Guidelines for Creation of Digital Collections

Digitization Best Practices

Karen Keehr, Archivist
Draft: February 2023

The goal of the American Historical Society of Germans from Russia (AHSGR) digitization program is to increase access and awareness of AHSGR's holdings. This document provides information about digital imaging or the scanning of records. It specifically refers to records that were originally in another format and are captured digitally, rather than those records that are born electronically. Topics cover will include advantages, basic principles, archival standards, modes of capture, image formats, file naming conventions, and combining multiple files into a single digital object. The appendix will include document handling guidelines, definitions, sample decision charts for collection selection and a sample digitization workflow.

Access versus preservation is an age-old problem in the museum and archive profession. As caretakers of history, we do all that we can to ensure our collections endure as long as possible. We keep our collections in the best environments that we can manage, wear gloves when handling objects, and house our collection in archival boxes. The other important part of archival work is creating access to the very objects we are trying to protect. This, of course, often goes against the principles of preservation. Digital images balance providing access to collections and reducing damage caused by overhandling of objects. Creating a digital derivative allows us to keep the originals safe and meet researchers' needs.

These guidelines explain the minimum requirements for digitization standards, as well as recommended and acceptable file types. Whenever possible, all newly acquired materials should adhere to these minimum requirements, although higher standards may be desirable for some materials. These guidelines are not intended to address all technical issues related to digitization.

Our standards are based on our need to balance the quality of the digital object with user accessibility. For that reason, we recommend two versions of digital files: one that is an archival quality master file and one that is an access copy intended for web delivery. These standards represent our current needs and are subject to change, given the future needs of the repository and our users.

Before starting any imaging project, knowing the project's mission, users, priorities (speed, image quality, and quantity), and functional goals (reference, web use, publications, other). Additionally, access staff expertise and availability (to scan, manage infrastructure, migrate data, and build metadata), and address content issues, such as physical condition, format, nature, and attributes to be captured.

There are many advantages and disadvantages to Digital Imaging:

Advantages:

- Ability to use very high-density electronic storage media instead of paper.
- Shorter retrieval time when the images are well indexed.
- Multiple users and access levels are possible.
- Low shipping costs and ease of information dissemination.
- Ease of use of imaged copies of records in vital records and disaster recovery plans.
- Ready access to digitized records may assist organizations needing to retrieve information efficiently during litigation and discovery.
- Ease of making copies of the imaged records.
- No loss of digital image quality from generation to generation. Well-made copies and derivatives can be as good as the original images.

Disadvantages:

- Digital images are not human-readable without computer equipment.
- Significant equipment costs, including hardware and software.
- Potential for hardware and software obsolescence. Generally, systems change every 18 months to 5 years, software changes every 2-3 years, and the life expectancy of media is relatively short.
- Indexing requirements may be more extensive than is required with other formats. Unless records are arranged in a logical sequence or clearly indexed, it may be difficult to identify a series or use groups of records as a series.
- Different types of scanners may be required to scan text, oversize items, photographic prints, slides, and other formats.
- Digital quality control and image and metadata capture and management are complex, time-consuming processes requiring expertise.

Other factors to consider:

- **Volume of records**. Imaging is generally used for large volumes of records.
- **Reference use**. Imaging is most effective on highly referenced collections where a short retrieval time is important or where there are multiple users accessing the same records. Combined with effective indexing, imaging records can facilitate retrieval.
- **Relationship to records on other media**. Consider whether the records to be imaged have to be used with records on other media.
- **Records and information usage**. Consider how the information is used and how long the record is needed. Required retention periods are specified in records schedules.
- **Legal acceptability**. Following established procedures and maintaining the documentation of audit trails and other business practices will ensure that information is kept that may be needed to document record authenticity and reliability.

- **Ease of maintenance**. Balance storage costs and capacity with indexing, conversion, quality control, and migration costs.
- **Staffing requirements**. Increased imaging and indexing of records and quality control procedures may require additional staff training.
- **Work process and information flow**. Would imaging facilitate the work process? Considerations include how records are routed, how information is added to records or files, and when records (finals or drafts) need to be captured.
- **Document preparation**. Determine how much work needs to be done to make the files ready for imaging. Document preparation for voluminous files may be significant.
- **Quality control issues**. Procedures must be instituted both while preparing documents for imaging and while verifying and validating imaged information.
- **Condition of original records**. The condition of the records will affect their handling during imaging as well as the quality of the imaged record that can be produced. This will particularly be a factor for records that are damaged, faded, or oversized.
- **In-house operation versus contracting operations with a service bureau**.
- **Image requirements** (resolution, compression, headers, etc.) will vary depending on how images will be used and the condition of the originals.
- **Indexing requirements** and metadata fields are determined by analyzing how users will access images.

Digital Imaging Basic Principles:

- Scan at the highest resolution appropriate.
- Scan at an appropriate level of quality to avoid re-scanning.
- Scan the original document rather than copy to capture the best quality image.
- Create an uncompressed master image file which can be used to produce access copies.
- Select equipment based on optical resolution as opposed to interpolated resolution as this will produce more accurate scans.
- Use non-proprietary components.
- Monitor and recopy files as needed.
- Implement a migration strategy to transfer data across generations of technology.
    - Transfer files to new media as it becomes widely available.
    - Do not let more than 5 years elapse before refreshing your data.
    - Longevity is less important than the ability to access.

Resolution:

Resolution is the number of pixels (in both height and width) making up an image. The more pixels in an image, the higher the resolution, and the higher the resolution of an image, the greater its clarity and definition (and the larger the file size). Resolution can also refer to the output device, such as a computer monitor or printer, used to display the image. Image file

resolution is often expressed as a ratio (such as 640x480 pixels), as is monitor resolution; however, resolution is also expressed in terms of pixels per inch (ppi). Image file resolution and output (print or display) resolution combine to influence the clarity of a digital image when it is viewed.

Modes of Capture:

- **1-bit or Bitonal**: Means a pixel can be black or white. Bitonal imaging is ideal for black and white images, such as line drawings and text. However, scanning in grayscale rather than bitonal may produce a better looking image.
- **8-bit grayscale**: Each pixel can be one of 256 shades of gray. Good for black and white photography, and handwritten documents.
- **Color:** Each pixel can be one of 16.8 million colors. Used for documents with continuous tone color information. Color images should be saved as RGB files instead of CMYK files, due to the limited color range of CMYK

Archival Records Standards for Digital Images:

Material	Bit Depth	Color Mode	Resolution	File Format
Printed Text	16-bit or 24-bit	Color	400-600 ppi	Uncompress Tiff or pdf
Handwritten Text	16-bit or 24-bit	Color	400-600 ppi	Uncompress Tiff or pdf
Newspapers	16-bit or 24 bit	Greyscale or Color	400-600 ppi	Uncompress Tiff or pdf
Photographs	24-bit	Color	5000 pixels across long dimension	Uncompress Tiff
Negatives and Transparencies	16-bit or 24-bit	Greyscale or Color	5000 pixels across long dimension	Uncompress Tiff
Oversized Materials (maps, surname charts, certificates, etc.)	16-bit or 24-bit	Color	400-600 ppi	Uncompress Tiff or pdf

Master File:

The digital master image represents, as accurately as possible, the visual information in the original objects. 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. A high-quality master image eliminates the need to re-digitize, and therefore re-handle, the same potentially fragile physical materials

again in the future. A master image should also support the production of a printed page facsimile that is a legible and faithful stand in for the original when printed at the same size.

Some general guidelines for creating digital master files:

- When scanning objects, the scanning resolution may need to be adjusted according to the size of the object. For example, when scanning documents with very small text, higher resolution and bit depth may be required. A higher resolution may offer increased accuracy for Optical Character Recognition (OCR) processing.
- 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.
- Where possible, scanning guidelines for the creation of digital master files should follow the specification outlined in the Federal Agencies Digitization Initiative (FADGI).
- AHSGR DI staff using PastPerfect should not upload full resolution TIFF files to PPS as a file storage solution. Archival image files should be stored on Google Drive. Low resolution derivatives should be made from Master Files and attached to PPS Records.

Derivative Images:

Derivative files are used for editing and enhancement, conversion to different formats, and presentation or transmission over networks. In the case of text works that comprise more than one-page, derivative images can be compiled into a single file that represents the entire work.

Derivative images can be created using image editing software such as Adobe Photoshop. Some applications, like Adobe Acrobat, can automatically downsize images when compiling a file made up of multiple page images, eliminating the need to create derivative copies by hand.

Access Images:

Access images represent the version of the image that users viewing the digital items online interact with. Access images should be of sufficient size and resolution to allow for detailed study, but not so large that they take too long to load in a browser. Access images may also be edited to improve the view experience for the user, through such processes as cropping, straightening, color correction, sharpening, or descreening. These edits can be made using image editing software such as Adobe Photoshop.

Recommendations:

	File Format	**Pixel Array & Resolution**	**Bit depth**
Master Image	TIFF or PDF	400-600 ppi for text or 5000 pixels across long dimension for photos	8-bit greyscale or 24-bit color
Access Image	JPG or PDF	72-200 ppi	8-bit greyscale or 24-bit color

File Naming Conventions:

Each digital object in a collection will 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 the collection's needs and specification. Archival collections file names should base on the Collection Number and its place within the collection (group and item numbers or series and folder, etc.)

Combining Multiple Files into a Single Digital Object

Text material often consists of many pages that collectively comprise a single work. Therefore, a digital facsimile of such a work must include a way to compile many separate scans and images into a single file format that maintains the order and structure of the original object, such as a pdf.

Document Handling Guidelines

- When scanning fragile materials, do not use automatic feed devices including the oversize scanner. If you are unsure, consult with the Archivist.
- Make sure the entire item being scanned is being supported at all times. Do not let part of the item hang loosely off the side of the flatbed scanner.
- Do not exert pressure on the top of the scanner to "flatten" an object or book
- Do not leave archival material unattended or uncovered on digitization equipment or elsewhere. When done scanning for the day, return the project you are working on to the box and store on processing shelves.
    - Be respectful of the shared workspace
- Keep material flat. Rolling, pulling, bending, or folding archival material is not permitted.
- Unroll or rewind archival material carefully.
    - Rewind any rolls of film with the emulsion side in.
- Hand bound volumes carefully and do not force them open or push on their spines to flatten them.
- Remove encapsulated material from their enclosures
- Report any material you feel cannot be safely digitized to the Archivist for further evaluation.
- Keep all materials in their original order and return them to their original jackets or containers.
    - Photographs and negatives should be placed in envelops with the emulsion facing away from the enclosure seems.
- No drinking or eating in the room where archival material or their containers are located
- Use clean dry hands when handling paper material and gloves when handling photographic material
    - Avoid using hand lotions or sanitizers while working with archival material.
- Do not use pens around archival material.
    - Do not write directly on archival material unless directed to do so by the Archivist.

    - Use pencils to label enclosures, folders and boxes.
- Do not lean on, sit on or otherwise apply pressure to archival material or their containers.
- AHSGR will provide staff and volunteers training in proper handing of artifacts and archival materials.

Definitions

- **Bitonal:** Means that each pixel can be either black or white.
- **Checksum:** A program which allows the user to see how many bytes were present before and after migration to detect information lost during migration.
- **Compression:** The reduction of image file size for processing, storage, and transmission. The quality of the image may be affected by the compression techniques used and the level of compression applied. There are two types of compression:
    - **Lossless** compression is a process that reduces the storage space needed for an image file without loss of data. If an image has undergone lossless compression, it will be identical to the image before it was compressed. Primarily used with bitonal images.
    - **Lossy** compression is another process that reduces the storage space needed for an image file, but it discards information that is "redundant" and not perceptible to the human eye. If an image that has undergone lossy compression is decompressed, it will differ from the image before it was compressed, even though the difference may be difficult for the human eye to detect.
- **Conversion:** The task of moving data from an existing format to a different format due to obsolescence.
- **Derived Image (Derivative Image):** An image that has been created from another image through some kind of automated process, usually involving a loss of information.
- **Dots per inch (dpi):** A measurement of the scanning resolution of an image or the quality of an output device. DPI expresses the number of dots a printer can print per inch or that a monitor can display both horizontally and vertically.
- **Interpolate:** To calculate or estimate intermediate values occurring between two known values. This function is acquired artificially through software.
- **Linear Tape Open (LTO):** An “open format” technology, which give the user multiple sources of product and media to choose from.
- **Metadata:** Data about data, or information known about the image in order to provide access to the image. Usually includes information about the intellectual content of the image, digital representation data, and security or rights management information.
- **Migration:** Preserving the integrity of digital images by transferring them across hardware and software configurations and across subsequent generations of computer technology. Migration includes refreshment (copying digital files from one media to another) as a means of preservation and access. However, migration differs from refreshment in the sense that it is not always possible to make an exact copy of a database or even an image file (as changes in hardware and software occur) and still maintain compatibility with the new generation of technology.

- **Pixels:** Often referred to as dot, as in "dots per inch". "Pixel" is short for picture elements, which make up an image, similar to grains in a photograph or dots in a half-tone. Each pixel can represent a number of different shades or colors, depending on how much storage space is allocated for it. Pixels per inch (ppi) is sometimes the preferred term, as it more accurately describes the digital image.
    - **Redundant Array of Independent Disks (RAID):** A redundancy subsystem of disk drives that improves performance, fault tolerance, and data recovery.
- **Resolution:** The number of pixels (in both height and width) making up an image. The more pixels in an image, the higher the resolution, and the higher the resolution of an image, the greater its clarity and definition (and the larger the file size). Resolution is expressed in pixels for image files or as dots per square inch (dpi) for prints. It can also refer to the output device, such as a computer monitor or printer, used to display the image, which is expressed as dots per inch.