The metadata challenge for libraries: a view from Europe

Michael Day[1]

UKOLN: The UK Office for Library and Information Networking,
University of Bath, Bath BA2 7AY, United Kingdom
http://www.ukoln.ac.uk/
m.day@ukoln.ac.uk

Abstract. The effective management of networked digital information - including resource discovery, the management of access based on rights information, long-term preservation, etc. - will increasingly rely on the effective development and use of systems that can collect and use appropriate metadata. This paper will outline an approximate typology of metadata formats and discuss the importance of metadata interoperability from the perspective of selected European metadata initiatives - including the BIBLINK, Cedars, DESIRE, MODELS and ROADS projects

1. Introduction

The effective management of networked digital information - including resource discovery, the management of access based on rights information, long-term preservation, etc. - will increasingly rely on the effective development and use of systems that can collect and use appropriate metadata. This paper will attempt to introduce some issues relating to the metadata challenge for libraries from the perspective of some European metadata initiatives.

2. A typology of formats

The first thing to note is that metadata formats are diverse in their nature and implementation. A *Review of metadata formats*, carried out for the European Union funded DESIRE (Development of a European Service for Information on Research and Education) project, identified and described over twenty formats that were in use

[1] Paper delivered at the conference: Metadiversity - A Call to Action. Responding to the Grand Challenge for Biodiversity Information Management through Metadata, Natural Bridge, VA., USA, 9-12 November 1998. Published in: *Metadiversity: responding to the grand challenge for biodiversity information management through metadata*, ed. Richard T. Kaser and Victoria Cox Kaser (Philadelphia, Penn.: National Federation of Abstracting & Information Services, 1999), pp. 131-140. ISBN: 0-942308-51-4

(or under development) in 1996 and additional formats are in use or under development [1]. Lorcan Dempsey and Rachel Heery point out that many subject communities and market sectors are strongly attached to their own 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. [2]

For these reasons, this 'format diversity' is likely to be perpetuated over time and, indeed, new metadata formats will periodically be developed to address the perceived needs of other subject domains and communities. There are tensions between this continuing drive for specialist formats and any requirement for a level of interoperability that would permit adequate resource discovery across subject domains and information types.

In order to analyse the different metadata formats in existence, the DESIRE study produced a typology of metadata based upon the underlying complexity of the various formats (Figure 1). According to this typology, there is a continuum from simple metadata like that used by Web search engines, through simple structured generic formats like Dublin Core to more complex formats which have structure and are specific to one particular domain or are part of a larger semantic framework. Examples of these more complex formats are the MARC formats used by libraries and formats based on the Standard Generalised Markup Language (SGML).

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 Dublin Core IAFA/Whois++ templates	FGDC MARC	TEI headers ICPSR EAD CIMI

Figure 1. Typology of metadata formats, adapted from Dempsey and Heery (1998)

Band One formats are relatively unstructured and are typically extracted automatically from resources by Web search services. There is no widely-used standard format. Band Two formats tend to have some structure but are simple enough to be created by non-specialist users. These formats do not tend to contain elaborate internal structure and do not easily represent hierarchical objects or complex relationships between objects. Formats in this band include ROADS templates used by some Internet subject services and simple Dublin Core (DC). Band Three formats

contain more descriptive information, both for resource discovery and for the larger task of documenting objects or collections of objects and their relationships. Formats contain more structure than those in Band Two. A variety of formats exist in this category, including the family of MARC formats used by the library cataloguing community and SGML-based initiatives like the Text Encoding Initiative (TEI) header, Encoded Archival Description (EAD) and the format being developed by the Consortium for the Computer Interchange of Museum Information (CIMI).

The DESIRE report concluded that format diversity would remain: "vested interests, competitive advantage, integration with legacy systems or custom and practice will always mean that there are differences of approach".

3. Interoperability

This existing diversity of formats has major implications for interoperability. One response to this problem is the production of metadata crosswalks (or mappings) between one or more formats. At a very basic level, crosswalks can be used as a means of comparing metadata formats and their potential for interoperability. They can also, however, be used as the basis for the production of a specific format conversion program or, potentially, for the production of search systems that would permit the interrogation of heterogeneous metadata formats.

Interoperability, in this context, requires the adoption of core metadata formats, like that proposed by the Dublin Core (DC) initiative, to act as intermediaries for semantic interoperability between heterogeneous resource description models [3, 4]. Stu Weibel suggests that the promotion of a "commonly understood set of core descriptors will improve the prospects for cross-disciplinary search by unifying related attributes" [5]. With specific reference to Dublin Core, Weibel suggests that one approach to interoperability in a heterogeneous resource description environment would be to map many description schemas into a common set (like DC) which would give users "a single semantic model for searching" [6]. Crosswalks from Dublin Core to a variety of other metadata formats, including ROADS templates, USMARC have been produced with interoperability issues in mind [7].

3.1 Heterogeneous resource discovery

3.1.1 The MODELS project

The MODELS (MOving to Distributed Environments for Library Services) project is funded by the Joint Information Systems Committee of the UK higher education funding councils under its Electronic Libraries (eLib) programme. MODELS was motivated by a recognised need to develop an applications framework to manage the rapidly multiplying range of distributed heterogeneous information resources and services being offered to libraries and their users [8]. It was felt that, without such a framework, networked information use would not be as effective as it could be. MODELS essentially provides a forum (primarily in the form of a series of

workshops) for exploring shared concerns, addressing design and implementation issues, initiating concerted actions, and working towards a shared view of preferred systems and architectural solutions. Resource discovery issues have featured widely in MODELS discussions. For example, the MODELS 2 workshop was simultaneously the OCLC/UKOLN Warwick Metadata Workshop (now known as DC-2) that developed the concept of the Warwick Framework [9].

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. At the highest layer, the system could utilise a simple generic metadata format (like Dublin Core) for basic resource discovery. At lower layers of resource discovery, the same system could be configured to use descriptive information from domain-specific metadata formats. Rosemary Russell characterises this as enabling a user, "in a single search environment, to 'drill down' or move progressively through a hierarchy of increasingly rich and specialist metadata as they ... [move] through a continuum from resource discovery to resource evaluation, access, and use" [10].

3.1.2 The Arts and Humanities Data Service resource discovery system

An example of this layered approach is a resource discovery system being developed for the Arts and Humanities Data Service (AHDS) in the UK [11, 12]. The AHDS consists of five subject-based service providers that (amongst their other responsibilities) need to operate within a resource description context specific to their own subject domain. For example, the Oxford Text Archive - the AHDS service provider for literary and linguistic texts - would normally describe resources using Text Encoding Initiative (TEI) headers [13, 14]. The AHDS are implementing a resource discovery system that will provide unified access to the resource description systems of the service providers using Dublin Core and a Z39.50 gateway [15].

3.2 European Dublin Core implementations

3.2.1 The Nordic Metadata Project

Interoperability issues have been to the fore in early European Dublin Core implementations. The Nordic Metadata Project (funded by the Nordic Council for Scientific Information - NORDINFO) has produced a variety of Dublin Core tools including the development of a metadata aware search service [16]. The Nordic Metadata toolkit includes a utility called d2m, a Dublin Core to MARC converter which will convert Dublin Core metadata embedded in HTML into various Nordic MARC formats and USMARC [17].

3.2.2 BIBLINK: linking publishers and national bibliographic services

A different approach to interoperability is embodied in the BIBLINK project, which is funded by the Telematics Applications Programme of the European Commission [18]. This project is concerned with the development of a custom-built software system (the BIBLINK Workspace or BW) that will convert metadata produced by publishers - in the form either of an extended DC known as BIBLINK Core (BC) or a SGML header - into the UNIMARC format [19]. UNIMARC records will then be converted

into the formats (usually MARC) used by the participating national bibliographic agencies who can then enhance them for inclusion in their national bibliography and (potentially) for returning to the publisher.

3.2.3 DC-dot: a Dublin Core generator

Another Dublin Core tool that has been developed is DC-dot [20]. DC-dot is a metadata generator that will retrieve a Web page and automatically generate Dublin Core metadata suitable for embedding in the <META> section of HTML pages. The tags can additionally be edited using the form provided and converted to various other formats (USMARC, SOIF, IAFA/ROADS, TEI headers or RDF), if required.

3.3 Internet subject services and the ROADS project

The ROADS (Resource Organisation and Discovery in Subject-oriented services) project is funded by the JISC under eLib [21]. The project provides software tools and support for the creation of Internet subject services or information gateways. Services that use ROADS use a simple metadata format (ROADS templates) adapted from Internet Anonymous FTP Archive (IAFA) templates but the software itself has been designed to work with interoperability as its primary focus [22].

A basic requirement for ROADS-based services is that they are able to interoperate amongst themselves. In project terms this is referred to as cross-searching. For this, ROADS (version 1) makes use of the Whois++ protocol - a means of making structured information available from physically distributed servers [23]. The ROADS software uses Whois++ to query (and retrieve information from) distributed servers containing structured descriptions (ROADS templates) of Internet resources.

In addition, ROADS (version 2) makes use of the *centroid* facility of Whois++ to facilitate query routing between servers. A ROADS 'index server' will periodically visit selected ROADS subject services 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 [24]. Demonstrations of ROADS cross-searching (*CrossROADS*) using Whois++ and *centroids* have been made available on the Web [25].

ROADS-based services have great freedom with regard to which software tools they choose to implement and they ways in which they can configure their interfaces. Services can even create new ROADS template-types based on their own requirements. In order to help preserve a minimum level of interoperability between ROADS-based services and to help cross-searching, the project has set up a metadata registry - the ROADS Template Registry - to record information about all template-types in use and their associated metadata elements [26]. In addition, the project has developed some generic cataloguing guidelines in an attempt to help ensure that the information content of ROADS templates remain broadly consistent [27].

The ROADS project also has an interest in wider interoperability issues. It is felt that in some situations it would be desirable to make ROADS databases available to end-user clients (and intermediate systems) that use the Z39.50 search and retrieve protocol. To this end the project developed an experimental Z39.50 to Whois++ gateway. The gateway functions as a Z39.50 server, accepting queries from Z39.50 client systems. It then converts them to Whois++ queries and passes them to the ROADS server. As results are returned by the ROADS server, they are converted into a suitable format for use by Z39.50 client systems and returned to the client as a Z39.50 result set. An alternative approach would involve copying records from a ROADS database into another database that has a Z39.50 interface [28].

4. Metadata and digital preservation

The library and information community, and professionals in other disciplines, have other challenges to which metadata solutions can contribute. Publishers and other rights owners, for example, are increasingly giving consideration to using metadata to manage access to digital objects [29]. This is one of the motives for publishers to develop and adopt a Digital Object Identifier (DOI). Similarly, the European Broadcasting Union (EBU) in association with the Society of Motion Picture and Television Engineers (SMPTE) have convened a Task Force to develop harmonised standards on the exchange of television programme material as bit streams - including metadata [30]. However, there is one other challenge for the library and information community that brings together metadata issues related to resource discovery, rights management and administrative issues and places them in a more complex, long-term context. This challenge is digital preservation [31].

There is an increasing awareness that digital preservation will depend upon the creation, capture and storage of all relevant information (metadata) that is required to support a chosen preservation strategy, whether it be technology preservation, emulation or migration. This metadata should include technical data about file formats, software and hardware platforms, etc. but could also record information about authenticity and rights management issues [32]. There are a variety of initiatives that have attempted to identify preservation metadata elements. For example, the recently published report of a working group constituted by the Research Libraries Group (RLG) has specified metadata elements that could serve the preservation requirements of digital images [33].

A UK funded project called Cedars (CURL Exemplars in Digital Archives) is beginning to address some of the strategic, methodological and practical issues relating to digital preservation. Cedars is funded by JISC under its eLib Programme and is managed by the Consortium of University Research Libraries - a group of research libraries in the British Isles whose mission is "to promote, maintain and improve library resources for research in universities". This project has recently produced a preliminary overview of preservation metadata issues that notes the

importance of adopting or creating data models for digital archives, including their metadata systems [34]. The National Library of Australia's PANDORA (Preserving and Accessing Networked DOcumentary Resources of Australia) project has, for example, developed a logical data model based on entity-relationship modelling which forms the basis of identifying the particular entities (and their associated metadata) that need to be supported by the PANDORA system [35]. Cedars is likely to broadly adopt the approach embodied in the Reference Model for an Open Archival Information System (OAIS) being co-ordinated by the Consultative Committee for Space Data Systems (CCSDS).

The OAIS model was originally developed for digital information obtained from observations of terrestrial and space environments but should be applicable to archives. OAIS defines a high-level reference model for an archive, defined as "an organisation of people and systems, that has accepted the responsibility to preserve information and make it available for one or more designated communities" [36]. The OAIS model has a 'taxonomy of archival information object classes' that includes:

Content Information:	This is the information that is the primary object of preservation. This contains the primary Digital Object and Representation Information needed to transform this object into meaningful information.
Preservation Description Information:	This would include any information necessary to adequately preserve the Content Information with which it is associated. It includes: **Reference Information** - (e.g. identifiers), **Context Information** (e.g. subject classifications), **Provenance Information** (e.g. copyright) **Fixity Information** (documenting authentication mechanisms).
Packaging Information:	The information that binds and relates the components of a package into an identifiable entity on a specific media.
Descriptive Information:	The information that allows the creation of **Access Aids** - to help locate, analyse, retrieve or order information from an OAIS.

The Cedars project will not just be adopting (or adapting) a high-level data model like OAIS. It will attempt to develop demonstrators that will implement selected aspects of digital preservation including those related to metadata. The precise nature of the metadata implementation has yet to be decided by the project but the Resource Description Framework (RDF) being developed under the auspices of the World Wide Web Consortium (W3C) is of potential interest - as it is also of interest to other metadata initiatives, including Dublin Core. RDF provides a data model for describing resources and proposes an Extensible Markup Language (XML) based

syntax based on this data model [37]. The need to aggregate multiple sets of metadata was noted at the second Dublin Core workshop and was the principle that underlay the formulation of the Warwick Framework container architecture [38, 39]. Similarly, RDF aims to facilitate modular interoperability among different metadata element sets by creating what Eric Miller calls "an infrastructure that will support the combination of distributed attribute registries" [40]. The modular principle of RDF means that Cedars-defined preservation metadata elements could be aggregated with metadata types defined for other purposes, e.g. Dublin Core for simple resource discovery or structured data about terms and conditions. This type of interoperability is likely to be a useful aspect of preservation metadata systems.

5. References

1. Lorcan Dempsey and Rachel Heery with contributions from Martin Hamilton, Debra Hiom, Jon Knight, Traugott Koch, Marianne Peereboom and Andy Powell, *Specification for resource description methods. Part 1, A review of metadata: a survey of current resource description formats*. DESIRE deliverable D3.2 (1). Bath: UKOLN, March 1997. <URL:http://www.ukoln.ac.uk/metadata/desire/overview/>
2. Lorcan Dempsey and Rachel Heery, 'Metadata: a current view of practice and issues'. *Journal of Documentation*, 54 (2), March 1998, pp. 145-172; here p. 155.
3. Dublin Core metadata: <URL:http://purl.oclc.org/metadata/dublin_core/>
4. S. Weibel, J. Kunze, C. Lagoze and M. Wolf, Dublin Core metadata for resource discovery. RFC 2413. September 1998. <URL:http://info.internet.isi.edu:80/in-notes/rfc/files/rfc2413.txt>
5. Stuart L. Weibel, 'The evolving metadata architecture for the World Wide Web: bringing together the semantics, structure and syntax of resource description'. In: Proceedings of ISDL '97: International Symposium on Research, Development and Practice in Digital Libraries 1997, Tsukuba, Japan, 18-21 November 1997. Tsukuba: University of Library and Information Science, 1997, pp. 16-22; here p. 18. <URL:http://www.dl.ulis.ac.jp/ISDL97/proceedings/weibe.html>
6. Ibid., p. 19.
7. Michael Day, Mapping between metadata formats. Bath: UKOLN, August 1996. <URL:http://www.ukoln.ac.uk/metadata/interoperability/>
8. MODELS project. <URL:http://www.ukoln.ac.uk/dlis/models/>
9. Lorcan Dempsey and Stuart L. Weibel, 'The Warwick Metadata Workshop: a framework for the deployment of resource description'. *D-Lib Magazine*, July/August 1996. <URL:http://www.dlib.org/dlib/july96/07weibel.html>
10. Rosemary Russell, 'UKOLN MODELS 4: evaluation of cross-domain resource discovery'. In: *Discovering online resources across the humanities: a practical implementation of the Dublin Core*, edited by Paul Miller and Daniel Greenstein. Bath: UKOLN on behalf of the Arts and Humanities Data Service, 1997, pp. 18-21; here p. 19. <URL:http://ahds.ac.uk/public/metadata/disc_04.html#ukoln>
11. Arts and Humanities Data Service. <URL:http://www.ahds.ac.uk/>
12. Paul Miller and Daniel Greenstein, 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, October 1997. <URL:http://ahds.ac.uk/public/metadata/discovery.html>

13. Richard Giordano, 'The documentation of electronic texts using Text Encoding Initiative headers: an introduction'. *Library Resources and Technical Services*, 38(4), October 1994, pp. 389-401.
14. Richard Giordano, 'The TEI header and the documentation of electronic texts'. *Computers and the Humanities*, 29 (1), 1995, pp. 75-84.
15. Lorcan Dempsey, Rosemary Russell and Robin Murray, 'The emergence of distributed library services: a European perspective'. *Journal of the American Society for Information Science*, 49 (10), 1998, pp. 942-951; here p. 950.
16. Juha Hakala, Preben Hansen, Ole Husby, Traugott Koch and Susanne Thorborg, The Nordic Metadata Project: final report. Helsinki: Helsinki University Library, July 1998. <URL:http://linnea.helsinki.fi/meta/nmfinal.htm>
17. d2m : Dublin Core to MARC converter: <URL:http://www.bibsys.no/meta/d2m/>
18. BIBLINK: Linking Publishers and National Bibliographic Services: <URL:http://hosted.ukoln.ac.uk/biblink/>
19. Michael Day, Rachel Heery and Andy Powell, 'National bibliographic records in the digital information environment: metadata, links and standards'. *Journal of Documentation*, 55 (1), 1999, pp. 16-32.
20. DC-dot. <URL:http://www.ukoln.ac.uk/metadata/dcdot/>
21. ROADS project. <URL:http://www.ilrt.bris.ac.uk/roads/>
22. Jon P. Knight and Martin Hamilton. Overview of the ROADS software. LUT CS-TR 1010. Loughborough: Loughborough University of Technology, Department of Computer Studies, 1995. <URL:http://www.roads.lut.ac.uk/Reports/arch/arch.html>
23. P. Deutsch, R. Schoultz, P. Faltstrom and C. Weider, Architecture of the WHOIS++ service. RFC 1835. August 1995. <URL:http://info.internet.isi.edu:80/in-notes/rfc/files/rfc1835.txt>
24. John Kirriemuir, Dan Brickley, Susan Welsh, Jon Knight and Martin Hamilton, 'Cross-searching subject gateways: the query routing and forward knowledge approach'. *D-Lib Magazine*, January 1998. <URL:http://www.dlib.org/dlib/january98/01kirriemuir.html>
25. CrossROADS. <URL:http://roads.ukoln.ac.uk/crossroads/>
26. ROADS Template Registry. <URL:http://www.ukoln.ac.uk/metadata/roads/templates/>
27. Michael Day, ROADS Cataloguing Guidelines. Bath: UKOLN, January 1998. <URL:http://www.ukoln.ac.uk/metadata/roads/cataloguing/cataloguing-rules.html>
28. Andy Powell, ROADS and Z39.50: searching ROADS servers using Z39.50 clients. Bath: UKOLN, June 1998. <URL:http://www.ukoln.ac.uk/metadata/roads/interoperability/roads-z3950.html>
29. Godfrey Rust, 'Metadata: the right approach. An integrated model for descriptive and rights metadata in e-commerce'. *D-Lib Magazine*, July/August 1998. <URL:http://www.dlib.org/dlib/july98/rust/07rust.html>
30. David Bradshaw, EBU-SMPTE Task Force on Metadata. Digital workflow through production and documentation: a seminar organised by the Documentation Commission of FIAT/IFTA, BBC Conference Centre, London, 7-8 May 1998.
31. Neil Beagrie and Daniel Greenstein, *A Strategic Policy Framework for Creating and Preserving Digital Collections*. London: Arts and Humanities Data Service, 14 July 1998. <URL:http://ahds.ac.uk/manage/framework.htm>
32. Michael Day, Issues and Approaches to Preservation Metadata. Joint RLG and NPO Preservation Conference: Guidelines for Digital Imaging, University of Warwick, 28-30 September 1998. <URL:http://www.rlg.org/preserv/joint/day.html>
33. RLG Working Group on Preservation Issues of Metadata, *Final report*. Mountain View, Calif.: Research Libraries Group, May 1998. <URL:http://www.rlg.org/preserv/presmeta.html>
34. Michael Day, Metadata for Preservation. CEDARS Project Document AIW01. Bath: UKOLN, 8 August 1998. <URL:http://www.ukoln.ac.uk/metadata/cedars/AIW01.html>

35. National Library of Australia, PANDORA Logical Data Model, Version 2. Canberra: National Library of Australia, 10 November 1997. <URL:http://www.nla.gov.au/pandora/ldmv2.html>
36. Consultative Committee for Space Data Systems, Reference Model for an Open Archival Information System (OAIS), ed. L. Reich and D. Sawyer. CCSDS 650.0-W-4.0. White Book, Issue 4, 17 September 1998. Latest version available from: <URL:http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html>
37. World Wide Web Consortium, Resource Description Framework (RDF) model and syntax specification, eds. Ora Lassila and Ralph R. Swick. W3C Recommendation, 22 February 1999. <URL:http://www.w3.org/TR/REC-rdf-syntax/>
38. Carl Lagoze, Clifford A. Lynch and Ron Daniel, The Warwick Framework: a container architecture for aggregating sets of metadata. Cornell Computer Science Technical Report TR96-1593. Ithaca, N.Y.: Cornell University, 1996. <URL:http://cs-tr.cs.cornell.edu:80/Dienst/UI/1.0/Display/ncstrl.cornell/TR96-1593/>
39. Stuart L. Weibel and Carl Lagoze, 'An element set to support resource discovery: the state of the Dublin Core, January 1997'. *International Journal on Digital Libraries*, 1(2), 1997, pp. 176-186.
40. Eric Miller, 'An introduction to the Resource Description Framework'. *D-Lib Magazine*, May 1998. <URL:http://www.dlib.org/dlib/may98/miller/05miller.html>

6. Acknowledgements

UKOLN is funded by the British Library Research and Innovation Centre (BLRIC), the Joint Information Systems Committee (JISC) of the UK higher education funding councils, as well as by project funding from several sources. UKOLN also receives support from the University of Bath, where it is based. The views expressed in this paper do not necessarily reflect those of UKOLN or its funding bodies.

The author would like to thank Lorcan Dempsey (UKOLN), Kelly Russell (Cedars Project Manager) and Rosemary Russell (UKOLN) for comments on an earlier draft of this paper.