# A Framework of Guidance for Building Good Digital Collections

2nd Edition (2004)

Links updated (2006-03-01)

NISO Framework Advisory Group

NISO Press, Bethesda, Maryland, U.S.A.

**Published by**
**NISO Press**
**4733 Bethesda Avenue, Suite 300**
**Bethesda, MD 20814**
**www.niso.org**



ISBN: 1880124-64-5

**Bibliographic citation for this document:**

NISO Framework Advisory Group. *A Framework of Guidance for Building Good Digital Collections*. 2nd edition. Bethesda, MD: National Information Standards Organization, 2004. Available from: http://www.niso.org/framework/framework2.pdf

## CONTENTS



## CASE STUDIES



## TABLES

## FOREWORD

This revision was produced by the NISO Framework Advisory Group:

- Grace Agnew, Rutgers University
- Liz Bishoff, OCLC, Inc.
- Priscilla Caplan (chair), Florida Center for Library Automation
- Rebecca Guenther, Library of Congress
- Ingrid Hsieh-Yee, Catholic University
- Assistants: Amy Alderfer, a graduate student at Catholic University, and Jen Childree, a student at Sante Fe Community College

With many thanks to Joan K. Lippincott and Peter Hirtle for their review and advice, and to our colleagues at the Illinois Digital Archives Project, the Don Hunter Archive Project and the New Jersey Digital Highway Project for sharing their experiences through case studies.

The first (2001) edition of this document was produced by members of the IMLS Digital Library Forum:

- Liz Bishoff, Colorado Digitization Alliance
- Priscilla Caplan (chair), Florida Center for Library Automation
- Tim Cole, University of Illinois Urbana-Champaign
- Anne Craig, Illinois State Library
- Daniel Greenstein, Digital Library Federation
- Doug Holland, Missouri Botanical Garden
- Ellen Kabat-Lensch, Eastern Iowa Community College
- Tom Moritz, American Museum of Natural History
- John Saylor, Cornell University.

A note on the sponsorship of this *Framework*:

The first edition of this document was funded by the Institute of Museum and Library Services (IMLS). IMLS supported the development of the *Framework* to encourage institutions to plan their digitization practices strategically in order to develop collections that will be accessible and useful for the long-term, and that can integrate with other digital collections to support a growing network of broadly accessible digital information resources. IMLS encourages use of this *Framework* to assist collection developers in planning and implementing good digital practices. However, use of the *Framework* is not linked to IMLS funding, which is determined by separate proposal, evaluation, and selection procedures that are documented in depth on the IMLS website. This second edition is now supported by the National Information Standards Organization (NISO), an organization that supports the development and maintenance of information standards and guidelines for the support of good practices in digital information development and sharing.

## INTRODUCTION

This *Framework* has two purposes. First, to provide an overview of some of the major components and activities involved in the creation of good digital collections. Second, to provide a framework for identifying, organizing, and applying existing knowledge and resources to support the development of sound local practices for creating and managing good digital collections. It is intended for two audiences: cultural heritage organizations planning projects to create digital collections, and funding organizations that want to encourage the development of good digital collections.

The use of the word ***good*** in this context requires some explanation. In the early days of digitization, projects were often justified as test beds—vehicles for the development of methods and technologies, experiments in technical or organizational innovation, or simply learning experiences to help the organization determine what role, if any, it would play in the emerging digital information space. As a test bed, a collection could be considered good if it provided proof of concept, even if the collection was of minimal usefulness to the organization's users or if it disappeared altogether at the end of the project period.

As the digital environment matured, standards and practices to support collection developers emerged from the test bed phase. The focus of collection building shifted toward the goal of creating relevant and responsive collections that served the needs of one or more communities of users. The bar of goodness was raised to include levels of usability, accessibility, and fitness for use appropriate to the anticipated user group(s).

Digital collection development has now moved to a third stage, where even serving information effectively to a known constituency is not sufficient. As the digital environment itself has matured to become a critical and often primary vehicle for delivering information to the vast majority of people, integration and trust have emerged as critical criteria for digital collection building. Web standards and technologies now support the integration of vast amounts of disparate information and users increasingly demand "one-stop shopping" for their information needs. Concomitantly, the vast amount of information available makes it increasingly difficult for users to find trusted information—information that is reliably available for the long-term and is known to be authentic. Objects, metadata, and collections must now be viewed not only within the context of the projects that created them but as building blocks that others can reuse, repackage, and build services upon. Indicators of goodness must now emphasize factors contributing to interoperability, reusability, persistence, verification, documentation, and support for intellectual property rights.

The *Framework* is organized around indicators of goodness for four core entities:

- **Collections** (organized groups of objects)
- **Objects** (digital materials)
- **Metadata** (information related to objects)
- **Projects** (initiatives to create or manage collections)

Note that services have been deliberately excluded as out of scope. It is expected that if quality collections, objects, and metadata are created, it will be possible for any number of higher level services to make effective use of these entities.

For each of these four entities, general principles related to quality are defined and discussed, and supporting resources providing further information are identified. These resources may be standards, guidelines, best practices, explanations, discussions, clearinghouses, case studies, or examples. Every effort has been made to select resources that are useful, current, and widely accepted as authoritative. However, the list is not exhaustive and, given the dynamic nature of digital information, can be expected to change over time. The resources listed will in some cases serve as a starting point to lead the reader to additional resources. This *Framework* is intended to be flexible enough to accommodate new principles, considerations, and resources, and to absorb the contributions of others. At the same time, it is intended to be a concise introduction to core

considerations for the building of good digital collections and to serve as a springboard to encourage further research and innovation by its readers.

There are no absolute rules for creating good collections, objects, or metadata. Every project is unique, with its own goals and needs. There are almost as many ways of categorizing collections as there are collections. Projects dealing with legacy collections or with born-digital materials, for example, have different constraints than projects embarking on new digitization to create entirely new collections. Museums, libraries, and school boards have different constituencies, priorities, institutional cultures, funding mechanisms, and governance structures. The key to a successful project is not to strictly follow any particular path but to plan strategically and make wise choices from an array of tools and processes to support the unique goals and needs of each collection. To use the *Framework* successfully, project planners should take into consideration their organizational goals, their audience, and the content available to them, and they should select the set of principles and resources that best meet their project's needs while ensuring content will be reusable in new and innovative contexts.

A number of excellent resources take a holistic view of digitization projects, covering topics ranging from selection, capture, and description to preservation and long-term access. The following are highly recommended:

- *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*, 2002. http://www.nyu.edu/its/humanities/ninchguide/
- Northeast Document Conservation Center. *Handbook for Digital Projects: A Management Tool for Preservation & Access*, 2000. http://www.nedcc.org/digital/dighome.htm
- Anne R. Kenney and Oya Y. Rieger. *Moving Theory into Practice: Digital Imaging for Libraries and Archives* (Research Libraries Group, 2000). An online tutorial serves as an introduction to topics covered more extensively in the printed volume. http://www.library.cornell.edu/preservation/tutorial/
- *The Arts and Humanities Data Service* (AHDS) publishes a series of guides to good practice that cover collection, description, and digitization for specific types of materials, such as GIS, performance resources, and virtual reality. http://www.ahds.ac.uk/creating/guides/index.htm
- Stephen Chapman. *Techniques for Creating Sustainable Digital Collections*. Library Technology Reports 40(5) (Sept./Oct.2004). Available from Library Technology Reports at http://www.techsource.ala.org/ltr/. Less comprehensive but more in-depth coverage of digitization projects.

## COLLECTIONS

A digital collection consists of digital objects that are selected and organized to facilitate their access and use. Good digital collections include metadata used to describe and manage them. Metadata may be provided for the collection as a whole, for individual items within the collection, or both. In addition to objects and metadata, a digital collection will include at least one interface that gives access to it. This interface often includes a way to search for objects, a way to browse through meaningful categories of objects, and methods for retrieving, rendering, and using objects. As such, the whole is greater than the sum of the parts. Digital collections are generally created by organizations or groups of cooperating organizations, often as part of a project.

Principles that apply to good digital collections are:

***Collections principle 1: A good digital collection is created according to an explicit collection development policy that has been agreed upon and documented before digitization begins.***

Of all factors, collection development is most closely tied to an organization's own goals and constituencies. Collection builders should be able to summarize the mission of their organization and articulate how a proposed collection furthers or supports that mission. Project managers should be able to identify the target audience(s) for the collection (both in the short term and in the future) and how the selected materials relate to their audience. The digital collection should fit in with the organization's overall collection policy, as digital collections should not stand in isolation from the original materials or from the collection as a whole.

There is an often unexamined assumption that digitization will dramatically increase the use or value of materials. This is not necessarily the case, and funding agencies in particular have learned to expect a more considered justification. If the materials exist in non-digital form, how heavily are they used? What factors specifically will influence their use or value when digitized?

The following documents are guidelines for selecting materials for digitization. The list does not include electronic collection development policies, which are documents drafted to guide libraries in their selection of commercially available resources.

- *Building the Digital Collection* (Library of Congress, 1999). http://memory.loc.gov/ammem/ndlpedit/handbook/digitizing.html
- D. Greenstein. *Strategies for Developing Sustainable and Scalable Digital Library Collections* (Digital Library Federation, May 2000). http://www.diglib.org/collections/collstrat.htm
- Dan Hazen, Jeffrey Horrell, and Jan Merrill-Oldham. *Selecting research collections for digitization* (CLIR, August 1998). http://www.clir.org/pubs/abstract/pub74.html
- Joint RLG and NPO Preservation Conference. *Guidelines for Digital Imaging: Guidance for selecting materials for digitisation*. http://www.rlg.org/preserv/joint/ayris.html
- *Moving Theory into Practice. Digital Imaging Tutorial: Selection.* http://www.library.cornell.edu/preservation/tutorial/selection/selection-01.html
- *Towards a Learning Nation: The Digital Contribution*. Recommendations proposed by the Federal Task Force on Digitization. Final Report. (December 31, 1997). Part B Issue 2: Selecting materials for digitization. Although this addresses Canadian collections, the principles are generalizable. http://www.collectionscanada.ca/8/3/r3-407-e.html

A report of the DLESE Collections Committee, *How to Identify the 'Best' Resources for the Reviewed Collection of the Digital Library for Earth System Education*, describes a distributed selection process that could be applied to other learning resources (http://www.ldeo.columbia.edu/DLESE/collections/CGms.html).

Below are some examples of local collection development policies. This is intended to be a selection and not comprehensive.

- Columbia University Libraries. *Selection Criteria for Digital Imaging*. http://www.columbia.edu/cu/libraries/digital/criteria.html
- Library of Congress. *Collections Policy Statements: Electronic Resources*. http://lcweb.loc.gov/acq/devpol/electron.html
- New Jersey Digital Highway. *Collection Development Policy*. http://www.njdigitalhighway.org/documents/njdh-coll-dev-policy.pdf
- North Carolina ECHO (Exploring Cultural Heritage Online) Portal *Collection Development Policy*. Criteria for selecting sites to include in a portal. http://www.ncecho.org/colldev.asp
- University of California *Selection Criteria for Digitization*. http://libraries.universityofcalifornia.edu/cdc/pag/digselec.html

Additional examples may be found in the Digital Library Federation's database of digital library documents, which includes the collection development policies of a number of DLF members (http://www.hti.umich.edu/cgi/b/bib/bib-idx?c=dlf). Some of these policies concern all electronic acquisitions while others focus on retrospective digitization. Browse by subject for *Collection development policies and practices*.

In some instances digitization may be a valid choice for reformatting paper and analog materials for preservation. In 2004 the Association of Research Libraries endorsed digitization as an acceptable preservation reformatting option. There are a number of guidelines for selecting materials for digitization specifically for preservation purposes:

- Joint RLG and NPO Preservation Conference. *Guidelines for Digital Imaging: Selection Guidelines for Preservation*. http://www.rlg.org/preserv/joint/gertz.html
- *Selection Criteria for Preservation Digital Reformatting*. Library of Congress Preservation Reformatting Division. http://lcweb.loc.gov/preserv/prd/presdig/presselection.html

Collection builders should be aware that special constraints may exist in relation to politically and culturally sensitive materials. Even items that are unexceptional in the context of a repository can be disturbing when taken out of context. Selection guidelines with particular attention to sensitivity are included in the Northeast Documentation and Conservation Center's *Handbook for Digital Projects*, chapter IV: *Selection of Materials for Scanning*, by Diane Vogt-O'Connor (http://www.nedcc.org/digital/iv.htm).

***Collections principle 2: Collections should be described so that a user can discover characteristics of the collection, including scope, format, restrictions on access, ownership, and any information significant for determining the collection's authenticity, integrity, and interpretation.***

Collection description is a form of metadata (see also METADATA). Such description serves two purposes: it helps people discover the existence of a collection, and it helps users of the collection understand what they are viewing. Describing collections in established catalogs and registries is also a way of establishing the authority of the content, helping users distinguish authoritative from informal information.

When possible, collections should be described in collection-level cataloging records contributed to a national union catalog such as the OCLC or RLIN databases. Websites and individual digital objects can be cataloged through OCLC Connexion (http://www.oclc.org/connexion/).

There are also a number of directories where collections can be registered. Registry entries can be created by non-catalogers, and can be both simpler and more descriptive than cataloging records. The December 2000 issue of RLG DigiNews gives a slightly out-of-date but still useful inventory of directories of Web-accessible collections.

(http://www.rlg.org/preserv/diginews/diginews4-6.html - faq). The registries listed below allow institutions to register their own collections, or to propose their collections for registration.

- The Association of Research Libraries maintains a database of digital initiatives that includes technical as well as collections information. (Only ARL members can contribute.) http://www.arl.org/did/index.html
- The Smithsonian Institution maintains a list of Library and Archival Exhibitions on the Web. http://www.sil.si.edu/SILPublications/Online-Exhibitions/
- The UNESCO/IFLA Directory of Digitized Collections lists major cultural heritage collections and programs worldwide. http://www.unesco.org/webworld/digicol/
- The University of Arizona maintains a Clearinghouse of Image Databases. http://www.library.arizona.edu/images/clearinghouse/clearinghouse.html

The Digital Collections and Content project, an IMLS-funded initiative at the University of Illinois at Urbana Champagne, is building a registry of all digital collections built with IMLS funds (http://imlsdcc.grainger.uiuc.edu/resources.htm - whatisregistry). Their website discusses the benefits of collection level description and gives examples of collection description schema. As is clear from their pages, there is no dominant metadata standard for describing collections, although at least two initiatives are working towards this goal:

- The Collection Description project of the UK's Research Support Libraries Programme (RSLP) links to many materials related to collection description including an RDF-based schema intended to be both human and machine-readable (http://www.ukoln.ac.uk/metadata/rslp/). The set of data elements included in this schema can be used as a checklist of information a project might want to provide about its collection.
- In the U.S., a task group has been formed as part of the NISO MetaSearch Initiative to foster standardization in the area of collection description in order to enable both people and software applications to select the most relevant collections to target in a federated metasearch of multiple collections (http://www.niso.org/committees/MS_initiative.html). Deliverables include a list of data elements to describe a collection, guidelines for maintaining and exchanging collection information, and recommendations on further steps including best practices and implementation guidelines.

After a user has discovered a relevant collection, collection description should help him or her understand the nature and scope of the collection and any restrictions that apply to the use of materials within it. If the collection has a website, it is good practice to incorporate a narrative description of the collection on the site in human readable prose. There should be a description of the materials comprising the collection, including how and why they were selected. The organization(s) responsible for building and maintaining the collection should be clearly identified, as organizational provenance is an important clue to the authenticity and authority of the collection. Terms and conditions of use, restrictions on access, special software required for general use, the copyright status(es) of collection materials, and contact points for questions and comments should be noted. Many project planners find a description of the methodologies, software applications, record formats, and metadata schemes used in building other collections helpful.

Good examples of collection-level terms and conditions of use are provided by the Library of Congress' American Memory (http://memory.loc.gov/ammem/index.html), JSTOR (http://www.jstor.org/about/terms.html), and the Ad*Access Project (http://scriptorium.lib.duke.edu/adaccess/copyright.html). For examples of websites with extensive technical and project documentation, see Ad*Access and Historic Pittsburgh (http://digital.library.pitt.edu/pittsburgh/). The PALMM (Publication of Archival, Library and Museum Materials) program of the state universities of Florida has a standard template for "sidebar" information with links such as "About the Collection," "Technical Aspects," "Related Sites," etc. (http://palmm.fcla.edu/strucmeta/guidelines.pdf).

When a digital collection is also an archival collection, special rules apply. Archival collections are generally described by curators according to established principles of archival description. Archival finding aids describe archival collections in an hierarchical way, starting with high-level description of the collection as a whole, and moving through established series and subseries, down to the contents of individual boxes and folders.

- The *General International Standard Archival Description* is a set of general rules for archival description developed by the International Council on Archives. http://www.ica.org/biblio.php?pdocid=1
- *Encoded Archival Description* (EAD) is a systematic representation of archival finding aids. EAD is most often expressed in XML. It provides a hierarchical structure map for the archival collection. http://www.loc.gov/ead/

***Collections principle 3: A collection should be sustainable over time. In particular, digital collections built with special internal or external funding should have a plan for their continued usability beyond the funded period.***

Sustainability at the collection level is related to, but not identical with, persistence at the object level (see OBJECTS). Certainly the collection-level archiving strategy should be tied to the preservation strategy at the object level. Managers of collections containing materials of long-term importance should take steps to ensure not only that the objects within them will be preserved in usable form over time, but also that collection-level access to the materials is maintained.

This implies, first and foremost, that some organizational commitment to the ongoing maintenance of the collection is established. Collection maintenance may take different sets of skills and different commitments of resources than the original collection building. The digital repository must be integrated into the institutional collections management workflow. Aspects of ongoing maintenance include such functions as maintaining the currency of locations, ensuring that access applications remain usable, data entry and data cleaning, logging and accumulating statistics, and providing some level of end-user support. They also include the system administration functions of upgrading server hardware and operating system software as required over time, maintaining server security, and ensuring that restoration of applications and data from backups is always possible. Institutions are beginning to adopt persistent naming policies that apply both at the collection and object level; at the collection level this may involve establishing a persistent namespace.

There is a growing body of literature on sustainability. Two particularly relevant resources are listed here:

- *Building and Sustaining Digital Collections: Models for Libraries and Museums* (CLIR, August 2001) reports on a meeting focusing on business models for sustainability. http://www.clir.org/pubs/reports/pub100/pub100.pdf
- Waters, Donald J. *Building on Success, Forging New Ground: The Question of Sustainability*. First Monday, v. 9: no 5, (May 2004). The program officer for scholarly communications at The Andrew W. Mellon Foundation discusses three factors that contribute to the sustainability of digital scholarly resources. http://firstmonday.org/issues/issue9_5/waters/index.html

As part of its goals to document and promote strategies for developing sustainable, scaleable digital collections, as well as to encourage the development of new collections and collection services, the Digital Library Federation is establishing the Registry of Digital Masters jointly with OCLC. The Registry is designed to record digital books and journals that are created in accordance with the DLF's Benchmarks for Digital Reproductions. The *Digital Registry Phase One Implementation Guidelines* are available at this time to testing institutions (http://www.diglib.org/collections/reg/reg.htm).

Two works on creating portals to third-party resources (rather than creating new digital content) that focus on sustainability are:

- The *DESIRE Information Gateways Handbook*, which contains generally useful information on link checking and related maintenance activities in a section on collection management. http://www.desire.org/handbook/
- Pitschmann, Louis A. *Building Sustainable Collections of Free Third-Party Web Resources*. (Washington, DC: Digital Library Federation, Council on Library and Information Resources, June 2001). http://www.clir.org/pubs/abstract/pub98abst.html

***Collections principle 4: A good collection is broadly available and avoids unnecessary impediments to use. Collections should be accessible to persons with disabilities, and usable effectively in conjunction with adaptive technologies.***

At this time, the World Wide Web is the vehicle for broad availability. Collections should be accessible through the Web, using technologies that are well-known among the target user community. There is often a tradeoff between functionality and general usability; the timing of the adoption of new features should be considered in light of how many potential users will be capable of using the technology and how many will find it a barrier. Bandwidth requirements are also a consideration, as some file formats or interfaces may not be usable by individuals on low bandwidth connections. The minimum browser version and bandwidth requirements for use should be documented as part of the collection description.

For general access collections, the webpages and search forms providing access to the collection, as well as the metadata and digital object displays, should be tested against a range of browser applications (e.g., Netscape, Internet Explorer, Opera) and browser versions. Different operating systems support different commands for manipulating screen information, such as selecting multiple items in a drop down menu on a search screen, so testing should include Windows, Mac, and Linux operating systems for at least the current and previous three years. Testing should include different screen resolutions (varying height and width pixel arrays). Look for particularly problematic items, such as color variations, display of non-English language characters, and rendition of XML.

There are several guides to style sheets and Web browsers. Some useful ones include the following:

- *WebReview Browser Compatibility Chart* is a useful quick reference, although not completely current as of this writing. http://www.afactor.net/toolbox/notes/HTML/info/browsersCompatability.html
- *NetMechanic Browser Compatibility Tutorial* is a tutorial to help solve problems with browser compatibility. http://www.netmechanic.com/browser-photo/tutorial.htm

The report *Performance Measures for Federal Agency Websites* by Chuck McClure et. al. addresses website design in terms of efficiency, effectiveness, service quality, impact, usefulness, and extensiveness (http://fedbbs.access.gpo.gov/library/download/MEASURES/measures.doc).

Accessibility is not only good policy, it is also the law as embodied in the *Americans with Disabilities Act of 1990*. The International Center for Disability Resources on the Internet publishes *An Overview of Law & Policy for IT Accessibility* (http://www.icdri.org/CynthiaW/SL508overview.html).

Current de facto accessibility standards are developed by the World Wide Web Consortium (W3C) Web Accessibility Initiative (WAI), which has issued *Web Content Accessibility Guidelines 1.0* (http://www.w3.org/TR/WAI-WEBCONTENT/). As of March 2004, version 2.0 was being drafted. (See also http://www.w3.org/WAI/.)

An example of how accessibility guidelines can be applied in an institutional context is given by the Yale University Library. Their document, *Services for Persons with Disabilities,* has a section

on *Web Accessibility Guidelines* which also lists other accessibility resources (http://www.library.yale.edu/Administration/SQIC/spd2.html - s3).

The Bobby application will check a webpage or website for barriers to persons with disabilities (http://bobby.watchfire.com/bobby/html/en/index.jsp). Bobby is a free service offered by the Watchfire Corporation.

Several clearinghouses focus on Web accessibility, among them:

- CPB/WGBH National Center for Accessible Media has a number of accessibility initiatives including projects focused on educational materials. http://ncam.wgbh.org/projects/
- University of Wisconsin. Trace Research and Development Center. *Designing More Usable Websites*. A clearinghouse of useful tools, initiatives, documentation, and websites. http://trace.wisc.edu/world/web/

***Collections principle 5: A good collection respects intellectual property rights. Collection managers should maintain a consistent record of rightsholders and permissions granted for all applicable materials.***

Intellectual property law must be considered from several points of view: what rights the owners of the original source materials retain in their materials, what rights or permissions the collection developers have to digitize content and make it available, what rights collection owners have in their digital content, and what rights or permissions the users of the digital collection have to make subsequent use of the materials. Viewed from any side, rights issues are rarely clear cut, and the rights policy related to any collection is more often a matter of risk management than one of absolute right and wrong.

There are a number of clearinghouses on law and policy related to copyright and intellectual property. The International Federation of Library Associations maintains a site with international scope (http://www.ifla.org/II/cpyright.htm). The Library of Congress Copyright Office maintains a site that combines both general and procedural information (http://www.copyright.gov/).

Particularly helpful publications on copyright include:

- Georgia Harper. *Copyright Crash Course*. A general introduction to virtually all copyright-related issues. There is a particularly useful section on the logistics of obtaining permission which takes the perspective of risk vs. benefit. http://www.utsystem.edu/ogc/intellectualproperty/cprtindx.htm
- Mary Minnow. *Library Digitization Projects and Copyright*. Comprehensive and entertainingly presented. http://www.llrx.com/features/digitization.htm
- *Intellectual Property Law Primer for MultiMedia Developers* and *Licensing Still Images: Some Basic Information for Multimedia Developers*. Two aging (1994) but highly praised sources for multimedia developers. http://www.timestream.com/stuff/neatstuff/
- Peter Hirtle. *When works pass into the public domain in the United States: Copyright term for archivists and librarians*. A handy chart for quick lookup of likely status by the date of publication. http://www.copyright.cornell.edu/training/Hirtle_Public_Domain.htm

If digitized materials do have restrictions on use, these must be documented and enforced. At this time there are few mechanisms for exercising programmatic control of resources. Rights Expression Languages (RELs) are seen by some as a solution and their continued development and application bear watching. *Rights expression languages: a report for the Library of Congress* (Karen Coyle, 2004) analyzes a sample of rights expression languages in terms of their impact on the selection, maintenance, and preservation of digital content (http://www.loc.gov/standards/relreport.pdf).

Intellectual property issues surrounding digital media, which are readily shared to a wide audience over the Internet, have resulted in a spate of legislation intended to extend the

provisions of the *Copyright Act* (Title 17 of the U.S. Code) to encompass digital media. The controversial *Digital Millennium Copyright Act* (DMCA) went into effect in October 2000, and explicitly addresses some issues of digital information as part of the *Copyright Act*. The American Library Association offers an excellent overview of DMCA (http://www.ala.org/ala/washoff/WOissues/copyrightb/dmca/Default2515.htm).

***Collections principle 6: A good collection has mechanisms to supply usage data and other data that allows standardized measures of usefulness to be recorded.***

Effective collection management includes standardized measures of collection usefulness, including consistent usage data that can be evaluated over time to gauge the continued relevancy of the collection to users, as well as measures to evaluate the continued usefulness of the collection in supporting the organization's mission. These are important for demonstrating success and ensuring continued institutional support.

A variety of research methods can be used to assess collection usefulness. Observation, surveys, interviews, experiments, and transaction log analysis are a few examples. Each method has its strengths and limitations. Transaction logs, for example, can provide data on the volume of use, what materials were used, and who used them ("this document was viewed by x number of users from y different domains") but by themselves shed limited light on how useful the collections really are to users. Collection evaluators, therefore, will need to employ a series of measures using several data collection methods to obtain a clear picture of the usefulness of a digital collection.

Since measures should be maintained over time and take some resources to support, the measures chosen should be designed to serve some purpose of the sponsoring project or organization. One common objective is to attempt to justify resources devoted to a collection by volume of use, either generally or within a certain user population. Other objectives may be to enlighten collection development policy or to improve the functionality of the retrieval and delivery system.

Metrics are also a tool in the evaluation of collections and projects (see PROJECTS). Evaluation is an iterative process. Results of evaluation should inform the design and improvement of a digital collection. For example, user recommendations can be codified and shared with other users, such as the teacher evaluations provided for digital objects in the MERLOT educational repository collection. Two CLIR reports provide good overviews of evaluation methods used by leading digital libraries to assess their digital collections and services:

- Carol Tenopir. *Use and Users of Electronic Library Resources*. http://www.clir.org/pubs/reports/pub120/contents.html
- Denise Troll Covey. *Usage and Usability Assessment: Library Practices and Concerns*. http://www.clir.org/pubs/abstract/pub105abst.html)

Standards are emerging for measuring use and effectiveness of digital content:

- Project COUNTER's *Code of Practice* is aimed at commercial vendors but can be used by cultural heritage collections as well. http://www.projectcounter.org/
- The *Report on the NISO Forum on Performance Measures and Statistics for Libraries* contains a useful "webography" of initiatives and resources. http://www.niso.org/news/reports/stats-rpt.html
- The Association of Research Libraries has an initiative to develop measures for electronic resources (e-metrics) that includes both commercial resources and local digital collections. http://www.arl.org/stats/newmeas/emetrics/index.html
- The National Information Standards Organization has revised Z39.7, a standard for library statistics, to include better measures for electronic resources. http://www.niso.org/emetrics/

***Collections principle 7: A good collection fits into the larger context of significant related national and international digital library initiatives. For example, collections of content useful to education in science, math, and/or engineering should be usable in the NSF-funded National Science Digital Library (NSDL).***

Collection developers should be cognizant of larger national and international priorities such as the Grand Challenges in science and computer science when defining the intellectual content of their collections. This is a way to expand the use and usefulness of digital collections and may help gain sustainable support for them.

Collection developers should also pay attention to interoperability issues, particularly the ability to contribute metadata to more inclusive search engines. They should be aware of and in contact with related efforts, follow widely accepted benchmarks for quality of content and of metadata, and provide adequate collection description for users to place one collection in the context of others.

Some tools supporting interoperability include:

- Dublin Core Metadata Initiative. In addition to the element set, DCMI is developing registries for metadata interoperability. http://dublincore.org/
- METS (Metadata Encoding and Transmission Standard), a standard developed by the library community for exchanging digital objects. http://www.loc.gov/standards/mets/
- MXF (Materials Exchange Format), a wrapper intended to facilitate the interchange of audiovisual information across organizations and applications. http://www.broadcastpapers.com/sigdis/Snell&WilcoxMXF01.htm
- The *Open Archives Initiative* (OAI), which has developed a protocol for contributing local metadata to larger metadata repositories. http://www.openarchives.org/

Topical collections may fit into broader clearinghouses or cooperative portals. Project planners should search for clearinghouses in their subject area; there is an increasing number of clearinghouses, particularly in areas related to scientific or environmental information. For example:

- The Geospatial Data Clearinghouse is a collection of over 250 spatial data servers that have digital geographic data primarily for use in Geographic Information Systems (GIS), image processing systems, and other modeling software. http://www.fgdc.gov/clearinghouse/clearinghouse.html
- The Global Biodiversity Information Facility aims for "compilation, linking, standardisation, digitisation and global dissemination of the world's biodiversity data." http://www.gbif.org/
- Moving Image Collections (MIC) is a union catalog of moving images held by a variety of organizations with an archive directory and links to other moving image resources. http://mic.imtc.gatech.edu/
- National Science Digital Library is an OAI-based collection of materials useful in science, technology, engineering, and mathematics education. http://www.nsdl.org/
- Sheet Music Consortium is building an open collection of digitized sheet music also using the OAI protocol. http://digital.library.ucla.edu/sheetmusic/
- The Research Libraries Group's Cultural Materials Initiative is a central repository of metadata and some content for cultural heritage digital objects. http://www.rlg.org/culturalres

Cooperative portals are gateways to existing websites and other resources maintained collaboratively by a group of institutions, each taking responsibility for selecting quality resources within some subtopic of a larger subject area. Some examples include:

- Agnic, a portal to agricultural information being developed by the National Library of Agriculture, land grant universities, and other partners. http://www.agnic.org/
- Healthweb, a cooperative project of about 20 health sciences libraries for health-related resources. http://healthweb.org/

### CASE STUDY: REFINING A SELECTION POLICY

**Illinois State Library, Illinois Digital Archives**

The Illinois Digital Archives (IDA), a project begun by the Illinois State Library (ISL) in 1998, operates as both a search engine and a development tool. As a search engine it indexes digital resources located on the Internet sites of libraries, museums and cultural institutions across the state. Among the sites indexed are sites that relate the story of Illinois through the history of local communities, women, coal mining, and events such as the Nazi march on Skokie. As a development tool the project offers grants to libraries for the imaging and cataloging of additional state history resources that will be added to the IDA.

Before the project developed explicit guidelines and selection criteria, proposals were submitted for everything from photograph collections to book jackets. When Alyce Scott became digital imaging coordinator in 2000, she consulted reference sources to develop a resource selection process for the IDA. Several documents helped Scott and project planners focus on the qualities of primary source materials and identify appropriate section criteria, among them the Joint RLG and NPO Preservation Conference Guidelines for Digital Imaging: *Guidance for Selecting Materials for Digitization*, Columbia University Libraries *Selection Criteria for Digital Imaging*, and Hazen, Horell and Merrill-Oldham's *Selecting Research Collections for Digitization*. Questions presented in Hazen on the evaluation of the intellectual content of primary materials, the current and potential users of digital materials, and relationships to other digitization projects were particularly helpful.

With a clearer goal in mind, Scott and her colleagues developed a program called "Planning Your Digital Imaging Project" which was presented several times around the state. A short checklist on selection criteria and issues included in this presentation has become the basis for a more formal checklist that will be distributed to prospective grant writers. In addition, ISL grant offerings now specify proposals to digitize original Illinois history-related materials only, and provide a clear definition of primary source materials. Alyce Scott reports that as a result "in the last round of grant reviews we began to see the fruits of this labor: much more primary source local history collections, with better statements of need. We are now receiving proposals for digitization projects that reflect more accurately the type of material we want to include in the Illinois Digital Archives."

## OBJECTS

This *Framework* is concerned with two kinds of digital objects: digitized objects produced as surrogates for materials in some analog format (e.g. printed books, manuscripts, museum artifacts, video tapes, etc.), and "born digital" objects originally produced in machine-readable form (some e-books, scientific databases, sensor data, digital photographs, websites, etc.).

An object may be complete in one file (e.g. a report issued as a PDF), or it may consist of multiple linked files (e.g. an HTML page and in-line images), or it may consist of multiple files and the structural metadata needed to tie them together (e.g. a book digitized as page images). In this sense, objects are equivalent conceptually to the items that may be found within library holdings, museum collections, and archival collections.

Within the context of this *Framework*, collections consist of objects (see COLLECTIONS). Obviously, no hard and fast line can be drawn between objects and collections. Our definition of an object extends to complex objects such as digitally reformatted books, but not as far as a collection (which in this case would include, for example, two or more digitally reformatted books). A digital object may belong to more than one digital collection.

When speaking of digital objects, it is often useful to distinguish between master or preservation copies and access or use copies. As their names imply, masters are the original first generation of a digital object, typically the highest quality versions that the production technique allows, while use or access copies are derivatives created for specific purposes, distribution scenarios, or users. Thus, a master copy of a digitally reformatted 35mm slide might be an uncompressed, 18 megabyte TIFF file, while the access copy derived from this image might be a 150 KB JPEG image allowing a reasonable download time for the average Web-based user. Where both master and service copies are created, the principals outlined below apply to the master copy, though some apply equally well to the service copy. Note that institutions may have different definitions of what they consider a preservation versus an access copy.

Among the advantages in reaching agreement about what constitutes good objects are the following:

- By agreeing to minimum level benchmarks for good objects, organizations that produce such objects can reduce the risk involved in producing and maintaining them while encouraging their use.
- Because good objects will be considered capable of meeting known current and likely future needs, organizations can invest in their creation secure in the knowledge that they will not be forced to re-create the objects at some future date, even as production techniques improve, thus preserving the original objects from excessive handling.
- Users of good objects will develop confidence in the objects because they will have a minimum level of well-known and consistent properties, and will support a variety of known uses.
- By building consensus around the characteristics of good objects, organizations will be able to more effectively write contracts with vendors who create such objects, to compare vendors' prices, and to define and narrow preservation options required for migrating or emulating the objects in the future.

The following principles apply to good objects:

***Objects principle 1: A good digital object will be produced in a way that ensures it supports collection priorities, while maintaining qualities contributing to interoperability and reusability.***

Decisions about how objects are produced and described should reflect and follow from those made about why they are being produced, for whom, and for what purpose. For that reason, the

guidelines for selection listed in COLLECTIONS are equally relevant to the creation of good objects.

Some examples of how decisions about production and description should follow naturally from strategic collection development decisions are available in Neil Beagrie and Daniel Greenstein, A *Strategic Policy for Creating and Preserving Digital Collections* (2001) (http://www.ahds.ac.uk/strategic.pdf).

***Objects principle 2: A good object is persistent. That is, it will be the intention of some known individual or institution that the good object will remain accessible over time despite changing technologies.***

Digital information is inherently impermanent. The lifespans of digital media, hardware and software platforms, and digital file formats are notoriously short. As Roy Rosenzweig points out, "The life expectancy of digital media may be as little as ten years, but few hardware platforms or software programs last that long."

There are many strategies being tested for use in the preservation of digital objects. Two of the most widely discussed are migration and emulation. Migration involves transforming objects so they can move between technical regimes as those regimes change. Migration occurs at all levels, as objects are moved:

- across media as media evolve (e.g. from diskette to CD, and from CD to optical disk or DAT tape);
- across software products as the products become outmoded (e.g. from one version of a word-processing or database application to another); and,
- across formats as formats evolve (e.g. from SGML to XML, or from JPEG to JPEG2000).

Emulation posits that in some cases, it is better (involves less expense and/or less information loss) to emulate on contemporary systems the computer environment in which digital objects were originally created and used. Emulation strategies may be particularly appropriate for complex multimedia objects such as interactive learning modules.

Another preservation strategy is to create master objects with digital preservation in mind. Although no digital file format will last forever, certain qualities will improve the chances that a digital object can be successfully carried forward into the future. When possible, choose file formats that are non-proprietary and do not contain patented technologies. Formats that are widely used and have published specifications are most likely to have migration paths. Preservation masters of retrospectively digitized materials should be as close to the analog version as possible, and should not contain access inhibitors like watermarks or encryption.

Some resources related to the assessment of digital file formats for preservation:

- U.K. National Archives. *Digital Preservation Guidance Note 1: Selecting file formats for long-term preservation*. http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_file_formats.pdf
- ERPANET. *File Formats for Preservation*. Papers and PowerPoint slides from a seminar on file formats for digital preservation held in Vienna in May 2004. http://www.erpanet.org/events/2004/vienna/index.php - papers
- Gregory W. Lawrence, William R. Kehoe, Oya Y. Rieger, William H. Walters, and Anne R. Kenney. *Risk Management of Digital Information: A File Format Investigation* (CLIR 2000). An investigation conducted by Cornell University Library to assess the risks to digital file formats during migration. http://www.clir.org/pubs/abstract/pub93abst.html
- PRONOM is an online source of information about file formats and software products maintained by the U.K. National Archives. http://www.records.pro.gov.uk/pronom/

There is a large and growing body of literature on the preservation of digital material. For more information on all aspects of this fascinating area, see PADI (Preservation Access to Digital Information), a comprehensive clearinghouse maintained by the National Library of Australia (http://www.nla.gov.au/padi/). See also materials related to preservation of audio and video on Stanford University's Conservation On-Line (CoOL) site (http://sul-server-2.stanford.edu/bytopic/audio/).

***Objects principle 3: A good object is digitized in a format that supports intended current and likely future use or that supports the derivation of access copies that support those uses. Consequently, a good object is exchangeable across platforms, broadly accessible, and will either be digitized according to a recognized standard or best practice or deviate from standards and practices only for well documented reasons.***

In almost every case, there is a direct correlation between the production quality of a digitized object and the readiness and flexibility with which that object may be used, reused, and migrated across platforms. As a result, the digitization of objects at the appropriate level of quality can pay off in the long run as the objects are rendered more useful and accessible over the longer term. Not all objects, of course, will have long-term value. A project needs to assess the value of the digital objects in its collections and make appropriate decisions about persistence and interoperability.

A number of file formats are presented in Table 1 below. They are organized according to a typology that recognizes data types, and within data types, applications to which objects of that type may be put. Note that there is an international effort to document digital representation formats by establishing a global registry that will associate persistent identifiers for digital formats and their syntactic and semantic properties. This is an important development and may be tracked at: http://hul.harvard.edu/gdfr/. See also: Stephen L. Abrams and David Seaman, *Towards a Global Digital Format Registry* (2003) (http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf).

**Table 1: A Typology of Formats**

| DATA TYPE | APPLICATIONS | FORMATS | GUIDELINES and REFERENCES |
|---|---|---|---|
| Alphanumeric data | Flat files; hierarchical or relational datasets | US-ASCII or UTF-8 text, or portable format files recognized as de facto standards (e.g. SAS or SPSS) with enough metadata to distinguish tables, rows, columns, etc. | For social science and historical datasets, see *Guide to Social Science Data Preparation and Archiving* (ICPSR, 2002) (http://www.icpsr.umich.edu/ACCESS/dpm.html) and *Digitising history, a guide to creating digital resources from historic documents* (HDS, 1999) (http://hds.essex.ac.uk/g2gp/digitising_history/index.asp). |
| Alphanumeric data | Encoded texts for networked presentation and exchange of text-based information | SGML, XML; use documented DTDs or Schema | |
| Alphanumeric data | Encoded texts for literary and linguistic content analysis | SGML, XML | *Text Encoding Initiative* (TEI) (http://www.tei-c.org). *Creating and documenting electronic texts* (OTA, 1999) (http://ota.ahds.ac.uk/documents/creating/) and *TEI text encoding* in *Libraries: Guidelines for Best Practice* (DLF, 1999) (http://www.diglib.org/standards/tei.htm). |

| DATA TYPE | APPLICATIONS | FORMATS | GUIDELINES and REFERENCES |
|---|---|---|---|
| Image data; bitonal, grayscale, and color page images of textual documents | Book or serial publication | Archival masters likely to be uncompressed baseline TIFF files or lossless compressed JPEG2000 at color depth and pixelation appropriate for application. Derivative formats for access likely to vary depending on use. | National Archives and Records Administration. *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* (June 2004) (http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf). *A consensus for minimum characteristics is Benchmark for faithful digital reproductions of monographs and serials*, Version 1 (DLF, 2002) (http://www.diglib.org/standards/bmarkfin.htm). An example of one institution's local benchmarks: California Digital Library. *Digital Image Format Standards* (http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf). |
| Image data; bitonal, grayscale, and color page images of textual documents | Newspapers | Grayscale raster formats for masters, almost always supplemented with PDFs and OCR text for access and use. | Library of Congress. *The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants* (http://www.loc.gov/ndnp/pdf/ndnp_techguide.pdf). OCLC Digitization and Preservation Resource Center. Click link to *Newspaper Digitization* (http://digitalcooperative.oclc.org/). |
| Scalable bit-mapped image data to support zooming and multiple resolution delivery from a single file | Maps, herbarium specimens, photographs, aerial photographs | Lossless compressed JPEG2000 files can be used as archival masters and lower quality/smaller size access copies can be derived from them. | |
| Audio | Music audio | Archival masters should consist of a linear PCM bit stream, which may be wrapped as an uncompressed WAVE or AIFF file. End-user delivery format options include MP3 (MPEG-1 Level 3), AAC, and RealAudio. | This brief technical introduction to Digital Audio by the National Library of Canada provides useful explanations although it suggests the use of cleanup tools, a practice eschewed by most preservation reformatting programs (http://www.collectionscanada.ca/9/1/p1-248-e.html). The Harvard University Library *Digital Initiative Audio Reformatting* site has links to industry standards and will include project guidelines in the future (http://hul.harvard.edu/ldi/html/reformatting_audio.html). Essays that discuss concepts and practices include Carl Fleischhauer's paper on the *Library of Congress Digital Audio Preservation Project* (http://www.arl.org/preserv/sound_savings_proceedings/fleischhauer.html) and *Sound Practice: A Report on the Best Practices for Digital Sound Meeting* (16 January 2001) at the Library of Congress (http://www.rlg.org/preserv/diginews/diginews5-2.html#feature3). A statement on ethics and practices has been published by the International Association of Sound and Audiovisual Archives (http://www.iasa-web.org/iasa0013.htm). |

| DATA TYPE | APPLICATIONS | FORMATS | GUIDELINES and REFERENCES |
|---|---|---|---|
| Audio | Spoken word (e.g. oral histories) | See music audio above. | HistoricalVoices.org (http://www.historicalvoices.org/), a project of the National Gallery of the Spoken Word (http://www.ngsw.org/), includes best practices and research papers on digital speech and an oral history tutorial.<br><br>The Spoken Word Project at Northwestern University is a good example of work on synchronizing transcripts and sound recordings (http://www.at.northwestern.edu/spoken/). |
| Video (A/V) | Moving image content originally created as video or transferred from film | High resolution video files are huge and digital file formats for preservation quality video are immature. Therefore, at this time most organizations maintain their best archival copies of video content in media-dependent form. Preferred media-dependent formats contain a minimally compressed or uncompressed signal, e.g., DigiBeta, D1, or D5 tape. In a high bandwidth LAN, access copies may be high-bit rate MPEG-2 or MPEG-4 files in larger picture sizes; for lower bandwidth applications and the Web, one may present lower-bit rate MPEG-4, RealVideo, or QuickTime formats with smaller picture sizes. | The *NINCH Guide to Good Practice in the Digital Representation of Cultural Heritage Materials* has a good chapter on audio/video capture and management (http://www.nyu.edu/its/humanities/ninchguide/).<br>The Video Development Group (ViDe) provides information about digital video file creation (http://www.vide.net).<br>Agnew, Grace. *Video on Demand: the Prospect and Promise for Libraries* in the *Encyclopedia of Library and Information Science* (New York: Marcel Dekker, 2004) gives an overview of digital video (http://www.dekker.com/sdek/issues~db=enc~content=t713172967).<br>Association of Moving Image Archivists. *Reformatting for Preservation: Understanding Tape Formats and Other Conversion Issues* (http://www.amianet.org/publication/resources/guidelines/videofacts/reformatting.html).<br>The Association of Moving Image Archivists (http://www.amianet.org/) is a non-profit professional association established to advance the field of moving image archiving. Many of the postings on the AMIA-L listserv (http://www.amianet.org/amial/amial.html) are relevant to video archiving; the archive for the listserv may also be consulted: (http://lsv.uky.edu/archives/amia-l.html). |
| Video (A/V) | Capturing live performances | See above. | Two draft guides from the Internet2/CNI Performance Archive and Retrieval Working Group: *Capturing Live Performance Events,* version 0.9 (2003) (http://arts.internet2.edu/files/performance-capture(v09).pdf).<br>*Current Practice in Digital Asset Management,* version 0.9 (2003) http://arts.internet2.edu/files/digital-asset-management(v09).pdf |

| DATA TYPE | APPLICATIONS | FORMATS | GUIDELINES and REFERENCES |
|---|---|---|---|
| Miscellaneous | GIS | Alphanumeric data (e.g. as required to record coordinates), vector, and raster graphics (e.g. to represent maps). | *GIS. A guide to good practice* (ADS, 1998) (http://ads.ahds.ac.uk/project/goodguides/gis/index.html) |

***Objects principle 4: A good object will be named with a persistent, unique identifier that conforms to a well-documented scheme. It will not be named with reference to its absolute filename or address (e.g. as with URLs and other Internet addresses) as filenames and addresses have a tendency to change. Rather, the stable identifier can be resolved (mapped) to the actual address.***

In the best of all possible worlds, locally assigned identifiers would conform to known national or international standards. Unfortunately, most standard identifiers point to classes of objects (e.g. the ISBN, which identifies all books in a particular edition), or can only be assigned by particular agencies, or cost a fee to register. For most digital collections, the object identifiers will have to be assigned locally, according to some local scheme. This is not a problem, so long as the scheme is documented and the documentation is accessible. It is also possible to incorporate standard identifiers into a local naming scheme. For example, in a digital collection of journal articles, the object identifier could consist of a prefix indicating the institution assigning the identifier followed by the SICI for the article. (There is a longstanding controversy over whether identifiers should be "smart" or "dumb," that is, whether they should carry meaning or not. We feel that neither method is universal best practice and that applications can have good reason to prefer one or the other.)

Actionable identifiers for Internet accessible objects should utilize a registry (or "resolver") that maps from the static persistent identifier to the current location of the object. Although the mapping tables must be updated when an object is moved, this degree of indirection facilitates maintenance because the location needs only be updated once in a central spot, no matter how many times the identifier occurs in references. Two registry based applications are Persistent URLs (PURLs) and Handles.

PURLs are URLs resolved to true locations by a PURL server which contains tables mapping the PURLs to the true URLs. OCLC runs a central PURL server that anyone can use. Alternatively, any organization can download and install the free PURL server application and manage its own PURL server locally. (See http://www.purl.org/.)

The Corporation for National Research Initiatives (CNRI) developed the Handle System, a resolver application for persistent identifiers called “handles.” CNRI maintains a global handle registry as well. Organizations wishing to utilize the Handle System must register a namespace with CNRI. As with the PURL server, organizations have the choice of using the resolver at CNRI or running their own application locally. The DOI (Digital Object Identifier) is a proprietary implementation of the Handle System. Many commercial and open-source digital repository applications, such as Fedora and DSpace utilize the Handle System for object identification. (See http://www.handle.net/.)

A new method of representing persistent identifiers is through the INFO URI scheme. This is a consistent way to represent and reference such standard identifiers as Dewey Decimal Classifications on the Web so that these identifiers can be "read" and understood by Web applications. INFO is a lightweight way to register public identifier schemes. Some that have been registered include: Library of Congress Control Number (LCCN), PubMed identifier, DDC number, and Digital Object Identifier (DOI). Other identifiers have been registered as URN namespaces, such as ISBN and ISSN. (See http://info-uri.info/.)

Regardless of the technology used, for identifiers to remain persistent, it is important that an institution take responsibility for the object and its identifier's maintenance.

The following sites contain information about standard numbers:

- International Standard Book Number system (ISBN). http://www.isbn-international.org/
- Digital Object Identifier. http://www.doi.org/
- International Standard Serial Number (ISSN). http://www.issn.org/pub/
- Serial Item and Contribution Identifier. http://sunsite.berkeley.edu/SICI/
- MPEG-21 Part 2 Digital Item Declaration http://www.chiariglione.org/mpeg/standards/mpeg-21/mpeg-21.htm - _Toc23297974

For more information about Uniform Resource Names (URNs), see http://www.ietf.org/html.charters/OLD/urn-charter.html.

For information about the application of naming schemes see:

- *Handle System: a general purpose global name service enabling secure name resolution over the Internet*. (Corporation for National Research Initiatives, 2003). http://www.handle.net/
- Harvard University Library Office for Information Systems. *Naming and Repository Services. An Introduction*. A detailed introduction to these services including a gentle explanation of the importance of good practices in the design of naming services. http://hul.harvard.edu/ldi/resources/nrsdrsservice.pdf
- Harvard University Library Office for Information Systems. *Name Resolution Service (NRS) Technical Overview* (2000). A technical overview of the Name Resolution Service (NRS) developed by the Harvard's Library Digital Initiative. http://hul.harvard.edu/ldi/resources/nrs-overview-public.html
- *IMS Persistent, Location-Independent, Resource Identifier Implementation Handbook*. http://imsglobal.org/implementationhandbook/imsrid_handv1p0.html
- California Digital Library. *Archival Resource Key (ARK)*. http://www.cdlib.org/inside/diglib/ark/

***Objects principle 5: A good object can be authenticated in at least three senses. First, a user should be able to determine the object's origins, structure, and developmental history (version, etc.). Second, a user should be able to determine that the object is what it purports to be. Third, a user should be able to determine that the object has not been corrupted or changed in an unauthorized way.***

Being able to authenticate an object is essential for a number of reasons. Research is predicated on verifiable evidence. Teaching and learning, as well as other forms of cultural engagement, also rely on a user's ability to assess an information object's veracity, accuracy, and authenticity. There are some cases where verification takes on additional significance, as for example with information that supplies evidence about important past or current events. In archives, the authenticity of records, including digital records, has legal significance.

Information documenting an object's origin and developmental history is known as "digital provenance" and is recorded as metadata associated with an object (see METADATA).

Determining the veracity of a digital object is likely to rely upon techniques whose reliability is still debated. Techniques appropriate to digital images include message digests, digital signatures, and watermarking.

More information may be found at

- *Authenticity in a Digital Environment* (CLIR, 2000). Report of a group of experts convened by CLIR to address the question: What is an authentic digital object? http://www.clir.org/pubs/reports/pub92/contents.html
- *The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections* (2001). Addresses the importance of verifying the authenticity of information objects. http://www.clir.org/pubs/reports/pub103/contents.html
- MD5 unofficial home page. http://userpages.umbc.edu/~mabzug1/cs/md5/md5.html
- David Youd. *What is a digital signature?* The simplest explanation of digital signatures. http://www.youdzone.com/signature.html
- *The information hiding homepage: Steganography and digital watermarking*. http://www.petitcolas.net/fabien/steganography/

***Objects principle 6: A good object will have associated metadata. All good objects will have descriptive and administrative metadata. Some complex objects will have structural metadata.***

The Philadelphia Art Museum reports some 300,000 unique items in its collection. Think how hard it would be to find an object, or to know what it was, who created it, or when it was created, if the museum did not provide this information in the form of metadata. The same holds true for digital objects. A good object must have embedded or associated information allowing humans or their agents to locate and identify it.

For more information see METADATA.

## METADATA

One of the most challenging aspects of the digital environment is the identification of resources available on the Web. The existence of searchable descriptive metadata increases the likelihood that digital content will be discovered and used. Collection-level metadata is addressed in the COLLECTIONS section of this document. This section addresses the description of individual objects and sets of objects within collections.

Metadata is structured information associated with an object for purposes of discovery, description, use, administration, and/or management. Metadata can be added at any stage of an information object's life cycle. For example, at the creation stage, metadata about an object's authors, contributors, and sources could be recorded by the original authors. At the organization stage, metadata about subjects, publishing history, intended audience, summary, and so on could be recorded by catalogers or indexers. At the access and usage stage, metadata on access privileges, reproduction rights, and preservation could be included by access managers. It is helpful to keep in mind that different types of metadata can be added by different people at various stages of an information object's life cycle.

It is common to distinguish between three basic kinds of metadata. Descriptive metadata helps users find objects, distinguish one object from another, and understand the subject or contents of objects. Administrative metadata helps collection managers keep track of objects for such purposes as file management, rights management, and preservation. Structural metadata documents relationships among objects, such as the relationship between articles, issues, and volumes of serial publications, or the pages and chapters of a book.
A primary reason for digitizing collections is to increase access to the resources held by the organization. Creating broadly accessible descriptive metadata is a way to maximize access by current users and attract new user communities. Examples of metadata-based access tools include library catalogs, archival finding aids, and museum inventory control systems. Over the years metadata schemes have been developed for describing a wide range of digital objects. Within this multiplicity of schemes, there is a degree of consistency that supports interoperability. For example, many schemes provide for a name, date, and identifier. While cultural heritage institutions explore the metadata standards that are being adopted within their field, they will want to consider the interoperability issue early in their metadata implementation to assure the greatest likelihood of interoperability. See Metadata Principle 2 for more information about interoperability.

There is usually a direct relationship between the cost of metadata creation and the benefit to the user: describing each item is more expensive than describing collections or groups of items; using a rich and complex metadata scheme is more expensive than using a simple metadata scheme; applying standard subject vocabularies and classification schemes is more expensive than assigning a few keywords, and so on. It should be noted however, that expenditures in development often result in greater efficiency and effectiveness for the end user. Use of a standardized subject thesaurus, for example, provides greater precision in searching for end users and can enable future functionality, such as structured subject browsing and dynamic subject portals. The decisions of which metadata standard(s) to adopt, what levels of description to apply, and so on must be made within the context of the organization's purpose for digitizing the collection, the users and intended usage, approaches adopted within the community, and the desired level of access.

Questions to consider include, but are not limited to:

- Purpose of the digital collection
  - What are the goals and objectives for building this collection?

- User needs & intended usage
  - Who are the targeted users? What types of information do they need to know about the collection and the individual items?
  - Are the materials to be accessed as a collection or will individual items be accessible?
  - Will users need to have access to the source object and its digital counterpart?
- Metadata standard selection & usage
  - Does the collection or its items have metadata before the digital collection is built? How useful are existing metadata for collection control, management, and access?
  - What type of cultural heritage institutions will be involved in the project? What are the metadata standards that are used by organizations in this domain? Which ones are most appropriate for this particular collection?
  - What subject discipline will be involved? What are the metadata standards that are commonly used by users of this discipline?
  - What is the format of the original resources?
  - How rich of a description is needed and does the metadata need to convey hierarchical relationships?
  - How will you distinguish between the source object and the digital surrogate available on the Web?

Several principles guide the selection and implementation of metadata standards.

***Metadata Principle 1: Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current, and likely future use of the digital object.***

There are a variety of published metadata schemes that can be used for digital objects, websites, and e-resources. The book *Metadata Fundamentals for All Librarians* (Priscilla Caplan, ALA Editions, 2002) describes more than fifteen schemes used by educational, scientific, and cultural institutions. There will often be more than one scheme that could be applied to the materials in a given collection. The choice of scheme will reflect the level of resources the project has to devote to metadata creation, the level of expertise of the metadata creators, the expected use and users of the collection, and similar factors. Organizations should consider the granularity of description, that is, whether to create descriptive records at the collection level, at the item level, or both, in light of the desired depth and scope of access to the materials. They should also consider which schemes are commonly in use among similar organizations; using the same metadata scheme will improve interoperability among collections.

The International Federation of Library Association site *Digital Libraries: Metadata Resources* is a clearinghouse of metadata schemes (http://www.ifla.org/II/metadata.htm). *Introduction to Metadata: Pathways to Digital Information* (Murtha Baca, ed.) is a good general introduction to metadata issues for cultural heritage institutions (http://www.getty.edu/research/conducting_research/standards/intrometadata/index.html). See also NISO's *Understanding Metadata* (http://www.niso.org/standards/resources/UnderstandingMetadata.pdf).

The following is a selection of metadata schemes used by many cultural heritage institutions.

**Table 2: Metadata Schemes**

| METADATA SCHEME | DESCRIPTION | APPLICATIONS AND GUIDELINES |
|---|---|---|
| Dublin Core<br>http://dublincore.org/ | A simple generic element set applicable to a variety of digital object types. Dublin Core has been adapted by a number of communities to suit their own needs (such as the CIMI application profile for the museum community), and incorporated into several domain-specific metadata schemes. | *The CIMI Guide to Best Practice for Museums using Dublin Core*. http://www.cimi.org/public_docs/meta_bestprac_v1_1_210400.pdf<br>GEM (Gateway to Educational Materials). http://www.thegateway.org/about/documentation/gem-2-element-set-and-profiles<br>Open Archives Initiative. http://www.openarchives.org/<br>Western States Dublin Core Metadata Best Practices. http://content.lib.utah.edu/cdm4/item_viewer.php?CISOROOT=/docs_regional&CISOPTR=1&REC=1 |
| Encoded Archival Description (EAD)<br>http://www.loc.gov/ead/ead2002.html | A set of rules for the representation of the intellectual and physical parts of archival finding aids. Often expressed in XML or SGML so that the information can be searched, retrieved, displayed, and exchanged. | SAA. EAD Working Group. *Encoded Archival Description Application Guidelines*. (SAA, 1999.) Guidelines for the latest (2002) version of the format are not yet available; watch http://www.loc.gov/ead/ for news of their release.<br>RLG. EAD Advisory Group. *RLG Best Practice Guidelines for Encoded Archival Description* (2002). http://www.rlg.org/rlgead/bpg.pdf<br>Online Archives of California. *OAC Best Practice Guidelines for EAD*, Version 2.0. http://www.cdlib.org/inside/diglib/guidelines/bpgead)<br>*The EAD Cookbook*. http://www.archivists.org/saagroups/ead/resources/ead2002cookbook/EAD2002cookbook.pdf |

| METADATA SCHEME | DESCRIPTION | APPLICATIONS AND GUIDELINES |
|---|---|---|
| Learning Object Metadata | Learning Object Metadata is used to describe educational resources in course management systems and learning management systems. The main standard is the IEEE *Standard for Learning Object Metadata* (1484.12.1-2002), also called the LOM, which must be ordered from IEEE (http://ltsc.ieee.org/wg12/par1484-12-1.html).<br>However, the LOM has been incorporated into a number of other standards, including the IMS Global Learning Consortium's *Meta-Data Specification* which is freely available from the IMS (http://www.imsglobal.org/specificationdownload.cfm). | *IMS Meta-data Best Practice Guide for IEEE 1484.12.1-2002 Standard for Learning Object Metadata*. Version 1.3 Public Draft. http://www.imsglobal.org/metadata/mdv1p3pd/imsmd_bestv1p3pd.html<br>*CanCore Guidelines for the Implementation of Learning Object Metadata (LOM) 2.0*. http://www.cancore.ca/documents.html<br>These guidelines were developed over three years by the Canadian educational community. |
| MARC21<br>http://lcweb.loc.gov/marc/ | A long established standard within the library community for exchanging cataloging information. MARC supports the *Anglo-American Cataloging Rules* and is maintained by the library community. Over the last several years, MARC has been enhanced to support descriptive elements for electronic resources. | Library of Congress. *Understanding MARC Bibliographic: Machine-Readable Cataloging*. 7th Edition. http://lcweb.loc.gov/marc/umb/<br>MARC documentation: Extensive documentation is available at the LC site and at OCLC http://oclc.org/ |
| Metadata Encoding and Transmission Standard (METS)<br>http://www.loc.gov/standards/mets | An XML schema for encoding structural metadata about compound objects, with placeholders for descriptive, administrative, and technical metadata. | *METS Implementation Registry*. http://sunsite.berkeley.edu/mets/registry/<br>METS profiles. A number of profiles were under development as of this writing; when completed they will provide implementation guidelines and will be made available at the METS website (http://www.loc.gov/mets). |
| MODS (Metadata Object Description Schema)<br>http://www.loc.gov/standards/mods | An XML schema for descriptive metadata compatible with the MARC 21 bibliographic format. | *MODS User Guidelines* are available at http://www.loc.gov/standards/mods/v3/mods-userguide.html. |
| VRA Core Categories Version 3<br>http://www.vraweb.org/vracore3.htm | A scheme developed by the Visual Resources Association for the description of art, architecture, artifacts, and other visual resources. The *Core Categories* were designed with the awareness that there are often multiple representations of a work of art, such as the original painting and a slide of the painting used in teaching. | *Cataloguing Cultural Objects* (VRA 2004). Comprehensive guidelines for describing cultural works and their images. http://www.vraweb.org/CCOweb/ |

| METADATA SCHEME | DESCRIPTION | APPLICATIONS AND GUIDELINES |
|---|---|---|
| MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938)<br>http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm<br>The standard can be purchased from the International Organization for Standardization (ISO):<br>http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=34232&ICS1=35 | MPEG-7 is a multimedia description and indexing system that combines XML-based content description with non-textual indexing of physical features (color, movement, shape, sound, etc.) via processing of the media bit stream for multimedia information--audio, video, and images. Part 5 of the standard (ISO/IEC 15938-5) provides descriptive, technical, and usage metadata. | The Moving Image Collections (MIC) project has published an application profile with user guide, PowerPoint tutorials, a crosswalk to Dublin Core, and a prototype MPEG-7 cataloging utility in Microsoft Access, available for free download at: http://gondolin.rutgers.edu/MIC/text/how/cataloging_utility.htm.<br>The IBM alphaWorks development team has released a downloadable MPEG-7 Annotation Tool to annotate video sequences with MPEG-7 metadata, available at http://www.alphaworks.ibm.com/tech/videoannex. |
| SMPTE Metadata Dictionary<br>http://www.smpte-ra.org/mdd | The *SMPTE Metadata Dictionary*, developed by the Society of Motion Picture and Television Engineers, also provides technical metadata for audiovisual formats. | None available at time of publication. |

***Metadata principle 2: Good metadata supports interoperability.***

Teaching, learning, and research today take place in a distributed networked environment. It can be challenging to find resources that are distributed across the world's libraries, archives, museums, and historical societies. To alleviate this problem, cultural heritage institutions must design their metadata systems to support the interoperability of these distributed systems.

The goal of interoperability is to help users find and access information objects that are distributed across domains and institutions. Use of standard metadata schemes facilitates interoperability by allowing metadata records to be exchanged and shared by systems that support the chosen scheme. Another way to achieve interoperability is to map elements from one scheme to those of another scheme. These mappings, or crosswalks, help users of one scheme to understand another, can be used in automatic translation of searches, and allow records created according to one scheme to be converted by program to another. If a locally created metadata scheme is used in preference to a standard scheme, a crosswalk to some standard scheme should be developed in anticipation of future interoperability need.

- The Getty Standards Program maintains crosswalks relevant to art, architecture, and cultural heritage information on their *Metadata Standards Crosswalks* page. http://www.getty.edu/research/conducting_research/standards/intrometadata/3_crosswalks/index.html.
- The Library of Congress maintains crosswalks to and from MARC21. http://lcweb.loc.gov/marc/marcdocz.html
- The NSDL Standards Working Group has a metadata resources page largely devoted to crosswalks. http://metamanagement.comm.nsdlib.org/IntroPage.html
- *A Spectrum of Interoperability: The Site for Science Prototype for the NSDL*. D-Lib Magazine, Vol. 8, no. 2 (Jan. 2002). The paper describes how NSDL achieves interoperability. http://www.dlib.org/dlib/january02/arms/01arms.html

- University of Washington Digital Collections has mapped various metadata standards to Dublin Core. http://www.lib.washington.edu/msd/mig/datadicts/default.html

Another way to increase interoperability is to support the metadata format and harvesting protocol of the Open Archives Initiative (OAI). Systems that support the OAI *Protocol for Metadata Harvesting* can expose their metadata to harvesters, allowing their metadata to be included in large databases and used by external search services.

- The Open Archives Initiative home page links to the *Protocol for Metadata Harvesting* and guidelines for implementers. http://www.openarchives.org/
- The University of Michigan's OAIster search service contains more than three million records for digitized cultural heritage materials harvested from nearly 300 collections. http://oaister.umdl.umich.edu/o/oaister/

Another way to increase interoperability is to support protocols for cross-system searching, also called “metasearch.” Under this model, the metadata remains in the source repository, but the local search system accepts queries from remote search systems. The best known protocol for cross-system search is the international standard Z39.50 (http://lcweb.loc.gov/z3950/agency/).

Many systems have implemented Z39.50 for metasearch. Two selected examples are:

- Searchlight (http://searchlight.cdlib.org/cgi-bin/searchlight)
- PHAROS (http://pharos.calstate.edu/webpac/pharosstart.html)

Z39.50 International Next Generation (ZING) is an initiative to modernize Z39.50 for the Web environment. Two emerging standards, SRW and SRU, allow Z39.50-like queries and responses to be exchanged using common Internet protocols (http://www.loc.gov/z3950/agency/zing/).

***Metadata principle 3. Good metadata uses authority control and content standards such as controlled vocabularies that are in line with user expectations to describe the content of objects and collocate related objects.***

Attributes of distributed objects should be expressed according to standard controlled vocabularies when possible. These include, but are not limited to, personal names, corporate names, place names, subjects, and genre headings. Classification schemes, a form of controlled vocabulary that groups related resources into a hierarchical structure, can be useful in providing online subject access.

As with metadata schemes, there are many published thesauri, taxonomies, and authority files, and there is no “one size fits all” solution. The choice of vocabularies to use will depend to some extent on factors such as the metadata scheme chosen and the resources of the institution. Authors and other untrained metadata creators can not generally be counted on to use controlled vocabularies successfully unless the authority list is very short and simply organized.

Other important factors include:

- The anticipated users of the digital collection. Will they be adults or children, specialists or generalists? What languages do they speak? What other resources are they likely to use, and what vocabularies are employed in those?
- Tools to support the use of the vocabulary. Is there an online thesaurus? Can it be incorporated into the collection’s search system? Are there cross-references and related terms?
- Maintenance. New terms come into use, and old terms become archaic or obsolete. Who maintains the vocabulary, and how are updates issued?

Whatever vocabularies are chosen, their use should be documented and guidelines should be provided to help metadata creators select terms consistently.

*Controlled vocabularies, thesauri and classification systems available in the WWW* lists several dozen web-accessible controlled vocabularies by subject area (http://www.lub.lu.se/metadata/subject-help.html). The High Level Thesaurus Project (HILT) is a clearinghouse of information about controlled vocabularies, including related resources, projects, and an alphabetical list of thesauri (http://hilt.cdlr.strath.ac.uk/Sources/index.html).

The Getty Vocabulary Program builds, maintains, and disseminates several thesauri for the visual arts and architecture:

- *Art & Architecture Thesaurus* (AAT). http://www.getty.edu/research/conducting_research/vocabularies/aat/
- *Union List of Artist Names* (ULAN). http://www.getty.edu/research/conducting_research/vocabularies/ulan/
- *Getty Thesaurus of Geographic Names* (TGN). http://www.getty.edu/research/conducting_research/vocabularies/tgn/

Some other controlled vocabularies are:

- *Revised nomenclature for museum cataloging: a revised and expanded version of Robert C. Chenhall's system for classifying man-made objects* (Nashville: American Association for State and Local History). Not available on the Web, this resource is used by many small museums and historical societies.
- *Library of Congress Authorities* files. Search subject, name, title, and name/title headings from LC's online catalog. http://authorities.loc.gov/
- *Medical Subject Heading List*. http://www.nlm.nih.gov/mesh/
- *Thesaurus for Graphic Materials I: Subject Terms*. http://lcweb.loc.gov/rr/print/tgm1/
- *Thesaurus for Graphic Materials II: Genre and Physical Characteristics Terms*. http://lcweb.loc.gov/rr/print/tgm2/
- *The Geographic Names Information System*. http://geonames.usgs.gov/

Classification systems available on the Web include:

- *Dewey Decimal Classification*. [Subscription required for access.] http://connexion.oclc.org/
- *Library of Congress Classification*. [Subscription required for access.] http://classweb.loc.gov/

### *Metadata principle 4. Good metadata includes a clear statement on the conditions and terms of use for the digital object.*

Terms and conditions of use include legal rights (e.g. fair use), permissions, and limitations on the exercise of permissions. The user should be informed how to obtain permission for restricted uses and how to cite the material for allowed uses. Special technical requirements, such as the required viewer or reader, should also be noted.

If this information is the same for all the materials in a collection, documenting it in collection-level metadata is adequate (see COLLECTIONS). Otherwise, metadata records for individual objects should contain information pertaining to the particular object. Many metadata schemes have designated places to put this information; if they do not, some locally defined element should be used.

For an example of collection-level copyright information, see the Florida Heritage Collection of the Florida state university libraries at http://palmm.fcla.edu/fh/ and click *Copyright information* on the sidebar.

***Metadata principle 5: Good metadata supports the long-term management of objects in collections.***

Administrative metadata is information intended to facilitate the management of resources. It can include data such as when and how an object was created, who is responsible for controlling access to or archiving the content, what control or processing activities have been performed in relation to it, and what restrictions on access or use apply. Technical metadata, such as capture information, physical format, file size, checksum, sampling frequencies, etc., may be necessary to ensure the continued usability of an object, or to reconstruct a damaged object.

Preservation metadata is a subset of administrative metadata aimed specifically at supporting the long-term retention of digital objects. It may include detailed technical metadata as well as information related to the rights management, record keeping, management history, and change history of the object. It should, therefore, be compatible with the collections management workflow of organizations that manage collections. In some cases, this may require a negotiation to resolve institutional workflow and digital object descriptions.

- An OCLC/RLG Working Group called PREMIS (Preservation Metadata: Implementation Strategies) is working to develop a core set of preservation metadata for release in 2004. Watch the website at http://www.oclc.org/research/projects/pmwg/. A predecessor to the Working Group drafted a state-of-the-art review of preservation metadata element sets and a framework for preservation metadata. http://www.oclc.org/research/projects/pmwg/wg1.htm
- The PADI (Preserving Access to Digital Information) clearinghouse at http://www.nla.gov.au/padi has a long annotated listing of resources related to preservation metadata at http://www.nla.gov.au/padi/topics/32.html.
- NISO Z39.87 (AIIM 20-2002), *Data Dictionary - Technical Metadata for Digital Still Images* is one of the few formal standards for technical metadata. It focuses on images created by scanning. http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

Archivists and records managers are particularly interested in record-keeping metadata.

- *Recordkeeping Metadata Standards*. Defines the metadata recommended by the National Archives of Australia for the recordkeeping systems of Australian government agencies. http://www.naa.gov.au/recordkeeping/control/rkms/summary.htm
- *Australian Recordkeeping Metadata Schema*. An effort to specify the whole range of recordkeeping metadata needed to manage records in digital environments. http://www.sims.monash.edu.au/research/rcrg/research/spirt/

Structural metadata relates the pieces of a compound object together. If a book consists of several page images, it is clearly not enough to preserve the physical image files; information concerning the order of files (page numbering) and how they relate to the logical structure of the book (table of contents) is also required. Three standards for packaging complex digital objects are the *Metadata Encoding and Transmission Standard (METS)*, the *IMS Content Packaging XML Binding*, and the MPEG-21 *Digital Item Declaration Language (DIDL)*. Of these, METS is most widely used in the cultural heritage community. METS is an XML schema that not only specifies how to represent structural metadata for an object, but also provides a framework for associating descriptive and administrative metadata (http://www.loc.gov/standards/mets/).

***Metadata principle 6: Good metadata records are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.***

Because metadata carries information that vouches for the provenance, integrity, and authority of an object, the authority of the metadata itself must be established. “Meta-metadata,” or stored information about the metadata, should include the identification of the institution that created it and what standards of completeness and quality were used in its creation. The institution should

provide sufficient information to allow the user to assess the veracity of the metadata, including how it was created (automatically or manually) and what standards and vocabularies were used.

The Dublin Core Metadata Initiative proposed but never finalized a simple set of data elements for describing metadata records (http://metadata.net/admin/draft-iannella-admin-01.txt). This was unfortunately called the *Administrative Core* (or "A-Core"), generating some confusion with the more prevalent understanding of administrative metadata. However, despite the unfinished and unapproved nature of the work, some implementers have found it useful.

Some metadata schemes include within them sets of metadata elements for describing the metadata records themselves. These include the IEEE LOM (in the section called "meta-metadata"), the EAD (in "eadheader"), and MODS (in “recordInfo”).

The problem of non-authentic and inaccurate metadata is real and serious. Many Internet search engines deliberately avoid using metadata embedded in HTML pages because of pervasive problems with spoofing (one organization supplying misleading metadata for a resource belonging to another organization) and spamming (artificially repeating keywords to boost a page's ranking). The same techniques used to verify the integrity and authenticity of digital documents (e.g. digital signatures) can also be applied to metadata.

Metadata should be documented in a registry that provides standardized information for the definition, identification, and use of each data element. A registry defines metadata characteristics and formatting requirements to ensure that a metadata schema and data elements in use by one organization can be applied consistently within the organization or community, reused by other communities, and interpreted by computer applications and human users. The ISO 11179 metadata registry standard, particularly part 3, *Basic attributes of data elements*, provides for the consistent definition, interpretation, and use of data elements. Core requirements of ISO 11179-3 include: data element name, data element label, data type, data element identifier and version number, repeatability, obligation for use (e.g., mandatory or optional), controlled vocabulary, and the context or information domain of use.

A prototype registry is under development by the Dublin Core Metadata Initiative (http://dublincore.org/groups/registry/index2.shtml). The Moving Image Collections (MIC) project utilizes a simple 11179 registry to document the MIC union catalog core data element set (http://gondolin.rutgers.edu/MIC/text/how/unioncat_registry_table_04_23.htm) and the moving image organizations directory database (http://gondolin.rutgers.edu/MIC/text/how/mic_directory_version_one.htm).

Metadata schemes should also be documented in a syntax that provides guidance and validation for metadata records creation. An XML (eXtensible Markup Language) schema (http://www.w3.org/XML/Schema) provides a structured syntax for indicating whether data elements are mandatory, the sources of the controlled vocabularies or the formatting principles used to create the information that populates the data element, and the relationship between data elements. The XML schema can be used to validate that metadata records have been created properly, which is critical for sharing metadata across organizations and for ensuring that applications that use the organization's metadata, such as search engines, can correctly and consistently interpret and use the metadata. XML Document Type Definitions (DTDs) can also be used to provide standardization and validation of metadata information (http://www.w3schools.com/dtd/default.asp).

## CASE STUDY: METADATA CHOICES

### The New Jersey Digital Highway Project

Preservation, access, and active use of New Jersey historical materials are the goals of the New Jersey Digital Highway (NJDH), a multi-faceted collaboration between Rutgers University, the New Jersey State Library, and other project partners. The project will develop a website featuring state and local history collections from cultural institutions across the state with specialized portals to support teaching and learning activities. Featured topics, such as the site's inaugural theme on the immigration experience, will incorporate such diverse resources as an oral history collection and materials from local PBS affiliate WNET-TV. In addition, the project will develop a digitization and description system that can be used by organizations of varying sizes to preserve and manage digital collections derived from original materials in a variety of formats.

NJDH design placed several constraints on metadata choices. The metadata had to accommodate many digital formats, be compatible with the METS schema, and allow OAI-compliant metadata harvesting from the NJDH site. NJDH Metadata Work Group members wrestled with the task of identifying metadata schemas that would be sufficiently flexible, rich enough to represent a variety of materials to diverse users, and simple to use. To guide their decision they studied the METS schema, observed the choices of developers of projects similar to the NJDH, made contact with MODS users through the MODS listserv, and analyzed the needs of museum collections.

Four schemes were chosen to accommodate the metadata needs of the project. Descriptive metadata will be carried by MODS, a bibliographic schema selected for its rich element set, MARC and OAI compatibility, XML format, and approval as a METS extension schema. Rights information will be conveyed by the draft METS Rights schema and technical metadata will be represented in a "lite" version of NISO's Technical Metadata for Digital Still Images (MIX). Both of these schemas are also in XML format and are METS-endorsed. The needs of museum and archival partners for structured data on source and physical condition that could not be accommodated by MODS led to the choice of the CIDOC framework for this data. A schema for coverage of digital provenance is still to be determined.

Metadata Work Group members are discovering the strengths and drawbacks of their choices. They rejected the use of Dublin Core for descriptive data out of concern that the schema would require significant extension to accommodate the variety of resources anticipated in the project. MODS, METS, and MIX are richer schema, but they are also more complex and less fully developed. As a result, the Work Group invested considerable time in analyzing MIX elements to determine what was necessary for preservation metadata. Further modifications were required to adapt the elements for use with other media such as audio and video resources. Potential project partners are unfamiliar with MODS and some lack the resources to handle complex metadata entry. To address this, additional time is being devoted to the development of work forms and databases to mask the complexity of MODS and make metadata entry easier for non-librarians.

Project developers are learning that an implementation that is not based on existing models and software involves complex coordination and effort shared between many people. Drawing on diverse areas of expertise, the project is encouraging closer working relationships among information technology professionals, library systems librarians, and catalog librarians. In addition, the overall deadlines of the three-year statewide initiative require that metadata development, workflow management, and the digitization itself occur on parallel tracks. As Ruth Bogan of Rutgers University Libraries and chair of the Metadata Work Group comments, work "must go ahead by anticipating the metadata end product without being certain of its final form…We hope our choices of schemas will extend gracefully to these other formats and objects, but we do not know that for sure."

## PROJECTS

Projects to build digital collections can be initiated by users such as scholars or students or by staff of a cultural heritage institution (curators, librarians, archivists). Projects have specific goals and are of finite duration. Project planning includes plans for the future of the digital collection after the project is concluded. Plans for continued access to collections and for some level of collection maintenance should be specified and responsibility for maintenance clearly designated.

Digital collection projects often involve individuals from various disciplines or institutions. In assembling a team, the project manager should consider the qualifications, competency, reliability, and compatibility of potential team members. From the very beginning of the project, the project manager should invest in team building for the team to benefit from the perspectives and background of all of its members.

The project manager should have the ability to recognize when additional expertise is needed. Projects should consider the use of consultants and build this into the project proposals and plans. The project manager should coordinate the work of all project participants and maintain the project plan and timeline. The project manager may report to a higher manager, to a board of directors, or to an advisory board. However, the project manager should have the authority to delegate work, make decisions, and take remedial actions within the parameters set by the higher agency.

The project manager should have a plan for keeping stakeholders informed of the progress of the project to demonstrate accountability, generate interest in the digital collections and ensure continuing support for the project. Promotion of the project and the digital collections should be considered at the project planning stage.

### *Projects principle 1: A good collection-building project has a substantial design and planning component.*

Project planning is crucial to the completion and success of a project. It encompasses all aspects of the project, from processing workflow to the ultimate look and feel of the collection website. Very early in the project, planners should specify the targeted audience for the digital collection and perform a needs assessment to ascertain the functional requirements of these users. After that, a written project plan can be prepared that covers all significant aspects of the project: short and long-term goals and objectives, project constraints (e.g. time, resources, political factors), selection, digitization, copyright issues, metadata and access, maintenance, dissemination, and evaluation.

Resources on needs assessment:

- The University of Arizona Library has an online needs assessment tutorial with examples from library projects, although not related to digital collections. http://digital.library.arizona.edu/nadm/tutorial/index.htm
- The Washington State Library *Digital Best Practices* site has a section on *Project Management* with a focus on market research as a tool for both design and promotion. http://digitalwa.statelib.wa.gov/newsite/projectmgmt/index.htm
- The Colorado Digitization Alliance provides an example of the identification of market segments and their varying needs. http://www.cdpheritage.org/resource/reports/rsrc_users.html

Guides to digitization project planning:

- IMLS. *NLG [National Leadership Grant] Project Planning: A tutorial*. Although specifically aimed at NLG applicants, this is a generally useful tutorial that includes the grant-writing stage. http://e-services.imls.gov/project_planning/
- *RLG/DLF Guides to Quality in Visual Resource Imaging: 1. Planning an Imaging Project*. http://www.rlg.org/visguides/visguide1.html

- Northeast Document Conservation Center. *Handbook for Digital Projects: A Management Tool for Preservation & Access. III: Considerations for Project Management*. http://www.nedcc.org/digital/dighome.htm
- *NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*, Chapter II, *Project Planning*, (2002). http://www.nyu.edu/its/humanities/ninchguide/

### *Projects principle 2: A good project has an evaluation plan.*

An evaluation plan demonstrates the commitment of a project to its stated goals and objectives. Evaluation can focus on the process and the outcome. Evaluation of process can involve assessment of a project's operations such as staffing and management, workflow, and procedures. Recent emphasis is on outcome-based evaluation. The goals and objectives of a project should help the project team specify desirable outcomes.

Products, services, and impact of a project are likely targets for evaluation. In assessing the product of a digital collection building project, evaluators may measure the digital collection's size, quality, and usage. In addition, the usability of the collection's website, users' experience with the collection and the service, and the impact of the collection on users are also good indicators of a project's success. Surveys, focus groups, interviews, transaction logs, and case studies can be used to collect data for analysis.

- The IMLS encourages outcomes-based evaluation for their funded projects and points to supporting resources. http://www.imls.gov/grants/current/crnt_obe.htm
- *Basic Guide to Outcomes-Based Evaluation for Nonprofit Organizations with Very Limited Resources* is aimed at all non-profit organizations. http://www.mapnp.org/library/evaluatn/outcomes.htm
- *Building Better Websites: Evaluative Techniques for Library and Museum Websites* was developed by the University of Texas with an IMLS grant. http://www.lib.utexas.edu/dlp/imls/index.html
- Thomas R. Bruce and Diane I. Hillmann. *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*. In *Metadata in Practice* (ALA Editions, 2004). Discusses completeness, provenance, accuracy, conformance to expectations, logical consistency or coherence, timeliness, and accessibility as measures of metadata quality.

### *Projects principle 3: A good project produces a project report and broadly disseminates information about the project process and outcomes.*

A good project will provide documentation to make its processes and outcomes clear to readers. If the project produces any models, tools, or prototypes, they should be made readily available to the public to encourage adaptation. If a project has local, regional, or national impact, that impact should be reported through publication, presentation, media, and other channels.

The primary goal of any project should be to accomplish its stated objectives within the time and budget allowed. However, the knowledge gained in implementing a digital collection should not be lost to other organizations. Although most funding agencies require some sort of report at the end of the project period, these are not always generally available. A project report providing a detailed description and honest assessment of work accomplished should be produced and remain accessible on the Web indefinitely.

Some examples of useful, comprehensive project reports:

- Library of Congress. *Manuscript Digitization Demonstration Project. Final Report.* October 1998. http://lcweb2.loc.gov/ammem/pictel/
- *Final Report, 2002/2003 LSTA/NC ECHO Digitization Grant: A Cultural History of the Physician Assistant Profession*. Less formal but includes sections such as “Positive or

negative surprises" and "What would we do differently?"
http://archives.mc.duke.edu/pahx/report/

- *Preserving and Digitizing Plant Images: Linking Plant Images and Databases for Public Access*. November 2000. Final report from the Missouri Botanical Garden to the IMLS. http://ridgwaydb.mobot.org/mobot/imls/final.asp

***Project principle 4: A good project considers the entire lifecycle of the digital collection and associated services developed through the project.***

The staff, equipment, software, and level of effort required to plan and develop a digital collection are generally very different from that required for long-term management and sustainability. Planning efforts should include projecting the use of the collection over time and the amount of updating of both the collection and the project website that will be required. There should also be a plan for maintaining master objects to ensure persistence over time, and for evaluating their continued quality. Objects, regardless of storage medium, should be periodically checked for accessibility and usability.

During project implementation special staff positions are often needed including project managers, metadata creators, and digitization technicians. Completed collections, or collections that grow steadily and incrementally over time, should ideally be subsumed into ongoing workflow within the organization. It is critical to estimate and plan for the level of ongoing effort.

A good digital project should result in collections and services that become important and trusted parts of the organization's information repertoire and must therefore be maintained to the same standards that the organization has set for its other collections and services.

### CASE STUDY: OVERVIEW FOR PROJECT PLANNING

**The University of Oregon—Museum of Natural History and University Library—The Don Hunter Archive Project**

A collaborative effort between the University of Oregon Museum of Natural History and the University Library, the Don Hunter Archive Project will preserve and provide access to slide and sound presentations of local photographer and audio archivist, Don Hunter. Project planners are consulting multiple resources to formulate their digitization plan because the project's source materials exist in a variety of media, including video, for which conversion guidelines are less well-developed.

Project planners consulted *Moving Theory into Practice: Digital Imaging for Libraries and Archives* as they were developing their initial proposal. The combination of overview and detailed information about digital imaging and digital preservation presented in this resource helped them understand and explain the processes that are involved in carrying out a digital project. With this knowledge, they were able to develop a proposal that clearly articulated for their institution the scope and budget implications of what was required for the project.

Among the resources they are consulting are *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials and Principles and Specifications for Preservation Digital Reformatting* from the Library of Congress (http://www.loc.gov/preserv/prd/presdig/presprinciple.html). The NINCH document's outline of potential issues in the capture and storage of image, audio, and video materials is proving to be a useful resource for the development of project-specific guidelines for the restructuring of multimedia analog resources to DVD and digital storage. In addition, the document provides significant resources for decision-making related to achieving a balance between protection of intellectual property and fair use access to digital collections. The Library of Congress document offers an overview of an authoritative and time-tested system as well as specific digital archives guidelines and procedures.


ISBN: 1880124-64-5