

Managing a Digitization Project

Visual Resource Association

2005 Annual Conference

March 7, 2005

Trudy Levy, Image Integration

Handouts for:

Planning, Standards and Procedures

New Tools

Project Management



Planning, Standards and Procedures

Importance of defining and setting goals

SPAM Statements - Specific, Positive Attainable Measurable Goals

Use Statement - critical to defining project standards

Standards apply to three components:

Creation

Retrieval

Access

Selection Standards - Getting the Best Source

Scanning Standards - Defining Master Images

Preservation Standards - No Originality please!

Planning, Standards and Procedures - Resources

Bookmarks for Scanning Standards

Standard Setting Bodies

Western State Digital Image Best Practices

http://www.cdpheritage.org/westerntrails/wt_bpscanning.html

California Digital Library's Digital Image Format Standards

<http://www.cdlib.org/about/publications/CDLImageStd-2001.pdf>

NARA Guidelines for Digitizing Archival Materials for Electronic Access January 1998 - http://www.archives.gov/research_room/arc/arc_info/guidelines_for_digitizing_archival_materials.pdf

The NARA- National Archives and Records Administration- guide may be a little old, but it is still quite good and clear.

National Information Standards Organization Technical Metadata for Digital Still Images - http://www.niso.org/standards/resources/Z39_87_trial_use.pdf

This is still a draft document and is not an easy read, but it will introduce to you to the best thinking on this subject.

The following two publications cover more than just good scanning standards and will be useful throughout the life of the your digitization project.

Handbooks

NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials - <http://www.ninch.org/guide.pdf>

NINCH - National Initiative for a Networked Cultural Heritage Handbook for Digital Projects - <http://www.nedcc.org/digital/dighome.htm>

This is produced by the Northeast Document Conservation Center, the original authority on digitizing in the Library sector. The handbook is also available in paper -Library of Congress Cataloging Number ISBN No. 0-9634685-4-5

Planning, Standards and Procedures - Resources

Other Good Sources for Expert Opinions and Advice:

Research Library Group Guides to Quality in Visual Resource Imaging -
<http://www.rlg.org/visguides>

These are a series of articles all of which you will find most helpful.
A few scanning tips - <http://www.scantips.com/>
Wayne Fulton presents a little bit lighter and easier read. Many people with a great deal of digital experience use it. Especially read "Evaluating Scanner Features and Performance"

Technical Advisory Service for Images, TASI - <http://www.tasi.ac.uk/>
From Great Britain. They provide this great resource their academic community about the digital creation, storage and delivery of image-related material.

Digital Image Guide - <http://www.dig-mar.com>
Image Integration's Resource on the web

New Tools

DAMS – input is not the same as output

DAMS developed for different purposes – archival, teaching, workflow

Level of Access – enter information, view data, use image

Color

Calibration and Profiles – language of color

Dynamic Range – range of tones from dark to light

Gamma – Contrast ratio

File Naming

What Meaning

Unique – beyond your borders

Meta tags

Project Management

Project Management

Through out the life of a project – Track and Assign responsibility

Critical Path Diagram

task	staff person	training?	hrs	week 1	week 2	week 3	week 4	week 5	week 6	week 7	week 8	week 9	week 10	w
Managing Your Digitizing Project* course	RA		8	2 hours	2 hours	2 hours	2 hours							
Museum manager (based on MDG course)	RA		4					4 hours						
receive approval to proceed	from DB		-					•						
management (5 hours per week * 16 weeks)	DB		80	5 hours	5 hours	5 hours	5 hours	5 hours	5 hours	5 hours	5 hours	5 hours	5 hours	
hand-written data on photo index cards	RA		8						4 hours	4 hours				
crosswalk to Dublin Core fields	RA		8							4 hours				
reflect fields (in-house museum software)	RA		8								4 hours			
develop 3-way crosswalk chart	RA		8									4 hours	4 hours	
Rudy Levy's scanning class (Infopeople)	RA		8							8 hours				
scanning (as thumbnails? Higher res?)	RA/DB		-								•			
inner & disk space (& redesign project!)	RA/DB		-								•			
mail only, get existing scanner functioning	DB		20									20 hours		
collection (300 records and photographs)	NS		12									12 hours		
check condition of photos selected	DB		6										6 hours	
records and photos as needed to reach 300	NS		4										4	
project team on data entry into PastPerfect	RA/NS/DB		4										4	
reflect at an average of 20 min per record	RA/NS/DB		100											
train project team on scanning methods	RA/NS/DB		4											
photos at an average of 8 minutes per scan	RA/NS/DB		40											
reflect custom reports to check data entry	RA		8											

Evaluate & Document Success – Digital Slide count = Web logs

Project Management

Monthly Statistics for February 2004		
Total Hits	14321	
Total Files	12498	
Total Pages	3196	
Total Visits	1925	
Total KBytes	71216	
Total Unique Sites	1461	
Total Unique URLs	292	
Total Unique Referrers	284	
Total Unique Usernames	4	
	Avg	Max
Hits per Hour	25	261
Hits per Day	622	933

New Tools Glossary

Glossary of Terms

Graphic Image Formats

There are too many formats [1] for communicating image information to computers and their peripherals to list here, but I will give you the highlights and some interesting and maybe even helpful information about them.

First thing to learn is that the Mac operating system uses 4 letters in its tags to define files. They are shown here as [tag]. More importantly, Macs do not bother to show you these tags or attach them to a file. In the past the Windows operating system used only three letters in their tags and it can not read a file if it does not have a tag.

Rule 1: Mac operators must always insert file tags on their files when saving them.

List of File Formats

Tag	Full name	Defined by
BMP	Windows Bit Map	Microsoft
EPS	Encapsulated Postscript	Patent owned by Adobe
GIF	Graphic Interchange Format	Patent currently own by Unisys
JPG [JPEG]	Joint Photographic Experts Group	ISO standard 10918
PCD	Kodak Photo CD	Patent owned by Kodak
PNG	Portable Network Graphics	Internet Committee - W3
PDF	Portable Document Format	Patent owned by Adobe
TIF [TIFF]	Tagged Image File Format	Adobe

BMP is the MS-Windows standard format for graphic files created within its programs. It is a raster image as it is a "mapping" of each bit of information. Raster imaging is not scaleable as vector imaging is, but it is able to show more subtle variations in color as it does map each pixel. It also makes a larger file size.

EPS is a Vector image and therefore scaleable. Vector images are equations or math formulas; the computer then translates the formulas to the monitor's or display's pixels. It is smaller and will show crisper edges.

New Tools Glossary

GIF is the grand daddy of image formats for the web. It originally was DJIF and therefore the soft “g” pronunciation. The LZW compression algorithm used in the GIF format is owned by Unisys, and companies that make products that exploit the algorithm (including the GIF format and PhotoShop) need to license its use from Unisys. There are two versions of the format, 87a and GIF89a. Version 89a (July, 1989) allows for the possibility of an animated GIF, which is a short sequence of images within a single GIF file. A GIF89a can also be specified for interlaced (progressive) GIF presentation.

JPEG or (Joint Photographic Experts Group), is the ISO/IEC group of experts that develops and maintains standards for a suite of compression algorithms for computer image files. They have released an updated version call JPG 2000 that is slowly being adopted.

A JPEG file is created by choosing from a range of compression qualities (actually, from one of a suite of compression algorithms). When you create a JPEG or convert an image from another format to a JPEG, you are asked to specify the quality of image you want. Since the highest quality results in the largest file, you can make a trade-off between image quality and file size.

The JPEG's correct name is JFIF for JPEG File Interchange Format. The digital camera manufacturers developed another version call JPEG EXIF. The difference is the metadata information that is stored within the image file. The JFIF stores information about the color profile, while the EXIF stores information about the camera.

Kodak's **Photo CD** uses a proprietary image format called an Imagepak that has its own unique image compression technology and color management. It is very good, but will only be found on Photo CDs. The 4.5 MB files expand to clear accurately colored 18mb files. As with any computer file, you can copy the PCD files from the CD-ROM to a server within your system. If you use this scanning method - usually only provided by photo vendors - you can keep them in their original format or convert them to a more useful image format, keeping the original files as archives.

PNG as with the previously mention JPEG 2000 are attempts to create a non-proprietary improved web image format. The move is toward interested compression formulas that do not lose information. It is extensible file format for

New Tools Glossary

the loss less, portable, well-compressed storage of raster images. Extensible means you can extend its specifications and develop other versions.

Another extensible format is the TIFF. It utilizes a structure that can store image data and attributes of that image. It has been extended into several versions, the most common of which is TIFF-Fx for the fax machine.

[1] For an example of a still only partial listing of formats, see <http://www.daubnet.com/formats/index.html>

Computer Hardware and Software terms

Algorithm A step by step Procedure for solving a problem

Bit The smallest unit of computer information, = 0 or 1, a binary digit

Byte a sequence of bits, usually 8 that is used as a unit of storage. The fact it is 8 bits rather than 10 is why 5 megabytes do not equal 5000 KB.

Display Technology:

CRT Cathode-ray tube - Like a TV an electronic discharge is shot at the back of its screen "excite" the phosphor coated back of the display. This is the older display technology and so that there are several image display management tools created for it. It does not use the same frequency as a TV though, which is why they are not interchangeable.

LCD Liquid Crystal Display - sometimes referred to as a Thin Film Transistor (TFT) display. A current is sent across two conductors on the grid to control the lighting of each pixel. These are the flat panel monitors, which use less space, and less electricity, but, as a relatively new technology, the color and other display management systems are in their infancy

OLED Organic Light Emitting Diodes - Newly developed by Kodak it is still only available on cell phones and PDAs. The diodes transmit light when a low current is passed through them.

Plasma This flat panel display uses charged gas and Phosphors to create its display. It is brighter then and LCD panel and has a broad viewing angle. They can be quite large such as in high definition TV.

OS Operating System - A set of programs used to control, assist or supervise all the other programs that run on that computer.

Resolution: For a display device, it is the quantity of pixels that the device can display.

New Tools Glossary

Bit-Depth: the number of "bits" of information about something, in imaging it usually describes color but can also describe sound in audio equipment. Note that each additional bit in a binary number system doubles the possibilities so that a 24-bit process is 2 to the 24th power or 16,777,216 available descriptions of information.

Bitonal: a bitonal, a.k.a. black & white or line art, image consists of pixels that can represent one of two tones.

Compression: a process used to reduce image file size. Image-compression algorithms may be either information-preserving, a.k.a. lossless, or non-information-preserving, known as lossy.

Image Types

Raster image is a bit map or grid of x/y coordinates specifying each pixel on a display. It is a large file

Vector image is a representation of a sequence of mathematical statements or sequence of commands that place lines or shapes on the display screen. It is a smaller file than a raster image, details "lines" or edges very well and is scaleable.

Pixel: the smallest logical unit of visual information that can be used to build an image. Short for picture element, pixels are the little squares that can be seen by enlarging an image.

equal to one 'bit' if black and white (monochrome, or 2-colors),

equal to one 'byte' (8 bits) for 'Indexed Color' (256 colors),

equal to 2 bytes for 16-bit color ('Thousands of Colors')

equal to 3 'bytes' (24 bits) for 'True Color', or ('Millions of Colors')

equal to 4 'bytes' (32 bits) for " " + 'alpha' (transparency) channel

RAID: (Redundant Array of Inexpensive Disks). Hard disc arrays enable very fast access to data + Much increased reliability through redundancy (information may be mirrored / striped across several discs. Range from Level 0-5 implementation.

Imaging Terms

CLUT: Color Lookup Table (or Palette) a table used to choose which of 256 colors are used in 8-bit, Indexed color. Every computer has a built-in ('System') CLUT or else the image has to provide it.

New Tools Glossary

Color Space mathematical model of color as a three dimensional space using color terms as the axis

Digital Watermark A digital watermark is an identification code carrying information about the copyright owner, the creator of the work, authorized consumers and so on. The watermark is invisible and permanently embedded into digital data for copyright protection and for checking if the data has been corrupted. By means of watermarking, the work is still accessible, but permanently marked.

Dither: The process of electronically arranging spatial clusters on a displayed or printed image. Black & white process used within the printing industry to represent color or grey tone images. Similar to Half-Toning as used by the newspaper industry.

DVD: (Digital Versatile Disc). These are, as yet, an unproven standard. Several variants currently exist – DVD-R, DVD-RAM, DVD-RW, DVD+RW. Backwards compatibility is a key concern currently. Current capacities exceed 9 GB (selected variants only).

Dynamic range: The range of information concerning the light to dark values in an image.

EDMS: (Electronic Document Management System). Term to describe the storage and management (including retrieval) of documents, which are input to a computer system via a variety of mechanisms (including scanning device, conventional software products such as Word Processors, faxes, email).

Greyscale: (8 bit). Usually refers to a continuous tone image (typically 256 shades of grey).

Histogram: plot of the number of pixels for each possible grayscale value. A histogram is a graph that shows the distribution of intensities in an image. The horizontal axis represents the full range of tonal values; the vertical axis indicates the number of pixels for each intensity value.

Highkey an image where the histogram is weighted toward the white point, thus resulting in a light image.

Lowkey an image where the histogram is weighted toward the black point, thus resulting in a dark image

Hue: color.

New Tools Glossary

Resampling: resampling occurs in two forms: down sampling and up sampling. Scanning often occurs at a higher resolution than is necessary and the required resolution is obtained by resampling the image. Aliasing can occur when the image data are downsized. To minimize its effects low-pass filtering can be applied to the image before it is downsized. Up sampling should be avoided because no additional image information can be created.

Saturation: the intensity of the color.

True color a.k.a. 24-bit color, is the specification of the color of a pixel on a display screen using a 24-bit value. Each RGB level is represented by the range of decimal numbers from 0 to 255. The total number of available colors is $256 \times 256 \times 256$, or 16,777,216 possible colors.

TWAIN: (Scanner driver). Generic device driver that allows the scanned image captured by the scanner to be passed to the host PC. The word TWAIN originates from Kipling's "The Ballad of East and West" ~ "...and never the twain shall meet...", reflecting the difficulty, at the time, of connecting scanners and personal computers. It was up-cased to TWAIN to make it more distinctive. This led people to believe it was an acronym, and then to a contest to come up with an expansion. None were selected, but the entry "Technology Without An Interesting Name" continues to haunt the standard. (Taken from the www.twain.org site)

Unsharp masking: an artificial transformation that can make images appear more in focus by exaggerating transitions that have been dulled by scanning.

Value: the amount of light to dark in the color.