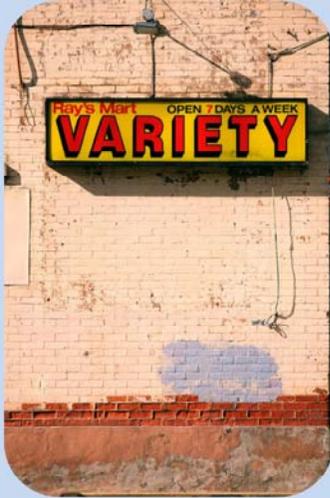


Variety, variety... it's the spice of life.

There are so many file formats – how do you know which ones are right for your situation?

The good news is you'll find lots of guidance and opinions to help you navigate the variety.

Why do we care?



- Initial digitization = effort + \$\$ + time
- Cultural/institutional responsibility
- Preservation over time takes work
- We already know formats can be a problem

But why do we care which file formats we use during digitization? Well, here are a few reasons...

1. Initial digitization, which you already know or may be finding out soon, takes effort, money, and time. It's an investment, and redoing it for any reason, including poor file format choices, can get you scrutinized by folks in charge or the public. Plus, I have yet to meet anyone who enjoys redoing work.
2. As members of cultural heritage institutions, we take our stewardship responsibilities seriously. Responsibly creating and managing these digital files is included in that stewardship.
3. It takes work to preserve digital files –more work than preserving most manuscripts, books, photographs or film. If we're going to go to that expense and effort, why not do it with files in the most robust formats we can choose?
4. Folks have already had experiences with obsolete formats, formats that don't hold up as they need them to. This isn't a brave new world. So why not do the best we can to avoid those known problems?



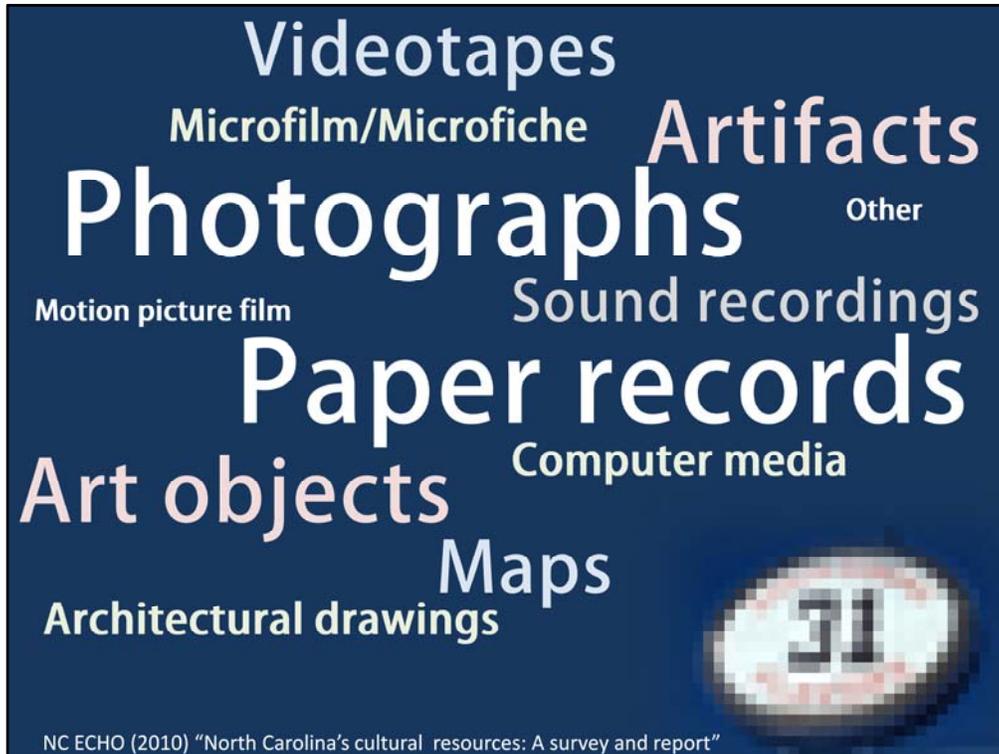
You may already know this, but I want to go over it anyway to make sure I'm clear about the types of file formats you might have after digitization.

The preservation format is your original image capture. It should be high quality, and you shouldn't be making changes to it through image manipulation. Keep that preservation format pristine, so you can go back to it again and again over time to create...

Your access formats. These are files that are meant for use. They are often smaller size, sometimes smaller resolution, usually easier to put online. Sometimes they've been cropped, color corrected, etc.

Today, I'm predominately talking about the preservation file format – the format you create when you digitize an item. I'll only touch on access format recommendations but can answer questions about those later.

Now, let's take a look at what North Carolina institutions might be digitizing.



The NC ECHO (Exploring cultural heritage online) initiative surveyed 761 of the 950 identified cultural heritage institutions in NC.

Here are the types of media found at those institutions. The larger the word, the more prevalent the medium.

of institutions holding types of media:

- 80.0% **Paper records**
- 78.4% **Photographs**
- 64.4% **Artifacts**
- 57.3% **Art objects**
- 50.8% **Maps**
- 45.3% **Videotapes**
- 36.6% **Sound recordings**
- 33.1% **Architectural drawings**
- 29.0% **Microfilm and microfiche**
- 26.5% **Computer media**
- 18.0% **Motion picture film**
- 15.0% **Other**

I'll be focusing on those big words – paper and still images - today. They're also the area in which I have the most experience. If your institution is doing audio and video, which is much more complex, I will touch on those formats but feel more comfortable directing you to folks who can give you more in-depth information if you need it.

Okay, so with soooo many different types of file formats out there - how do you decide what types of file formats to use?



Choosing file formats from the multitude of options

- What will your hardware produce ?
- What makes a file format a good choice?

That same NC ECHO survey found that 90.5% of digitization is being done in house, where you have some control over your procedures.

You will need to look at what your hardware and software will produce – don't rely on just the default format or the format that comes up first in the settings box. Look at all of the options and do some homework. The good news is that most of the common machines used for digitization will have the recommended preservation formats as options.

Next, you have to consider what characteristics make a file format a good choice.



Here are good choices to make for your file format diet:

Non-Proprietary – For some formats, a company or organization has ownership of or control over how a file can be displayed because they’ve restricted access to the code. Think of Microsoft products, Adobe Photoshop, etc. Those are called proprietary formats. Where you can, choose non-proprietary formats, because of the ability for anyone with programming experience to help display that file.

Common(ish) – Any type of format that isn’t commonly used, either because it’s a format that’s fallen out of use or because it’s associated with a piece of software that isn’t broadly adopted, is considered more of a risk.

Lossless – When a program saves a file in a lossy format, it uses an algorithm to determine parts of a file that it feels can be considered redundant or extraneous. This is helpful for making a file smaller, but you don’t want to lose any of your scan data for your preservation copy of a file – you want it to contain as much information as possible. So go with a lossless format for preservation files.

Recommended preservation formats?



Images: TIFF (.tif) or JPEG2000 (.jp2)*



Audio: WAVE (.wav) w/BWF header



Video: MPEG-2 (.mpg, .mpeg, .mp2) or MPEG-4

Here are some generally accepted recommendations for digitizing different types of media. Cultural heritage institutions all over the world have settled on these types of formats because of their losslessness, their ability to withstand corruption, and their ubiquity.

TIFF is the recommended format for an image file. Many also use JPEG2000, which is a newer format. However, I've put an asterisk next to JPEG 2000 because of an ongoing debate related to whether or not it's a good choice as a preservation copy. It is more easily corrupted than other formats, however it can offer a significant size savings, and if you have redundant and managed preservation storage, corruption in one spot may not be an issue.

For audio, WAVE format with a BWF header is recommended. BWF stands for "broadcast wave format," and it basically adds additional information to the .wav file header.

MPEG-2 and MPEG-4 (which is just a more recent version of MPEG-2) are recommended for video.

Keep in mind – some of the recommended preservation formats are QUITE large in size. I'm talking about the difference between a few hundred KB vs. a few hundred MB. You need to plan for storage needs up front.

Recommended access formats?

Supported YouTube formats

YouTube supports a wide variety of the range of **video and audio formats** in use. You can get a quick overview of what may be an issue with your video's formatting by taking a look at the warnings and error messages that we communicate to you while you're uploading your video.

Here is a list of some well-known formats that YouTube supports:

- **WebM files** - VP8 video codec and Vorbis Audio codecs
- **.MPEG4, 3GPP and MOV files** - Typically supporting h264, mpeg4 video codecs, and AAC audio codec
- **.AVI** - Many cameras output this format - typically the video codec is MJPEG and audio is PCM
- **.MPEGPS** - Typically supporting MPEG2 video codec and MP2 audio
- **.WMV**
- **.FLV** - Adobe-FLV1 video codec, MP3 audio



I'd like to upload a book. What format should it be in? How do you do your sponsored scanning for Contributing Libraries?

Probably the simplest way to contribute a text item currently is as a pdf. That way, the entire set of images can be submitted as a single file, and there are no special naming requirements, beyond ending the filename with ".pdf". If the pdf has no hidden text layer (i.e., isn't searchable), then after doing OCR, Archive.org creates a second pdf with a text layer.

Items can also be submitted as a stack of image files, one image per page. The files can be in JPEG2000, JPG, or TIFF format. We plan to provide a more flexible intake procedure, but at present, there are rather strict requirements for how the files in an image stack are to be named, and the stack needs to be packed into a single .zip or .tar file before submission.

Are there limits on file sizes or file types for uploads?

With a free account, you can upload photos up to 15MB in size, or 2 videos per month- up to 100MB for each video. If you have a Pro account, you can upload photos up to 20MB and videos up to 500MB in size. For more on video files and types supported, please visit the [Video FAQ](#).

Flickr officially supports JPEGs, non-animated GIFs, and PNGs. You can also upload TIFFs and some other file types, but they will automatically be converted to and stored in JPEG format.

As you publish photos, they're compressed and resized by Flickr (if necessary) in the following sizes:

- 75x75 pixels
- 100 pixels (on the longest side)
- 240 pixels
- 500 pixels
- Large (which will be 1024 pixels if it exceeds that length)
- The original size (if you have a pro account)

What about access formats?

I personally recommend that folks look to the upload suggestions of familiar media sharing sites like those shown here. These guys are experts in presenting easily-accessible content online that most people can use with the software on most computers, so why not take their word for it?

Regardless, be sure to consider file size and how long it might take to download the file for the average user. The Department of Commerce reports that, while broadband use in cities and rural areas is growing, it still only makes up 70 and 60 percent of the total. Not everyone can download files quickly.



I know it's not what this presentation is directly about, but I also wanted to briefly discuss born-digital file formats. Many cultural heritage institutions are starting to find donors interested in submitting digital files. You don't have any control over what they want to contribute, which is a challenge. These should also be treated with some planning.

First, relational files (files who either must be used in conjunction with each other or whose meaning is enhanced when they operate together) present an extra layer of difficulty because it's not simply one file whose format you need to worry about. I'm thinking of specialized situations like GIS data, certain types of databases, etc.

Commonly used but proprietary formats can also present challenges. We've had trouble with .ppt and .pub, which are common but hard or impossible to convert to an open format without some data loss.

Although very few people enjoy writing policies, it really is in your best interest, if you need to collect and manage born-digital files, to establish policies about what you will take in the first place, as well as what you will promise about what you collect. It's not a bad idea to do that for all of your digital file management. 😊

Resources

Sustainability of digital formats
(Library of Congress)

<http://www.digitalpreservation.gov/formats>

PADI file formats bibliography

<http://www.nla.gov.au/padi/topics/612.html>

State Library of North Carolina Digital
Information Management Program

Email us: Digital.info@ncdcr.gov

FOR DEPTH:

File formats for preservation
technology watch report
(Digital Preservation Coalition)

http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation

Presentation photos by flickr users cheesy42, fullres,
lusterbr, cuttlefish, lordog, phatman, thenkv, and Bart Heird

