

# Digital Library Center Digitization Standards and Procedures

## University of Tennessee Knoxville

This document outlines digitization standards and procedures that will be implemented in the University of Tennessee's Digital Library Center (UT DLC). These standards are meant to serve as benchmarks for digitization projects conducted in the DLC. The benchmarks and procedures are recommendations for minimum standards and best practice and cannot account for all variables involved in a digitization project.

### Rationalization for Standards

The UT DLC was created in 2001 to address the university community's needs for electronic publishing, digital collection acquisition, digitization of unique research materials, access to digital collections, and facilitation the integration of digital resources into core teaching and learning (*Proposal for a University of Tennessee Digital Library Center, 2001*). Each DLC project presents unique and varied issues that staff must address. The minimum digitization standards in this document will reduce project-planning time, enable staff to concentrate more on the specifics of a project, and increase the consistency of the digital objects created in the DLC.

### Creating Digital Images

The goal of the DLC is to create digital objects that capture the original as well as possible, serve as surrogates or archival copies for preservation purposes, function across platforms, and allow for quality derivative image creation.

#### Archival/Master image

An archival or master image is one that is scanned or otherwise captured at a high resolution. The master image is uncompressed and unedited; thus as similar to the original as possible. The DLC may save the master image as an archival copy or surrogate of physical object, since many items are on loan to UT for digitization purposes only and may not be returned for rescanning. The Tiff format (Tag Image File Format) is chosen for master images because of its interoperability, large data capture, and non-proprietary nature. Tiff files are large, so adequate storage space for a collection should be considered.

#### Derivative images

Derivative file(s) are created from the archival/master image. They are often called Access images and/or Thumbnail images. Using image-editing software, such as Adobe Photoshop, DLC staff can alter the master image, resize it, and save it in a format that is more suitable for viewing on the Web or another electronic medium. The level of quality in a derivative file will not be as high as the master image, but the file size will be smaller and more manageable.

### Imaging Basics

Factors such as the quality and type of the originals, the needs of the users and how the collection will be used affect what imaging decisions should be made in the DLC. Each collection will present different needs, so testing using the minimum guidelines and DLC equipment should be conducted.

#### Image Capture

There are three modes for capturing an image digitally:

- **Bitonal** - best suited to high contrast documents such as printed text. Used for documents such as reports, memos, or other items with black and white only.

- **Grayscale** - best suited to continuous tone documents such as b&w photographs. Used for black and white photographs, although for older photos determination should be made whether color will capture the current state of the photograph more accurately (e.g., sepia tones).
- **Color** - suited to documents with continuous tone color information. Used for color photos and other documents with color representation.

### Spatial Resolution

Dots per inch (dpi) or pixels per inch (ppi) relates to spatial resolution. The term dpi refers to the dot frequency in a printer, while the term ppi refers to the sampling frequency of a scanner or camera. The DLC uses the term ppi, because most digital capture is done by scanner and there is little demand for printing.

A higher resolution will represent the original more accurately because it captures more information or pixels. Since the physical size of collection items will vary, it is important to adjust the spatial resolution to get optimal results. To ensure consistent results, determine the necessary resolution using the following formula:

Desired long side in pixels /	Length of item's long side in inches =	Resolution in ppi
-------------------------------	--	-------------------

*For example:* To determine the resolution an 8x10 image will need to be scanned at to get a digital image with the long side of 5,000 pixels:

Desired long side in pixels /	Length of item's long side in inches =	Resolution in ppi
5,000 /	10 =	500

### Bit Depth

Bit depth is determined by the number of bits used to define each pixel. The greater the bit depth, the greater the number of tones (grayscale or color) that can be represented.

Digital images may be:

- 1 bit or bitonal is usually black and white
- 8 bit is 256 colors/shades of gray
- 24 bit is 16 million colors/shades of gray

## Scanner Basics

The philosophy of a standard scanning procedure can range from zero image manipulation within the scanner software to complete manipulation of the image's color balance, dynamic range, etc. by the scanning software. The standard scanning procedure for DLC is a middle ground between these two. That is, while we recognize the need to preserve the "raw" image without manipulation, there is also a practical advantage in workflow efficiency to automating some of the necessary routine image adjustment during the scan. Therefore, it is the DLC standard procedure to do a conservative amount of adjustment to the image during the scan in order to minimize the amount of post-scan processing required, while preserving as much of the original scan without manipulation. The details of this procedure are outlined in the *Scanner Workflow* document.

### Scanner and Imaging Software

The DLC uses Silverfast scanning software for both flatbed and film scanners, and Photoshop for post-scan processing of images. In addition, Photoshop plugin software is used for sharpening and compression of derivatives.

## Color Management

All hardware in the DLC workflow must be calibrated in order to ensure integrity of color throughout the process. Monitor calibration must be performed, using the Eye-one system from Gretag Macbeth, quarterly as well as before the beginning of a new project. Scanners are calibrated using the color targets provided with the Silverfast software. The color management section of Silverfast must then be properly set to incorporate the scanner profile.

The DLC uses the Adobe RGB (1998) as the working colorspace during the scan and post-processing in Photoshop. This colorspace is imbedded in the saved .tif images. It is also standard procedure to convert derivative images (i.e. web-based view and thumbnail images) to the sRGB colorspace, which is the most suitable for viewing images on web pages. See the *DLC Color Management Guidelines* document for details.

## Workspace Standards

In this document, the term “workspace standards” refers to room lighting and monitor conditions that enable consistent and predictable color perception during the digitization process. The human eye is very susceptible to being distracted by poor lighting, computer screensavers, even the color of the walls in the workspace.

The DLC standard workspace calls for the computer screensaver set to plain black and the desktop background set to neutral gray. Lighting in the standard workspace should be from Solux bulbs, which are capable of simulating 4700K daylight conditions ([www.soluxtli.com](http://www.soluxtli.com)). If possible, the walls should be of a neutral color, without distracting colorful pictures or other objects.

## Creating Derivative Images

The steps for creating a derivative image vary depending on the level of image processing required. The two main scenarios for processing are item and batch processing.

Item processing is necessary when images require processing that is specific to each individual image. These include cropping, levels adjustment and the repair of flaws (tears). This level of treatment requires a significant investment of time and resources and so should only be used when necessary.

Batch processing is possible when a group of images all require the same general manipulations. These include sharpening, size reduction and file format change. For specific instructions on how to run a batch process using Photoshop see the instructions for *Creating an Action* and *Batch Processing*.

Currently JPEG is the file format used for most derivative images created for DLC projects. For the majority of collections we will create multiple delivery files, these include:

Purpose	Resolution	Long side
Thumbnail	72ppi	200ppi
Default Access	72ppi	600ppi
Large	72ppi	800ppi
Extra Large*	72ppi	1000ppi

\*If control of print versions of a collection is important, extra large images will not be provided online.

JPEG-2000 will also be investigated as a possible alternative for delivery images.

## Quality Control for Images

The basic goal of a quality control process is to insure that consistent, high quality images are produced. The DLC implements this process in the following steps:

1. Controlling the scanning environment, including hardware, software, and viewing conditions (see **Workspace Standards** and **Color Management** sections above).
2. Establishing clear production procedures to ensure that consistent digital objects are created.
3. Before beginning production, pilot test procedures and settings to verify that the digital images meet benchmark requirements.
4. Review output to insure consistent, quality results.

## Digitization Guidelines

Material Type	Bit-depth (Color Depth)	Pixel dimensions / Spatial resolution	Processing allowed	File Format
Printed Text only	8-bit grayscale	400 ppi		Uncompressed TIFF, IBM PC byte order
Artifactual text / manuscript	24-bit color possibly 8-bit grayscale in some situations pending investigation	400 ppi minimum		Uncompressed TIFF, IBM PC byte order
Negative / transparency - B&W	8-bit grayscale	4000-6000 on the long side	Sharpening, expansion of dynamic range	Uncompressed TIFF, IBM PC byte order
Negative / transparency - Color	24-bit color	4000-6000 on the long side	Sharpening, expansion of dynamic range	Uncompressed TIFF, IBM PC byte order
Photograph - B&W	8-bit grayscale or 24-bit color	4000-6000 on the long side	Sharpening, expansion of dynamic range	Uncompressed TIFF, IBM PC byte order
Photograph - Color	24-bit	4000-6000 on the long side	Sharpening, expansion of dynamic range	Uncompressed TIFF, IBM PC byte order

Glossary
----------

- **Archival Image:** (also referred to as the *Master image*). Images of high resolution that contain the greatest fidelity to the original. Typically the master image is a TIFF, from which JPEGs and GIFs are derived for on-screen viewing. Usually kept in a secure place, and stored off-line. While close to the original, the archival image still serves as a *surrogate* rather than an exact replica, as currently, even the best scanner technology loses some information.
- **Artifacts:** Visual modifications or effects (not contained in the original), introduced to an image during scanning. Artifacts may include: moiré, pixellation, regularly repeated patterns, and dotted or straight lines.
- **Bi-tonal image:** printed text or line art that is restricted to black and white values. Black dots have a value of "0" (absence of light); white dots have a value of "1" (presence of light). Bitonal images are 1-bit.
- **Bit depth:** potential number of bits used to define each pixel (the greater the bit depth, the greater the number of gray scale or color tones that can be represented). Common bit depths range from 1 bit to 24 bits. In an *eight-bit image*, each dot may have one of 256 gray or color values (calculated by raising 2 to the 8<sup>th</sup> power). In a *twenty-four bit image*, high tonal resolution color image, each dot may have one of 16.7 million color values (2 raised to the power of 24). Usually, the 24 bits are divided into 8 bit values, each representing red, green, or blue.
- **Born digital:** an image or document that was originally created or captured in digital form. For example: (1) an image of a three dimensional object in the physical world that is taken by a digital camera; (2) a document that was originally created electronically, like an electronic thesis or dissertation. Born digital images or documents do not go through developing or scanning processes.
- **Color management:** term that describes a technology that translates the colors of images, graphics, or text from a given color space, to the color space of an output device (printer or monitor).
- **Color space:** way of describing color that is independent of the device and materials used to reproduce it. A *color space* is a model for representing colors numerically by three or more coordinates (for example, the RGB color space uses red, green, and blue as coordinates). The coordinates are independent of the device(s) used in reproduction (printers or monitors), and allow for predictable reproduction across these devices.
- **Compression:** reduction of file-size for storage, processing, or transmission. Different compression techniques may affect the quality of the image. *Lossless* compression reduces the file size without loss of data when the image is compressed, or *decompressed*. The retrieved image will be nearly identical to the way it was before compression. *Lossy* compression reduces the file size but discards information in the process (for example, lossy compression occurs when creating a JPEG). The retrieved image will be slightly different than the original. Compression techniques may also be standard or proprietary. Standard techniques are generally preferable. Some compression techniques are intended for compressing pictures; others are for compressing text.
- **Continuous tone:** an image, such as a black and white photograph, that involves more tonal values than simple bi-tonal (black or white) reproductions.

- **CMYK:** a color space model used in printing. The letters stand for the colors: (C) cyan, (M)magenta, (Y)yellow, and (K)black.
- **Derivative image:** an image that is created from another image through an automated process. These images may be created by sampling to a lower resolution, using lossy compression techniques, or using image processing techniques. Usually information is lost when creating a derived image. Derivative images are typically used when fidelity to the original is not the primary concern (for example: for display on the Web).
- **Digital Object:** refers to the digital representation of an original document or three dimensional item. The original may be the hard copy of a photograph, painting, book, article, letter or non-graphic piece (such as a fossil, flower, plate, etc.).
- **Digitizing:** converting images into binary code by scanning or photographing.
- **Dots per inch (dpi):** term used to describe the resolution of an image. While sometimes used interchangeably with pixels per inch (ppi), the UT-DLC uses dpi to describe the output of a printer.
- **Dynamic range:** range between the lightest and darkest tones of an image. Although not always an indicator of the number of tones reproduced, a higher dynamic range allows for more potential shades.
- **File Size:** expressed in bytes and proportional to the resolution of an image (the higher the resolution, the bigger the file size). File size is calculated by multiplying the following document characteristics: (pixel dimensions x bit depth/ 8 (the number of bits in a byte). For example: a 24 bit image with pixel dimensions of 2,000 by 3,000, works out in this way:  $(2,000 \times 3,000 \times 24) = (144 \text{ million})/8 = 18 \text{ million bytes}$  (expressed as 18 MB). File size is conventionally expressed in KB, MB, GB, or TB.
- **File Formats:** both the bits that comprise the image and header information about a file. File formats vary in resolution, bit depth, color capabilities, and support for compression and metadata. Some common file formats: TIFF, JPEG and GIF.
- **GIF:** (Graphics Interchange Format). File format extensively used on the Internet because browsers don't need plug-ins or viewers. Gifs are best suited for simple graphics.
- **Grayscale image:** composed of pixels of multiple bits (2 to 8 or more bits), grayscale images (typically black and white photographs), are the next step up from bitonal images (strictly black or white values used to represent line art or printed text).
- **Halftone image:** image in which the size and density of dots convey an impression of gray shades (large and dense groupings convey dark gray or black; small and sparse groupings convey light gray or white).
- **Image Capture:** use of a scanner, digital camera, or other device to create a digital representation of an image. Current practice involves capturing an image at the highest resolution (TIFF), and storing it as an archival image on a CD-ROM. Derivative images can then be made through lossy compression or subsampling techniques.
- **Image Manipulation or Alteration:** Making modifications to an image using processing software, such as cropping, sharpening, or tonal adjustments.

- **JPEG:** (Joint Photographic Experts Group) compressed file format achieved by dividing the picture into tiny pixel blocks. Main format for derivative images used on the Web by digital libraries.
- **Long side:** refers to the longer side of a document to be digitized.
- **Optical Character Recognition (OCR):** computer method of converting scanned pages of text into an electronic document. OCR creates a file that can be opened in any word processor or searched by keyword.
- **Noise:** Marks picked up during image capture or data transfer that do not correspond to the original.
- **PhotoShop:** standard post-scanning software currently being utilized by most digital libraries to process images.
- **Pixels:** Elemental units that comprise an image. Each can represent a different shade or color whose range depends upon the storage allocation.
- **Pixel dimensions:** horizontal and vertical measurements of an image expressed in pixels (determined by multiplying the width and height by the dpi.) For example, an 8 x 10 object scanned at 500 dpi has the following pixel dimensions: horizontal (10 x 500 = 5,000); vertical (8 x 500 = 4,000).
- **Pixels per inch (ppi):** term the UT-DLC uses to describe the resolution of achieved by using a scanner or digital camera. Note: the digital library does not use ppi to describe the resolution of printer output (which is better expressed in dpi).
- **Quality Control:** ensuring that best practices are maintained through the stages of a process, and that documented steps and procedures are followed.
- **Resolution:** ability to distinguish fine spatial detail. Commonly expressed in dpi or ppi.
- **RGB (red, green, blue):** a color space that uses three coordinates (red, green, and blue) for the purpose of reproducing a graphic on a printer or monitor in a predictable way.
- **sRGB:** developed by the International Color Consortium, sRGB is a color space solution well suited to CRT monitors, television, scanners, digital cameras, and printing systems.
- **Resolution (spatial):** measure of the clarity of an image as expressed by the number of dots, or pixels that comprise it (high resolution=high number of pixels). High resolution images have greater clarity, detail, and file size. Image file resolution is sometimes expressed as a ratio (for example 400 x 800 pixels). Resolution can also apply to the output device (monitor display or printer), expressed in dpi or ppi.
- **Resolution:** measure of the tonal range of an image as expressed by the number of bits per pixel in a color or gray scale.
- **Sharpening:** Process of returning the sharpness of an image after capture. Sharpening is commonly done using Photoshop, Unsharp, Mask, or Plug-in software to correct for fuzziness introduced to an image during scanning.
- **Short side:** refers to the shorter side of a document to be digitized.
- **Silverfast:** Scanning software currently being used by the UT-DLC.

- **Spotting:** use of *band-aid* tool in PhotoShop to cover defects (such as dust marks) on a scanned image.
- **TIFF** (Tagged Image/Interchange File Format): standard file format for master images. TIFFS can be compressed or uncompressed. Group IV compression allows for a lossless compression of TIFFS that retrieves all the information when the image file is opened. TIFFs are typically used to create derivative images and for print-outs that retain fidelity to the original.