be value in the intellectual creation of metadata, and, sometimes, the use of controlled vocabularies. Work in this area tends to be driven by domain specialists: archivists, librarians, etc. It is apparent that for certain categories of material - individual images in large image databases, for example - less full approaches may be taken.

2.2.2 Web-based information management

After a first phase of rapid development, attention is now turning to issues of more controlled web management. A number of management packages and techniques are appearing, and consensus and standardisation activities are being driven through the World Wide Web Consortium (W3C 1997c). This is an area of vendor push and pragmatic consideration: server, browser and web management software vendors want to support user requirements in consistent ways.

Of interest in this area is work on Web Collections (a way of aggregating web objects at different levels of granularity), on digital signatures, on distributed authoring and versioning, on PICS-ng (Platform Independent Content Selection, next generation), and on transparent content negotiation. In different ways, these all touch on metadata issues and, at the time of writing, the W3C has just announced the setting up of a Metadata Co-ordination Group which is charged with moving a 'Resource Description Framework' forward. This initiative subsumes PICS activities, noted elsewhere in this volume. PICS is a strategic technology because it will become integral to the web and be supported in commonly used software.

Work in this area is being driven by the diverse community of web users. There is an increasingly strong commercial interest, coordinated through the World Wide Web Consortium, as vendors seek a collaborative context for development.

2.2.3 Network resource discovery

A diverse range of directory and search services has grown up in the last couple of years which broadly support discovery of network resources. These include the 'vacuum cleaner' Internet search engines, and more selective services. An example of the latter is the eLib subject gateways, which are based on intellectually created, simple resource descriptions (eLib 1997). Several of the subject gateways use these description records within a distributed framework provided by the WHOIS++ protocol. This service is supported by the ROADS (Resource Organisation And Discovery in Subject-based services) project and is being extended in Europe through the DESIRE project (Dempsey 1996).

This will be an area of further intense work, with a mix of commercial, research and other offerings emerging based on mixed requirements. There was a strong initial impetus from the computer science and computing services area but it is increasingly diverse.

2.3 A variety of metadata models (Dempsey and Heery 1997)

Metadata frameworks are developed to meet different functional requirements, for different types of resources, and different approaches are taken to syntax, semantics, and data content issues. Nevertheless, a loose taxonomy could be presented for descriptive purposes.

Unstructured text	**Intermediate structured formats**	**Structured**
Location/ discovery	Discovery/ web management	Discovery/ documentation/ preservation/ etc.
Web indexes	Dublin Core; Directory formats (e.g. WHOIS++); RFC1807; GILS	TEI CIMI EAD ... MARC
	Attribute value pairs, ...	SGML
Comprehensive	Generic	Domain specific
No semantics	Dublin Core - semantics; Others have syntax focus with particular semantics defined	Syntax and Semantics
Free text searching	Fielded searching	

This table presents a simple taxonomy of metadata formats along an approximate spectrum of increasing fullness, structure, and specialisation.

In the right hand column are domain specific formats, typically SGML Document Type Definitions (with the exception of MARC), which are used and developed by particular curatorial traditions. The influence of the Text Encoding Initiative's Header has been marked here. Examples are the Encoded Archival Description (EAD 1997) and CIMI (1997), for interchange of museum information. These formats are often part of a larger enterprise looking at the encoding of 'content' also. Typically the creator is concerned to describe a collection, however defined, and its components, rather than discrete, unrelated items.

In the middle column are a range of initiatives which often support directory and search services. Again, typically, they form part of a larger framework. For example, the GILS format has been developed in the context of the Government Information Locator System initiative, an approach to the search and retrieval of description of government materials and other objects using the Z39.50 protocol. As noted above, several of the eLib subject based services use WHOIS++ records (Falstrom et al. 1997), based on templates developed in the context of the WHOIS++ protocol.

2.4 A variety of creation models

Continuing the schematic approach we have taken here, we can identify three categories of metadata creator.

The first is the author or creator of a resource. The author of a web page may embed some metadata in the <HEAD> area to be picked up by Alta Vista or another robot. Several Dublin Core aware robots are in development. For example, UKOLN is involved in the DESIRE and NewsAgent projects, both of which have an interest in harvesting Dublin Core records (Powell 1997).

The second is a resource or repository manager. This will become more common as there is more focus on managed information repositories and tools appear to support it. So, for example, a University might provide a managed environment in which the variety of outputs on the campus are consistently disclosed by ensuring the proper association of metadata with resources. Products like Netscape's Catalog Server will support this process. Managers of data archives, electronic text archives, and so on, will also ensure that their resources are appropriately disclosed, a process which is at the core of the interests of this volume.

The third is by third party creators. Again, we can point to the eLib subject gateways as an example here, but there are others emerging from commercial, research, and other sources.

In due course, it is likely that resource discovery metadata might be enhanced or modified as it traverses a use chain that links more than one of these types of 'creators'. For example, an author might embed some metadata which describes a particular local resource. This might be harvested into a subject gateway, where it forms the basis of a fuller record with added data and structure. Similarly, core or generic data from richer metadata stores may be exported in various ways and be used by discovery services.

2.5 A variety of search and retrieve models

Different approaches have been taken in different sectors and domains, and, inevitably, there is a general lack of maturity. One can note some convergences and trends, but fuller discussion of this area is outside the scope of this short piece. There is interest in Z39.50 and in the Z39.50 Profile for access to digital collections within the curatorial traditions (Library of Congress 1997d). Directory services (LDAP and WHOIS++) will become more widely used to support discovery services. A variety of other distributed models are in use - Harvest, for example. Whatever scenarios emerge, and whatever the level of commercial or other control exercised over the data, it is likely that there will be much increased flow and conversion of data in support of a variety of services.

A couple of potentially related issues deserve some comment.

The first relates to the positioning of the Dublin Core to support semantic interoperability between richer metadata models. Much of the discussion of this assumes the creation of parallel metadata stores with links back into richer databases. However, as the work on Dublin Core has been to develop a core semantics for simple resource discovery, it is also possible to support other interoperability models.

For example, the Dublin Core can be used to develop 'attribute sets' which define the semantics of queries within a search protocol - Z39.50, for example. A service might support such an attribute set, undertaking appropriate mappings onto indexes in the background. These issues are currently the subject of technical research and development: as always, several models will co-exist.

The second concerns 'centroids'. The ROADS project has recently been developing some experience with centroids and the Common Indexing Protocol. A centroid is an inverted index-style representation of database content. Centroids are collected in index servers, as a way of supporting 'forward knowledge' of available resources and query routing. In partnership with others, UKOLN will be exploring the use of centroids in other contexts also. Without such forward knowledge parallel searching may wastefully consume human and system resources.

In this context, it will be interesting to explore possible points of contact between the two approaches suggested above (the export of Dublin Core records to a parallel store and the generation of Centroids) as each involves in different ways the generation of surrogates for richer metadata stores.

3. Granularity and aggregation

Information objects can be considered at various levels of granularity and aggregation and are interrelated in various ways. Consider, for example, in an informal way, the various objects of interest in the library domain.

Typically, library practice has been to describe objects at the 'title' level: individual book or serial titles. It has not described more fine-grained objects: serial articles are described in a parallel set of indexing services, and book contents are not usually described. It has sought to provide access to various aggregations of titles in various ways (to works, to titles by same author, to series, etc.).

Libraries have collections (all the material they hold) which contain other collections (the lending stock, the serials collection, the reference collection, the standards collection, the EU documentation collection, 'special collections', slide collections, map collections, the archival collection and the various collections within it, and so on). These contained collections will themselves often contain other collections.

It should be clear that 'collection' is a term without any precise referent, but whose meaning changes with frame of reference. A collection may be a purely administrative category, or may be characterised by various other attributes (medium, thematic unity, provenance, etc.). Acknowledging this vagueness, it is a still a useful term at this level of discussion.

We can also introduce a systems view. We could start with a 'clump', a notion recently introduced within the MODELS project to describe a federation of catalogues (Dempsey and Russell 1997). A clump consists of servers, each of which may make one or more databases available. There is no predictable mapping between catalogue databases and collections.

In the distributed library environment towards which we are moving we want metadata for at least three types of 'object', which we could imperfectly name as 'items' (books, images, etc.), collections, and servers (making databases of item and collection descriptions available). Our current apparatus really only provides access at the item level. Collection description would be useful for several reasons. Whether at the 'whole collection' level, or at component collection level, it supports navigation, grouping (for item level searching), and informational functions. It may also be useful in providing some level of access to collections for which no item level descriptions exist (e.g. image databases, special collections, etc.). Clearly, much work remains to be done in identifying which collections to describe and in deciding what such a description should look like. Catalogue description or representation would be useful to provide 'forward knowledge' to assist in query routing in distributed environments. One will also want information about particular technical and access characteristics. Other metadata may, of course, be required as we move beyond resource discovery.

When we step into the cross domain environment, we recognise that different communities have different requirements, traditions and funding. We suggested that libraries have focused on item level description; archivists are rather more concerned with 'collection' level description, and operate within a different frame of reference. Similarly, as we look towards other domains, not only are there different descriptive practices but the objects of interest themselves may vary.

3.1 Cross domain 'item' description: Dublin Core

The development of the Dublin Core is motivated by several intended uses:

- a simple interchange format for descriptive metadata;
- content self-description for networked objects;
- semantic interoperability across domains.

In the light of the foregoing discussion one can see how Dublin Core has attracted interest, as its target uses coincide with strategic needs in the different sectors we have identified. It provides a basis for a simple resource description format which can be

deployed by directory services. It provides a basis for associating descriptions with web pages and other objects. It provides a way for richer domain-specific metadata formats to interoperate. It is interesting in that it has shown that it is possible to reach a level of consensus about item level description across domains, an achievement that is due in no small part to the eloquence and consensus-making activities of Stuart Weibel.

However, as the results described in Chapter 3 of this volume show, the Dublin Core will benefit from emerging implementation experience. It is still not widely deployed, and will evolve further. It will be especially interesting to see how the balance between simplicity and fuller structure will be maintained.

3.2 Cross-domain collection description and database representation

We have noted how 'generic' collection-level description is much less well developed than generic item-level description. This is an area in which UKOLN plans to do more work over the next year, and we are currently co-ordinating an eLib study to identify existing approaches and to highlight implementation issues. It will focus on approaches in libraries and archives, and also look at directory services, the experiences of the eLib subject gateways, and the work of the W3C on web collections. One aspect of the work will be to examine whether the concept of collection is sufficiently general to benefit from a generic approach.

Areas to be investigated include:

- The Z39.50 profile for access to digital collections. This profile has already been mentioned. Its principal aim was to exploit the various descriptive aids that have been used to describe collections, thus enabling users to navigate and select descriptive aids of interest. It also provides a framework for companion profiles which extend the use of the profile to specific applications: examples so far are the CIMI profile for access to museum objects and a profile for access to digital library objects. The profile only addresses the Z39.50 protocol, although it does not preclude the multiprotocol clients or gateways.
- A draft specification for the implementation of Z39.50 for access to archival collections has been compiled by Fretwell Downing and a UK national networking demonstrator project is being set up under the Archives Sub-Committee of the Humanities Non-Formula Funding (NFF) Committee. The project's aims include demonstrating services using Z39.50, ISAD(G) and data encoded in MARC (UK and AMC) and SGML (TEI and EAD) (UK National Networking Demonstrator Project 1997).
- Conspectus. This is a tool for comparing and analysing existing library collection strengths and current collection policies. Codings are allocated for depth and scope of coverage to a large number of subject subdivisions. It was developed in the early eighties by the Research Libraries Group and has been used throughout the world. An interesting current development is the joint BUBL/SCURL Collections Group initiative to convert the SCURL (Scottish Confederation of University and Research Libraries) Conspectus database to a form suitable for world wide web/ Z39.50 access, where it is seen as a possible support for clump-type applications.

- ISO 2146. This is an international standard, designed to assist in compiling directories of libraries, archives, information and documentation centres and their databases. The second (current) edition dates from 1988 and requires some updating to take account of electronic resource developments. There is little evidence that the standard is being widely used for collection description, though it is used in the French university sector.
- ISAD(G) and EAD. A variety of traditions for describing resources has existed within the archives communities. In response to the need for international standardisation the International Council on Archives appointed an Ad Hoc Commission on Archival Description Standards in 1989. This Commission co-ordinated the development of the eventual ISAD(G) standard which was adopted in 1993 (ICA 1997), and assimilates previously existing archival description standards. The EAD Document Type Definition (DTD) is a standard for encoding archival finding aids for collections of material using SGML. The standard is maintained in the Network Development and MARC Standards Office of the Library of Congress (LC) in partnership with the Society of American Archivists (EAD 1997).
- Approaches to database and collection description in 'directory services'. The study will look at GILS, the eLib subject services, the LIRN project (defining formats within an X.500 context), and other approaches.
- Centroids and Common Indexing Protocol (Allen and Falstrom n.d.). These were mentioned above. Although developed and implemented in a WHOIS++ environment, these are designed to support application in multiple domains.
- Web Collections (W3C 1997d). W3C has also been carrying out some work on collections. This was motivated by the feeling that it would be useful to be able to transfer and manipulate groups of Web documents as if they were either a single document, an ordered list of documents, or a hierarchical structure of documents. A group of members has submitted a draft specification to W3C. Web Collections are an application of XML.
- This selection is not to deny the interest of other approaches, but rather to circumscribe an initial investigation.

4. In at the shallow end - swimming towards the deep

The AHDS gateway will be interesting in that it will be one of the first significant attempts to provide genuine cross domain access. The service providers occupy different subject, curatorial and technical domains, and have evolved appropriate, individual practices. The provision of unified access will be an intriguing challenge.

The gateway needs to provide a generic level of access - to allow the general user to enter at the shallow end of the pool. We have suggested that in due course there needs to be support for item-level searching, for navigation of collections, and to allow sensible use of databases. The initiated may plunge in at the domain-specific deep end!

However, 'shallow' is also appropriate to suggest that we are only beginning to work through the issues involved in such system development, and that actual future

deployment and implementation choices need to be further informed by experiment. We look forward to working with the AHDS in deepening our understanding of these issues.

5. Acknowledgements

UKOLN is based at the University of Bath. It is funded by the Joint Information Systems Committee of the Higher Education Funding Councils of the UK and by the British Library Research and Innovation Centre, as well as by project funding from the EU. This work has been supported by the results of the ROADS, DESIRE, BIBLINK, and MODELS projects.

References

Allen, J. and P. Faltstrom, no date. "The Common Indexing Protocol (CIP)" <draft-ietf-find-new- cip-00.txt> Internet draft, expired

CIMI, 1997. Consortium for the Interchange of Museum Information World Wide Web home page. Consortium for the Interchange of Museum Information
<URL: http://www.cimi.org/>

Dempsey, L., 1996. "ROADS to Desire: some UK and other European Metadata and Resource Discovery Projects", D-Lib Magazine
<URL: http://www.dlib.org/dlib/ july96/07/dempsey.html>

Dempsey, L. and R. Heery, 1997. "Metadata: An Overview of Current Resource Description Practice". Peer review draft of deliverable for Work Package 3 of Telematics for Research project DESIRE
<URL: http://www.ukoln.ac.uk/metadata/DESIRE/overview/>

Dempsey, L. and R. Heery, forthcoming. "Metadata: A Current Review of Practice and Issues", Journal of Documentation

Dempsey, L. and R. Russell, 1997. "Clumps... or organised access to the printed scholarly record", Program, 31/3, 239-249

EAD, 1997. "The Encoded Archival Description Document Type Definition"
<URL: http://lcweb.loc.gov/ead/>

eLib, 1997. Electronic Libraries Programme World Wide Web home page. Electronic Libraries Programme
<URL: http://www.ukoln.ac.uk/elib/>

Falstrom, P., M. Hamilton, L. Daigle and J. Knight, 1997. "WHOIS++ Templates"
<Filename: draft-ietf-asid-whois-schema-01.txt>. UKOLN Internet draft.

International Council on Archives Ad Hoc Commission on Descriptive Standards, 1997. "ISAD(G): General International Standard Archival Description"
<URL: http://www.archives.ca/ica/dds/isad(g)e.html>

Library of Congress, 1997d. "The Z39.50 Profile for Access to Digital Collections". Library of Congress
<URL: http://lcweb.loc.gov/z39.50/agency/profiles/collections.html>

Powell, A., 1997. "Metadata management". Presentation at "Metadata - What is it?", Church House, Westminster, London, 18 June 1997
<URL: http://www.ukoln.ac.uk/metadata/presentations/metadata-june1997/ap/>

UK National Networking Demonstrator Project, 1997. UK National Networking Demonstrator Project
<URL: http://www.niss.ac.uk/education/src/demo_spec.html>

W3C, 1997d. "Web Collections Using XML". Working draft document. World Wide Web Consortium
<URL: http://www.w3.org/TR/NOTE-XMLsubmit.html>