METADATA MADE SIMPLER

by Gail Hodge

ACKNOWLEDGEMENTS

NISO extends its thanks to Priscilla Caplan, Janet Gertz, and Rebecca Guenther for reviewing early drafts of this publication. Their insights and contributions were very useful in defining the scope of this booklet.

This booklet is available for free on the NISO website (www.niso.org) and in hardcopy from NISO Press.

To order, contact:
NISO Press Fulfillment
P.O. Box 451
Annapolis Junction, MD
20701-0451
T: 301-362-6904
Fax: 301-206-9789

Published by
NISO Press
National Information Standards Organization
4733 Bethesda Avenue, Suite 300
Bethesda, MD 20814 USA
Email: nisohq@niso.org
T: 301-654-2512
Fax: 301-654-1721
url: www.niso.org

ISBN: 1-880124-50-5

Metadata Made Simpler: A guide for libraries

by Gail Hodge

What Is Metadata?

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. Metadata is often called data about data or information about information.

The term *metadata* is used differently in different communities. Some use it to refer to *machine understandable* information, while others use it only for records that describe electronic resources. However, in the library environment, *metadata* is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Traditional library cataloging is a form of metadata, and MARC 21 and the rulesets used with it such as AACR2 are metadata standards. Other metadata schemes have been developed to describe various types of textual and non-textual objects such as archival materials, visual materials, geographic information, and science and social science datasets.

There are several different types of metadata, including descriptive, administrative, and structural. *Descriptive* metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords. *Administrative* metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. *Rights management* metadata is a form of administrative metadata dealing with intellectual property rights. *Structural* metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.

Metadata can describe resources at any level of aggregation. It can describe a collection, a unitary resource, or a component part of a larger resource (for example, a photograph in an article). Just as catalogers make decisions about whether a catalog record should be created for a whole set of volumes or for each particular volume in the set, so the metadata creator makes similar decisions. Metadata can also be used for description at any level of the information model laid out in the IFLA (International Federation of Library Associations and Institutions) Functional Requirements for Bibliographic Records (http://www.ifla.org/VII/s13/frbr/frbr.pdf): work, expression, manifestation, or item. For example, a metadata record could describe a report, a particular edition of the report, or a specific copy of that edition of the report.

Metadata can be embedded in a digital object or it can be stored separately. Metadata is often embedded in HTML documents and in the headers of image files. Storing metadata with the object it describes ensures the metadata will not be lost, obviates problems of linking between data and metadata, and helps ensure that the metadata and object will be updated together. However, it is impossible to embed metadata in some types of objects (for example, artifacts). Also, storing metadata separately can simplify the management of the metadata itself and facilitate search and retrieval. Therefore metadata is commonly stored in database systems and linked to the objects described.

Metadata *schemes* (also called schema) are sets of metadata elements designed for a particular purpose, for example, to describe a particular type of information resource. The definition or meaning of the elements themselves is known as the *semantics* of the scheme. The values given to metadata elements are the content. Metadata schemes generally specify names of elements and their semantics. Optionally, they may specify *content rules* for how content must be formulated (for example, how to identify the main title) and/or *representation rules* for how content must be represented (for example, capitalization rules). There may also be *syntax rules* for how the elements and their content should be encoded. A metadata scheme with no prescribed syntax rules is called *syntax independent*.

Metadata can be encoded in MARC, in "keyword=value" pairs, or in any other definable syntax. Many current metadata schemes use SGML or XML. XML (Extensible Mark-up Language) is an extended form of HTML which allows for locally defined tag sets and the easy exchange of structured information. SGML (Standard Generalized Mark-up Language) is a superset of both HTML and XML and allows for the richest mark-up of a document.

Metadata is key to ensuring that resources will survive and continue to be accessible into the future.

What Does Metadata Do?

An important reason for creating descriptive metadata is to facilitate discovery of relevant information. In addition to resource discovery, metadata can help organize electronic resources, facilitate interoperability and legacy resource integration, support digital identification, and support archiving and preservation.

Resource Discovery

Metadata serves the same functions in resource discovery as good cataloging does by:

- allowing resources to be found by relevant criteria;
- identifying resources;
- bringing similar resources together;
- distinguishing dissimilar resources;
- giving location information.

Organizing Electronic Resources

As the number of Web-based resources grows exponentially, aggregate sites or portals are increasingly useful in organizing links to resources based on audience or topic. Such lists can be built as static web pages, with the names and locations of the resources "hardcoded" in the HTML. However, it is more efficient and increasingly more common to build these pages dynamically from metadata stored in databases. Software tools such as ColdFusion® can be used to automatically extract and reformat the information for web applications (http://www.allaire.com/Products/coldfusion/).

Another method of organizing Web information is through channels. Channels are preselected Web sites that automatically "push" streams of information to a user's browser, commonly used for continuously updated information such as stock quotes and news. The dominant metadata scheme for webcasting is the Channel Definition Format (CDF) developed by Microsoft and its partners (http://www.w3.org/TR/NOTE-CDFsubmit.html, http://msdn.microsoft.com/workshop/delivery/cdf/reference/CDF.asp).

Interoperability

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. Interoperability is the ability of multiple systems, with different hardware and software platforms, data structures, and interfaces, to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

Two approaches to interoperability are cross-system search and metadata harvesting. The Z39.50 protocol is commonly used for cross-system search (http://www.loc.gov/z3950/agency/). Z39.50 partners do not share metadata but map their own search capabilities to a common set of search attributes. A contrasting approach taken by the Open Archives Initiative (http://www.openarchives.org) is for all partners to translate their native metadata to a common core set of elements and expose this for harvesting. A search service then gathers the metadata into a consistent central index to allow cross-repository searching regardless of the metadata formats used by participating repositories.

Digital Identification

Most metadata schemes include elements such as standard numbers to uniquely identify the work or object to which the metadata refers. The location of a digital object may also be given using a file name, URL, or some more persistent identifier such as a Persistent URL (PURL) or the Digital Object Identifier (DOI). Persistent identifiers are preferred because file locations change frequently, making the URL (and therefore the metadata record) invalid. In addition to the actual elements that point to the object, the metadata can be combined to act as a set of identifying data, differentiating one object from another for validation purposes.

Archiving and Preservation

Most current metadata efforts center around the discovery of recently created resources. However, there is a growing concern that digital resources will not survive in usable form into the future. Digital information is fragile; it can be corrupted or altered, intentionally or unintentionally. It may become unusable as storage media and hardware and software technologies change. Format migration and perhaps emulation of current hardware and software behavior in future hardware and software platforms are strategies for overcoming these challenges.

Metadata is key to ensuring that resources will survive and continue to be accessible into the future. Archiving and preservation require special elements to track the lineage of a digital object (where it came from and how it has changed over time), to detail its physical characteristics, and to document its behavior in order to emulate it on future technologies. Many organizations internationally are working on defining metadata schemes for digital preservation, including the National Library of Australia (http://www.nla.gov.au/padi/topics/32.html), the British Cedars Project (CURL Exemplars in Digital Archives) (http://www.leeds.ac.uk/cedars/metadata.html), and a joint Working Group of OCLC and the Research Libraries Group (RLG) (http://www.oclc.org/digitalpreservation/presmeta_wp.pdf). Many of these initiatives are based on or compatible with the ISO Reference Model for an Open Archival Information System (OAIS) which incorporates preservation metadata along with descriptive, administrative, and rights management metadata (http://www.ccsds.org/RP9905/RP9905.html).

Metadata Element Sets Used in Library Environments

Many different metadata schemes are being used in library environments. A few of the most common ones are mentioned below.

Dublin Core

The Dublin Core Metadata Element Set arose from discussions at a 1995 workshop sponsored by OCLC and the National Center for Supercomputing Applications (NCSA). As the workshop was held in Dublin, Ohio, the element set was named the Dublin Core. The continuing development of the Dublin Core and related specifications is managed by the Dublin Core Metadata Initiative (DCMI) (http://dublincore.org/).

The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. Faced with a proliferation of electronic resources and the inability of the library profession to catalog all these resources, the goal was to define a few elements and some simple rules that could be applied by noncatalogers. The original 13 core elements were later increased to 15: title, subject, description, source, language, relation, coverage, creator, publisher, contributor, rights, date, type, format, and identifier.

The Dublin Core was developed to be simple and concise, and to describe Web-based documents. However, Dublin Core has been used with other types of materials and in applications demanding some complexity. There has historically been some tension between supporters of a "minimalist" view, who emphasize the need to keep the elements to a minimum and the semantics and syntax simple, and supporters of a "structuralist" view who argue for finer semantic distinctions and more extensibility for particular communities.

These discussions have led to a distinction between qualified and unqualified (or simple) Dublin Core. Qualifiers can be used to refine (narrow the scope of) an element, or to identify the encoding scheme used in representing an element value. The element "Date", for example, can be used with the refinement qualifier "created" to narrow the meaning of the element to the date the object was created. "Date" can also be used with an encoding scheme qualifier to identify the format in which the date is recorded, for example, following the ISO 8601 standard for representing date and time.

All Dublin Core elements are optional and all are repeatable (see sidebar). The elements may be presented in any order. While the Dublin Core description recommends the use of controlled values for fields where they are appropriate (i.e., controlled vocabularies for the Subject field), this is not required. However, working groups have been established to discuss authoritative lists for certain elements such as Resource Type. While Dublin Core leaves content rules to the particular implementation, the DCMI encourages the adoption of *application profiles* (domain-specific rules) for particular domains such as education and government. An application profile for libraries is being developed by the Libraries Working Group.

Thanks in part to its simplicity, the Dublin Core is now used by many outside the library community — researchers, museum curators, and music collectors to name only a few — because it does not require knowledge of highly specialized descriptive systems like AACR2. There are hundreds of projects worldwide that use the Dublin Core either for cataloging or to collect data from the Internet; more than 50 of these have links on the DCMI website. The subjects range from cultural heritage and art to math and physics. Meanwhile the Dublin Core Metadata Initiative has expanded beyond simply maintaining the Dublin Core Metadata Element Set into an organization that describes itself as "dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems."

Global (Government) Information Locator Service (GILS)

GILS is a U.S. Federal Information Processing Standard (FIPS Pub192) supported by various guidelines and memoranda from the Office of Management and Budget (http://www.dtic.mil/gils/documents/naradoc/fip192.html). GILS grew out of the U.S. government requirement for public access to government information, and it is authorized by the Paperwork Reduction Act of 1995. Originally called the "Government Information Locator Service", GILS in various forms has been adopted by other governments and for international projects, leading to its current designation, "Global Information Locator Service".

GILS itself does not formally define metadata elements, rules for

Dublin Core Elements for this Report

Title: Metadata Made Simpler

Creator: Hodge, Gail

Subject: metadata

Description: Describes metadata standards and projects for librarians.

Publisher: National Information Standards Organization (NISO)

Date: 20010601

Type: Text.Report

Format: text/html

Identifier: http://www.niso.org/metadatasimple/

Language: en

representation, and syntax. Rather, GILS specifies a profile of the Z39.50 protocol for search and retrieval, specifying which attributes must be supported. The Core GILS elements for the U.S. Federal GILS have been defined by the National Archives and Records Administration (http://www.dtic.mil/gils/documents/naradoc/).

The original goal of GILS was to provide high-level locator records for government resources, both electronic and nonelectronic. GILS records were intended to describe aggregates such as catalogs, publishing services and databases. The emphasis is on availability and distribution rather than on description. Therefore, a GILS record may have data elements such as the name and address of the distributor and the order process (see sidebar). However, some organizations use GILS at the individual item (journal article or technical report) level. Since GILS was an early metadata scheme, evaluations of its implementation and use are available and very valuable in developing other metadata systems.

A GILS Core Record for this Report

Title: Metadata Made Simpler

Originator: Gail Hodge

Local Subject Term: Metadata

Abstract: Describes metadata standards and projects for librarians.

Purpose: To serve as an educational aid to librarians.

Availability:

Distributor :

Name: NISO Press Fulfillment

Street Address: P.O. Box 451

City: Annapolis Junction

State: MD

Country: USA

Zip Code: 20701-0451

Telephone: 301-362-6904

Fax: 301-206-9784

Order Process: Available free on the NISO website or in hardcopy from NISO Press Fulfillment for $20

The Text Encoding Initiative (TEI) Header

The Text Encoding Initiative (http://www.tei-c.org/) is an international project to develop guidelines for marking up electronic texts such as novels, plays, and poetry, primarily to support research in the humanities. This SGML markup becomes part of the electronic resource itself. In addition to specifying how to encode the text of a work, the TEI Guidelines also specify a header portion, embedded in the resource, that consists of metadata about the work. The TEI header, like the rest of the TEI, is defined as an SGML DTD, a set of tags and rules defined in SGML syntax that describe the structure and elements of a type of document. Since the TEI DTD is rather large and complicated in order to apply to a vast range of texts and uses, a simpler subset of the DTD, known as "TEI Lite", is commonly used in libraries.

It is assumed that TEI-encoded texts are electronic versions of printed texts. Therefore the TEI Header can be used to record bibliographic information about both the electronic version of the text and about the non-electronic source version. The basic bibliographic information is not dissimilar to that recorded in library cataloging and can be mapped to and from MARC. However, there are also elements defined to record details about how the text was transcribed and edited, how markup was performed, what revisions were made, and other non-bibliographic facts. Libraries tend to use TEI headers when they have collections of SGML-encoded full text. Some libraries use TEI headers to derive MARC records for their catalog systems, while others use MARC records for the published source texts as the basis for creating TEI header descriptions.

The Encoded Archival Description (EAD)

In archives and special collections, the finding aid is an important tool for resource description. Finding aids differ from catalog records by being much longer, more narrative and explanatory, and highly structured in a hierarchical fashion. They generally start with a description of the collection as a whole, indicating what types of materials it contains and why they are important. If the collection consists of the personal papers of an individual there can be a lengthy biography of that person. The finding aid describes the series into which the collection is organized, such as correspondence, business records, personal papers, and campaign speeches, and ends with an itemization of the contents of the physical boxes and folders comprising the collection.

The Encoded Archival Description (EAD) was developed as a way of marking up the data contained in a finding aid, so that finding aids can be searched and displayed online. The EAD standard is maintained jointly by the Library of Congress and the Society of American Archivists (see http://www.loc.gov/ead/). Like the TEI Header, the EAD is defined as an SGML DTD. It begins with a header section that describes the finding aid itself (for example, who wrote it) which could be considered metadata about the metadata; it then goes on to the description of the collection as a whole and successively more detailed information. If individual items being described exist in digital form, the EAD can include pointers to the digital objects.

The EAD is particularly popular in academic libraries with large special collections. Although it is easier to put finding aids on the Web by simply marking them up in HTML than to follow the EAD specification, libraries and archives investing in EAD creation hope that using this metadata scheme will encourage consistency in encoding and give them some measure of search interoperability.

The Visual Resources Association (VRA) Core Categories

The VRA Core Categories is a metadata element set developed to describe visual materials such as buildings, photographs, paintings and sculptures (http://www.gsd.harvard.edu/~staffaw3/vra/vracore3.htm). Typically, visual resources collections used in teaching art history and similar subjects do not contain original art works but rather slides or photographs of the original art. Metadata for these materials therefore has to accommodate the description of multiple levels of related resources; for example, an original painting, a slide of the painting, a digitized image of the slide. Version 3.0 of the VRA Core Categories consists of 17 metadata elements which can be used as applicable to describe each of these versions and relate them to each other: record type, type, title, measurements, material, technique, creator, date, location, ID number, style/period, culture, subject, relation, description, source, and rights. Like the Dublin Core, the VRA Core scheme does not specify any particular syntax or rules for representing content. However the development of the VRA Core has served as impetus for the visual resources community to continue to develop shared vocabularies and classification schemes. Curators of visual resources collections hope that use of the VRA Core Categories will allow them to share descriptions of original works as well as to describe images held in their own collections.

ONIX International

ONIX (Online Information Exchange) International is an XML-based metadata scheme being developed by publishers under the auspices of a number of book industry trade groups in the United States and Europe (http://www.editeur.org/onix.html). The original ONIX specification was a direct response to the enormous growth in online book sales and the realization that books described with images, cover blurbs, reviews, and similar information significantly outsold books without this information. Therefore ONIX has elements to record a wide range of evaluative and promotional information as well as basic bibliographic and trade data. Although initially focused on the communication of book trade information to booksellers and distributors, ONIX is being expanded to accommodate other publication types and media, including journals and journal articles, conference proceedings, and electronic books.

Although libraries are not currently creating ONIX format data directly, ONIX may play a role in the processing stream for Cataloging in Publication (CIP) and in the creation of "provisional" or order-level bibliographic records. It is likely that additional library uses of ONIX for books and for serials will be found as ONIX becomes more pervasive within the publishing community. Mappings between ONIX and both USMARC and UNIMARC exist and are available from the ONIX website.

Metadata for Datasets

Metadata schemes for datasets are particularly significant for libraries that specialize in subjects where numeric and statistical data are of great importance.

One of the most well developed element sets is the Federal Geographic Data Committee's (FGDC) Content Standard for Digital Geospatial Metadata (CSDGM), officially known as FGDC-STD-001-1998 (http://www.fgdc.gov/metadata/contstan.html). Geospatial datasets include topographic and demographic data, GIS (geographic information systems), and computer-aided cartography base files. They are used in a wide variety of areas, including soil and land use studies, biodiversity counts, climatology and global change tracking, remote sensing and satellite imagery. The FGDC Content Standard is required for use with resources created and funded by the U.S. government and many state governments, and is also being used internationally.

A metadata scheme becoming well established in the social and behavioral sciences is the Data Documentation Initiative (DDI) standard for describing social science datasets (http://www.icpsr.umich.edu/DDI/codebook.html). The DDI is defined as an XML DTD, and allows for top down hierarchical description of a social science study, the data files resulting from

Metadata in Action

A county land planner is studying the impact of new zoning laws on a particular bird species. The study team is composed of an ecologist, hydrologist, civil engineer and environmental protection specialist. Remote sensing data for the last 20 years provides a trend analysis of the decrease in wetlands, the bird's habitat. These datasets have FGDC metadata. The biologists on the study team need to document the results of a field inventory. Using a biological profile to extend the FGDC element set, the biologists add the genus-species name and taxonomic hierarchy. The ecologists are concerned with collection methods and modeling tools. The data related to the changes in human population are documented using a metadata set developed by the Census Bureau. This study results in a technical report which is assigned Dublin Core metadata by the author. When the technical report is cataloged into the organization's repository, the Dublin Core elements are used as the basis for automatic generation of a MARC cataloging record. This record is enhanced by the cataloger and included in the library's online public access catalog.

that study, and the variables used in the data files. There is also a header area that uses Dublin Core elements for a high-level description of the DDI document itself.

A Dublin Core description represented in RDF

```
<?xml version="1.0"?>

<!DOCTYPE rdf:RDF SYSTEM "http://purl.org/
      dc/schemas/dcmes-xml-20000714.dtd">

<rdf:RDF xmlns:rdf="http://www.w3.org/1999/
      02/22-rdf-syntax-ns#"

   xmlns:dc="http://purl.org/dc/elements/1.1/
      ">

   <rdf:Description about="http://
      www.niso.org/metadatasimple/">

      <dc:title>Metadata Made Simple</
         dc:title>

      <dc:creator>Hodge, Gail</dc:creator>

       <dc:subject>metadata</dc:subject>

      <dc:description>Explains metadata
         standards and concepts for
         libraries.</dc:description>

      <dc:publisher>National Information
         Standards Organization (NISO)
         </dc:publisher>

      <dc:date>2001-06-01</dc:date>

      <dc:format>text/html</dc.format>

   </rdf:Description>

</rdf:RDF>
```

Using Metadata

In real-world applications, the formal metadata specification is just the starting point. Guidelines for applying the specification to a particular set of resources are invariably required, as are mechanisms for interoperating with other collections of metadata, and tools to aid in metadata creation.

Extensions and Profiles

Despite the recent development of many of these metadata schemes, most have already been subject to the changes brought about by implementing them in real world situations. These modifications are of two types: extensions and profiles.

An *extension* is the addition of elements to an already developed scheme to support the description of an information resource of a particular type or subject or to meet the needs of a particular interest group. Extensions increase the number of elements. *Profiles* are subsets of a scheme that are implemented by a particular interest group; profiles can constrain the number of elements that will be used, refine element definitions to describe the specific types of resources more accurately and specify values that an element can take.

In practice, many applications use both extensions and profiles of base metadata schemes. For example, the National Biological Information Infrastructure (NBII), with support from the Biological Resources Division of the U.S. Geological Survey, has developed a biological profile to the FGDC Content Standard for use with biological information resources (http://www.nbii.gov/datainfo/metadata/standards/index.html). The effort began by extending the elements to add important elements for the description of biological resources, such as the taxonomic (scientific or Latin) name of the organism and its classification in the taxonomic hierarchy. After the additional elements were agreed to, the group recommended a specific subset or profile that would be most useful to biologists.

The U.S. Department of Education's Gateway to Educational Materials (GEM) project has similarly based their own metadata scheme on the Dublin Core (http://www.geminfo.org/Workbench/Metadata/index.html). The GEM profile limits elements to be used (for example, Contributor is not allowed) and makes some elements mandatory. GEM also defines additional elements such as Audience, Grade, Quality and Standards, extending the base Dublin Core set for educational use.

Frameworks for Interoperability and Exchange

Many people ask: Do we need so many metadata standards and initiatives? Can't one standard serve the purpose? Are extensions and profiles really necessary?

It is important to remember that different schemes serve distinct needs and audiences (sidebar). Complementary schemes can be used to describe the same resource for multiple purposes serving a number of user groups. For example, a technical report could have a MARC metadata set in a library's online catalog, an FGDC description as part of the National Spatial Data Infrastructure Clearinghouse Mechanism, and an embedded set of Dublin Core elements. Practical aspects of this complex environment are being investigated by several groups, some of which are mentioned below.

The SCHEMAS project of the UK Office for Library and Information Networking (UKOLN) is a forum for the implementers of metadata schemes that aims to provide information about new and emerging metadata standards and to promote "good-practice guidelines for adapting multiple standards or metadata modules for local use in customized

schemas" (http://www.ukoln.ac.uk/metadata/schemas/).

The Resource Description Framework (RDF), developed by the World Wide Web Consortium (W3C), is a data model for the description of resources on the Web that also provides a mechanism for integrating multiple metadata schemes (http://www.w3.org/RDF/). In RDF a namespace is defined by a URL pointing to a Web resource that describes the metadata scheme that is used in the description. Multiple namespaces can be defined, allowing elements from different schemes to be combined in a single resource description. Multiple descriptions, created at different times for different purposes, can also be linked to each other. RDF is generally expressed in XML (sidebar).

Another project related to the interoperability of metadata is OCLC's Cooperative Online Resource Catalog (CORC) project (http://www.oclc.org/corc/). Beginning in January 1999, participating libraries have been using CORC to build a database of shared cataloging for Web resources. The CORC system supports both MARC and Dublin Core and can map between them to display, import and export records in either format. Some metadata elements can be extracted automatically by the system from the Web page being cataloged. Key to the success of the CORC project is the CORC Cataloging Guidelines and the availability of authority files for certain fields.

The Interoperability of Data in E-Commerce Systems (<indecs>) Framework (http://www.indecs.org) was an international collaborative effort supported by the European Commission's info2OOO Programme. The collaborators were major rights owners, such as publishers and members of the recording industry, who wanted to develop a framework for metadata standards to support network commerce in intellectual property. The foundation of the <indecs> work is a data model for intellectual property and its transfer. Rather than developing a new metadata scheme, <indecs> sought to develop a common framework to allow various schemes for transactions related to different genres such as music, journal articles, and books to be able to interchange information, particularly that related to intellectual property rights. In order to support this common framework, <indecs> has developed a minimal kernel of required metadata. Several pilot projects are under way to validate the metadata kernel. The framework and other metadata are being implemented in several major projects involving books and audiovisual materials.

Metadata Crosswalks

The interoperability and exchange of metadata is facilitated by *metadata crosswalks*. A crosswalk is a mapping of the elements, semantics and syntax from one metadata scheme to those of another.

A crosswalk allows metadata created by one community to be used by another group that employs a different metadata standard. The degree to which these crosswalks are successful at the individual record level depends on the similarity of the two schemes, the granularity of the elements in the target scheme compared to that of the source, and the compatibility of the content rules used to fill the elements of each scheme.

Crosswalks are important for virtual libraries where resources are being collected from a variety of sources and are expected to act as a whole, perhaps with a single search engine applied. While these crosswalks are key, they are also labor intensive to develop and maintain. The mapping of schemes with fewer elements (less granularity) to those with more elements (more granularity) is problematic.

Table 1 shows a crosswalk between Dublin Core and both MARC 21 and GILS for the Title and Author elements. In this case, there is no attempt to map at the content level. The first element name and the

Table 1. Example of Metadata Crosswalk Mapping

Dublin Core	GILS	USMARC
Title: The name given to the resource by the CREATOR or PUBLISHER.	Title	245$a (Title Statement/Title proper) (1st indicator=0) If repeated, all titles after the first: 246$a (Varying Form of Title/Title proper)
Author or Creator: The person(s) or organization(s) primarily responsible for the intellectual content of the resource. For example, authors in the case of written documents; artists, photographers, or illustrators in the case of visual resources. Qualifier possible: type.	Originator	$100a (Main Entry-Personal Name) 720$a (Added Entry–Uncontrolled Name/Name) (with $e=author) If type=personal: 700$a (Added Entry-Personal Name) If type=corporate: 710$a (Added Entry-Corporate Name)

definition are from Dublin Core. Mappings are then made to MARC 21 and GILS.

Metadata in Action

An oral historian makes tape-recordings of interviews with members of a particular ethnic group. Interviewees sign a paper release form giving intellectual property rights to the historian. Most interviewees grant permission to disseminate the interviews in print and electronically, but several restrict publication and dissemination until 25 years after death.

Information about each interview is kept in a database: interviewer, interviewee, date, place, etc. Each interview follows a questionnaire format. The questionnaire exists as a text file. The tapes, release forms, database, and text file are donated to a library that has a special collection focusing on the particular ethnic group.

The tapes are digitized. Since each interview runs over several tapes, technicians record structural metadata to keep component parts of each interview together. Technicians record administrative metadata such as file names, location of each interview in the files, equipment used, the methods of digitizing and assuring quality and completeness, file formats, etc. Different segments of this metadata allow the audio files to be automatically tracked, accessed, stored, refreshed, and migrated.

An archivist expands the database to include the persistent identifier of each interview, thereby linking the audio file to the descriptive metadata. The names of the data elements are revised to match Dublin Core terminology, including qualifiers used specifically for audio materials. Information on rights and permissions is entered.

An archivist creates an EAD finding aid for the audio collection using the database as the core. Portions of the questionnaire text file are incorporated as a rich source of subject keywords. A MARC record is derived from the EAD finding aid and added to OCLC and RLIN.

A webpage is created where researchers can access the finding aid, search the database, and listen to the audio files. Interviews coded as restricted are invisible to the search program until the date when they become open to the public. Administrative, structural, and descriptive metadata is created for the webpage to hold all the pieces together, allow them to be managed, and allow them to be accessed.

The library participates in a metadata harvesting protocol to provide extracts of local metadata in a common format to a service provider so that information about the collection is automatically included in a number of relevant tools such as catalogs and portals.

The webpage is linked to the library's website dedicated to resources about the ethnic group, where it is available to researchers in context with archival and visual materials, digitized secondary sources, etc. Administrative, structural, and descriptive metadata for this website also has been created to hold all of its pieces together, allow them to be managed, and allow them to be accessed.

Metadata Registries

Registries are an important tool for managing metadata. Metadata registries can provide information on the definition, origin, source, and location of data. Registration can apply at many levels, including schemes, usage profiles, metadata elements, and code lists for element values. Registries can document the meaning and use of the elements in a single metadata scheme as they change over time, or the way the same elements have been used in different applications.

Registries can also document element meanings in multiple schemes or databases, particularly within a specific field of interest such as health care, aeronautics, or environmental science. A good example is the U.S. Environmental Protection Agency, Environmental Data Registry (http://www.epa.gov/edr/) that provides information about thousands of data elements used in current and legacy EPA databases. The metadata registry provides an integrating resource for legacy data, acts as a look-up tool for designers of new databases, and documents each data element.

Standards relevant to metadata registries include ISO/IEC 11179 Specification and Standardization of Data Elements (a joint standard of the International Organization for Standardisation and the International Electrotechnical Commission) and the ANSI X3.285, Metamodel for the Management of Shareable Data.

Metadata Creation

Who creates metadata? The answer to this varies by discipline, the resource being described, the tools available, and the expected outcome, but it is almost always a cooperative effort.

Much basic structural and administrative metadata is supplied by the technical staff who initially digitize or otherwise create the digital object. For descriptive metadata also, in some situations it is best if the originator of the resource provides the information. This is particularly true in the documentation of scientific datasets where the originator has significant understanding of the rationale for the dataset and the uses to which it could be put, and for which there is little if any textual information from which a library cataloger could work. However, many projects have found that it is more efficient to have library catalogers or other information professionals create the descriptive metadata, because the authors or creators of the data do not have the time or the skills. In other cases, a combination of researcher and information professional is used. The researcher may create a skeleton, completing the elements that can be supplied most readily. Then results may be supplemented or reviewed by the cataloger for consistency.

Two major projects providing metadata tools and services for the Dublin Core are the Nordic Web Project and MetaWeb in Australia. The Nordic Web provides metadata creation software and Dublin Core to MARC conversion software which is free within the European Union. MetaWeb has developed a metadata editor called "Reggie."

There are a number of FGDC-compliant metadata creation tools, including Metamaker which was developed by the U.S. Geological

Survey, Biological Resource Division, to create FGDC-compliant metadata. Some FGDC-compliant products have been developed by geographic information system (GIS) vendors to support the documentation of geospatially referenced datasets stored within their products. While many of these are proprietary, there are efforts under way through the Open GIS Consortium to support an open metadata tool.

Extract tools which analyze the resource and automatically create a metadata record are also available. The Nordic Web Project has developed software to extract metadata from a selected Web site and create preliminary Dublin Core records. CORC also creates an initial metadata record by extracting key information from the resource itself. It then builds a "pathfinder" which brings together links to resources.

The creation of metadata both automatically and by people such as researchers who are not familiar with intellectual control raises several key issues. While attempts have been made for originators to provide metadata for their resources, in some cases the quality is less than desirable due to inconsistency, omission of important elements, or lack of controlled vocabulary. This can be solved by a review cycle by information professionals. However, these additional procedures increase the cost of the metadata creation. For any given project, there must be a balance between optimal quality and the resources available.

There are two keys to solving some of these challenges. The first is adequate training and awareness among metadata creators and data originators. Originators should be encouraged to use metadata creation tools and to think about entering consistent information. The other solution comes from metadata tool developers. Both commercial and proprietary software tools are addressing the need for quality. The tools are beginning to support improved validation rules, pick lists that limit the selection in a particular field, and the use of authority files and controlled vocabularies. Software may support templating and other customization to streamline data entry.

Metadata and the Standards Process

Many of the metadata schemes described here were developed by consensus within specific communities. Some of these are now seeking acknowledgment from national and international standards bodies such as NISO.

Metadata work is ongoing across a number of technical committees (TC) of the International Organization for Standardization (ISO). In ISO TC46 (NISO's counterpart at the international level), Subcommittee 4 Working Group 7 is addressing metadata development for bibliographic applications. The ISO-IECJTC1 (Information Technology) SC 32/WG2 is studying standards needs for the specification and management of metadata. This includes metadata elements, classification and coding schemes, and metadata management and exchange mechanisms. ISO TC211 (Geographic information/Geomatics) is studying metadata for applications in geographic information systems.

Next Steps

The World Wide Web is presenting a multitude of new challenges and opportunities for applying the analytic and organizational skills that are the hallmark of the librarian and information scientist. The development and application of metadata is one. Best practices and patterns are now emerging. The following resources will give you a head start in tracking developments.

More Information on Metadata

General Resources

Candy Schwartz's Metadata Resource List
http://web.simmons.edu/~schwartz/mymeta.html

International Federation of Library Associations (IFLA)
http://www.ifla.org/II/metadata.htm

"Metadata: Cataloging by Any Other Name," by Jessica Milstead and Susan Feldman
http://www.onlineinc.com/onlinemag/metadata/

Metadata Information Clearinghouse Interactive (MICI)
http://domino.wileynpt.com/NPT_Pilot/Metadata/mici.nsf

Metadata Schema Registry (Australia)
http://metadata.net/

UK Online Library Network (UKOLN) Site
http://www.ukoln.ac.uk/metadata/resources

Schemes, Initiatives, and Related Sites

CDF (Channel Definition Format)
http://msdn.microsoft.com/workshop/delivery/cdf/reference/channels.asp

Cedars (CURL exemplars in digital archives)
http://www.leeds.ac.uk/cedars/metadata.html

CORC (Cooperative Online Resource Catalog)
http://www.oclc.org/oclc/research/projects/corc/

DCMI (Dublin Core Metadata Initiative)
see Dublin Core

DDI (Data Documentation Initiative)
http://www.icpsr.umich.edu/DDI/

DOI (Digital Object Identifier)
http://www.doi.org/

Dublin Core
http://dublincore.org

EAD (Encoded Archival Description)
http://lcweb.loc.gov/ead/

EPA (Environmental Protection Agency) Metadata Registry
http://www.lbl.gov/~olken/epaintro.html

FGDC Content Standard for Digital Geospatial Metadata (CSDGM)
http://www.fgdc.gov/metadata/contstan.html

Functional Requirements for Bibliographic Records (IFLA)
http://www.ifla.org/VII/s13/frbr/frbr.pdf

GCMD (Global Change Master Directory)
http://gcmd.nasa.gov/difguide/difman.html

GEM (Gateway to Educational Materials)
http://www.geminfo.org/

GILS (Global Information Locator Service)
http://www.usgs.gov/gils/index.html

IDF (International DOI Foundation)
see DOI (Digital Object Identifier)

IFLA Functional Requirements for Bibliographic Records
see Functional Requirements for Bibliographic Records (IFLA)

<indecs> interoperability of data in e-commerce systems
http://www.indecs.org/

MARC (Machine-Readable Cataloging)
http://lcweb.loc.gov/marc

MCF (Meta Content Framework)
http://www.textuality.com/mcf/NOTE-MCF-XML.html

MetaWeb Project
http://www.dstc.edu.au/RDU/MetaWeb

NBII (National Biological Information Infrastructure)
http://www.nbii.gov/

NSDI (National Spatial Data Infrastructure)
http://www.fgdc.gov/nsdi/nsdi.html

Nordic Metadata Projects
http://linnea.helsinki.fi/meta/

OAI (Open Archives Initiative)
http://www.openarchives.org/

OAIS (Open Archival Information System)
http://www.ccsds.org/RP9905/RP9905.html

ONIX (Online Information Exchange)
http://www.editeur.org/

PADI (Preserving Access to Digital Information)
http://www.nla.gov.au/padi/topics/32.html

PURL (Persistent URL)
http://purl.org

RDF (Resource Definition Framework)
http://www.w3.org/RDF/

TEI (Text Encoding Initiative)
http://www.tei-c.org/

W3C (World Wide Web Consortium)
http://www.w3.org/

VRA (Visual Resources Association) Core Categories
http://www.gsd.harvard.edu/~staffaw3/vra/vracore3.htm

XML (Extensible Markup Language)
http://www.w3.org/XML/

Z39.50
http://www.loc.gov/z3950/agency/

Crosswalks and Lists of Crosswalks

Dublin Core to MARC and GILS
http://www.loc.gov/marc/dccross.html

FGDC to MARC
http://www.alexandria.ucsb.edu/public-documents/metadata/fgdc2marc.html

MARC 21 to Dublin Core
http://www.loc.gov.marc/marc2dc.html

Metadata: Mapping between Metadata Formats (UKOLN)
http://www.ukoln.ac.uk/metadata/interoperability/

Tools for Metadata Creation

BlueAngel Technologies
http://www.blueangeltech.com/

ColdFusion®
http://www.allaire.com/Products/coldfusion/

Dublin Core tools
http://dublincore.org/tools/

ESRI ArcInfo
http://www.esri.com/software/arcinfo/index.html

Metadata Software Tools
http://ukoln.bath.ac.uk/metadata/software-tools/

MetaPackager™
http://www.hisoftware.com/metapackager.htm

U.S. Army Corp of Engineers-Corpsmet95
http://www.nysl.nysed.gov/gis/training/tools.htm

Glossary

administrative metadata - metadata such as owner and accession date, provided to help manage a resource catalog - a searchable set or collection of separate metadata records.

channel - a preselected Web site that automatically "pushes" streams of information to a user's browser; for example, news or stock quotes.

crosswalk - a mapping of the elements, semantics, and syntax from one metadata scheme to another.

descriptive metadata - metadata that describes a work for purposes of discovery and identification, such as creator, title, and subject.

DOI - Digital Object Identifier, an identifier used by publishers which can be resolved to a location for a digital object.

DTD - document type definition, a way of defining an SGML or XML structure to be used for a specific kind of document, for example, a journal article or a finding aid.

element set - information segments of the metadata record, often called semantics or content.

encoding rules - the syntax or prescribed order for the elements contained in the metadata description.

extension - an element that is not officially part of a metadata scheme, which is defined for use with that scheme for a particular application.

interoperability - the ability of multiple systems, using different hardware and software platforms, data structures, and interfaces, to exchange and share data.

metadata - structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource.

metadata harvesting - a technique for extracting metadata from individual repositories and collecting it in a central catalog to facilitate search interoperability.

namespace - in RDF, a way to tie a specific use of a metadata element to the scheme where the intended definition is to be found.

profile - a subset of a scheme defined and used by a particular interest group to customize the scheme for its purposes.

PURL - Persistent URL, an identifier used by a name resolution system developed by OCLC to provide a means of indirection for URLs.

RDF - Resource Description Framework; a data model developed by the World Wide Web Consortium for the description of resources on the Web.

scheme - a metadata element set and rules for using it.

semantics - the names and meanings of metadata elements.

SGML - Standard Generalized Mark-up Language, a way of tagging elements and values used as a syntax for certain metadata schemes.

syntax - rules for how metadata elements and their content are encoded.

structural metadata - metadata that indicates how compound objects are structured, provided to support use of the objects.

XML - Extensible Mark-up Language, a subset of SGML gaining currency in Web applications.

Z39.50 - a NISO and ISO standard protocol for cross-system search and retrieval. Officially, International Standard, ISO 23950: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification" and ANSI/NISO Z39.50-1995.