Digitization Guidelines for Small Historical Institutions and Repositories

Updated October 25, 2013

Prepared on behalf of the Missouri Historical Records Advisory Board
(MHRAB) by the archives and records staff of the Secretary of State

Missouri Historical Records Advisory Board

Digitization Guidelines for Small Historical Institutions and Repositories

Purpose
During the 2008-09 Missouri Historical Records Advisory Board (MHRAB) strategic planning process, seven regional meetings were held across the state, drawing participants from 89 institutions. The purpose of these meetings was to obtain information on issues faced by local records repositories, as well as strategies to address those issues. Of the myriad topics raised, a constant was the desire for the MHRAB to provide guidance on various best practices, including digitization.

This document is intended as a guide for small cultural heritage institutions wishing to undertake digitization projects. Many sources were surveyed in the development of this text and are listed at the end of the document. These guidelines are not comprehensive but reflect the recommendations of the MHRAB, based on the aforementioned survey.

Digitization projects are costly and time-consuming, as anyone who has undertaken such an effort can attest. For this reason, significant time should be spent in planning. Institutions need to ask:

- Do we have the funding and staff to digitize materials?
- For whom are we digitizing the content; what are the needs of the user?
- Are the documents so fragile or accessed at such a high rate that it is worth the cost to digitize?
- Beyond actually having physical custody, do we own the rights to digitize and place these materials online?
- Do we possess the equipment to effectively digitize and present material?
- How will we manage digital content?

Selection
Determining what to digitize will be unique to each institution. Large libraries, archives or historical societies may attempt to digitize all or most of their collections, while small institutions may only wish to digitize select collections or portions of collections to publicize their holdings and invite the public to visit. Bigger institutions may have the budget and resources to take on large scale digitization projects while small institutions may have to be

more judicious in their selections. Although all cultural heritage institutions have similar missions, they are not identical and small institutions should not feel that digitization is an all or nothing proposition. Your digitization strategy needs to be in line with your institutional mission and capabilities.

How do you determine what to digitize? Let's look at what you have.

> 1. Prospective records for digitization should relate to the institution's collection policy (for more information on collection policies, see the [MHRAB's Best Practices for Local Archives and Repositories](#)).
> 2. Records should be unique or rare; if what you are considering for digitization is already available online, there is no reason to duplicate the existing effort.
> 3. Is there a noticeable demand to use the records; if no one cares about them in their original format, why would that change if the records are digitized?
> 4. Conversely, if there is a hidden jewel of a collection that few people know about, do you anticipate enough usage to justify the cost of digitization?
> 5. If there is demand to use a collection, could access be improved through digitization? Are the originals too fragile to stand up to regular handling by researchers? Would online or onsite digital images make the collection easier to use?

Generally, the best candidates for digitization on a budget are high-use collections. Digitization will improve access and improve your ability to preserve the original records, as they will be handled less.

Copyright
Potentially the most important question to ponder is whether you have the right to digitize your holdings. Just having physical possession of an item, object or collection does not give your organization the legal right to reformat and make that item available. Intellectual property rights must be granted in writing, unless they are already considered to be in the public domain.

Copyright protections have always been considered important in the United States, with Congress authorized to grant copyright protection under Article I, Section 8 of the U.S. Constitution. These protections are meant to encourage the production of new works, products and innovations. Copyright law is very complicated and the laws regarding copyright have changed numerous times with 14 statutes passed between 1790 and 2002. Over time, various types of works have been extended copyright protection beyond texts and manuscripts, including public performances; photographs and negatives; dramatic works; visual arts; foreign

works; motion pictures; sound recordings; architecture works; and digital files. Under the current statute, whether items are in the public domain relies on multiple factors.

To illustrate the complexities of copyright law, consider the following, drawn from *Copyright and Cultural Institutions* by Peter Hirtle, et al.:

Type	Term/Condition	Copyright or Public Domain 2013
Unpublished works	Life of author +70 years	Public domain, for authors who died before 1943
Unpublished anonymous works and works made for hire	120 years from date of creation	Public domain for works prior to 1893
Unpublished works created before 1978 that were published after 1977 but before 2003	Life of the author +70 years or December 31, 2047, whichever is greater	Still under copyright
Unpublished works created before 1978, but published after December 31, 2002	Life of the author +70 years	Public domain for authors who died before 1943
Unpublished works when author's date of death cannot be ascertained	120 years from date of creation	Public domain for works prior to 1893
Published before 1923		Public domain, copyright expired
Published 1923-1977	Published without copyright notice	Public domain, technicality
Published 1978 to March 1, 1989	Published without copyright notice and without registration	Public domain, technicality
Published 1978 to March 1, 1989	Published without copyright notice, but with registration	70 years after death of author or 95 years for publication with corporate author
1923-1963	Published with notice but not renewed	Public domain, technicality
1923-1963	Published with notice and renewed	95 years after publication date
1964-1977	Published with notice	95 years after publication date

1978 to March 1, 1989	Published with notice	70 years after death of author or 95 years for publication with corporate author
After March 1, 1989		70 years after death of author or 95 years for publication with corporate author
Government Publications		Generally in public domain, but works that were contracted out may be under copyright

This is just a short example of the varied copyrights associated with manuscripts and published works and does not pertain to photographs, maps, sound and video records, paintings or any of the other varied items protected by copyright.

Institutions do have the right to digitize copyrighted material for preservation, provided the images are only accessed at the institution. If you are planning to place items online, you will need to determine who owns the intellectual property rights (copyright, trademark, patent, publicity rights, performance rights etc.).

Copyright does not automatically transfer through the purchase or donation of a collection. To ensure you receive intellectual property rights, the bill of sale or deed of gift should contain language formally transferring copyright to the institution.

Institutions are strongly urged to research intellectual property rights of their holdings before undertaking a digitization project.

Metadata
Metadata documents the identification, management, nature, use and location of a resource. The Dewey Decimal and Library of Congress cataloging systems are examples of metadata—generally providing information about the "what" and "where" of a given library resource.

While a small collection of images may not need an index, if you plan to make the images searchable online or in-house, or if you undertake a large project, then file naming conventions and standardization are absolute musts for searching, retrieving, storage and management of the digital images.

Metadata generally falls into four categories: descriptive, structural, administrative and preservation.

- Descriptive metadata is just what it seems—the "who, what, where and when" of the item.
- Structural metadata describes the actual item—number of pages, collection information or other information about the physical nature of the item.
- Administrative metadata—when the item was digitized, rights and reproduction information, location etc.
- Preservation metadata—technical support information, such as media migration dates. Preservation metadata is sometimes considered administrative metadata.

The types of data you record will really be up to you. It may vary by collection and ultimately relies on your institutional needs. That said, it may not be necessary to create metadata from thin air; descriptive metadata may already exist in a database, spreadsheet or finding aid.

One issue that should be considered is your image naming standard. Rather than just relying on automatic numbering generated by your scanning software, you should consider creating unique identifiers based on the name of the collection being scanned. For instance, rather than naming the images 000001.tif, 000002.tif, etc., "XYZ Business Collection box 1, folder 2, image 1" might be rendered as "xyzbusinessB001F002_00001.tif," or any other unique name making it easier to locate the images in the future. "Leading zeros" before numbers are used to make sure file names are the same length and stay in file name order when sorted.

Metadata does not have to be complex. For instance, basic metadata may consist of just the following (taken from a photo scanning project):

Element	Explanation	Example
Collection	Name of collection	XYZ Business Collection
Volume/Box #	Location identifier	B001
Folder/Envelope #	Location identifier	F002
Image #	Photo identifier (for connecting metadata with photo for online searching)	_00001.tif
Print	Check box if scanned original was a photo print	X
Negative	Check box if scanned original was a photo negative	X
Folder Title	Name of folder	Old Thresher's Association

Image Summary/Title	What the photo depicts	Wheat Threshing
Year/Date	Of the original photo, if known	1920
County	Provides information about where	Lincoln
Subjects/Notes	Space to identify specific things in the photograph: people, streets, buildings, activities etc.	John Smith, Joseph Doe, operating an "X-type" threshing machine

Scanning Standards

There are many "standards" and guidelines available, generated by respected institutions and collaborative groups (National Archives and Records Administration, Library of Congress, Wisconsin Heritage Online, Western States Digital Standards Group etc.) and none of them wholly agree. In part, this is because of the varying goals each is trying to accomplish. If you are scanning an item to "preserve" it digitally, then the scanning requirements will be much more rigorous than if you are merely attempting to make information available to a wider audience (preservation vs. access). As this document's audience is not planning to dispose of original records and manuscripts, these recommendations are primarily aimed at scanning for access, rather than digital preservation.

A Note on PPI vs DPI

When scanning, you must choose a minimum image resolution. For these guidelines, we will use the term PPI (pixels per inch). PPI is often used interchangeably with DPI (dots per inch), but DPI really refers to the capacity of a printer to reproduce colors and tones from a given image file. PPI affects both the print size and the quality of the image. Too few pixels (low PPI) and the image becomes "pixelated" and unclear. This is when pictures look jagged or distressed and details cannot be made out. If you have a map or photo with a lot of minute detail, you should choose a higher PPI. A lower setting is acceptable for imaging text because pixilation is generally less noticeable. You may need a higher PPI, however, if you plan to make your text scans searchable through Optical Character Recognition (OCR).

Why not scan everything at the maximum PPI? Scanning at a high PPI creates larger files because they contain more data. If you are scanning for access, the attitude to take is "this is good enough." The same goes for scanning in color instead of bitonal (black and white) or grayscale; color scans are larger than grayscale scans, which are larger than bitonal.

An example of minimum scanning resolutions (below), drawn from the Library of Congress for "access" scanning, shows they can be fairly straightforward. It is your decision whether to scan in black and white, grayscale or color. Typically you will only choose color if there is significant color in the object (color photographs, slides or maps). For basic text, particularly typed material, black and white (bitonal) is appropriate. For most applications, grayscale should be the default choice.

Original	Use	Resolution	Bit-depth	Notes
Printed text: books, pamphlets, typed material etc.	Access image	300 ppi minimum	1-bit bitonal; 8-bit grayscale	
Script text: handwritten material	Access image	300 ppi minimum	8-bit grayscale	Use 24-bit color if color is an important attribute of the document
Maps: printed or hand drawn	Access content	300 ppi minimum	24-bit color	
Photographs	Access content	300 ppi minimum	8-bit grayscale	Use 24-bit color if color is an important attribute of the document
	Reproduction	Device maximum	24-bit color	
Slides	Access content	600 ppi	8-bit grayscale; 24-bit color	

File Types

Determining which file format to use for your project depends in large measure on your goals. If you are creating files for preservation, then you will want to select a file format utilizing lossless compression. Lossless formats compress the file using a compression algorithm, but all data is retained. These files tend to be large, thus requiring significantly more storage space, but there is no loss in quality. If your primary goal is providing access to a particular collection, such as for placement on the Internet, then a lossy file format is acceptable. Lossy formats use a different compression algorithm to reduce the file size, but the process discards part of the file information each time it is saved. For this reason, lossy formats are generally not considered acceptable for long-term preservation.

Media Type	Preservation	Access	Notes
Images/Photographs	TIFF/.tif	JPEG/.jpg	
Documents	TIFF/.tif; PDF/A	JEPG/.jpg; PDF/A	
Audio	Waveform Audio File/.wav; Audio Interchange File Format/.aiff	MP3	Uncompressed formats are too large for many users to access over the Internet; a lossy format should be used.
Video	No official standard; Library of Congress has adopted JPEG2000 with an MXF wrapper	MPEG-2; MPEG-4; AVI; MOV	Preservation quality video is expensive to store; a single hour can take up to 72 GB.

Managing Digital Files

It is best to maintain multiple copies of digital files on various storage media. The most secure storage solution for repositories involves online "cloud" storage. Files stored in this manner reside on an active, remote server and are regularly "refreshed."

You may also wish to make back-up copies on DVD/CDs, Flash drives, portable hard drives or any of a number of media types. These solutions are economical but not ideal for long-term storage. Keep in mind that DVD/CDs have a limited shelf life (less than 10 years); the disks should be checked regularly and the images migrated as necessary.

Do not store files only on a local computer. As many have experienced, one computer crash and all of your work is lost. Better to be safe and have plenty of back-up options than to lose all of your work.

Placing Digital Files Online

There are numerous options for institutions wanting to place digital content online.

If the bulk of your digitized items are photographs or maps, the simplest option is to create a free account through an online service such as Flickr. This photo sharing website allows each user one terabyte of cloud storage space and unlimited uploads. Photographs can be organized into sets and metadata can be added for each image. Similar websites include Fotki, PBase, SmugMug and many others. Some are free; some have associated fees.

For institutions with IT resources and large collections to be placed online, there are a number of options, both free and fee based.

- **ContentDM** is digital collection management software that allows you to upload multiple media types and organize items into collections. There are annual costs for maintenance and implementation.
- **Omeka** is a web publishing platform. It can be downloaded for free and claims to be designed for use by non-IT specialists. Omeka requires installation on servers, but also allows users to upload various file types and to create highly customizable online exhibits.
- **Omeka.net** is a slightly less customizable version of Omeka that allows users to purchase various storage levels and other customizable options. This is an easier option for institutions without their own servers. Plan levels run from free (500 MB storage) to $999/year (25 GB storage).
- **Collective Access** is free web-based software designed to catalog, manage and publish collections. It supports numerous file formats and metadata standards, but requires some programming ability to install and customize for your needs.

This is just a short list of the available options. You should understand that free products may require greater technical expertise to set up, manage and fix problems. Pay services typically offer technical support, but there are on-going fees—like a subscription.

For more information, please consult the resources listed on the following pages.

Supplemental Information

Association for Library Collections & Technical Services, Minimum Digitization Capture Recommendations (2013): www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations

Canadian Museum of Civilization Corporation, Digitization Standards (2006): www.pro.rcip-chin.gc.ca/sgc-cms/coursenligne-onlinecourses/content/29/CMCdigitization.pdf

Cornell University Library, Report of the Digital Preservation Policy Working Group on Establishing a Central Repository for Preserving Digital Image Collections (2001): www.library.cornell.edu/imls/image%20deposit%20guidelines.pdf

De Stefano, Paula, et al., Digitizing Video for Long-Term Preservation: An RFP Guide and Template (2013): http://library.nyu.edu/preservation/RFP_20131023.pdf

Hirtle, Peter B., et al., Copyright and Cultural Institutions: Guidelines for Digitization for U.S. Libraries, Archives and Museums (2009): http://ecommons.cornell.edu/bitstream/1813/14142/2/Hirtle-Copyright_final_RGB_lowres-cover1.pdf

Library of Congress, Technical Standards for Digital Conversion of Text and Graphic Materials (2007): http://memory.loc.gov/ammem/about/techStandards.pdf

Library of Congress, Sustainability of Digital Formats Planning for Collections (2013): www.digitalpreservation.gov/formats/

National Archives and Records Administration, Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files-Raster Images (2004): www.archives.gov/preservation/technical/guidelines.html

Northeast Document Conservation Center, Preservation Leaflet 6.6 Preservation and Selection for Digitization (2007): www.nedcc.org/resources/leaflets/6Reformatting/06PreservationAndSelection.php

University of Maryland Libraries, Best Practice Guidelines for Digital Collections (2007): http://ourdigitalworld.org/wp-content/uploads/2012/04/DigitizationBestPractices_Schreibman.pdf

University of Massachusetts-Amherst, Guidelines for Digitization (2011):
www.library.umass.edu/assets/aboutus/attachments/UMass-Amherst-Libraries-Best-Practice-Guidelines-for-Digitization-20110523-templated.pdf

University of North Carolina, Digital Library Services Digitization Guidelines [Quick Reference] :
www.unc.edu/projects/diglib/docs/dig_guidelinesQR011.pdf

Western States, Digital Imaging Best Practices (2003):
www.mndigital.org/digitizing/standards/imaging.pdf

Wisconsin Heritage Online, Digital Imaging Guidelines (2009):
http://recollectionwisconsin.org/wp-content/uploads/2013/01/WHO-Digital-Imaging-Guidelines-2009.pdf

Online platforms
Collective Access. (2013): www.collectiveaccess.org/about/overview

Omeka.net. (2013): http://info.omeka.net/about/

Omeka. (2013): http://omeka.org/

ContentDM. (2013): www.contentdm.org/

Flickr. (2013): www.flickr.com/