Resource discovery, interoperability and digital preservation: some aspects of current metadata research and development

by Michael Day, Research Officer, UKOLN: The UK Office for Library and Information Networking, University of Bath

Introduction

Metadata is a term that is increasingly being used by the library and information communities and others to refer to structured data that describes or otherwise documents other data in order to support one or more specified functions. These functions may include, for example, resource discovery and access, collection management and resource evaluation, rights management and digital preservation.

Metadata is therefore an important area of research and development, much of it carried out by the library community in conjunction with colleagues working in the cultural heritage sector (primarily archives and museums), research institutions and publishing. UKOLN, the UK Office for Library and Information Networking, has been involved in a number of metadata-based projects and initiatives over the past five years. This paper will introduce some of the metadata-related issues raised by these. Particular, but not exclusive, attention will be made to projects and initiatives in which UKOLN has had some participation.

Internet resource discovery

The Internet is being increasingly used as the medium of choice for the dissemination of a wide (and growing) range of digital information. New dissemination media require new resource discovery tools. One consequence of this has been the development of robot-based Web index services like *AltaVista* and *Lycos*. Web index services are constantly under development, but have been criticised for their poor coverage of the Web space that exists and their imprecision, especially when compared with services based on structured information or metadata. Their main problem is the current state of the Web, where search engines are mostly restricted to making keyword searches of what can be seen as almost featureless, full-text files (Jackson and Gilstrap 1999, p. 316).

An alternative approach to Internet resource discovery might be to add catalogue records for Internet resources to traditional library catalogues or abstracting and indexing services. The library catalogue approach was taken, for example, by OCLC's Internet Cataloging (InterCat) project that was designed to test the use of the USMARC format and AACR2 cataloguing rules for describing Internet resources. However, there is a general awareness that the MARC formats may not be the best 'fit' for the dynamic and fugitive resources that inhabit the Web environment (e.g. Weber 1999, p. 301).

There is a feeling that some of the most scalable solutions to Internet resource discovery would be the embedding of descriptive metadata in Web for harvesting by metadata-aware Web index services or the creation of the specialised services based on selection and cataloguing that are known as information gateways. For both of these approaches, it would be useful to have a standardised, simple 'core' metadata format. This was one of the original motivations of the Dublin Core initiative.

The Dublin Core Metadata Initiative

The Dublin Core Metadata Initiative (DCMI) is an international and interdisciplinary attempt to define a 'core' set of descriptive metadata elements for resource discovery. The element set was initially developed through a series of workshops sponsored by the Online Computer Library Center (OCLC) and other organisations, the first workshop being held at OCLC's US headquarters in Dublin, Ohio in March 1995. More recently, however, the development of Dublin Core (DC) element set has become more formalised with the creation of a Dublin Core Directorate (hosted by the OCLC Office of Research), an Executive

Committee, and an Advisory Committee. The resolution of particular issues has been devolved to a series of open working groups who report back to the DC community and the two advisory committees.

UKOLN has been involved in Dublin Core from almost the beginning, helping to organise the second workshop - the OCLC/UKOLN Warwick Metadata Workshop - now known as DC-2 (Dempsey and Weibel 1996). Members of UKOLN belong to both the DC executive and advisory committees and also chair several working groups. In addition, UKOLN has developed an useful Dublin Core metadata generation tool known as DC-dot (Powell 1999).

The first workshop, now known as DC-1, aimed to develop a set of metadata elements that would be simple enough to allow authors and other information providers to describe their own resources, but which also would be able to facilitate semantic interoperability among resource discovery tools (Weibel *et al*. 1995). By the end of 1996, the initiative had identified and defined fifteen core metadata elements. The semantics of these elements are described in RFC 2413 - the reference description of version 1.0 of the Dublin Core element set (Weibel *et al*. 1998).

Within particular implementations, the fifteen 'core' elements are intended to be both optional and repeatable. They also can be augmented by extension to include additional elements and refined by the use of qualifiers. DC qualifiers were extensively discussed at the DC-4 Canberra workshop in 1997, and take three main forms (Weibel, Iannella and Cathro 1997):

- TYPE (sub-element) - a qualifier that narrows the semantics of an element name. For example, the value of a DC creator element could be specified as being either a personal name or a corporate name.
- LANGUAGE - a qualifier that specifies the language of an element value.
- SCHEME - a qualifier that notes the inclusion of an element value taken from an externally defined scheme or standard. Examples would include; for example, subject classification codes or titles used for authority control.

The initial focus of DC was the Web, so the initiative produced guidelines for the encoding of DC elements in HTML Meta-tags (Kunze 1999). In this way, Dublin Core metadata can be embedded into existing Web documents where it can then be harvested by metadata-aware indexing robots like that developed for the Nordic Web Index (NWI).

Dublin Core and the Resource Description Framework (RDF)

Recent developments in DC have concentrated on implementing the element set using the Resource Description Framework (RDF), which is a set of conventions for expressing metadata using the Extensible Markup Language (XML). Like XML, RDF is an initiative of the World Wide Web Consortium (W3C) - the organisation that overlooks the development of open Web standards. RDF provides a data model and a means of expressing this model in XML (Miller 1998). The framework has been under development since 1997 and the *RDF Model and Syntax specification* was released as a W3C Recommendation in February 1999 (Lassila and Swick 1999).

Unlike HTML META tags, RDF/XML has been specifically designed to be a container for metadata and it permits the unambiguous use of multiple metadata element sets on the Web by means of the XML Namespace concept. This allows RDF statements to reference a particular RDF schema - especially important where the same headings might be used to refer to quite different things. Potential conflicts are avoided because an element tag can contain a code that indicates the particular RDF application to which that tag belongs. For example, Dublin Core metadata encoded in RDF/XML might contain a statement like the following (Miller, Miller and Brickley 1999):

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:dc="http://purl.org/dc/elements/1.0/">
```

Any element tag containing a "dc" namespace can then be assumed to conform to the semantics of the Dublin Core metadata element set as defined in the reference description of DC version 1.0. Additional elements from any number of different namespaces can be added into any RDF statement. In this way, RDF facilitates modular interoperability among different metadata element sets by creating what Eric Miller (1998) calls "an infrastructure that will support the combination of distributed attribute registries".

The Dublin Core community has been active in the development of RDF and has also used the tools provided by the RDF model to help refine the Dublin Core element set and its data model. In particular, the Dublin Core Data Model working group is working on identifying a common structural expression of DC qualifiers. The use of qualifiers within DC implementations has always been inconsistent and it has been suggested (Weibel 1999) that some standardisation of the semantics and methods for qualification of the basic elements would be necessary if qualified DC is to be widely interoperable. The underlying RDF model has informed much of this work on the qualification of Dublin Core, although implementations of qualified DC will not of necessarily have to be based on RDF/XML.

Internet information gateways

A different, but complimentary, approach to solving the Internet resource discovery problem is the development of specialised Web-based services known as information gateways (Worsfold 1998). These services aid resource discovery by allowing human-generated descriptions of selected Internet resources to be searched or browsed.

The simplest gateways may just consist of Web pages containing lists of hypertext links to resources, but more mature services tend to be based on databases of human-generated resource descriptions (metadata). Firstly, resources are selected according to some pre-defined criteria - usually based on the particular subject coverage of the gateway and (possibly) some measure of 'quality' - and then metadata for that resource is created. This metadata typically contains bibliographic-type information together with contact details and subject classification codes that can be used to form the basis of a hierarchical browsing interface.

UKOLN has been involved in a number of collaborative projects that concern the development of information gateways and the tools that support them. Chronologically, the first of these was the ROADS (Resource Organisation And Discovery in Subject-based services) project.

The ROADS project

In 1995, the Joint Information Systems Committee (JISC) of the UK higher education funding councils began to fund a number of gateways under its Electronic Libraries Programme (eLib). The rationale was to explore the validity of the information gateway approach to Internet resource discovery (Russell 1998, p. 368). These gateways covered a range of subject areas, for example: the social sciences (SOSIG), medicine (OMNI), urban design (RUDI) and history (History). As part of the same strand of eLib, JISC also funded the ROADS project to develop software and other tools that would support the development of the eLib gateways and contribute to the wider resource discovery environment.

ROADS was a collaborative project involving the Institute of Learning and Research Technology (ILRT) at the University of Bristol, the Department of Computer Studies at Loughborough University and UKOLN. The project has resulted in the development of an open-source software toolkit that is used by an increasing number of gateways - both within the UK and elsewhere. Gateways based on the ROADS software toolkit give access to a database of resource description for selected Internet resources that are stored in a metadata format known as ROADS templates; a simple text-based format based on attribute-value pairs. A number of different template-types are available allowing the description of a number of different resource-types. ROADS-based services have great freedom with regard to which particular tools they choose to implement and they ways in which they can configure their search and browse interfaces.

The ROADS project, however, was not only concerned with providing tools to create standalone gateway-type services. Rather, the project partners wanted to develop tools that would promote interoperability and allow the easy cross-searching of one or more distributed gateways. ROADS (version 1) uses the Whois++ search and retrieve protocol (Deutsch *et al*. 1995) to allow cross-searching between one or more information gateways. This may be useful in one of several situations. For example, a user may wish to search two (or more) different gateways for information on the same topic. Alternatively, a gateway might be able to geographically distribute its service over a number of servers - say on a regional or national basis - but retain a unified searching and browsing interface.

ROADS (version 2) makes use of the centroid facility of Whois++ to facilitate query routing between servers. If this is implemented, an 'index server' would periodically visit chosen ROADS-based

information gateways and generate an index summary (or centroid) for each. This centroid will contain all relevant index terms in that database so that an initial search of the index server will determine which of the subject services will have information that matches a given query. If desired, the query can automatically be passed on to all of the subject services whose centroids indicate the existence of relevant index terms and the relevant templates returned for display to the end user. Demonstrations of ROADS cross-searching using Whois++ and centroids have been made available on the Web (ROADS project 1998). A discussion of the technologies that underlie cross-searching and query routing using centroids can be found in a paper by Kirriemuir *et al*. (1998).

In order to help preserve a minimum level of interoperability between ROADS-based services and to help facilitate cross-searching, the project created a number of resources that it was hoped might help promote interoperability (Day 1999). The project first set up a simple metadata registry - the *ROADS Template Registry* - to record information about all template-types in use and their attributes. In addition, the project developed generic cataloguing guidelines in an attempt to help ensure that the information content of ROADS templates would remain broadly consistent (Day 1998a).

The ROADS project also investigated the possibility of interoperability with other metadata formats and search protocols. For example, the project looked at semantic interoperability issues by developing several metadata crosswalks, mapping ROADS templates to both USMARC and DC (Day 1996). The project partners also developed test implementations that permitted ROADS-based services to interact with the Z39.50 search and retrieve protocol. For example, the project developed an experimental Z39.50 to Whois++ gateway called ZEXI (Powell 1998). The gateway functions as a Z39.50 server, accepting queries from Z39.50 client systems and then converts them to Whois++ queries and passes them to a ROADS server. As the ROADS server returns results, they are then converted into a suitable format for use by Z39.50 client systems and returned to the client as a Z39.50 result set.

The DESIRE project

The success of the information gateway approach to Internet resource discovery has meant a steady increase in the number of gateways being made available. It has also resulted in an ongoing series of research projects that are devoted to investigating the issues that surround them. For example, back in 1996, the three ROADS partners joined together with the Koninklijke Bibliotheek (the National Library of the Netherlands) and NetLab (the Lund University Library research and development department), to participate in the DESIRE project. DESIRE (Development of a European Service for Information on Research and Education) is funded by the European Union under the Telematics for Research area of its Fourth Framework Programme, and is primarily concerned with improving European researchers' use of information networks. The original project (1996-98) - now known as phase 1 - covered a very wide range of topics, including cataloguing and indexing, caching, security issues and training, and was co-ordinated by SURFnet in the Netherlands.

Work Package 3 (WP3) of DESIRE concerned itself with what it called 'cataloguing and indexing', and took a two-strand approach to the Internet resource discovery problem. The first of these built upon the experiences of NetLab and the National Technological Library of Denmark (DTV) in the development of distributed, robot-based Web indexes as part of the Nordic Web Index. As part of this strand, NetLab and DTV produced a state-of-the-art review of indexing and data collection methods used in Web search engines (Koch *et al*. 1996) and a functional specification for a European Web Index (Lundberg *et al*. 1996). The project also resulted in the development of a metadata-aware harvesting robot called *Combine* (NetLab 1999). *Combine* was especially developed for services that need to specify rules for the URLs or servers that need to be targeted, e.g. for services that need to cover a particular country or organisation.

The second strand of WP3 was mainly concerned with information gateways and built on the experiences of the ROADS project and a number of existing gateways. These included the eLib-funded SOSIG and Biz/ed gateways (both based at ILRT), the Lund-based EELS (Engineering Electronic Library, Sweden) and the Nederlandse Basisclassificatie Web (NBW) hosted by the Koninklijke Bibliotheek. Initial work within DEISRE WP3 centred on the production of a three-part *Specification for resource description methods*. This deliverable included a detailed state-of-the-art survey of selected metadata formats (Dempsey *et al*. 1997) and an evaluation of the use of subject classification schemes for providing access to Internet resources (Koch *et al*. 1997). A study of selection criteria in use by gateways (Hofman *et al*. 1997)

led to the development by ILRT of a Web-based tutorial called *Internet Detective*; an interactive tutorial designed to help users evaluate the quality of Internet resources.

In 1998, DESIRE entered a second phase and began to focus on four main areas: distributed Web indexing, information gateways, directory services, and caching. DESIRE 2 builds upon the work carried out in phase 1 of the project but is primarily concerned with helping to facilitate an organisational framework in which both index services and information gateway can begin to operate more efficiently. The project is, therefore, continuing to explore interactions between the two different types of service. Specific topics being considered as part of WP3 of DESIRE 2 include frameworks for metadata registries, the implementation of quality ratings information in RDF and the production of a generic information gateways handbook (Belcher, Knight and Place 1999).

Other projects and services

As the number and variety of gateways grow, there is a growing awareness of the need for gateways to collaborate. At a meeting in 1998, an informal organisation known as IMesh was set up to help encourage international collaboration amongst information gateways.

This collaboration has led to a number of new information gateway projects and initiatives. The IMesh Toolkit project has been funded under the National Science Foundation (NSF) and JISC International Digital Libraries Initiative to develop a configurable, reusable and extensible toolkit for subject gateway providers. At the same time a research strand will consider issues of relevance for information gateways, including metadata sharing and reuse. Project partners include ILRT, UKOLN and the Internet Scout Project based at the University of Wisconsin-Madison.

The UK Resource Discovery Network (RDN) is an important attempt to build upon the experience of the eLib gateways (and other initiatives) in order to create a single point of entry for the UK learning and research communities to a range of resources - including those distributed through the Internet. The RDN is funded by JISC, the Economic and Social Research Council (ESRC) and the Arts and Humanities Research Board (AHRB) and consists of a Resource Discovery Network Centre (RDNC) and a number of independent service providers called 'hubs'. The RDNC is run jointly by UKOLN and the Arts and Humanities Data Service (AHDS) based at King's College London and will set service standards, create collection development policies and explore strategic partnerships. Hubs that have been currently agreed include one for the social sciences, business and law (SOSIG), another for engineering, computing and maths (EMC, including EEVL) and a third for the biomedical sciences (BIOME, based on OMNI). A fourth hub will be based on Humbul and cover the humanities. Further hubs to cover the remaining subject areas will shortly be agreed.

Interoperability

Format conversion

Information gateways are only one of an increasing number of digital information resources that are being made available through the Internet. These resources are based on a large number of different data (or metadata) formats and made available through a number of different protocols. For example, the DESIRE study of metadata (Dempsey *et al*. 1997) described over twenty different metadata formats or standards that had been produced (or were under development) in 1996. The editors of the review thought that format diversity would persist, as it was unlikely that any one single element set would be perceived to be applicable to all types of digital resources. They also noted that different subject communities and market sectors have spent - and continue to spend - considerable effort in developing specialised metadata formats and the associated systems designed to provide services based on them (Dempsey and Heery 1998, p. 155).

It is no accident, therefore, that many metadata initiatives and projects have had to concern themselves at some point with interoperability issues. We have already seen, for example, that the ROADS project provided some resources that would help promote consistency between different gateways (for improving cross-searching), produced some metadata crosswalks, and experimented with developing tools that would help integrate services using different search protocols.

Promoting interoperability was also one of the original objectives of the Dublin Core initiative. Weibel (1997) has suggested that a core set of metadata elements could act as an intermediary for semantic interoperability between heterogeneous resource description models. One consequence of this has been the development of a growing number of crosswalks that map the Dublin Core elements to a variety of other metadata formats, including the USMARC format, the FGDC metadata standard and ROADS templates (Day 1996). Interoperability issues have also provided a focus for some Dublin Core implementations. For example, the Nordic Metadata Project produced a number of software tools, including a utility called *d2m*, a Dublin Core to MARC converter that converts Dublin Core metadata embedded in HTML into various Nordic MARC formats and USMARC (Hakala, *et al*. 1998). Format conversion tools have also been developed for the BIBLINK project.

The BIBLINK project

The BIBLINK project is funded by the European Union under the Telematics for Libraries area of its Fourth Framework Programme. Within the project, the British Library leads a consortium of partners that include the national libraries of France, the Netherlands, Norway and Spain, the Universitat Oberta de Catalunya and UKOLN. The project aims to promote electronic links between the publishers of electronic material and national bibliographic agencies.

The first stage of the project involved the production of a series of studies that looked at metadata formats, the feasibility of format conversion, digital identifiers and authenticity. The main project deliverable has been the development of a custom-built software system known as the BIBLINK Workspace (Day, Heery and Powell 1999). This demonstrator system has been designed to test the conversion of metadata produced by publishers into formats that can be used by national bibliographic agencies for inclusion in national bibliographies. The BIBLINK Workspace demonstrator, therefore, takes metadata records in one of two designated formats - an extended form of DC known as the BIBLINK Core (BC), and a selected type of SGML header - and converts them into the Universal MARC (UNIMARC) format. These UNIMARC records can then be converted into the formats - usually a different flavour of MARC - used by the participating national bibliographic agencies, who can then enhance them for inclusion in the national bibliography or for returning to the publisher.

Distributed library systems

Metadata crosswalks and format conversion tools, however, only solve a small part of a much larger interoperability problem. One of the biggest challenges facing those who are attempting to develop digital libraries at the present time is attempting to integrate access to the wide range of distributed and heterogeneous information resources and services that are available. The successful integration of these resources and services is perceived as of being of great benefit to libraries and their end users. Dempsey, Russell and Murray (1999, p. 35) point out that resources are typically differently presented, accessed and structured, and that users, for example, may have to interact with a number of quite different information systems in order to carry out a full search. They suggest the development of an additional service layer - here described as 'middleware' - that would shield the user from any underlying complexity and heterogeneity. This middleware - a broker service - would need to provide "a higher level interface, creating a federated resource from underlying heterogeneity and mediating access to it" (Dempsey, Russell and Murray 1999, p. 38).

Several projects and initiatives have tried to address these issues. For example, the Stanford Digital Library project developed an infrastructure known as an information bus (InfoBus) based on CORBA (the Common Object Request Broker Architecture) that should be able to translate formats, broker services and support financial transactions (Paepcke *et al*. 1996; Baldonado *et al*. 1997). UKOLN's involvement in these types of interoperability issues has centred mainly upon the MODELS initiative, but has recently branched out into the Agora project and the hosting of the JISC and Library and Information Commission-funded Interoperability Focus post.

The UKOLN MODELS initiative

The MODELS (MOving to Distributed Environments for Library Services) project is an UKOLN initiative that has gained additional support from JISC (through eLib) and the British Library, with Fretwell

Downing Informatics (FDI) as technical consultants. MODELS provides a forum - primarily through a series of workshops - that can allow relevant stakeholders to explore shared concerns about distributed and heterogeneous resources and services. The initiative has attempted to address design and implementation issues, initiate concerted actions, and work towards a shared view of preferred systems and architectural solutions.

Resource discovery and metadata issues have featured widely in MODELS workshop discussions. For example, the MODELS 2 workshop was also DC-2, the second Dublin Core workshop that developed the Warwick Framework (Dempsey and Weibel 1996). The MODELS 3 workshop concerned 'National resource discovery', and introduced the concept of 'clumps' - groups of metadata resources which can be searched together to facilitate discovery (Dempsey and Russell 1997). The MODELS 4 workshop concerned integrating access to resources across domains (defined as institutions, disciplines or regions) and identified a systems framework that would use a 'layered' approach to cross-domain resource discovery (Russell 1997). Further MODELS-facilitated deliberations have led to the development a logical framework for information management in a distributed environment known as the MODELS Information Architecture (MIA). Attempts have been made to implement an MIA-type broker service in the Agora project.

Agora

Agora is funded under phase III of eLib. The University of East Anglia leads the project, with UKOLN, FDI and the Centre for Research in Library in Information Management (CERLIM) at Manchester Metropolitan University as the other partners. Agora is a 'hybrid-library' project in that it attempts to integrate the technologies developed for new digital services with those used to give access to traditional library collections (Russell 1998). The project builds upon work carried out within MODELS - especially the MIA - and is developing a Hybrid Library Management System (HLMS) that will be an MIA-type broker (Dempsey, Russell and Murray 1999, p. 58). Through this, the project is experimenting with providing integrated access to a variety of services that use different protocols and have different interfaces, including library catalogues, Web index services, information gateways and document supply services.

The Arts and Humanities Data Service gateway

Another example of a MODELS-influenced system that provides integrated access to distributed and heterogeneous resources is a resource discovery system developed for the Arts and Humanities Data Service (AHDS). The AHDS consists of five subject-based service providers that have five distinct resource discovery systems based on a number of different technologies and metadata formats. After extensive consultation into the cross-domain use of the Dublin Core (Miller and Greenstein 1997), AHDS commissioned the production of a gateway that would be able to form a 'virtual union catalogue'. The test implementation provides unified access to the five different service provider catalogues through Dublin Core and a Z39.50 gateway (Beagrie 1999).

Metadata for digital preservation

Preservation strategies and metadata

We began this paper with an acknowledgement that metadata can be used to support a variety of different functions within the digital library context. It is becoming increasingly clear that one of the most important of these functions will be the use of metadata to aid the long-term preservation of digital information.

The successful long-term preservation of digital information will be dependent upon relevant organisations identifying and implementing suitable preservation strategies (Beagrie and Greenstein 1998). Currently, there are three main preservation options: technology preservation, software emulation and data migration. None of these options provide a perfect solution and, as Seamus Ross (1997, p. 331) argues, selecting any one strategy will require trade-offs to be made.

Technology preservation - the preservation of an information object together with all of the software and hardware needed to interpret it - may have an important short-term role for the recovery of data from obsolete storage media and platforms, but is unlikely become a viable long-term strategy. Mary Feeney

(1999, p. 42) points out that collection managers who relied only upon this approach would soon end up with "a museum of ageing and incompatible computer hardware". As a result, most current approaches to digital preservation tend not to be concerned with the preservation of physical artefacts (hardware, media, etc.), but concentrate instead upon the preservation of the information objects themselves in some disembodied digital form (Lynch 1999). Both emulation and migration strategies are examples of this general approach.

Emulation strategies are based on the premise that the best way to preserve the functionality and 'look-and-feel' of digital information objects is to preserve its original software and run this on emulators that can mimic the behaviour of obsolete hardware and operating systems. Emulation strategies would involve encapsulating a data object together with the application software used to create or interpret it and a description of the required hardware environment - i.e., a specification for an emulator. It is suggested that these emulator specification formalisms will require human readable annotations and explanations (metadata). Jeff Rothenberg (1999, p. 27) says that the emulation approach requires "the development of an annotation scheme that can save ... explanations [of how to open an encapsulation] in a form that will remain human-readable, along with metadata which provide the historical, evidential and administrative context for preserving digital documents".

Migration - the periodic migration of digital information from one generation of computer technology to a subsequent one - is currently the most tried-and-tested preservation strategy. However, as Ross (1997, p. 331) points out, data migration inevitably leads to some losses in functionality, accuracy, integrity and usability. In some contexts, this is likely to be important. David Bearman (1994, p. 302), for example, has pointed out that if electronic records are migrated to new software environments, "content, structure and context information must be linked to software functionality that preserves their executable connections". If this, however, cannot be done, he suggests that "representations of their relations must enable humans to reconstruct the relations that pertained in the original software environment". Successful migration strategies will, therefore, depend upon metadata being created to record the migration history of a digital object and to record contextual information so that future users can either reconstruct or - at the very least - begin to understand the technological environment in which a particular digital object was created.

So, regardless of whether emulation-based or migration-based preservation strategies are adopted, the long-term preservation of digital information will involve the creation and maintenance of metadata. Clifford Lynch (1999) describes the function of some of this metadata.

> Within an archive, metadata accompanies and makes reference to each digital object and provides associated descriptive, structural, administrative, rights management, and other kinds of information. This metadata will also be maintained and will be migrated from format to format and standard to standard, independently of the base object it describes.

Preservation metadata has, therefore, become a popular area for research and development in the archive and library communities. Specific examples include:

- The metadata specification for evidence developed as part of the University of Pittsburgh Recordkeeping Functional Requirements Project, funded by the US National Historic Publications and Records Commission (Bearman and Sochats 1996).
- The *Recordkeeping Metadata Standard for Commonwealth Agencies* developed by the National Archives of Australia (1999).
- The final report of the Research Libraries Group (RLG) Working Group on Preservation Issues of Metadata (1998) that defined the semantics of metadata elements that could serve the preservation requirements of digital images.
- The logical data model (based on entity-relationship modelling) developed by the National Library of Australia (NLA) to help identify the particular entities (and their associated metadata) that needed to be supported within its PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) proof-of-concept archive (Cameron and Pearce 1998). The model has been recently revised for use within the NLA's Digital Services Project (National Library of Australia 1999).

An Open Archival Information System (OAIS)

One major recent development has been the production of an ISO Reference Model for an Open Archival Information System (OAIS) - an initiative co-ordinated by the Consultative Committee for Space Data Systems (CCSDS). This defines a high-level reference model for an OAIS, which is defined as an organisation of people and systems "that has accepted the responsibility to preserve information and make it available for a designated community" (CCSDS 1999, p. 1-11). The OAIS model defines a range of functions that are applicable to any archive - whether digital or not. These functions include ingest, archival storage, data management, administration, and access. Amongst other things, the OAIS model aims to provide a common framework that can be used to help understand archival challenges and especially those that relate to digital information.

Accordingly, the OAIS model identifies and distinguishes between the various types of metadata that will need to be recorded by such an archive. An Information Package is seen as encapsulating two types of information - Content Information and the associated Preservation Description Information (PDI) that will allow the understanding of the Content Information over an indefinite period of time (CCSDS 1999, p. 4-25). The PDI contains information that can unambiguously reference the Content Information (e.g. identifiers) and which can also document information about context, provenance and fixity (authenticity). The Content Information itself is divided into a Data Object - which would typically be a sequence of bits - and the Representation Information that gives meaning to this.

Several digital library projects are currently attempting to implement parts of the OAIS model. These include; the European Union-funded NEDLIB (Networked European Deposit Library) project that is developing a deposit system for electronic publications based on OAIS (Werf-Davelaar 1999) and the UK-based Cedars project.

Cedars project

Cedars (CURL Exemplars in Digital Archives) is a three-year project, funded under Phase III of eLib and managed by the Consortium of University Research Libraries (CURL). The lead sites in Cedars are the universities of Cambridge, Leeds and Oxford, with expertise being drawn from both computing services and libraries within the three organisations. The project's aim is to address some of the strategic, methodological and practical issues relating to digital preservation. These issues are being addressed in three main project strands; one looking at digital preservation strategies and techniques (including emulation); another concerned with collection development and rights management issues; and a third interested in the metadata required to adequately preserve digital information objects.

The metadata work within Cedars is being co-ordinated by a working group based at the University of Oxford with some assistance from UKOLN. In l998, this group produced a preliminary review of preservation metadata developments (Day 1998b) and is currently involved in developing a Cedars metadata schema for testing within the project demonstrators. The development of this schema has been informed by the OAIS model and will be broadly structured according to the taxonomy of information object classes that it identifies.

Conclusions

This paper has attempted to give a flavour of some recent metadata developments in the general areas of Internet resource discovery, interoperability and digital preservation. Inevitably, it has glossed over much that could be of interest and completely ignored other interesting issues like rights management or current awareness services. The focus of metadata research is subtly changing as projects gradually transform themselves into services and the importance of the user comes to the fore. This is an area where the library and information communities have much to contribute. It certainly seems to bear out Clifford Lynch's (1997, p. 44) perceptive comment that something very much like traditional library services will be needed to organise, access and preserve networked information if the Internet is to continue to thrive as a means of communication.

References

Baldonado, M., Chang, C.C.K., Gravano, L. and Paepke, A., 1997, The Stanford Digital Library metadata architecture. *International Journal on Digital Libraries*, 1 (2), pp. 108-121.

Beagrie, N., 1999, Convergence and integration online: the Arts and Humanities Data Service gateway and catalogues. Paper delivered at Museums and the Web 1999, New Orleans, La., 11-14 March. http://www.archimuse.com/mw99/papers/beagrie/beagrie.html

Beagrie N. and Greenstein, D., 1998, *A strategic policy framework for creating and preserving digital collections*. London: Arts and Humanities Data Service. http://ahds.ac.uk/manage/framework.htm

Bearman, D., 1994, *Electronic evidence: strategies for managing records in contemporary organizations*. Pittsburgh, Pa.: Archives and Museum Informatics.

Bearman, D. and Sochats, K., 1996, *Metadata requirements for evidence*. Pittsburgh, Pa.: University of Pittsburgh, School of Information Science. http://www.lis.pitt.edu/~nhprc/BACartic.html

Belcher, M., Knight, V. and Place, E., eds., 1999, *DESIRE Information Gateways Handbook*. DESIRE project deliverable. http://www.desire.org/html/subjectgateways/handbook/

Cameron, J. and Pearce, J., 1998, PANDORA at the crossroads: issues and future directions. In: *Sixth DELOS Workshop: Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998*. Le Chesnay: ERCIM, 1998, pp. 23-30. http://www.ercim.org/publication/ws-proceedings/DELOS6/index.html

Consultative Committee for Space Data Systems, 1999, *Reference Model for an Open Archival Information System (OAIS)*, Red Book, Issue 1. CCSDS 650.0-R-1. Latest version available from: http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

Day, M., 1996, *Mapping between metadata formats*. Bath: UKOLN, UK Office for Library and Information Networking. http://www.ukoln.ac.uk/metadata/interoperability/

Day, M., 1998a, *ROADS cataloguing guidelines*. Bath: UKOLN, UK Office for Library and Information Networking. http://www.ukoln.ac.uk/metadata/roads/cataloguing/cataloguing-rules.html

Day, M., 1998b, *Metadata for preservation*. CEDARS project document AIW01. Bath: UKOLN, UK Office for Library and Information Networking. http://www.ukoln.ac.uk/metadata/cedars/AIW01.html

Day, M., 1999, *ROADS interoperability guidelines*. Bath: UKOLN, UK Office for Library and Information Networking. http://www.ukoln.ac.uk/metadata/roads/interoperability-guidelines/

Day, M., Heery, R. and Powell, A., 1999, National bibliographic records in the digital information environment: metadata, links and standards. *Journal of Documentation*, 55 (1), pp. 16-32.

Dempsey L. and Heery, R., 1998, Metadata: a current view of practice and issues. *Journal of Documentation*, 55 (2), pp. 145-172.

Dempsey, L. and Russell, R., 1997, Clumps or ... organised access to printed scholarly material. *Program*, 31(3), pp. 239-249.

Dempsey, L., Russell, R. and Murray, R., 1999, A utopian place of criticism? Brokering access to network information. *Journal of Documentation*, 55 (1), pp. 33-70.

Dempsey L. and Weibel, S.L., 1996, The Warwick Metadata Workshop: a framework for the deployment of resource description. *D-Lib Magazine*, July/August. http://www.dlib.org/dlib/july96/07weibel.html

Dempsey L., Heery R., Hamilton, M., Hiom, D., Knight, J., Koch, T., Peereboom, M., and Powell, A., 1997, *A review of metadata: a survey of current resource description formats*. DESIRE project deliverable. http://www.ukoln.ac.uk/metadata/desire/overview/

Deutsch, P., Schoultz, R., Faltstrom, P. and Weider, C., 1995, *Architecture of the WHOIS++ service*. RFC 1835.
http://www.ietf.org/rfc/rfc1835.txt

Feeney, M., ed., 1999, *Digital culture: maximising the nation's investment*. London: National Preservation Office.

Hakala, J., Hansen, P., Husby, O., Koch, T. and Thorborg, S., 1998, *The Nordic Metadata Project: final report*. Helsinki: Helsinki University Library.
http://linnea.helsinki.fi/meta/nmfinal.htm

Hofman, P., Worsfold, E., Hiom, D., Day, M. and Oehler, A., 1997, *Selection criteria for quality controlled information gateways*. DESIRE project deliverable.
http://www.ukoln.ac.uk/metadata/desire/quality/

Jackson, J. and Gilstrap, D.L., 1999, XML and better Web searching. *Library Hi Tech*, 17 (3), pp. 316-320.

Kirriemuir, J., Brickley, D., Welsh, S., Knight, J. and Hamilton, M., 1998, Cross-searching subject gateways: the query routing and forward knowledge approach. *D-Lib Magazine*, January.
http://www.dlib.org/dlib/january98/01kirriemuir.html

Koch, T., Ardö, A., Brümmer, A. and Lundberg, S., 1996, *The building and maintenance of robot based Internet search services: a review of current indexing and data collection methods*. DESIRE draft project deliverable
http://www.ub2.lu.se/desire/radar/reports/D3.11/

Koch, T., Day, M., Brümmer, A., Hiom, D., Peereboom, M., Poulter, M. and Worsfold, E., 1997, *The role of classification schemes in Internet resource description and discovery*. DESIRE project deliverable.
http://www.ukoln.ac.uk/metadata/desire/classification/

Kunze, J., 1999, *Encoding Dublin Core Metadata in HTML*. IETF Internet-Draft, 15 September.
ftp://ftp.ietf.org/internet-drafts/draft-kunze-dchtml-02.txt

Lassila O. and Swick, R.R., eds., 1999, *Resource Description Framework (RDF) model and syntax specification*. W3C Recommendation, 22 February.
http://www.w3.org/TR/PR-rdf-syntax/

Lundberg, S., Ardö, A., Brümmer, A. and Koch, T., 1996, *The European Web Index: an Internet search service for the European higher education, research and development communities*. DESIRE deliverable D3.1.
http://www.nic.surfnet.nl/surfnet/projects/desire/deliver/WP3/D3-1.html

Lynch, C., 1997, Searching the Internet. *Scientific American*, 276 (3), March, pp. 44-48.
http://www.sciam.com/0397issue/0397lynch.html

Lynch, C., 1999, Canonicalization: a fundamental tool to facilitate preservation and management of digital information *D-Lib Magazine*, 5 (9), September.
http://www.dlib.org/dlib/september99/09lynch.html

Miller, E., 1998, An introduction to the Resource Description Framework. *D-Lib Magazine*, May.
http://www.dlib.org/dlib/may98/miller/05miller.html

Miller, E., Miller, P. and Brickley, D., 1999, *Guidance on expressing the Dublin Core within the Resource Description Framework (RDF)*. Dublin Core Metadata Initiative, Working Draft.
http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/

Miller, P. and Greenstein, D., eds., 1997, *Discovering online resources across the humanities: a practical implementation of the Dublin Core*. Bath: UKOLN, on behalf of the Arts and Humanities Data Service.
http://ahds.ac.uk/public/metadata/discovery.html

National Archives of Australia, 1999, *Recordkeeping metadata standard for commonwealth agencies*, version 1.0. Canberra: National Archives of Australia, May.
http://www.naa.gov.au/govserv/techpub/rkms/intro.htm

National Library of Australia, 1999, *Request for Tender for the provision of a Digital Collection Management System. Attachment 2 - Logical data model*. RFT 99/11. Canberra: National Library of Australia, 23 August.
http://www.nla.gov.au/dsp/rft/index.html

NetLab, 1999, *The Combine harvesting robot*. Lund: Lund University Library, NetLab.
http://www.lub.lu.se/combine/

Paepke, A., Cousins, S.B., Garcia-Molina, H., Hassan, S.W., Ketchpel, S.P., Röscheisen, M. and Winograd, T., 1996, Using distributed objects for digital library interoperability. *IEEE Computer*, 29 (5), pp. 61-68.

Powell, A., 1998, *Using ZEXI to provide Z39.50 access to ROADS servers*. Bath: UKOLN, UK Office for Library and Information Networking.
http://www.ukoln.ac.uk/metadata/roads/interoperability/zexi.html

Powell, A., 1999, *DC-dot Dublin Core generator*. Bath: UKOLN, UK Office for Library and Information Networking.
http://www.ukoln.ac.uk/metadata/dcdot/

RLG Working Group on Preservation Issues of Metadata, 1998, *Final report*. Mountain View, Calif.: Research Libraries Group, May.
http://www.rlg.org/preserv/presmeta.html

ROADS project, 1998, *CrossROADS*. Bath: UKOLN, UK Office for Library and Information Networking.
http://roads.ukoln.ac.uk/crossroads/

Ross, S., 1997, Consensus, communication and collaboration: fostering multidisciplinary co-operation in electronic records. In: *Proceedings of the DLM-Forum on Electronic Records, Brussels, 18-20 December 1996*. INSAR: European Archives News, Supplement II. Luxembourg: Office for Official Publications of the European Communities, pp. 330-336.

Rothenberg, J., 1999, *Avoiding technological quicksand: finding a viable technical foundation for digital preservation*. Washington, D.C.: Council on Library and Information Resources.
http://www.clir.org/pubs/reports/rothenberg/contents.html

Russell, K., 1998, The JISC Electronic Libraries Programme. *Computers and the Humanities*, 32, pp. 353-375.

Russell, R., 1997, UKOLN MODELS 4: evaluation of cross-domain resource discovery. In: Miller, P. and Greenstein. D., eds., *Discovering online resources across the humanities: a practical implementation of the Dublin Core*. Bath: UKOLN on behalf of the Arts and Humanities Data Service, pp. 18-21.
http://ahds.ac.uk/public/metadata/discovery.html

Weber, M.B., 1999, Factors to be considered in the selection and cataloguing of Internet resources. *Library Hi Tech*, 17 (3), pp. 298-303.

Werf-Davelaar, T. van der, 1999, Long-term preservation of electronic publications: the NEDLIB project. *D-Lib Magazine*, 5 (9), September.
http://www.dlib.org/dlib/september99/vanderwerf/09vanderwerf.html

Weibel, S., 1997, *The evolving metadata architecture for the World Wide Web: bringing together the semantics, structure and syntax of resource description*. Paper delivered at ISDL'97, Tsukuba, Japan, 18-21 November.
http://www.dl.ulis.ac.jp/ISDL97/proceedings/weibe.html

Weibel, S., 1999, The state of the Dublin Core, April 1999. *D-Lib Magazine*, 5 (4), April.
http://www.dlib.org/dlib/april99/04weibel.html

Weibel, S., Iannella, R. and Cathro, W., 1997, The 4th Dublin Core Metadata Workshop report. *D-Lib Magazine*, June.
http://www.dlib.org/dlib/june97/metadata/06weibel.html

Weibel, S., Godby, J., Miller, E. and Daniel, R., 1995, *OCLC/NCSA Metadata Workshop report*. Dublin, Ohio: OCLC.
http://purl.org/DC/workshops/dc1conference/report.htm

Weibel, S., Kunze, J., Lagoze, C. and Wolf, M., 1998, *Dublin Core metadata for resource discovery*. RFC 2413.
http://www.ietf.org/rfc/rfc2413.txt

Worsfold, E., 1998, Subject gateways: fulfilling the DESIRE for knowledge. *Computer Networks and ISDN Systems*, 30 (12-18), pp. 1479-1489.
http://www.desire.org/html/research/publications/tnc98gateways/

Projects and initiatives mentioned in the text

Agora project.
http://hosted.ukoln.ac.uk/agora/

Arts and Humanities Data Service Gateway.
http://ahds.ac.uk:8080/ahds_live/

BIBLINK: Linking Publishers and National Bibliographic Services.
http://hosted.ukoln.ac.uk/biblink/

Interoperability Focus.
http://www.ukoln.ac.uk/interop-focus/

Cedars project.
http://www.leeds.ac.uk/cedars/

DESIRE project.
http://www.desire.org/

Dublin Core Metadata Initiative.
http://purl.org/dc

IMesh Toolkit project.
http://www.imesh.org/toolkit/

Internet Detective
http://www.sosig.ac.uk/desire/internet-detective.html

MODELS project.
http://www.ukoln.ac.uk/dlis/models/

Nordic Web Index.
http://nwi.lub.lu.se/?lang=uk

Resource Discovery Network (RDN).
http://www.rdnet.ac.uk/

ROADS project.
http://www.ilrt.bris.ac.uk/roads/

Contact Details

Michael Day
Research Officer
UKOLN: The UK Office for Library and Information Networking
University of Bath
Bath BA2 7AY

Email: m.day@ukoln.ac.uk
URL: http://www.ukoln.ac.uk/

UKOLN is funded by the Library and Information Commission, the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.