

California Digital Library
Digital Image Format Standards

July 9, 2001

Reviewed and updated annually

Table of Contents

Introduction.....	3
Definition of a Digital Image.....	3
Digital Masters	3
Digital Image Storage Formats.....	4
Specific Minimum Resolution and File Formats	5
Quality Control	7
Standards Development Process	9
Distribution	10

Introduction

This document addresses the standards for *archival quality*, digital image collections for the California Digital Library. In conjunction with the companion document, *California Digital Library Digital Object Collection Standards* (July 9, 2001), these standards describe image quality considerations, file formats, storage and access standards for digital images created by or incorporated into the CDL as part of its permanent collections¹. They attempt to balance adherence to industry standards, reproduction quality, access, potential longevity and cost. Adherence to these standards is required for all CDL contributors and may also serve University of California staff as guidelines for digital image creation and presentation.

At a minimum, digital image collections incorporated into the CDL must have the following components.

- A set of archival quality digital image files as defined in this document.
- The required metadata defined in this document and encoded in an XML DTD.

These standards are not intended to address all of the administrative and technical issues surrounding the creation of digital image collections and they do not describe operational procedures for digitization. "Best Practices for Image Capture," a companion to these standards, describes best practices for libraries, archives, and museums; it does not define standards, but instead provides an overview of the issues that need to be addressed when initiating a digital imaging project.

Definition of a Digital Image

A digital image is defined for the purposes of this document as a raster based, 2-dimensional, rectangular array of static data elements called pixels, intended for display on a computer monitor or for transformation into another format, such as a printed page.

Digital Masters

Digital master files are created as the direct result of image capture. The digital master should represent as accurately as possible the visual information in the original object.² Digital images should be created through the direct scanning or imaging of the original object. However, if the original object can not be digitized directly due to its size or other attributes, it may be necessary to use a photographic intermediary. Care should be taken

¹ Together these two documents update and replace *CDL Digital Image Collection Standards* (September 1999), which may now be considered obsolete.

² Note: this is similarity in terms of technical fidelity using objective technological measurements; this is not accuracy as determined by comparing the object scanned to an image on a display device. This type of accuracy is obtained by the manipulation of settings before scanning rather than by image processing after scanning.

that the photographic intermediary is well documented and represents the original object as accurately as possible.

The primary function of digital master files are to serve as a long-term archival record and as a source for derivative files. A digital master file may serve as a surrogate for the original, may completely replace originals or may be used as security against possible loss of originals due to disaster, theft and/or deterioration. Derivative files are created from digital master images for editing or enhancement, conversion of the master to different formats, and presentation and transmission over networks.

Long term preservation of digital master files requires a strategy of identification, storage, and migration to new media, as well as policies about image use and access to them. It is essential that master files remain unaltered over time. Lossy compression techniques, such as GIF and JPEG, should not be applied to master files and migration procedures should include quality control procedures to ensure that the integrity of the files is maintained throughout the entire process.

The specifications for derivative files used for image presentation may change over time; digital masters can serve an archival purpose, and can be processed by different presentation methods to create necessary derivative files without the expense of digitizing the original object again. Because the process of image capture is so labor intensive, the goal should be to create a master that has a useful life of at least fifty years. Therefore, collection managers should anticipate a wide variety of future uses, and capture at a quality high enough to satisfy these uses. In general, decisions about image capture should err towards the highest quality.

Digital Image Storage Formats

Digital image collections intended for long term storage and presentation should store from three to four images for each original item: an archival image, derivatives for viewing and a thumbnail for browsing. The master or archival image should capture as much information as possible to preserve the investment in the capture process. Masters should use color rather than grayscale when color is an integral part of the information conveyed by the original object, and any compression applied to the file should be lossless. Viewing files can be created at any time from the archival image and should be created to provide reasonable access by standard viewers. It is recommended that at least two viewing files be created, a preview or thumbnail file for the fastest access during initial search and retrieval process and a service or reference image for more detailed viewing.

The following image formats are supported by the CDL, some for archival storage and some for presentation purposes. See the following table for the appropriate use of each format.

- **TIFF ITU-T.6** - A 24-bit storage format commonly used by Adobe PhotoShop and other bitmap editors, this TIFF format may be used to store color images. This format is also suited for bitonal text documents; it provides a high level of detail (up to 600 dpi, or 4,800 x 6,600 pixels for a letter-sized page). TIFF ITU-T.6 format should be used for archival files. With lossless compression, the data that results from compressing and then uncompressing the image is exactly the same as the original, uncompressed file. CCITT Group 4 compression is lossless and, therefore, permissible for masters and derivatives.

While compression is allowed for archival files, it is discouraged, as it adds complexity to the format migration issues of long-term preservation. When compression is used, it must be lossless and not proprietary.

- **JPEG** - A 24-bit, lossy (some information lost) compression format which is well-suited for screen and print presentation. JPEG is supported by all major computer platforms and by Internet web browsers. With lossy compression, the picture quality of the compressed file is reduced when compared to the original file, and can not be restored, except by going back to the original. The advantage is that the file sizes are much smaller, and image quality is acceptable in most cases. It is not acceptable as an archival file format.
- **JFIF JPEG File Interchange Format** - A specific implementation of the JPEG standard, commonly used by bitmap editing programs, viewers, and Web browsers.
- **GIF** - An 8-bit, lossless compression format which is well-suited for low resolution screen display of images. GIF is often used for image thumbnails and screen versions of text documents, and is supported by all major computer platforms and Internet web browsers.
- **PNG** - The Portable Network Graphic format is expected to provide a higher-quality replacement for the GIF, particularly for images delivered to World Wide Web browsers. Information on PNG can be found at <http://www.cdrom.com/pub/png/>
- **PDF** - Adobe Acrobat Portable Document Format provides a convenient way to view and print images at high resolution, and may also be used to group several files into chapters and books. PDF can provide additional navigational tools such as hyperlinks among pages within a document, and from one PDF document to another. Although this is a proprietary Adobe format, both the file specifications and the viewer software are freely distributed. Plug-ins are available for major web browser to enable them to view PDF files without launching an external viewer.

Specific Minimum Resolution and File Formats

The intent of the following table is to offer guidelines for scanning various types and sizes of original documents, so that the digital master files as captured will record all of the significant visual features in the original item. Capture resolutions in the table are based upon the assumption that a scanning resolution of 600 ppi will be sufficient to meet this requirement for most originals in most collections, apart from negatives and transparencies. Digital master files which fail to capture some of the visual information present in the original will presumably become obsolete as image capture techniques improve over time.

The reflective formats, such as photographic prints and illustrations, are based on 8.5" x 11" originals scanned at 600ppi. The 35mm format has a resolution standard of 4200 pixels in the longest dimension, as this is about as

much data that most 35mm films can capture. Scanning the 35mm format, which is 1.5" on the longest side, at 2800ppi will result in compliance with the 4200 pixel standard. Note, if you plan to create a film intermediary of the object and then digitize the 35mm intermediate, remember to consider the size of the original. Filming an original larger than 5" x 7" with the 35mm process will not capture all the original's detail. For example, 4200 pixels spread along the 7" inch side yields 600ppi (4200 pixels / 7"). If the original was 12" long, the image would be only 350ppi (4200 pixels / 12"), which is not archival quality.

Other transmissive formats, such as negatives and slides have a standard of 6000 pixels on the longest side, based on a 8.5" x 11" original, which yields an image just under 600ppi image (6000 pixels / 11"). Note again, if you plan to create a film intermediary and then digitize, you must consider the size of the original. For example, creating a 4" x 5" negative at 1200dpi from a 10" long original original yields the 6000ppi standard, or a 600dpi image (6000 pixels / 10"). Creating an image 6000 pixels on the longest side for a 12" long original would be digitizing at 500ppi and therefore, would lose detail from the original image.

Oversize originals such as posters and maps can be especially difficult to scan at the recommended resolution of 600 ppi. Few libraries own flatbed scanners capable of scanning originals larger than 11" x 17," and even if they do, the problems of handling image files larger than about 120MB are daunting. These problems may lead to the use of a lower standard of capture resolution, such as 300 ppi or 3000 pixels in the longest dimension (the "alternative minimum"), with the understanding that the useful life of the files may be limited and digital image capture for these objects will need to be repeated in the future.

Specific Resolutions and File Formats

	Reflective Originals (Prints, Manuscript, Text, Paintings, Maps & Drawings)	Transmissive Originals (Negatives & Transparencies)
Master Files	<ul style="list-style-type: none"> • 600 pixels per inch (PPI) capture resolution; less for oversize originals (see Alternative Minimum below) • TIFF, lossless compression • 8-bit greyscale; 24-bit color; (bitonal for typeset pages with typeface 7 pt and above) 	<ul style="list-style-type: none"> • Adjust capture to achieve 600 PPI when the Master is restored to the size of the original object being digitized³ • TIFF, lossless compression • 8-bit greyscale; 24-bit color
Alternative Minimum for Originals Larger than 8.5"x11"	<ul style="list-style-type: none"> • 300 PPI (otherwise, same as main guidelines above) 	<ul style="list-style-type: none"> • Adjust capture to achieve 300 PPI when the Master is restored the size of the original object being digitized.
Access Files	<ul style="list-style-type: none"> • 800, 1500 or 3000 pixels across the long dimension • JPEG, medium quality compression • 8-bit greyscale; 24-bit color 	Same as Reflective
Alternative Access File Format for Better Text Legibility	<ul style="list-style-type: none"> • 100 PPI (i.e., resample image master file to 100 pixels per inch of original document) • GIF; 4-bit greyscale; 8-bit color; or • JPEG, higher quality compression 	Same as Reflective
Thumbnail Files	<ul style="list-style-type: none"> • 200-400 pixels across the long dimension • GIF • 4-bit grayscale, 8-bit color 	Same as Reflective
Print	<ul style="list-style-type: none"> • 300-600 DPI • PDF or TIFF w/LZW compression for B&W materials, otherwise, JFIF - medium to high quality • 8-bit greyscale; 24-bit color 	Same as Reflective

Note: These guidelines are an attempt to guide one through general digitization projects. In all cases, the over-riding factor should be the need to capture all relevant elements that users will need.

³ E.g., when a 35mm negative is shot of a 4"x5" photograph (i.e., the original object), the resulting negative should be scanned at 2000 PPI. Since the 35mm negative is 1-1/2" long, this will provide 3000 pixels along the 1.5" (longest) side of the negative. Therefore, when an image of the slide is created at its original size, it will be 600 PPI. That is, we will have 3000 pixels spread over the 5" of the new image (3000/5" = 600 PPI). The formula to determine the correct scanning PPI is:

$$\frac{\text{Longest Side of the Original} * 600 \text{ PPI}}{\text{Longest side of the negative}}$$

Quality Control

It is important to establish a digitization workflow capable of producing consistent, reproducible results in a cost effective production environment. This is aided by the use of standard bars and rulers and quality control procedures. Elements of these procedures should be noted in the metadata associated with each digital image. For more information on quality control procedures, please see the references in "Best Practices for Image Capture" under "Scanning and Image Capture," with special attention to, "Steven Puglia and Barry Roginski. NARA Guidelines for Digitizing Archival Materials for Electronic Access, College Park: National Archives and Records Administration, January 1998. <http://www.nara.gov/nara/vision/eap/digguide.pdf>."

Standards Development Process

This is the second version of the CDL Digital Image Standard. This version is based upon the first (September 1, 1999) version of the CDL's Digital Image Standard, which included recommendations of the Museum Educational Site Licensing Project (MESL), the Library of Congress and the MOA II participants.

The Museum Educational Site Licensing Project (MESL) offered a framework for seven collecting institutions, primarily museums, and seven universities to experiment with new ways to distribute visual information--both images and related textual materials.

The Fowler Museum of Cultural History
The George Eastman House
Harvard University Art Museums
The Library of Congress
The Museum of Fine Arts, Houston
The National Gallery of Art
The National Museum of American Art
American University

Columbia University
Cornell University
University of Illinois at Urbana-Champaign
University of Maryland
University of Michigan
The University of Virginia
The Getty Information Institute
MUSE Educational Media

The Making of America (MoA II) Testbed Project is a Digital Library Federation (DLF) coordinated, multi-phase endeavor to investigate important issues in the creation of an integrated, but distributed, digital library of archival materials (i.e., digitized surrogates of primary source materials found in archives and special collections). The participants include Cornell University, New York Public Library, Pennsylvania State University, Stanford University and UC Berkeley.

The Library of Congress white papers and standards are based on the experience gained during the American Memory Pilot Project. The concepts discussed and the principles developed still guide the Library's digital conversion efforts, although they are under revision to accommodate the capabilities of new technologies and new digital formats.

The CDL Technical Architecture and Standards Workgroup includes the following members with extensive experience with digital object collection and management:

- Howard Besser, MESL and MOA II digital imaging testbed projects
- Diane Bisom, University of California, Irvine
- Bernie Hurley, MOA II, University of California, Berkeley
- Greg Janee, Alexandria Digital Library
- John Kunze, University of California, San Francisco
- Reagan Moore and Chaitanya Baru, San Diego Supercomputer Center, ongoing research with the National Archives and Records Administration on the long term storage and retrieval of digital content.
- Terry Ryan, University of California, Los Angeles
- David Walker, California Digital Library

Distribution

Draft standards are provided to each UC campus for distribution to staff for review and comment. The final copy of these standards are submitted to the CDL University Librarian annually by the CDL Technical Architecture and Standards Workgroup after review and updates based upon changes in technical standards and current practice.