Digitization Standards and Best Practices for
Indiana Memory and Hoosier State Chronicles Digitization Projects
March 1, 2025

This document provides information on the application of published standards and best practices for digital imaging to determine specifications for individual projects. It includes general principles based on best practices, minimum digitization guidelines for various material types, and provides links to other published standards and best practices at the end.

Indiana Memory and grant-funded projects must adhere to these guidelines.

GENERAL PRINCIPLES

Capture once, use many times

Digitization is expensive, time-consuming, and requires extensive handling of original materials. Any digitization project should thus focus on creating high-quality master files from which many derivative files can be created for specific uses (e.g., web delivery). The master file should capture all "important" information from the original material, which should be explicitly defined for each digitization project. The master file should also be flexible enough to allow derivatives to be created meeting a wide variety of current and future needs. Therefore, no manipulation should be done to the master file. The best practices described in this document are designed to achieve this goal of flexibility.

Create a faithful reproduction of the original

Access files displayed via Indiana Memory should look as close as possible to the original material from which they were derived. Many cultural heritage materials are aged and may show signs of damage from years of wear and tear. For example, black and white photographs can yellow over time, or there may be breakage around the edge of the pages of a diary from the 1800's. It is tempting to eliminate those signs of aging by digitally erasing a tear in the paper, cropping items so as to hide damaged areas, or color-correcting images to get rid of evidence of sun damage or discoloration. It is important not to take these measures, as these "imperfections" are often of interest to artists, historians, archaeologists, archivists, and researchers for a variety of reasons. There are some red, orange, and fluorescent hues that will not look exactly like the original no matter what digital imaging software is used to achieve a color balance, but digital scanning technicians should do their best to ensure that digital surrogates are faithful to the look and condition of the original material.

Digitize from earliest generation practical

As copies are made of analog materials, each generation loses some detail. From a photographic negative to a print to a copy negative, from a book to microfilm, there is generational loss of information. To capture the most information in a scanned image, always use the earliest generation of the original material that it is practical to use. In general, scanning from negatives rather than prints and scanning from original printed material rather than microfilm or photocopies is preferable. However, there are cases where practical considerations dictate using a second- or third-generation original as the source of a scanned image. A set of cracked or broken glass-plate negatives might benefit from professional printing, then scanning the prints. A large series of bound volumes that have been microfilmed would be considerably cheaper to scan from microfilm rather than to unbind the volumes for scanning or invest expensive face-up scanning equipment. In these cases, a determination must be made if images created from later-generation originals can still meet the flexibility goals of master files for the project.

Technical issues

When setting technical specifications for digitization projects, higher is not always better. There is no advantage to scanning at a resolution higher than what is needed to capture the amount of detail on the original. In fact, there is a large disadvantage to this practice in that this excess

resolution adds file size without adding detail to the digital image. The guidelines in this document are designed to help determine appropriate specifications and ensure files are as large as they need to be, but no larger.

A digitization program should employ some sort of color management solution to ensure scanners, monitors, and printers all represent image color accurately. Using “canned” International Color Consortium (ICC) profiles for each imaging device is a low-cost, somewhat effective mechanism, while using professional profiling software is a much more accurate but higher-cost solution.

Using digitization equipment appropriate to the materials being scanned is essential to an effective digitization project. Unfortunately, there are no one-size-fits-all digitization equipment solutions. For example, flatbed scanners are useful for unbound textual materials and photographic prints, while transparencies and negatives are much better imaged with dedicated film scanners. Never use a scanner at a resolution setting above its listed optical resolution (known as an interpolated resolution).

Quality control
A structured quality control program is essential to a good digitization project. An effective program might combine automated checking of objective criteria such as image resolution, file size, dimensions, and bit depth for all images with manual checking of subjective criteria such as color fidelity on a subset of scanned images.

Technical metadata
Recording adequate technical metadata about scanned images is essential for long-term maintenance of master files. The NISO draft standard Technical Metadata for Digital Still Images in its XML Schema form from the Library of Congress at <http://www.loc.gov/standards/mix/> offers guidance on what sorts of technical metadata are appropriate to record.

Copyright
When digitizing materials that are faithful reproductions of the originals, an institution cannot claim copyright on said digitized materials, unless they hold the copyright for the originals. Please do not add digital watermarks or banners to materials that are not under copyright. Materials before January 1, 1930 are in the public domain and be digitized without permission. Materials between January 1, 1930 and December 31, 1977 may be out of copyright and available for digitization, but research to determine copyright status is needed. Materials after January 1, 1978 are almost always copyrighted, unless the creator of the work explicitly placed it in the public domain. To learn more about copyright, please visit https://rightsstatements.org/en/.

INDIANA MEMORY: MINIMUM DIGITIZATION GUIDELINES

In general, the specifications below conform to the guidelines put forth in the June 2008 edition of the Western States Best Practices found at <https://sustainableheritagenetwork.org/system/files/atoms/file/bcrcdpImagingBP.pdf>.This document also expounds more on the general principles outlined above and contain a great deal of additional helpful information.

Originals as TEXT-based materials (Books, pamphlets, etc.)

	Master	Access	Thumbnail
File Format	TIFF	JPEG	JPEG
Bit Depth	1 bit bitonal 8 bit grayscale 24 bit color	1 bit bitonal 8 bit grayscale 24 bit color	1 bit bitonal 8 bit grayscale 24 bit color
Spatial Resolution	300 - 600 ppi (400 ppi and up for OCR purposes)	150 ppi	96 ppi
Spatial Dimensions	100% of original	600 pixels across the long dimension	150-200 pixels across the long dimension

Originals as PHOTOGRAPHS [i]

	Master	Access	Thumbnail
File Format	TIFF	JPEG	JPEG
Bit Depth	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color
Spatial Resolution	300-800 ppi, or 3000 to 5000 pixels across the long dimension	150 ppi	96 ppi
Spatial Dimensions	100% of original	600 pixels across the long dimension	150-200 pixels across the long dimension

Originals as MAPS

	Master	Access	Thumbnail
File Format	TIFF	JPEG	JPEG
Bit Depth	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color
Spatial Resolution	3000 pixels across the long dimension, or 300-400 ppi	150 ppi	96 ppi
Spatial Dimensions	100% of original	600 pixels across the long dimension	150-200 pixels across the long dimension

Originals as GRAPHIC MATERIALS (Broadsides, sheet music, etc.)

	Master	Access	Thumbnail
File Format	TIFF	JPEG	JPEG

Bit Depth	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color	8 bit grayscale 24 bit color
Spatial Resolution	300-600 ppi, or 3000 pixels across the long dimension	150 ppi	96 ppi
Spatial Dimensions	100% of original	600 pixels across the long dimension	150-200 pixels across the long dimension

Originals as PHYSICAL AUDIO-based materials (Vinyl records, audio cassettes, CDs, etc.)

	Master	Access
File Format	WAV	MP3
Bit Depth	24 bit	16 bit
Sample Rate	44.1 kHz	44.1 kHz

Originals as DIGITAL AUDIO-based materials (CD, digital audio file, etc.)

	Master	Access
File Format	WAV	MP3
Bit Depth	24 bit	16 bit
Sample Rate	44.1 kHz	44.1 kHz

Originals as PHYSICAL VIDEO-based materials (35mm film, video cassette, DVD, etc.)

	Master	Access
File Format	AVI (.avi) or QuickTime (.mov)	MPEG-4 (.mp4)
Codec	MPEG-4 AVC (H.264) or DV	MPEG-4 AVC (H.264)
Resolution	640 x 480 resolution (assuming 4:3 original aspect ratio)	320 x 240 resolution (assuming 4:3 original aspect ratio)
Data Rate	30 MiB/s data rate	256-600 kbps data rate

Originals as DIGITAL VIDEO-based materials (DVD, DV camera, smartphone, etc.)

	Master	Access
File Format	AVI (.avi) or QuickTime (.mov)	MPEG-4 (.mp4)
Codec	MPEG-4 AVC (H.264) or DV encoding	MPEG-4 AVC (H.264) encoding

Resolution	1920 x 1080 resolution (assuming 16:9 original aspect ratio)	1280 x 720 resolution (assuming 16:9 original aspect ratio)
Data Rate	30 MiB/s data rate	256-600 kbps data rate

[i] Photographic prints are generally not scanned at a fixed resolution but instead at a fixed number of pixels across the long side, resulting in two differently-sized prints from one negative yielding similarly-sized digital files. The appropriate resolution is determined by dividing the desired number of pixels (e.g. 3000) by the number of inches of the long side of the photograph (e.g. 10" for an 8x10 photo). In this case 3000 / 10 = 300, so an 8x10" print should be scanned at 300ppi.

In general, color photographs should be scanned as 24-bit RGB color and black & white photographs in 8-bit grayscale. There are many cases, however, when black & white photographs would benefit from color scanning, for example, when they are sepia-toned or badly faded.

Technical Guidelines for Hoosier State Chronicles

General Principles: Newspaper digitization projects are required to conform to the technical specifications of Hoosier State Chronicles (www.hoosierstatechronicles.org), which follows the Library of Congress' National Digital Newspaper Program guidelines. An effort should be made to digitize a complete run of the newspaper whenever possible, whether scanning from original paper or microfilm. The resulting digital images should be made freely available for research purposes. Newspapers printed prior to 1930 are in public domain. Please obtain the permission of the newspaper publisher, or research the copyright status, if the paper is more recent.

Newspaper selection factors:

1. Quality of original text & microfilm capture (poorly prepared originals = poor results)
2. Reduction ration used when microfilming the original newspaper (below 20x better)
3. Camera master negative microfilm duplicated should have resolution test patterns readable at 5.0 or higher – can estimate if test pattern not available on film

Scanning & Master file format Scanning

guidelines:

1. Scan from a clean second-generation duplicate microfilm.
2. Capture specifications are 8-bit grayscale at the maximum resolution possible relative to the physical dimensions of original (400 dpi preferred; 300 dpi acceptable)
3. Provide the master page images as uncompressed TIFF 6.0
4. Image naming convention should follow NDNP guidelines (LCCN\reel\issuedate)
5. Newspapers microfilmed with 2 sheets per frame should be split into 2 separate image files
6. Images with more than 3 degrees of skew should be deskewed.
7. Page image files cropped to the page edge – retain the actual edge and up to ¼ inch beyond.

Create derivative files from master TIFF files:

- searchable PDF with hidden text for each page image
- a compressed JPEG 2000 image file

OCR & Associated Information

Summary Guidelines:

1. One OCR text file per page image
2. Each OCR text file name corresponds to the page image it represents
3. Text in UTF-8 character set
4. No graphic elements saved with OCR text
5. OCR text ordered column-by-column (natural reading order)
6. OCR text file with bounding-box coordinate data at word level

PDF requirements:

1. PDF image with Hidden Text for each page image
2. Each searchable PDF file name corresponds to the page image it represents

3. The PDF will not contain bookmarks, links, named destinations, comments, forms, Javascript actions, external cross references, alternate images, embedded thumbnails, annotations, or private data.

Metadata

Each newspaper digitized must be supported by coherent metadata, to provide intellectual access

Check to see if your newspaper title has been cataloged using the U. S. newspaper cataloging guidelines maintained by the Cooperative Online Serials Cataloging CONSER program and included in CONSER database hosted within WORLDCAT. Search the US Newspaper Directory, 1690-Present (http://chroniclingamerica.loc.gov/search/titles/) to find the bibliographic record(s) for your newspaper. If your newspaper is not included, please contact the State Library for further information.

Metadata fields:

Field	Description	Example
Source Repository	owner of source that was digitized: city and state postal abbreviation	Library of Congress; Washington DC
Digital Responsible Institution	Organization responsible for making digital copy; city and state postal abbreviation	Library of Congress; Washington DC
LCCN	An LCCN is an identifier assigned by the Library of Congress for a metadata record (e.g., bibliographic record, authority record).	
Issue Date	Actual date issued, corrected if necessary; Use YYYY-MM-DD format	1908-03-21
Section Label	If present, as printed	C, IV, 3, Business
Page Number	Exactly as printed; if no numbers, supply them.	
Page Physical Description	Valid values are: microfilm, microfiche, print	Microfilm
Reel number	Reel number as printed	
Title		National forum (Washington, D.C.) Washington
County		
City		
Publisher		
Language		

This document is based on the guidelines provided by the Library of Congress for the National Endowment for the Humanities National Digital Newspaper Program: https://www.loc.gov/ndnp/guidelines/NDNP_202426TechNotes.pdf.

Vendors familiar with guidelines:

Apex Covantage - https://www.apexcovantage.com/
4045 Sheridan Avenue, 266
Miami Beach, FL 33140
703.709.3000

Backstage Library Works – http://www.BSLW.com
9 S Commerce Way
Bethlehem, PA 18017
610.758.8700

Canon Solutions America, Inc. - www.csa.canon.com
630 W Carmel Dr.
Carmel IN 46032
317.316.2921

Creekside Digital - http://www.creeksidedigital.com/
5200 Glen Arm Road
Suite Q
Glen Arm, MD 21057
443.213.0335

Crossroads Industrial Services - https://www.crossroadsindustrialservices.com/
8302 East 33rd Street
Indianapolis, IN 46226
317.897.7320

Crowley Company - https://thecrowleycompany.com/
5111 Pegasus Court, Suite M
Frederick, MD 21704
240.215.0224

Digital Divide Data - http://www.digitaldividedata.com/
266 West 37th Street
Suite 803
New York, NY 10018
212.461.3700

George Blood LP - https://www.georgeblood.com/
502 West Office Center Drive
Fort Washington, PA, 19034
215.248.2100

Media Preserve - https://ptlp.com/en/mediapreserve/overview/about-us/
111 Thomson Park Drive
Cranberry Township, PA 16066
800.416.2665

NEDCC - https://www.nedcc.org/audio-preservation/about
100 Brickstone Square
Andover, MA 01810
978.470.1010

Preserve South - https://www.preservesouth.com/
5023 B.U. Bowman Dr.
Buford, GA 30518
770.932.9801

Scene Savers - https://www.scenesavers.com/digitization
424 Scott Street
Covington, KY 41011
800.978.3445

PUBLISHED STANDARDS AND BEST PRACTICES

- CPD Digital Audio Group. Digital Audio Best Practices. Version 2.1. October 2006.
  <https://sustainableheritagenetwork.org/system/files/atoms/file/Audio_Best_Practices.pdf>.
- Digital Library Federation Benchmark Working Group. Benchmark for Faithful Digital Reproductions of Monographs and Serials, Version 1. December 2002
  <http://www.diglib.org/standards/bmarkfin.htm>.
- Kenney, Anne R., and Oya Y. Rieger. Moving Theory into Practice: Digital Imaging for Libraries and Archives. Mountain View, CA: Research Libraries Group, 2000.
  <http://preservationtutorial.library.cornell.edu/contents.html>.
- TEI Text Encoding (TEI): Guidelines for Electronic Text Encoding and Interchange, Version 4.8.1. November 1, 2024.
  <https://tei-c.org/Vault/P5/2.9.1/doc/tei-p5-doc/en/Guidelines.pdf>
- Mountain West Digital Library, Digital Imaging Working Group. BCR's CDP Digital Imaging Best Practices, Version 2.0. June 2008.
  <https://sustainableheritagenetwork.org/system/files/atoms/file/bcrcdpImagingBP.pdf>
- Library of Congress, National Digital Newspaper Program. Overview of Technical Guidelines, November 2019.
  <https://www.loc.gov/ndnp/guidelines/NDNPTechSpecs_Overview.pdf>
- CARLI, Guidelines for the Creation of Digital Collections: Digitization Best Practices for Moving Images, January 9, 2017.
  <https://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_video.pdf>