DCC | Digital Curation Manual

Instalment on

"Metadata"

http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata

---

Michael Day

UKOLN

University of Bath, Bath BA2 7AY

http://www.ukoln.ac.uk/

November 2005

Version 1.1

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Metadata
Creator	Michael Day (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the humanities.
Description	Instalment on the role of metadata within the digital curation life-cycle. Describes the increasingly important role of metadata for digital curation, some practical applications for metadata, issues of interoperability between metadata schemes, the topic's place within the OAIS reference model and the issues associated with preservation metadata.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	November 2005 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.2
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© Michael Day, UKOLN, University of Bath

Citation Guidelines

Day M, (November 2005), "Metadata", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from
http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Seamus Ross
Director, HATII, University of Glasgow (UK)

Michael Day
Research Officer, UKOLN, University of Bath (UK)

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Table of Contents

> *Metadata (information about data) provides a means for discovering data objects as well as providing other useful information about the data objects such as experimental parameters, creation conditions, etc.* (Rajasekar & Moore, 2001)

> *In order to exploit and explore the petabytes of scientific data that will arise from ... high-throughput experiments, supercomputer simulations, sensor networks, and satellite surveys, scientists will need assistance from specialized search engines, data mining tools, and data visualization tools that make it easy to ask questions and understand answers. To create such tools, the data will need to be annotated with relevant "metadata" giving information as to provenance, content, conditions, and so on; and, in many instances, the sheer volume of data will dictate that this process be automated.* (Hey & Trefethen, 2005, p.818)

1. Introduction and scope

This instalment will introduce the key topic of metadata and attempt to highlight just why it is considered critically important for the ongoing stewardship and curation of digital data and information.

Metadata can be defined simply as any "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage" any other resource (NISO, 2004). Unfortunately, the term is used in so many different contexts and applied to so many different things that it sometimes seems to convey very little meaning. For example, Duff (2004) has written that data about data can seemingly refer to everything, and concomitantly, nothing. Despite this, it is perhaps worth persisting with the term for now, partly because it remains a useful way of promoting cross-domain communication.

While many metadata initiatives have focused on the development of standards to facilitate the discovery of objects, there has also been a growing awareness of the role that metadata can play in supporting the reuse, management, and long-term preservation. This last has directly led to the development of projects and initiatives focused on the identification of that metadata specifically required to support long-term preservation, perhaps most definitively through the international working group known as PREMIS (http://www.oclc.org/research/projects/pmwg/). Despite this, developing a preservation metadata standard that can be easily implemented has proved difficult. One of the major challenges has been addressing the distinctive metadata requirements of the many different players in the preservation process. For example, referring to the simple taxonomy of users developed by the Digital Curation Centre for its requirements analysis (Carpenter, 2005), it is clear that the metadata needs of data creators may be quite different from those of curators or the re-users of data.

This first *Digital Curation Manual* instalment on metadata will attempt to provide a general introduction to the subject from a digital curation perspective. It will first attempt some definitions and try to explain why metadata is

being seen as increasingly important for supporting reuse and long-term preservation. A section highlighting some of the main uses (or functions) of metadata will be followed by a more detailed introduction to interoperability. Because of its direct relevance to digital curation, the instalment will then consider in slightly more detail the development of preservation metadata standards and the role of packaging formats like the Metadata Encoding and Transmission Standard (METS).

The curation manual will contain further instalments that will consider specific metadata issues and domains in more detail. Those already commissioned or planned (November 2005) include introductions to preservation metadata, interoperability, workflows and the automated extraction of metadata, and reviews of metadata initiatives relevant to learning objects, scientific data and archival records.

2. Definitions

While the term itself has rapidly become ubiquitous, literal definitions of metadata as "data about data" are perhaps now less than helpful. Instead we must try to define metadata in relation to its use, chiefly the functions that it is intended to support.

There have been a number of attempts to categorise these functions. For example, Haynes (2004, pp. 15-17) consolidated older categorisations into a five-point model, covering resource description, information retrieval, management, documenting ownership and authenticity, and interoperability. One of the most popular categorisations was first developed in the 1990s by a digitisation initiative called the Making of America II Testbed Project (Hurley, *et al*., 1999). This defined categories for descriptive, structural, and administrative metadata types, a broad structure that has to a large extent been inherited by the influential Metadata Encoding and Transmission Standard (METS). In this simple typology, *descriptive metadata* is that used for the discovery and identification of objects, *structural metadata* supports the display and navigation of objects, and *administrative metadata* includes any management information needed for the object, including information on the creation process, storage formats, the source and provenance of objects, and the intellectual property rights held in them.

What is missing from this categorisation is any specific acknowledgement of the importance of context. Gilliland-Swetland (1998) has noted that a large part of the activity of archives and museums has traditionally been focused on elucidating and preserving the context of records and artefacts. So, for example, archivists have long been aware that archival records are highly contingent upon what the InterPARES project refers to as their juridical-administrative, procedural, provenancial, documentary and technological contexts (Gilliland-Swetland & Eppard, 2000). The importance of context - and other archival principles like authenticity - is evident in the well-known definition of 'recordkeeping metadata' first developed at a working meeting held in the Netherlands in June 2000 (Wallace, 2001, p 255):

> *Structured or semi-structured information which enables the creation, management and use of records through time and within and across domains in which they are created. Recordkeeping metadata can be used to identify, authenticate, and contextualise records; and the people, processes and systems that create, manage, and maintain and use them.*

This is the understanding of metadata that underpins initiatives like the draft records management metadata standard (ISO/FDIS 23081-1:2005) currently under development by the ISO archives/records management subcommittee (ISO/TC46/SC11).

While the word 'metadata' is a fairly recent invention, the *idea* of metadata is much older, with its roots in library catalogues (and similar)

dating back to the *Pinakes* of Callimachus (a systematic bibliography of Greek literature compiled in the third century BC, probably based on the contents of the Library of Alexandria) and beyond to the record-keeping systems of the ancient near east (Casson, 2001). The term 'metadata' was first used in the context of database management systems to give a generic name for all the various additional data needed to describe and control the management and use of data (Mark & Roussopoulos, 1986). The increasing importance of computer networking had two main effects. On the one hand, it has led to the development of a bewildering array of new metadata standards, each focused on a particular subject domain, content type, function or application. Conversely, this very diversity led to the increased recognition of the importance of metadata in supporting interoperability between systems, both technical and semantic (e.g., Johnston, 2001).

Metadata is now seen as an essential part of the digital world that we live in now, facilitating the discovery, management and reuse of all kinds of digital and non-digital object. Gilliland-Swetland (2004) has observed that metadata "is recognised as a critically important, and yet increasingly problematic and complex concept with relevance for information objects of all types as they move through time and space." As already noted, metadata standards have been developed to support an extremely wide range of activities. These include facilitating the discovery of objects, the management of access and integration, and the documentation of object origins, life cycles and contexts - all at multiple levels of aggregation and focused on particular subject domains. Correspondingly, the world of metadata can look extremely complicated, with multiple domain-specific projects, initiatives and standards. This diversity makes providing generic advice on the use of metadata standards extremely difficult.

3. The growing importance of metadata

The importance of metadata is directly related to the roles they play in supporting the discovery, management and stewardship of digital resources. However, there are a number of general trends that are now making metadata even more crucial to digital curation and stewardship.

The first of these is the vast (and rapidly increasing) amounts of information becoming available in digital form, as reflected in the University of California at Berkeley's periodic analyses of the amount of information being created. These suggest that, even when ignoring the (greater) amount of information that flows through electronic channels (e.g., telephone, radio, television, the Internet), the amount of new information being created and stored on all types of media effectively doubled between 1999 and 2002 (Lyman & Varian, 2003). This 'information explosion' or 'data deluge' is evident in many contexts, e.g. in commerce, public administration and healthcare, but is becoming increasingly important in the research domain.

Scientists and other researchers are becoming increasingly dependent on the production and analysis of vast amounts of data, typically that generated by high-throughput instruments and computer simulations, or streamed from sensors and satellites (Hey & Trefethen). A few examples may suffice. In astronomy it has been suggested that the volume of observational data produced by telescopes and sky surveys doubles each year, with a consequent need to federate access to data across distributed multi-terabyte repositories (Szalay & Gray, 2001). In particle physics, it has been estimated that experiments on the Large Hadron Collider (LHC) currently under construction at CERN will, when operational, generate in the region of 12-14 petabytes of data per year, which will then need to be stored and managed across multiple sites through the LHC Computing Grid (LCG) project (http://lcg.web.cern.ch/lgc/). These examples from 'Big Science' domains may have the most extreme requirements, but related developments in bioinformatics, the environmental sciences and medicine (e.g., neuroinformatics) indicate that many other subject disciplines need to respond to the curation challenges of rapid data growth.

In addition to the growing amount of data being generated, there is an increasing focus in science policy on encouraging open access to data. For example, in January 2004, government ministers from all OECD member states (and some others) endorsed a declaration based on the principle that publicly funded research data should be openly available to the maximum extent possible (Arzberger, *et al*., 2004). An OECD working group is currently working on the development of a set of guidelines that would facilitate open access to digital research data (http://dataaccess.ucsd.edu/). However, as in other contexts, open access does not just depend on the willingness of scientists to share

data, or on the existence of appropriate intellectual property rights regimes (e.g., Waelde & McGinley, 2005), but on the ability of scientists to be able to *find* appropriate data and to be able to *understand* it sufficiently in order to reanalyse it or to integrate it with other data sources. The existence of sufficient good-quality metadata is a prerequisite for the reuse of data. For example, Deelman, *et al*., (2004) comment that it "is impossible to conduct a correct analysis of a data set without knowing how the data was cleaned, calibrated, what parameters were used in the process, etc." It can be argued that this need for metadata is even more important in the data-rich research environments that are characteristic of e-science. Hey and Trefethen (2005, p. 818) argue that metadata will be a necessary condition for the next generation of scientific tools, e.g. giving scientists assistance "from specialized search engines, data mining tools, and data visualization tools that make it easy to ask questions and understand answers." It is, therefore, not surprising that the US National Science Foundation's Blue-Ribbon Advisory Panel on Cyberinfrastructure (2003) argue that the creation and maintenance of metadata is essential for the ongoing stewardship and curation of data.

4. Some uses of metadata

As indicated before, metadata can be used to support a range of functions, from discovery, managing access, to recording sufficient descriptive and contextual information to enable the preservation or reuse of objects over time. Before elaborating a few of these functions in more detail, it is perhaps first worth noting that one of the fundamental characteristics of metadata is that, while it can be made human-readable, it is primarily intended to be processed by machines, e.g. for searching, sorting or display. This basic ability has been supplemented in recent years by the vision of a Semantic Web that facilitates the integration and reuse of data across applications and domains (Berners-Lee, Hendler & Lassila, 2001). We will return to this topic in our discussion of interoperability in section five.

4.1 Resource discovery and retrieval

Historically, a major focus of metadata development has been supporting discovery and retrieval. For example, this has long been one of the primary roles of the metadata held in library catalogues and one of the functions of archival finding aids. A large number of metadata standards have been developed to support resource discovery, although most of these tend to be focused on particular types of object or subject domain. A smaller number of metadata initiatives exist to promote resource discovery across domains. Perhaps the most well known of these is the Dublin Core Metadata Initiative (DCMI), which maintains a fifteen-element core metadata set together with definitions of other metadata terms that can be used to help build interoperability within and across domains (http://dublincore.org/). The element set has been widely implemented, e.g. in cross-domain services like the US National Science Digital Library (Arms & Arms, 2004) but also adapted for use in domain-specific areas like linguistics (Bird & Simons, 2003) or for distributed image collections (e.g., http://www.pictureaustralia.org/). It also underlies a number of metadata standards designed to facilitate access to government information, e.g. the Australian AGLS Metadata Standard (http://www.agls.gov.au/) and the metadata standard defined as part of the UK e-Government Interoperability Framework (e-GIF) (Cabinet Office, Office of the e-Envoy, 2004).

The types of information required for resource discovery tends to differ according to the type of digital object being described. For document-like-objects, there tends to be a strong focus on the types of information traditionally used by library catalogues or abstracting and indexing services, e.g. author and editor names, titles, abstracts, subject headings, etc. The MARC (Machine-Readable Cataloguing) formats traditionally used by libraries have translated well into the metadata world through the provision of things like the MARC21 XML Schema (http://www.loc.gov/standards/marcxml/) and mappings to formats like Dublin Core and

ONIX (http://www.loc.gov/marc/marcdocz.html), also through the creation of simplified formats like the XML-based Metadata Object Description Schema (http://www.loc.gov/standards/mods/). Metadata standards supporting the discovery of images or multimedia tend to include information describing semantic content as well as a range of relevant technical characteristics. Metadata standards for scientific datasets tend to include additional information about the producers of the data, access provisions, and transfer protocols (e.g., Kim, 1999).

4.2 The management of resources

Another area where metadata has a potential important role is in supporting the management of digital resources. This may, for example, record key aspects of the production or curation process (e.g. reasons for selection, preservation actions undertaken) as well as information about intellectual property rights that could be used to manage end-user access. This is the main focus of the 'administrative metadata' section defined by the Metadata Encoding and Transmission Standard (METS) (http://www.loc.gov/standards/mets/). Other types of administrative metadata have been identified by a Digital Library Federation initiative that has produced a data structure that can be used to support the management of dynamic collections of digital resources within library management systems and similar (Jewell, *et al*., 2004). The ONIX metadata standards for books and serials provide publishers with a way to share product information with each other and with suppliers, in part built on a generalised framework developed to facilitate rights metadata transactions in e-commerce contexts (http://www.editeur.org/).

Supporting the long-term management and reuse of digital objects brings us to the realm of digital preservation. Since the mid-1990s, a number of projects and initiatives, mostly originating in the library domain, have attempted to identify the precise role of metadata in supporting digital preservation activities. In recent years, much of the focus of this activity has been on the international working group on Preservation Metadata: Implementation Strategies (PREMIS) (http://www.oclc.org/research/projects/pmwg/), the outcomes of which will be described in more detail in section six (below) and in a separate instalment of this curation manual that will deal specifically with preservation metadata.

4.3 The management of archival records

Also focused on the longer-term is the important work being undertaken by archivists and records managers in identifying the metadata needed to ensure the preservation of the value of archival records as evidence. Research initiatives like the seminal Pittsburgh Project (Functional Requirements for Evidence in Recordkeeping) (Bearman & Duff, 1997; Duff, 2001), both phases of InterPARES (http://www.interpares.org/), and the Australian Recordkeeping Metadata Schema (RKMS)

(McKemmish, *et al*., 1999) have done much to facilitate a better understanding of the role of metadata in the archives and records domain.

In addition, a number of archives have developed specific standards to support the capture (and presentation) of metadata from electronic records management systems (ERMS). For example, the functional requirements for ERMS published by the National Archives in the UK identifies not only the metadata required to support records management functions but also that intended to fulfil external requirements like the e-Government Interoperability Framework, with which it is aligned (National Archives, 2002). Similar standards that specify the metadata that records management software should be able to capture include the influential "Design Criteria Standard for Electronic Records Management Software Applications" issued by the US Department of Defense (DoD 5015.2-STD) and the "Model Requirements for the Management of Electronic Records" (MoReq) (http://www.cornwell.co.uk/moreq.html). This type of records management metadata is primarily designed to support standardisation within organisations, and it is not yet clear yet how much of this metadata will prove useful in supporting long-term preservation. The National Archives metadata standard (and the related e-Government Metadata Standard) contains a specific section for preservation information (e.g. for recording format information), but includes a note that the area is subject to further development. It is also perhaps worth making the point that such metadata will not entirely remove the need for more traditional forms of archival description, which is much better for reflecting the context of a given body of records and their complex interrelationships.

Currently, the archives and records management sub-committee of the International Organization for Standardisation (ISO/TC46/SC11) is working on the development of a standard for records management metadata, building on the metadata requirements identified by the earlier ISO Records Management standard (ISO 15489-1:2001), which defined metadata as "data describing the context, content and structure of records and their management through time." The new standard - ISO 23081 - is made up of three parts. The part now under development is a general outline of the principles of records management metadata, currently a draft standard (ISO/FDIS 23081-1:2005). Further parts will look at implementation issues and provide some methods of assessment. Building on a popular definition first developed at a working meeting in 2000 (Wallace, 2001), the draft standard refines the ISO 15489 clause to define records management metadata as "structured or semi-structured information that enables the creation, registration, classification, access, preservation and disposition of records through time and within and across domains," adding that it "can be used to identify, authenticate and contextualize records and the people, processes and systems that create, manage, maintain and use them and the policies that govern them" (ISO/FDIS 23081-1:2005).

This definition highlights the need for recordkeeping systems to capture metadata about many different types of entity, e.g. the records themselves and their business context, the underlying policies of archives, records management processes and the agents that undertake them.

This diverse metadata would be expensive to create manually, so the viability of records management metadata will depend upon the possibility of automatically capturing the desired information from recordkeeping systems, existing metadata, and other sources. Both the Clever Recordkeeping Metadata Project and InterPARES 2 are investigating the extent to which records management metadata can be captured from business processes and systems and are exploring the potential roles of metadata registries (Evans & Lindberg, 2004; Evans, McKemmish & Bhoday, 2004).

A separate instalment in this curation manual will deal with archival and records management metadata in more detail.

4.4 Facilitating data sharing and reuse

In research domains where data needs to be shared, the creators of data have long recognised the need to maintain contextual and other information about data that allow it to be correctly interpreted or analysed by other researchers. For example, in a paper on ecological metadata, Michener, *et al*. (1997) noted that "highly detailed instructions or documentation may be required for scientists to accurately interpret and analyze historic or long-term data sets, as well as data resulting from unfamiliar research or complicated experimental designs." Helly, Staudigel and Koppers (2003) view this type of documentation as application metadata, "describing the content, context, quality, structure, accessibility and so on of a specific data set." Large-scale data sharing depends to a large extent upon two things. Firstly, it depends upon the existence of some kind of data sharing infrastructure - e.g. databases, repositories or data centres - that can store, curate and provide continued access to data. Secondly, large-scale data sharing requires standardised forms of data and metadata so that users are able to correctly process the retrieved data. Many scientific disciplines and sub-disciplines, therefore, have been involved in developing standards that can facilitate the exchange of data and metadata (Wouters & Reddy, 2003; Ball, Sherlock & Brazma, 2004). These standards tend to be specific to one sub-discipline or type of data.

Metadata sharing is of particular importance in the geosciences, where a number of standardisation initiatives exist (Kim, 1999). Perhaps the most prominent of these is the Content Standard for Digital Geospatial Metadata (CSDGM), developed by the US Federal Geographic Data Committee for the sharing and dissemination of geospatial data (http://www.fgdc.gov/metadata/contstan.html). It is widely used by federal agencies, local government and universities, especially in the United States. Domain-specific profiles of

CSDGM have also been developed for biological data, shoreline data and remote sensing metadata. A technical committee of the International Organization for Standardization (ISO/TC 211) has also developed a metadata standard for describing geographical information and services (ISO 19115:2003).

Another domain where data sharing is important is the social sciences, especially for quantitative data. The Data Documentation Initiative (DDI) is an attempt to develop an international standard for the exchange and preservation of social and behavioural science datasets (http://www.icpsr.umich.edu/DDI/). The standard is currently based on XML and is being used by a growing number of projects and data centres.

The principle of reuse is also a motive behind the development of metadata standards that describe learning objects, most prominently the Learning Object Metadata (LOM) standard developed by the Learning Technology Standards Committee of the IEEE Computer Society (IEEE Std 1484.12.1-2002).

5. Metadata interoperability

A key issue in the networked world is interoperability, the ability of heterogeneous data and metadata to be shared across different systems, e.g. for data aggregation or federated searching. While it is not *exclusively* a technical issue, the main focus has been on the technical and semantic aspects of interoperability (Johnston, 2001). At the technical level, interoperability is dependent on the existence of standard syntaxes, e.g. based on the Extensible Markup Language (XML), and the use of common communication protocols. Popular protocols include the Z39.50 standard (ANSI/NISO Z39.50-2003), typically used for searching distributed collections of bibliographic data like library catalogues, and the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze, *et al*., 2002). Once working, these aspects of interoperability are usually hidden from the user.

Once technical interoperability has been achieved, there is then a need to consider the greater problem of semantic interoperability, e.g. dealing with differences in terminology and meaning across domains. This can be very problematic. In their book *Sorting things out: classification and its consequences*, Bowker and Star (1999, p. 287) remind us that all forms of classification reflect a particular point of view, "that categories are historically situated artifacts and, like all artifacts, are learned as part of membership in communities of practice." Reflecting on differences of meaning, Harvey, *et al*. (1999) argue that true semantic interoperability requires the means "to resolve [the] complex differences that lurk behind apparently consensual terminology and procedures."

The simplest solutions to the semantic interoperability problem involve a combination of metadata transformations based on human-generated mappings (or crosswalks) and the use of cross-domain metadata standards like Dublin Core. The transformation of one metadata schema to another using mappings is a fairly common activity, e.g. when organisations adopt new data formats or systems, but can be far from a straightforward task in practice, with many opportunities for 'mistranslation' (e.g., Woodley, 1998; Godby, Smith & Childress, 2003). The underlying problem, as Duff (2001) reminds us, is that metadata standards are most often developed to address a specific set of needs or requirements and are usually based on quite different conceptual models.

Beyond the Dublin Core, many communities of practice have developed their own standardised formats for facilitating interoperability within particular domains or with relation to particular object types, e.g. the IEEE Standard for Learning Object Metadata (IEEE Std 1484.12.1-2002). For facilitating access to scientific datasets, the Council for the Central Laboratory of the Research Councils (CCLRC) has investigated the development of a generic model for all types of scientific metadata as part of its Data Portal project (Sufi & Matthews, 2004;

Drinkwater & Sufi, 2004).

There are some deeper aspects of semantic interoperability. Heflin & Hendler (2000) comment that in order to achieve it, "systems must be able to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands." This brings us firmly into the domain of ontologies and the Semantic Web. The latter is a vision of a World Wide Web where the meaning of information can be processed by machines. Berners-Lee and Hendler (2001) stress that the concept of machine-processability is not based on artificial intelligence techniques, but "solely on the machine's ability to solve well-defined operations on well-defined data." What this means in practice is that resources are described or annotated with semantic markup (metadata) that means that they can be processed by software agents. Semantic Web technologies like the Resource Description Framework (RDF) and ontology languages have many potential applications, e.g. for the integration of data and information (Hendler, 2003; Staab, 2003; Wroe, *et al*., 2004), and for supporting collaborative and interdisciplinary e-science (e.g., De Roure & Hendler, 2004).

6. The OAIS model and preservation metadata

Preservation metadata and the Reference Model for an Open Archival Information System (OAIS) will be dealt with in more detail in other manual instalments. However, their importance to digital curation means that short introductions to both may be useful here.

Since the mid-1990s, those responsible for the long-term preservation of digital objects have realised that all digital preservation strategies depend - to some extent - upon the capture, creation and maintenance of appropriate metadata (e.g., Day, 2004). This 'preservation metadata' is understood to be all of the *various types of data* that allows the re-creation and interpretation of the structure and content of digital data over time (Ludäsher, Marciano and Moore, 2001). Understood in this way, it is clear that such metadata needs to support an extremely wide range of different functions, including discovery, the technical rendering of objects, the recording of contexts and provenance, to the documentation of repository actions and policies. Conceptually, therefore, preservation metadata spans the popular division of metadata into descriptive, structural and administrative categories. Lynch (1999), for example, has noted that within digital repositories, metadata should accompany and make reference to digital objects, providing associated descriptive, structural, administrative, rights management, and other kinds of information.

The wide range of functions that preservation metadata is expected to support means that the definition (or recommendation) of standards is not a simple task. The situation is complicated further by the knowledge that different kinds of metadata will be required to support different digital preservation strategies and that metadata standards themselves need to evolve over time.

6.1 Types of preservation metadata

The OAIS information model (CCSDS 650.0-B-1, 2002) has been very influential on the development of preservation metadata. This section will briefly outline the general types of metadata that it suggests are necessary to support the preservation of digital objects and note, where possible, work being undertaken in related areas.

The OAIS standard defines an information model for the objects that are managed by an archive. This model built around an entity called an *information package*, which conceptually links into a single entity the object that is the focus of preservation together with all of the additional information types (metadata) necessary to support its continued use. Of the three information packages defined in the OAIS model, the *Archival Information Package* (AIP) may perhaps be understood as the most important for preservation purposes, "defined to provide a concise way of referring to a set of information that has, in principle, all the qualities needed for permanent, or indefinite,

Long Term Preservation of a designated Information Object" (CCSDS 650.0-B-1, 2002, 4-33). The other two information packages defined by the model emphasise that there are likely to be differences between the objects held within an OAIS - the AIP - and those submitted by producers or disseminated to consumers (Lavoie, 2004). However, in the OAIS information model, the AIP is the *key* information package that needs to be preserved.

As with all OAIS information packages, an AIP is a conceptual container of two types of information, called Content Information and Preservation Description Information. Both of these are encapsulated and identified by Packaging Information and discoverable through Descriptive Information, package-level metadata that can be used to create finding aids (Figure 1).

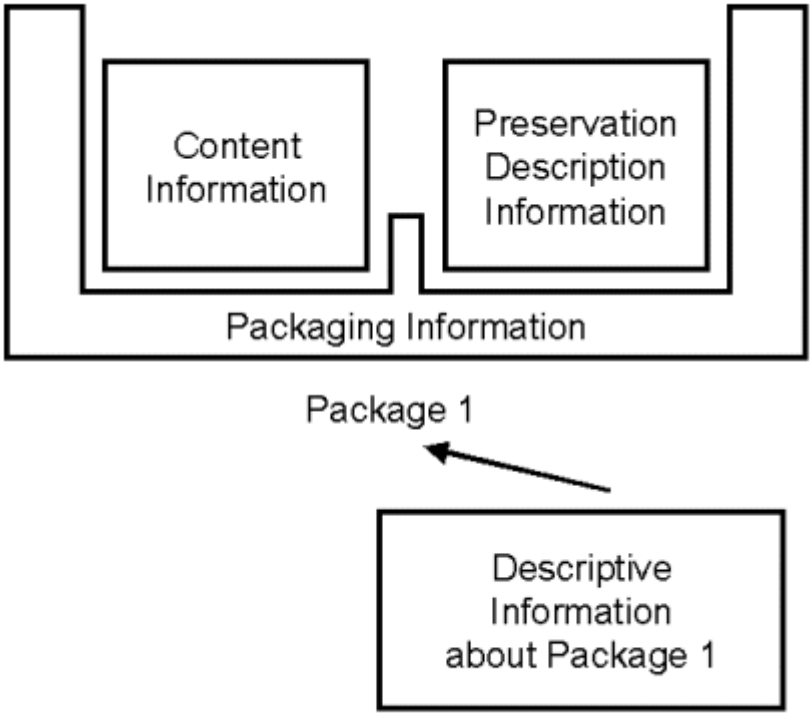

Figure 1. Information Package Concepts and Relationships (from OAIS CCSDS 650.0-B-1, 2002, *Fig. 2-3)*

6.1.1 Technical and structural metadata

Content Information has two components: the Content Data Object, i.e. the object needing preservation (for digital resources this is typically a bit stream), and the associated

Representation Information required to make that object understandable to the users of the OAIS. The OAIS model defines Representation Information as "the information that maps a Data Object into more meaningful concepts" (CCSDS 650.0-B-1, 2002, 1-13), but for digital resources it is essentially the technical information (or metadata) needed to render the bit sequences into something meaningful. Typically, Representation Information might include descriptions of the formats, character sets, etc. in use, possibly with descriptions of hardware and software environments (Structure Information). It might also include any additional information that is required to establish the particular meaning of data content, e.g. that raw numbers should be understood as dates or as temperatures in degrees Celsius (Semantic Information). The OAIS information model understands that Representation Information can be recursive, i.e. that it may itself may need some Reference Information, resulting in what the model defines as a Representation Network. While Representation Information is conceptually part of the Content Information, in practice it could just link to centralised information held elsewhere within the OAIS or in third party registries. A start has been made with developing registries of information about file formats, but similar approaches could be used for other types of Representation Information. The Digital Curation Centre is itself experimenting with the development of a prototype registry of Representation Information (http://dev.dcc.ac.uk/dccrrt/).

6.1.2 Descriptive, administrative and contextual metadata

In addition to the Content Data Object and its Representation Information, the OAIS model suggests that an AIP would also typically include some Preservation Description Information (PDI). This is the type of information that will allow the continued *understanding* of the Content Information over time. The OAIS model document says that PDI is "specifically focused on describing the past and present states of the Content Information, ensuring that it is uniquely identifiable, and ensuring that it has not been unknowingly altered (CCSDS 650.0-B-1, 2002, 4-27). It then defines four classes of PDI, based on categories defined in the seminal 1996 report of a Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group (Garrett & Waters, 1996). This report noted that these four categories, together with the definition of content at different levels of abstraction, were the key features for determining information integrity in the digital environment and argued that they deserved special attention. The following paragraphs will introduce the four categories in more detail.

Fixity - The users of digital resources need to have confidence that they are what they claim to be and that their integrity has not been compromised. Digital information is relatively

easy to manipulate, enabling producers to change or withdraw information released previously (Lynch, 1996). This problem is particularly acute for continuously updated databases, such as those that now play an increasingly important role in scientific research and in commerce. While metadata by itself cannot solve the integrity problem, the OAIS model suggests the inclusion of Fixity Information that can support data integrity checks at the level of Content Data Objects. These might include the use of cryptographic techniques like checksums that can help protect the bit-level integrity by highlighting any changes made to individual data objects.

Reference - Another aspect of the integrity of digital resources identified by the Task Force on Archiving of Digital Information was the need for objects to be identified and located over time. Their report said that for an object "to maintain its integrity, its wholeness and singularity, one must be able to locate it definitively and reliably over time among other objects" (Garrett & Waters, 1996, p. 15). This brings us to the traditional realm of descriptive metadata, e.g. that used in bibliographies, catalogues, and finding aids, but also highlights a key role for persistent identifiers. Identifiers feature highly in the OAIS model's definition of Reference Information, although the practical examples make it clear that other types of descriptive metadata could also be included. There is a separate category in the OAIS information model for descriptive metadata about information packages (Descriptive Information) that can be used to facilitate discovery and access, although it acknowledges that at least some Reference Information will often be replicated in these Package Descriptions (CCSDS 650.0-B-1, 2002, 4-28)

Context - Many resources cannot properly be interpreted without some understanding of their context. Digital objects do not often exist in isolation, but interact with other objects and their wider environment. The context might - in part - be technical, e.g. recording dependencies on particular hardware or software configurations. It might also reflect less tangible realities, e.g., a scientific dataset might be part of a set produced from one experiment, investigation or exploration. In the OAIS model, Context Information is defined as documenting the relationships of the Content Information to its environment (CCSDS 650.0-B-1, 2002, 4-28).

Provenance - Provenance refers to a longstanding principle of the archives profession and embodies the concept that a key part of the integrity of an object is being able to trace its origin and chain of custody. For example, Cook (1993) has written that when archivists adhere to the principles of provenance and original order, "the evidential character of archives is protected, whereby the records inherently reflect the functions, programmes and activities of the person or institution that created them, and the transactional processes by which that actual creation took place." Knowing the provenance or lineage of data is also becoming increasingly important in scientific contexts,

where there is a need to be able to trace the origin and subsequent processing history of datasets to facilitate their reuse (Bose & Frew, 2005). This is especially critical in research disciplines where data can be reprocessed many times by different software applications and services. In bioinformatics, for example, Zhao, *et al*. (2004) have noted the importance of provenance data, understood as the records of where, how and why results were generated, "in order to help e-Scientists to verify results, draw conclusions and test hypotheses." The UK e-Science project myGrid (http://www.mygrid.org.uk) has investigated the development of workflow tools that enable the automatic capture of provenance data, including both information about the organisational context of experiments and their life cycle (Wroe, *et al*, 2004; Stevens, *et al*., 2004). The OAIS model views Provenance Information as a special type of context information that documents the history of the Content Information. This might include information about its creation and provide a record of custody and preservation actions undertaken.

It is perhaps worth noting that the traditional descriptive practices adopted by archivists have been particularly good at providing contextual information. Archival description serves to locate archival records in their relationships to other records (documentary context), to the activities that created them (procedural or business context) and to the entities that created, used, and maintained them over time (provenancial context).

6.2 Preservation metadata initiatives

National and research libraries began to develop preservation metadata standards in the late 1990s with the publication of a number of draft element sets. The National Library of Australia produced the first of these (Phillips, *et al*., 1999), quickly followed by the Cedars and NEDLIB projects (Russell, *et al*., 2000; Lupovici & Masanès, 2000). An international working group sponsored by OCLC Online Computer Library Center and the Research Libraries Group (RLG) then built upon these (and other) proposals to produce a unified *Metadata Framework to Support the Preservation of Digital Objects* (Working Group on Preservation Metadata, 2002). While the earlier initiatives had all been informed by the (then) evolving Reference Model for an Open Archival Information System (OAIS) (CCSDS 650.0-B-1, 2002; ISO 14721:2003), the OCLC/RLG Metadata Framework was *explicitly* structured around its information model.

Following publication of the Metadata Framework, OCLC and RLG commissioned another international group to investigate the issues of implementing preservation metadata in more detail. The resulting Working Group on Preservation Metadata: Implementation Strategies (PREMIS) (http://www.oclc.org/research/projects/pmwg/), co-chaired by Priscilla Caplan and Rebecca Guenther, had the twin objectives of producing a 'core' set of preservation metadata elements and evaluating alternative strategies for

encoding, storing, managing and exchanging such metadata. The group first undertook a survey of the practices of existing and planned preservation repositories. The responses to the survey (PREMIS Working Group, 2004) revealed that most repositories were capturing or planning to capture many different types of metadata. Of individual schemes, the Metadata Encoding & Transmission Standard (METS) was the most popular, with over half of respondents using or planning to use it in some way. The next most popular schemes were the ANSI/NISO Z39.97 standard (Data dictionary -- Technical metadata for digital still images) and OCLC's Digital Archive Metadata Elements. Many repositories were developing custom-built local schemes based on other standards. The working group, however, acknowledged that the relatively small number of respondents (48) meant that it was hard to know exactly how representative the results were.

The working group issued its proposal for core preservation metadata elements in May 2005 with the publication of the PREMIS *Data Dictionary for Preservation Metadata* (PREMIS Working Group, 2005; Lavoie & Gartner, 2005). While this is intended to be a translation of the earlier Metadata Framework into a set of implementable semantic units, the Data Dictionary developed its own data model and is not afraid to diverge from the OAIS model in its use of terminology. The Data Dictionary defines preservation metadata as "the information a repository uses to support the digital preservation process," specifically that "metadata supporting the functions of maintaining viability, renderability, understandability, authenticity, and identity in a preservation context" (p. ix). The Data Dictionary itself defines elements (called semantic units) for describing four of the entities identified by the PREMIS data model: objects (at different levels of aggregation), events, agents, and rights, the latter two in no real detail. The working group also limited the scope of the Data Dictionary by excluding categories of metadata deemed not directly relevant to preservation (e.g. descriptive metadata) or outside the expertise of the group (e.g. technical metadata, information about media and hardware).

6.3 Metadata packaging and METS

The Information Package concept as developed by the OAIS reference model suggests that digital objects should be packaged with both the technical data (Representation Information) needed to convert those bits into meaningful information and all of the other information needed to find, understand and interpret the object (PDI). The model itself does not propose any particular packaging mechanism.

Various models have been proposed for the packaging of data and metadata. For example, the need for some kind of packaging mechanism for different types of metadata and data was realised at the second Dublin Core workshop, held at the University of Warwick in 1996. The outcome of this was the Warwick Framework, a conceptual architecture for the logical

aggregation of multiple types of metadata (or data) in packages called containers (Lagoze, Lynch & Daniel, 1996). This, in turn, influenced the development of the active digital object model that is now a key part of the FEDORA repository architecture (http://www.fedora.info/).

One important recent trend has been the development of the Metadata Encoding & Transmission Standard (METS) (http://www.loc.gov/standards/mets/), a standard maintained by the Library of Congress's Network Development and MARC Standards Office. METS is an attempt to provide an XML Schema for encoding metadata that can support the management and exchange of digital library objects. Essentially, it is an XML-based framework in which different types of metadata can be packaged together. Beedham, *et al*. (2005, p. 70) say that METS "uses XML to provide a vocabulary and syntax for identifying the components that together comprise a digital object, for specifying the location of these components, and for expressing their structural relationships." A METS document consists of seven sections: a METS Header for brief descriptive information about the METS document itself, Descriptive Metadata, Administrative Metadata, a File Section listing all of the files that make up the object, Structural Map and Structural Links sections that enable individual files and metadata to be mapped to the structure of the object, and a Behavior section that provides information on how particular components should be rendered. The administrative metadata section is intended to store technical information about the file, as well as information about intellectual property rights held in the resource, the source material, and provenance metadata that records relationships between files and migrations. The modular design of METS means that objects can also include metadata from 'extension schemas' - i.e. from standards defined elsewhere. For example, the descriptive metadata could include or link to records conforming to standards like the Encoded Archival Description (EAD), the Metadata Object Description Schema (MODS), or Dublin Core. Technical information about still images could be taken from ANSI/NISO Z39.87 or its XML encoding in MIX (http://www.loc.gov/standards/mix/).

METS evolved from an XML Document Type Definition developed for the Making of America II digitisation project (Hurley, *et al*., 1999) and it is perhaps true to say that the standard has been most widely implemented to date in similar contexts (Gartner, 2002). It has been used, for example, in the Oxford Digital Library (http://www.odl.ox.ac.uk/) to provide integrated access to digitised image files with searchable texts. However, there has also been some interest in the potential for METS as a container for preservation metadata. Much of this interest has focused on the potential of METS for object exchange and has been linked with the OAIS concept of Information Packages. For example, Harvard University Library (2001) experimented with METS for defining a Submission Information Package in its Mellon-

funded E-Journal Archiving Project. METS could also be used within a digital repository to package all of the data and metadata required for an Archival Information Package. Despite this interest, however, a recent study of the potential use of METS by the UK Data Archive and The National Archives concluded that the "potential for aggregating the metadata required for different purposes, such as resource discovery, rendering, processing and preservation, into one METS document to act as an OAIS information package, has not been realised sufficiently in practice" (Beedham, *et al.*, 2005, p. 75).

Other potential packaging formats exist. For example, the Los Alamos National Laboratory Digital Library has experimented with the MPEG-21 Digital Item Declaration (DID) specification from ISO/IEC 21000-2:2003 for the packaging of complex digital objects (e.g., Bekaert, Hochstenbach & Van de Sompel, 2003). One part of a standard originally developed for the expression and communication of intellectual property rights information about multimedia objects, the MPEG-21 DID abstract model and its XML syntax (the MPEG-21 Digital Item Declaration Language) has provided the Los Alamos team with a standards-based way of representing compound objects and their associated metadata. Some consideration has also been given to aligning the MPEG-21 DID with the OAIS model. For example, Bekaert, De Kooning & Van de Walle (2005) defined a OAIS-based model for the systematic comparison of object packaging formats and applied this to METS and the MPEG-21 Digital Item Declaration (DID) specification.

Other candidate packaging frameworks might include the Resource Description Framework (RDF) or packaging models developed for use with learning objects, e.g. the Advanced Distributed Learning's Sharable Content Object Reference Model (SCORM) (http://www.adlnet.org/scorm/) or the IMS Content Packaging XML Binding (http://www.imsglobal.org/). These are areas that need more investigation from a digital preservation perspective.

6.4. Some open questions

The last decade has seen an increased awareness of the role of metadata in supporting the preservation and reuse of digital resources and the start of some progress on developing the necessary standards and schemas. The two working groups sponsored by OCLC and RLG have played a major role in this, as has widespread acceptance of the OAIS model. However, there exist a number of questions that suggest that much work remains to be done in developing our understanding of the role of metadata in preservation and curation contexts.

A first question that needs to be asked is whether creating and maintaining metadata on the scale needed to support preservation is either achievable or sustainable. The PREMIS Data Dictionary assumes that repositories will have to capture and maintain information about *at least*

three different entities, the objects needing preservation (which themselves can be complex), the actions undertaken on them, and the people, organisations or software programs controlling these actions. The human generation of metadata is expensive and time consuming, so it will be important for repositories, where possible, to make use of automatic means of capturing this information, be it from objects themselves, already existing metadata, third-party repositories of Representation Information, or from repository processes. Combining all of this into a coherent whole will be a far from trivial task, and it will only ever be possible to check manually a very small proportion of the objects in a repository. Projects like PAWN (Producer - Archive Workflow Network) have now begun to experiment with the automatic capture of metadata as part of repository ingest processes, collecting administrative, preservation and chain of custody (provenance) metadata, and encapsulating it in METS (JaJa, *et al*., 2004; Smorul, *et al*., 2004).

Quality control of metadata will be a potential second problem. The importance of consistent metadata has already been recognised by those trying to develop services that combine data from more than one repository using the OAI-PMH (e.g., Hillmann, Dushay & Phipps, 2004). While there are some ways of supporting the creation of consistent metadata in these contexts (e.g. Guy, Powell & Day, 2004), it is difficult to be certain that these have always been adhered to in practice. Completeness is also likely to be a problem, as some types of metadata will typically not be available for capture. Van Ossenbruggen, Nack and Hardman (2004, p. 39) comment that the editing information for multimedia products is often discarded after production. Also, Vogel (1998) has noted that there are not always sufficient incentives for researchers to fully document their data, although this does vary from discipline to discipline. In preservation contexts, inconsistent, incomplete and misleading metadata are likely to persist for long periods of time.

A related issue is that of the hidden subjectivity and cultural bias of metadata, especially when it will be maintained over long periods of time. Van Ossenbruggen, Nack and Hardman (2004, p. 46) note that contexts of use will most likely be radically different from anything the human creators of metadata might have imagined. In a thought-provoking paper, Bowker (2000, p. 645) has argued that the creators of databases need to historicise data and its organisation "so as to create flexible databases that are as rich ontologically as the social and natural worlds they map." He provides examples from the history of science to show that even relatively fixed things like measurement standards can change over time. The OAIS model tries to solve this problem by saying that an OAIS "must understand the Knowledge Base of its Designated Community to understand the minimum Representation Information that must be maintained," adding that it could also decide to maintain additional Representation Information to enable understanding by a

broader community (CCSDS 650.0-B-1, 2002, 2-4). It remains to be seen whether this is a viable way of solving Bowker's concerns.

A final thing to be considered is the need for metadata itself to be preserved. Metadata is itself digital and will need to be migrated into new forms when necessary, although Rothenberg, et al. (2005, p. 26) note that metadata tend not to be highly application-dependent, meaning that they "are not as vulnerable to loss as more general online information." The OAIS principle of encapsulating Content Object and metadata Information Packages is another possible way of ensuring metadata longevity. In terms of the OAIS model, Preservation Description Information is itself understood to be an Information Object, needing its own associated Representation Information (CCSDS 650.0-B-1, 2002, 2-5). In practice, however, the overhead associated with processing the metadata encapsulated in Information Packages may mean that implementations choose to store metadata in separately managed databases. A related issue is that of the ongoing evolution of metadata standards and the need to modify existing metadata to conform with them.

7. Conclusions

This instalment has attempted to introduce the concept of metadata and indicate its general relevance to digital curation and preservation topics. It has provided some definitions, outlined some of the functions that metadata are intended to support, and introduced in more detail the role that metadata plays in supporting the preservation and reuse of digital objects.

It has been outside the scope of this introductory instalment to cover all relevant topics. For example, it does not include a detailed discussion of metadata designed for specific types of object (e.g. government information, scientific data, learning objects, multimedia) or for functions like rights management. Other instalments in this manual will provide more detailed introductions to some of these topics.

Acknowledgments

I would like to thank Hans Hofman (Nationaal Archief), Terry Eastwood (University of British Columbia), Andy Powell (UKOLN, University of Bath) and my colleagues in the DCC services team for their comments on earlier drafts of this instalment.

References

ANSI/NISO Z39.50-2003. Information retrieval (Z39.50): application service definition and protocol specification. Betheda, Md.: NISO Press.

ANSI/NISO Z39.87 AIIM 20. (2005). Data dictionary -- Technical metadata for digital still images [draft standard]. Retrieved January 30, 2006, from the National Information Standards Organization Web site: http://www.niso.org/standards/resources/Z39-87-200x-forballot.pdf

Arms, C R., & Arms, W. Y. (2004). "Mixed content and mixed metadata: information discovery in a messy world," in: Diane I. Hillmann and Elaine L. Westbrooks, eds., *Metadata in practice*. Chicago, Ill.: American Library Association, 223-237.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhlir, P., & Wouters, P. (2004). "An international framework to promote access to data." *Science*, 303, 1777-1778.

Ball, C. A., Sherlock, G., & Brazma, A. (2004). "Funding high-throughput data sharing." *Nature Biotechnology*, 22(9), 1179-1183.

Bearman, D., & Duff, W. (1997). "Grounding archival description in the functional requirements for evidence." *Archivaria*, 41, 275-303.

Beedham, H., Missen, J., Palmer, M., & Ruusalepp, R. (2005). *Assessment of UKDA and TNA compliance with OAIS and METS standards*. Colchester: UK Data Archive. Retrieved January 30, 2006, from http://www.data-archive.ac.uk/news/publications/oaismets.pdf

Bekaert, J., De Kooning, E., & Van de Walle, R. (2005). "Packaging models for the storage and distribution of complex digital objects in archival information systems: a review of MPEG-21 DID principles." *Multimedia Systems*, 10(4), 286-301.

Bekaert, J., Hochstenbach,P., & Van de Sompel, H. (2003). "Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library." *D-Lib Magazine*, 9(11), November. Retrieved January 30, 2006, from http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

Berners-Lee, T., & Hendler, J. (2001). "Publishing on the Semantic Web." *Nature*, 410, 1023-1024.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). "The Semantic Web." *Scientific American*, 284(5), 28-37.

Bird, S., & Simons, G. (2003). "Extending Dublin Core metadata to support the description and discovery of language resources." *Computers and the Humanities*, 37(4), 375-388.

Bose, R., & Frew, J. (2005). "Lineage retrieval for scientific data processing: a survey." *ACM Computing Surveys*, 37(1), 1-28.

Bowker, G. C. (2000). "Biodiversity

datadiversity." Social Studies of Science, 30(5), 643-683.

Cabinet Office, Office of the e-Envoy. (2004). *e-Government Metadata Standard*, v. 3.0, 29 April. Retrieved January 30, 2006, from http://www.govtalk.gov.uk/schemasstandards/metadata.asp

Carpenter, L. (2004). "Taxonomy of digital curation users." Retrieved January 30, 2006, from the Digital Curation Centre Web site: http://www.dcc.ac.uk/docs/Taxonomy-dc-users.pdf

Casson, L. (2001). *Libraries in the ancient world*, New Haven, Conn.: Yale University Press.

CCSDS 650.0-B-1. (2002). Reference model for an Open Archival Information System (OAIS), Washington, D.C.: Consultative Committee on Space Data Systems. Retrieved January 30, 2006, from http://www.ccsds.org/documents/650x0b1.pdf

Cook, T. (1993). "The concept of the archival fonds in the post-custodial era: theory, problems and solutions." *Archivaria*, 35, 24-27.

Day, M. (2004). "Preservation metadata," In: G. E. Gorman and Daniel G. Dorner, eds., *Metadata applications and management*, International Yearbook of Library and Information Management, 2003-2004, London: Facet Publishing, 253-273.

Deelman, E., Singh, G., Atkinson, M. P., Chervenak, A., Chue Hong, N. P., Kesselman, C., Patil, S., Pearlman, L., & Su, M. -H. (2004). "Grid-based metadata services." 16th International Conference on Scientific and Statistical Database Management (SSDBM04), Santorini Island, Greece, 21-23 June 2004. Retrieved January 30, 2006, from http://www.isi.edu/~annc/papers/deelman_final1.pdf

De Roure, D., & Hendler, J. A. (2004). "E-science: the Grid and the Semantic Web." *IEEE Intelligent Systems*, 19(1), 65-71.

DoD 5015.2-STD. (2002). "Design criteria standard for electronic records software applications." Washington, D.C.: Department of Defense, 19 June. Retrieved January 30, 2006, from http://www.dtic.mil/whs/directives/corres/html/50152std.htm

Drinkwater, G., & Sufi, S. (2004). "CCLRC Data Portal," UK e-Science All Hands Meeting 2004 (AHM2004), Nottingham, UK, 31 August - 3 September 2004. Retrieved January 30, 2006 from http://www.allhands.org.uk/2004/proceedings/papers/161.pdf

Duff, W. M. (2001). "Evaluating metadata on a metalevel." *Archival Science*, 1, 285-294.

Duff, W. (2004). "Metadata in digital preservation: foundations, functions and issues." In: Frank M. Bischoff, Hans Hofman, and Seamus Ross, eds., *Metadata in preservation: selected papers from an ERPANET Seminar at the Archives School Marburg, 3-5 September 2003*. Veröffentlichungen der Archivschule

Marburg, Institut für Archivwissenschaft, 40, 27-38.

Evans, J., & Lindberg, L. (2004). "Describing and analyzing the recordkeeping capabilities of metadata sets." International Conference on Dublin Core and Metadata Applications 2004 (DC2004), Shanghai, China, October 11-14, 2004. Retrieved January 30, 2006 from http://purl.org/metadataresearch/dcconf2004/papers/Paper_27.pdf

Evans, J., McKemmish, S., & Bhoday, K. (2004). "Create once, use many times: the clever use of recordkeeping metadata for multiple archival purposes." 15th International Congress on Archives, Vienna, Austria, 23-29 August 2004. Retrieved January 30, 2006 from http://www.wien2004.ica.org/

Garrett, J., & Waters, D., eds., (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group*, Washington, D.C.: Commission on Preservation and Access, 1996. Retrieved January 30, 2006, from http://www.rlg.org/en/page.php?Page_ID=114

Gartner, R. (2002). *Metadata Encoding and Transmission Standard (METS)*. JISC Techwatch Report TSW 02-05. Retrieved January 30, 2006, from http://www.jisc.ac.uk/index.cfm?name=techwatch_report_0205

Gilliland-Swetland, A. J. (1998). "Setting the stage." In: Murca Baca, *Introduction to metadata: pathways to digital information*. Los Angeles, Calif.: Getty Information Institute. Retrieved January 30, 2006, from http://www.getty.edu/research/conducting_research/standards/intrometadata/

Gilliland-Swetland, A. (2004). "Metadata - where are we going?" In: G. E. Gorman and Daniel G. Dorner, eds., *Metadata applications and management*, International Yearbook of Library and Information Management, 2003-2004, London: Facet Publishing, 16-33.

Gilliland-Swetland, A. J., & Eppard, P. B. "Preserving the authenticity of contingent digital objects: the InterPARES project." D-Lib Magazine, 6(7/8), July/August. Retrieved January 30, 2006, from http://www.dlib.org/dlib/july00/eppard/07eppard.html

Godby, J., Smith, D., & Childress, E. (2003). "Two paths to interoperable metadata." International Conference on Dublin Core and Metadata Applications 2003 (DC2003), Seattle, Wa., United States of America, September 28-October 2, 2003. Retrieved January 30, 2006 from http://purl.oclc.org/dc2003/03godby.pdf

Guy, M., Powell, A., & Day, M. (2004). "Improving the quality of metadata in eprint archives." *Ariadne*, 38. Retrieved January 30, 2006, from http://www.ariadne.ac.uk/issue38/guy/

Harvard University Library. (2001). *Submission Information Package (SIP) specification*, v. 1.0

draft. Retrieved January 30, 2006, from the Digital Library Federation Web site: http://www.diglib.org/preserve/harvardsip10.pdf

Harvey, F., Kuhn, W., Pundt, H., Bishr, Y., & Riedemann, C. (1999). "Semantic interoperability: a central issue for sharing geographic information." *Annals of Regional Science*, 33, 213-232.

Haynes, D. (2004). *Metadata for information management and retrieval*. London: Facet.

Helly, J., Staudigel, H., & Koppers, A. (2003). "Scalable models of data sharing in earth sciences." *Geochemistry, Geophysics, Geosystems*, 4(1), 1010.

Hendler, J. (2003). "Science and the Semantic Web." *Science*, 299, 520-521.

Heflin, J., & Hendler, J. (2000). "Semantic interoperability on the Web." Extreme Markup Languages 2000, Montreal, Canada, August 15-18, 2000. Retrieved January 30, 2006, from http://www.cs.umd.edu/projects/plus/SHOE/pubs/extreme2000.pdf

Hey, T., & Trefethen, A., (2003). "The data deluge: an e-science perspective." In: Fran Berman, Geoffrey Fox and Anthony J. G. Hey, eds., *Grid computing: making the global infrastructure a reality*. Chichester: Wiley, 809-824. Retrieved January 30, 2006, from http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf

Hey, T., & Trefethen, A. E. (2005). "Cyberinfrastructure for e-science." *Science*, 308, 817-821.

Hillmann, D., Dushay, N., & Phipps, J. "Improving metadata quality: augmentation and recombination." International Conference on Dublin Core and Metadata Applications 2004 (DC2004), Shanghai, China, October 11-14, 2004. Retrieved January 30, 2006 from http://purl.org/metadataresearch/dcconf2004/papers/Paper_21.pdf

Hurley, B. J., Price-Wilkin, J., Proffitt, M., & Besser, H. (1999). *The Making of America II Testbed Project: a digital library service model*, Washington, D.C.: Council on Library and Information Resources. Retrieved January 30, 2006, from http://www.clir.org/pubs/abstract/pub87abst.html

IEEE Std 1484.12.1-2002. IEEE Standard for Learning Object Metadata. New York: Institute of Electrical and Electronics Engineers.

ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model, Geneva: International Organization for Standardization.

ISO 15489-1:2001. Information and documentation -- Records management -- Part 1: General, Geneva: International Organization for Standardization.

ISO/IEC 21000-2:2003. Information technology -- Multimedia framework (MPEG-21) -- Part 2: Digital Item Declaration, Geneva: International Organization for Standardization.

ISO/DIS 23081. Information and documentation -- Records management processes -- Metadata

for records -- Part 1: Principles, Geneva: International Organization for Standardization.

JaJa, J., McCall, F., Smorul, M., Moore, R., & Chadduck, R. (2004). "Digital archiving and long term preservation: an early experience with Grid and digital library technologies." Retrieved January 30, 2006, from the National Archives and Records Administration Web site: http://www.archives.gov/era/papers/thic-04.html

Jewell, T. D., Anderson, I., Chandler, A., Farb, S. E., Parker, K., Riggio, A., & Robertson, N. D. M. (2004). *Electronic resource management: report of the DLF Initiative*. Washington, D.C.: Digital Library Federation, August. Retrieved January 30, 2006, from http://www.diglib.org/pubs/dlfermi0408/

Johnston, P. (2001). "Interoperability: supporting effective access to information resources," Library and Information Briefings, 108, London: South Bank University.

Lagoze, C., Lynch, C. A., & Daniel, R. (1996). "The Warwick Framework: a container architecture for aggregating sets of metadata." Cornell University Technical Report TR96-1593, Ithaca, N.Y.: Cornell University Library, 28 June. Retrieved January 30, 2006, from http://techreports.library.cornell.edu:8081/Dienst/UI/1.0/Display/cul.cs/TR96-1593

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). The Open Archives Initiative Protocol for Metadata Harvesting, v. 2.0, 14 June, Retrieved January 30, 2006, from http://www.openarchives.org/OAI/openarchivesprotocol.html

Lavoie, B. F. (2004). *The Open Archival Information System Reference Model: introductory guide*, DPC Technology Watch Report 04-01, Digital Preservation Coalition. Retrieved January 30, 2006, from http://www.dpconline.org/docs/lavoie_OAIS.pdf

Lavoie, B. F., & Gartner, R. (2005). *Preservation metadata*, DPC Technology Watch Report 05-01, Digital Preservation Coalition. Retrieved January 30, 2006, from http://www.dpconline.org/docs/reports/dpctw05-01.pdf

Ludäsher, B., Marciano, R., & Moore, R. (2001). "Preservation of digital data with self-validating, self-instantiating knowledge-based archives," *SIGMOD Record*, 30(3), 54-63. Retrieved January 30, 2006, from http://www.acm.org/sigmod/record/issues/0109/SPECIAL/ludaescher8.pdf

Lupovici, C., & Masanès, J. (2000). *Metadata for the long term preservation of electronic publications*, The Hague: Koninklijke Bibliotheek. Retrieved January 30, 2006, from http://www.kb.nl/coop/nedlib/results/NEDLIBmetadata.pdf

Lyman, P., & Varian, H. R. (2003). "How much information? 2003." Berkeley, Calif.: University of California at Berkeley, School of Information Management and Systems. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/

Lynch, C. (1996) "Integrity issues in electronic publishing," in Robin P. Peek and Gregory B. Newby, eds., *Scholarly publishing: the electronic frontier*. Cambridge, Mass.: MIT Press, 133-145.

Lynch, C. (1999). "Canonicalization: a fundamental tool to facilitate preservation and management of digital information." *D-Lib Magazine*, 5(9), September. Retrieved January 30, 2006, from http://www.dlib.org/dlib/september99/09lynch.html

Mark, L., & Roussopoulos, N. (1986). "Metadata management," *Computer*, 19(12), 26-36.

McKemmish, S., Acland, G., Ward, N., & Reed, B. (1999). "Describing records in context in the continuum: the Australian Recordkeeping Metadata Schema." *Archivaria*, 48, 3-43. Retrieved January 30, 2006, from http://www.sims.monash.edu.au/research/rcrg/publications/archiv01.htm

Michener, W. K., Brunt, J. W., Helly, J. J., Kirchner, T. B., & Stafford, S. G. (1997). "Nongeospatial metadata for the ecological sciences." *Ecological Applications*, 7(1), 330-342.

National Archives. (2002). *Functional requirements for electronic records management systems*, 2002 revision. Kew: The National Archives. Retrieved January 30, 2006, from http://www.nationalarchives.gov.uk/electronicrecords/reqs2002/

National Information Standards Organization. (2004). *Understanding metadata*. Bethesda, Md.: NISO Press. Retrieved January 30, 2006, from http://www.niso.org/standards/resources/UnderstandingMetadata.pdf

National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure,. (2003). *Revolutionizing science and engineering through cyberinfrastructure*. Arlington, Va.: National Science Foundation, Directorate for Computer & Information Science & Engineering (CISE). Retrieved January 30, 2006, from http://www.nsf.gov/cise/sci/reports/toc.jsp

Phillips, M., Woodyard, D., Bradley, K., & Webb, C. (1999). *Preservation metadata for digital collections: exposure draft*, 1999. Retrieved January 30, 2006, from the National Library of Australia Web site: http://www.nla.gov.au/preserve/pmeta.html

PREMIS Working Group. (2004). *Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage community*. Dublin, Ohio: OCLC Online Computer Library Center. Retrieved January 30, 2006, from http://www.oclc.org/research/projects/pmwg/

PREMIS Working Group. (2005). *Data dictionary for preservation metadata*. Dublin, Ohio: OCLC Online Computer Library Center. Retrieved January 30, 2006, from

http://www.oclc.org/research/projects/pmwg/

Rajasekar, A. K., & Moore, R. W. (2001). "Data and metadata collections for scientific applications." 9th International Conference on High-Performance Computing and Networking (HPCN 2001), Amsterdam, Netherlands, 25-27 June 2001, Lecture Notes in Computer Science, 2110, Berlin: Springer, 72-80. Retrieved January 30, 2006, from http://www.sdsc.edu/dice/Pubs/Data-management_moore.pdf

Rothenberg, J., Graafland-Essers, I., Kranenkamp, H., Lierens, A., Oranje, C. van, & Schaik, R. van. (2005). *Designing a national standard for discovery metadata*, RAND Corporation Technical Report TR-185-BZK. Retrieved January 30, 2006, from http://www.rand.org/publications/TR/TR185/

Russell, K., Sergeant, D., Stone, A., Weinberger, E., & Day, M. (2000). *Metadata for digital preservation: the Cedars project outline specification*. Retrieved January 30, 2006, from the University of Leeds Web site: http://www.leeds.ac.uk/cedars/metadata.html

Smorul, M., JaJa, J., Wang, Y., & McCall, F. (2004). "PAWN: Producer - Archive Workflow Network in support of digital preservation" University of Maryland, Institute for Advanced Computer Studies Technical Report UMIACS-TR-2004. Retrieved January 30, 2006, fromhttp://www.umiacs.umd.edu/~joseph/pawn-july2-2004.pdf

Staab, S. (2003). "The Semantic Web: new ways to present and integrate information." *Comparative and Functional Genomics*, 4(1), 98-103.

Stevens, R., McEntire, R., Goble, C., Greenwood, M., Zhao, J., Wipat, A., & Li, P. (2004). "myGrid and the drug discovery process." *Drug Discovery Today: BIOSILICO*, 2(4), 140-148.

Sufi, S., & Matthews, B. (2004). *CCLRC Scientific Metadata Model, version 2*, CCLRC Technical Report, DL-TR-2004-001. Retrieved January 30, 2006, from http://epubs.cclrc.ac.uk/work-details?w=30324

Szalay, A., & Gray, J. (2001). "The world-wide telescope." *Science*, 293, 2037-2040.

Van de Sompel, H., Bekaert, J., Liu, X., Balakireva, L., & Schwander, T. (2005). "aDORe: a modular, standards-based digital object repository." *Computer Journal*, 48(5), 514-535.

Van Ossenbruggen, J, Nack, F., & Hardman, L. (2004). "That obscure object of desire: multimedia metadata on the Web, part 1," *IEEE Multimedia*, 11(4), 38-48.

Vogel, R. L. (1998). "Why scientists haven't been writing metadata." Retrieved January 30, 2006, from the Joint Committee on Antarctic Data Management Web site: http://www.jcadm.scar.org/Articles/why_scientists_dont_write_metadata.htm

Waelde, C. & McGinley, M. (2005). "Public domain; public interest; public funding: focussing on the 'three Ps' in scientific

research." *SCRIPT-ed*, 2(1), March. Retrieved January 30, 2006, from http://www.law.ed.ac.uk/ahrb/script-ed/vol2-1/3ps.asp

Wallace, D. (2001). "Archiving metadata forum: report from the Recordkeeping Metadata Working Meeting, June 2000." *Archival Science*, 1(3), 253-269.

Woodley, M. (1998). "Crosswalks: the path to universal access?" In: Baca, M., ed., *Introduction to metadata: pathways to digital information*, Los Angeles, California: Getty Information Institute. Retrieved January 30, 2006, from http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/

Working Group on Preservation Metadata. (2002). *A metadata framework to support the preservation of digital objects*. Dublin, Ohio: OCLC Online Computer Library Center. Retrieved January 30, 2006, from http://www.oclc.org/research/projects/pmwg/pm_framework.pdf

Wouters, P., & Reddy, C. (2003). "Big science data policies." In: P. Wouters and P. Schröder, eds., *Promise and practice in data sharing*. Amsterdam: Koninklijke Nederlandse Akademie van Wetenschappen, Nederlands Instituut voor Wetenschappelijke Informatiediensten (NIWI-KNAW), 13-40. Retrieved January 30, 2006, from http://www.virtualknowledgestudio.nl/en/

Wroe, C., Goble, C., Greenwood, M., Lord, P., Miles, S., Papay, J., Payne, T., & Moreau, L. (2004). "Automating experiments using semantic data on a bioinformatics grid." *IEEE Intelligent Systems*, 19(1), 48-55.

Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., & Greenwood, M. (2004). "Using Semantic Web technologies for representing e-science provenance." In: Proceedings of the Third International Semantic Web Conference (ISWC2004), Hiroshima, Japan, November 2004, Lecture Notes in Computer Science, 3298, Berlin: Springer-Verlag, 92-106.

Further reading

There is an extremely extensive (and growing) literature on metadata, much of it freely available on the Web.

The best short general introduction to metadata remains the chapter by Gilliland-Swetland in Baca (1998), which should perhaps now be supplemented by the more recent textbook treatment of the topic by Caplan (2003). Of the older introductions, the paper by Dempsey and Heery (1998) and the chapter by Lagoze and Payette in *Moving theory into practice* (Kenney & Rieger, 2000) remain useful. An interesting recent paper by researchers based at OCLC Research looks at the potential role of modular metadata services in the constantly changing contexts of research and learning (Dempsey, *et al*., 2005).

Three edited volumes provide interesting overviews. The chapters in the books edited by Gorman and Dorner (2004) and Hillmann and

Westbrooks (2004) provide up-to-date accounts of metadata developments in different cultural heritage domains. The book edited by Jones, Aronheim and Crawford (2002) is based on papers delivered at an event held in 2000 and, although they are of uneven quality, collectively they do provide a good flavour of metadata thinking in the library domain at the end of the 1990s.

A good collection of papers on the various roles of metadata in digital preservation contexts can be found in the book edited by Bischoff, Hofman and Ross (2004), the outcome of an ERPANET training seminar held in late 2003. On the subject of preservation metadata itself, the OAIS model (CCSDS 650.0-B-1, 2002) is fundamental, although this should be supplemented by a reading of the various reports issued by the two working groups on preservation metadata commissioned by OCLC and the Research Libraries Group (http://www.oclc.org/research/projects/pmwg/). The chapter by Day in Gorman and Dorner (2004) is a review of the state-of-the-art at about the time the second of these groups (PREMIS) started its deliberations. Lavoie and Gartner (2005) provide an overview of PREMIS developments and METS from a preservation perspective.

A good deal of work has been undertaken in the archives and records management domain on the identification of metadata that supports the preservation of the authenticity and integrity of archival records. Unfortunately, there is not a lot of freely available information on ISO 23081 (the draft standard can be purchased from ISO and other national standards bodies) except for a summary provided by the Government of Quebec (http://www.autoroute.gouv.qc.ca/publica/normes/introduction.htm). Papers by McKemmish, *et al*. (1999), Cunningham (2000), and Hedstrom (2001) all provide interesting overviews of particular initiatives in the recordkeeping domain.

Issues around the management of multimedia resources are introduced in two papers published in the *IEEE Multimedia* journal (van Ossenbruggen, Nack & Hardman, 2004; Nack, van Ossenbruggen & Hardman, 2005).

Nilsson, Palmér and Naeve (2002) provide insight into the role of metadata and the Semantic Web in e-learning contexts. A more detailed assessment of the potential role of the Semantic Web in UK higher and further education contexts can be found in Matthews (2005).

Further references

Baca, M., ed. (1998). *Introduction to metadata: pathways to digital information*, Los Angeles, California: Getty Information Institute. Retrieved January 30, 2006, from http://www.getty.edu/research/conducting_research/standards/intrometadata/

Bischoff, F. M., Hofman, H., & Ross, S., eds. (2004). *Metadata in preservation: selected papers from an ERPANET seminar at the Archives School Marburg, 3-5 September 2003*,

Veröffentlichungen der Archivschule Marburg, Institut für Archivwissenschaft, Nr. 40.

Caplan, P. (2003). *Metadata fundamentals for all librarians*, Chicago, Illinois: American Library Association.

CCSDS 650.0-B-1. (2002). Reference model for an Open Archival Information System (OAIS), Washington, D.C.: Consultative Committee on Space Data Systems. Retrieved January 30, 2006, from http://public.ccsds.org/publications/archive/650x0b1.pdf

Cunningham, A. (2000). "Dynamic descriptions: recent developments in standards for archival description and metadata," *Canadian Journal of Information and Library Science*, 25(4), 3-17.

Dempsey, L., & Heery, R. (1998). "Metadata: a current view of practice and issues," *Journal of Documentation*, 54(2), 145-172.

Dempsey, L., Childress, E. R., Godby, C. J., Hickey, T. B., Houghton, A., Vizine-Goetz, D., & Young, J. (2005). "Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape," in Shapiro, D. (ed.), *EScholarship: a LITA guide*. Chicago, Illinois: American Library Association. Preprint Retrieved January 30, 2006, from http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf

Gorman, G. E., & Dorner, D. G., eds. (2004). *Metadata applications and management*, International Yearbook of Library and Information Management, 2003-2004, London: Facet.

Hedstrom, M. (2001). "Recordkeeping metadata: presenting the results of a working meeting." *Archival Science*, 1(3), 243-251.

Hillmann, D. I., & Westbrooks, E. L., eds. (2004). *Metadata in practice*, Chicago, Illinois: American Library Association.

Jones, W., Aronheim, J. R., & Crawford, J., eds. (2002). *Cataloging the Web: metadata, AACR, and MARC21*, Lanham, Maryland: Scarecrow Press.

Kenney, A. R., & Rieger, O. Y., eds. (2000). *Moving theory into practice: digital imaging for libraries and archives*, Mountain View, Calif.: Research Libraries Group.

Lavoie, B. F., & Gartner, R. (2005). *Preservation metadata*, DPC Technology Watch Report 05-01, Digital Preservation Coalition. Retrieved January 30, 2006, from http://www.dpconline.org/docs/reports/dpctw05-01.pdf

Matthews, B. (2005). *Semantic Web technologies*. JISC Technology and Standards Watch Report TSW0502. Retrieved January 30, 2006, from http://www.jisc.ac.uk/index.cfm?name=techwatch_ic_reports2005_published

Nack, F., van Ossenbruggen, J., & Hardman, L. (2004). "That obscure object of desire: multimedia metadata on the Web, part 2," *IEEE Multimedia*, 12(1), 54-63.

Nilsson, M., Palmér. M., & Naeve, A. (2002). "Semantic Web meta-data for e-learning: some architectural guidelines." 11th International World Wide Web Conference, Honolulu, Hawaii, USA, 7-11 May 2002. Retrieved January 30, 2006, from http://kmr.nada.kth.se/papers/SemanticWeb/p744-nilsson.pdf

Van Ossenbruggen, J, Nack, F., & Hardman, L. (2004). "That obscure object of desire: multimedia metadata on the Web, part 1," *IEEE Multimedia*, 11(4), 38-48.

Author information

Michael Day is a Research Officer at UKOLN, based at the University of Bath, United Kingdom (http://www.ukoln.ac.uk/). Since joining UKOLN in 1996, he has worked on a range of metadata-related research projects, which have mostly concerned the development of Internet subject gateways, interoperability, and digital preservation. His most recent completed projects include ePrints UK, concerned with the development of services that give access to the content of multiple institutional repositories, and phase one of eBank UK, an attempt to develop the open access repository paradigm for research data with an initial focus on crystallography. At present, he is mostly working for the UK Digital Curation Centre (http://www.dcc.ac.uk/), a national focus for research and expertise relating to digital preservation issues, and the European Union-funded DELOS Network of Excellence on Digital Libraries (http://www.delos.info/).