Digitization Basics

10:45 – 11:45 am
August 18, 2016

John Sarnowski
Director
The ResCarta Foundation

Creating digital images of analog objects seems almost trivial in this age of cell phone cameras. Those tasked with creating digital analogs of culturally important materials will need to know how to produce digital objects that can stand the test of time. This introductory session will cover the process of scanning and digital photography from the librarian, historian or archivist's prospective. Attendees will come away with a working knowledge of file formats, resolution, bit depths and the considerations of selecting materials for digitization.

Contents

Digitization Programs

When selecting materials for digitization, consider your audience. Is it K12 students, K12 teachers, genealogist, or research scientists? Copyright is not your most pressing issue.

Organize and research the physical collection before digitizing. What have you got? How can you capture it?

Tackle a smaller group of materials at first. Take time to make mistakes and learn what the best work flow might be. Restart the process and tweak it before you tackle the “boxes in the basement”.

Digitizing

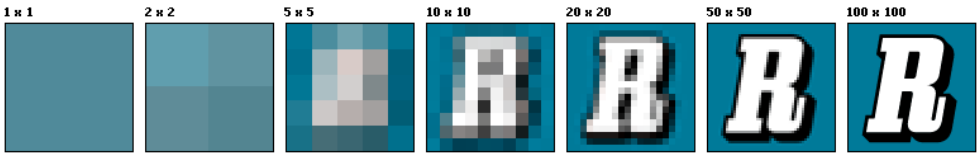
The term digitizing is often used when text, images, or sounds are converted into a single binary code (Zeros and Ones). Converting analog continuously variable objects into digital (on/off) bits and placing those bits into a file on a computer system.

Bits/Bytes

Bit is short for 'binary digit.' It's a single digit in a binary number, and it can be either 1 or 0.
A byte is 8 bits. That's the definition. With 8 bits you can store any number between 0 and 255, since there are 256 different combinations of 1 and 0 to choose from.

Resolution

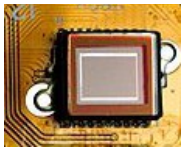
Resolution is a measurement of the output quality of a digital image or sound. Common units to measure resolution include: PPI (pixels per inch), DPI (dots per inch). Resolution quality is the ability to “resolve” details in the digital object that exist in the original analog object.



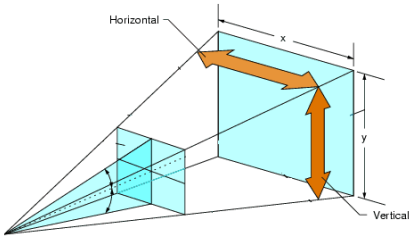
https://en.wikipedia.org/wiki/Image_resolution

Linear and Area Arrays

Scanners normally have a linear arrays meaning they have a single line of sensors that they move over an object using motors. This gives a **fixed** number of sensors in the horizontal to create digital values and a variable length that the motor can drive the array.



Cameras normally have area arrays meaning they have a rectangular area filled with sensors which have a fixed number of senors in the horizontal and a fixed number of vertical lines. The camera lens can apply this array over a **variable** portion of its view.

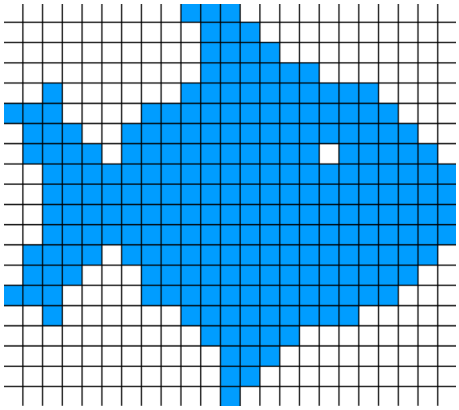


Bits/Bytes and colors

Sensors in scanners and cameras output values based on the light striking them. Storing the resulting values in binary numbers gives us the ability to capture color. The number of colors captured results in the number of bits stored per pixel.

Bit Depths bit (on/off) byte (8bits 01010101)
1bit = Black/White **Bi-tone** (2^1)
2bit = 4 colors/values (2^2)
4bit = 16 colors/values (2^4)
8bit = 256 colors **Gray Scale** (2^8)
24bit = Three 8bit chunks 256 RED, 256 Green, 256 Blue **TruColor** 16,777,216 Colors (2^{24})
Human eye differentiation limitation approx. 10 million colors
48bit = Three 16bit chunks 65536 RED, 65536 Green, 65536 Blue 2.8 trillion colors (2^{48})

Storing the results of the light sensor in an array gives us data that can be stored in a memory or saved to a file. This is usually called a raster image. Each pixel stored is stored as an element of the array in the bit depth specified. The raster image size can be determined by the number of pixels in the array row times the number of rows times the number of bits used to store the color. So for our fishy friend here we have an array of 24pixels by 20 lines or 480 total pixels.



Stored as Bi-tone	24×20×1 =	480 bits or	60 bytes
Stored as Gray Scale	24×20×8 =	3,840 bits or	480 bytes
Stored as TruColor	24×20×24 =	11,520 bits or	1,440 bytes
Stored as 48bit Color	24×20×48 =	23,040 bits or	2,880 bytes

Doing this with an 8.5×11 inch page scanned at 300 dots per inch

Stored as Bi-tone	2550×3300×1 =	1,051,875 bytes
Stored as Gray Scale	2550×3300×8 =	8,415,000 bytes
Stored as TruColor	2550×3300×24 =	25,245,000 bytes
Stored as 48bit Color	2550×3300×48 =	50,490,000 bytes

File Formats

To store and reuse digitized information, the bits/bytes of digital data are written out to files held by some form of storage media. In order to decipher the content of the digital data , additional technical information must be stored along with the raw digital data. This is done by storing the technical information and digital data in a known file format. Common file formats for image data are TIFF (Tagged Image File Format), JPEG (Joint Photographic Experts Group), PNG (Portable Network Graphics), and PDF (Portable Document Format). Common audio formats are WAV (Waveform Audio File Format), BWF(Broadcast Wav Format) and MPEG (Moving Picture Experts Group)

Each of the formats listed above have a variety of sub formats. Each supports various compression schemes and extensions making it difficult to predict useability by various hardware/software decoders.

Storage and Compression

As we noted digital files produced by scanners and cameras can contain large amounts of data. A single newspaper page scanned in trucolor would be 11in x 400dpi x 23in x 400dpi x 24bits /8 or 121,440,000 byes of raw data or about 122 megabytes per page. With current storage capacity a Terabyte drive could hold only 8000 pages.

We can manage storage concerns in two ways. The first is to reduce the bit depth and the second is to compress the images data.

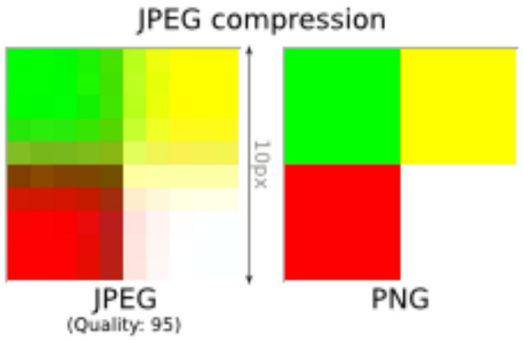
If we scan our newspapers using a bi-tone or single bit we can then store 192,000 uncompressed pages on a Terabyte disk.

Compression comes in two main flavors, loss-less and lossy. Loss-less compression systems can store and retrieve every pixel in the original raster. Lossy compression systems make averages of areas in the raster and store average values, they cannot restore every pixel of the original raster.

Using loss-less compression like Group 4 on a newspaper page scanned in bi-tone can reduce the storage requirements by an average ratio of 20:1. So we could fit nearly four million pages of compressed bi-tone newspapers images on a Terabyte disk. Group 4 compression is always loss-less.

There are also loss-less compression systems for 8bit gray and 24bit color. These include JPEG 2000 (JP2), PNG (Portable Network Graphics), Flate(ZIP), and Lempel–Ziv–Welch (LZW). The reduction in size varies with the content and in some cases a compressed gray/color file may actually be larger than the original uncompressed file. In many cases the defaults for JPEG 2000 compression is for lossy compression by default.

Lossy compression systems like JPEG, JPEG 2000, and Fractal cannot uncompress back to the original raster. JPEG averages a block of pixels at a time so most edges (like text) get averaged and blurred. Most lossy compression loss is invisible to the human eye since the blocks of pixels used are small (16 x16).



For culturally important materials it is best to save your digital data using uncompressed or loss-less compression formats. Uncompressed TIFF is best for long term storage and reuse in an archive.

Uncompressed map 28 x24in @300dpi 24bit color 179,696,012 bytes

JPEG compressed 28 x24in @300dpi 24bit color 2,084,027 bytes

Scanners vs Cameras

Scanners

Scanners come in various configurations like flatbed, rotary, and slide/film. The advantage to scanners is primarily in their self contained lighting systems. The light source (fluorescent, led etc.) is part of the system and the linear CCDs are matched to the temperature of the lamps. The linear CCDs tend to have a higher number of sensors than cameras since linear arrays (a single line of sensors) is easier to manufacture.

Most scanners come with software that allows an operator to select from a range of resolutions, color depths and output file formats. This is more flexible and can allow for a more controlled work flow.

Scanners can deliver consistent resolution outputs like 300 dpi, 400 dpi, 600 dpi, etc. and fixed bed sizes like 8.5 in x 14 in, 18 in x 24 in and so on.

Cameras

Cameras come in various configurations as well; with removable and non removable lenses, adjustable and fixed aperture sizes, variable or fixed speeds, variable or fixed focus, and more. Due the variety of features and settings it can be a more complex tool for digitization.

Cameras also come with a fixed area array, with a fixed aspect ratio. Camera arrays are often measured in megapixels (width x height). Lighting is necessary when using cameras and controlling the quality and angle of the lighting is important to the resulting digitization.

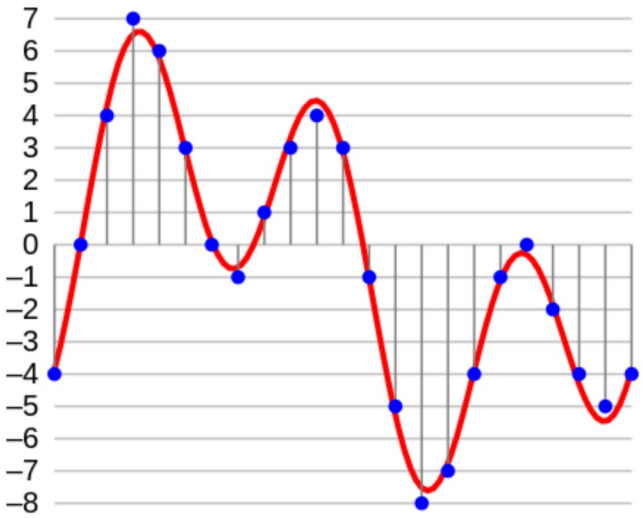
Cameras	Scanners	
2048×1536 = 3 Megapixel		
2560×1920 = 5 Megapixel		
3264×2448 = 8 Megapixel (iPhone 6)	11×8.5 @ 300 dpi 3300×2550	8 MP
4256×2832 = 12 Megapixel		
6016×4000 = 24 Megapixel (Nikon 3200)	11×8.5 @ 600 dpi 6600×5100	33 MP
6708×8956 = 60 Megapixel (Hasselblad $44k)	11×8.5 @1200 dpi 13200×10200	134 MP

Sample scan sizes

Photograph	5 in x 7 in at 400 dpi = 2000×2800	
	5 in x 7 in at 600 dpi = 3000×4200	Suggested best practice (4000 pixel on long edge)
	8 in x 10 at 400 dpi = 3200x4000	Suggested best practice (4000 pixel on long edge)
Letter	11×8.5 in at 400 dpi = 4400×3400	
	11x8.5 in at 600 dpi = 6600x5100	Suggested best practice for text (600dpi 3pt text)
Book page	9 in x6 in at 600 dpi = 5400×3600	Suggested best practice for text (600dpi 3pt text)

Audio

Digitization of audio can be accomplished with modern computers without special equipment. Most portable computers have 1/8 in stereo audio jacks which are connected to the computer's audio analog-to-digital converter. And like scanners it can save the input at various resolutions. Audio resolutions are measured in kilohertz per second. This is like the motor on the scanner as it allows the sampling of the input at various rates. And like color depth in images there is a bit per sample "depth" in audio. The audio stored on CDROMs is normally Stereo (two tracks) 44100hz (44.1khz) and 16bits per sample. Most modern computer's audio analog-to-digital converters can handle 16, 24 and 32bits per sample internally.



WAV is the most common storage format and can be compressed like images using lossy compression systems like MPEG. The reduction in size is again offset by a reduction of the ability to reconstruct all the data originally recorded. Broadcast WAV is a format which allows for additional technical and descriptive information to be stored along with the raw audio streams.

Video

Video Formats

Video file formats are containers that hold an assembly of video segments, related audio segments and metadata. The contents of video format files are normally compressed. Video files can contain video and audio in almost any format, and have file extensions named after the container type

AVI (.avi)

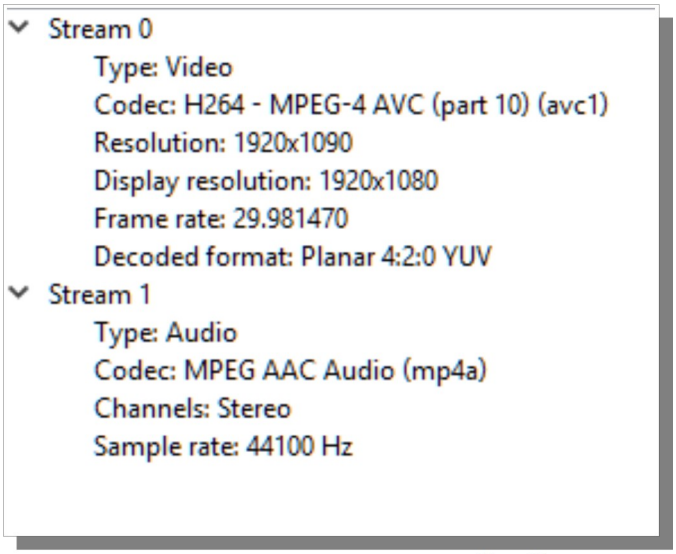
Quicktime (.mov)

Standardized video formats have restrictions on the type of video and audio content and on the types of compression allowed.

Ogg Video (.ogv)

WebM (.webm)

M4v (.m4v)



Video frame sizes

480×320	VHS tape
700×480	NTSC Television Limit
1920×1090	Apple iPhone
4096×2160	4k Digital Cinema

Metadata

Descriptive

Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and subjects.

Structural

Structural metadata indicates how compound objects are put together, for example, how pages are named or gathered to form chapters.

Administrative

Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information.

OCR/AAT/Textural Metadata

Optical Character Recognition

OCR converts images of typed or printed text into machine-readable text format. OCR returns the extracted text, along with information about the location of the detected text that can be used to search for the images file.

AAT

Automatic Audio Transcription

AAT converts audio into machine-readable text format. AAT returns the extracted text, along with information about the time offset of the detected text that can be used to search for the related audio file.

Metadata storage

MARC – Binary machine readable with extensive field definitions

METS/MODS/MIX/audioMD/videoMD – XML machine and human readable with extensive fields

Dubin Core – XML with reduced set of fields

Metadata can be stored within the digital object file. It can also be stored in an external file or a database.

Digital Preservation

To digitize something is to convert something from an analog into a digital format. To digitally preserve something is to maintain it over a long period of time. How do you know that your digital file is not deteriorating like the analog paper object?

Step one. **Add checksum** signatures to each digital file as they are created.

Step two. **Verify** those signatures against your digital files on a regular basis.

Step three. **Make copies** of digital files/checksums and store them in areas away from each other.

Step four. **Evaluate** digital file formats as viable/readable/viewable by current software.

Links to more...

Resolution
https://en.wikipedia.org/wiki/Image_resolution
https://www.archives.gov/preservation/products/definitions/dig-img-qc.html

Color depth
https://en.wikipedia.org/wiki/Color_depth

Indexed color
https://en.wikipedia.org/wiki/Indexed_color#Image_file_formats_supporting_indexed_color

Library of Congress Sustainability of Digital Formats
http://www.digitalpreservation.gov/formats/fdd/descriptions.shtml

Checksum
https://en.wikipedia.org/wiki/Checksum
https://www.avpreserve.com/tools/fixity/
http://www.rescarta.org/index.php/sw/45-checksum-verification-tool

Cornell University Tutorial – Conversion
https://www.library.cornell.edu/preservation/tutorial/conversion/conversion-01.html

JPEG 2000 Truly Loss-less
http://dltj.org/article/lossless-jpeg2000/

SOBEKcm Video
https://www.youtube.com/watch?v=fg2KGf617c8

ResCarta Toolkit http://www.rescarta.org