NISO
How the information world CONNECTS

A Framework of Guidance for Building Good Digital Collections

3rd edition
December 2007

A NISO Recommended Practice

Prepared by the
NISO Framework Working Group
with support from the
Institute of Museum and Library Services

INSTITUTE of Museum and Library SERVICES

About NISO Recommended Practices

A NISO Recommended Practice is a recommended "best practice" or "guideline" for methods, materials, or practices in order to give guidance to the user. Such documents usually represent a leading edge, exceptional model, or proven industry practice. All elements of Recommended Practices are discretionary and may be used as stated or modified by the user to meet specific needs.

This recommended practice may be revised or withdrawn at any time. For current information on the status of this publication contact the NISO office or visit the NISO website (www.niso.org).

Published by
National Information Standards Organization (NISO)
One North Charles Street, Suite 1905
Baltimore, MD 21201
www.niso.org

Printed in the United States of America
ISBN (10): 1-880124-74-2
ISBN (13): 978-1-880124-74-1

National Information Standards Organization
Baltimore, Maryland

CONTENTS

FOREWORD

The 3rd edition of *A Framework of Guidance for Building Good Digital Collections* was produced by the National Information Standards Organization (NISO) Framework Working Group, with the generous support of the Institute of Museum and Library Services (IMLS). Working Group members are:

Grace Agnew, Rutgers University
Murtha Baca, Getty Research Institute
Priscilla Caplan (Chair), Florida Center for Library Automation
Carl Fleischhauer, Library of Congress
Tony Gill, Center for Jewish History
Ingrid Hsieh-Yee, Catholic University
Jill Koelling, Northern Arizona University
Christie Stephenson, American Museum
Karen A. Wetzel, NISO liaison

The Working Group is grateful to the following individuals for taking time to read the review draft and offer their very helpful and often extensive comments and suggestions. The *Framework of Guidance* document is improved immeasurably by their review.

Stuart Dempster, Director, Strategic e-Content Alliance, JISC
Jane Greenberg, Associate Professor, School of Information and Library Science, University of North Carolina-Chapel Hill
Martin R. Kalfatovic, Head New Media Office, Smithsonian Institution Libraries
Elizabeth O'Keefe, Director of Collection Information Systems, Morgan Library & Museum
Matthias Razum, Head ePublishing and eScience, FIZ Karlsruhe
Matthew Walker, Digital Collections Architect, National Library of Australia
Maureen Whalen, Senior Counsel, Intellectual Property & Alliances, J. Paul Getty Trust
Karla Youngs, Director, Technical Advisory Service for Images
Marcia Zeng, Professor, School of Library and Information Science, Kent State University

The 2nd (2004) edition of the *Framework of Guidance for Building Good Digital Collections* 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 Santa Fe Community College

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 *A Framework of Guidance for Building Good Digital Collections* 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.

INTRODUCTION

This *Framework of Guidance for Building Good Digital Collections* has three purposes:

1. To provide an overview of some of the major components and activities involved in creating good digital collections.
2. To identify existing resources that support the development of sound local practices for creating and managing good digital collections.
3. To encourage community participation in the ongoing development of best practices for digital collection building.

It is intended for two audiences:

- cultural heritage organizations planning and implementing initiatives 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, a collection could be considered good if it provided proof of concept or resulted in new institutional capabilities—even if the resulting collection itself was short-lived or of minimal usefulness to the organization's users.

As the digital environment matured, the focus of digital collection-building efforts shifted toward the creation of useful and relevant 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 evolved and matured to a third stage, where simply serving useful digital collections effectively to a known constituency is not sufficient. Issues of cost/value, sustainability, and trust have emerged as critical success criteria for good digital collections. 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, repurpose, and build services upon. "Goodness" now demands interoperability, reusability, persistence, verification, documentation, and support for intellectual property rights.

In edition three of this Framework we acknowledge that digital collections increasingly contain born-digital objects, as opposed to digital objects that were derived through the digitization of analogue source materials. We also acknowledge that digital collection development has moved from being an ad hoc "extra" activity to a core service in many cultural heritage institutions.

Digital collections must now intersect with the user's own context—within the course, within the research process, within the leisure time activities, and within the social networks that are important to the end user.

Users—in particular the younger generations of users—have integrated digital technologies so completely into their lives that they are ready and even eager to move into a role as creators and collaborators. The rise of shared information spaces such as YouTube and Flickr; the popularity of social networking sites such as MySpace, Facebook, and LinkedIn; and the growth of the "mash-up" as the vehicle for new creativity demonstrate that good digital collection-building has become an active collaboration between the information professional and the user, resulting in collections that are reliable and authoritative, yet also compelling and useful to a wide range of users wherever they live, work, and play.

The *Framework of Guidance* provides criteria for goodness organized around four core types of entities:

- Collections (organized groups of objects)
- Objects (digital materials)
- Metadata (information about objects and collections)
- Initiatives (programs or projects to create and 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 and reuse of them.

For each of these four types of 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, or examples.

How to Contribute
Every effort has been made to select resources that are useful and current, and to provide helpful annotations. However, the list of resources is not exhaustive and, given the dynamic nature of the digital information environment, can be expected to change rapidly over time.

With the third edition of the Framework, we open the document up for ongoing contributions from the community of librarians, archivists, curators, and other information professionals. We encourage you to contribute your own ideas and experiences, suggest resources, and evaluate those that have been suggested.

Please see the Community Version on the Web at:
http://purl.fcla.edu/fcla/NISOCommunityFramework

How to Use
There are no absolute rules for creating good digital collections. Every digital collection-building initiative is unique, with its own users, goals, and needs. Initiatives dealing with legacy collections, for example, have different constraints than projects embarking on new digitization efforts, which in turn have different constraints than projects building collections of born-digital materials. Museums, libraries, archives, and schools have different constituencies, priorities, institutional cultures, funding mechanisms, and governance structures.

The key to a successful project is not to strictly and unquestioningly 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.

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:

- UKOLN, *Good Practice Guide for Developers of Cultural Heritage Web Services* (2006) http://www.ukoln.ac.uk/interop-focus/gpg/.
- Anne R. Kenney and Oya Y. Rieger, *Moving Theory into Practice: Digital Imaging for Libraries and Archives* (2000) http://www.library.cornell.edu/preservation/tutorial/. An online tutorial of imaging basics in English, French and Spanish.
- Northeast Document Conservation Center, *Handbook for Digital Projects: A Management Tool for Preservation & Access* (2000) http://nedcc.org/oldnedccsite/digital/dighome.htm.
- Arts and Humanities Data Service (AHDS), *Guides to Good Practice* website http://www.ahds.ac.uk/creating/guides/index.htm. A series of guides to covering collection, description, and digitization for specific types of materials, such as GIS, performance resources, and virtual reality.
- Washington State Library, *Digital Best Practices* website http://digitalwa.statelib.wa.gov/newsite/best.htm.
- Susan Schreibman (editor), *Best Practice Guidelines for Digital Collections at University of Maryland Libraries*, 2nd ed. (2007) http://www.lib.umd.edu/dcr/publications/best_practice.pdf.

COLLECTIONS

A digital collection consists of digital objects that are selected and organized to facilitate their discovery, access, and use. Objects, metadata, and the user interface together create the user experience of a collection.

Principles that apply to good digital collections are:

Collections Principle 1: A good digital collection is created according to an explicit collection development policy.

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.

Collections Principle 3: A good collection is curated, which is to say, its resources are actively managed during their entire lifecycle.

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.

Collections Principle 5: A good collection respects intellectual property rights.

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

Collections Principle 7: A good collection is interoperable.

Collections Principle 8: A good collection integrates into the users own workflow.

Collections Principle 9: A good collection is sustainable over time.

COLLECTIONS PRINCIPLE 1

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

Of all factors, collection development is most closely tied to an organization's own goals and constituencies. Collection builders should be able to refer to the mission statement of their organization and articulate how a proposed collection furthers or supports that mission. The institution should be able to identify the target audience(s) for the collection but also think about unexpected uses and users. If the institution collects print, artifacts or other non-digital materials, the digital collection should fit in with the organization's overall collection policy.

There are a few cases where a selection policy may not be required: digitization on demand, when an organization is creating digital content based on end-user requests, and mass digitization programs, which are often indiscriminate. Even these efforts require planning and should follow principles for building good collections as appropriate. Disciplinary or institutional repositories that encourage users to deposit their own intellectual property present an interesting case. These still benefit from a published collection policy, but it may have to be fairly flexible in acknowledgement that the users may be the best judges of relevance.

The following documents are general guidelines for selecting materials for digitization.

- Technical Advisory Service for Images (TASI), *Selection and Preparation of Materials* (2003) http://www.tasi.ac.uk/advice/creating/selection.html.
- Technical Advisory Service for Images (TASI), *Selection Procedures* (2003) http://www.tasi.ac.uk/advice/creating/selecpro.html.
- Northeast Document Conservation Center, *Handbook for Digital Projects*, chapter IV: Selection of Materials for Scanning (2000) http://nedcc.org/oldnedccsite/digital/iv.htm.
- Joint RLG and NPO Preservation Conference, *Guidelines for Digital Imaging: Guidance for selecting materials for digitisation* (1999) http://eprints.ucl.ac.uk/archive/00000492/01/paul_ayris3.pdf.
- Anne R. Kenney and Oya Y. Rieger, *Moving Theory into Practice*: Chapter 2, Selection (2000) http://www.library.cornell.edu/preservation/tutorial/selection/selection-01.html. A short, general guide with pointers to library selection policies and a bibliography.
- California Digital Library, *Collection Development Framework* website http://www.cdlib.org/inside/collect/framework.html. Covers both commercially licensed and locally digitized resources.

Local policies on selecting materials for digitization:

- Columbia University Libraries, *Selection Criteria for Digital Imaging* (2001) http://www.columbia.edu/cu/libraries/digital/criteria.html.

- University of California Libraries, *Selection Criteria for Digitization* (2005) http://libraries.universityofcalifornia.edu/cdc/pag/digselec.html.

National library policies for digitized and born digital materials:

- Digital National Library of Scotland, *Strategic Plan 2005-2008* (2005) http://www.nls.uk/professional/policy/docs/nls_digital_library_strategy.pdf.
- Library and Archives Canada, *Digital Collection Development Policy* (2006) http://www.collectionscanada.ca/collection/003-200-e.html.
- Library of Congress, *Electronic Resources Selection Guidelines* (2004) http://www.loc.gov/acq/devpol/electronicselectionguidelines.html.
- National Library of Australia, *Collection Digitisation Policy* (2006) http://www.nla.gov.au/policy/digitisation.html.

Criteria for inclusion in portals:

- North Carolina ECHO (Exploring Cultural Heritage Online), *Portal Collection Development Policy* (2000) http://www.ncecho.org/colldev.asp.
- Digital Library for Earth System Education (DLESE), *Collection Scope and Policy Statement* (2004) http://www.dlese.org/documents/policy/CollectionsScope_final.html. Selection for inclusion in a topical library of learning objects.
- National Science Digital Library, *Collections Policy* http://nsdl.org/about/?pager=collection_policy. Selection policy for a collection of collections.
- New Jersey Digital Highway, *Collection Development Policy* (2004) http://www.njdigitalhighway.org/documents/njdh-coll-dev-policy.pdf. Selection for a statewide collaborative including libraries, museums, archives, and other cultural heritage organizations.

Selecting materials for digitization specifically for preservation purposes:

- Library of Congress Preservation Reformatting Division, *Selection Criteria for Preservation Digital Reformatting* http://lcweb.loc.gov/preserv/prd/presdig/presselection.html.
- National Library of Medicine, *Selection Criteria for Digital Reformatting* (2006) http://www.nlm.nih.gov/psd/pcm/digitizationcriteria.pdf.

Selecting born-digital content for preservation:

- Mary Ide and Leah Weisse, *Recommended Appraisal Guidelines for Selecting Born-digital [Television] Master Programs For Preservation and Deposit with the Library of Congress* (2006) http://www.ptvdigitalarchive.org/docs/Selection/Recommended Appraisal Guidelines.pdf.

COLLECTIONS PRINCIPLE 2

> **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.

Collection descriptions should help users understand the nature and scope of the collection and any restrictions that apply to the use of materials within it. It is good practice to incorporate a narrative description of the collection, description of the scope and extent of the collection, names and contacts for the organization(s) responsible for building and maintaining the collection (as organizational provenance is an important clue to the authenticity and authority), 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. Many project planners find a description of the methodologies, software applications, record formats, and metadata schemes used in building other collections helpful.

There is no dominant metadata standard for describing collections, although in the last few years there has been substantial progress towards this goal.

- IMLS, *Digital Collections and Content: Resources* website http://imlsdcc.grainger.uiuc.edu/resources.asp. Discusses the benefits of collection level description and gives examples of collection description schema.
- Research Support Libraries Programme, *RSLP Collection Description* website http://www.ukoln.ac.uk/metadata/rslp/. An early effort to develop a standard collection description schema.
- *NISO Z39.91 Collection Description Specification* (2005) http://www.niso.org/standards/resources/Z39-91-DSFTU.pdf. A draft standard that builds on the RSLP effort and work in the Dublin Core community, developed by the NISO MetaSearch Initiative (http://www.niso.org/committees/MS_initiative.html).
- *UKOLN Collection Level Description* (2001) http://www.ukoln.ac.uk/metadata/cld/.

Archival description can also be thought of as a form of collection description.

- *ISAD(G): General International Standard Archival Description* (2000) http://www.ica.org/en/node/30000. Set of general rules for archival description developed by the International Council on Archives.
- *Encoded Archival Description (EAD)* website http://www.loc.gov/ead/. The EAD provides an XML representation of archival finding aids.

Good examples of collection-level terms and conditions of use:

- Library of Congress, *Prints and Photographs Online Catalog* website http://www.loc.gov/rr/print/catalog.html.
- National Maritime Museum (U.K.) Search Station, *Copyright Notice, Disclaimer And Terms Of Use* http://www.nmm.ac.uk/searchbin/searchs.pl?return=copyright.
- AdAccess Project, *Copyright and Citation Information* (1999) http://scriptorium.lib.duke.edu/adaccess/copyright.html.

Examples of websites with informative information about the collection and/or project:

- *Historic Pittsburgh* website http://digital.library.pitt.edu/pittsburgh/.
- *Yiddish Children's Books* website http://palmm.fcla.edu/ycb/index.shtml. This site follows the PALMM (Publication of Archival, Library and Museum Materials) program's standard template for "sidebar" information with links such as "About the Collection," "Technical Aspects," "Related Sites," etc. (http://palmm.fcla.edu/strucmeta/guidelines.pdf).
- *Histpop: The Online Historical Population Reports website, Project Histpop*website http://histpop.org/ohpr/servlet/Category?page=Project&path=Project&active=yes&treestate=expandnew.

When possible, collections should be described in collection-level cataloging records contributed to a union catalog such as OCLC's WorldCat (http://www.oclc.org/worldcat/).

The registries listed below allow institutions to register their own collections, or to propose their collections for registration. Unfortunately most registries appear to be poorly maintained.

- Digital Library Federation, *Digital Collections Registry* website http://dlf.grainger.uiuc.edu/DLFCollectionsRegistry/browse/.
- Smithsonian Institution Libraries, *Library and Archival Exhibitions on the Web* website, http://www.sil.si.edu/SILPublications/Online-Exhibitions/. Web exhibitions only.
- *UNESCO/IFLA Directory of Digitized Collections* website http://www.unesco.org/webworld/digicol/. Particularly useful for the international focus.
- IMLS, *Digital Collections and Content* website http://imlsdcc.grainger.uiuc.edu/collections/GemTopPlusSubs.asp. Registry of all digital collections built with IMLS funds.
- *Imagelib and the Clearinghouse of Image Databases* website http://elearn.arizona.edu/imagelib/.
- Technical Advisory Service for Images (TASI) *Image Sites* website http://www.tasi.ac.uk/imagesites/index.php. Actively maintained.

COLLECTIONS PRINCIPLE 3

> **Collections Principle 3**: A good collection is curated, which is to say, its resources are actively managed during their entire lifecycle.

Digital curation is concerned with the lifecycle management of a resource from the time it is created or obtained until it is purposely disposed of. Curation encompasses a set of activities that include active data management, archiving, and digital preservation.

Active data management is required to ensure that objects in a collection can be used and reused over time. It can include creating, correcting, and enhancing metadata; correcting or enhancing the data itself; and adding annotations, linkages to other materials, or other enriching information. It can involve working with the creators of the digital objects to ensure they are appropriately transferred to the custody of the curator, and appropriately described and documented.

- *Digital Curation Centre* website http://www.dcc.ac.uk/. The U.K.'s Digital Curation Center promotes digital curation by sponsoring events like workshops and conferences and collecting or commissioning publications and tools. Their website links to a wealth of information, much of it focused on active data management. The DCC is also publishing a comprehensive Curation Manual http://www.dcc.ac.uk/resource/curation-manual/in a series of installments. More than 45 chapters have been commissioned so far, covering a wide range of topics from appraisal and selection to technological obsolescence. Although only a handful of chapters have been published so far, this is likely to become a definitive resource on digital curation.
- Philip Lord and Alison Macdonald, *E-Science Curation Report* (2003) http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf. Details the requirements of data curation in the sciences and database-intensive social science and humanities disciplines.

There are industry standard practices applicable to all mission-critical data and are not specific to digital collections. Data center and IT staff should be aware of these good general resources:

- ISO/IEC 27002:2005, *Information technology – Security techniques – Code of practice for information security management* (June 2005)http://en.wikipedia.org/wiki/ISO/IEC_27002. This Wikipedia article describes the standard and links to purchase information.
- *Web Application Security Consortium* website http://www.webappsec.org/. The Consortium produces and releases technical information, articles, guidelines, and documentation for best practice security standards.

Capture of born-digital materials can present special challenges, particularly ephemeral materials and works with distributed authorship such as websites and emails. Many academic institutions have established institutional repositories for content generated by students, faculty, and staff, but it is difficult to convince authors to deposit their own materials.

- *Creating an Institutional Repository: LEADIRS Workbook* (2004) http://www.dspace.org/implement/leadirs.pdf. Covers all angles of planning, policy and implementation. Written by MIT for a British audience.
- *Australian Partnership for Sustainable Repositories (APSR)* website http://www.apsr.edu.au/. APSR supports the implementation and use of institutional repositories at universities in Australia, and promotes linkages among them.

While some usages equate digital preservation with archiving, preservation is more properly thought of as that subset of archiving concerned with the application of active preservation strategies to ensure an object remains usable despite hardware and software obsolescence. Preservation strategies generally involve format transformation, hardware/software emulation, or combinations of the two. The long-term archiving and preservation of digital materials is a difficult and expensive undertaking that requires substantial resources and serious institutional commitment. Resources are now available that continue to move the discussions forward toward best practice for preservation of digital content.

- *Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist* (2007) http://www.crl.edu/PDF/trac.pdf. These metrics will likely become the basis on an international standard for assessing trustworthy digital repositories.
- *Digital Repository Audit Method Based on Risk Assessment (DRAMBURA)* website, http://www.repositoryaudit.eu/. Toolkit and supporting tutorials are designed to help a repository do a self-audit against the TRAC criteria.
- *PREMIS Preservation Metadata Maintenance Activity* website, http://www.loc.gov/standards/premis/. Includes a Data Dictionary for preservation metadata, supporting materials and a forum for the PREMIS Implementors' Group.
- National Library of Australia, *Preserving Access to Digital Information (PADI)* website http://www.nla.gov.au/padi/. Comprehensive clearinghouse of current and historical materials related to digital preservation and curation.

COLLECTIONS PRINCIPLE 4

> **Collections Principle 4**: A good collection is broadly available and avoids unnecessary impediments to use.

This principle encompasses three attributes: availability, usability, and accessibility.

Availability means that the collection is accessible and usable upon demand by an authorized person. This implies that collections should be accessible through the Web, using technologies that are well known among the target user community. They should be "up" as close to 24/7 as possible, which has implications for system security and maintenance. Availability does not require that use of all materials be free and unrestricted; charging for use and limiting access may be appropriate and even necessary in some circumstances. But it does require an attempt to make the materials as widely available as possible within any required constraints.

- American Library Association, *Principles for Digital Content* (2007) http://www.ala.org/ala/washoff/oitp/Principlesfinalfinal.pdf. These recently adopted principles emphasize commitment to equitable access.

Usability refers to ease of use. 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 web pages and search forms providing access to the collection, as well as the metadata and digital object displays, should be tested against various browsers 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.

- *Usability.gov* website http://www.usability.gov/. An excellent source of information on usability and user-centered design for websites and other communication systems.
- U.S. Department of Health and Human Services, *Research-Based Web Design & Usability Guidelines – Current Research-Based Guidelines on Web Design and Usability Issues* (2006) http://www.usability.gov/pdfs/guidelines.html.
- Technical Advisory Service for Images (TASI), *Developing Effective Interfaces for Online Image Collections* (2006) http://www.tasi.ac.uk/advice/delivering/interfaces.html.
- Technical Advisory Service for Images (TASI), *Developing Usable and Accessible Interfaces for Online Image Collections* (2006) http://www.tasi.ac.uk/advice/delivering/usability.html.

Accessibility is the property of being usable by people with disabilities. Collection interfaces should be designed to maximize usability for people with visual impairments, loss of hearing, loss of mobility (for example, trouble using a mouse) and even cognitive impairments.

Legislation and de facto standards define web accessibility:

- World Wide Web Consortium (W3C), *Web Accessibility Initiative (WAI)* website http://www.w3.org/WAI/. The most important single site for accessibility issues. Includes links to W3C accessibility standards.
- W3C Web Accesssibility Initiative, *Policies Relating to Web Accessibility* website http://www.w3.org/WAI/Policy/. Links to accessibility legislation in 17 countries plus the United Kingdom and European Union.

Several clearinghouses focus on web accessibility, among them:

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

There is a large body of literature on accessible web design:

- Utah State University Center for Persons with Disabilities, *WebAIM* (Web Accessibility in Mind) website http://www.webaim.org/. An excellent introduction to Web accessibility issues and evaluation tools.
- Massachusetts Institute of Technology, *Adaptive Technology for Information and Computing* website http://web.mit.edu/atic/www/accessibility/index.html. Shows how accessibility guidelines can be applied in an institutional context.
- *Audio Illinois* website http://www.alsaudioillinois.net/. A model site using audio narration to describe pictures for the sight impaired.

COLLECTIONS PRINCIPLE 5

Collections Principle 5: A good collection respects intellectual property rights.

The collection development policy should reference the organization's copyright policy and/or incorporate principles of support for copyright and the copyright status of the organization's collections.

Intellectual property rights 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/or make it available; and
- what rights or permissions the users of the digital collection are given, to make subsequent use of the materials.

Rights management is facilitated by good recordkeeping. Collection managers should maintain a consistent record of rights holders (including contact information when possible) and permissions granted for all applicable materials.

Rights management is complicated by the fact that a work may include contributions from many creators. The underlying rights of complex multimedia works can be challenging to untangle and can involve contract law as well as copyright.

Many useful collections lack a deed of gift that clearly permits the digital distribution of resources. When resources have uncertain provenance, current best practices suggest a practical approach:

- actively solicit information about the creator—from the donors or their heirs or from the audience that utilizes the digital collection; and
- develop a risk management strategy that balances the educational value of the collections against principles of fair use, the potential commercial exploitation of the collection, and the organization's ability to identify and solicit permissions from copyright holders.

A risk assessment will enable the library to make practical, defensible choices among collections of uncertain provenance—a critical concern for digital collection building. 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. A policy statement, posted prominently on the web portal to the collection, can articulate the organization's reasons for making works of uncertain provenance available in digital form to a wide audience.

Rights in the International Arena
It is important to realize that intellectual property rights are an international protection that is governed by treaty. Current intellectual property law derives from international treaty. As

nations sign and ratify an intellectual property rights treaty, they must develop laws to reflect the provisions of the treaty.

Intellectual property rights treaties provide the minimum requirements to which all signatory states must adhere. Treaties generally provide guidance to signatory states on the areas where member states have flexibility to customize their laws to reflect regional and local needs.

The fundamental treaties governing international copyright are:

> Berne Convention (Paris, 1971)
> The primary international copyright treaty is the *Berne Convention for the Protection of Literary and Artistic Works*. This treaty provides minimum enforceable standards intended to harmonize across the laws and standards of its signatory countries. The most significant minimum standard is the term of protection granted to a work: life of the author plus fifty years.
>
> WIPO Copyright Treaty (WCT) (Geneva, 1996)
> The WCT is intended to bring copyright into the digital era and is the genesis of many provisions of the *Digital Millennium Copyright Act*, the U.S. copyright law enacted to enable compliance with the WCT. Among other provisions, the WCT recognizes computer programs as copyright protected literary works, as well as providing copyright status for compilations of data that exhibit creativity in the arrangement or selection of the data.
>
> WIPO Performances and Phonograms Treaty (WPPT) (Geneva 1996)
> The WPPT updates the *Rome Convention of 1961*, particularly describing the rights of performers and producers for authorizing the fixing of performances and phonograms and making them available to the public by wired or wireless means.

The WCT and WPPT are noteworthy for two controversial provisions. The treaties require signatory countries to provide legal protection and remedies against the circumvention of technical measures that protect the exercise of authors' rights or that restrict acts of use that are not authorized by authors or permitted under law. This provision represents a critical change in the treatment of infringement by providing for the prevention of infringing uses of a work. Previously, the Berne treaty provided for legal remedies against infringing use after the fact within the country of origin for the work. For the first time, the WCT and WPPT provide authors with the legal ability to technically prohibit or restrict infringing uses of a work in advance of such use, and legally bind the signatory countries to protect against the circumvention of these technological measures.

The WCT and WPPT also require that signatory countries provide legal remedies against any party that knowingly removes or alters rights management information, where this information is defined as "information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information is

attached to a copy of a work or appears in connection with the communication of a work to the public." (WIPO Copyright Treaty, *art. 12*)

Legally enforced metadata is a new concept for metadata that is discussed further in Metadata Principle #4.

Four key issues that are important to understand about international copyright are:

1. Copyright is viewed by treaty as a "natural right" that obtains to a work as soon as it is fixed in tangible form. No country requires registration for copyright to take effect. The United States is notable for applying copyright requirements on works based on the law in effect when the work was first fixed in tangible form. Thus some works require copyright registration or renewal or copyright notice conforming to specifications. Most nations apply the current provisions of copyright law to all existing works.

2. International copyright originates in treaty, which establishes minimum standards for compliance. These minimum standards may only be referenced in national law, so when dealing with works of an international character or that will be distributed internationally, a basic understanding of at least the three treaties referenced above is important.

3. Copyrighted works are accorded "national treatment" by all the signatory member states of a treaty, which means that a copyright holder who is a foreign national is accorded the same treatment, with respect to copyright and copyright infringement within the member state, as the citizens of that member state.

4. Every country may make exceptions to copyright to further the common good of its citizens. These exceptions, generally called fair use or fair dealing, differ markedly across countries but must meet the minimum conditions imposed by treaty. They:

    - must represent special cases rather than the normal use of resources,
    - must not conflict with "normal exploitation of the work," and
    - must not "unreasonably prejudice the legitimate interests of the author."

The best source for information about international copyright is the World Intellectual Property Organization, which has oversight for intellectual property treaties: http://www.wipo.int/portal/index.html.en.

The IFLA Committee on Copyright and Other Legal Matters (CLM) provides news and information by country as well as background papers, etc. http://www.ifla.org/III/clm/copyr.htm.

Most countries have government agencies tasked with intellectual property right management. Many countries also have non-governmental organizations that advise on copyright. The national library or national library association is a good place to find the relevant agency, as

well as a good copyright bibliography, if you are unfamiliar with the authoritative copyright agency or organization for a specific country.

"How-to" guides for digital collections:

- New Jersey Digital Highway, *Copyright Issues for Digital Collections* website http://www.njdigitalhighway.org/copyright_issues_libr.php.
- Washington State Library, *Digital Best Practices, Rights and Permissions to Publish Digital Collections* website http://digitalwa.statelib.wa.gov/newsite/collection/rights.htm.
- Mary Minow, *Library Digitization Projects and Copyright* (2002) http://www.llrx.com/features/digitization.htm. Comprehensive and entertainingly presented.

Helpful general publications on copyright:

- Georgia Harper, *Copyright Crash Course* website http://www.utsystem.edu/ogc/intellectualproperty/cprtindx.htm. A general introduction to virtually all copyright-related issues. There is a useful section on the logistics of obtaining permission that takes the perspective of risk vs. benefit.
- Canadian Heritage Information Network (CHIN), *Copyright Guide for Museums and Other Cultural Organizations* website http://www.chin.gc.ca/English/Intellectual_Property/Copyright_Guide/index.html.
- Peter Hirtle, *Copyright Term and the Public Domain in the United States* (2007) http://www.copyright.cornell.edu/training/Hirtle_Public_Domain.htm. A handy (possibly indispensable) chart for quick lookup of likely status by the date of publication.
- JISC Legal Information Service, *Intellectual Property Rights Overview* (2006) http://www.jisclegal.ac.uk/ipr/IntellectualProperty.htm. Covers "the landscape of copyright law and its application to Further and Higher Education."
- Electronic Information for Libraries (eIFL), *Handbook on Copyright and Related Issues for Libraries* (2006) http://www.eifl.net/cps/sections/services/eifl-ip/issues/eifl-handbook-on.

Particular material types:

- Daniel I. Cohen and Roy Rosenzweig, *Digital History: A Guide to Gathering, Preserving and Presenting the Past on the Web*, Chapter "Owning the Past: Images, Music and Movies" (2005) http://chnm.gmu.edu/digitalhistory/copyright/6.php. Digital multimedia.
- Technical Advisory Service for Images (TASI), *Copyright FAQ* (2006) http://www.tasi.ac.uk/advice/managing/copyright_faq.html. Digital images.
- AHDS, *Creating Digital Audio Resources: A Guide to Good Practice*, Chapter 2 "Working with Copyright" http://ahds.ac.uk/creating/guides/audio-resources/GGP_Audio_2.1.htm. Digitizing audio.

Clearinghouses on law and policy related to copyright and intellectual property:

- International Federation of Library Associations (IFLA), *Committee on Copyright and Other Legal Matters* website http://www.ifla.org/III/clm/copyr.htm.
- *United States Copyright Office* website http://www.copyright.gov/.
- American Library Association, *Copyright* website http://www.ala.org/ala/washoff/woissues/copyrightb/copyright.cfm.
- *Section 108 Study Group* website http://www.loc.gov/section108/study.html. Section 108 of U.S. Copyright Law provides limited exceptions for libraries and archives. The study group was convened to recommend changes to the law for current technologies.
- University of Maryland University College, *Center for Intellectual Property* website http://www.umuc.edu/distance/odell/cip/cip.shtml. Resource center for the higher education community on the topics of copyright and intellectual property.

In the UK, the Technical Advisory Service for Images (TASI), a JISC-funded project that advises the UK's further and higher education community on digitization initiatives, provides a number of useful guides that address each aspect of rights, from digitization to use:

- Technical Advisory Service for Images (TASI), *Copyright and Digital Images* (2006) http://www.tasi.ac.uk/advice/managing/copyright.html.
- Technical Advisory Service for Images (TASI), *Roles and Responsibilities for Staff Involved in Building Digital Image Collections* (2006) http://www.tasi.ac.uk/advice/managing/copyright-creators.html.
- Technical Advisory Service for Images (TASI), *Roles and Responsibilities for Staff Involved in Using Images for Teaching and Research* (2006) http://www.tasi.ac.uk/advice/managing/copyright-users.html.

Although treaties have an important role to play in the international arena by providing the minimum standards to which all signatory countries adhere, every country is different in its approach to copyright law. This edition of the *Framework of Guidance* uses Web 2.0 technologies to encourage readers to share information. The editors encourage readers to share guidance in copyright adherence for digital collection building for their countries.

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 not widely in use in cultural heritage environments, but their continued development and application bear watching.

- California Digital Library Rights Management Framework, *Copyright Data Elements for the CDL* http://cdlib.org/inside/projects/rights. Includes a schema for recording information relevant to the copyright status of a work.
- Karen Coyle, *Rights Expression Languages: A Report for the Library of Congress* (2004) http://www.loc.gov/standards/relreport.pdf. Analyzes a sample of rights expression

languages in terms of their impact on the selection, maintenance, and preservation of digital content.

- Rutgers University Libraries, *Rucore Rights Metadata - Draft* (2006) http://rucore.libraries.rutgers.edu/collab/ref/doc_mwg_rights_md_draft.pdf. Provides copyright information and an event subschema for documenting rights events, such as rights transfer, rights research, etc.

Examples of digital collections with copyright or legal policies on their web pages:

- Library of Congress, *Legal* website http://www.loc.gov/homepage/legal.html.
- University of Washington, *College of Architecture and Urban Planning Copyright Policy* website http://www.caup.washington.edu/vrc/policies/vrccopyright.html.
- University of Arkansas Digital Collections, *Copyright and Permissions* website http://0-digitalcollections.uark.edu.library.uark.edu/copyright.asp. Includes a form to request permission to publish.
- Tate Online, *Copyright and the Reproduction of Works* website http://www.tate.org.uk/home/copyright.htm.

COLLECTIONS PRINCIPLE 6

> **Collections Principle 6**: A good collection has mechanisms for collecting data that measure use and usefulness.

Digital collections should be evaluated periodically to monitor usage, assess service effectiveness, demonstrate return on investment, inform collection development, inform strategic planning, and support funding requests. The criteria, methods, and metrics for evaluating collections will vary by the objectives of the collections and the purposes of the evaluation. For example, the collections of the National Science Digital Library are designed to support teaching and learning, so it is appropriate that evaluation measures focus on the educational impact of these collections.

Effective collection management employs a variety of research methods to assess collection usefulness. Observation, surveys, focus groups, interviews, experiments, case studies, and transaction log analyses have been used by digital libraries to assess usage and usability. Each method has its strengths and limitations. To obtain a clear picture of the value of a digital collection is to answer the question: "Who is using what, how, and why?" It is often necessary for collection evaluators to use a combination of methods and measures to answer this question effectively.

The use of the digital collection is closely related to the collection's content, functionality, usability, and accessibility. Establishing benchmarks for use, collecting usage data over time, and following international standards for measuring use of digital content will enable collection managers to conduct longitudinal collection assessment and compare collection services with those provided by peers. Evaluation is an iterative process. Results of evaluation should inform the design and improvement of a digital collection.

Frameworks and guidance for evaluating digital collections:

- Fourth DELOS Workshop, *Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics* (2002) http://www.sztaki.hu/conferences/deval/presentations.html. Offers a promising evaluation scheme by identifying Users, Data/Collection, System/Technology, and Usage as four dimensions of digital libraries and developing evaluation metrics for each.
- Christine Borgman, *Evaluating the Uses of Digital Libraries* (2004) http://www.delos.info/files/pdf/events/2004_Ott_4/Borgman.pdf. A useful framework for evaluating several dimensions of digital library usage.
- Roxanne Missingham, *What Makes Libraries Relevant in the 21st Century? Measuring Digital Collections from Three Perspectives* (2003) http://www.nla.gov.au/nla/staffpaper/2003/missingham2.html.
- Thomas C. Reeves, Xornam Apedoe, and Young Woo, *Evaluating Digital Libraries: A User-Friendly Guide* (2003) http://eduimpact.comm.nsdl.org/evalworkshop/UserGuideOct20.doc.

- Tefko Saracevic, *How Were Digital Libraries Evaluated?* (2004) http://www.scils.rutgers.edu/~tefko/DL_evaluation_LIDA.pdf. An overview of previous digital library evaluation efforts.

Currently the COUNTER standards are dominant for measuring the use of digital collections, but they focus more on vendor-provided data than on collections produced by institutions.

- *COUNTER (Counting Online Usage of NeTworked Electronic Resources)* website http://www.projectcounter.org/code_practice.html. The COUNTER Codes of Practice are standards for measuring use of electronic journals, databases, and e-books.
- NISO, *Standardized Usage Statistics Harvesting Initiative (SUSHI)* website http://www.niso.org/committees/SUSHI/SUSHI_comm.html. Defines Web services-based harvesting of COUNTER usage data from different vendor platforms.

Usage data are somewhat limited when considered alone. When combined with input measures, output measures, or instructional data, they can help shed light on the effectiveness of a digital collection or digital library. Google Analytics (http://www.google.com/analytics/) provides tools to track where users come from and how they use a website.

Resources on collection evaluation methods, standards, and tools:

- Denise Troll Covey, *Usage and Usability Assessment: Library Practices and Concerns* (2002) http://www.clir.org/pubs/abstract/pub105abst.html. A good overview of research methods for studying collection use.
- International Coalition of Library Consortia (ICOLC), *Revised Guidelines for Statistical Measures of Usage of Web-Based Information Resources* (2006) http://www.library.yale.edu/consortia/webstats06.htm. Endorses COUNTER and SUSHI and provides guides for recording usage statistics.
- *Managing Electronic Collections: Strategies from Content to User* website http://www.niso.org/news/events_workshops/Collections-06-Agenda.html. Presentations from a NISO workshop held September 28-30, 2006 at Denver, Colorado. Presentations from day one "Understanding users and usage" and day two "Usage statistics wrap-up; practical collection and repository management" are especially useful.
- Association of Research Libraries, *MINES for Libraries: Measuring the Impact of Networked Electronic Services* website http://www.arl.org/stats/initiatives/mines. A web-based survey on user demographics and their reasons for using networked electronic resources.
- ANSI/NISO *Z39.7-2004, Information Services and Use: Metrics & statistics for libraries and information providers--Data Dictionary* (2004) http://www.niso.org/emetrics/. A data dictionary of terms pertaining to use metrics and statistics, includes measures for electronic resources; the main focus is on usage of resources in libraries.

Some collection assessment studies:

- Johan Bollen and Rick Luce, "Evaluation of Digital Library Impact and User Communities by Analysis of Usage Patterns," *D-Lib Magazine,* v. 8, no. 6 (2002) http://www.dlib.org/dlib/june02/bollen/06bollen.html. Using server logs to assess the impact of a digital collection and understand the user community of a digital library.
- Christine Borgman, *Evaluating a Digital Library for Undergraduate Education: A Case Study of the Alexandria Digital Earth Prototype* (2002) http://www.sztaki.hu/conferences/deval/presentations/borgman.ppt. How users make use of digital collections and the effect of digital collections on students and instructors.
- Casey Jones et al, "Developing a Web Analytics Strategy for the National Science Digital Library," *D-Lib Magazine*, v. 10, no. 10 (2004) http://www.dlib.org/dlib/october04/coleman/10coleman.html. A summary of NSDL evaluation efforts.
- Susan Musante, *Evaluating MicrobeLibrary on Many Levels: Library Use, User Needs, Accessibility Issues, and Educational Impact* (2004) http://nsdl.comm.nsdl.org/meeting/archives/2003/wiki/uploads/36/MicrobeLibrary_NSDL_2003_Presentation.ppt.
- Chris Neuhaus, *Digital Library Evaluation: Measuring Impact, Quantifying Quality, or Tilting at Windmills?* (2003) http://nsdl.comm.nsdl.org/meeting/archives/2003/wiki/uploads/36/nsdlevaluation_101503_eerl_2.1.ppt.
- Michael Organ, "Download Statistics: What Do They Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia," *D-Lib Magazine*, v. 12, no. 11 (2006) http://www.dlib.org/dlib/november06/organ/11organ.html.
- Kristen Fisher Ratan, *Applications of Usage Statistics* (2006) http://www.niso.org/presentations/MEC06-05-Ratan.pdf. How usage statistics are used for collection management purposes.

COLLECTIONS PRINCIPLE 7

Collections Principle 7: A good collection is interoperable.

Collection developers should design their services to support interoperability, particularly the ability to share their metadata with external search engines. At an early stage in the collection design process, collection developers should scan the landscape for related efforts. Collection developers 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. This is a way to expand the use and usefulness of digital collections and may help gain sustainable support for them.

Making collections interoperable will also open up new opportunities for re-purposing the contents of a collection. We can assume that current delivery and access services may evolve into mechanisms that are inconceivable thus far. The National Library of Australia's *IT Architecture Project Report* (http://www.nla.gov.au/dsp/documents/itag.pdf) alludes to this in both its content and recommendations.

The Google *Webmaster Tools* website (https://www.google.com/webmasters/tools/docs/en/about.html) provides free tools to improve Google indexing of any website.

For more on collection description, see Collections Principle 2.

For more on sharable metadata, see Metadata Principle 2.

For more on interoperable objects, see Objects Principle 3.

COLLECTIONS PRINCIPLE 8

> **Collections Principle 8**: A good collection integrates into the workflows of staff and end users.

When digital collection building represents a significant new service for an organization, it presents an opportunity to review existing workflows, and possibly reallocate resources, responsibilities and tasks. In order to successfully add digital collections to an organization's service suite, it is important to integrate digital collection building into staff workflows. When digital collection building is an established activity, it is useful to periodically review staff workflows for improvement. The operational staff performing these activities will generally be the best advisors as to how to make them more intuitive and less time consuming.

End users find information most useful when it integrates smoothly with their own patterns of work. A faculty member looking for research articles and a recreational genealogist building an electronic family tree will work in different places, at different times, and using different tools. However each will use a digital collection more comfortably if they can access it from the environment with which they are familiar.

In the emerging digital landscape, digital collection building is increasingly a collaboration with the end user. Some sites allow users to add keywords to the metadata ("social tagging"). Some allow them to contribute digital resources, such as personal stories or family photographs to a local history web portal or preprints or postprints to an academic institutional repository. Integrating with the user's workflow enables the user to contribute without significant additional effort. An example might be the ability, in a collaboration space such as the Sakai learning management system, to simply "save to the repository" when the faculty author has completed his work. A local history portal might ask, at the end of a search on a topic, "Do you have any resources on this topic to share?" and provide a simple menu-driven process to upload digital resources.

Workflow examples for digital collection building:

- Jessica Williams, Sandra Paske and Stephen Dast, *Audio Procedures and Workflow for the University of Wisconsin Digital Collections Center* (2004) http://uwdcc.library.wisc.edu/documents/AudioWorkflow.pdf.
- Technical Advisory Service for Images (TASI), *Managing the Workflow* website http://www.tasi.ac.uk/advice/managing/workflow.html.
- Sharon Favaro, *Metadata Workflow for Digital Collections* (2006) http://www.njdigitalhighway.org/documents/njdh-metadata-workflow.pdf.

End user workflow:

- Herbert Van de Sompel, et al, "Rethinking Scholarly Communication: Building the System that Scholars Deserve," *D-Lib Magazine*, v. 10, no. 9 (2004) http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.

- Tim O'Reilly, *What is Web 2.0?* (2005) http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. The classic article defining Web 2.0 as a new platform harnessing the power of the community.
- Michael E. Casey and Laura C. Savastinuk, "Library 2.0: Service for the Next Generation Library," *Library Journal*, v. 131, Issue 14 (2006) http://www.libraryjournal.com/article/CA6365200.html.

COLLECTIONS PRINCIPLE 9

Collections Principle 9: A good collection is sustainable over time.

Digital collections containing resources of long-term value should be sustained and archived permanently to ensure access. Sustainability needs to be addressed from an organizational, financial, and technical perspective. Organizational commitment requires buy-in from administrators. There must be a clear understanding of the long-term obligations necessary to ensure a sustained digital collection.

Sustaining the collection may take different sets of skills and different commitments of resources than the original collection building. Aspects of ongoing maintenance include 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.

In particular, digital collections built with special funding should have a plan for their continued availability, maintenance and support beyond the funded period. Optimally, regardless of how it was initiated, the digital collection will be integrated into the institutional collections management workflow.

Sustainability at the collection level is related to, but not identical with, persistence at the object level (see OBJECTS Principle 2). Certainly the collection-level archiving strategy should be tied to the object-level preservation strategy. However, 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 content is maintained.

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

- Technical Advisory Service for Images, *Sustainability of Digital Resources* (2006) http://www.tasi.ac.uk/advice/managing/sust.html.
- Donald Waters, "Building on Success, Forging New Ground: The Question of Sustainability," *First Monday*, v. 9, no. 5 (2004) http://firstmonday.org/issues/issue9_5/waters/index.html.
- Liz Bishoff and Nancy Allen, *Business Planning for Cultural Heritage Institutions* (2004) http://www.clir.org/PUBS/reports/pub124/contents.html.
- *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*, Chapter XI, Sustainability: Models for Long-term Funding (2002) http://www.nyu.edu/its/humanities/ninchguide/XI/.

OBJECTS

OBJECT PRINCIPLE 1

> **Objects Principle 1**: A good object exists in a format that supports its intended current and future use.

Consequently, a good object is exchangeable across platforms, broadly accessible, and formatted according to a recognized standard or best practice.

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 creation of digital 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. An object intended to have long-term value should be formatted to render it exchangeable across platforms and broadly accessible. 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.

When speaking of digital content, the word "format" carries multiple meanings. Several of these are discussed in an introductory essay to the *Sustainability of Digital Formats: Planning for Library of Congress Collections* website (http://www.digitalpreservation.gov/formats/). In the context of this document, two of the most important meanings pertain to file formats and bitstream encodings. File formats are generally identified by file extensions (e.g., .mp3) or MIME (e.g., text/html). Bitstream encodings underlie certain file formats, e.g., the linear pulse code modulated (LPCM) waveforms that may be found in WAVE or AIFF files, or the H.264 video that may be found in QuickTime or MPEG-4 files. Those two encodings are specific to content category (in these cases, audio and video), while others are generic, e.g., LZW (Lempel-Ziv-Welch compression encoding). There are few strict correlations between file formats and encodings.

A variety of international efforts have been launched to document digital formats and to provide tools to help manage them. Important examples include the following:

- *Global Digital Format Registry* website http://hul.harvard.edu/gdfr/. A model and implementation of interoperating distributed format registries, initiated in the United States. See also Stephen L. Abrams and David Seaman, *Towards a Global Digital Format Registry* (2004) http://www.ifla.org/IV/ifla69/papers/128e-Abrams_Seaman.pdf.
- *PRONOM* and *DROID* website http://www.nationalarchives.gov.uk/pronom/. Two tools developed by The National Archives (United Kingdom). PRONOM is an online registry of technical information about file formats, software products, and related topics. DROID is an automatic file format identification tool.

- *Automatic Obsolescence Notification System* (AONS) website http://pilot.apsr.edu.au/wiki/index.php/AONS_II. Now under development by the National Library of Australian (NLA) and the Australian Partnership for Sustainable Repositories (APSR), AONS will be a platform-independent, downloadable tool that automatically provides information from authoritative international registries informing users when file formats in their repositories are obsolete or at risk of becoming obsolete.

Information about file formats and encodings is provided in the two tables below: one for reformatting activities and one for the acquisition of born-digital content. The understanding and application of digital technology to library and archive content and its preservation has moved unevenly across the various content types (and sometimes subtypes) that are listed in the tables below. The resulting variation is reflected in the number and quality of references cited, and in the confidence that the compilers of this document bring to each content type.

TABLE 1: REFORMATTING NON-DIGITAL SOURCE MATERIALS

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
Printed matter and manuscripts, not oversize (images of pages)	Master images as uncompressed TIFF files (well established) or lossless compressed JPEG2000 (emerging practice). Service-image formats for access vary according to delivery system, generally favoring formats supported natively in browsers or via free, widely available plug-ins, e.g., JPEG and PDF.	Quality factors include bit depth and spatial resolution; these vary from project to project, with most selecting grayscale or color images in the 300-600 ppi range, at 8 or 24 bits per pixel, for master images, and few selecting bitonal. Special use cases can motivate a project to scan at higher levels of resolution; certain classes of old manuscripts, for example, have been digitized at levels as high as 2,400 ppi for study by scholars. The most thorough general treatment of raster imaging is the 2004 document from the National Archives and Records Administration: *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files — Raster Images* (http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf). Another example of general coverage of raster imaging is the California Digital Library: *Digital Image Format Standards* (http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf). A useful document limited to printed matter is *Benchmark for Faithful Digital Reproductions of Monographs and Serials* (http://www.diglib.org/standards/bmarkfin.htm), Version 1 (DLF, 2002). Meetings of the JPEG 2000 in Archives and Libraries Interest Group (http://j2karclib.info/) have highlighted growing interest in employing JPEG 2000 images as masters or archival formats in reformatting projects. Although particular to newspapers (and related to the scanning of microfilm rather than paper), guidelines potentially of broad utility are offered by the Library of Congress National Digital Newspaper Program (NDNP) *Technical Guidelines for Applicants* (http://www.loc.gov/ndnp/pdf/ndnp_techguide.pdf – click link to *Newspaper Digitization*). Meanwhile, a new activity within the NDIIPP project at the Library of Congress is bringing together several federal agencies, including NARA and the Government Printing Office, to develop guidelines and standards for use within the

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
		federal government. A public website is expected before the end of 2007.
Printed matter and manuscripts (machine-readable texts)	Master files as marked-up texts, generally within an established XML schema or DTD, e.g., TEI or TEI-lite. Formats for access vary according to the requirements of indexing and/or delivery systems.	Most online presentations depend upon images for the authoritative representation of content, with text accuracy considered as satisfactory if sufficient to support search and retrieval. Quality factors generally focus on the degree and correctness of markup. General information on markup is provided by *Creating and Documenting Electronic Texts* (http://ota.ahds.ac.uk/documents/creating/) (OTA, 1999). The Text Encoding Initiative (http://www.tei-c.org/) (TEI) is an important initiative; see *TEI Text Encoding in Libraries: Guidelines for Best Encoding Practices* (http://www.diglib.org/standards/tei.htm) (DLF, 1999). Some additional links (including ALTO, the Analyzed Layout and Text Object) are provided by the National Digital Newspaper Project (http://www.loc.gov/ndnp/metadatalinks.html). Many online presentations of manuscripts do not include machine-readable texts.

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
Pictorial materials (reflected light)	Master images as uncompressed TIFF files (well established) or lossless compressed JPEG2000 (future practice). Derivative formats for access vary according to delivery system, often formats supported natively in browsers, most often JPEG.	Quality factors include bit depth and spatial resolution; these vary from project to project, with most selecting grayscale or color images in the 300-600 ppi range, at 8 or 24 bits per pixel, for master images. Some projects may scan at higher levels because their designated community wishes to examine very small or subtle features of the original. The most thorough general treatment of raster imaging is the 2004 document from the National Archives and Records Administration: *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images* (http://www.archives.gov/research_room/arc/arc_info/techguide_raster_june2004.pdf) (June 2004). Another example of general coverage of raster imaging is the California Digital Library *Digital Image Format Standards* (http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf). New approaches are under consideration; see recent papers by NARA's Steve Puglia and Erin Rhodes (examples: http://www.rlg.org/en/page.php?Page_ID=21033 - article2 and http://www.imaging.org/conferences/archiving2007/details.cfm?pass=50). Meetings of the JPEG 2000 in Archives and Libraries Interest Group (http://j2karclib.info/) have highlighted growing interest in employing JPEG 2000 images as masters or archival formats in reformatting projects.

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
Pictorial materials (negatives, other transmitted light)	Same as above today, practice may vary in future.	Today's practices are generally very similar to the preceding, except for an emerging preference for higher bit depth ("extended data range") to accommodate all of the information in a photographic negative. The Library of Congress frequently scans black-and-white negatives as 16-bit grayscale images (for example, see http://memory.loc.gov/ammem/collections/anseladams/aambuild.html). Spatial resolution at the level of a negative (often much smaller than a print) is often expressed in terms of overall pixel count, e.g., "5,000 pixels on the long side." Some discussion of emerging practices will be found in recent papers by NARA's Steve Puglia and Erin Rhodes (examples: http://www.rlg.org/en/page.php?Page_ID=21033 - article2and http://www.imaging.org/conferences/archiving2007/details.cfm?pass=50).
Oversize typographic or pictorial materials	Master images generally employ the same formats as the preceding. Delivery to users, however, often exploits the scaling or tiling functionality of formats like JPEG 2000, MrSID, or DjVu, with on-the-fly creation of GIF or JPEG images for end users.	Scant information is provided by pages like http://memory.loc.gov/ammem/help/mrsid.html at the Library of Congress and http://www.delamare.unr.edu/maps/digitalcollections/nvmaps/siteinfo.html at the University of Nevada, Reno.
Sound recordings, no synchronized transcription (music or speech)	Masters should consist of a linear PCM bitstream, which may be wrapped in a WAVE, AIFF, or Broadcast WAVE file,	INTRODUCTORY DISCUSSIONS: (1) International Association of Sound and Audiovisual Archives (IASA). *Guidelines on the Production and Preservation of Digital Objects*, IASA-TC 04 (ISBN 87-990309-1-8), August 2004. (2) *The NINCH Guide to Good Practice in the Digital Representation of Cultural Heritage Materials* has a good chapter on audio/video capture and

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
	which can include a bit more metadata. End-user delivery formats are typically MP3, QuickTime, WindowsMedia, and RealAudio.	management. (http://www.nyu.edu/its/humanities/ninchguide/VII/) (3) Carl Fleischhauer, *Library of Congress Digital Prototyping Project, 1998-2003* (http://www.arl.org/preserv/sound_savings_proceedings/Digital_audio.shtml). PRACTICAL GUIDELINES: (1) The forthcoming report (expected in 2007) from the *Sound Directions* project (Harvard and Indiana Universities, http://www.dlib.indiana.edu/projects/sounddirections/) will offer many useful ideas and pointers to tools. Some related information is available via the *Harvard University Library Digital Initiative Audio Reformatting* website (http://hul.harvard.edu/ldi/html/reformatting_audio.html). (2) Detailed information on the playback of analog sound media for digitization is provided in *Capturing Analog Sound for Digital Preservation* from an experts' roundtable discussion organized in 2004 by the Library of Congress (http://www.clir.org/PUBS/execsum/sum137.html). (3) The National Library of Canada provides useful explanations in its brief technical introduction to Digital Audio (http://epe.lac-bac.gc.ca/100/202/301/netnotes/netnotes-h/notes49.htm), although it suggests the use of cleanup tools, a practice eschewed by most preservation reformatting programs. NOTES ON SPOKEN WORD CONTENT: Useful information pertaining to the digitization of speech is offered by the National Gallery of the Spoken Word (NGSW) projects, based at Michigan State University (http://www.historicalvoices.org/papers/audio_digitization.pdf and http://www.historicalvoices.org/papers/sounds.rtf). The transcription and translation of spoken word content (as of 2002) is described in this report from a working group supported in the US by the NSF and in the EU by DELOS (http://www.dcs.shef.ac.uk/spandh/projects/swag/). See also the next table row

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
		for information about the synchronization of sound content and transcriptions (which may be musical as well as textual).
Sound recordings with synchronized transcriptions (music or speech, e.g., oral histories)	For sound formats, see the preceding table row. Synchronized music notation and text formats represent emerging practices; see the examples cited in this row.	INTRODUCTORY DISCUSSIONS: The Spoken Word Project (http://www.at.northwestern.edu/spoken/) at Northwestern University features information on synchronizing transcripts and sound recordings. PRACTICAL GUIDELINES: See the Northwestern University resources cited above. Examples of projects with synchronization: the *OYEZ* multimedia archive devoted to the Supreme Court of the United States (http://www.oyez.org/); see examples like the William O. Douglas interviews (http://www.oyez.org/justices/william_o_douglas/interview-tapes/), which uses Adobe Flash to present synchronized content to end users.
Moving images, video recordings on conventional tangible media (analog and digital videotapes, DVDs)	See comments at right and list of resources in next table row.	CURRENT PRACTICE, HYBRID APPROACH: For the reformatting of videotapes, most archives continue to produce a new videotape as a preservation master, typically a Beta SX (DigiBeta); some archives may use the more expensive D1, D5, or other types. All of these magnetic tape formats are obsolete, however, and may require re-reformatting within a decade. Service copies are generally digital files: in a high-bandwidth LAN, high-bit-rate MPEG-2 or MPEG-4 files in larger picture sizes; for lower bandwidth applications and the Web, lower-rate MPEG-4, RealVideo, or QuickTime formats with smaller picture sizes. A good introduction is provided by the Association of Moving Image Archivists (AMIA) in *Reformatting for Preservation: Understanding Tape Formats and Other Conversion Issues* (http://www.amianet.org/new/resources/guides/fact_sheets.pdf). EXPLORING FILE-BASED MASTERS: Little in the way of fully realized, experience-based documentation exists for this approach; much must be gleaned from e-mail discussion lists and personal communication. One useful guideline for making files containing uncompressed

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
		video streams is *Standards Analysis for Video Objects: Recommended minimum requirements for preservation sampling of moving image objects*, by Isaiah Beard for the Rutgers University RUcore project (http://rucore.libraries.rutgers.edu/collab/ref/dos_avwg_video_obj_standard.pdf). Meanwhile, several experts advocate preservation masters that employ a "frame-by-frame" approach; individual frame images may be uncompressed or encoded as JPEG 2000 (lossless or lossy), within a suitable wrapper (MXF, Motion JPEG 2000, AVI, others); or as MPEG-2 or MPEG-4 "all I frame" encodings; or even as DV. For the MPEG and DV lossy encodings, higher data rates (e.g., 50 mbps) are preferred to lower. Reformatting (to tapes as well as files) often requires transcoding, e.g., from composite to component color space and, for compressed formats, to compress the signal. In contrast, it is possible to extract the native digital signal from formats like DVDs (MPEG-2) of DV/DVC/DVCPRO videotapes (DV), but there seems to be no established practice for this. Making a file entails placing the encoded digital essence in a wrapper, e.g., MXF, Motion JPEG 2000, AVI, QuickTime, MPEG-4, but again, the community has not yet established practices. REGARDING SOUNDTRACKS: Sound may be interleaved with the video in the "stream," or may be managed as a separate element within several wrapper formats (e.g., MXF, Motion JPEG 2000, AVI). Audio encoding may be uncompressed linear PCM or compressed (usually lossy) in an encoding that is accepted by the wrapper.
Moving images, video recordings on conventional tangible media (analog and digital videotapes, DVDs). Continued from	List of resources at right.	As the preceding row indicates, a number of players with a variety of ideas (conventional and cutting-edge) are exploring the conservation of older videotapes and best practices for reformatting them. (1) A historical overview of digital video is offered by Grace Agnew's "Video on Demand: the Prospect and Promise for Libraries" (http://www.dekker.com/sdek/issues~db=enc~content=t713172967) in the *Encyclopedia of Library and Information Science* (New York: Marcel Dekker,

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
preceding row.		2004). (2) The Association of Moving Image Archivists (AMIA) (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 discussion list (http://www.amianet.org/participate/listserv.php) are relevant to video archiving; see also the listserv's archive (http://lsv.uky.edu/archives/amia-l.html). (3) *Magnetic Tape Storage and Handling*, CLIR report, 1995 (http://www.clir.org/pubs/reports/pub54/). (4) A useful wiki from the European broadcasting PrestoSpace project beginning in 2006 (http://wiki.prestospace.org/), includes a tutorial titled "Why Digitise." (5) The Video Preservation page at Stanford University offers a menu of links to other sites (http://palimpsest.stanford.edu/bytopic/video/). (6) A British overview report from the Arts and Humanities Data Service is linked from this page: http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/project_movingimagesound.aspx. (7) *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/VII/). (8) Regarding file-based approaches, the vendor Media Matters offers some useful white papers (http://www.sammasystems.com/whitepapers.html). (9) The Video Development Group (http://www.vide.net/) (ViDe) has a special interest in video that supports higher education and provides information about digital video file creation.
Moving image (film)	See comments at right	CURRENT PRACTICES: Virtually all archives today employ the well understood and well established approach of traditional photochemical reproduction. The original film is printed onto

CONTENT CATEGORY	TARGET FORMATS	REFERENCES AND COMMENTS
		an appropriate film stock, which is developed in the conventional way, yielding archival masters. Depending upon the configuration of the starting material (i.e., negative or positive, sound or silent) the nature and number of the archival preservation masters varies. One or two additional generations may be printed and developed, yielding duplicating copies and a print for projection. Prints may also be made directly from the original materials when appropriate. Guideline documents include the 2004 *Film Preservation Guide: The Basics for Archives, Libraries, and Museums* (http://www.filmpreservation.org/preservation/film_guide.html) and the *Film Preservation Handbook* from Australia's National Film and Sound Archive (http://www.nfsa.afc.gov.au/preservation/film_handbook/). EMERGING IDEAS: The extensive use of digital technology by commercial filmmakers will lead to changes in reformatting in archives during the next few years. Here are two likely approaches: (1) Film-to-digital-to-film. Original film is scanned (and may be subsequently corrected) to produce a "digital master" consisting of high resolution frame images, which are "recorded" back to film stock, in turn developed in the traditional manner. The resulting film can either be a master for archival storage or a print that can be projected. (2) Film-to-digital-to-digital. Original film is scanned to produce a "digital master" consisting of high-resolution frame images (same as b). This master is used to produce a digital-projection element, which might confirm to the Digital Cinema Initiative Distribution Master standard. The master may also be used to produce lower resolution proxies for distribution via DVD or the Web.

The information in Table 2 is tentative. Many of the preferred formats listed are those suggested by the Library of Congress *Sustainability of Digital Formats* website (http://www.digitalpreservation.gov/formats/index.shtml), which emphasizes that its recommendations are provisional. This table, like that website, is written from the perspective of institutions that are likely to receive content from creators not under their control. The Library of Congress, for example, receives content in many ways, ranging from copyright deposit to the donation of personal papers with boxes of floppy disks. Thus the table below allows for the possibility that incoming content may take the form of, say, PDF files or even word-processing files. However, where an institution has any control over born digital content, it is highly desirable to encourage authors to create their works in specified formats. For example, in some scholarly projects, it may be possible to insist that authors create their documents in XML, following guidelines like those from the Text Encoding Initiative (http://www.tei-c.org/). Similarly the graduate schools of many universities compel students to submit electronic theses and dissertations in library-approved formats.

Three topics have general applicability and are not articulated row-by-row in the table below. The first is the strong preference of libraries and archives to receive content that includes metadata, whether embedded in a file or as an associated "sidecar." The second concerns digital works that arrive at an archive in a hard-to-sustain format, prompting the archive to transcode the content into an easier-to-sustain format. Several archivists argue that such content should be kept in both the original form (even though it may be very hard to read in the future) and in the migrated, easier-to-sustain form.

The third general topic has to do with technological protections, the "locking" of content associated with Digital Rights Management (DRM) regimes. Some formats have embedded capabilities to restrict use, say, by time period, to a particular computer or other hardware device, or by requiring a password or active network connection. Since the exploitation of the technical protection mechanisms for a given format is usually optional, this consideration arises when a format is used in a particular business context, e.g., the sale of downloadable music from entities like Apple iTunes. To preserve digital content and provide service to users and designated communities decades hence, custodians must be able to replicate the content on new media, migrate and normalize it in the face of changing technology, and disseminate it to users at a resolution consistent with network bandwidth constraints.

The preceding paragraph offers a problem statement but does not cite examples of library or archive practice that address the matter. In fact, the compilers of this document believe that libraries and archives have only slight experience with DRM or any other aspects of born digital content. Few of us have direct experience with many of the formats listed below. Therefore, we strongly encourage our readers to enrich this table with fresh information or links to relevant resources.

TABLE 2: BORN DIGITAL MATERIALS

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
Textual content, monographic (emphasizing layout or typography)	PDF, PDF/A, various word-processing formats, other	Text formats should represent the underlying text in a way that is accessible to search engines. Preferred are PDF/A or other PDF subtypes created from machine-readable text (as opposed to page images). HTML (hierarchy or network of linked pages) are acceptable if published/disseminated only in this form. Proprietary binary formats used by word-processing and desktop-publishing software are not a good choice for long-term management; text documents in such formats should be printed to PDF (preferably PDF/A) and/or converted to a transparent non-proprietary format such as OpenOffice, which is XML-based.	*Guidelines for Creating Archival Quality PDF Files*, Florida Center for Library Automation (http://www.fcla.edu/digitalArchive/pdfs/PDFGuideline.pdf). *PDF/A –Format – Status and Practical Experiences*, Presentation at the European Document Lifecycle Management (DLM) Forum, 2006 (http://www.aiimhost.com/DLM/DLMHelsinki_MarcStraat.pdf)
Textual content, monographic (marked up)	XML, SGML, HTML	Preferred: XML or SGML using standard or well-known DTD or schema appropriate to a particular textual genre.	Open eBook Forum (http://www.openebook.org/). Supporting Documentation for *ANSI/NISO Z39.86, Specifications for the Digital Talking Book*

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
		Examples: *Open eBook Publication Structure*, Version 1.2 (for novels, text-books, scholarly monographs, etc.); *Digital Talking Book*, ANSI/NISO Z39.86 with full transcript of text (for novels, text-books, scholarly monographs, etc.); *Journal Archiving and Interchange Document Type Definition* (DTD) from NML.	(http://www.loc.gov/nls/z3986/). For the NLM DTD, see http://dtd.nlm.nih.gov/. The Text Encoding Initiative (TEI; http://www.tei-c.org/) is most often associated with reformatting projects but could also play a role with newly created texts.
Textual content, serial	See comment	A place to start for articles and e-journals: Journal Archiving and Interchange Document Type Definition (DTD; http://dtd.nlm.nih.gov/).	More resources pertaining to e-journals are sought from readers. However, one important implementation of the NLM DTD mentioned in the previous row is by the PORTICO preservation archiving service (http://www.portico.org), which ingests marked-up and PDF-formatted serial content. An excellent description of their practices and systems is Evan Owens' 2006 article "Automated Workflow for the Ingest and Preservation of Electronic Journals" (http://www.portico.org/news/Archiving2006-Owens.pdf).

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
Harvested websites	Prior to harvesting: HTML and other formats	ARC or WARC	Web archiving activity typically collects web pages and embedded images, sounds, and the like, in as complete a manner as possible, including the link structure. At this writing, most harvesting employs one of two related formats designed for web archiving: ARC and WARC. The former was developed by the Internet Archive to support its work; WARC is a refined and extended format that is based on ARC and, in 2006-2007, under consideration as a standard by ISO. For a description of practices at the Library of Congress, see http://www.loc.gov/webcapture/technical.html.
Alphanumeric data (datasets)	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 the Data Documentation Initiative (DDI) (http://www.icpsr.umich.edu/DDI/), from the Inter-university Consortium for Political and Social Research (ICPSR). The creation of databases associated with historical research is the topic of a 1999 document from the British Arts and Humanities Data Service: *Digitising History, a Guide to Creating Digital Resources from Historic Documents* (http://hds.essex.ac.uk/g2gp/digitising_history/index.asp).
Digital photographs and other born-digital	TIFF (various RGB encodings), JPEG, various camera raw formats, JPEG 2000, HD Photo	Retain TIFF and JPEG; see comment below regarding RGB encoding. Convert camera raw to DNG, JPEG	CAMERA RAW: Although proprietary, camera raw formats permit the creator-editor to interpret the image as the various post-processing steps are applied,

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
bitmapped graphics		2000, or TIFF; retain HD Photo in native format until more experience has been gained.	tailoring an outcome that suits the subject matter and intended use. This may be a desirable characteristic for future users of an archive who wish to adjust a picture for a particular application. DNG currently represents the only format to contain normalized raw data with minimal loss of malleability. (As is noted below, the need for normalized formats that retain malleability also arises in other categories, e.g., GIS and music-notation.) Conversion to JPEG 2000 or TIFF entails some level of "rendering," but in many circumstances may be a reasonable compromise. COLOR INFORMATION PREFERENCES: Specified color space preferred over unspecified or unknown color space. RGB or luminance-chrominance color space, e.g., YUV, YCC, YCrCb, preferred for images originating from scanners or cameras. For RGB formats, inclusion of creation-device ICC color profile or equivalent preferred to omission. For RGB formats, sRGB color space preferred when profiles or other color management tools have not been employed. Color-specifying color space, e.g., CMYK and CIELab, preferred for bitmapped images originating in paint or other graphic arts software. CMYK images that comply with Specifications for Web Offset Publications (SWOP) or Specifications for Newsprint

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
			Advertising Production (SNAP) preferred over non-compliant images.
Digital vector graphics, "desktop" categories	Files produced by Adobe Illustrator (AI), CorelDraw (CDR), Corel Exchange (CMX), Micrografx Draw (DRW), Windows Metafile (WMF). Also Scalable Vector Graphics, Version 1.1 or Scalable Vector Graphics, Version 1.2 (both referred to as SVG), and AutoCad Drawing Interchange Format (DXF)	Preferred formats include both versions of Scalable Vector Graphics (SVG) and DXF.	Information is requested from readers about this category, including information about additional "normalizing" formats like Initial Graphics Exchange Specification (IGES) and Computer Graphics Metafile (CGM and WebCGM).
Professional CAD-CAM, engineering, manufacturing applications	See comment	See comment	Comments are requested from readers that pertain to the potential value of (or guidelines for the use of) the Standard for the Exchange of Product Model Data (STEP; ISO 10303). STEP is intended to provide a complete computer-interpretable definition of the characteristics of a product throughout its life cycle. Meanwhile, a special interest group has been formed under the rubric Long Term Sustainment of Digital Information for Science and Engineering (LTKR); the most recent meeting (2007) took place at the National Institute for Standards and Technology (http://digitalpreservation.wikispaces.com/LTKR+2007+Call+for+participation). Experts associated with LTKR see value in converting

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
			proprietary CAD-CAM data to STEP, although they report imperfections in the conversion process. As a safeguard, these experts recommend the additional creation of a "mesh file" with relatively transparent information about the shapes and volumes represented.
Geospatial, GIS	*Vector formats*: ArcInfo Coverages (ESRI), ESRI Export file (.e00), Shapefiles (ESRI), MapInfo MID/MIF, TIGER, Spatial Data Transfer, Standard (SDTS), Digital Line Graphs (DLG) *Raster formats*: TIFF/GeoTIFF, BIP/BIL/BSQ, JPEG, JPEG2000, MrSID, ESRI Grid	Reader advice requested	An overview for librarians is offered by Steve Morris (NCSU) in *Library Trends* (2006; http://muse.jhu.edu/journals/library_trends/v055/55.2morris.html – access by subscription only). Highlighting the use of GIS in archeology but offering broader insights and guidelines is the 1998 Arts and Humanities Data Service (UK) document *GIS: A Guide to Good Practice* (http://ads.ahds.ac.uk/project/goodguides/gis/index.html). GIS data can generally be "frozen" and output as a bitmapped image, which can be archived for the long term. Most future GIS researchers, however, will wish to have access to historical data in a malleable form. Experts in the field identify two relevant open specifications, neither of which offers a perfect path to data normalization at this time: Geography Markup Language (GML), for which there are diverse and not always compatible applications, and the Spatial Data Transfer Standard (SDTS), which is not widely supported at this time.

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
Digital sound, single waveform bitstream (may be mono, stereo, surround; for commercial distribution, e.g., iTunes)	*Encodings*: Linear PCM, MP3, AAC, DSD, others *File formats*: MP3, AAC, WAVE, AIFF, QuickTime, Windows Media Audio, others	*Encoding preferences*: Linear PCM (uncompressed) preferred over compressed (lossy or lossless); AAC lossy compression preferred over MP3. *File format preferences*: Broadcast WAVE Audio File Format, WAVE Audio File Format with LPCM Audio, AIFF File Format with LPCM Audio, MP3 File Format, MP4 file format (with AAC).	QUALITY PREFERENCES: Higher sampling preferred over lower; 24-bit sample word-length preferred over shorter; higher data rate (e.g., 128 kilobits per second) preferred over lower for same compression scheme and sampling rate.
Digital sound (multi-track waveform project sets)	ProTools project file or stack equivalent produced by other applications	Reader comments sought: AES31-3, AAF, others?	
Digital sound (note-based and mixed formats)	MIDI, various MODs or Tracker formats, eXtensible Music Format, SoundFont sf2	MIDI sequence data preferred; consider transcoding digital compositions in other note-based formats to audio waveform files. DLS standardized downloadable sounds preferred to proprietary samples.	This category includes formats that provide data to support dynamic construction of sound through combinations of software and hardware. The most prominent note-based formats are associated with MIDI, the Musical Instrument Digital Interface, although there are many devotees of formats called MODs, from modules, sometimes called tracker files. The sound elements may be short segments of waveform sound (sometimes called samples or loops) or data elements that characterize a sound so that a

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
			synthesizer (which may be in software or hardware) or sound generator (usually hardware) can produce the actual sound.
Music notation formats	Output from applications such as Finale, Sibelius, others	Comments from readers desired.	Content from these applications can be output in proprietary files (which remain malleable) or as PDF or other bitmapped formats (which "freeze" the data). Like the camera raw formats discussed above, some future users of archived content will wish to be provided with malleable data; an effective and widely adopted normalized format is needed. (Readers are asked for information on this topic.)
Digital moving images, single bitstream (for commercial distribution)	MPEG-4 (various encodings), QuickTime (various encodings), AVI (various encodings), Windows Media (various versions and encodings)	File format and encoding preferences: MPEG-2, MPEG-4 (AVC coding aka H.264), MPEG-4 (Video coding aka H.263), wrappers like AVI and QuickTime (with H.263 or H.264).	Quality preferences: Larger picture size preferred over smaller; high definition preferred standard definition, assuming picture size is equal or greater; higher bit rate preferred over lower for same compression scheme.
Digital video (multi-track projects sets)	Formats provided by applications like Avid, FinalCut Pro, and others; some are proprietary, some are not	File format and encoding preferences: DPX with picture information together with suitable format for sound information; Digital Cinema Distribution Master; MXF containing uncompressed images, JPEG 2000 frame images, or MPEG-2 streams;	Quality preferences: Larger picture size preferred over smaller; content from high definition sources preferred over standard definition, assuming picture size is equal or greater; encodings that maintain frame integrity preferred over formats that use temporal compression; uncompressed or lossless compressed preferred over lossy; higher bit rate preferred over lower; extended dynamic range

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
		Motion JPEG 2000 with lossless encoding; uncompressed or lossless compressed in wrappers like AVI and QuickTime.	(scene brightnesses) preferred over "normal" dynamic range for such items as Digital Cinema or scanned motion picture film.
Video created by a library or archive to document a live performance	This specialized row is included since some organizations create their own video.	Generally speaking, the preferences stated above apply; if videotape is produced, see the notes about reformatting video in the first table.	Two draft guides from the Internet2/CNI Performance Archive and Retrieval Working Group are: (1) *Capturing Live Performance Events* (http://arts.internet2.edu/files/performance-capture(v09).pdf), version 0.9 (2003), and (2) *Current Practice in Digital Asset Management* (http://arts.internet2.edu/files/digital-asset-management(v09).pdf), version 0.9 (2003).
Animation, interactive	FLASH, SMIL, others	Reader comments encouraged; possible preferences: Adobe (Macromedia) Flash Project File or SWF File; Scalable Vector Graphics, Version 1.1 or Version 1.2; files may also be saved-as bitmapped video (see preceding row).	This category includes files that contain encoding for dynamically generating animations and/or moving image interactive programs, e.g., animated shorts for web delivery or for playback on personal computers. Such animations may be produced by specialized software, e.g., Macromedia Flash, or by certain kinds of Computer Aided Design or Computer Aided Manufacturing (CAD-CAM) systems, especially for three-dimensional drawings that may be rotated to simulate viewing from various points of view.

CONTENT CATEGORY	SOURCE FORMATS	PREFERRED FORMATS	REFERENCES AND COMMENTS
Games, various genres. Reader comments on categories and details requested	Reader comments requested		No experience-based guidelines known to the compilers of this guideline.

OBJECTS PRINCIPLE 2

Objects Principle 2: A good object is preservable.

That is, the object will not raise unnecessary barriers to remaining accessible over time despite changing technologies.

Anyone who has ever used a Wang Word Processor or a 5.25" floppy disk knows the life spans of media, hardware, software platforms, and digital file formats are notoriously short. A digital object that is perfectly usable today may be unusable in the future unless some preservation action is taken. 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 CD to DVD);
- across software products as the products become outmoded (e.g., from one version of a spreadsheet to another); and,
- across formats or encodings as new standards emerge (e.g., from SGML to XML, or from JPEG to JPEG2000).

Emulation involves reproducing on contemporary systems the computer environment in which digital objects were originally created and used. Programs can be written to emulate obsolete hardware, device drivers, and operating systems. Emulation strategies may be particularly appropriate for executables and complex multimedia objects such as interactive learning modules.

Migration and emulation should be seen as complementary approaches. Neither is appropriate for all types of materials, and many preservation strategies combine aspects of emulation and migration.

Master objects should be created whenever possible with digital preservation in mind. Although no digital 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 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. Prefer formats that allow embedded metadata and have few external dependencies.

Preservation masters of retrospectively digitized materials should be as close to the source version as possible. This generally means using a high resolution or sampling rate. Master files should not contain access inhibitors like watermarks or encryption, and should not be compressed with proprietary or lossy compression schemes. Whenever possible, embed everything needed to render the object in the object itself; for example, PDF files should always embed their fonts.

Selecting file formats for preservability:

- U.K. National Archives, *Digital Preservation Guidance,* Note 1: Selecting file formats for long-term preservation (2003) http://www.nationalarchives.gov.uk/documents/selecting_file_formats.rtf. A straightforward explanation of criteria for evaluating digital formats.
- Library of Congress, *The Sustainability of Digital Formats: Planning for Library of Congress Collections* website http://www.digitalpreservation.gov/formats/. Comprehensive guidelines for evaluating formats that balance fitness-for-purpose with preservability.
- Carl Fleischhauer, *Sound Savings: Preserving Audio Collections, The Library of Congress Digital Audio Preservation Prototyping Project* http://www.arl.org/preserv/sound_savings_proceedings/Digital_audio.shtml. An illuminating case study on selecting the digital format and parameters for audio conversion.

The actual provision of migration or emulation services is expected to require detailed information about the format characteristics and the hardware and software environment required to support it. Because the process of gathering and maintaining this information is so complex, a distributed network of authoritative format information registries is expected to emerge.

Format registries:

- The National Archives (U.K.), *PRONOM* website http://www.nationalarchives.gov.uk/pronom/. PRONOM is an online registry of information about file formats, and the hardware platforms and software applications that support them.
- *Global Digital Format Registry* website http://hul.harvard.edu/gdfr/about.html. The GDFR, still under development, "will provide sustainable distributed services to store, discover, and deliver representation information about digital formats." The website links to the project's wiki, with detailed working documents.
- Digital Curation Center, *Representation Information Registry Repository* website http://registry.dcc.ac.uk/omar/.

There is a large and growing body of literature on the preservation of digital material. Here are a few starting points:

- National Library of Australia, *PADI (Preserving Access to Digital Information)* website http://www.nla.gov.au/padi/. A comprehensive clearinghouse covering all aspects of digital preservation. Entries are annotated and new references are added regularly.
- Stanford University, *COOL (Conservation Online)* website http://sul-server-2.stanford.edu/. Covers both analog and digital preservation, including an excellent section on preservation of audio and video.
- Digital Curation Centre, *Digital Curation Manual* website http://www.dcc.ac.uk/resource/curation-manual/. A practitioners' guide to curation, archiving, and preservation, being released in single-topic installments.

- Arts and Humanities Data Service, *Moving Images and Sound Archiving Study* (2006) http://www.ahds.ac.uk/about/projects/archiving-studies/moving-images-sound-archiving-final.pdf.
- Arts and Humanities Data Service, *Digital Images Archiving Study* (2006) http://www.ahds.ac.uk/about/projects/archiving-studies/digital-images-archiving-study.pdf.

OBJECTS PRINCIPLE 3

Objects Principle 3: A good object is meaningful and useful outside of its local context.

A good digital object should be coherent, meaningful, and usable outside of the context in which it was created. Depending on the discipline, objects with these properties may be called "portable," "reusable," or "interoperable."

Assumptions about accessing and using the object that are valid locally may no longer hold in the wider networked environment. This means that the object must be both portable and self-explanatory:

- The object's metadata should be self-contained, include all pertinent information about the object, and comply with a standard metadata schema, so that the object's metadata can be more readily mapped from one schema to another depending on the context of use. See METADATA Principle 2.
- The object's format and any technical requirements necessary for its use should be readily apparent.
- The object must carry with it a clear statement of acceptable users and uses to encourage use by authorized users.

In education, there is an emphasis on reusable learning objects, which are defined as chunks of instruction designed to teach a stand-alone learning objective. The more granular the object, the more easily it can be embedded within different pedagogical streams.

- *The Centre for Excellence in Teaching and Learning in Reusable Learning Objects* website http://www.rlo-cetl.ac.uk/.
- e-Learning Centre, *Learning Objects and Standards* website http://www.e-learningcentre.co.uk/eclipse/Resources/contentmgt.htm.

Interoperable objects are also the focus of efforts to link distributed digital libraries or repositories.

- *Australian Partnership for Sustainable Repositories (APSR)* website http://www.apsr.edu.au/. Many efforts of this initiative are devoted to interoperability of repositories of scholarly assets.
- *UKOLN Interoperability Focus* website http://www.ukoln.ac.uk/interop-focus/. Encompasses libraries, museums, archives, and other aspects of the cultural heritage, as well as government and community information.

OBJECTS PRINCIPLE 4

> **Objects Principle 4**: A good object will be named with a persistent, globally unique identifier that can be resolved to the current address of the object.

An identifier is a name assigned to an object according to a formal standard, an industry convention, or a local system providing a consistent syntax. Good identifiers will at minimum be locally unique, so that resources within the digital collection or repository can be unambiguously distinguished from each other. Global uniqueness can then be achieved through the addition of a globally unique prefix element, such as a code representing the organization.

Locally unique identifiers should be:

- scalable, so that many identifiers can be assigned without danger of running out or duplication;
- consistent, having a construction that can be easily applied over time;
- actionable, or capable of taking one to the object with a single "click" or action; and
- persistent, such that the identifier does not change when the location of the object changes.

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 name resolvers, software that uses a registry to map from the static persistent identifier to the current location of the object. Although the registry 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. Some identifier schemes utilizing name resolvers include PURLs, handles, and ARKs.

PURLs (Persistent URLs) are URLs resolved to true locations by a PURL server. OCLC runs a central PURL server that anyone can use. Alternatively, any organization can download and

install the free PURL server application (http://www.purl.org/) and manage its own PURL server locally.

The Corporation for National Research Initiatives (CNRI) developed the Handle System (http://www.handle.net/), 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 together with a local Handle application or running their own Handle application locally. The DOI (Digital Object Identifier) is a proprietary implementation of the Handle System (http://www.doi.org/). Use of DOI requires an annual membership fee to the International DOI Foundation to support maintenance of the DOI registry, metadata, and policy framework. Many commercial and open-source digital repository applications, including DigiTool, Fedora, and DSpace, can use the Handle System for object identification. Many electronic publishers, national libraries, and information consortia use DOI.

The Archival Resource Key (ARK) is a globally unique, actionable identifier scheme developed by the California Digital Library (http://www.cdlib.net/inside/diglib/ark/). CDL also provides an open source utility, NOID, which can be used to generate both ARK and handle identifiers (http://www.cdlib.net/inside/diglib/noid/). NOID can also be set up as a name resolver.

URLs and other Internet identifiers are types of Uniform Resource Identifiers (URI) (http://gbiv.com/protocols/uri/rfc/rfc3986.html). The INFO URI scheme provides a consistent way to represent and reference legacy identifiers so that they can be used by web applications (http://info-uri.info/). Some that have been registered to date include the Library of Congress Control Number (LCCN), PubMed identifier, DDC number, and OCLC WorldCat Control Number. The INFO URI scheme provides a lightweight method of registration that can be used instead of the more formal URN namespace registration process. A small number of legacy identifiers have been registered as URN namespaces, such as ISBN and ISSN.

Two emerging identifier specifications are XRI (eXtensible Resource Identifier) and IRI (Internationalized Resource Identifier). The IRI is a form of URI that supports internationalization by extending the character set to UNICODE characters and allowing up/down and right/left scanning in addition to left/right. The IRI specification is being developed by the W3C (http://www.w3.org/International/O-URL-and-ident.html).

The XRI builds on the IRI to identify resources independent of any specific physical network path, location, or protocol. Interestingly, XRI can be used for people as well as objects, and it can incorporate cross-references, such as an email address or website. The IRI specification is being developed by OASIS (http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xri).

It is important to understand that no identifier scheme or resolver system can guarantee persistence. Regardless of the technology used, for identifiers to remain persistent an institution must take responsibility for both the object and for the maintenance of its identifier.

Useful resources on developing an identifier strategy:

- Hans-Werner and Jochen Kothe, *Implementing Persistent Identifiers: Overview of Concepts, Guidelines and Recommendations* (2006) http://www.knaw.nl/ecpa/publ/pdf/2732.pdf. Written for the European Commission on Preservation and Access, this report explains the principle of persistent identifiers and helps institutions decide which scheme would best fit their needs.
- Harvard University Library Office for Information Systems, *Naming and Repository Services: An Introduction* http://hul.harvard.edu/ldi/resources/nrsdrsservice.pdf. Includes a gentle explanation of the importance of good practices in the design of naming services.
- *IMS Persistent, Location-Independent, Resource Identifier Implementation Handbook* (2001) http://imsglobal.org/implementationhandbook/imsrid_handv1p0.html. Using URNs for learning objects.
- International DOI Foundation, *DOI Handbook* (2006) http://www.doi.org/hb.html. Although all about the DOI, includes general explanations of many practical aspects of naming and name resolution.

OBJECTS PRINCIPLE 5

Objects Principle 5: A good object can be authenticated.

Authenticity refers to the degree of confidence a user can have in the integrity and trustworthiness of an object. Authentication is the act of determining that the object conforms to its documented origin, structure, and history, and that the object has not been corrupted or changed in an unauthorized way.

It is important to note that authenticity does not refer to the accuracy of the content or meaning of the object. As Clifford Lynch noted, "An authentic document may faithfully transmit complete falsehoods." (*Authenticity and Integrity in the Digital Environment*, http://www.clir.org/pubs/reports/pub92/lynch.html) Nonetheless, research and scholarship rely upon the ability to verify the authenticity of materials in order to use them appropriately. For archives, authenticity is an important component of the evidentiary value of records and has legal significance.

In the non-digital realm, the authenticity of documents is often determined through forensics such as paleography, examination of physical characteristics, and comparison of handwritten signatures. For digital objects, such physical clues do not exist, and the importance of documentation increases proportionately. The user wants to know the origin of the digital object, whether or not the object has been altered since its creation, and if so, how and by whom. Some methods of providing this information include documentation of digital provenance, watermarking, and fixity checking.

The digital provenance of an object is its origin and change history, which can be recorded as metadata (see METADATA). Origin information can be provided internally, often in the file header. Change history is most often recorded externally. The METS schema defines a placeholder section (digiProvMD) for digital provenance, but does not define any metadata elements to use within it. However, the Event Entity in the PREMIS *Data Dictionary for Preservation Metadata* defines semantic units that document digital provenance (http://www.oclc.org/research/projects/pmwg/premis-final.pdf). An XML schema for the PREMIS event entity can be used as a METS extension schema under digiprovMD (http://www.loc.gov/standards/premis/schemas.html).

Digital watermarking is a technique for adding a visible or invisible message to an object. Digital watermarks are most often used to assert copyright or ownership. Although watermarks may provide useful information similar to embedded origin information, they should be viewed cautiously as documentation of authenticity. (See Clifford Lynch, *Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust* http://www.clir.org/pubs/reports/pub92/lynch.html.)

The fixity of an object can be verified by comparison of message digests (often called checksums) generated from the object at different points in time. A message digest is a string created by applying an algorithm called a “hash function” to the bits comprising the object. The

message digest is saved and compared to a message digest created by the same algorithm at a later date. If they are the same, the object is bit-wise unchanged.

Context can also provide clues to authenticity. A good object will be related to other versions of the object, to other objects within a collection, and to host objects and/or contained objects. The archival profession has done both theoretical and practical work in preserving context and original order in the digital environment.

About authenticity:

- CLIR, *Authenticity in a Digital Environmen*t (2000) http://www.clir.org/pubs/reports/pub92/contents.html. Although getting dated, some of the essays in this compilation are still among the best on the topic, particularly Clifford Lynch's.
- DigiCULT, *Integrity and Authenticity of Digital Cultural Heritage Objects* (2002) http://www.digicult.info/downloads/thematic_issue_1_final.pdf. DigiCULT monitors, discusses, and analyses the impact of new technology on cultural and scientific heritage organizations. This publication gathers an eclectic but interesting set of primarily European perspectives.
- *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES Project* (2005) http://www.interpares.org/book/index.cfm. This report from the first stage of the international InterPARES project focuses on the preservation of the authenticity of records created and/or maintained in databases and document management systems in the course of administrative activities.
- National Library of Australia, *PADI (Preserving Access to Digital Information): Authenticity* website http://www.nla.gov.au/padi/topics/4.html. Well-maintained webliography of resources.

About message digests and watermarking:

- Fred Mintzer, Jeffrey Lotspiech, and Norishige Morimoto, *Safeguarding Digital Library Contents and Users: Digital Watermarking* (1997) http://www.dlib.org/dlib/december97/ibm/12lotspiech.html. A good basic explanation with illustrations.
- Richard Entlich, "A Little Bit'll Do You (In): Checksums to the Rescue," *RLG DigiNews*, v. 9, no. 3 (2005) http://www.rlg.org/en/page.php?Page_ID=20666#article3. General introduction to checksums and message digests.
- Wikipedia, *MD5* http://en.wikipedia.org/wiki/MD5. Might be more than you want to know about Message-Digest algorithm 5, but maybe not.
- Audrey Novak, *Fixity Checks: Checksums, Message Digests and Digital Signatures* (2005) http://www.library.yale.edu/iac/DPC/AN_DPC_FixityChecksFinal11.pdf. Best practices from Yale University Library.

OBJECTS PRINCIPLE 6

Objects Principle 6: A good object has associated metadata.

A good object will have descriptive and administrative metadata, and compound objects will have structural metadata to document the relationships between components of the object and ensure proper presentation and use of the components.

Metadata can often be embedded within an object and can be harvested for resource discovery and management purposes. Metadata can also be stored separately and linked to the resources described. Best practice is to encourage object creators to provide metadata at the time of object creation and to embed as much metadata in the object as feasible to increase portability and preservability. Examples of embedded metadata include META tags in web pages, XMP packets in PDF files, and UUID boxes in JP2 files. Whether embedded or not, metadata accessibility is critical and users must be able to read and understand metadata for them to be of value.

A good object may have more than one set of metadata associated with it, each reflecting the purposes of the individuals or organizations associated with it. For example, a creator may provide descriptive metadata at the time of object creation, while a publisher may supply administrative and structural metadata for managing and displaying the object. All of these metadata can be embedded in the object or stored separately and linked, directly or indirectly, to the object.

Objects and metadata can be packaged together in standardized containers, essentially creating new objects. Container standards used for digital collections and digital preservation include:

- *MPEG-21 Digital Item Declaration Language* http://xml.coverpages.org/mpeg21-didl.html or http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=41112&COMMID=&scopelist=.

- *Metadata Encoding and Transmission Standard (METS)* http://www.loc.gov/standards/mets/.

- *Sharable Content Object Reference Model (SCORM)* http://www.adlnet.gov/scorm/. Used primarily for learning objects.

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 (see COLLECTIONS Principle 2). 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, management, and preservation.

Metadata creation is an incremental process that should be a shared responsibility among various parts of an institution. Different types of metadata can be added by different people at various stages of an information object's life cycle. For example, at the creation stage, metadata about an object's authors, contributors, source, and intended audience could be provided by the original authors. At the organization stage, metadata about subjects, publishing history, and access rights could be recorded by catalogers or indexers. At the access and usage stage, evaluative information such as reviews and annotations could be added by the user. Creators of digital objects should be encouraged to embed as much metadata as possible within the object before it is shared or distributed. On the life cycle of information objects, see the article by Gail Hodges, and the first chapter of Baca, *Introduction to Metadata*, both cited below under "General introductions to metadata issues."

It is common to distinguish between three basic kinds of metadata. Descriptive metadata helps users find and obtain objects, distinguish one object or group of objects from one another, and discover the subject or contents. Administrative metadata helps collection managers keep track of objects for such purposes as file management, rights management, and preservation. Structural metadata documents relationships within and among objects and enables users to navigate complex objects, such as the pages and chapters of a book.

A primary reason for building digital 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, museum inventory control systems, and search utilities such as Google.

Over the years, various metadata schemes have been developed for describing different types of objects. Within this multiplicity of schemes, there is a degree of consistency that supports interoperability. For example, most schemes provide for a creator or contributor name, date, title, and identifier. As 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 ensure the greatest likelihood of interoperability (see Metadata Principle 2 and Objects Principle 3). Institutions must carefully consider not only which metadata schemes and information protocols are best suited to their collections; they must also give considerable thought to which controlled vocabularies and thesauri they should implement (see Metadata Principle 3), and which data content (i.e., cataloging) standards are

most suitable for the objects in their collections. There are long-established cataloging guidelines such as AACR (*Anglo-American Cataloguing Rules*), and recent, new, and emerging standards such as DACS (*Describing Archives: A Content Standard*), CCO (*Cataloging Cultural Objects)*, and RDA (*Resource Description and Access*). The cataloging standard that an institution chooses to follow, or the adaptation or combination of cataloging standards selected, is a key factor for providing good end-user access and creating sharable metadata records that work well in aggregated collections. See "Guidelines for Use" in the chart under Metadata Principle 1.

The following table, taken from Anne Gilliland's essay in *Introduction to Metadata* (revised edition, 2008, cited below), provides a typology of data standards and how they should work together, with examples.

TABLE 3: TYPOLOGY OF DATA STANDARDS

Type of Data Standard	Examples
Data *structure* standards (metadata element sets, schemas). These are "categories" or "containers" of data that make up a record or other information object.	*the set of MARC (Machine-Readable Cataloging format) fields, Encoded Archival Description (EAD), Dublin Core Metadata Element Set (DCMES), Categories for the Description of Works of Art (CDWA), VRA Core Categories*
Data *value* standards (controlled vocabularies, thesauri, controlled lists). These are the terms, names, and other values that are used to populate data structure standards or metadata element sets.	*Library of Congress Subject Headings (LCSH), Library of Congress Name Authority File (LCNAF), LC Thesaurus for Graphic Materials (TGM), Medical Subject Headings (MeSH), Art & Architecture Thesaurus (AAT), Union List of Artist Names (ULAN), Getty Thesaurus of Geographic Names (TGN), ICONCLASS*
Data *content* standards (cataloging rules and codes). These are guidelines for the format and syntax of the data values that are used to populate metadata elements	*Anglo-American Cataloguing Rules (AACR), Resource Description and Access (RDA), International Standard Bibliographic Description (ISBD), Cataloging Cultural Objects (CCO), Describing Archives: A Content Standard (DACS)*
Data format/technical interchange standards (metadata standards expressed in machine-readable form). This type of standard is often a manifestation of a particular data structure standard (type 1 above), encoded or marked up for machine processing.	*MARC21, MARCXML, EAD XML DTD, METS, MODS, CDWA Lite XML schema, Simple Dublin Core XML schema, Qualified Dublin Core XML schema, VRA Core 4.0 XML schema*

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, 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 uncontrolled 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 or other controlled vocabulary, for example, can provide greater precision and recall in searching, and can enable future functionality, such as faceted subject browsing and dynamic searching of subject matter.

The decisions about which metadata standard(s) to adopt and what levels of description to apply must be made within the context of the organization's purpose for creating the collection, the available human and technical resources, the users and intended usage, and approaches adopted within the particular field of inquiry or knowledge domain.

Questions to consider include, but are not limited to:

- What is the purpose of the digital collection?
- What are the goals and objectives for building this collection?
- Who are the targeted users? What information do they need, and what is their typical information-seeking behavior?
- Are the materials to be accessed at the collection level or as individual items, or both?
- Do multiple versions or manifestations of the object need to be distinguished from each other?
- Will the collection or its objects have metadata before the digital collection is built?
- What subject discipline will be involved? What are the metadata standards that are commonly used within this discipline?
- What metadata standards are used by organizations in this domain? Which ones are most appropriate for this particular collection?
- How rich a description is needed, and does the metadata need to convey hierarchical relationships?

Institutions should be aware that, depending upon the nature of their collections, a single metadata scheme may not suffice for all their needs. Thus a judicious combination of metadata schemes may be the best solution for some materials—for example, using EAD as the scheme at the collection level for archival collections with a common provenance, and MODS, VRA Core 4.0, CDWA Lite or another appropriate scheme at the item level. Likewise, a well-thought out selection of controlled vocabularies, published and collection-specific, should be applied as the data values to populate key access elements within the selected schemes.

Metadata Principle 1: Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.

Metadata Principle 2: Good metadata supports interoperability.

Metadata Principle 3: Good metadata uses authority control and content standards to describe objects and collocate related objects.

Metadata Principle 4: Good metadata includes a clear statement of the conditions and terms of use for the digital object.

Metadata Principle 5: Good metadata supports the long-term curation and preservation of objects in collections.

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.

METADATA

METADATA PRINCIPLE 1

> **Metadata Principle 1**: Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.

It is essential to conform to, or at the very least map to, known community standards for metadata, rather than using proprietary or homegrown schemes. However, simply because a particular metadata scheme is considered a standard does not necessarily mean that it is the appropriate standard for any given collection. For example, EAD is a well-established standard for describing intact archival collections with a common provenance, but it is not the best scheme for describing heterogeneous cultural heritage collections composed of objects that all have a different provenance.

One of the very first steps in implementing a metadata strategy is to analyze and identify the most appropriate metadata standard—or set of standards—for your collections. The metadata scheme(s) and controlled vocabularies and thesauri that have been developed for specific communities and types of materials should be carefully researched and analyzed, and the most appropriate ones selected and implemented.

There are a variety of published metadata schemes that can be used for digital objects. The book *Metadata Fundamentals for All Librarians* (P. Caplan, 2003, see below) 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 nature of the collections themselves, the level of resources that the institution has to devote to metadata creation, the level of expertise of the metadata creators, the expected use and users of the collection, the goal of enabling interoperability and sharing digital collections, and other factors.

Organizations should consider the granularity of description, that is, whether to create descriptive records at the collection level, at the series or group levels, at the item level, or at multiple levels, 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.

In some cases, the best strategy may be to utilize two or more metadata schemes in an integrated manner. For example, MARC or EAD might be used at the collection or group level, and MODS or CDWA Lite or VRA Core might be used to describe individual items within those collections or groups. METS could be used as a metadata "wrapper" to associate metadata expressed in various schemes.

Simply identifying the appropriate metadata scheme(s) for your collections is not sufficient; in most cases, institutions also need to develop and implement their own local "application

profile" for the selected scheme, specifying exactly what will be done in those areas where the scheme allows for various options. Once a community metadata standard has been selected for application to a particular collection or group of collections, a detailed profile that specifies how that scheme should be implemented locally should be developed and clearly documented.

Application profiles make it possible to combine metadata elements from multiple existing metadata schemes. Definitions, requirements, best practices, and qualifiers from the original schema may be modified or added as needed for the particular application profile. Profiles also make it possible to add local elements to an existing standard scheme.

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

TABLE 4: METADATA CHART

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
CDWA LITE http://www.getty.edu/research/conducting_research/standards/cdwa/cdwalite.html	An XML schema for core records for works of art and material culture based on *Categories for the Description of Works of Art* (http://www.getty.edu/research/conducting_research/standards/cdwa/).	The CDWA site includes the XML schema as well as the specification/data dictionary. Cataloging examples are at http://www.getty.edu/research/conducting_research/standards/cdwa/examples.html. The CDWA Lite schema assumes the use of *Cataloging Cultural Objects (CCO)*, comprehensive guidelines developed by the art information, visual resources, and museum communities for describing cultural works, including art, architecture, objects of material culture, and their images: http://www.vraweb.org/ccoweb/cco/index.html. As of this writing, OCLC/RLG Programs is hosting a working group to help museums implement the CDWA Lite XML schema: http://www.rlg.org/en/page.php?Page_ID=335.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
CIDOC CRM http://cidoc.ics.forth.gr/	A conceptual reference model or "reference ontology" that provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. The CIDOC CRM has been an official ISO standard (ISO 21127) since late 2006: http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=34424&scopelist=PROGRAMME	A variety of tools for implementation and mapping are available at http://cidoc.ics.forth.gr/tools.html.
copyrightMD http://www.cdlib.org/inside/projects/rights/schema/	An XML schema for rights metadata developed by the California Digital Library (CDL); designed for incorporation with other XML schemas for descriptive and structural metadata (e.g., CDWA Lite, MARC XML, METS, and MODS).	Full record examples for materials with various types of rights metadata are at http://www.cdlib.org/inside/projects/rights/schema/examples.html.
Darwin Core http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome	A metadata element set developed to provide for the geographic occurrence of species and the existence of specimens in collections	The Darwin Core wiki site http://wiki.tdwg.org/twiki/bin/view/DarwinCore/WebHome describes the Darwin Core elements and extensions, and hosts discussions on revisions. The Mammal Networked Information Systems (MaNIS) utilizes Darwin Core in its data portal: http://manisnet.org/index.html.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
Dublin Core http://dublincore.org	A relatively simple, generic metadata 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 (e.g. GEM, for K-12 education metadata: http://www.thegateway.org/about/documentation/schemas), and has been incorporated into several domain-specific metadata schemes). As of this writing, Dublin Core XML is the required basic XML schema for OAI harvesting, and is often used as the "lowest common denominator" in metadata crosswalks.	Encoding guidelines from the Dublin Core Metadata Initiative (DCMI) are at http://dublincore.org/resources/expressions/. See also *Collaborative Digitization Program (CDP) Dublin Core Metadata Best Practices* http://www.cdpheritage.org/cdp/documents/cdpdcmbp.pdf Dublin Core XML records can be harvested via the OAI Protocol for Metadata Harvesting (OAI/PMH): http://www.openarchives.org/pmh/ (OAI home page); http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/Main_Page (Digital Library Federation best practices for OAI data providers). For other OAI-harvestable XML schemas that may be more appropriate for specific types of collections, see: http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/MultipleMetadataFormats/

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
Encoded Archival Description (EAD) http://www.loc.gov/ead/	A set of elements and 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. RLG, EAD Advisory Group, *RLG Best Practice Guidelines for Encoded Archival Description* http://www.rlg.org/en/page.php?Page_ID=450. (http://www.rlg.org/en/pdfs/bpg.pdf) Online Archive of California, *OAC Best Practice Guidelines for EAD*http://www.cdlib.org/inside/diglib/guidelines/bpgead. *The EAD Cookbook*, version 2.0 http://www.archivists.org/saagroups/ead/ead2002cookbook.html.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
IPTC Core Schema for XMP http://www.iptc.org/IPTC4XMP/	A metadata schema intended for use within Adobe's eXtensible Metadata Platform (XMP) framework (see http://www.adobe.com/products/xmp/). Files created using Adobe's Creative Suite of software tools (e.g., Photoshop) contain embedded XMP metadata, thus making it possible to automatically capture and embed technical metadata in image files.	User guidelines can be found at http://www.iptc.org/std/Iptc4xmpCore/1.0/documentation/Iptc4xmpCore_1.0-doc-CpanelsUserGuide_13.pdf. The IPTC Core specification can be found at http://www.iptc.org/IPTC4XMP/
Learning Object Metadata http://ltsc.ieee.org/wg12/	Learning Object Metadata is used to describe educational resources in course management systems and learning management systems. Learning objects are also collected in institutional and statewide repositories. The main standard is the *IEEE Standard for Learning Object Metadata* (1484.12.1-2002) (http://ieeexplore.ieee.org/iel5/8032/22180/01032843.pdf?arnumber=1032843), also called the LOM, which must be ordered from IEEE. However, the LOM has been incorporated into a number of other standards, including the IMS Global Learning Consortium's *Meta-Data Specification* (http://www.imsproject.org/metadata/), which is freely available from the IMS.	*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. CanCore is the official site for documents, presentations and other resources related to the CanCore Learning Resource Metadata Initiative, which uses LOM: http://www.cancore.ca/en/index.html.
MARC 21 http://lcweb.loc.gov/marc/	A long-established standard for exchanging bibliographic records, developed and maintained by the library community. Over the last several	Library of Congress, *Understanding MARC Bibliographic: Machine-Readable Cataloging*, 7th Edition http://lcweb.loc.gov/marc/umb/.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
	years, MARC has been enhanced to support descriptive elements for electronic resources. There is a MARC Lite scheme (http://www.loc.gov/marc/bibliographic/lite/), as well as a MARC XML schema (http://www.loc.gov/standards/marcxml/).	OCLC's *Bibliographic Formats and Standards* (http://www.oclc.org/bibformats/) provides tagging guidelines and rules for inputting MARC records into WorldCat. Most libraries that use MARC21 use AACR (currently evolving into Resource Description and Access, RDA). *Anglo-American Cataloguing Rules, second edition, 2002 revision* (Chicago: ALA Editions, 2005. However, MARC21 is language-neutral and data-content-standard-neutral, and can also be used in conjunction with DACS, CCO, and descriptive cataloging codes developed for non-English-language catalogs. Information on RDA, still an evolving cataloging standard at the time of publication, can be found at http://www.collectionscanada.ca/jsc/rdaprospectus.html.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
Metadata Encoding and Transmission Standard (METS) http://www.loc.gov/standards/mets	An XML schema for encoding structural metadata about complex digital objects. METS also acts as a container with places to insert descriptive, administrative, and technical metadata.	The METS Implementation Registry includes projects that have been fully implemented, as well as projects in the planning and implementation stages: http://www.loc.gov/standards/mets/mets-registry.html. Registered METS profiles are available at http://www.loc.gov/standards/mets/mets-registered-profiles.html.
MIX (NISO Metadata for Images in XML) http://www.loc.gov/standards/mix/	An XML schema comprising a set of technical data elements required to manage digital image collections. The schema provides a format for interchange and/or storage of data.	An example of a MIX XML document is at: http://www.loc.gov/standards/mix/instances/test_mix10.xml.
MODS (Metadata Object Description Schema) http://www.loc.gov/standards/mods	An XML schema for descriptive metadata compatible with the MARC 21 bibliographic format.	*MODS User Guidelines* are at: http://www.loc.gov/standards/mods/v3/mods-userguide.html.
MPEG-7, Multimedia Content Description Interface (ISO/IEC 15938) http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm. The standard can be purchased from the International Organization	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 (http://gondolin.rutgers.edu/MIC/text/how/cataloging_utility.htm). The IBM alphaWorks development team has

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
for Standardization (ISO) http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34232		released a downloadable MPEG-7 Annotation Tool (http://www.alphaworks.ibm.com/tech/videoannex) to annotate video sequences with MPEG-7 metadata.
Object ID http://icom.museum/object-id/	An international standard for describing cultural objects, primarily with a few to documenting and protecting them as cultural property and protecting them from illicit traffic. Maintained and disseminated by the International Council of Museums (ICOM) in collaboration with UNESCO.	The Object ID checklist is available at: http://icom.museum/object-id/checklist.html. "Introduction to Object ID" (http://icom.museum/object-id/guide/guide_index.html) provides detailed guidelines for the implementation of this relatively simple set of metadata elements.
PBCore Public Broadcasting Metadata Dictionary Project http://www.pbcore.org/	A metadata dictionary for television, radio, and web activities.	Implementation just beginning during 2007.
PREMIS Data Dictionary for Preservation Metadata http://www.oclc.org/research/projects/pmwg/premis-dd.pdf	A set of core preservation metadata elements developed by an international working group, Preservation Metadata: Implementation Strategies (PREMIS).	The PREMIS working group's final report (http://www.oclc.org/research/projects/pmwg/premis-report.pdf) includes details on methodology and implementation. A registry of institutions and vendors that are implementing this standard is at http://www.loc.gov/standards/premis/premis-registry.php.
Society of Motion Picture	A registry of metadata element descriptions for use	None available at time of publication.

METADATA SCHEME	DESCRIPTION	GUIDELINES FOR USE, AND APPLICATIONS
and Television Engineers SMPTE Metadata Dictionary http://www.smpte-ra.org/mdd	with video, audio, or other data.	
Spectrum http://www.mda.org.uk/specfaq.htm	The UK standard for museum documentation, consisting of procedures and information requirements.	Guidelines for the use of Spectrum are included in the downloadable version, which is available for licensing fee-free: http://www.mda.org.uk/spectrum.htm
VRA Core Categories Version 4.0 http://www.vraweb.org/projects/vracore4/index.html	An XML schema developed by the Visual Resources Association for the description of art, architecture, works of material culture, with emphasis on visual surrogates of such works. The VRA Core Categories were designed with the awareness that there are often multiple representations and views of a work of art, architecture, or material culture.	Like CDWA Lite, VRA Core 4.0 assumes the use of CCO for cataloging guidelines: http://www.vraweb.org/ccoweb/cco/index.html Descriptions of the metadata elements and tagging examples are at http://www.vraweb.org/projects/vracore4/VRA_Core4_Element_Description.pdf.

General introductions to metadata issues:

- Murtha Baca, editor, *Introduction to Metadata: Pathways to Digital Information* (version 2.1 available online; version 3.0 forthcoming in 2008) http://www.getty.edu/research/conducting_research/standards/intrometadata/.
- Priscilla Caplan, *Metadata Fundamentals for All Librarians* (Chicago: ALA Editions, 2003).
- CHIN (Canadian Heritage Information Network), *Metadata Standards for Museum Cataloguing* website http://www.chin.gc.ca/English/Standards/metadata_intro.html.
- Jane Greenberg, *Understanding Metadata and Metadata Schemes* (2005). http://www.ils.unc.edu/mrc/pdf/greenberg05understanding.pdf.
- Gail Hodges, "Best Practices for Digital Archiving: An Information Life Cycle Approach," *D-Lib Magazine*, v. 6, no. 1 (2000) http://www.dlib.org/dlib/january00/01hodge.html.
- National Information Standards Organization (NISO), *Understanding Metadata* (2004) http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.
- Technical Advisory Service for Images (TASI), *Metadata Overview* (2006) http://www.tasi.ac.uk/advice/delivering/metadata.html.
- Technical Advisory Service for Images (TASI), *Getting Practical with Metadata* (2006) http://www.tasi.ac.uk/advice/delivering/metadata-practical.html.

Best practices:

- DLESE (Digital Library for Earth System Education), *Metadata Best Practices* website http://www.dlese.org/Metadata/collections/metadata-best-practices.htm.
- DLF (Digital Library Federation), *Best Practices for OAI Data Provider Implementations and Shareable Metadata* http://webservices.itcs.umich.edu/mediawiki/oaibp/?PublicTOC.

Portals to metadata resources:

- International Federation of Library Associations (IFLA), *Digital Libraries: Metadata Resources* website http://www.ifla.org/II/metadata.htm.
- UKOLN *Metadata* website http://www.ukoln.ac.uk/metadata/.

Application profiles:

- Thomas Baker, et al., *Dublin Core Application Profile Guidelines* (2005) http://dublincore.org/usage/documents/profile-guidelines/.
- R. Heery, and M. Patel, "Application Profiles: Mixing and Matching Metadata Schemas" in *Ariadne* 25 (2000) http://www.ariadne.ac.uk/issue25/app-profiles/intro.html.
- *International Conference on Dublin Core and Metadata Applications DC-2007* website http://www.dc2007.sg/. The conference theme is "Application Profiles: Theory and Practice."
- Western States Digital Standards Group Metadata Working Group, *Western States Dublin Core Best Practices* (2003) http://content.lib.utah.edu/cgi-

bin/showfile.exe?CISOROOT=/docs_regional&CISOPTR=1. An example of a general application profile for a digital collections project with many contributing institutions.

- Victoria Online, *Metadata Application Profile and Taxonomy Guidelines* (2006) http://egov.vic.gov.au/pdfs/VomapGuidelinesTaxv4.1_Final-Dec2006.pdf. A profile for an e-government portal.

METADATA PRINCIPLE 2

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.

Good metadata should be coherent, meaningful, and useful in global contexts beyond those in which it was created. This means that it must include all pertinent information about the object, since assumptions about the context in which it is accessed locally may no longer be valid in the wider networked environment. For example, a photo archive may not indicate in each record that the object being described is a photograph. However, in the wider network context, form and genre information becomes important. Digital collections with a topical focus are notorious for creating non-interoperable metadata when they assume that users know the main topic of the collection. When this metadata is shared in larger aggregations, descriptions that made sense in the context of the original collection can be mystifying. This has been dubbed the "on a horse" problem, from the description of a photograph in Harvard's Teddy Roosevelt collection, where the title assigned to the photograph did not indicate who was sitting on the horse, since all the materials in the collection related to Roosevelt.

The creation of accessible, meaningful shared collections implies responsibilities on both the part of the data providers (organizations that create metadata records and contribute them to federated collections) and service providers (aggregators that provide access to federated collections or union catalogs). Data providers should strive to create consistent, standards-based metadata, to use appropriate controlled vocabularies and thesauri, and to follow appropriate data content (i.e., cataloging) standards. Service providers must implement metadata normalization, remediation, and enhancement, and should, as their name implies, provide additional "value-added" services such as vocabulary-assisted searching, subject clustering, terminology mapping, and other enhancements. Adherence to appropriate standards and collaboration between data providers and service providers are crucial elements of effective aggregated digital collections.

- Sarah Shreves, Jenn Riley and Liz Milewicz, "Moving Toward Shareable Metadata," *First Monday*, v. 11, no. 8 (2004) http://www.firstmonday.org/issues/issue11_8/shreeves/index.html. How local descriptions fail in aggregations.
- Scottish Museums Council, *Metadata Interoperability Project* website http://cms.cdlr.strath.ac.uk/about.html.
- William Y. Arms et. al., "A Spectrum of Interoperability: The Site for Science Prototype for the NSDL," D-*Lib Magazine*, v. 8, no. 2 (2002) http://www.dlib.org/dlib/january02/arms/01arms.html.
- Technical Advisory Service for Images (TASI), *Metadata Standards and Interoperability* website http://www.tasi.ac.uk/advice/delivering/metadata-standards.html.

- Roy Tennant, "Metadata's Bitter Harvest," *Library Journal* (2004) http://www.libraryjournal.com/article/CA434443.html?display=Digital+LibrariesNews&industry=Digital+Libraries&industryid=3760&verticalid=151.

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.

Ideally, metadata schemes 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 scheme 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 as well as human users.

- *ISO/IEC 11179-3: 2003(E), Information Technology – Metadata Registries (MDR) – Part 3: Registry metamodel and basic attributes* http://standards.iso.org/ittf/PubliclyAvailableStandards/c031367_ISO_IEC_11179-3_2003(E).zip. The metadata registry standard 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.
- The Moving Image Collections (MIC), *Core Data Element Registry* website http://gondolin.rutgers.edu/MIC/text/how/unioncat_registry_table_04_23.htm shows how one project used a simple 11179 registry.

When different metadata schemes must be used, one way to achieve interoperability is to map elements from one scheme to those of another. 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 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 needs.

- Getty Research Institute, *Metadata Standards Crosswalks* website http://www.getty.edu/research/conducting_research/standards/intrometadata/crosswalks.html. Crosswalks relevant to art, architecture, and cultural heritage information maintained by The Getty Standards Program.
- Library of Congress, *MARC Standards: MARC21 Formats* website http://www.loc.gov/marc/marcdocz.html. Mappings (crosswalks) to and from MARC 21.
- National Science Digital Library (NSDL), *NSDL Metadata Resources Page* website http://metamanagement.comm.nsdlib.org/IntroPage.html. A metadata resources page largely devoted to crosswalks.

- University of Washington, *Metadata Implementation Group* website http://www.lib.washington.edu/msd/mig/datadicts/default.html. The library maintains mappings to Dublin Core from the data dictionaries used by each of its digital collections.

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

- *Open Archives Initiative* website http://www.openarchives.org/. Links to the *Protocol for Metadata Harvesting* and guidelines for implementers.
- *OAIster* website http://www.oaister.org/. The University of Michigan's OAIster search service contains millions of records for digitized cultural heritage materials harvested from hundreds of collections via the OAI-PMH.
- *Best Practices for OAI Data Provider Implementations and Shareable Metadata* website http://webservices.itcs.umich.edu/mediawiki/oaibp/index.php/Main_Page. A joint initiative between the Digital Library Federation and the National Science Digital Library.

Yet 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, which is being modernized for the web environment.

- Library of Congress, *SRU: Search/Retrieve via URL* website http://www.loc.gov/standards/sru/. A standard protocol for passing Z39.50-like search queries in a URL, utilizing a Common Query Language. This site also links to the SRW (Search/Retrieve Web Service) specification, in which queries are passed not via URL as in SRU, but by using XML over HTTP using SOAP (Simple Object Access Protocol).

METADATA PRINCIPLE 3

> **Metadata Principle 3**: Good metadata uses authority control and content standards to describe objects and collocate related objects.

Attributes of distributed objects should be expressed using standard controlled terms whenever possible. These include, but are not limited to, personal names, corporate names, place names, titles of works, subjects, and genre headings. Names and titles should be formulated according to standard descriptive cataloging rules; subject and genre terms should be taken from controlled vocabularies and thesauri. 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, the nature of the collections being described, the resources of the institution, and user expectations. Factors to consider 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?

To enable the most effective end-user access, the implementation of local, collection-specific authorities and vocabularies in addition to the use of terms and names from standard published authorities is often the best strategy. Whatever combination of vocabularies is chosen, their use should be carefully documented and in-house guidelines should be provided to help metadata creators select terms consistently. Authors and other untrained metadata creators cannot generally be counted on to use controlled vocabularies successfully unless the authority list is very short and simply organized.

The High Level Thesaurus Project (HILT, http://hilt.cdlr.strath.ac.uk/Sources/index.html) is a clearinghouse of information about controlled vocabularies, including related resources, projects, and an alphabetical list of thesauri.

Some organizations maintain suites of thesauri for use within specific domains:

- *The Getty Vocabulary Program* website http://www.getty.edu/research/conducting_research/vocabularies/aat/. The Getty builds, maintains, and disseminates several thesauri for the visual arts, architecture, and

material culture. The *Art & Architecture Thesaurus* (AAT) is also available in Spanish (http://www.aatespanol.cl/) and Dutch (http://www.aat-ned.nl/).
- MDA, *Terminology Bank* website http://www.mda.org.uk/spectrum-terminology/termbank.htm. The MDA (formerly known as the Museum Documentation Association) builds, maintains and disseminates thesauri for museum objects, including vocabularies for describing archaeological objects, waterways, railways, costumes, and aircraft types.
- *Library of Congress Authorities* website http://authorities.loc.gov/. The Library of Congress builds, maintains, and disseminates authority files for bibliographic description, including a controlled list of subject headings and a file containing authorized forms of personal and corporate names, titles, and name/title headings.

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, 1988). Not available on the Web, this resource is used by many small museums and historical societies. All of the terminology from Chenhall's "Nomenclature" that falls within the scope of the *Art & Architecture Thesaurus* has been included in the AAT.
- *ICONCLASS* website http://www.iconclass.nl/. A classification system, consisting of alphanumeric notations, textual correlates, and related keywords, for describing the narrative and iconographic content of works of art and other visual materials. The master version is in English; German, Italian, French, and Finnish translations are also available.
- *Thesaurus for Graphic Materials (TGM) I: Subject Terms* (1995) http://lcweb.loc.gov/rr/print/tgm1/.
- *Thesaurus for Graphic Materials (TGM) II: Genre and Physical Characteristics Terms* (2004) http://lcweb.loc.gov/rr/print/tgm2/.
- U.S. Geological Survey, *U.S. Board on Geographic Names' Geographic Names Information System* website http://geonames.usgs.gov/.

Classification systems available on the Web include:

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

OCLC's *Terminologies Service* website http://www.oclc.org/terminologies/. Provides metadata-building tools combined with access to a range of controlled vocabularies and thesauri, including the Getty vocabularies, MeSH, and TGM I and II.

METADATA PRINCIPLE 4

> **Metadata Principle 4**: Good metadata includes a clear statement of the conditions and terms of use for the digital object.

Terms and conditions of use include the copyright status of the object—whether it is in the public domain or is copyright protected—and any restrictions on use. The user should be informed how to obtain permission for restricted uses and how to cite the material for allowed uses. The institution should also document whether the resource is published or unpublished, and whether the creator or rights holder is known. Contact information for rights holders should be maintained.

If this information is the same for all the materials in a collection, documenting it in collection-level metadata is adequate (see COLLECTIONS). Otherwise, it should be recorded at the object level.

Many metadata schemes have designated places to put this information; if they do not, an external scheme or locally defined element set should be used.

Rights metadata is a rapidly evolving area. Rights metadata is generally expressed in XML (eXtensible Markup Language) and may serve three complementary but distinct purposes:

- **Rights description**, which may include the description of the copyright status of works, rights holder requirements for use of the resource, and user attributes required for authorized use of a resource and agreements between both parties for resource use. PREMIS rights metadata, <indecs>rdd (rights data dictionary) and Creative Commons licenses are examples of rights description. The California Digital Library's *copyrightMD Schema* is a rights description schema for recording detailed copyright information that may evolve into a standard (http://www.cdlib.org/inside/projects/rights/schema/).

- **Rights licensing** is an emerging area of rights management within the library environment focused on the development and exchange of license information for resources. ONIX-PL (ONIX for Publications Licenses) and the PLUS License Data Format are metadata schema for communicating license terms for library subscriptions and picture images respectively.

- **Rights workflow** – rights expression languages support rights transactions between the rights holder and the user. They are designed to be actionable within a suite of standards and protocols to manage the digital workflow of rights management, whether it is the authorization of users, enforcement of rights agreements, control of resource access, tracking of resource use, or all of the above. Rights workflow generally incorporates licenses but goes beyond simple license messaging to providing an end-to-end actionable platform for managing agreements between parties. R XrML, the core technology within the MPEG-21 rights expression language, ORDL (Open Rights Description Language) and XACML (eXtensible Access Control Markup Language) are examples of rights expression languages for workflow.

Rights metadata has the distinction of being the only legally enforceable type of metadata. The WIPO Copyright Treaty (WCT) and WIPO Performances and Phonograms Treaty (WPPT) are international copyright treaties that have been ratified and incorporated into the national law of most signatory countries, including the United States and members of the European Union. The WCT and WPPT treaties require that signatory countries provide legal remedies against any party that knowingly removes or alters rights management information, where this information is defined as "information which identifies the work, the author of the work, the owner of any right in the work, or information about the terms and conditions of use of the work, and any numbers or codes that represent such information, when any of these items of information is attached to a copy of a work or appears in connection with the communication of a work to the public" (*WIPO Copyright Treaty, art. 12,* Geneva, SZ: World Intellectual Property Organization, http://www.wipo.int/treaties/en/ip/wct/trtdocs_wo033.html - P66_786\5).

Institutions are likely to encounter this metadata in file headers, particularly file headers for images and documents that utilize XMP, the extensible metadata platform, that utilizes RDF (Resource Description Framework) to provide data and storage models for incorporating and handling metadata within file headers. Adobe introduced XMP in 2001, and its adoption along the digital object creation and management chain has been steadily increasing. XMP is predominantly used for documents and images but is extensible to most digital file formats.

Other avenues for incorporating metadata within digital objects include the metadata track in the MPEG-4 multimedia file format, and metadata support within MP-3 and ogg container format for digital multimedia. Institutions can expect to encounter rights metadata, which cannot be removed or altered by law, in many commercially distributed digital objects, such as resources that are licensed from publishers or distributors. The metadata that is integrated into digital objects may contain useful information about the creation and provenance of the object as well as permissions and restrictions on use that can be useful to populate metadata databases through automatic data capture, as long as the data capture does not modify, delete or interfere with the actionability of the metadata.

METADATA PRINCIPLE 5

> **Metadata Principle 5**: Good metadata supports the long-term management, curation, and preservation of objects in collections.

Administrative metadata is information intended to facilitate the management of resources. It includes information such as when and how an object was created, who is responsible for controlling access to or archiving the content, what processing activities have been performed in relation to it, and what restrictions on access or use apply.

Technical metadata and preservation metadata are particular types of administrative metadata. Technical metadata describes digital files and includes capture information, format, file size, checksum, sampling frequencies, and similar characteristics. Technical metadata may be necessary to ensure the continued usability of an object, or to reconstruct the object if it is damaged.

Preservation metadata supports the long-term retention of digital objects. It may include detailed technical metadata as well as information related to the object's context and relationships, custody and change history, processing, storage and status. It should, therefore, be compatible with the collections management workflow of the archiving institution. In some cases, this may require a negotiation to resolve institutional workflow and digital object descriptions.

Recordkeeping metadata documents and facilitates the systematic creation, use, maintenance, and disposition of records to meet administrative, programmatic, legal, and financial needs and responsibilities. It is of primary interest to archivist and records managers.

Structural metadata relates the pieces of a compound object together and/or bundles related objects into a package. For example, if a book is digitized as individual page images, structural metadata can record 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.

Preservation metadata:

- Library of Congress, *PREMIS Preservation Metadata Maintenance Activity* website http://www.loc.gov/standards/premis/. The PREMIS Data Dictionary is a core set of metadata elements for preservation, with "core" being defined as "what most preservation repositories will need to know, most of the time." PREMIS has become the de facto standard for basic preservation metadata in the English-speaking world. It has an active maintenance activity and implementers group.
- Deutsche Nationalbibliothek, *LMER Long-term Preservation Metadata for Electronic Resources* website http://www.ddb.de/eng/standards/lmer/lmer.htm. A schema used in Germany in preference to PREMIS.
- *Preserving Access to Digital Information (PADI)* website http://www.nla.gov.au/padi/. Includes an extensive annotated listing of resources related to preservation metadata.

Technical metadata:

- *ANSI/NISO Z39.87-2006, Data Dictionary – Technical Metadata for Digital Still Images*http://www.niso.org/standards/standard_detail.cfm?std_id=731. One of the few formal standards for technical metadata. It focuses on images created by scanning. The XML expression of this data set is the MIX schema (http://www.loc.gov/standards/mix/).
- In development are two AES standards for administrative metadata (roughly speaking, the equivalent of the NISO imaging data dictionary and MIX): AES-X098B, *Audio Object Schema,* and AES-X098C, *Process History Schema*.
- *JHOVE - JSTOR/Harvard Object Validation Environment* website http://hul.harvard.edu/jhove/. JHOVE is an open source tool for automated extraction of technical metadata which focuses on open audio, video, image, and text formats.
- National Library of New Zealand, *Metadata Extraction Tool* (2007) http://meta-extractor.sourceforge.net/. This is an open source tool for automated extraction of technical metadata that includes handling formats created by common office applications.

Recordkeeping metadata:

- Commonwealth of Australia, *Recordkeeping Metadata Standard for Commonwealth Agencies* (1999) http://www.naa.gov.au/Images/rkms_pt1_2_tcm2-1036.pdf.
- Minnesota Historical Society, *Minnesota Recordkeeping Metadata Standard* (2003) http://www.mnhs.org/preserve/records/metadatastandard.html. An example of a state standard.

Structural metadata:

- Library of Congress, *Metadata Encoding and Transmission Standard (METS)* website http://www.loc.gov/standards/mets/. METS is the most widely used packaging standard in the cultural heritage community. METS specifies how to represent structural metadata for an object, and also provides a framework for associating descriptive and administrative metadata.
- *ISO/IEC 21000-2:2005 Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration* http://standards.iso.org/ittf/PubliclyAvailableStandards/c041112_ISO_IEC_21000-2_2005(E).zip. The Digital Item Declaration Language (DIDL) is also used to package cultural heritage objects.
- IMS Global Learning Consortium, Inc., *IMS Content Packaging Information Model,* version 1.1.2 (2001) http://www.imsglobal.org/content/packaging/cpv1p1p2/imscp_infov1p1p2.html. Used primarily in the education community.

METADATA PRINCIPLE 6

> **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.

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 web 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.

Automated controls on data entry and/or data values help ensure quality metadata. Many metadata schemes today have standard representations as XML schema (http://www.w3.org/XML/Schema). XML schema language can define characteristics such as repeatability and obligation, and can enforce these properties when metadata records are validated against the schema. XML Document Type Definitions (DTDs) (http://www.w3schools.com/dtd/default.asp) can also be used to provide standardization of metadata information, but they are less effective than XML schemas, because they do not support as many editorial controls over the data.

INITIATIVES

Digital programs provide the framework that pulls together people, policies and tools. Projects are activities within programs that have specific goals and are of finite duration. Project planning and program planning have common principles, and both must include plans for ongoing sustainability. For this reason, we refer to both projects and programs together as "digital initiatives."

Digital collection-building programs have become a core part of many organizations' missions, but this may not be reflected in the organizational structure and funding. A key component of the digital program manager's job is ensuring that the core nature of digital collection building is explicit at every level of the organization.

Building a digital collection often involves assembling a team of individuals from various disciplines, departments, and/or institutions. From the very beginning, the manager should invest in team building to encourage all members to benefit from each other's perspectives and backgrounds.

Initiatives Principle 1: A good digital initiative has a substantial design and planning component.

Initiatives Principle 2: A good digital initiative has an appropriate level of staffing with necessary expertise to achieve its objectives.

Initiatives Principle 3: A good digital initiative follows best practices for project management.

Initiatives Principle 4: A good digital initiative has an evaluation component.

Initiatives Principle 5: A good digital initiative markets itself and broadly disseminates information about the initiative's process and outcomes.

Initiatives Principle 6: A good digital initiative considers the entire lifecycle of the digital collection and associated services.

INITIATIVES PRINCIPLE 1

> **Initiatives Principle 1**: A good collection-building initiative has a substantial design and planning component.

Planning is crucial to the completion and success of any program or project. It encompasses all aspects of the initiative, from processing workflow to the ultimate look and feel of the collection interface. Early on, 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, or political factors), selection, digitization, copyright issues, metadata and access, maintenance, dissemination, and evaluation.

Resources on needs assessment:

- Institute of Museum and Library Services, *NLG [National Leadership Grants] Project Planning: A Tutorial* website http://www.imls.gov/project_planning/. Includes sections on organizational needs and on needs analysis of target audiences.
- Collaborative Digitization Program, *Market Segments and Their Information Needs* (1999) http://www.cdpheritage.org/project/rsrcusers.html.

General guides to planning for digital collections initiatives:

- JISC, *Funding Opportunities: Project Planning* website http://www.jisc.ac.uk/fundingopportunities/proj_manguide/projectplanning.aspx. A very thorough set of guides focused on preparing applications for JISC funding but generalizable to other projects; includes links to a Word project plan template.
- Technical Advisory Service for Images (TASI), *Generic Image Workflow: TASI Recommended Best Practice for Digitisation Projects* (2004) http://www.tasi.ac.uk/advice/managing/workflow_generic.html.
- Northeast Document Conservation Center, *Handbook for Digital Projects, III: Considerations for Project Management* (2000) http://nedcc.org/oldnedccsite/digital/iii.htm. Focuses on pre-project planning, despite the title.
- Technical Advisory Service for Images (TASI), *Risk Assessment* (2006) http://www.tasi.ac.uk/advice/managing/risk.html.
- Linda Serenson Colet, *RLG/DLF Guides to Quality in Visual Resource Imaging: 1. Planning an Imaging Project* (2000) http://web.archive.org/web/20060707235539/www.rlg.org/legacy/visguides/visguide1.html.

INITIATIVES PRINCIPLE 2

> **Initiatives Principle 2**: A good digital initiative has an appropriate level of staffing with necessary expertise to achieve its objectives.

There are many staff roles, each requiring different skills and abilities, that must work together to build a successful digital collection. At some point in time every digital initiative will require some expertise in management and project management, budget and finance, programming and systems administration, content selection, metadata creation and more. Some roles may be filled by the same person, while other roles may require multiple people. Every initiative has different requirements, and an organization may choose to emphasize different roles at different times, particularly as an organization's digital collection initiative matures. In some cases, building or enhancing an organization's capacity to create good digital collections can be an explicit goal of a digital initiative.

There are three strategies for an organization to accommodate the different roles and skills needed: in-house staffing, outsourcing, and collaboration with one or more partner organizations. Each of these strategies has advantages and drawbacks, and all three are generally used in combination by successful digital collection initiatives.

Some useful resources on staffing and managing digital initiatives:

- Grace Agnew, *Staffing Roles for Digital Collection Building* (2006) http://www.njdigitalhighway.org/documents/staffing-roles-for-digital-collection-building.pdf. Comprehensive summary of staffing roles and organizational strategies.
- Stephen Chapman, "Chapter III: Considerations for Project Management" in NEDCC, *Handbook for Digital Projects* (2003) http://nedcc.org/oldnedccsite/digital/iii.htm.
- North Carolina ECHO, *Project Management,* Revised edition (2007) http://www.ncecho.org/guide/management.asp.
- Technical Advisory Service for Images (TASI), *Advice – Managing Digitisation Projects* website http://www.tasi.ac.uk/advice/managing/managing.html.
- Technical Advisory Service for Images (TASI), *Digitisation: To Outsource or Not?* (2006) http://www.tasi.ac.uk/advice/managing/outsourcing.html.
- Technical Advisory Service for Images (TASI), *Staff Training* (2006) http://www.tasi.ac.uk/advice/managing/staff_training.html.
- Harvard College Library Preservation and Imaging Services, *Planning Digitization Projects: A Brief Bibliography* (2005) http://preserve.harvard.edu/bibliographies/digitalplanning.pdf.

INITIATIVES PRINCIPLE 3

> **Initiatives Principle 3**: A good digital initiative follows best practices for project management.

Digital initiatives, whether they are projects of finite duration or ongoing programs, share many of the same characteristics as projects in any other field, and so should follow industry standard project management practices.

There are many different methodologies for effective project management, to the extent that project management has become a discipline in its own right, but most project management methodologies share a small number of key common components:

1. Project Planning Stage

    - Clearly articulate the goals and deliverables of the project.
    - Conduct formative evaluation to validate the initial goals and deliverables of the project.
    - Identify what work needs to be done to accomplish the goals and deliverables of the project.
    - Break down the work into manageable sub-tasks, and identify dependencies between the sub-tasks.
    - Estimate and allocate the time and resources required to successfully complete each sub-task.
    - Create a project plan that includes an estimated timetable for the completion of the sub-tasks, estimates the resource requirements for the completion of each sub-task, and identifies key milestones and deliverables in the project.

2. Project Implementation Stage

    - Once the project has begun, monitor completion of tasks, sub-tasks, and milestones on the project plan.
    - Regularly review and update the project plan as new and more detailed information about scheduling and resource allocation becomes available.
    - Conduct additional formative evaluation to revalidate and, if necessary, modify the project's goals, deliverables, and the project plan.

3. Project Review Stage

    - After the final milestone in the project has been reached, review and document the project's progress, and identify any changes that were required to the project plan, goals, or deliverables.
    - Conduct summative evaluation to determine the success of the project.
    - Articulate the findings of the summative evaluation in a report that captures the lessons learned from the project.

A wide variety of software products are available to facilitate project management, including many open source, free, and/or low-cost tools. The commercial Microsoft Project (http://office.microsoft.com/en-us/project/default.aspx) and the open source dotProject (http://www.dotproject.net/) are among the popular general project management applications in cultural heritage organizations. Both of these projects can generate Gantt charts for schedule management (http://en.wikipedia.org/wiki/Gantt_chart).

Project management for digital initiatives is a popular topic for pre-conferences and seminars. Watch for announcements by professional associations and membership organizations. Some web-accessible resources include:

- *ProjectSmart* website http://www.projectsmart.co.uk/. A clearinghouse of information about project management, including many introductory materials.
- *PRINCE2 (PRojects IN Controlled Environments)* website http://www.prince2.com/. A de facto UK standard methodology for information technology project management:
- Stephen R. Toney, *Automating Your Museum*. Part 2: Managing the Project (2000) http://www.systemsplanning.com/mnc2.asp. While nominally about implementing a new content management system, the information applies equally well to all types of projects.

INITIATIVES PRINCIPLE 4

Initiatives Principle 4: A good initiative has an evaluation plan.

Whether the initiative is short-term or long-term, project managers should use an evaluation plan to identify and refine project goals, assess progress toward project goals, determine the quality of project results, measure the impact of the project, show accountability, and demonstrate the value of the project to funding agencies.

Evaluation can focus on the process and/or the outcome. Evaluation of process can involve assessment of a project's operations—such as staffing and management, workflow, and procedures—and focuses on input measures. While output measures such as the number of items digitized can be useful, recent emphasis is on outcome assessment, which is concerned with how people, collections, organizations, and systems have been affected by the project. The evaluation plan should emphasize the importance of an ongoing two-way dialogue with key stakeholder communities. Outcomes should be closely related to project goals and objectives and should be measurable.

Output measures for a digital collection building initiative may focus on the digital collection's size, quality, and usage. Other dimensions of the project, such as the functionality and usability of the collection's website, and users' experience with the collection and the service, are also output measures. The impact of a digital collection is the best indicator of a project's value, but it is most challenging to measure because it often involves many factors that are hard to quantify, and it demands considerable input from users. Surveys, interviews, and transaction logs are good for measuring inputs and outputs, while focus groups, interviews, and case studies are good for outcome and impact assessment. It is often necessary to combine various research methods to obtain quality data on a project's outcomes and impact.

Project managers should begin with clear evaluation objectives and have a plan for analyzing, reporting and implementing evaluation results. Results can be used to improve an ongoing project or to initiate follow-up efforts. A good evaluation plan will provide solid data to sustain a project over time.

Information on developing and implementing evaluation plans:

- Institute of Museum and Library Services, *Outcomes Based Evaluation* website http://www.imls.gov/applicants/obe.shtm. The IMLS encourages outcomes-based evaluation for their funded projects; this site has a webliography and points to supporting resources.
- Carter McNamara, *Basic Guide to Outcomes-Based Evaluation for Nonprofit Organizations with Very Limited Resources* (1997-2007) http://www.mapnp.org/library/evaluatn/outcomes.htm.
- *Building Better Websites: Evaluative Techniques for Library and Museum Websites* website http://www.lib.utexas.edu/dlp/imls/index.html. Developed by the University of Texas with an IMLS grant.

- *Usability.gov* website http://usability.gov/. Includes information on how to plan for usability testing, conduct usability tests, and analyze test results.
- Thomas C. Reeves, Xornam Apedoe, and Young Woo, *Evaluating Digital Libraries: A User-Friendly Guide* (2003) http://eduimpact.comm.nsdl.org/evalworkshop/UserGuideOct20.doc. A very useful project evaluation guide.

Case studies and examples:

- Joanne Evans, Andrew O'Dwyer, and Stephan Schneider, *Usability Evaluation in the Context of Digital Video Archives* (2002) http://www.sztaki.hu/conferences/deval/presentations/schneider.ppt.
- *Formative Evaluation of 5/99: The EDNER Project* (2002) http://www.cerlim.ac.uk/edner/dissem/brophy-nott-2002.ppt. Provides a framework for designing evaluation projects, with helpful illustrations.
- Michael Mabe, *DL Classification & Evaluation: A Publisher's View of the Work of the DELOS Evaluation Forum* (2002) http://www.sztaki.hu/conferences/deval/presentations/mabe.ppt. Digital collection managers may appreciate a publisher's perspective on the evaluation of digital libraries and resources.

INITIATIVES PRINCIPLE 5

> **Initiatives Principle 5**: A good digital initiative has a marketing strategy and broadly disseminates information about its progress and outcomes.

A good digital initiative, whether a short-term project or an ongoing program, will document and actively communicate its processes, progress, and outcomes to its stakeholder communities. This is called marketing when aimed at the community of potential users and dissemination when aimed at other information professionals.

A good digital initiative communicates its activities and broadcasts the availability of its deliverables as widely as possible. If the initiative produces any models, tools, or prototypes, they should be made available to the public to encourage adoption. If the initiative has local, regional, or national impact, that impact should be reported through publications, presentations, media, and other channels. "Trade" meetings of library, archive and/or museum professionals can be excellent venues for disseminating information about content, technologies and lessons learned. The Institute of Museum and Library Services' annual WebWise conferences, for example, are designed to showcase digital collections and projects funded by the IMLS.

Good collection description and good interoperability features like support for the Open Archives Initiative *Protocol for Metadata Harvesting* can help users find collections, but good marketing and communications are essential. Marketing should not be an add-on, but an essential part of building good digital collections, and funds for anticipated marketing expenses should be included in project and program budgets.

Modern marketing techniques aim to promote collections where the users are, on Facebook, YouTube, and other social networking sites as well as Google and Wikipedia. The New Jersey Digital Highway created a collection-level description entry in Wikipedia to promote their collections (http://en.wikipedia.org/wiki/New_jersey_digital_highway). The University of Washington Libraries have gone a step further by creating links to their digital collections in individual Wikipedia articles:

- Ann M. Lally and Carolyn E. Dunford, "Using Wikipedia to Extend Digital Collections," *D-Lib Magazine*, v. 13, no. 5/6 (2007) http://www.dlib.org/dlib/may07/lally/05lally.html.

Other resources on publicity and promotion:

- Washington State Library, *Digital Best Practices: Marketing* website http://digitalwa.statelib.wa.gov/newsite/projectmgmt/marketing.htm
- James Andrew Buczynski, *Referral Marketing Campaigns: 'Slashdotting' Digital Library Resources* (2007) http://hdl.handle.net/1853/13617. An audio, PowerPoint and auxiliary materials from a presentation advocating word-of-mouth marketing.

- University of North Texas Libraries, *Portal to Texas History* (2007) http://www.youtube.com/watch?v=rlDx9n4wFb0. A YouTube video promoting digital collections at UNT.
- Darren Kornblut, *Online Primetime: Promoting via Portals and Other Traffic Building Tricks*, presentation at Museums and the Web (2000) http://www.archimuse.com/mw2000/papers/kornblut/kornblut.html.

The primary goal of any project or program should be to accomplish its stated objectives within the time and budget allowed. However, the knowledge gained in the process should not be lost to other organizations. Most funding agencies require interim and final reports at the end of the project period, but internally funded programs should also issue reports at least annually.

Web-accessible reports should provide a detailed description and honest assessment of work accomplished, and should always include a section on "lessons learned."

Some examples of useful, comprehensive project reports:

- Library of Congress, *Manuscript Digitization Demonstration Project, Final Report* (1998) http://lcweb2.loc.gov/ammem/pictel/. Although the recommendations are dated, this remains a classic example of a good report.
- *Colorado's Historic Newspaper Collection: Final Report* (2005) http://www.cdpheritage.org/collection/chncfinalreport.html. Lacks only dates to be a model report.
- Peabody Museum, Harvard University, *Final Report Carnegie Institution of Washington Collection of Maya Archaeological Photographs: Phase 1 and 2* (2005) http://hul.harvard.edu/ldi/resources/Maya_Final_Report.pdf. Lacks lessons learned but includes useful illustrations.
- *Preserving and Digitizing Plant Images: Linking Plant Images and Databases for Public Access, Final Report from the Missouri Botanical Garden to the IMLS* (2002) http://www.mobot.org/mobot/imls/. A report designed for the Web.
- Library of Congress, *Ameritech National Digital Library Competition (1996-1999): Lessons Learned* website http://memory.loc.gov/ammem/award/lessons/lessons.html. A compilation of awardee reports on format issues, workflow and project management, staffing and skills, intellectual access, publicity, and other outcomes.

INITIATIVES PRINCIPLE 6

> **Initiatives Principle 6**: A good digital initiative considers the entire lifecycle of the digital collection and associated services developed.

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

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. Completed collections, and collections that grow steadily and incrementally over time, should be subsumed into the ongoing workflow of the organization. Essentially, digital projects are continued by digital programs, which are ideally part of the routinely funded business of the organization.

- JISC, *Exit and Sustainability Plans* website http://www.jisc.ac.uk/fundingopportunities/proj_manguide/projectplanning/exit.aspx. The JISC Project Guidelines provide practical frameworks for planning for "project exit" as well as sustainability.
- *LIFE (Life-cycle Information For E-literature)* website http://www.ucl.ac.uk/ls/life/. A collaboration between University College London and the British Library, the LIFE Project has developed a methodology to model the digital lifecycle and calculate the costs of preserving digital information for the next five, ten, or 100 years.
- Tom Clareson, "NEDCC Survey and Colloquium Explore Digitization and Digital Preservation Policies and Practices," *RLG DigiNews*, v. 10, no. 1 (2006) http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000070519&reqid=84280. Includes among its findings that "the lack of budget for acquisition and maintenance of digital materials was most clearly evident among the archive, public library, and ethnological/anthropological museum respondents."