Follow up Information

Caring for Digital Materials:
Digitization and file conversion

Jake Nadal, instructor

General Digitization and Conversion

Q. *We have over 500 audio cassettes from a 1970s oral history project that need to be converted. What should we be aware of?*

Start with your metadata plan. You'll want to work that out with your vendor (or youself, if digitization will be in-house) in advance. The actual digital capture will be similar from tape-to-tape, but knowing how to describe the files that are created before you create them is important.

Q. *I'm quite curious about the best strategy involved for conversion of microfilm/fiche to searchable PDF. Will/can we address that at some point in this series?*

There are a number of vendors and if you are purchasing a scanner, it should include software packages that handle this conversion. Basically, your scanning will create image files that will be passed to an OCR program to create text. This text and the scanned image are embedded in the PDF as layers in the page.

Q. *If you bump up the resolution, aren't you actually "adding" image data? In other words, not necessarily the "true" image capability if the scanner capability is only optimized for 300?*

Yes, the scanner will have an *optical* resolution and an *interpolated* resolution. The interpolated resolution uses a software algorithm to estimate color values between the optically scanned pixels. Read the specs carefully.

Resource: Three part series on the Library of Congress' blog "The Signal":

- http://blogs.loc.gov/digitalpreservation/2012/12/what-resolution-should-i-use-part-1/
- http://blogs.loc.gov/digitalpreservation/2013/01/what-resolution-should-i-use-part-2/
- http://blogs.loc.gov/digitalpreservation/2013/03/what-resolution-should-i-use-part-3/

Q. *Can you please provide a link to the visual color spectrum graphic?*

http://en.wikipedia.org/wiki/Visible_spectrum

Q. *How do I work with ICC profiles?*

You'll need a calibration setup, which should include printed color targets (which you scan or photograph), a spectrometer (which measures colors from your screen or printer) and calibration

software (which creates an icc profile based on these measurements). Searching for "color calibration system" will turn up lots of options for you, and a visit to a camera shop (online or in-person) can be helpful.

Q. *Does the ICC profile need to be downloaded or is it usually packaged with a computer?*

There are standard ICC profiles (e.g. "Adobe RGB") that will be packaged with your imaging software. Color calibration creates a device specific profile for your equipment.

Q. *Is it necessary to calibrate a scanner with a color correcting sample before beginning a new scanning project?*

--See above--

PDFs and PDF/A

Resources:

- http://www.adobe.com/enterprise/standards/pdfa/
- http://www.cvisiontech.com/reference/pdf-a/pdf-archive-format.html

Q. *What equipment/software is needed to convert to PDF/A?*

Adobe Acrobat Pro is the standard, but recent versions of Word also save to the format and there are many third-party PDF conversion tools just a web search away.

Q. *How do you know which pdf format you are saving in? Is there something in the filename that would tell you if a document is PDF or PDF/A?*

The filename will not, but opening the file with a type of program called a PDF validator will let you read the PDF header to be sure. JHOVE is a program that is often built into digital preservation systems to validate files of many types.

Q. *What kinds of objects in PDFs might cause problems later?*

Images (which may be compressed), audio and video, and interactive forms, for example. PDF is a wrapper that can contain many types of digital object.

Q. *Is pdf 1.4 still widely available and how compatible is it with later versions?*

PDF is a backwards-compatible format, so if you program works in 1.7, it should still read 1.4.

Q. *Is it better for preservation to convert .doc or .docx to PDF if don't have PDF/A?*

.doc/x are so widely used that you should be safe with them. If you do want to convert to have a fallback, there are a number of free or inexpensive .doc(x) to PDF/A converters available. Current versions of Word actually can save to PDF/A automatically.

Q. *Should you keep the not-preferred format as well as the copy you converted to a pdf/a?*

Always wise to keep the source format, even if it's a preservation problem.

Q. *Is PDF is a format to preserve music scores or dance scores which consist of symbols?*

Potentially, but be aware that PDF may display the score well visually, but may not actually encoded what it means musically, the way a MIDI file or MusicXML file could.

Microsoft Word and related formats

Q. *Is docx more stable than doc?*

It's more open, to be precise. .docx is an XML package, so it can be read in any XML reader, although that XML itself is pretty obtuse. .doc is a binary file that only Word (or a similar program) knows how to read.

Q. *How do you deal with older software like Word Perfect vs. Microsoft Word?*

WordPerfect is a tough one. It's wise to keep the original files, in the hopes that a good WordPerfect reader/emulator comes along. That said, it's also wise to migrate them to Word, ODF (Open Document Format), or PDF as well.

Q. *Should Microsoft documents created on older Apple machines be converted to .docx, .pptx, etc., for easier preservation?*

Yes, migration from older .doc versions to newer versions is a good idea, no matter what platform they were originally created on.

Q. *So are you recommending Microsoft Word as a preservation format? What about backwards compatibility issues or restricting the ability to modify such documents?*

Microsoft Word presents a low preservation risk, though it is not itself a great preservation format.

Q. *We have a number of born digital documents in Word 3.1 or Word Perfect. Our shop currently runs Windows 7. How do we make the jump?*

First step is to make sure your original files are safely stored. You should keep them as "masters" in the event that good conversion/emulation becomes available.

Recent versions of Office have conversion software for WordPerfect 5.x and 6.x. (You may need to activate these through the windows "add program" control panel.) Older Word versions should open in Office natively.

Images

Q. *What about the DNG format for image preservation?*

DNG is a viable preservation format for born-digital photography.

Q. When *accessing JEPG's vs. PDF's, is there memory loss?*

This is an apples and oranges comparison. PDF is a wrapper that can contain JPEG, TIFF, and many other image formats. You'll want to check you PDF creation program's settings carefully to see how it is storing images.

Q. *Is it ok to transfer JPEG images into TIFF without rescanning the item?*

You won't actually gain anything by doing this. The JPEG has as much data as it will ever have, and decompressing it to save as TIFF just eats disk space. If you want higher resolution that the JPEG contains, you'll need to rescan.

Q. *What about using LZW loss-less compression on TIFF files. Does this cause any concerns for preservation?*

It's a very small risk, but the potential problems is that you have a break in the metadata chain so that later users (or software programs) do not know that the file is compressed, and assume that it is in fact corrupted. Given that it's a TIFF file, LZW compression would be a normal guess at the problem, and an easy one to correct.

Q. *Is there anything different/special you need to do when digitizing images for use in high definition media -- TV, etc.?*

High definition can mean many things – there are several specifications that are branded "HD". Each one specifies pixel dimensions for the HD image. As long as your digital master has that many pixels or more, it will display fine as HD, and most digital images that meet library and archives standards actually exceeds HS standards.

Other formats

Participant-supplied resources for formats

- Library and Archives Canada (LAC) Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-term Access: http://www.collectionscanada.gc.ca/obj/012018/f2/012018-2200-e.pdf

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

Q. *Can you offer any resources on preserving complex vector graphics such as CAD?*

Familiarizing yourself with the SVG standard and the CAD data exchange is a good starting place.

Q. *This likely outside the bounds of this general discussion, but are there resources for preserving architectural, engineering, & construction (AEC) 3-D CAD models (outside of the MIT FACADE project)?*

This is getting out of the scope of the discussion. I'd suggest contacting the people at the MIT FACADE (Future-proofing Architectural Computer-Aided Design, an IMLS-funded project which ran from 2006-2009; see http://facade.mit.edu/) to ask about additional resources.

Q. *What about PNG format?*

Useful for delivery, but not an ideal master format, due to its compression.

Q. *We have thousands of items scanned in JPEG, should we be concerned?*

Generally, preserve what you have. If you still have the original photos, it's important to preserve them. The JPGs are your first-line backup. You don't need to rescan unless your patrons or researchers ask for higher-quality digital copies.

Audio

Participant-supplied resources for audio

- International Association of Sound and Audiovisual Archives (IASA): http://www.iasa-web.org/listserv
- Video Format Identification Guide: http://videopreservation.conservation-us.org/vid_id/index.html
- Museum of Historic Video Equipment: http://videopreservation.conservation-us.org/museum/index.html

- Library of Video History, Science and Technology: http://videopreservation.conservation-us.org/library/index.html

Q. *Is .wav same as bwav? Can we convert .wav to bwav files?*

BWAV is WAV plus a small metadata header. The sound data is the same in both.

Q. *Will you talk about converting 44.1 16-bit to 96 25 bit?*

Generally you can't "add" information after you've done the original digitization. Try to capture at a higher resolution, sampling rate, or and then convert to a lower (smaller) file for access. Sometimes with these media you get one good shot at them, particularly bad reel audio tapes.

Q. *Is any of the conversion equipment (audio recordings to digital) that is available to a home owner able to "do the job"...asking as a small low funded museum.*

See "Recommendations for DAC" below

Q. *What about born digital audio? We initially converted to Quicktime and Real Player.*

MP3 is probably a better delivery format, unless you have rights-management needs, and BWAV is far and away a better choice for your master files.

Q. *How about material that is provided in mp3 format?*

Always keep the original source material. MP3 is well-documented and low-risk, but also fairly compressed. If the provider has a higher-resolution source, it's worthwhile to try to get a copy.

Q. *Recommendations for DAC?*

Mike Casey, "Sound Directions: Best Practices" has a very, very good example of how to set up a capture sound lab based on what they did at Indiana University (see pages 14-22): http://www.dlib.indiana.edu/projects/sounddirections/papersPresent/sd_bp_07.pdf

Websites

Q. *Can anyone link to resources on recommendations for archiving entire web pages?*

Wikipedia has a fairly good page on archiving websites: http://en.wikipedia.org/wiki/Web_archiving. The Library of Congress also has a Frequently Asked Questions page about web archiving: http://www.loc.gov/webarchiving/faq.html, which includes information about permissions. They, like most organizations interested in this activity, focus on preserving entire web sites. The International Internet Preservation Consortium (IIPC) (http://www.netpreserve.org/) is the main body involved with setting standards for this activity.

Other questions? Email info@heritagepreservation.org and include "Caring for Digital Materials" in the subject. Or join the community at http://www.connectingtocollections.org/ and ask your questions there!