Digitization Guidelines

This document provides recommendations for the format and resolution of digitized content based on length of time the item will be kept and its purpose. Items that are destined to come to the State Archives or held at a state agency for permanent retention will have specific formats and higher resolutions than content that will only be held for a short period of time at the agency before deletion according to records schedules or are intended to be user copies of a permanent record. Also included here are brief summaries of digitization terms that will be helpful for you to understand in developing your project and filling out the Digitization Project form. This is not meant to be an in-depth explanation of these terms which can be found in more detail in resources listed in the bibliography of the main "Digitization Project Guidance" document.

RESOURCES FOR RECOMMENDATIONS

This document contains recommendations for scanning the most common types of records which you may encounter during your digitization projects. The below recommendations primarily come from the National Archives and Records Administration (NARA) as they have a comprehensive guide of scanning specifications for a wide range of items that are intended for permanent retention and access. This is supplemented by BCR's CDP Digital Imaging Best Practices and FADGI's Technical Guidelines for Digitizing Cultural Heritage Materials. If you have records that don't meet any of the below categories, you may refer to one of these resources or contact the State Archives staff at the Wisconsin Historical Society for additional support.

DIGITIZATION RECOMMENDATIONS

For each of the below categories, the highest recommended resolutions are for items that will be held permanently at a state agency or transferred to the State Archives for permanent retention according to records schedules. For records that will be held at the agency and eventually deleted according to records schedules, the decision as to whether the resolution is High or Minimal will be determined by the agency through an evaluation of the records to be scanned. The PPI resolutions listed in each category below are the minimal recommendations for that document. An agency can go higher if they so choose.

High Resolution

- Records that will be transferred to the State Archives at the end of their retention period according to record schedules.
- Records that will remain permanently with the state agency based on approved record schedules.
- Agency records that will remain with the agency for some period of time prior to deletion according to records schedules and are of significant value to the agency's mission and reputation. The high resolution option should be used if the act of not having a high-quality scan could have legal ramifications or have a negative impact on the agency's reputation.
- Because of the high capture specifications, this file could confidently take the place of the original record if the original record no longer exists. This option should be used where there is a need to ensure that all significant visual information is carried forward.
- Files must be uncompressed with no application/product dependencies or extensions.

Minimal Resolution

- Agency records that will remain with the agency for some period of time prior to deletion according to records schedules and are **not** of significant value to the agency's mission and reputation. These are records that are traditionally of low use by staff and the public after they have reached the inactive stage.
- Because the capture/transfer specifications are closer to or at the minimum recommended level, this file would not be an optimal substitute for the original record if the original record is no longer viable.

Access Copies

- Agency records that will remain in paper for its retention period but will be distributed electronically.
- Agency records that have a high or mid resolution master copy, but need a lower resolution, smaller sized copy for distribution.
- This file is retained only as long as needed to satisfy a specific "on demand" distribution need. It is not retained after delivery to the requesting customer and is not actively managed over a long term.

TEXT DOCUMENTS

Scanned text is a photograph of a printed page produced by digital camera or scanner.

- Bitonal (1-bit black and white) images must be scanned at 300-600 ppi. Scanning at 600 ppi is recommended. This is appropriate for documents that consist exclusively of clean printed type possessing high inherent contrast (e.g., laser printed or typeset on a white background).
- Gray scale (8-bit) must be scanned at 300-400 ppi. Scanning at 400 ppi is recommended. This is appropriate for textual documents of poor legibility because of low inherent contrast, staining or fading (e.g., carbon copies, thermofax, documents with handwritten annotations or other markings), or that contain halftone illustrations or photographs.
- Color (24-bit RGB) must be scanned at 300-400 ppi. Scanning at 400 ppi is recommended. Color mode (if technically available) is appropriate for text containing color information important to interpretation or content.

Content Description	Minimum Scan Resolution	Format	Notes
High Resolution *(NOTE: smallest character being larger than 1.00mm (or 2.835 point/font))*			
Bi-tonal (1–bit black and white)	600 ppi	TIF JPEG2000 PDF/A	Should be uncompressed Should have OCR
Gray scale (8-bit)	400 ppi 8-bit grayscale or 24-bit color		
Color (24-bit RGB)	400 ppi where original is up to 34” x 55” 300 ppi where original is between 34”x55” and 46”x73”		
Mid Resolution			
Bi-tonal (1-bit black and white)	300 ppi for records with smallest significant character of 2.0 mm or larger	TIF PDF/ A PDF JPEG (high resolution) PNG	Uncompressed or lossless compression Should have OCR
Gray scale (8-bit)	300 ppi for records with smallest significant character of 1.5 mm or larger		
Color (24-bit RGB)	300 ppi for records with smallest significant character of 1.5 mm or larger		
Access Copies *(NOTE: OCR works best at 300ppi or higher)*			
Bi-tonal (1–bit black and white)	150 ppi bitonal	TIF JPEG2000 PDF/A PDF JPEG PNG GIF	Can be lossy OCR recommended but depends on use
Gray scale (8-bit)	200 ppi bitonal or grayscale		
Color (24-bit RGB)	300 ppi 24-bit color		

STILL PHOTOGRAPHS-PRINTS

When converting analog materials, agencies should digitize to standards appropriate for the accurate preservation of the original image. Photograph digitization standards are more complex than paper scanning standards since the intent is to maintain the smallest detail of the photo. Size of the photo, quality, medium, and color all play a role in determining the best resolution for the scan.

300 ppi is generally considered the bare minimum to reproduce a photo well at the size of the original. As such, the lowest resolution for scanned images is listed at 400 ppi. The below recommendations assume you match the original size with no magnification or reduction of the physical photo. The minimum resolution section can be used by an agency to produce a decent quality image when responding to reference and reproduction requests. The access section is best used for web distribution of agency photos.

For the categories below, pixel array category defines the minimum resolution of pixels across the long dimension of the object to be scanned. Generally speaking the larger the image, the larger the pixel array.

Photo Size	Bit Depth	Pixel Array	Resolution	Format
High Resolution (agency copy)				
Still photo prints (B&W, color, and monochrome) with a size range of 8" x 10" or smaller (square area smaller than or equal to 80 square inches)	B&W 8-bit grayscale Color & Monochrome 24-bit RGB	4000	Examples: 4” x 5” or 3.5”x 5” 800 ppi 5” x 7” 570 ppi Up to 8” x 10 “ 400 ppi	TIF
Still photo prints (B&W, color, and monochrome) with a size range of larger than 8" x 10" and up to and including 11" x 14" (square area larger than 80 square inches or smaller than 154 square inches)		6000	Examples: 8” x 10 “ 570 ppi Up to 11” x 14” 430 ppi	
Still photo prints (B&W, color, and monochrome) with a size range of larger than 11" x 14" (square area equal to or larger than 154 square inches)		8000	Examples: 11” x 14” 570 ppi Larger sizes will range down to around 400 ppi	
Minimum Resolution (reference and reproduction requests)				
Still photo prints (B&W, color, and monochrome). Match original size with no magnification/reduction	B&W 8-bit grayscale Color & Monochrome 24-bit RGB	3000	Examples: 4” x 5” 600 ppi 8” x 10 “ 300 ppi	TIF JPEG PNG
Access (web copy)				
Still photo prints (B&W, color, and monochrome). Match original size with no magnification/reduction	B&W 8-bit grayscale Color & Monochrome 24-bit RGB	800 - 1200	72 ppi – 200 ppi	GIF (for smaller originals) JPEG (for larger originals, medium →high compression)

ARCHITECTURAL AND ENGINEERING DRAWINGS

Records in this category cover a broad range of technical drawings that are intended for construction or mechanical purposes. Since these items are typically oversized, the scans of the originals have the potential to be very large in size. It is highly likely you will need a large format scanner and specialized handling to legibly capture small details which are prevalent in these materials.

NARA's scanning recommendations for Architectural and Engineering Drawings follow the same rules as the text documents in the above section.

MAPS AND CHARTS

Records in this category are graphic representations of selected geographical features and include such items as atlases, relief models, photomaps, hydrographic/nautical charts and cartograms. These items are typically oversized, and the scans of the originals have the potential to be very large in size. It is highly likely you will need a large format scanner and specialized handling to legibly capture small details which are prevalent in these materials.

NARA's scanning recommendations for Maps and Charts follow the same rules as the text documents in the above section.

TERMINOLOGY

PPI and DPI

Pixels per inch (PPI) refers to the number of pixels captured in a given inch and is used when discussing scanning resolution and on-screen display. When referring to digital capture, PPI is the preferred term, as it more accurately describes the digital image. Dots per inch (DPI) is more often used when discussing the optical resolutions for images and hardware. This refers to how many dots per inch a printer puts on paper when printing it out or across a computer monitor. For the purposes of this document, PPI will be used for the recommended capture resolutions of various objects.

Resolution

Resolution refers to the quality of an image. The higher the PPI, the more accurate rendering of the original document is created. Images at low PPI may look fine on a computer screen, but will be illegible when printed out. High resolution images capture more information about the original document and therefore take up more server space in terms of storage. PPI will also vary depending on the object being scanned. An 8x10 photograph will most likely require a different resolution than an 8x10 text document.

Compression

Digitized image files have the potential to take up a lot of room on a drive. Uncompressed files generally take up the most space and are often considered the "raw" image that comes out of the scanning process or camera. Many scanning operations will perform some sort of image compression to reduce the file size for storage, processing and transmission. There are two key compression schemes that you will encounter during the digitization process but they have very different results.

Lossless compression

Lossless compression abbreviates the underlying binary code without discarding any of the information so when it opens, it can reconstruct the code and the image will be a perfect copy of the original. Using this technique, every single bit of data that was originally in the file remains after the file is uncompressed. While items stored with lossless compression tend to be larger than those stored with lossy compression, this is preferable for items that will be remain a long time at the agency or transferred to the archives for permanent retention.

Lossy compression

With lossy compression, the file size is reduced during the compression process by permanently eliminating information from the image. When the file is opened (uncompressed), only a part of the original information is still there. Although the discarded information may be invisible to the human eye when viewing the image, a loss of quality has occurred. Each time a lossy image is manipulated or edited, the quality of the image further decreases. Generational loss over time is the primary reason that lossy compression may be a good option for user copies or documents of a short-term duration, but not for an object that will be the official record for long periods of time or are of permanent duration.

FORMATS

The use of the digital images you are creating will determine which formats to choose when digitizing. While there are seemingly endless formats to choose from, only a small handful are currently viewed as being sustainable for preserving digital information as an authentic resource for future generations. These formats tend to be widely adopted; well documented, non-proprietary and has no external hardware or software dependencies. It is considered best practice to use these formats as they tend to be supported across most systems and and less likely to become obsolete over time. Choosing the correct format is part of a strong strategy for future preservation actions on the records such as the adoption of new technologies or the migration to new formats. A list of recommended formats for digitization projects and their uses are below.

GIF: GIFs are best used to store screen-quality images that do not contain many colors. GIF files are typically very small, but cannot reproduce the range of colors necessary to reproduce photographic images like JPEG. As such, this format is recommended for user copies of text records

JPEG: This lossy image file format is commonly used for photographs and other complex images. It is a great option for user copies on a website since you can create smaller sized images by reducing the image quality. This format is not recommended for text or line drawings.

JPEG 2000: JPEG 2000 is another lossless format that is increasingly being used as a long-term / archival format and has been adopted as such at the National Archives and has been published as an ISO standard. A downside to this format is that access to JPEG 2000 files may require a special reader on the user-end but access is becoming easier as JPEG 2000 becomes more ubiquitous. Also large in file size, this will take up more server space than most other formats. This format is acceptable for items being transferred to the State Archives for permanent preservation although you may prefer one of the other options for content that will stay with the agency. JPEG 2000 images can be lossy or lossless depending on the purpose of the final document. This format is not widely supported by web browsers and is not generally used on the internet.

PNG: PNG (Portable Network Graphics) is an open standard graphics file format that allows accurate rendering of greyscale and RGB color objects. PNG format can be used to store high-color images, which means it is also suitable for storing photographic content. This format is not widely implemented.

PDF: PDF (Portable Document Format) is best used to store vector-based graphics (i.e. graphics drawn using lines and curves rather than pixels). Vector graphics stored in PDF format will be much smaller, will read more cleanly, and any included text will be searchable. Equations, charts, and diagrams that combine text with vector-graphics are particularly appropriate to store in PDF format.

PDF/A: PDF/A is a special type of PDF format meant for documents needing to be preserved for long-periods of time. This format is an ISO standard which helps guarantee that it will be accessible for the foreseeable future as technology advances.

TIFF: TIFF (Tagged Image File Format) is a stable, well-documented, widely adopted, uncompressed file format and is a preferred format for archival and master images. This is an excellent choice for digitized content that will be replacing the paper records in a record schedule and will be preserved for a long-period of time → permanent. The downside, of course is that they will take up more server space than most of the other options. TIFF images can be lossy or lossless depending on the purpose of the final document. Its file extension is .TIF

MODES OF CAPTURE

Most imaging equipment offers four modes for capturing a digital image. The mode you choose is dependent on the content that will be scanned.

Mode of Capture	What it is	Best for
Bitonal	One bit per pixel representing black and white.	Best suited to high-contrast documents such as printed text
Greyscale	Multiple bits per pixel representing shades of gray. Grayscale has 1 color channel – typically black or white.	Grayscale is suited to continuous tone items, such as black and white photographs. Standard is either 8-bit or 16-bit grayscale images (more shades)
RGB (color mode)	Multiple bits per pixel made by combining 3 color channels: Red, green or blue.	Color capture is suited to items with continuous tone color information. Standard is either 24-bit color or 48 bit color images (more shades)
CMYK (color mode)	Short for Cyan-Magenta-Yellow-Black and is the standard model used in offset printing for full color documents. This mode is best for printing and should never be used for archival purposes.	Best when printing color images (not archival)

BIT DEPTH

Bit depth quantifies how many unique colors are available in an image's color palette in terms of the number of 0's and 1's, or "bits," which are used to specify each color. The lower the bit depth, the fewer colors are available to represent a scanned object.

So for an object with a bit depth of 1, each pixel in the image will have two possible values – either white or black. As the bit depth increases for both grayscale and colored objects, more colors are added so highly detailed images with lots of color variations, gradients, shadows, etc. become clearer and a truer representation of the original object. A higher bit depth also means that more information is being captured and stored for the image which can lead to a larger file size.

Bit Depth	# of Colors	Colors Seen	How it works
1 bit Bitonal	2^1 or 2 colors	Black and White	An image with a bit depth of 1 has pixels with two possible values: black and white
8 bit Grayscale	2^8 or 256 colors	Black, white , gray	An image with a bit depth of 8 has pixels with 256 possible values: black, white and various shades of gray
24 bit Color	2^{24} or 16,777,216 colors	Colors across the Red, Green Blue Spectrum	RGB images are made of three color channels. An 8 bit per pixel RGB image has 256 possible values for each channel

Resources referenced for this document

CDP Digital Imaging Best Practices Working Group. BCR's CDP Digital Imaging Best Practices V.2.0. 2008. http://sustainableheritagenetwork.org/digital-heritage/bcrs-collaborative-digitization-program-digital-imaging-best-practices-version-20 (accessed June 2017).

FADGI - Still Image Working Group. *Technical Guidelines for Digitizing Cultural Heritage Materials*. 2016. http://www.digitizationguidelines.gov/guidelines/digitize-technical.html (accessed June 2017).

Library of Congress. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. March 2017. https://www.loc.gov/preservation/digital/formats/index.shtml (accessed June 2017).

National Archives and Records Administration (NARA). *Preservation -- Original Record Type*. 2017. https://www.archives.gov/preservation/products/definitions/original.html (accessed June 2017). (Congress 2017)