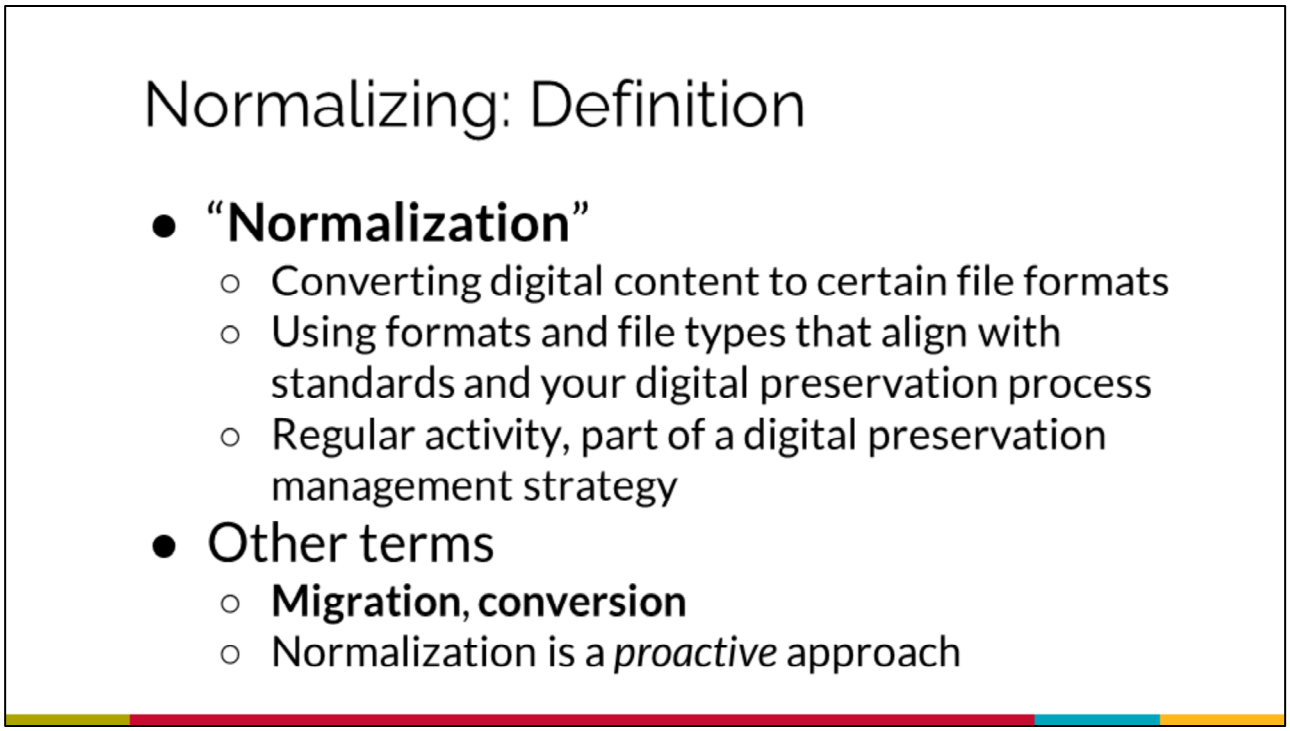
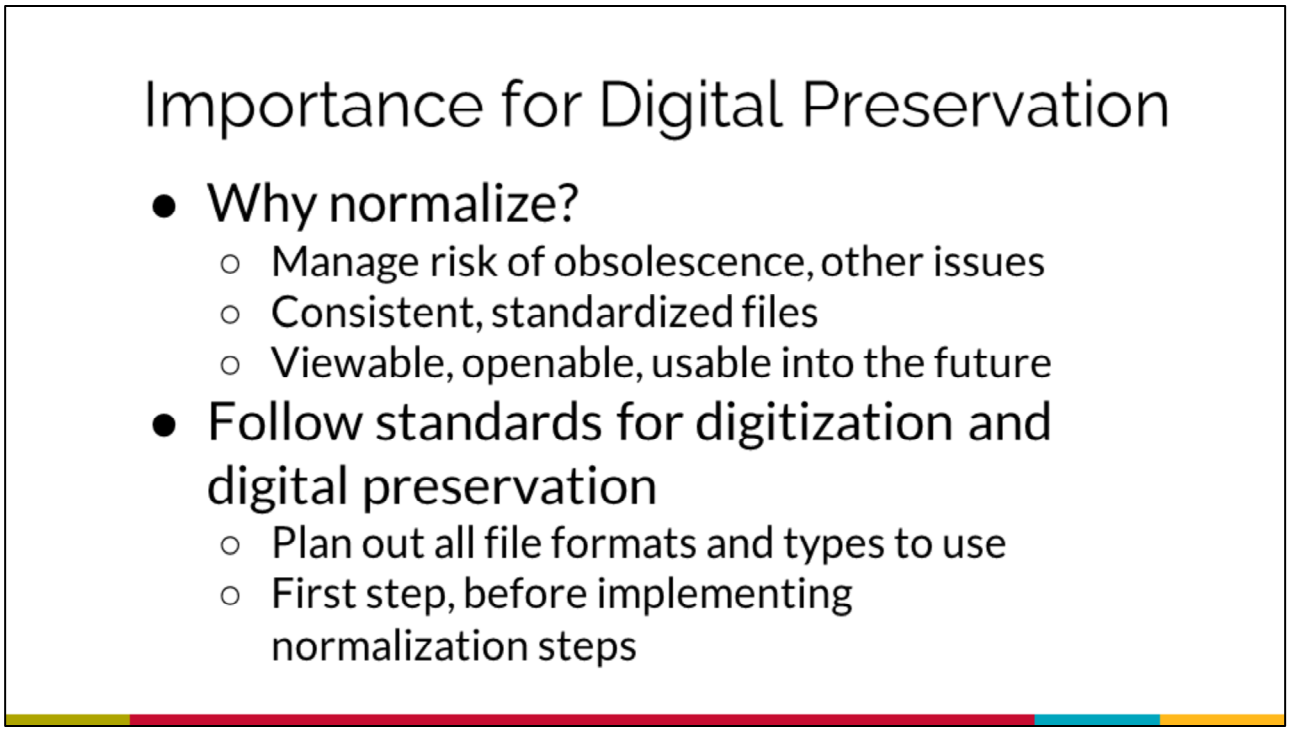
Normalizing Digital Files

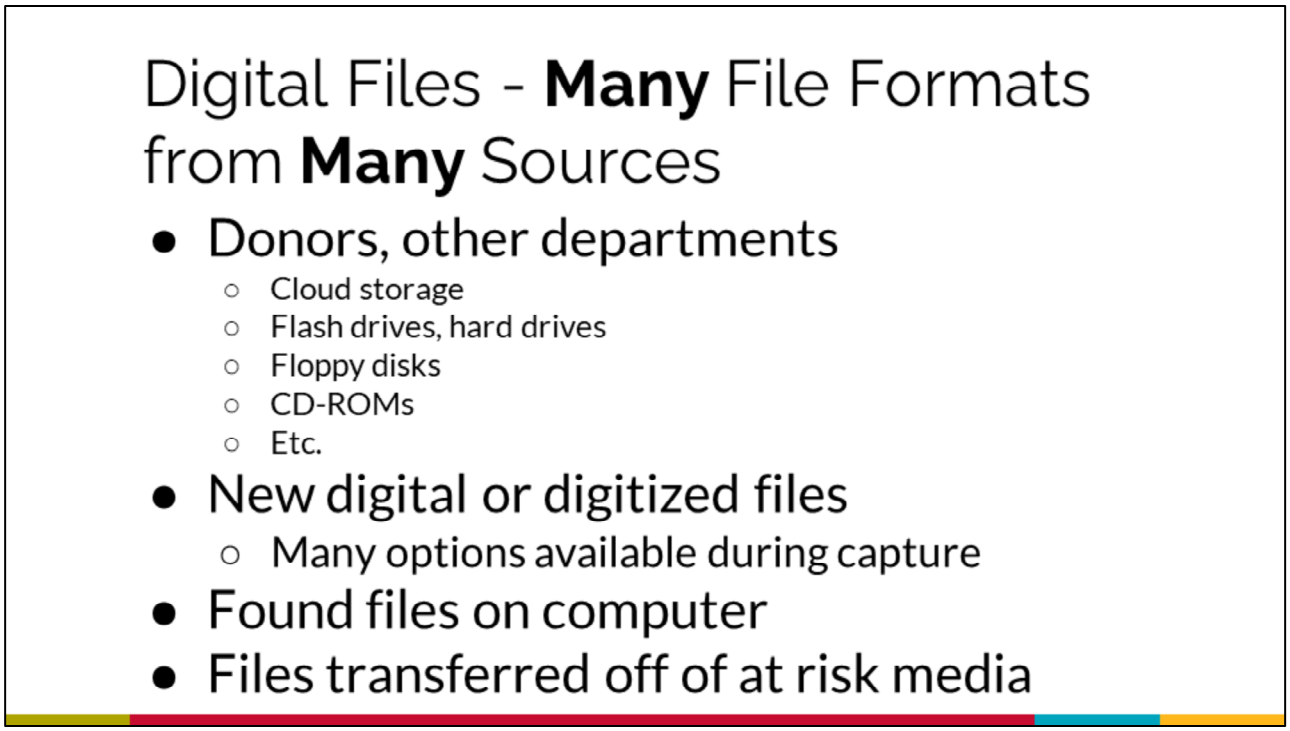
Digital Stewardship Curriculum



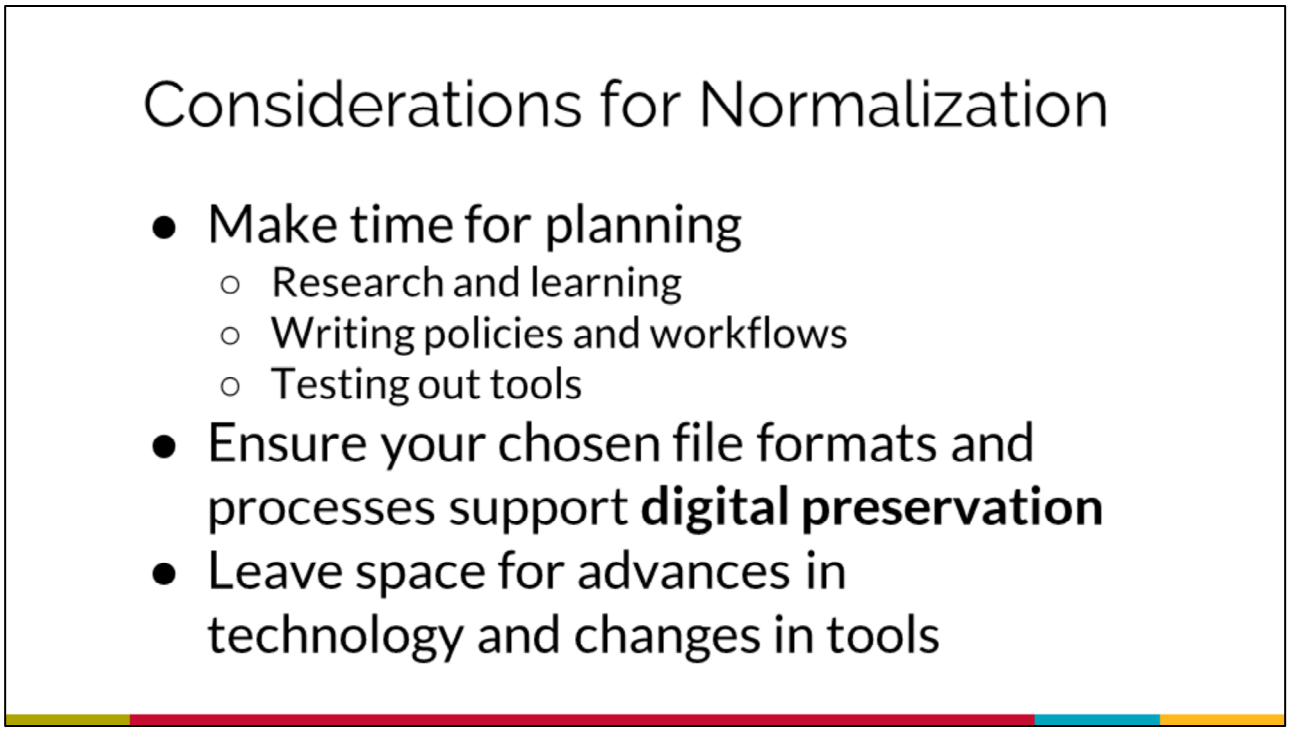
- Normalization means converting files from one format and type to another - with the new format being part of a set list of a small number of formats that are acceptable
    - For example, instead of having all sorts of image file types in a folder (.jpg, .tiff, .gif, .png) - you would go through and convert all the files to (.tiff) or whatever the standard for file types is in your department.
    - This allows you to focus on preserving ONLY the file types on your set list, instead of any file type that gets donated or created
- Make sure that your list of acceptable formats are in line with current standards and the ways that you do digital preservation management at your institution
- Migration, conversion and other terms may refer to similar processes
    - Migration sometimes refers to migrating content to a new format WHEN it becomes obsolete and not usable (or when describing copying files from old media to new storage)
        - Waiting and examining files on a case by case basis is time consuming, and the risks are higher that you might not be able to recover the files
- Normalization is proactive and is a step to perform on all digital content in your care once you establish how you will accomplish it



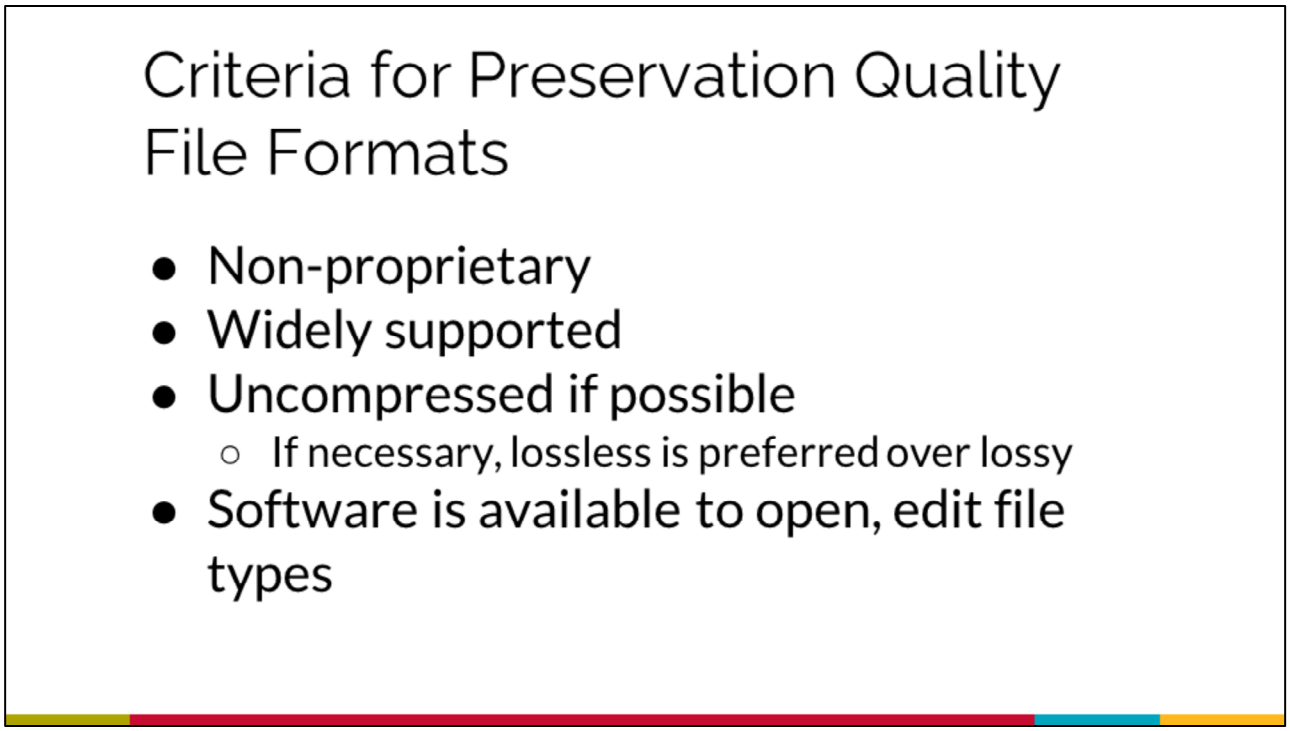
- Why?
  - A way to manage risks of file types becoming obsolete or not openable
  - Consistency and same files to work with
- You should have some standards picked out
  - FADGI standards, resources at Library of Congress and professional organization websites
  - SHN resources: Standards and Specifications Slides
- These standards will apply to files that come into your collections in different ways.
- You will have to keep up on the most sustainable formats - for national standards, but also as things change in your institution



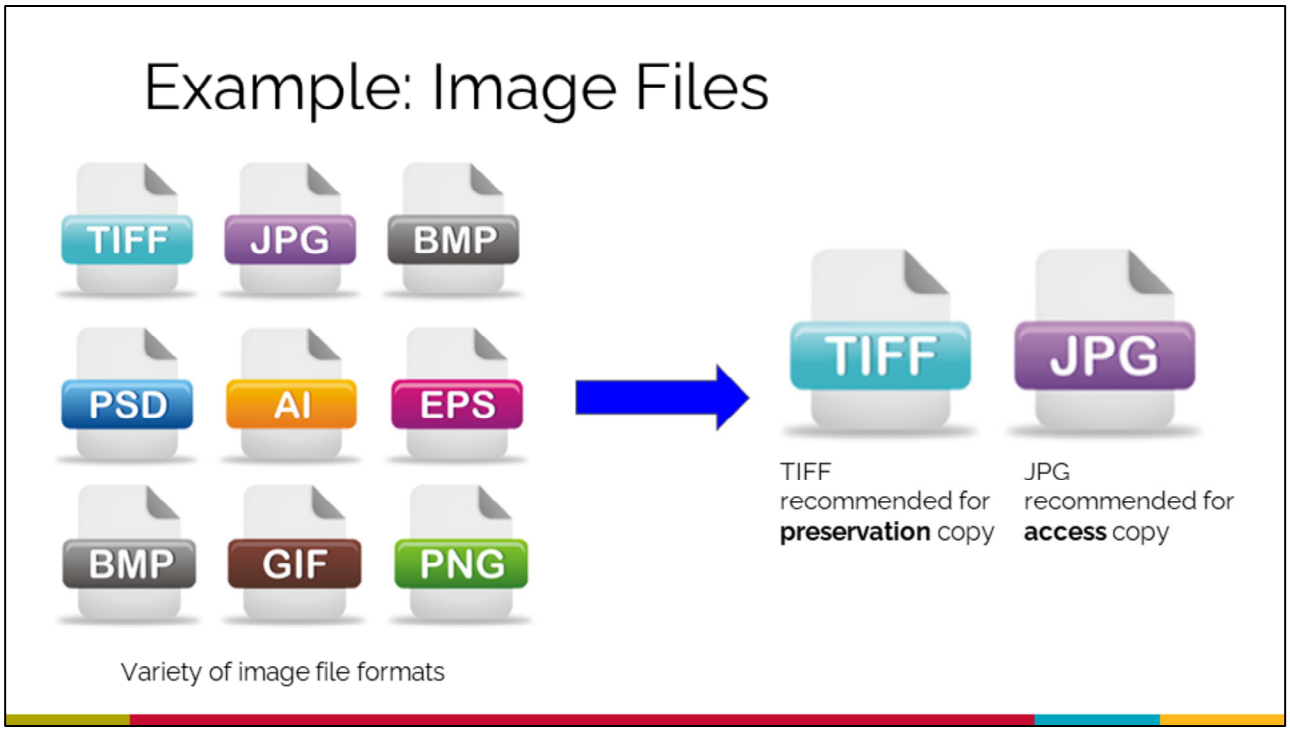
- File may come in with many formats and properties -
    - Donated and transferred files from people in your community or elsewhere
        - Working with donors becomes incredibly important. Easy to donate lots of unneeded or unorganized files. Sitting down with them and going through the files before or during donation will be very helpful.
    - Other department files that get sent to the archives
        - Work with department contacts to keep files to a standard
- You may have many options when creating digital files through capturing new audio/video/images/text, or digitizing
    - Will need to have standard ways of creating new files
- You may also find files in your own department either on a computer
- Or on media that is not regularly used and at risk (for example files that were burned on CDs) - priority to retrieve off of media and copy to a new location



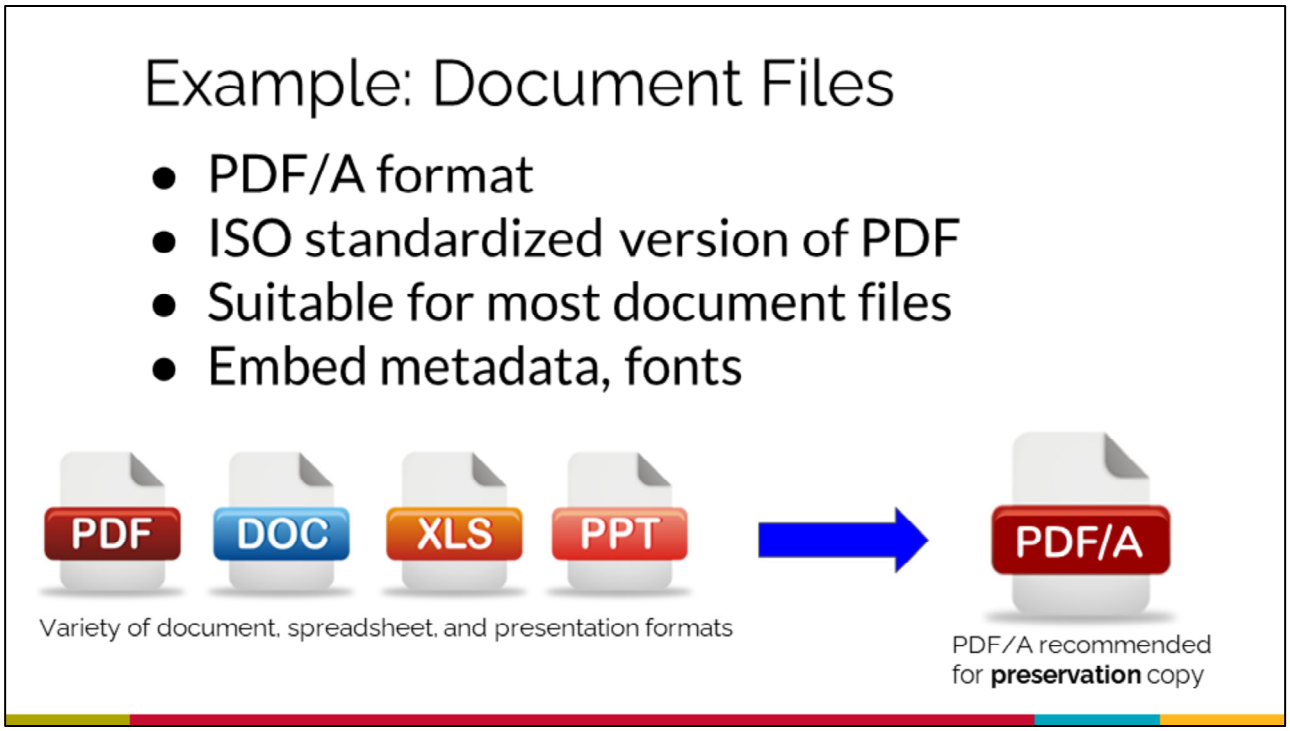
- Must take time to plan out normalization in your policies and workflows'
- May need to adjust and try different formats and different ways to get a result that works best
- As technology changes, and you and others learn more about digital preservation, it is important to be able to adapt your normalization process



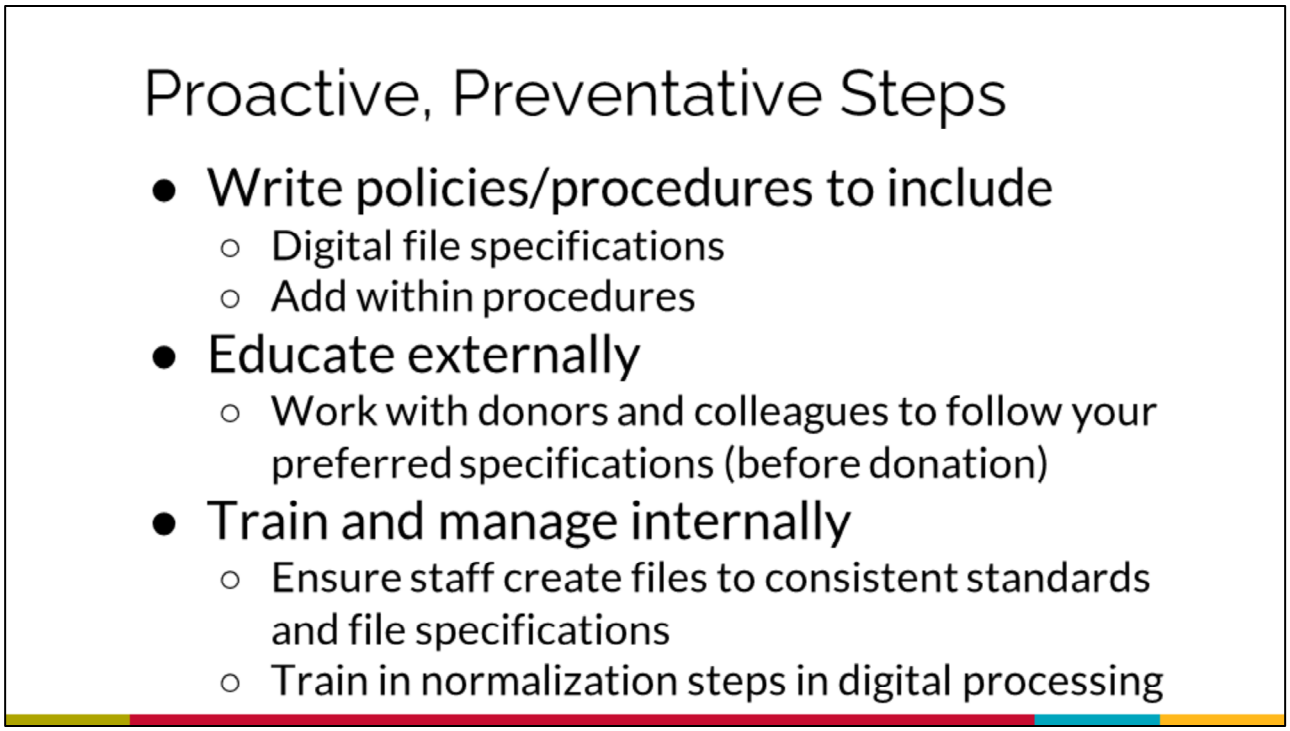
- Proprietary file types are tied to a company, organization, or person -
  - Can be CLOSED designed and encoded in a way that is not transparent or openly documented to protect trade secrets (i.e. the specifications of the file type are not publically available)
  - Proprietary formats are within the company/organizations control
  - Proprietary formats sometimes require special software to open, which might cost money
  - Can also be OPEN proprietary formats, or formats in between
  - For example if using Adobe Photoshop, you can save a .psd file - this is a POOR choice for format for digital preservation, because
- Open formats (non-proprietary) should be used in most cases
- Do some research to make sure file formats are well supported and won't be unusable any time soon
- Uncompressed file formats are *always* the preferred compression - all possible data is preserved
  - Sometimes video files are much too large to store as uncompressed, so a file format that uses *lossless* compression should be used (never lossy)
- Finally, make sure you understand the software needed to open files, render, edit, and any other necessary functions



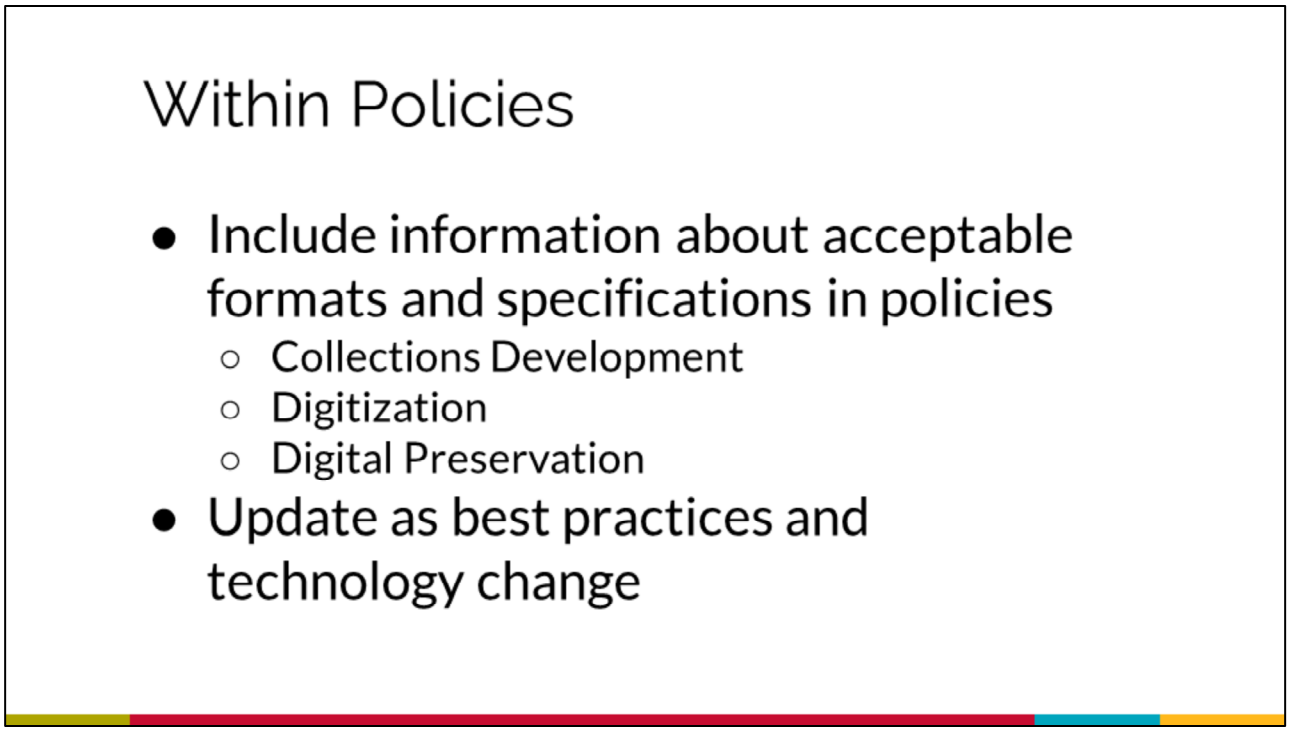
- Example
  - Image files can come in many different varieties (and often do!)
  - In a workflow at Washington State University (informed by FADGI guidelines) we would convert files to to TIFFs for preservation, then JPEGs for access
  - Always paying attention to quality standards (not just the extension, other factors)



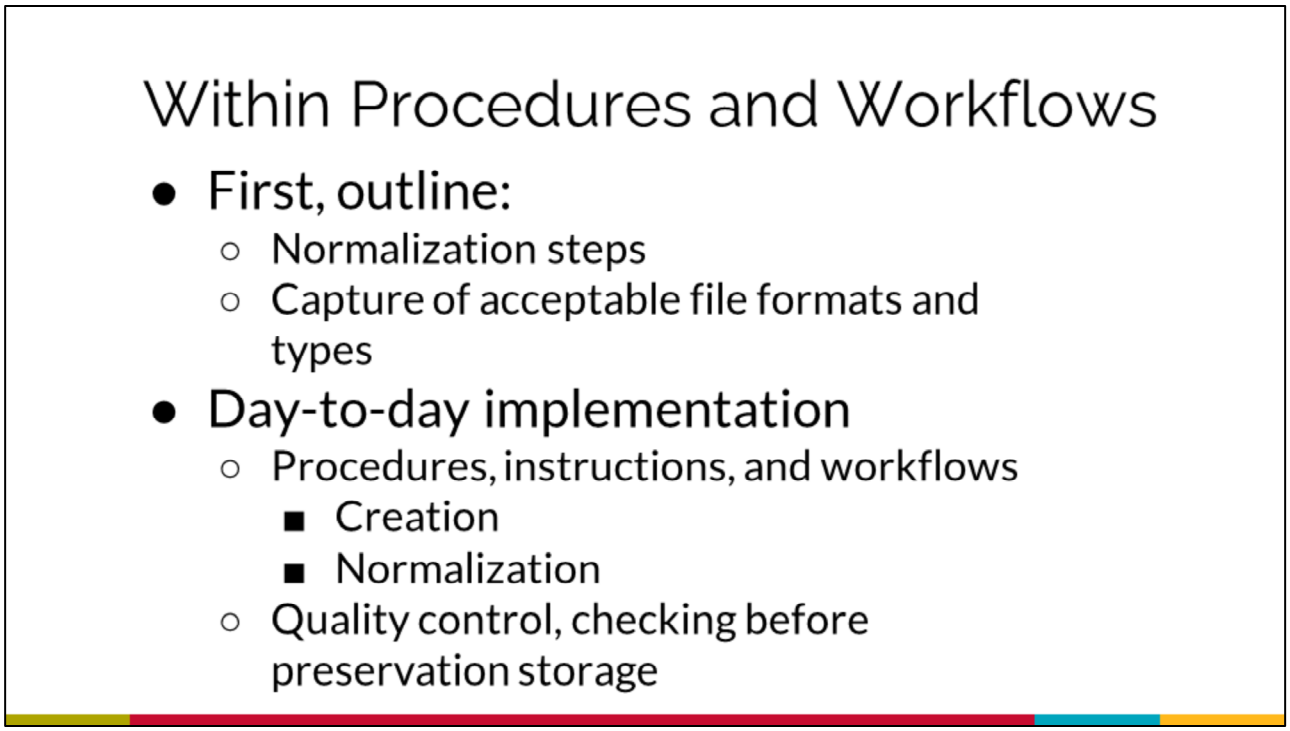
- Documents
  - PDF/A is an ISO-standardized version of the Portable Document Format (PDF) specialized for use in the archiving and long-term preservation of electronic documents. PDF/A differs from PDF by prohibiting features unsuitable for long-term archiving, such as font linking (as opposed to font embedding) and encryption.[1] The ISO requirements for PDF/A file viewers include color management guidelines, support for embedded fonts, and a user interface for reading embedded annotations.
  - PDF/A is suitable for archiving and long term preservation
  - Things that you don't need users to interact with, only view
  - Can include embedded metadata, save fonts along with files (rather than the option for linked fonts in a regular PDF)



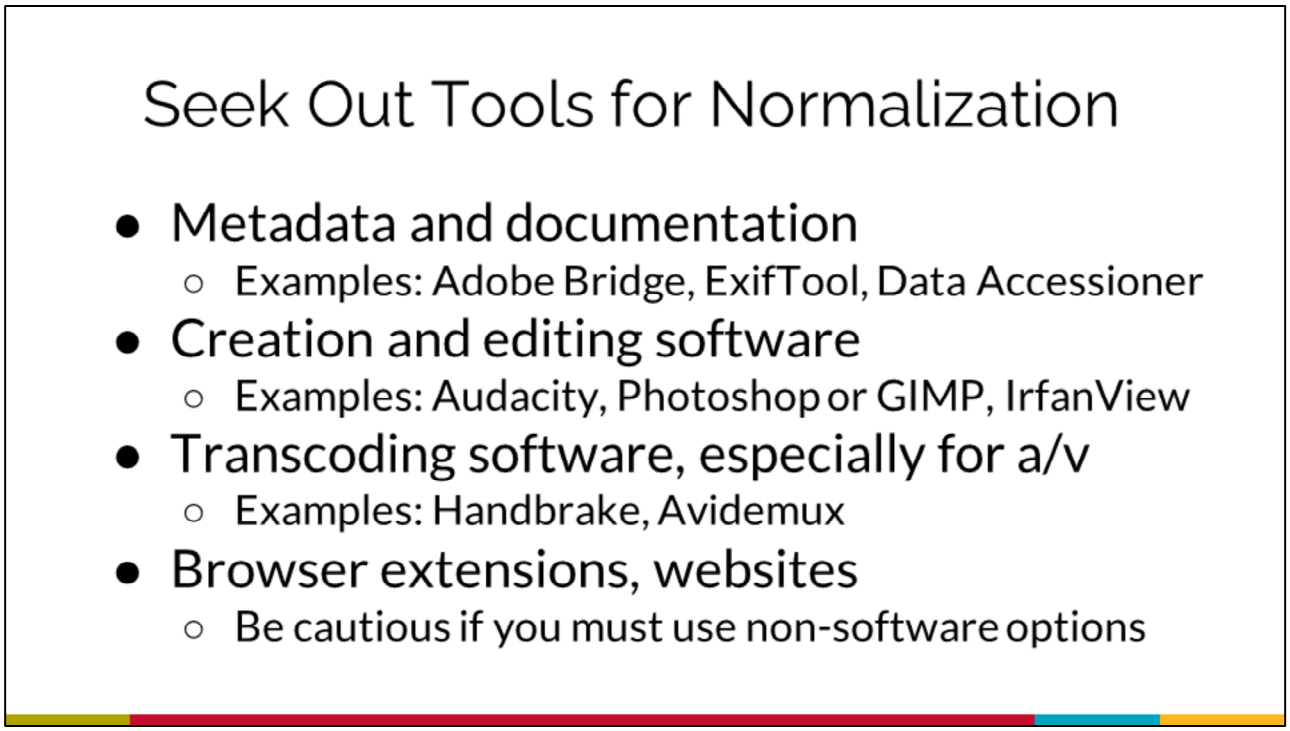
- If you can encourage use of acceptable file formats at all levels, you may decrease the files that need to be normalized
- To do these proactive steps
    - Write policies and procedures to include acceptable formats
    - Educate donors about why they should be donating their highest quality, uncompressed digital files
    - Educate internally - make sure your staff is trained
        - When creating digital files
        - Also for normalization steps before a project specifically calls for it
- However, you can't catch everything before it is created or accessioned - which is the reason for creating a normalization process



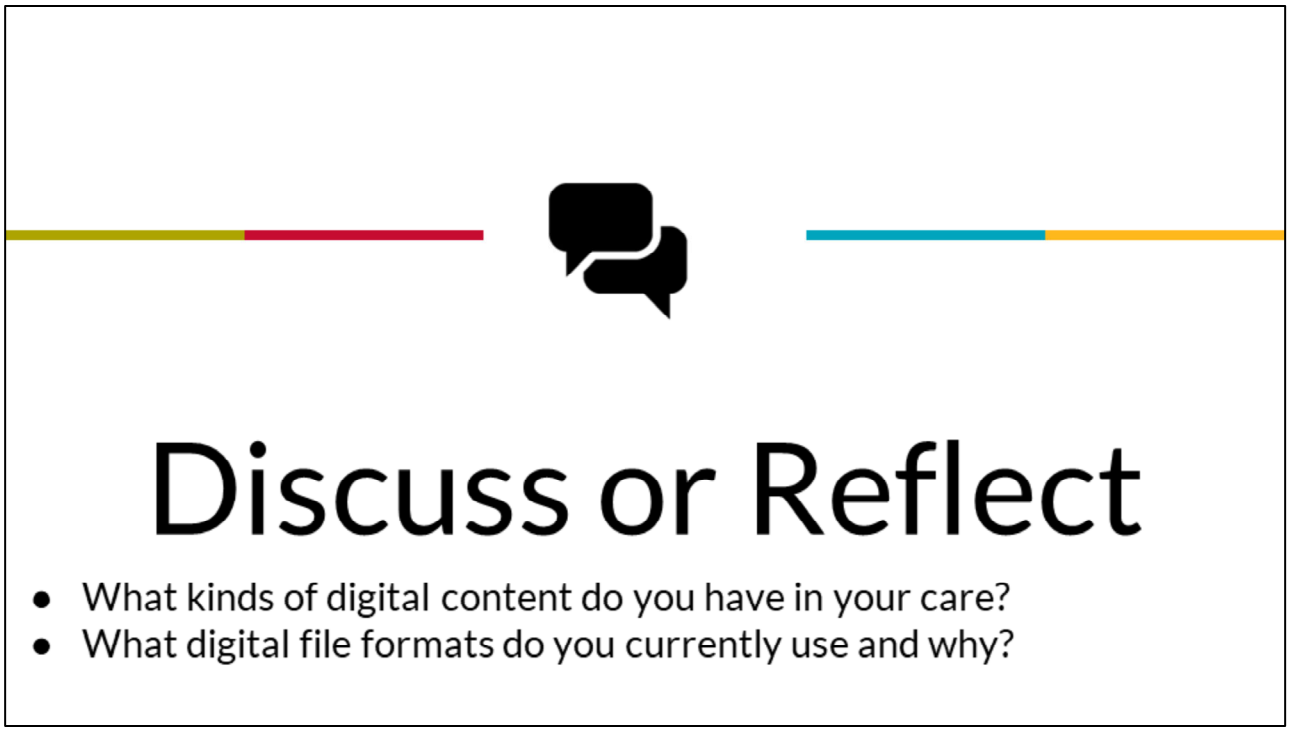
- Within policies
  - Collections development or donation guidelines - tell donors what you accept
  - Digitization - can include the specifications and standards you follow - either generally or specifically if preferred
  - Digital preservation - include information about normalization steps, file formats for preservation, versions of files, etc (again can be as general or as specific as is appropriate for your style of policy)



- Procedures are where all information about accepted file types should be documented thoroughly
- Make sure any process you want to implement is explained clearly
- Level of detail can also vary here, but your work/reasons for normalization and ways of capture should be documented
- Implementing
    - May have other instructions or training



- Tools to convert (will depend on media type)
  - Decide if you will need to be able to convert in bulk
  - Note on video - tried a few different software programs for one specific type (.webm), none of them worked! So I ended up using a browser extension



- Take 20-30 minutes and discuss with others, or reflect by yourself and take notes
- What kinds of digital content do you have in your care?
    - For example: audio files, video files, spreadsheets, word documents, outdated formats like CDs or DVDs, software or apps created specifically for your Tribe or department?
    - Do you have any concerns about any of these formats?
- What **digital file formats** do you currently use and why?
    - What are the TYPES of files
    - Can you write up a list of all the different file types within your collection?
    - Look at file extensions (the three characters after the filename - for example photograph.jpg)
    - Then think about why you have or why you create different types of files.
    - Do you follow standards? What standards?

Further Resources

- Digital POWRR https://digitalpowrr.niu.edu/
- NCDCR http://digitalpreservation.ncdcr.gov/
- The Signal blog https://blogs.loc.gov/thesignal/
- Digital Preservation Q&A https://qanda.digipres.org/
- Digital Preservation Coalition http://dcponline.org
- National Digital Stewardship Alliance http://ndsa.org
- Sustainable Heritage Network resources
  - Standards and Specifications Slides
  - Resources related to Digital Preservation Access

Credits

- Presentation template by SlidesCarnival.
- Minicons by Webalys
- *This template is free to use under Creative Commons Attribution license.*
- These slides contain changes to color scheme and content.

Using this Resource

The Digital Stewardship Curriculum is an Open Educational Resource created by the Center for Digital Scholarship and Curation.