UMass Amherst **Libraries**

Guidelines for Digitization

Authored by the Digital Creation and Preservation Working Group
Meghan Banach (chair)
Brian Shelburne
Kelcy Shepherd
Aaron Rubenstein

May 2011

Contents

Introduction

The UMass Amherst Libraries Guidelines for Digitization were developed by members of the Digital Creation and Preservation Group for the use of all library digitization projects. These guidelines are designed to provide digital project managers with a set of *minimum* specifications for preservation-quality digitization of printed text, manuscripts, photographs, slides, rare books, sheet music, graphic arts, and maps. They provide a baseline for creating digital images that are of sufficient quality for long-term preservation. These specifications should not be seen as a replacement for project-specific analysis of factors such as purpose and use that might dictate higher level capture standards.

This document provides an introduction to the following aspects of your digitization project:

- Deciding whether your project will produce content for long-term preservation
- Identifying various uses of the materials and types of images needed
- Developing file naming conventions
- Choosing hardware and software for scanning
- Establishing quality control procedures
- Determining storage needs
- Understanding the project's metadata needs

The guidelines also include specifications for the file format, resolution, bit depth, dimensions, and compression for a variety of types of two dimensional works. These specifications are provided for preservation/service masters; full size high resolution; full size low resolution; medium size; and thumbnail images. Guidelines for audio and video will be developed as necessary.

These guidelines are designed to provide an overview of the issues that project managers will need to consider when starting a digitization project and to present the context necessary to understand the digitization specifications given. They are not intended to be an exhaustive treatment of all aspects of digitization or to provide a tutorial on digitization. For more detailed information about the topics introduced here, please see the Resources section.

Are you digitizing materials for short or long term

Short term	Long term
Small scale: A few scans, limited in scope	**Large scale:** Thousands of scans for a curated collection
Ephemeral: Content that changes, is updated frequently, or only used for a limited time	**Enduring:** Content that is stable and will have enduring value
Limited audience: Material for use solely within the Library or a Library department	**Wide audience:** Content is significant and has research interest beyond the University.
Short term examples: - Library web pages - LibGuides - Staff wikis - Content scanned for ILL or Reserves - Training materials - Blogs - Library publicity	**Long term examples:** - W. E. B. Du Bois Papers Digitization Project - Material digitized by the Open Content Alliance - UMass Faculty Pre-prints - Image collections from Image Collection Library - Electronic dissertations and theses

Are you scanning for the web or for print?

PRINT
Follow suggestions for Full-High Resolution

WEB
Follow suggestions for Full-Low Resolution

Review UMass Selection Criteria Policy

Determine what scanning specifications best fit your project (see tables on pp.13-14). To ensure long-term preservation, meet or

Overview of Scan Types

When planning a scanning project, it should be remembered that images are used in a variety of ways, even within one specific project. A well-constructed database of images, for example, may have as many as four different copies of each image that were created in order to provide an ideal level of display and access. Each copy is distinguished by its size, something dictated by the specific use of the image.

- *Preservation/Service image* – this is considered the master image. It should be at the highest size and quality; probably beyond what your anticipated need requires. The preservation image will be stored once finished and will be kept safe to serve as a backup to the images in active use. The service image is a working copy of the preservation image. The service image serves as the source image for all of your other copies of this image.

- *Full size image* – the full size image is the best quality image that you intend for your users to have. Depending on policies and rights you may choose to allow the user to have a high-resolution or low-resolution image. The difference generally is based on the ability to print the image at a large size. High resolution full-size images can produce a photo quality print at dimensions up to 8” x 10”. Low resolution full-size images will be suitable for projecting, screen viewing, and small prints. At the larger sizes they will not print well.

- *Medium size images* are also known as screen-size images. They are intended to serve as an image that may be viewed on a computer screen but will not project well or print well at a large size. The advantage to using medium size images is that they display well on a computer screen and take far less storage than full size images. They can provide a very modest degree of protection against misuse of the image if your intention is for the image to be viewed solely on screen.

- *Thumbnail images* are used as very small surrogates for a larger version of the same image. Thumbnails are often used as links to the larger size images. They are a convenient way to display many images on one page, allowing a user to see the contents of a database, a web page, or a folder at a glance. Thumbnails are not good as a print source, nor are they suitable for detailed onscreen viewing. They are large enough to be recognizable and to lead a viewer to a larger image.

When planning a scanning project, it is wise to anticipate your intended use for the images as much as possible and be certain that your digital image will be appropriate for your anticipated need. You should create your preservation/service images at the largest size that you can store and maintain. Technologies and intended usages may change with time. You may find additional uses for the images that require larger sizes, and it would be far preferable to re-use your original scan rather than scanning the same materials multiple times.

File naming conventions

Effective file names are essential for the stability and sustainability of digital storage and access systems and for ensuring interoperability between systems. A good file naming scheme can help you connect the various parts of a digital object, tie together metadata and images, and track individual files throughout the digitization process.

Some things to keep in mind when designing a file naming scheme:

- Each file name must be unique
- Think long term: how will this name scale as you add digital material to your collections?
- File names should provide context: names could include codes for department or collection.
- Keep file names simple for readability
- Self-explanatory file names make it easier to understand the context of files as they make their way through digitization work flows
- The more complicated the file name, the higher likelihood of human error when entering the name.
- Consider including the system's unique digital object ID in the name of the individual files that make up that object
- File names are not metadata: let your metadata describe the digital object. Use file names to connect metadata to digital images
- File names will outlast the current project staff

File names should at least:

- ☐ Be unique
- ☐ Use lowercase letters of the Latin alphabet and the numerals 0-9
- ☐ Have no spaces between characters
- ☐ Avoid punctuation marks other than hyphens and underscores
- ☐ Have no more than 31 characters (the fewer the better)
- ☐ Have a single period between the file name and the three letter extension

Scanning hardware and software

The hardware and software used to capture and manage digital images is critical to the success of digitization projects. It is essential to communicate the project's needs to the Library Systems Department to receive guidance on appropriate equipment and to ensure that the project's needs are met.

The following are some basic things to consider when assessing hardware and software tools for digitization:

Hardware

1. *Workstations*

The workstation used for the digitization process can have a large influence on the efficiency and accuracy of digitization work. Here are some questions to ask when selecting workstations:

- Does the workstation computer have enough memory to process large images? The more random access memory (RAM), the faster images can be processed.
- Does the workstation computer have high-speed data ports to connect to the scanner? USB 2.0 is the most common high-speed data connection.
- Does your monitor support at least 32-bit color quality?

2. *Scanners*

The most appropriate scanner for a project depends on the type of materials to be digitized. For many manuscript materials, an off-the-shelf flatbed scanner will suffice. When scanning stacks of modern 8.5 x 11" documents, a scanner with an Automatic Document Feeder (ADP) might be the most efficient. Large format or extremely delicate materials have special concerns that might require special equipment. For projects that include film or negatives, a Transparent Media Adapter (TMA) for the scanner will be necessary. A detailed plan of the project's goals and needs will be an important tool to assess which digitization hardware would be appropriate. Be sure to consider the following when selecting a scanner:

- What type of material(s) will be digitized?
- Can the scanner's output quality fulfill the minimum requirements laid out in these best practices?
- Does the scanner's output quality allow for higher quality images if it is necessary for the goals and needs of the project?

Software

1. *Scanning software*

Each scanner comes with its own software for interfacing with the scanner. Since this software is bundled with the scanner by the manufacturer, there is usually little choice once a scanner has been selected. Nonetheless, it is still worth ensuring that the scanning software has a few basic features:

- Does the scanning software support the TWAIN protocol? This will allow you to scan images directly into Photoshop and other image editing applications.

- Does the scanning software give you control over image resolution, bit-depth, and the option to turn off automatic adjustment features?

- How easy is it to use? Will it provide efficiency hurdles for the scanner operators?

2. *Image editing software*

Depending on the project needs, it may be necessary to edit images after they have been created. The industry standard for image editing is Adobe Photoshop though there is an open source alternative called GIMP. A good list of criteria for image editing software is available on the BCR's CDP Digital Imaging Best Practices on page 21.

Quality control

Quality control (QC) is an important part of any digitization project. QC encompasses procedures and techniques to verify the quality, accuracy, and consistency of digital images. The goal of any scanning project should be to "capture once, use many times." Digitization is expensive, time-consuming, and requires extensive handling of original materials. A digitization project should therefore focus on creating high-quality master images from which many derivative images can be created for specific uses.

Some things you may want to think about when developing a quality control program

- **Consider the goals of the project**

  The first step is to define the goals of the project as the quality control criteria will depend on them. For example, if the goal is to create a faithful reproduction of the original, the digital images should look as close as possible to the original material. However, if the goal of the project is to create the best quality digital images regardless of the condition of the originals, then it will not be part of your quality control criteria to make sure that the digital images accurately represent the originals.

- **Identify Your Products**

  Identify the products to be evaluated. These might include master and derivative images, printouts, accompanying metadata, and converted text or OCR'ed files.

- **Develop a Consistent Approach**
  To measure quality and judge whether the products are satisfactory, clearly define baseline characteristics for "acceptable" and "unacceptable" digital products.

- **Control the QC Environment**
  The impact of image-display conditions on perceived quality is often underestimated. Given an improper environment, even a high-quality image may come across as unsatisfactory. Factors that may affect on-screen image quality include viewing conditions, human characteristics, monitor calibration, and color management.

Assessing Image Quality

The key factors in image quality assessment are resolution, color and tone, and overall appearance.

- **Resolution**
  Resolution is the key factor in determining image quality for textual materials and other distinct, edge-based representations. Resolution attributes to inspect are legibility, completeness, darkness, contrast, sharpness, and uniformity.

- **Color and Tone**
  For color, grayscale, and some monochrome images, color and tone reproduction are significant indicators of quality, complementing the "detail" provided by resolution. The goal behind assessing color and tone appearance is to determine the extent to which a digital image conveys the same appearance as the color and tone ranges of the original document (or intermediate used). Tone and color assessment may be highly subjective and changeable according to the viewing environment and the characteristics of monitors and printers.

- **Overall Appearance**
  Image quality is cumulative, affected by a range of individual factors--capture system performance, resolution, dynamic range, and color accuracy. The final evaluation should be made on the overall image, appreciating all the individual factors that contribute to quality.

Quality Control Checklist

	Master digital image is a faithful representation of the original (if that is the goal)
	File name is correct
	File format is correct
	Bit depth is correct See: file, properties, details
	Image is correct size/resolution in long dimension
	Image is not rotated or backwards
	Image is not skewed or off centered
	Image has clean edges, clear contrast, and legible text
	No broken figures (illustrations, maps, etc.)
	No moiré patterns (wavy lines or swirls, usually found in areas where there are repeated patterns)

	No presence of digital artifacts (such as very regular, straight lines across picture)
	No pixellation (individual pixels are apparent to the naked eye)
	Not too light or too dark
	No loss of detail in highlight or shadows
	No errors in OCR
	If image has accompanying metadata check for accuracy and completeness

Storage & Access

Before undertaking a digitization project, project managers should consider how the files will be stored and how access to the files will be provided to the end user both now and into the future. A well thought out project plan should be developed in consultation with the Digital Strategies Group, the Digital Creation and Preservation Working Group, and the Metadata Working Group. If the project is related to research data, also consult with the Data Working Group.

Some points to consider about storage

- Think about how much space will you need to store your files
- Back up your files, preferably at a remote location or at least on a network drive.
- Store your preservation materials separate from files you (and others) access and work with on a daily basis so that files aren't accidentally deleted
- Depending on project needs and goals you may need either online or offline storage, or you may need some combination of both.
- Offline storage: storage of digital data outside the network in daily use (eg, on backup tapes) that is only accessible through the offline storage system, not the network.
    - Optical
        - CD-R (CD-Recordable)
        - DVD-R (DVD-Recordable), DVD+R, DVD+RW, DVD-RW, …
    - Tape
        - DLT, LT03, LT04, etc.
    - Pros: Relatively inexpensive media, producing a tangible item
    - Cons: Lower capacity, physical space requirements, unknown longevity, migration, potential format obsolescence, need to refresh the media every 5-10 years
- Online storage systems: storage of digital data as fully accessible information on the network in daily use.
    - Combine disk and automated tape storage with software to keep track of where files are located
    - Locally managed or remote provider
    - Pros: higher capacity, migration can be handled by software,
    - Cons: expensive, complex, network bandwidth issues, potential single point of failure, must trust service provider, if outsourcing must have clear Service Level Agreement (SLA) in place
- Digital library “objects” have many parts

    - Metadata
    - Preservation/archival files
    - Delivery files
- To keep them connected:
    - Good practice in file naming, directory organization, project documentation
- Make sure to have automated methods of checking files for errors
- Accessibility of the files will need to be maintained through migration, emulation, or some other preservation strategy.
    - Migration involves converting a file from its current format to another format before obsolescence, e.g., Excel 2.0 (from 1987) to Excel 2007.
    - Emulation involves recreating the hardware and software environment required to open an obsolete file, e.g., playing Pac-Man on your computer.
- Before beginning a new digitization project consult with the Library Systems Department and the Digital Strategies Group

Some points to consider about access

- How will end users access the materials you have digitized? Possibilities include:
    - Local web server
    - Commercially-hosted web site
    - Consortial service provider
    - Digital repository system (locally or commercially hosted):
        - Complex systems
        - Integrate data from databases, full-text search engines, file systems, and other sources
        - Allow for cross-collection searching through the use of OAI-PMH
        - Examples of commercial digital repository systems:
            - ContentDM, Luna Insight, Bepress Digital Commons
        - Examples of open source digital repository systems:
            - Eprints, DSpace, Fedora
- Additional considerations:
    - What access tools and interfaces will you need? These may include:
        - Searching
            - Metadata
            - Fulltext
        - Browsing
            - By subject, date, author, …
        - Navigation
            - Page turning, image panning/zooming, …
        - Streaming
            - For audio/video
        - Access controls for restricted materials

- Persistent URLs (a persistent URL or link is a web address that will consistently point to a specific information source)

Metadata

Metadata is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource." (NISO, Understanding Metadata) Good metadata is essential for the management and preservation of digital objects. Without metadata, information about a digital object can only be obtained from its filename, file extension, and perhaps its directory structure. This data will not provide enough information for the creator, user, or manager of the digital object to discover and identify the object, much less understand what the object is, how it was created, who can use it, and what computing environment is necessary to access it.

There are many different types of metadata that together support the discovery, evaluation, selection, access, navigation, management, and preservation of digital objects:

- **Descriptive metadata** provides information about the intellectual content and physical format of the object. It supports identification, discovery, and selection. Examples of descriptive metadata elements include title, author, abstract, extent, and subject terms.
- **Structural metadata** is data about the components (e.g., individual files) that make up a complex digital object, and the relationships among those components. For example, a four page pamphlet may be represented by four thumbnail images, four full-size access images, four archival master images, and an XML file of the encoded text. Structural metadata provides a means of identifying each of these files and their roles, as well as tracking their sequence and/or hierarchical relationships. It supports navigation and reconstruction of the object in an online environment.
- **Administrative metadata** supports the short- and long-term management of a digital object. It includes the following subtypes:
    - **Technical metadata** includes information about the format and creation of a digital object, such as data on hardware and software, compression ratios, and encryption keys. Technical metadata is necessary for current and future access to the object. As such, it is an important part of preservation metadata.
    - **Preservation metadata** is "the information necessary to carry out, document, and evaluate the processes that support the long-term retention and accessibility of digital materials." (PREMIS) It includes information about the object's creation, any changes to it, its chain of custody, and technical requirements for access.
    - **Use metadata** is information about the use of the digital object, for example usage statistics and search logs.
    - **Rights metadata** is data about intellectual property rights, such as copyright details, terms of use statements, and license agreement information.

Good, well-structured metadata that conforms to community-based standards and best practices is an essential part of any digital object, whether that metadata be stored separately or embedded in the digital object itself. Providing guidance on selecting appropriate standards and creating good metadata is beyond the scope of these guidelines. Project managers should consult on these issues with the Metadata Working Group at the start of any new digitization project. At a minimum, metadata should conform to the libraries' *Shareable Metadata Guidelines*.

Minimum Scanning guidelines

Manuscripts, Printed Text, Photographs, 35 mm. Slides, & Sheet Music

	Preservation/Service	Full Size -High resolution	Full Size -Low resolution	Medium	Thumbnail
File format	TIFF	JPEG/PNG	JPEG/PNG	JPEG/PNG	JPEG/PNG
Resolution	300-600 ppi	300-600 ppi	150 ppi	150 ppi	150 ppi
Bit depth	24 bit color or 8-bit grayscale	24 bit color or 8-bit grayscale	24 bit color or 8-bit grayscale	24 bit color or 8-bit grayscale	24 bit color or 8-bit grayscale
Dimensions	3000-6000 pixels across the long dimension	3000-6000 pixels across the long dimension	3000-6000 pixels across the long dimension	600 pixels across the long dimension	150-200 pixels across the long dimension
Compression	Preservation copy uncompressed/LZW lossless compression okay for service	LZW lossless compression	LZW lossless compression	LZW lossless compression	LZW lossless compression

Rare books

	Preservation/Service	Full Size -High resolution	Full Size -Low resolution	Medium	Thumbnail
File format	TIFF	TIFF	JPEG/PNG	JPEG/PNG	JPEG/PNG
Resolution	400-600 ppi	400-600 ppi	150 ppi	150 ppi	150 ppi
Bit depth	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale
Dimensions	3000-6000 pixels across the long dimension	3000-6000 pixels across the long dimension	3000-6000 pixels across the long dimension	600 pixels across the long dimension	150-200 pixels across the long dimension
Compression	Preservation copy uncompressed/LZW lossless compression for service	LZW lossless compression	LZW lossless compression	LZW lossless compression	LZW lossless compression

Graphic Art

	Preservation/ Service	Full Size -High resolution	Full Size -Low resolution	Medium	Thumbnail
File format	TIFF	JPEG/PNG	JPEG/PNG	JPEG/PNG	JPEG/PNG
Resolution	600-800 ppi	600-800 ppi	150 ppi	150 ppi	150 ppi
Bit depth	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale
Dimensions	6000-8000 pixels across long dimension excluding mounts & borders	6000-8000 pixels across long dimension excluding mounts & borders	6000-8000 pixels across long dimension excluding mounts & borders	300 pixels across long dimension excluding mounts & borders	150-200 pixels across long dimension
Compression	Preservation copy uncompressed/LZW lossless compression for service	LZW lossless compression	LZW lossless compression	LZW lossless compression	LZW lossless compression

Maps

	Preservation/Service	Full Size -High resolution	Full Size -Low resolution	Medium	Thumbnail
File format	TIFF	TIFF	JPEG/PNG	JPEG/PNG	JPEG/PNG
Resolution	**Maps less than 36 inches on the longest dimension:** 600 ppi **Maps greater than 36 inches on the longest dimension:** 300 ppi -400ppi?	**Maps less than 36 inches on the longest dimension:** 600 ppi **Maps greater than 36 inches on the longest dimension:** 300 ppi -400ppi?	150 ppi	150 ppi	150 ppi
Bit depth	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale	24 bit color or 8 bit grayscale
Dimensions	6000-8000 pixels across the long dimension	6000-8000 pixels across the long dimension	6000-8000 pixels across the long dimension	1078 pixels across the long dimension	150-200 pixels across the long dimension
Compression	Preservation copy uncompressed/LZW lossless compression okay for service	LZW lossless compression	LZW lossless compression	LZW lossless compression	LZW lossless compression

Resources

Arkansas State Archives Scanning Quality Control Checklist, www.arkives.com/photo/images/pdf/QC.pdf (accessed February 24, 2011)

Baca, Murtha, editor, *Introduction to Metadata*, http://www.getty.edu/research/conducting_research/standards/intrometadata/ (accessed October 13, 2010)

BCR's CDP Digital Imaging Best Practices Version 2.0, http://www.bcr.org/dps/cdp/best/digital-imaging-bp.pdf (accessed February 24, 2011)

Dunn, Jon, "Digital Collections: Storage and Access" www.dlib.indiana.edu/education/workshops/alioct03/storageaccess.ppt (accessed February 24, 2011)

Library of Congress Technical Standards for Digital Conversion of Text and Graphic Materials, http://memory.loc.gov/ammem/about/techStandards.pdf (accessed February 24, 2011)

Moving theory Into Practice: Digital Imaging Tutorial, Cornell University Library/ Research Department, http://www.library.cornell.edu/preservation/tutorial/contents.html (accessed February 24, 2011)

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/preservation/technical/guidelines.html (accessed February 24, 2011)

NISO, *Understanding Metadata,* http://www.niso.org/standards/resources/UnderstandingMetadata.pdf (accessed October 13, 2010)

NISO, *A Framework of Guidance for Building Good Digital Collections,* http://framework.niso.org/node/24 (accessed October 13, 2010)

Oya Y, Rieger, "Establishing a Quality Control Program," in *Moving Theory into Practice: Digital Imaging for Libraries and Archives*, Mountain View, CA: Research Libraries Group, 2000; pp. 61-83. http://library.oclc.org/cdm4/item_viewer.php?CISOROOT=/p267701coll33&CISOPTR=269 (accessed February 24, 2011)

UMass Amherst Libraries, *Shareable Metadata Guidelines*, http://www.library.umass.edu/wikis/dlgr/doku.php?id=metadata_guidelines (accessed October 13, 2010)