

# **Technical Standards for the Creation of Digital Collections**

*The School of the Art Institute of Chicago*

*John M. Flaxman Library*

This document sets forth the technical standards for digitizing SAIC digital library collections. The issues described concern image quality, file formats, storage and access.

## **Creating Digital Images**

Although no universal standards for quality image capture exist and technical standards are constantly evolving, the SAIC Digital Libraries 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.

### **Number of 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)

A total of three types of images should be generated when an object is digitized:

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

from Western States Digital Standards Group, Digital Imaging Working Group, *Digital Imaging Best Practices*, [http://www.cdpheritage.org/digital/scanning/documents/WSDIBP\\_v1.pdf](http://www.cdpheritage.org/digital/scanning/documents/WSDIBP_v1.pdf), January 2003.

## Master Images

Due to the stress of digitizing unique materials, a digital master must be generated for every object created. 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 dpi (dots per inch) or ppi (pixels per inch). The files are saved to a designated server or other long term storage device.

Master images should be scanned at an appropriate level of quality 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 based on end user needs, sizes and types of original objects, software specifications, available file storage space, etc.
- Each SAIC digital collection will develop specific scanning guidelines based on individual collection needs and requirements.
- Where possible, scanning guidelines for 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>

### Derivative Images

Derivative files are used for editing and enhancement, conversion to different formats, and presentation or transmission over networks. For each master image, two derivative files are created: an access image (for more detailed onscreen viewing) and a thumbnail image (for searching and browsing).

General Guidelines for Creation of Derivative Files:

	<b>File Format</b>	<b>Pixel Array</b>	<b>Resolution and Bit Depth</b>
<b>Access Image</b>	JPEG or JPEG2000	800-3000 pixels across the long dimension	24-48 bit color; 72-300 dpi
<b>Thumbnail Image</b>	GIF	100-200 pixels across the long dimension	4-bit grayscale, 8-bit color; 72 dpi

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 must be assigned a unique identifier.

Unique Identifiers in the SAIC digital collections will follow a consistent naming format to ensure ongoing identification and retrieval of digital files.

- Each collection in the SAIC digital collections will establish a unique collection identifier. Identifiers should be relatively short (no more than 6-10 characters long) and should not contain uppercase letters or symbols.
- Unique identifiers for digital objects in SAIC collections will be created using the following formula:

**collectionID\_original object ID number\_view**

Example:

A book object from the Joan Flasch Artists' Book Collection, Accession Number "21.6", has three image files associated with it – one view of the cover, and two other views of interior pages. A set of digital files (masters and derivatives) is created for each image.

The corresponding image files for this object would be represented in the following way:

**Joan Flasch Collection ID** = jfabc\_

**Joan Flasch Object ID Number** = original Accession Number with an underscore (\_) replacing any periods (.)

**File Names:**

Original	Digital Master	Access Image	Thumbnail
21.6 – cover	jfabc_21_6_cover.tif	jfabc_21_6_cover.jpg	jfabc_21_6_cover.gif
21.6 – view 1	jfabc_21_6_view1.tif	jfabc_21_6_view1.jpg	jfabc_21_6_view1.gif
21.6 – view 2	jfabc_21_6_view2.tif	jfabc_21_6_view2.jpg	jfabc_21_6_view2.gif

Any points or periods (.) in original ID numbers should be replaced with an underscore (\_), and identifiers should contain only lowercase letters.

Collection ID and Object ID numbers should be defined within the guidelines of each specific SAIC digital collection.

## 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 should be selected with the help of CRIT and will vary depending on department budgets, equipment and software specifications.

## Text Collections

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

When scanning text documents, spatial resolutions should be based on the size of text included in the document and resolutions should be 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 image capture:

	<b>File Format</b>	<b>Pixel Array</b>	<b>Resolution and bit depth</b>
<b>Master Image</b>	TIFF	4000-6000 pixels across the long dimension	1-bit bitonal mode or 8-bit grayscale: 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 <a href="#">NARA guidelines</a> .
<b>Access Image</b>	JPEG or JPEG2000	800-3000 pixels across the long dimension	1-bit bitonal or 8-bit grayscale: 72-200 dpi

based on: CDL Guidelines for Digital Images, <http://www.cdlib.org/inside/diglib/guidelines/bpgimages/>, June 7, 2005. NARA Guidelines: <http://www.archives.gov/research/arc/digitizing-archival-materials.pdf>

## **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.

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

- **Optical Character Recognition**

OCR is a system that reads text and translates the image into a form the computer can manipulate. The process transforms a bitmapped image of printed text into text code, thereby making it machine readable.

- **Transcriptions**

Text that is difficult to read, especially handwritten manuscripts, is an example of material that should be considered for transcription. Transcribed text, especially when encoded with markup languages, helps the researcher navigate and search long documents.

Transcription presents its own problems – it can be labor intensive and cost prohibitive.

- **Character Encoding**

Character encoding is the assignment of a computer code to each of the letters in the document. A text encoded with a markup language provides searchability. Recognized text in access copies may be delivered in a variety of text formats, including HTML, ASCII, XML in EAD, TEI or other accepted standard depending on the needs of the project. Participants in the American Memory Project at the Library of Congress, use SGML in a DTD (Document Type Definition) based on the TEI (Text Encoding Initiative) Guidelines. Since SGML viewers are not yet freely available for viewing SGML over the Internet, an HTML version can be derived from the SGML version for widespread viewing online.

SAIC digital text collections may be handled in various ways and methods will depend on factors such as departmental resources, quality of the original materials, software requirements, and end user needs.

## MINIMUM GUIDELINES FOR DIGITAL IMAGE CREATION

ORIGINAL MATERIAL	DIGITAL MASTER			SCREEN DISPLAY			THUMBNAIL DISPLAY		
	Pixel array	Resolution & bit depth	file format	Pixel array	Resolution & bit depth	file format	Pixel array	Resolution & bit depth	file format
text document	4000-6000 across the long dimension	8 bit grayscale 24-48 bit color: 400-600 ppi	TIFF	800-3000 across the long dimension	8 bit grayscale; 24-48 bit color: 73-300 ppi	JPG or JP2	100-200 across the long dimension	4 bit grayscale 8 bit color: 72 ppi	GIF
Illustrations, Maps, Manuscripts, Mixed Formats, etc.	4000-6000 across the long dimension	8 bit grayscale; 24-48 bit color: 400-600 ppi	TIFF	800-3000 across the long dimension	8 bit grayscale; 24-48 bit color: 73-300 ppi	JPG or JP2	100-200 across the long dimension	4 bit grayscale 8 bit color: 72 ppi	GIF
Film, slides & negatives: 35 mm and medium format up to 4x5 in.	4000 across the long dimension	8 bit grayscale; 24-48 bit color: 2800 ppi for 35mm; 800 ppi for 4x5 in. originals	TIFF	800-3000 across the long dimension	8 bit grayscale; 24-48 bit color: 73-300 ppi	JPG or JP2	100-200 across the long dimension	4 bit grayscale 8 bit color: 72 ppi	GIF
Photographic Materials: 8X10 in. or smaller	4000 across the long dimension	8 bit grayscale; 24-48 bit color: from 400 ppi for 8x10 in. ranging down to 570 ppi for 5x7 in., 800 ppi for 4x5 in. or 3 ½ x5 in. originals	TIFF	800-3000 across the long dimension	8 bit grayscale; 24-48 bit color: 73-300 ppi	JPG or JP2	100-200 across the long dimension	4 bit grayscale 8 bit color: 72 ppi	GIF
Photographic Materials: Equal to or larger than 8x10 in. up to 11x14 in.	6000 across the long dimension	8 bit grayscale; 24-48 bit color: from 600 ppi for 8x10 in. ranging down to 430 ppi for 11x14in. originals	TIFF	800-3000 across the long dimension	8 bit grayscale; 24-48 bit color: 73-300 ppi	JPG or JP2	100-200 across the long dimension	4 bit grayscale 8 bit color: 72 ppi	GIF

\*Guidelines are based on the CDL Guidelines for Digital Images: <http://www.cdlib.org/inside/diglib/guidelines/bpgimages/>, June 7, 2005