Digitization Guidelines

Introduction

This document contains the Maryland State Archives recommendations for digitizing the most common types of records that can be reproduced as still images (printed text, manuscripts, maps, photo prints, etc). If you have a unique item not covered in this document please feel free to contact the Maryland State Archives for further information.

These digitization guidelines adhere to standards and best practices developed by the Federal Agencies Digitization Guidelines Initiative (FADGI). Updated in 2017 FADGI'S "Technical Guidelines for Digitizing Cultural Heritage Materials" provides an in depth guide of scanning specifications for a wide range of media types. A link to these guidelines is in the resources section of this document.

If there are concerns that digitization will cause damage to the original due to its condition or equipment limitations please refer to the resources section to seek additional information or contact the Maryland State Archives Conservation Lab for assistance via email at msa.helpdesk@maryland.gov.

Contents

Digitization Terms
Below are terms that you will encounter frequently with any digitization project.

***Digitization*-**Digitization is a process by which a document or photo is scanned and converted to a digital format. After scanning , the original document or photo is represented by a series of pixels. The image can then be kept on a network or transferred onto a variety of storage options

Pixels- A pixel is the smallest unit of a digital image or graphic that can be displayed and represented on a digital display device.A pixel is the basic logical unit in digital graphics. Pixels are combined to form a complete image, video, text or any visible thing on a computer display.

PPI/DPI- Technically speaking, PPI (pixels-per-inch) is the way that image resolution is properly described; it affects the size and quality of the image. DPI (dots-per-inch) is better suited to describing the resolution of printers and printed output. PPI and DPI are often used interchangeably.

Resolution- The quality of a digital image is partially dependent on the initial scanning resolution. Resolution is expressed in the number of pixels used to represent a image (DPI/PPI) . The higher your image resolution the larger your file size will be.Optical resolution is the actual resolution that digitization equipment (scanner, digital camera) is capable of capturing. Interpolation is the computer filling in or guessing to make up the resolution between what can actually be captured and what is being requested. Interpolation is rarely recommended when scanning, but works well for printing images for large posters.

Compression- Compression is 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 and Lossy Compression. In selecting a compression technique, it is necessary to consider the attributes of the original object. Some compression techniques are designed to compress text; others are designed to compress pictures.

Color Depth - (also known as Bit Depth) The number of possible shades or tonal gradations that a color can have from black to white. A bitonal image is 1-bit (2^1, or 2 colors). Grayscale images are typically 8-bit (2^8 or 256 values). Color images are typically 24-bit (3 colors, 8-bits per color, 16 million values).

- ***Bitonal - (bilevel, binary, or 1-bit)*** Bitonal means that each pixel in the image file can only have one of two tonal values, black or white (the tonal value can be stored in one bit of digital data, hence 1-bit or binary). Bitonal images are easier for OCR software to interpret. Because of the limited color range, bitonal images are dramatically smaller than a grayscale or color files.
- ***Grayscale*** - a black-and-white form of continuous tone imagery. Unlike bitonal images, where one two tonal values can be described, grayscale images are (typically) composed of 256 shades of gray (2^8 or 8-bit), varying from black at the weakest intensity to white at the strongest. High-end scanners are capable of capturing 12-bit (2^{12}) and 16-bit (2^{16}) grayscale. Grayscale images are also called monochromatic, as they only capture one channel of color.

- ***Color - ('truecolor')*** The representation of color images on a monitor is done with the RGB (red-green-blue) color model. Whereas grayscale uses one color channel, color images use 3 channels (one each for red, green, and blue). Typically, each color channel has 8 bits or 256 values from darkest to lightest, resulting in 24-bit color. On Macintosh computers, 24-bit color is referred to as "millions of colors" because 256 x 256 x 256 = 16,777,216 possible color combinations. As with grayscale, high-end

scanners can also capture 36-bit (12 bits per channel) and 48-bit color (16 bits per channel).

Metadata- Usually defined as "data about data" is used to describe an object (digital or otherwise), its relationships with other objects, and how the object has been and should be treated over time. A structured format and a controlled vocabulary, which together allow for a precise and comprehensible description of content, location, and value, are its basic elements. Metadata often includes items like file type, file name, creator name, date of creation, and the record's classification.

OCR- Is the process of electronically translating a scanned image of text material into machine readable text. A program will read the character content within the image and creates a digital version of the text.This allows the text to be searched and indexed, or used in other processes. The accuracy of the OCR depends a number of factors including image quality.

File Format (s)

The specific way digital information is made and stored by the computer. Not all programs are compatible with all file formats. Images created with the intent of replicating an original document are considered a master image or copy. Master copies should be of high quality and follow recommend standards when possible. Master copies are not to be used on a regular basis. Access images are generally copies of master files whose main purpose is to provide access to users. Access files are normally lower in quality resulting in smaller file size which allows easier sharing/access.

Below are a few common formats:

- ***TIFF*** files, are widely usable in many different programs. TIFF files utilize lossless compression and are commonly used for master copies. Files in TIFF format end with a .tif extension.
- ***JPEG*** is a lossy compression for color and grayscale images. Depending on the degree of compression, the loss of detail may or may not be visible to the eye. Files in JPEG format end with a .jpg extension.
- ***JPEG 2000*** uses a image compression to produce both lossy and lossless digital files. Lossless images may compete with TIFF files for archival quality masters. Files in JPEG2000 format use .jp2, .jpf and other extension's.
- ***PDF*** contain a image of a page, including text and graphics. PDF files are widely used for read-only file sharing. Adobe Acrobat is, by far, the most popular PDF file application.
- ***PDF/A*** is a standard file format for long-term archiving of electronic documents , is a subset of PDF, Files are 100% self contained and do not rely on outside sources for document information.

Minimum Digitization Guidelines

This is a simplified chart of common scanning guidelines. For a more in depth review see the resources section.

Manuscripts, Printed Text, Books, Photographs

Master File Format	TIFF, JPG 2000
Access File Format	Any (TIFF, JPG 2000, Jpg, PDF/A)
Resolution	300 ppi
Bit Depth	8 Bit Grayscale or 24 Bit Color where appropriate
Pixel Dimensions	30000 pixels across

Photograph Negatives, 35mm to “4x5”

Master File Format	TIFF, JPG 2000
Access File Format	Any (TIFF, JPG 2000, Jpg, PDF/A)
Resolution	“8 x 10” 300 ppi “4 x 5” 600 ppi 35 mm neg 2100 ppi
Bit Depth	8 Bit Grayscale or 24 Bit Color where appropriate
Pixel Dimensions	30000 pixels across

In House vs Outsourcing

Most agencies do not have the appropriate scanning equipment, software, or staff expertise to execute a digitization project. Evaluation of your resources will help determine if your

digitization process should be done in-house or outsourced to a vendor who specializes in digital imaging.

Vendors provide digitizing services, technical advice, and sometimes the long-term maintenance of the resulting files. Before talking to vendors, be familiar with digitization technology and have a clear idea of your project and its goals.

Questions to ask internally before contacting a vendor:

- How much material will be digitized? What type of materials will be digitized?
- Can the materials leave your site? What precautions are necessary to ensure the security of the materials?
- What is the physical condition of the materials? Do they need to be prepared for scanning (removing staples and paperclips)? Do they have any special handling requirements that would keep them from being outsourced? Can they be transported easily?
- What is the required quality of the digital images? High or low resolution? Black and white or color?
- What is the desired end product? A document management system? A searchable online collection? Who is the intended audience? Staff members? Researchers? The general public?
- Why are you digitizing the materials? What file format(s) fit your requirements? Do you need both master and access copies? How will each be created? And when? Do the access copies need to be watermarked?
- What will happen to the original paper documents that were imaged? Do they need to be kept for any reason? Local access? Retention schedules? If not, how will they be properly disposed of?

The Northeast Document Conservation Center highlights issues relating to working with vendors. A link can be found in the resources section.

Resources

SERI Digitization Projects Guidelines
https://www.statearchivists.org/files/6015/0272/2035/COSA_DigitizationProjects_final.pdf

FADGI Digitization Guidelines
http://www.digitizationguidelines.gov/guidelines/digitize-technical.html

NARA
https://www.archives.gov/preservation/technical/guidelines.html

ITCP Metadata
https://iptc.org/standards/photo-metadata/

Preservation Guidelines for Digitizing
https://www.loc.gov/preservation/care/scan.html

External Digitization Standards/ Society of American Archivists
https://www2.archivists.org/standards/external/123

Northeast Document Conservation Center
:https://www.nedcc.org/free-resources/preservation-leaflets/6.-reformatting/6.7-outsourcing-and-vendor-relations

National Archives: Table of File Formats
https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html#digitalstillimages