

GUIDELINES FOR THE CREATION OF DIGITAL COLLECTIONS

Digitization Best Practices for Images and Text

This document sets forth guidelines for digitizing images and text for the CARLI Digital Collections. Topics covered include: image quality, file formats, storage and access. It was created by the CARLI [Digital Collections Users' Group](#) (DCUG), standards subcommittee.

For questions about this document, please contact CARLI at support@carli.illinois.edu

Image Collections

Although no universal standards for quality image capture exist and technical standards are constantly evolving, the CARLI Digital Collections will adhere to the best practices adopted by recognized leading institutions.

Digital Images

A digital image is a two-dimensional array of small square regions known as pixels. In the case of a monochrome image, the brightness of each pixel is represented by a numeric value. Gray-scale images typically contain values in the range from 0 to 255, with 0 representing black, 255 representing white and values in between representing shades of gray. A color image can be represented by a two-dimensional array of Red, Green and Blue triples, where 0 indicates that none of that primary color is present in that pixel and 255 indicates a maximum amount of that primary color.

Creating Images

At least one copy of a digital master or archival image file should be created for each object photographed or scanned. From that master file, at least two derivative files will be created:

- An access image (an image used for detailed on-screen viewing)
- A thumbnail image (for fast access during search, browse and retrieval)

Therefore, when digitizing an image for inclusion in a digital library, at least three types of images should be generated:

Guidelines for the Creation of Digital Collections

Master Image	Access Image	Thumbnail Image
<ul style="list-style-type: none"> • Represents as closely as possible the information contained in the original • Uncompressed, or lossless compression • Unedited • Serves as long term source for derivative files and print reproductions • Can serve as surrogate for the original • High quality • Large file size • Stored in the TIFF file format 	<ul style="list-style-type: none"> • Used in place of master image for general web access • Generally fits within viewing area of average monitor • Reasonable file size for fast download time; does not require a fast network connection • Acceptable quality for general research • Compressed for speed of access • Usually stored in JPEG or JPEG2000 file format 	<ul style="list-style-type: none"> • A very small image usually presented with the bibliographic record • Designed to display quickly online; allows user to determine whether they want to view access image • Usually stored in GIF or JPEG file formats • Not always suitable for images consisting primarily of text, musical scores, etc.; user cannot tell what content is at so small a scale

from Western States Digital Standards Group, Digital Imaging Working Group, *Digital Imaging Best Practices*, http://www.bcr.org/cdp/best/wsdibp_v1.pdf, January 2003.

Master Images

Because digitization requires handling materials, often of a fragile or unique nature, a digital master should be generated for every object digitized, thereby eliminating the need to re-digitize, and therefore re-handle, the same physical material again in the future. The digital master image represents, as accurately as possible, the visual information in the original object. This image's primary function is to serve as a long-term archival record, as well as a source for derivative files and printed materials. Digital master files are measured in pixels. Scanning is measured in ppi (pixels per inch), also commonly referred to as dpi (dots per inch). Master files are most often saved to a designated server or other long-term storage device (such as DVDs or CD-Rs).

	File Format	Pixel Array and Resolution	Bit depth
Master Image	TIFF	4000-6000 pixels across the long dimension. Scan at a minimum of 300 dpi, more if necessary, to acquire an image	1-bit bitonal mode, 8-bit grayscale, or 24-bit color

Guidelines for the Creation of Digital Collections

		conforming to 4000-6000 pixels across the long dimension.	
--	--	---	--

Master images should be high-quality scans to avoid re-handling of any original materials. Scanned master images should not be edited for any specific output or use, and should be saved as large TIFF files with lossless or no compression.

Creating digital master files:

- Guidelines for file size and resolution of digital master files will vary by collection and be based on end-user needs, sizes and types of original objects, software specifications, available file storage space, etc.
- Each library should develop specific scanning guidelines based on individual collection needs and requirements.
- Where possible, scanning guidelines for the creation of digital master files should follow the specifications outlined in the CDL Guidelines for Digital Images: Guidelines for Digital Master Files:
<http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#guidelinesmaster>
- CARLI member libraries using CONTENTdm **should not** upload full resolution TIFF files to the CARLI server. Archival image file storage is the responsibility of each contributing institution and must be managed locally. Please note that the CONTENTdm Full Resolution Manager does not upload TIFFs to the server, so use of this CONTENTdm feature is acceptable.

Derivative Images

Derivative files are used for editing and enhancement, conversion to different formats, and presentation or transmission over networks. For each master image, at least two derivative files are created: an access image (for more detailed onscreen viewing) and a thumbnail image (for searching and browsing). Most digital asset management systems will automatically generate a thumbnail image for each item loaded into the software. In the case of collections using CONTENTdm, the software can also be configured to automatically generate access images from the master file.

Guidelines for the Creation of Digital Collections

General Guidelines for Creation of Derivative Files:

	File Format	Pixel Array and Resolution	Bit Depth
Access Image	JPEG or JPEG2000	1024-3000 pixels across the long dimension (72 – 300 ppi)	8 bit grayscale or 24 bit color
Thumbnail Image	GIF or JPEG	100-200 pixels across the long dimension (72 ppi)	4 - 8 bit grayscale or 8 – 24 bit color

from CDL Guidelines for Digital Images: Guidelines for Derivative Files, <http://www.cdlib.org/inside/diglib/guidelines/bpgimages/reqs.html#guidelinesderiv>, March 10, 2005

File Naming Conventions

Each digital object in a collection should be assigned a unique identifier. Unique identifiers should follow a consistent naming format to ensure ongoing identification and retrieval of digital files.

Guidelines for file names will vary by collection and will be based on local needs and specifications. Each library should develop specific file naming conventions based on individual collection needs and local requirements.

Monitor Calibration

Monitors used for image editing and color correction should be calibrated according to the following specifications:

- Set to 24 millions of colors
- Set monitor Gamma at 2.2
- Color temperature at 6500 degrees K

Monitor calibration software can be selected and purchased by member libraries and will vary depending on local budgets, equipment and software specifications.

Guidelines for the Creation of Digital Collections

Text Collections

Text materials include printed matter, photocopies, typed or laser printed documents, and may include some line drawings, graphic illustrations, manuscripts, music scores, blueprints and plans.

When scanning text documents, the scanning resolution should be based on the size of text included in the document and adjusted accordingly. Documents with smaller printed text may require higher resolutions and bit depths than documents that use large typefaces.

The following chart specifies basic guidelines for text document capture:

	File Format	Pixel Array and Resolution	Bit depth
Master Image	TIFF	4000-6000 pixels across the long dimension. Adjust the scan resolution to produce a Quality Index (QI) measurement of 8 for the smallest significant character. For more information about QI, see the NARA guidelines (PDF). The guidelines are also available in html here .	1-bit bitonal mode, 8-bit grayscale, or 24-bit color
Access Image	JPEG or JPEG2000	1024-3000 pixels across the long dimension (72 – 200 ppi)	1-bit bitonal or 8-bit grayscale

based on: CDL Guidelines for Digital Images, <http://www.cdlib.org/inside/diglib/guidelines/bpgimages/>, June 7, 2005. NARA Guidelines: <http://www.archives.gov/preservation/technical/guidelines.pdf> and <http://www.archives.gov/preservation/technical/guidelines.html>.

Machine Readable Text

Machine-readable text results either from a scanning and conversion process performed on textual materials or from manually transcribing text with a word processor, both of which produce some form of text file. A plain text file is common.

In digital library collections, these text files are often stored in such a way that they can be displayed on-screen, and they are often processed and indexed so that the content is searchable. Many options exist for digitizing and indexing text. Among them are:

- **Optical Character Recognition (OCR)**
OCR is the process of electronically scanning a bitmapped image for text and extracting that text for further manipulation, often by placing the text in a separate machine-readable file.

Guidelines for the Creation of Digital Collections

- **Transcriptions**

Text that is difficult to read, especially handwritten manuscripts, should be considered for transcription. Transcribed text, particularly if it is encoded with markup languages, such as XML or XHTML, helps the researcher navigate and search long documents. Transcription presents its own problems – it can be labor intensive and cost prohibitive.

- **Character & Document Encoding**

Character encoding is the assignment of a computer code to each of the letters in the document. Many are familiar with, or at least heard of, ASCII (American Standard Code for Information Interchange). ASCII was a character set based on the English alphabet. It was later expanded to include accented characters (e.g., á è ü) and other (Latin language-based) characters, such as the German scharfe S (ß). Since then, and currently, UTF-8, or the 8-bit Unicode Transformation Format, is the generally accepted standard (it is backwards compatible with ASCII). UTF-8 encoding, importantly, can accommodate not only Latin-based language characters, but also Greek, Cyrillic, Hebrew, Arabic, and much more. For these reasons, it is recommended that all textual documents be encoded as UTF-8.

How a document is encoded (document encoding) refers to the character set encoding used. All programs, or very nearly all of them, can interpret a UTF-8 encoded document. Most computer programs can save text-based documents (plain text files, XML, or HTML) as a UTF-8 encoded document, if they do not already. Additionally, some document formats, such as XML and HTML, provide a way to declare the file as UTF-8 encoded, which a parser can then use to interpret the rest of the document. In XML, this can be seen easily in the first line of the file, where the type of file is declared (XML) and so is its encoding (UTF-8).

```
<?xml version="1.0" encoding="UTF-8"?>
```

Note: declaring a file's encoding UTF-8 does not guarantee the file will actually be UTF-8 encoded. Before saving the text file, check the software's save options.

Text based materials in the CARLI Digital Collections may be handled in various ways. Methods will depend on factors such as library resources, quality of the original materials, software requirements, and end user needs.

Appendix

Sample Workflow

A sample workflow will demonstrate how to apply the image digitization requirements and best practices outlined in this document. For example purposes, the item to be digitized is a drawing done on an 8 ½ x 11 inch standard sheet of paper. It is possible to divide the workflow into two sections:

Creating the Digital Master Image

- 1) On the computer, open the scanning application and place the sheet to be scanned on the scanner's scan bed.
- 2) Verify that the scan settings are set to scan the image in color (24-bit) at 400 dpi (ppi). This will ensure that digital master file is roughly 4400 pixels on the long side (400 dpi x 11 in = 4400 pixels). 4000 pixels on the long side is the minimum for a good archival master file.
- 3) If possible, preview the item to be scanned and use the scanning software to crop the previewed image to the proper size. Leave a little margin beyond the items borders to ensure that nothing is accidentally missed when scanning.
- 4) Scan the image.
- 5) Once the image is scanned, immediately save the image as an uncompressed TIFF image, giving it a unique name.
- 6) Optional: Repeating steps 1 to 5, continue to scan images and save them as archival masters.

Creating the two Derivative Images – Access Image and Thumbnail Image

- 7) Open one of the scanned digital master TIFF images in an image-editing program, such as Photoshop or The Gimp (The Gimp is a open source image editing program).
- 8) Crop, straighten, and color correct the image, as is necessary.
- 9) Create the access image: Resize the image, maintaining the image's proportions, to about 2000 pixels on the long side (any size between 1024 pixels and 3000 pixels on the long side is acceptable).
- 10) Optional: "Sharpen" the image, if appropriate (in Photoshop one could also use the Unsharp Mask option for this).
- 11) Use "Save as" to save the corrected/manipulated image as a JPEG file at high quality. **Be careful not to overwrite the digital master TIFF image.** Save the access image in a location that separates it from the digital masters and/or append an additional letter or number, for example, to the filename to identify it as the access image.
- 12) Create the thumbnail image: Maintaining the image's proportions, resize the image, for a second time, so that the image is about 150 pixels on the long side (any size between 100 pixels and 200 pixels on the long side is fine).
- 13) Optional: "Sharpen" the image, if appropriate (bear in mind, it has been "sharpened" once already).
- 14) Use "Save as" to save the corrected/manipulated image as a JPEG file at high quality. **Be careful not to overwrite the digital master TIFF image or the access image.** Save the thumbnail image in a location that separates it from the digital masters and/or append

Guidelines for the Creation of Digital Collections

an additional letter or number, for example, to the filename to identify it as the access image.

NOTE: Your digital asset management system (e.g., CONTENTdm) may have a feature that automatically generates thumbnail images for you. If so, it may be unnecessary to create them.