Scanning Best Practices

By Pat Vince and Michael D. Emery
December 2005 - Version 1

The Claremont Colleges Digital Library

Table of Content

Purpose of the Scanning Best Practices

The Claremont Colleges Digital Library's (CCDL) Scanning Best Practices seeks to provide fundamental guidelines to the seven Claremont Colleges for creating digital archival master and access files destined for dissemination in the CCDL. Although fundamental guidelines provided in this document are broad enough to apply to a majority of material, it is important to remember that each resource has its own characteristics, requiring a unique approach to scanning the material. Therefore we recommend that each collection be considered on a case by case basis.

Additionally, this document only addresses the more standard formats of material such as text and image. As we progress in growing the CCDL, we will add best practices for encoding audio and video.

Developing these best practices according to standards will:

1. Increase interoperability and accessibility across all collections created by the Claremont Colleges.
2. Increase long term preservation of the digital files.
3. Ensure image quality across all collections created by the Claremont Colleges.
4. Allow for multi-purposing scans down the road.

Copyright

U. S. copyright laws protect creators of original works of ownership and grants exclusive rights to such creators to display or perform work publicly, to reproduce the work and to distribute the work. The Library of Congress' Copyright Office Web site provides more detail information in their Circular 1 "Copyright Basics." http://www.loc.gov/copyright/circs

There are created original works that are not copyright protected. These works are in the **Public Domain** and may be freely used by everyone. Laura N. Gassaway, director of the Law Library & Professor of Law, University of North Carolina, Chapel Hill provides a helpful chart on when works pass into the public domain. http://www.unc.edu/~unclng/public-d.htm

When determining materials to digitize, begin with materials you know are not protected by copyright, or materials where you own the copyright or materials that are in the public domain. Remember to also clear the materials from any copy and/or distribution restrictions placed on the materials by the donor.

If the materials are not free of copyright, then detailed documentation must show clear intent to obtain permission to digitize and disseminate for educational purposes. The Library of Congress' Copyright Office Web site provides Circular 22 on "How to Investigate the Copyright Status of a Work." http://www.copyright.gov/circs/

The Library of Congress also provides links to Copyright Licensing Organizations & Publications Rights Clearinghouses at: http://lcweb.loc.gov/copyright/resces.html. The University of Texas prepared a helpful list of Collective Rights Organizations that you can contact for help at http://www.utsystem.edu/OGC/IntellectualProperty/permissn.htm#coll.

Special Situations - Music

Almost anything to do with music is protected by intellectual property rights in one form or another. Distribution of music must have the artist's permission and if requested by the artist a royalty payment may be required. The artists and/or their rights organizations may grant permission to distribute their music. "A Guide to Copyright for Music Librarians" is available at http://www.lib.jmu.edu/org/mla/.

Digital Imaging Basics: A Technical Primer

Introduction

The intent of the Digital Imaging Basics section of this document is to provide a clear starting point for the development an understanding of the many concepts related to the digitization of resources for online access. The hope is that this document will take these many interrelated concepts and provide clear, straight-forward, and contextualized definitions that are as concise as possible but still complete enough to allow actual understanding of the term and how it relates to digitization and digital libraries.

Binary Numbers

Binary code or numbering is the basic language of computers. A binary numeral system is based on two potential numbers, ones and zeros. Another way to think of it would be to think in terms of on and off, black and white, yes and no, or true and false. This differs significantly from the decimal system with ten possible numbers zero through nine. Where in a decimal system each column is ten times larger than the column to its right, in a binary system, each number is only twice as large.

Binary and Decimal Comparison

Binary			Decimal		
Example: 5 (in decimal)			Example: 472		
4s	2s	1s	100s	10s	1s
1	0	1	4	7	2
There is one 4s		= 4	There are four 100s		= 400
There are zero 2s		= 0	There are seven 10s		= 70
There is one 1		= 1	There are two 1s		= 2
Total		= 5	Total		= 472

As a result, binary numbers are long strings of ones and zeros, even when representing fairly small numbers. For example, the number twenty-five is fairly small, requiring a two in the ten's place and a five in the one's place, for a total of 20 + 5 = 25. However, as a binary number, it requires a one in the sixteen's place, a one in the eight's place, a zero in the four's and two's places, and another one in the one's place, for a total of 16 + 8 + 1 = 25.

Binary and Decimal Numbers

Binary Places						Decimal
16s	8s	4s	2s	1s		
				0	=	0
				1	=	1
			1	0	=	2
			1	1	=	3
		1	0	0	=	4
		1	0	1	=	5
1	1	0	0	1	=	25

Bits & Bytes

At their most basic level, computers work with bits of data, which is another way of saying computers work in binary code, or that they work in a binary number system. Briefly, binary numbers can be thought of in terms of a one or a zero, or maybe in terms of on and off, black and white, yes and no, or true and false. Each choice between a one and zero is a single bit of data. Strings of eight bits are then grouped together and called bytes. Each byte of data is therefore a series of eight ones or zeros. Since there are eight sets of two choices (or 2^8) there are 256 (0-255) possible combinations of ones and zeros in a single byte.

Binary Number	Binary Size	Decimal Number
0	1 bit	0
10	2 bits	2
11011011	8 bits or 1 byte	219
10101001 11010010	16 bits or 2 bytes	43,474

A single byte of data can represent a single character, since 256 possible combinations is enough room for the twenty-six letters of the alphabet, both lower and upper case, the numbers, and punctuation marks as well as letters with some basic diacritical marks. Obviously, a ten letter word would then require ten bytes of data, and beyond individual words, hundreds or thousands of bytes are required to save documents that contain hundreds or thousands of words.

Beyond text documents, images files are significantly larger because for an image file each pixel must be saved separately and might require multiple bytes of data to store color information. When files get larger, they are no longer measured in bytes, but rather in kilobytes, megabytes, gigabytes, or potentially even terabytes. Each one of these is 1,000 times larger than the previous step.

1 Kilobyte (KB)	=	1,000 bytes
1 Megabyte (MB)	=	1,000 KB
1 Gigabyte (GB)	=	1,000 MB
1 Terabyte (TB)	=	1,000 GB

A note on potential size confusion:

It is important to note however that there is a difference between whether the measurement is in binary or decimal. For example, measuring in binary numbers, a Kilobyte is 2^{10}, or 1,024 bytes. However measured in decimal numbers, this is rounded down to 10^3, or 1,000 bytes. While this is fairly insignificant when comparing binary bytes to decimal bytes, it becomes much more significant at each increase in magnitude.

File Size Comparison - Binary and Decimal

Binary			Decimal		
1 bit	=	8 bytes		=	
1 Kilobyte (KB)	=	1,024 bytes	1 Kilobyte (KB)	=	1,000 bytes
1 Megabyte (MB)	=	1,024 KB	1 Megabyte (MB)	=	1,000 KB
1 Gigabyte (GB)	=	1,024 MB	1 Gigabyte (GB)	=	1,000 MB
1 Terabyte (TB)	=	1,024 GB	1 Terabyte (TB)	=	1,000 GB

The problem is actually more complicated because the binary is also an abbreviation. Each level up compounds the problem.

File Size Comparison - Binary and Decimal

Size	Binary	Binary Abb.	Actual # of Bytes	Decimal	Decimal Abb.
Kilobyte (KB)	2^{10}	1,024 bytes	1,024	10^3	1,000 bytes
Megabyte (MB)	2^{20}	1,024 KB	1,048,576	10^6	1,000 KB
Gigabyte (GB)	2^{30}	1,024 MB	1,073,741,824	10^9	1,000 MB
Terabyte (TB)	2^{40}	1,024 GB	1,099,511,627,776	10^{12}	1,000 GB
Petabyte (PB)	2^{50}	1,024 TB	1,125,899,906,842,624	10^{15}	1,000 TB

For more information on how large certain types of data are to store, see both Bit Depth and File Size.

Bit Depth

Bit depth, also known as color depth and less frequently as pixel depth, is the measurement of how many colors, or shades of grey, that an image has. The number is based on how many bits of data are used to store color data for each pixel of an image. Obviously, as the number of bits used to store color information increases, the number of colors that can be presented also increases, but it does so exponentially.

For example, a one bit image would use a single bit of data for each pixel, creating an image that was made up of only two colors, usually black and white. An

eight-bit (or one byte) image would have 256 (2^8) possible colors, which might be 256 shades of gray in a grayscale image or a limited 256 color palate in an image saved in the GIF file format.

Although not used with images, 16-bit color depth would indicate 65,536 possible colors. But again, images are not generally saved as 16-bit images, instead, where you are likely to find 16-bit color depth is in some older monitors might only be able to produce 16-bits worth of color data.

A 24-bit (or three byte) would have a total of 16,777,216 possible colors. Most 24-bit images are made up of three eight-bit channels as part of the RGB color model. RGB images are made up of 256 shades of red, 256 shades of green, and 256 shades of blue for a total of 16,777,216 color combinations (256 x 256 x256 or 2^{24}). When the bit depth is 24-bit or higher, it is also known as Truecolor (called Millions on a Macintosh) because it represents a significant portion of the range of colors visible to the human eye.

Comparison of 1-bit, 8-bit, and 24-bit images

"Aerostation out at Elbows or the Itinerant Aeronaut," from the Aviation Collection - Special Collections at the Libraries of The Claremont Colleges

Beyond 24-bit images, there are also 32-bit and 48-bit images. The difference between the two is considerable. An RGB 32-bit image usually has a fourth channel used as an alpha channel to aid with transparency for the additional 8-bits of data. In other words, there are four 8-bit channels, a red, green, blue, and an additional alpha channel. However, CMYK images are 32-bit images because they have four channels of color.

Generally the highest bit-depth available is a 48-bit image, which is usually used with RGB images and has 16-bits of data reserved for each of the three channels. This means there are 16-bits of data, or 65,546 possible shades each, of red, green, and blue. While most software and hardware is not able to display this much data, there is software that can make use of this data, particularly when manipulating the images via

a histogram or curves. Although the additional color data is not displayable on screen, images that have the additional information can be manipulated with much less degradation of the image because of the additional information.

A note on potential naming confusion:

While usually bit depth refers to the overall depth of color for the image, there are times when instead, the bit depth is used to refer to the bit depth of each individual channel. In other words, a 24-bit image might be referred to as an 8-bit image, because it has 8-bits of data for each of the three channels: red, green, and blue. And, sometimes a 48-bit image might be referred to as a 16-bit image because there are 16-bits of data for each of the three channels: red, green, and blue. Photoshop, in particular, does this.

Color Depth - see Bit Depth

Color Model

A color model is a way of representing color by dividing it into multiple components. There are usually three or four components to a color model, and they are often associated with a particular color range. These color models are then mapped to certain reference colors creating a new color space. This color space can then be used to create a range of specific colors. Common color models include, RGB, CMYK, LAB, and HSB. However, not all color spaces using the same color model are the same. For example, the Adobe RGB color space and the sRGB color space are slightly different, although both are based on the same RGB color model.

Color Models

CMYK

The CMYK color model is divided into four channels of cyan, magenta, yellow, and black. CMYK is the basis of most full color printing and works on a subtractive color basis. Because the combination of cyan, magenta, and yellow do not actually create a true black, black is added to the mix for a range of darker colors.

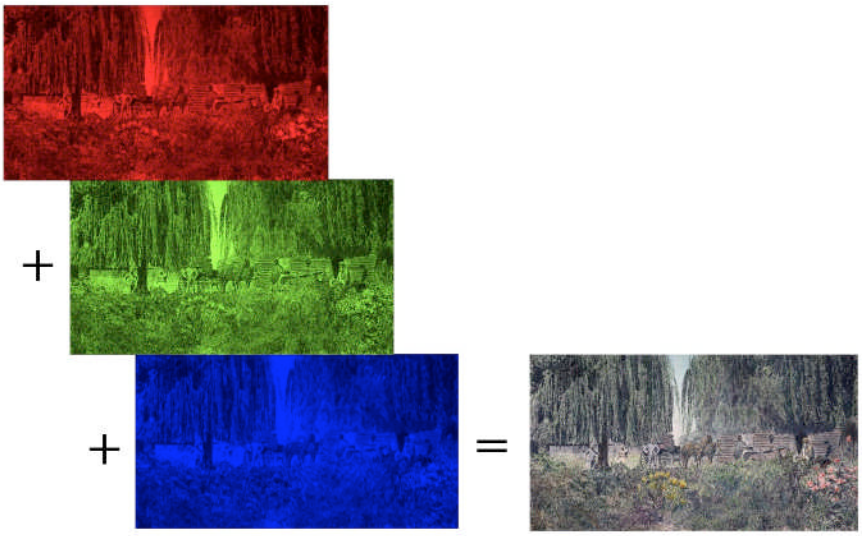
LAB

The LAB color model is made up of three components which are not strictly limited to color channels. The L, or lightness, channel controls how bright or dark a pixel is. The A channel controls the color by shifting between green and red colors, and the B channel which controls the color by shifting between blue and yellow colors.

RGB

The RGB color model is divided into red, green, and blue channels and is an additive color model. RGB stores separate values for each of the three colors to create the range of possible colors in the color model. Two of the more common color spaces in the RGB model are sRGB and Adobe RGB.

RGB Color Model



"Claremont in 1884," photo of El Alisal from the Wheeler Scrapbooks, Book 2, Page 100, Item 1 - Special Collections at the Libraries of The Claremont Colleges.

El Alisal was named by owner H. A. Palmer to the land between 8th and 10th Street on the north and south and Yale Avenue and Indian Hill Boulevard in the east and west.

HSB

The HSB color model is divided into hue, saturation, and brightness and, like the LAB color model, is another way of viewing colors that functions differently than combining multiple color channels. The H channel controls the hue, or color, of the pixel. The S, or saturation, channel controls the tone or purity of the color by shifting from pure color to gray (where the colors are equal). The B, or brightness, channel controls the shade from full color to black.

Variations on HSB

There are two variations on the HSP color model, the HSV and the HSL. For each of these, the hue and saturation channels work the same way. In the HSV color model V, or value, replaces brightness is identical to the HSB color model. However, the HSL color model is a slight variation where the L, or

lightness, channel which controls lightness replaces the B, or brightness, channel. The difference between them is that in the brightness/value model, the brightness or value of a pure color is equal to white, but in the HSL color model, the lightness of a pure color is the equivalent of a medium gray.

Additive vs. Subtractive Color

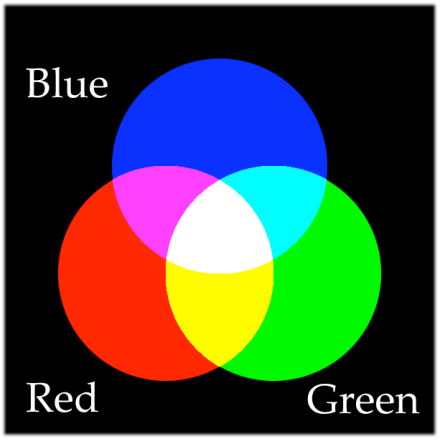
Additive Color

The concept of additive color is based on the idea of color being created by direct light, or in other words, the color is derived from the color of the light itself. The absence of light would be black and pure light would be white, so for a particular color to exist there must be particular color of light creating the color. For example, in the RGB color model, red would require (the addition of) red light and yellow would require a combination (or addition) of red and green.

Subtractive Color

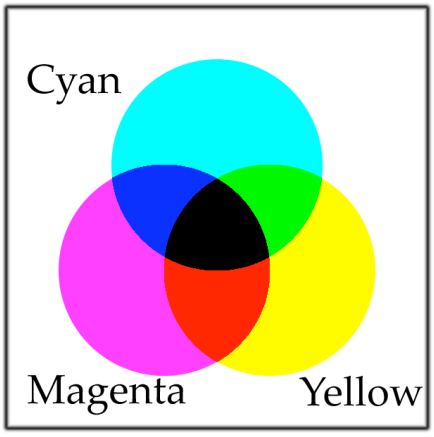
The idea of subtractive color is based on the reflection of light where the unwanted colors are absorbed and the desired colors are reflected. Black is then when all the light is absorbed (or subtracted) and white when no light is absorbed. To create a red color in a subtractive color environment, the cyan must be subtracted so that magenta and yellow are reflected creating red.

Additive Color



An additive color model starts with black, or no light, and as each new color is added, the light gets closer to white.

Subtractive Color



In a subtractive color model, each new color is absorbing more of the color spectrum and reflecting less and less color, until finally, it reflects no color and is black.

Additional Color Terms

Hue

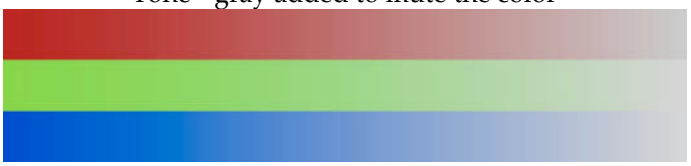
A color's hue is its basic color. In other words, it is the description of the basic color whether it is orange or blue or yellow or any other color.

Hue - the range of color

Tone

A color's tone is its hue plus either gray or the opposite color to mute or tone done the color.

Tone - gray added to mute the color



Tint

A color's tint is its hue plus white which lightens the color.

Tint - white added to lighten the color

Shade

A color's shade is its hue plus black to darken the color.

Shade - black added to darken the color

Color Space - see Color Model

Compression

Applying compression to a digital image reduces the size of a file by either abbreviating the data or throwing away data that can, in theory, be recreated later. The advantage of compression is that it decreases the amount of storage space needed to store files, and, when serving files over the Internet, this smaller file size allows for faster downloads for the end-user. The disadvantage comes with compression systems that throw data away because if the data thrown away cannot be recreated accurately, there is a potential for loss of image quality. There are two types of compression: lossy compression and lossless compression.

Lossy Compression

Lossy compression works by throwing away, or losing, data. This data is permanently lost, and although it is recreated on the fly when the file is viewed, it is important to realize that this recreation of data is not identical to the original information which is permanently gone.

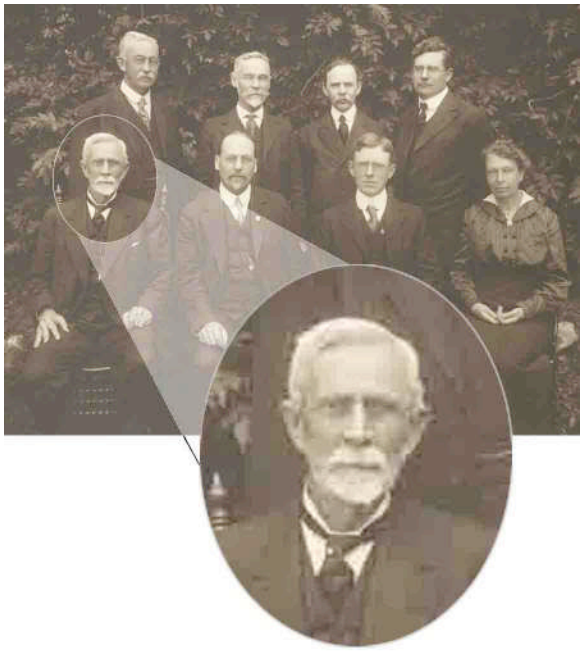
Ideally, the data that is thrown away and then recreated does not negatively impact the image, however, because there are varying levels of compression, this is not always the case. The most common lossy compression format is the JPEG format which allows a range of compression options from high to medium to low with each greater level of compression resulting in a increasingly negative impact on the image quality.

However, when a limited lossy compression save is performed, the image usually looks remarkably similar to the human eye, even if there are numerous changes on a pixel by pixel level. On the other hand, in even a medium compressed lossy image there might be a notable degradation to the human eye.

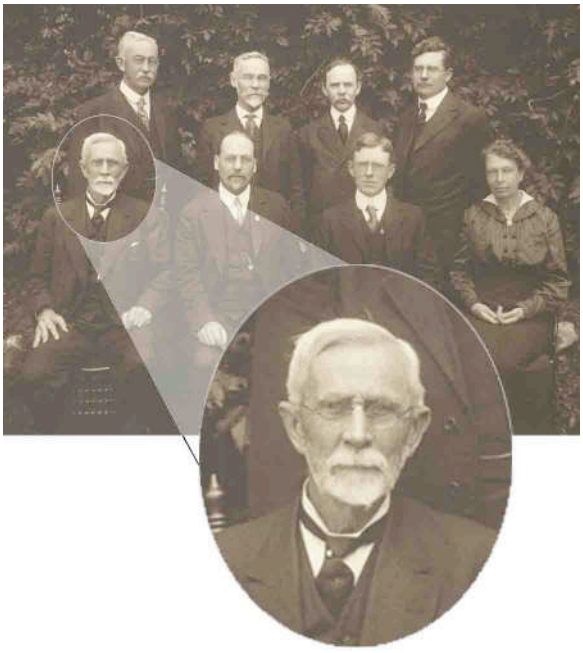
It is also important to note that while repeatedly viewing a file that is saved with lossy compression does not further degrade the image, resaving the image numerous times will slowly continue to degrade the image quality even when only minimally compressed. Of course, the rate of this degradation will depend on the level of compression as well as the specific image being compressed.

High Compression and Low Compression JPEG Comparison

High Compression JPEG



Low Compression JPEG



"The 'Old Guard' - May 24, 1917," showing original faculty of Pomona College from the Wheeler Scrapbooks, Book 1, Page 5, Item 2 - Special Collections at the Libraries of The Claremont Colleges.

Back row: Prof. D. H. Colcord, Prof. A. D. Bissell, Prof. G. G. Hitchcock, and Dr. G. S. Sumner.
Front row: Rev. C. B. Sumner (in focus), Dean E. C. Norton, Prof. F. P. Brackett, and Dr. P. E. Spalding.

Lossless Compression

Lossless compression works by removing repeated information in the binary code of the file. As a result, lossless compression has different levels of success with different images. Images that contain large areas of identical information or areas that have repeated information tend to compress well.

Images that have been saved with lossless compression look identical to the original to the human eye because they are identical on a pixel by pixel basis, having been recreated with no loss of data. Because of this, unlike lossy compression formats, files that are saved with lossless compression can be resaved indefinitely with no loss of quality because all the data is retained.

Digital Image

Digital images are electronically encoded images that can be created from a wide range of sources. A digital image can be a scanned electronic version of a photograph or a born digital photograph. A digital image can be a scanned or digitally captured image of a manuscript, painting, map, or any physical thing that can be photographed. A digital image can be an image created on a computer using drawing or related software.

There are two general ways of saving digital images. The first, and far more common, is as raster image and the second is as a vector image.

Raster or Bitmap Images

A raster, or bitmapped, image is made up of a rectangular grid of individual pixels with each pixel assigned a particular color. The range of colors that an individual pixel can display depends on the bit depth of the image. The larger the number of colors, the larger the file size as each pixel must store a greater and greater number of bits of data to represent those colors.

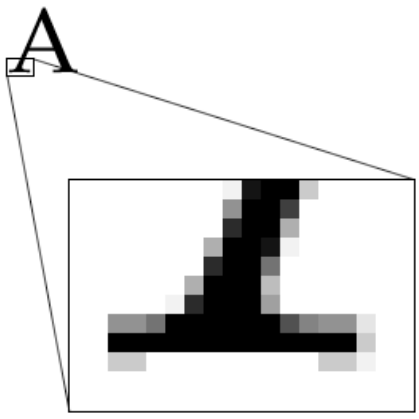
Vector Images

Vector images vary significantly from raster images in that vector images are derived from mathematical formulas. Rather than a line being made up of a particular number of pixels of a particular size that have each been assigned a particular color as it is in a bitmapped image, a line in a vector file is derived from a mathematical formula which is drawn on the fly.

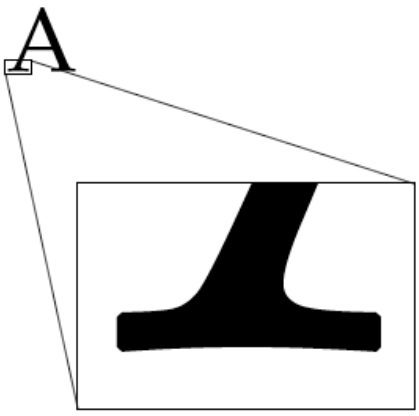
Vector graphics are generally smaller and are far more scalable than bitmapped images. However, because they are made up of mathematical formulas, they cannot represent complex images like photographs. Vector graphics are great for line drawings, text, and cartoon-like images.

Raster and Vector Image Comparison

Raster Image of the Letter A



Vector Image of the Letter A



Dynamic Range

Dynamic range is the measurement of the range of optical densities that a scanner can capture from the lightest to the darkest areas of an image from zero to four. While bit depth determines a theoretical range of data that can be saved, the actual

measure of what can be captured by a scanner is the dynamic range. Dynamic range is the difference between the dMin and the dMax of an image. The dMin is the measurement of how close to white at 0.0 a scanner can capture and the dMax is the measurement of how close to pure black which is a little over 4.0.

File Formats

Digital images can be saved in several different file formats. There are a range of issues to consider with file formats including the type of image, and the intended usage of the image, and whether the file format is an open standard or proprietary format.

For example, the uncompressed TIFF format lends itself well to long term archival storage, but these files are far too large to be used online. On the other hand, a the JPEG format is great for compressing images for online display, but its lossy compression method makes it a poor choice for long term archival storage of the image. Because of different intended usages, it is not uncommon for an image to exist in multiple instances in different file formats.

For example, a large map might be scanned and saved as a TIFF file for archival storage. A JPEG2000 version of the map might be saved for a display version, allowing users to zoom and pan through the image. Finally, smaller JPEGs might be created as a preview and thumbnail images. The particular needs of the image and its intended usage will play a significant role in deciding which file formats are appropriate.

GIF

The Graphics Interchange Format (GIF) can display up to 256 colors that are compressed via LZW lossless compression. Although there is significant flexibility in which 256 colors are chosen, the 256 color limit on GIFs makes them generally unsuitable for photographs except possibly as thumbnails.

JPEG

The term JPEG, or Joint Photographic Experts Group, actually only indicates how a picture is converted into a stream of bytes. What we tend to think of as JPEGs are really JFIF (JPEG File Interchange Format) files. JPEG files have a bit depth of 24, allowing for over 16 million colors, and they are saved via a lossy compression mechanism. JPEG files can be saved with a minimum compression mechanism, or they can be saved at very high levels of compression resulting in images with very low file sizes but considerable degradation of the image.

JPEG2000

JPEG 2000 is a newer version of the JPEG format that compresses images via wavelets and provides for zooming and panning of images. It currently

requires either a server to be set up to provide access to the images or a user side plug-in that must be installed on the end-users machine.

PNG

The PNG, or Portable Network Graphic, is a file format developed by the World Wide Web Consortium as a patent-free replacement for the GIF. Although not fully supported by current internet browsers (specifically Internet Explorer), the PNG format offers alpha channel transparency (providing up the 254 levels of partial transparency) versus the binary transparency offered in the GIF format.

TIFF

Tagged Image/Interchange File Format (the current version is the TIFF ITU-T.6) is the standard file format for master file formats (both archive and display). It can handle 24-bit images, and although there are compression options available for use with the TIFF files master images should be saved only as uncompressed TIFFs.

File Sizes

Images usually have a very large file size. The reason for this is that because most images are saved as raster files with each individual pixel needing to be saved separately. Additionally, not only is every pixel of the image saved separately, but the higher the bit depth, or the more colors, the more data is required to save each individual pixel.

The file size of an image file can be determined by multiplying its height and width in pixels time the images bit depth. The height and width can each be multiplied by the pixels per inch, or the height and width can be multiplied by each other and then by the pixels per inch squared. Regardless, this result is then multiplied by the bit depth of the image. It is probably easier to convert the bit depth to the total number of bytes (8 bits to a byte) first.

Calculating File Size

Option 1:

(height x ppi) x (width x ppi) x bit depth (in bytes) = file size in bytes

Option 2:

height x width x ppi^2 x bit depth (in bytes) = file size in bytes

For example, using the first formula, the file size of a five by seven photo saved at 300 dpi, with a bit depth of 24 could be calculated by:

Height	Width	Bit Depth	File Size
(5 in. x 300 ppi) x	(7 in. x 300 ppi) x	24 bits =	
(1500 pixels) x	(2100 pixels) x	3 bytes =	9,450,000 Bytes, 9,450 Kilobytes, or 9.45 MB

How big is a byte or a megabyte or a gigabyte? The following chart gives approximate equivalents to various file size.

Approximate Equivalents to Various File Sizes

Size	Approximate Equivalent
Bits	
1 bit	A one or zero
Bytes	
1 byte	A single character
10 bytes	A ten letter word
Kilobytes	
4 KBs	A single page of text
50-100 KBs	A low resolution compressed image (JPEG)
Megabytes	
1 MB	A small novel
2-3 MBs	A high resolution compressed image (JPEG)
3-5 MBs	An MP3 file (compressed music file)
10 MBs	A high resolution uncompressed 5x7 image at 300 dpi (TIFF)
90 MBs	A high resolution uncompressed 8x10 image at 600 dpi (TIFF)
650 MBs	The contents of a CD
Gigabytes	
1 GB	A symphony in high quality sound
1 GB	Approximately 2 CDs worth of uncompressed music
17 GBs	The contents of a DVD
50 GBs	A floor of books
Terabytes	
1 TB	2,000 audio CDs in original uncompressed format
1 TB	1 million books
1 TB	160 DVD movies
20 TBs	The printed collection of the US Library of Congress
Petabytes	
1 PB	2 million CDs worth of uncompressed music
1 PB	160,000 DVD movies
10 PBs	All US academic research libraries

Halftones

A halftone is created when a grayscale image is printed in black and white in such a way to still look like a grayscale image. The process is accomplished by creating dots of varying sizes to create areas of lighter and darker color. Halftones are made up of different patterns and often are described as having a certain number of lines per inch or a specific line screen.

Line Screen

When an image is printed, it is usually divided into a preset number of lines depending on the format. These lines are made up of dots of varying sizes to create the desired colors. For example, a large black area would have large black dots, versus a white area which might lack any dots at all.

Pixels

Pixels, or picture elements, are the tiny dots that make up a picture on a computer monitor in a way that grains make up a photograph or dots of varying sizes make an image in a newspaper half-tone. Each pixel or dot and represent any number of colors which is a measurement of bit depth.

Pixels per Inch versus Dots per Inch

The measurement of pixels per inch is a count of the number of pixel elements per inch and is usually used in reference to on screen viewing. The measurement of dots per inch is virtually identical except that it is usually used in reference to printing. Both of these terms are slightly misleading because of the way that images are created both on screen and via printing.

When viewing an image on screen, the images pixels per inch is irrelevant because monitors have fixed resolutions. Whether an image has 300 pixels per inch or 25 pixels per inch, a 100 pixel image will be the same size on screen and take up 100 pixels.

By contrast, the dots per inch of an image will control the size of an image that is printed. However, additional factors such as the line screen the image is printed in will also impact the final printed image.

Best Practices for Scanning

Introduction

The intent of these best practices for scanning recommendations is to provide a starting point, or a baseline, for the minimum level of quality needed for scanning materials. Beyond providing this baseline, the goal is to provide an explanatory framework for why certain decisions should be made and, beyond this, to provide the user with the ability to make informed decisions about when the needs of the collection or particular item surpass the baseline parameters suggested in this document.

General Guidelines

While there are many specific guidelines to digitizing resources, there are a few general guidelines which help steer more specific guidelines and digitization work in general.

1. Scan items once at the highest level possible (and appropriate)
2. Create and save an archival master of all files from which all derivatives are derived
3. For extremely degraded images, create a display master to be offered only in addition, as an alternative, but treated as an archival master in all other respects
4. Avoid proprietary formats and use standards based file formats
5. Gather appropriate and as complete as possible metadata regarding all digital images
6. Back up all archival and display masters
7. Monitor backup copies and recopy, migrate, and/or transfer data as necessary depending on data degradation, file format/software obsolescence, and hardware obsolescence

Specific Guidelines for Digitization

Bit Depth

When considering the bit depth that an object should be scanned in, the three basic options are black and white (1-bit), grayscale (8-bit), and full color (24-bit). It is recommended that all objects be scanned in full, 24-bit color. Although many objects such as text and black and white photographs initially appear to be candidates for grayscale scanning, most have enough color information to benefit from full color scanning.

For example, while a black and white photograph could be scanned in grayscale and might look fine, older images often have some coloration due to the aging process. A full color scan would help to accurately represent the current condition of the photograph. In particular, manuscript pages often provide significant contextual information beyond the content of the words alone. Because of this, full color images often provide additional contextual information about the item being scanned and help limit the potential need to rescan objects.

Spatial Resolution

While there is no one perfect size for all images, the goal should be to scan images as close as possible to 600 pixels per inch at original size. When this is not possible, such as when scanning oversized objects like maps, the items should be scanned as close to 600 pixels per inch at original size as possible with a minimum resolution of 300 pixels per inch. Beyond this, certain images such as high resolution maps will potentially require higher resolution than this. These larger images often need to be dealt with on a case by case basis depending on the exact size of the object in question.

Additionally, transparent media such as negatives and slides are often intended to be enlarged and should be scanned accordingly. Smaller 35 mm slides and negatives should be scanned at 3000 pixels along the longest edge at 600 pixels per inch. Negatives and slides that are larger than this should be scanned at 6000 pixels along the longest edge.

File Formats

Archival Masters

Every image created should have an extremely high-quality archival master created and saved in an uncompressed TIFF format. These uncompressed archival masters should be saved unedited as a preservation master from which all derivative files are created. The archival master should be as high a quality image as possible to limit the potential need for rescanning later.

Display Masters

When necessary due to the extreme degradation of the original archival master, there might be times it is appropriate to create a modified display master that corrects some of the damage to the original. This display master should be offered only in conjunction with the original and always be properly identified as a modified version of the original. This display master should be treated as a second instance of the object and be saved accordingly as an uncompressed TIFF format.

Overview of Best Practices by Type of Material

Type of Object	Attributes	Archival Master
Text	File Format	TIFF
	Bit Depth	24-bit color
	Spatial Resolution	600 dpi
	Spatial Dimensions	100% of original
Photograph	File Format	TIFF
	Bit Depth	24-bit color
	Spatial Resolution	600 dpi
	Spatial Dimensions	100% of original
Negatives - Small (35 mm neg. & slides)	File Format	TIFF
	Bit Depth	24-bit color
	Spatial Resolution	3000 pixels along longest edge (@ 600 dpi)
	Spatial Dimensions	varies
Negatives - Large (> 35 mm film)	File Format	TIFF
	Bit Depth	24-bit color
	Spatial Resolution	6000 pixels along longest edge (@ 600 dpi)
	Spatial Dimensions	varies
Maps & oversized objects	File Format	TIFF
	Bit Depth	24-bit color
	Spatial Resolution	600 dpi (in pieces if necessary)
	Spatial Dimensions	100% of original
	Note	Should capture one full image in as high a resolution as possible (300 dpi minimum)

Glossary

Archival Master

The archival master is the master file that functions as the preservation master and from which all derivatives are created. Archival masters should be unedited and saved as uncompressed TIFFs.

Artifacts

Artifacts are any unwanted digital information that is not part of the original image. Artifacts might be introduced during the scanning process, and compression artifacts are the result of a compression scheme that recreates faulty data that varies significantly enough from the original to be noticeable.

Binary Numbers

Binary numeral systems are based on two potential numbers, ones and zeros. Because each place is represented by one of two numbers, each increase in place in a binary system is twice as large the previous. This is quite different from the ten times increase in a decimal numbering system.

Bits and Bytes

Computers work with bits of data. Each bit of data in a single zero or one. Bits are collected together into groups of eight which are called bytes. Each byte of data, or strong of eight ones and zeros can represent a single character.

Bit Depth

Bit depth is the measurement of how many bits of data are used to store color information for each pixel of an image.

Color Depth

See Bit Depth

Color Model

Color models are a way of representing color by dividing it into multiple components. When a color model is mapped to specific colors, it forms a color space which can be used to create a range of colors. Common color models are RGB, CMYK, HSB, and LAB.

Compression

Image files can be compressed to allow for easier online access. There are two primary types of compression, lossy which deletes data to be recreated later and lossless which removes redundant or repeated information with no loss of data.

Digital Image

A digital image is an electronically encoded image and can be anything that can be scanned or photographed.

Display Master

The display master is a companion to an archival master for those images that are extremely degraded to the point of being almost unusable in their original state. A display master is created to visually correct the image but used only in conjunction with the original archival master and always labeled accordingly.

Dynamic Range

Dynamic range is the measurement of the range of optical densities that a scanner can capture from the lightest to the darkest areas of an image.

Hexadecimal Numbers

Like binary and decimal numbering systems, hexadecimal code is another numbering system. While binary is based on two and decimal on ten, hexadecimal numbering is based on sixteen. Each place is represented by 0-F, with "A" functioning as the decimal equivalent of ten, "B" functioning as the decimal equivalent of eleven, and so on. Thus the number 7F is the equivalent of (7 X 16 = 112) + (F = 15) = 127.

Histogram

A histogram of a digital image shows the distribution of pixels by creating a graph showing the count of pixels of certain colors. Images saved in the RGB color model usually offer a full color histogram as well as a histogram of each color channel.

Moiré Patterns:

Moiré patterns are the result of scanning halftone images where the halftone pattern does not match up with the raster pattern created during scanning.

Noise

Noise is the addition of artifacts into an image as a part of scanning or digital photographing process. Usually it is the result of low light and present in the darker areas of a digital image. It usually appears somewhat similar to the graininess seen in photographic film. Noise is most often found when an image has its contrast adjusted.

Interpolation

Interpolation is the process by which software digitally creates pixels based on the existing pixels. There are different levels of interpolation, and some programs are better than others, but in the end, all interpolation creates pixels that are not actually there and should be avoided. Interpolation contributes heavily to fuzzy lines in images.

Halftones

A halftone is created when a grayscale image is printed in black and white in such a way to still look like a grayscale image.

Line Screen

Line screen is the number of lines of varying sizes of dots that create images that are printed.

Pixels

Pixels, or picture elements, are the dots that make up an image on screen.

Spatial Resolution

The spatial resolution of an image is the measurement of the number of pixels per inch. A low resolution image would have a resolution of under 100 pixels per inch, and a high resolution image would have a resolution of 600 pixels per inch or even higher.