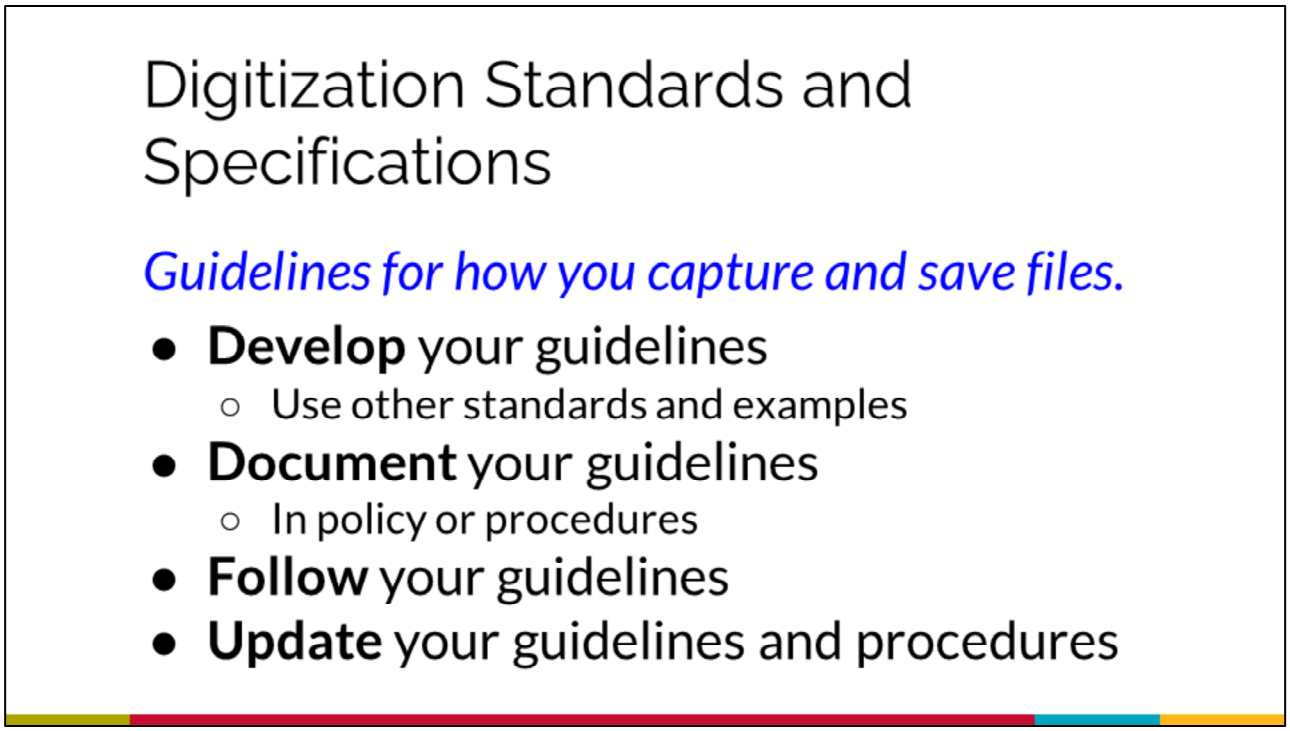
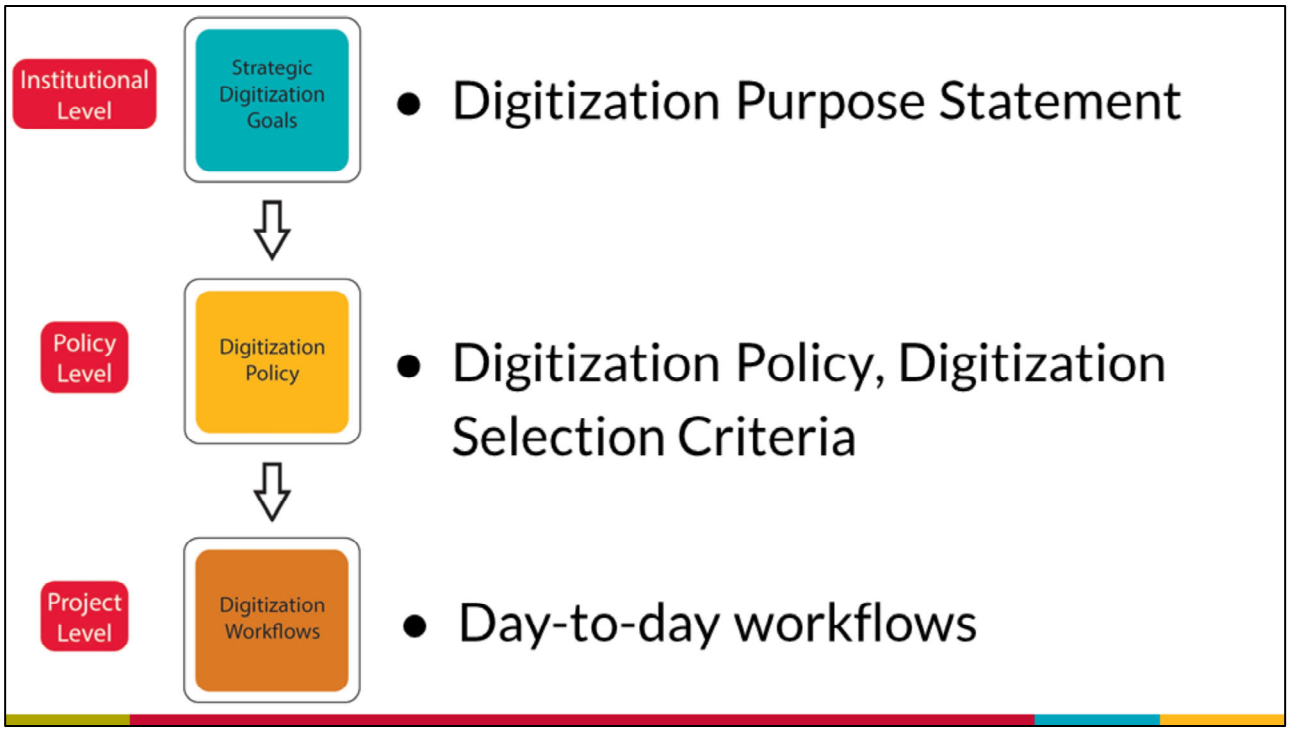
Digitization Standards and Specifications

Digital Stewardship Curriculum

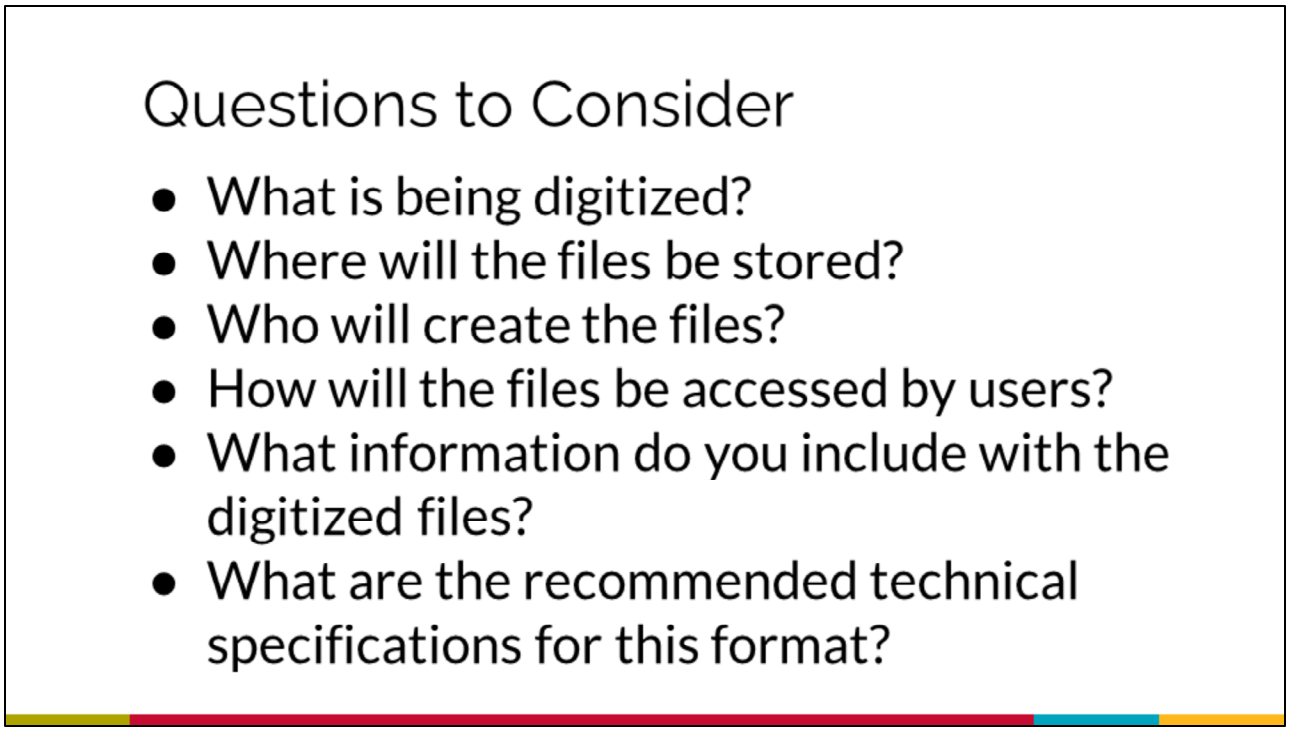


- Standards and specifications for digitization are important to include in your policies, procedures, and day-to-day documentation
    - They are GUIDELINES for how you capture and save digital files
- 4 stages to choosing and including this information in your digital projects
    - Develop guidelines using other standards and examples in the field
    - Document your guidelines in ways that are going to be helpful and useful
    - Follow guidelines - do so yourself and train staff to use them
    - Update guidelines and procedures regularly, build in time to learn about new standards and best practices, and regularly check for areas that can be updated

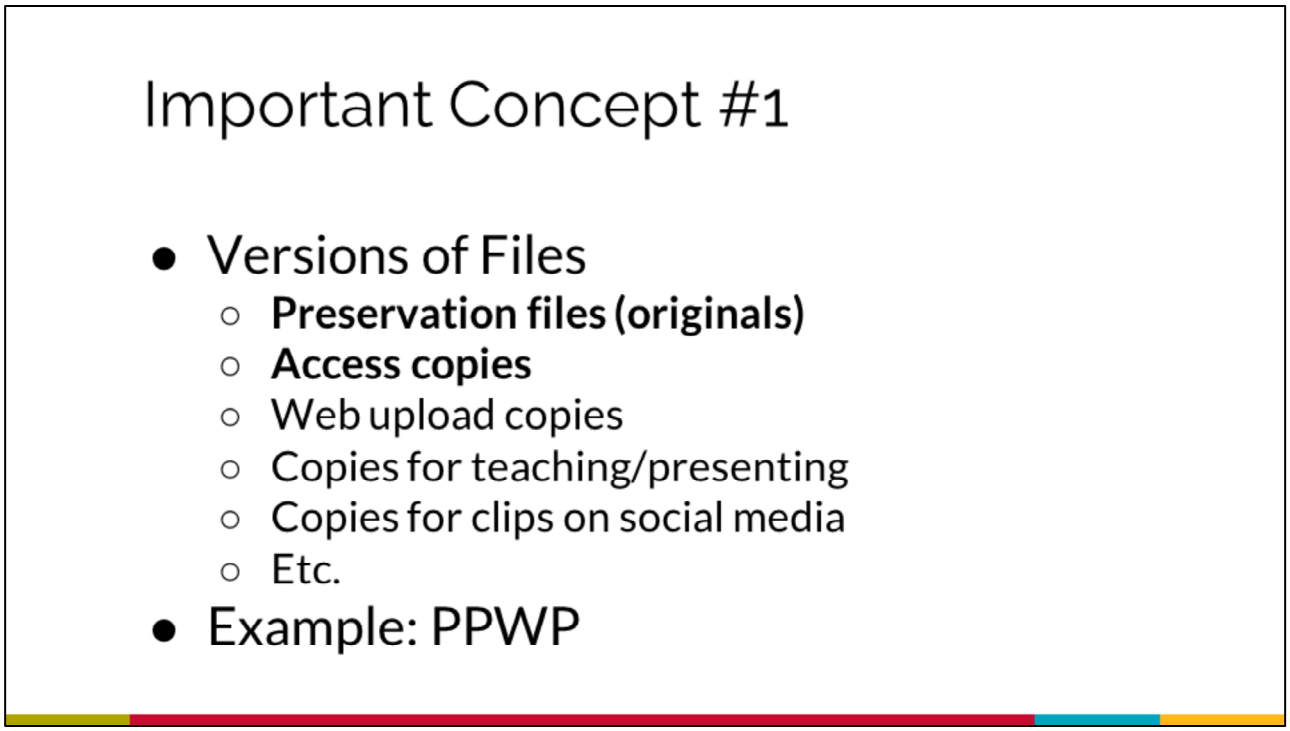
- Why use standards and Specifications?
  - Deliver same result over and over again
  - Put thought into making files helpful for your audiences
  - Use the most sustainable formats to help in preservation
  - Use formats that are SHAREABLE
- Make sure your program and project planning matches up with your TECHNICAL results. Very specific standards will help you manage temporary staff, volunteers, and others working on digital projects. Consistency over time will save reduce future metadata clean-up work



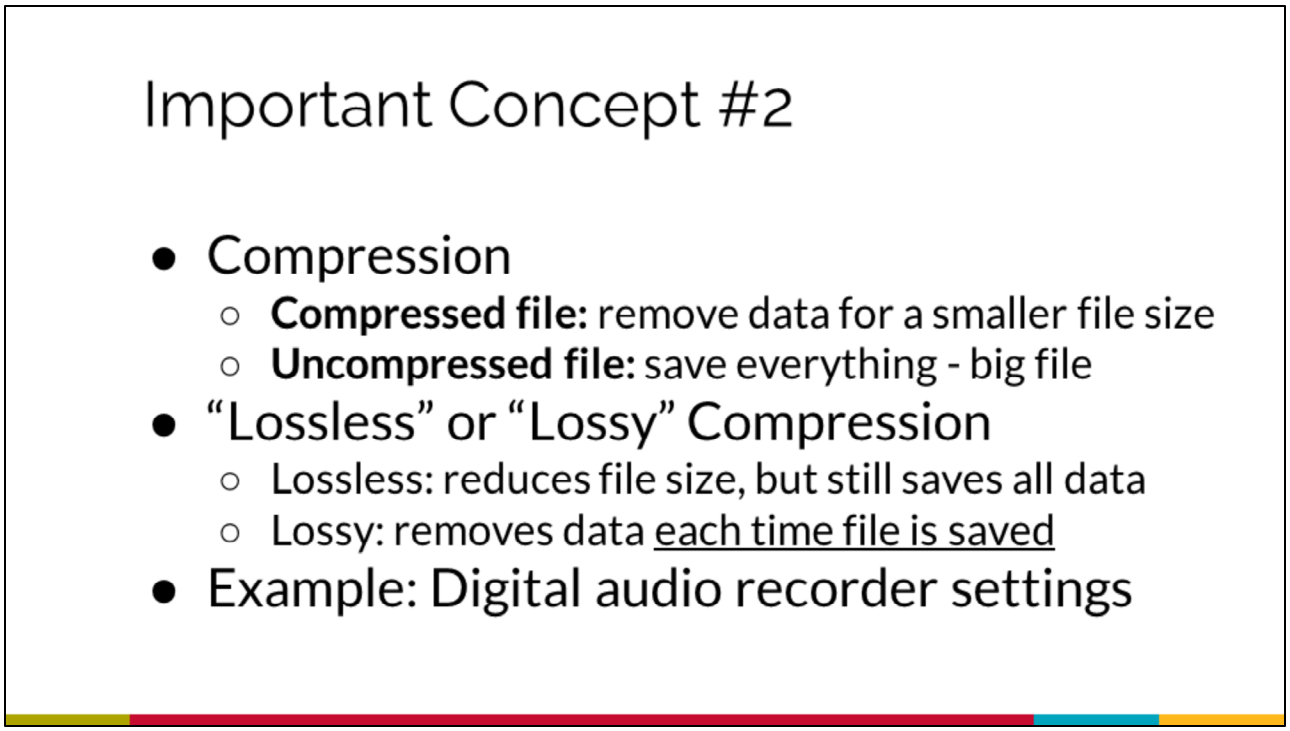
- Standards and specifications can be included at every level of your planning and work
    - Institutional level - keep it general at this level, don't need specifics - a digitization purpose statement is written at a high level - showing your vision and intention for digitization at your institution
        - However, words like "safely preserve digital content" or "share knowledge with our community" which might be in a purpose statement are based in an understanding of digitization processes and needs
    - Policy level - may need some specifics within a digitization policy or selection criteria, depending on how detailed your policy documents are
        - Here is where you might cite international standards, explain how you will carry out digitization, what best practices you will follow, and other important information - policies are a wonderful place for information on standards and specifications.
    - Project level - this level of your day-to-day workflows is where you will actually implement the standards and specifications
        - For example, by making sure the scanner settings are set correctly, or when copying and converting an original file to make an access copy with the specified file settings



- If you do not currently digitize, or are unsure of what standards, specifications, and best practices you follow, consider these questions
- The answers to all of these questions inform the decisions about how (equipment, throughput, staff, expertise, etc) and what (type of material, how much of it, how diverse it is - flatbed scanner will only work for flat, 2D paper material for instance)
    - What is being digitized?
        - What formats are you working with?
    - Where will the files be stored?
        - What is in place for digital storage, management and long term preservation?
    - Who will create the files?
        - What staff works in digitization?
    - How will the files be accessed by users?
        - How do we provide access to our community for these types of digitized files?
    - What information do you include with the digitized files?
        - What is the metadata? Do we have a specific metadata format or scheme that we use? Do we input the info into a specific platform or collections management software?
        - Seek out resources on metadata on the SHN and other sources to learn more - it absolutely ties in to your standards and specifications.
    - What are the recommended technical specifications for this format?
        - What settings should you be using when you scan a photograph, digitize an audio cassette, or convert a VHS tape? Etc.
        - Each format has special needs and considerations

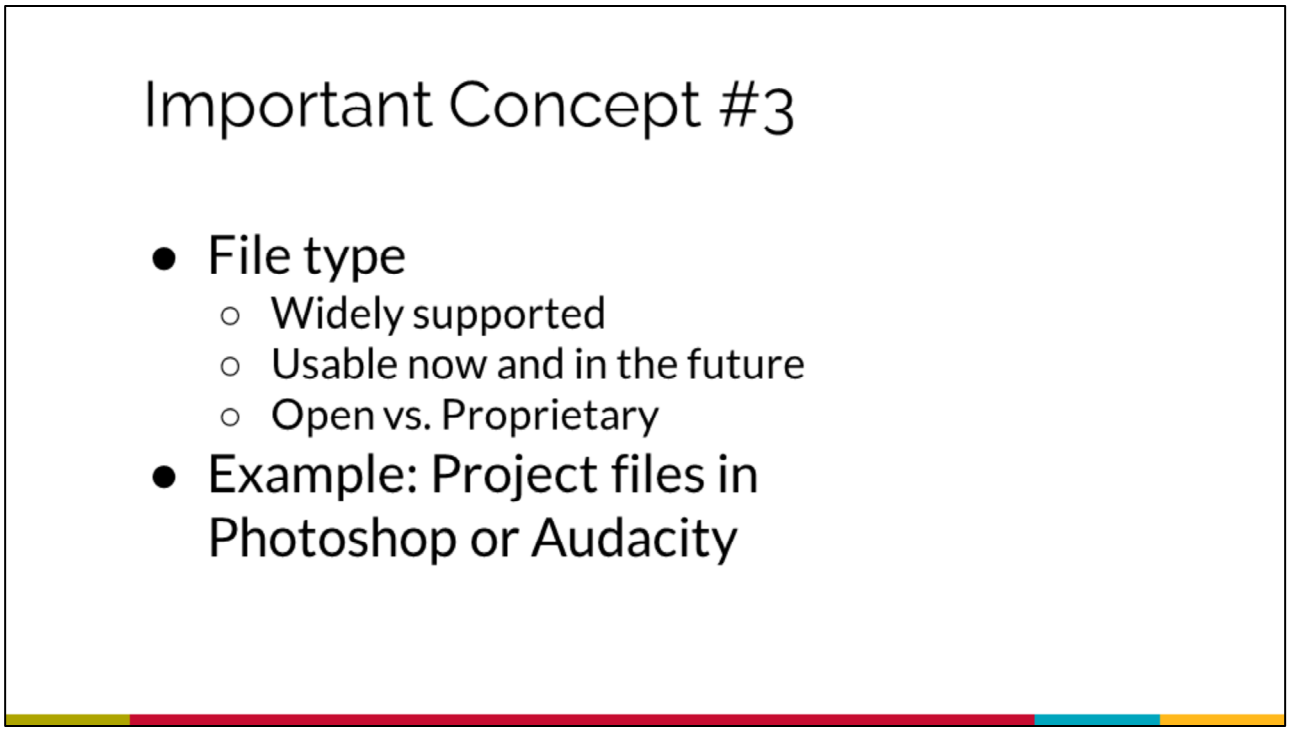


- 5 important concepts that goes for ANY digital format
- The first concept is Versions of Files
    - That you will have a Preservation or Original file that is first created
        - Highest quality you can afford to capture, large size
    - Then you might create other versions, copies of that original preservation file
        - Smaller file size, lower quality file, easier to share, upload, email, etc.
        - Access copes, web copies, many other uses
    - The different versions have different uses, for example: Preservation quality image at institution location, access copy uploaded to website, thumbnail for community on front page
- The Plateau Peoples' Web Portal is an example for the use of access copies. In this collaborative project, one partner - the WSU Archives, Manuscripts, and Special Collections digitizes and uploads files to the shared website based on the wishes of the Tribal partner representative. The WSU staff make a high quality digitized file of (for example) a photograph, make sure that is safely stored on University servers as a TIFF file, and then make a JPG copy of the file, which is smaller, and upload to the Portal website. This smaller file takes less time to upload, will take less time to load for users accessing the Portal.

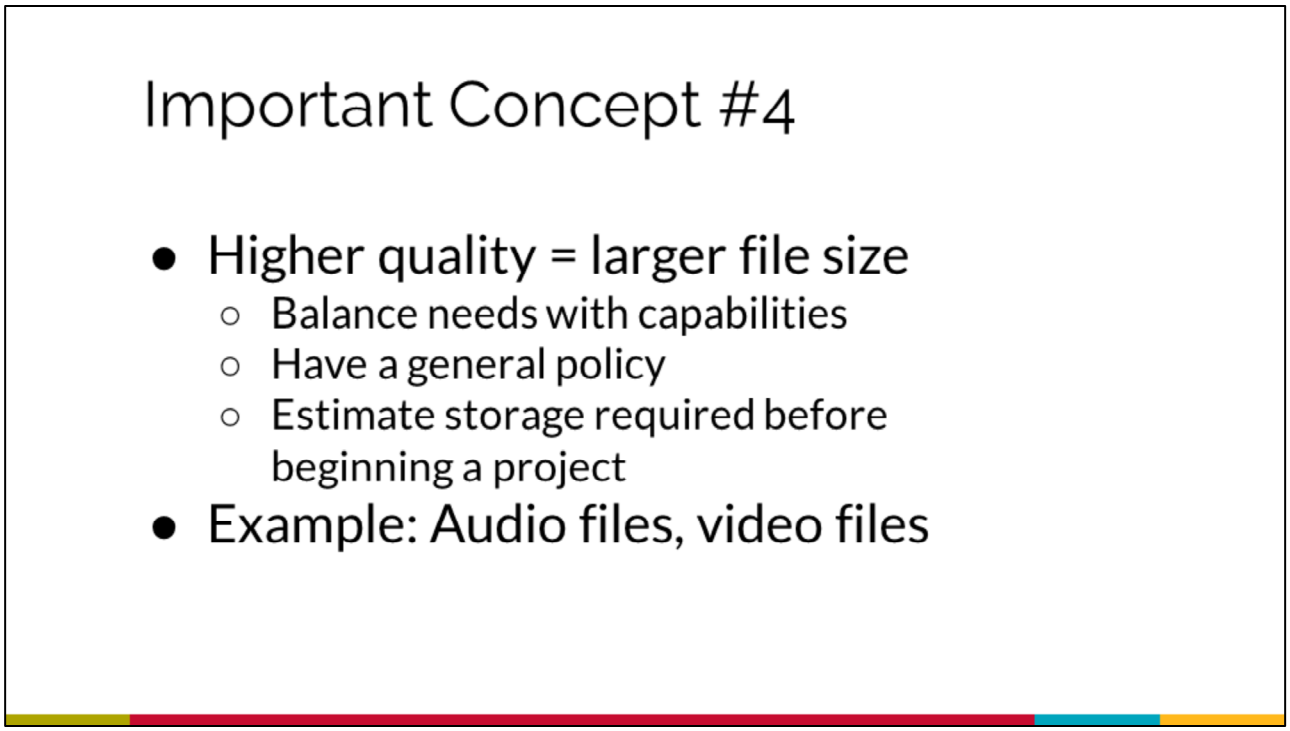


- Compression means = stripping out information to make the file smaller. Using different algorithms.
- A file can be COMPRESSED or UNCOMPRESSED.
- When saving files, you can use compressing to make files smaller. You'll want to choose if you do this for masters, access copies, and other versions of files.
- Lossless compression reduces the file size but saves all the data in the file, which is restored when the file is uncompressed.
- Lossy compression compresses files by throwing data out every time the file is saved.
    - If you copy or change a file that uses lossy compression, the file will be compressed all over again when you save it. The file size may not change much, but the quality of the file will decrease as more data is lost each time you save it.

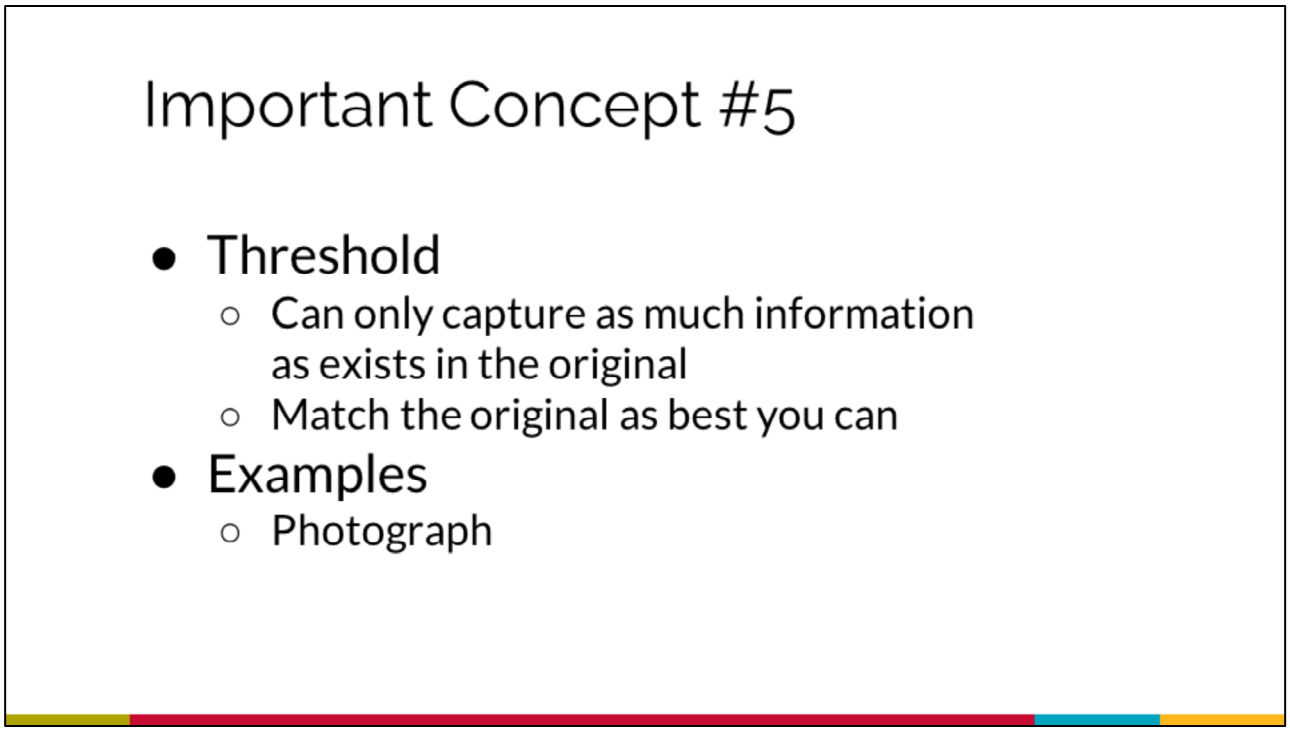
- Images: RAW, BMP, and PNG are all lossless image formats. JPEG and WebP are lossy image formats.
- Audio: WAV is a container file often used to contain lossless audio, although it is also capable of containing lossy audio. FLAC is a lossless audio format, while MP3 is a lossy audio format.
- Video: Few lossless video formats are in common consumer use, as they would result in video files taking up a huge amount of space.
    - Common formats like H.264 and H.265 are all lossy.
    - H.264 and H.265 can provide smaller files with higher qualities than previous generations of video codecs because it has a "smarter" algorithm that's better at choosing the data to throw out.



- #3 is file type
- You want to use file types that are most widely supported.
- Ask yourself: Is the format going to be openable by viewers and editors that you use now, and many years from now?
- Also ask: Will the company supporting the format go away?
- Open source is typically better, won't get tied up w/ a proprietary format. Doesn't mean you CAN'T use something proprietary - just be aware of what you are using.
    - An example of file types that would be a poor choice for preservation are Photoshop project files (.psd). Instead, export the file to something that can be opened by multiple types of programs, and is recommended within the archives field, such as TIFF JPEG2000 or PDF.



- Sometimes what you are scanning or the needs around the digital files will call for higher quality.
- For example, if you are scanning drawings and anticipate the files may be enlarged for an exhibit… a higher resolution will be called for.
- If you do need to save something at a higher quality, the file size will be larger.
- For this reason, try to balance the quality level you want to capture, with the resources that you realistically have.
- It is important to plan ahead, and estimate the storage space that you need.
- This concept is a call back to compressed and uncompressed file type, but also other QUALITY measures, depending on the format you are dealing with.
- For example, Documents and images are relatively small sizes, and you can usually afford to capture these at their highest possible quality. However, when it comes to audio sometimes, and video especially, size can quickly become a concern. Not all institutions can save a FULL uncompressed video file (.avi format for example), and will choose to save an .mp4 file, compromising on quality to be able to afford storage space.

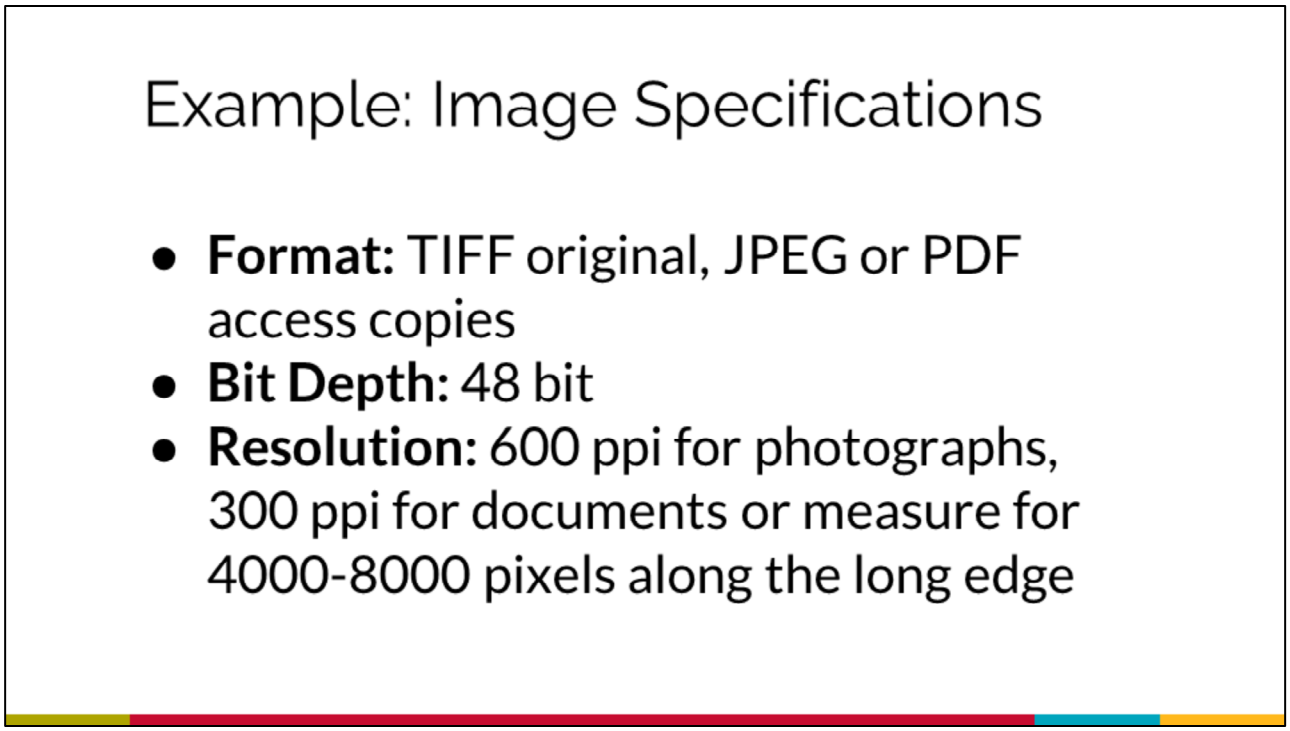


- What if the item you want to digitize is not that high quality to begin with?
- Can only capture as much as exists in the original.
- After that threshold, you run into a situation of diminishing returns, where the space needed for a larger file is not justified by the quality of the original content.
- Match your capture settings to the original material -
    - For example, if you are digitizing a clipping of a grainy and blurry photo printed in a newspaper, you will not be able to make the image look any better by capturing at a higher resolution than recommended for the size
        - See SHN resources around the topic of Image Resolution for more information
- Save as lossless or uncompressed format to preserve data - most often a good idea to use uncompressed file types.

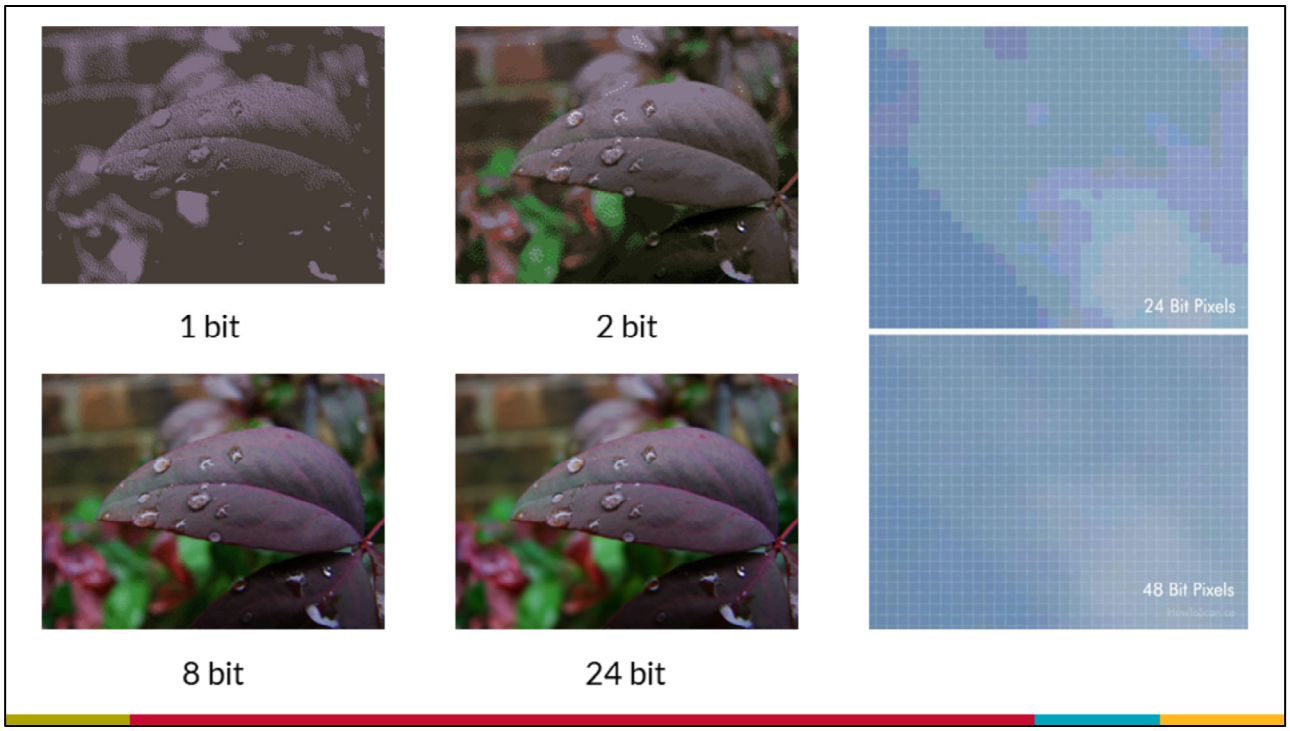
- The next slides will look specifically at image files, and how the steps of -
  - Develop guidelines
  - Document guidelines
  - Follow guidelines
  - And update your guidelines
- Can apply when you are looking at a specific format, taking into account the important concepts for ANY file type -
  - Versions of files
  - Compression
  - File type
  - With higher quality, larger size
  - And threshold

Images: Develop Your Guidelines

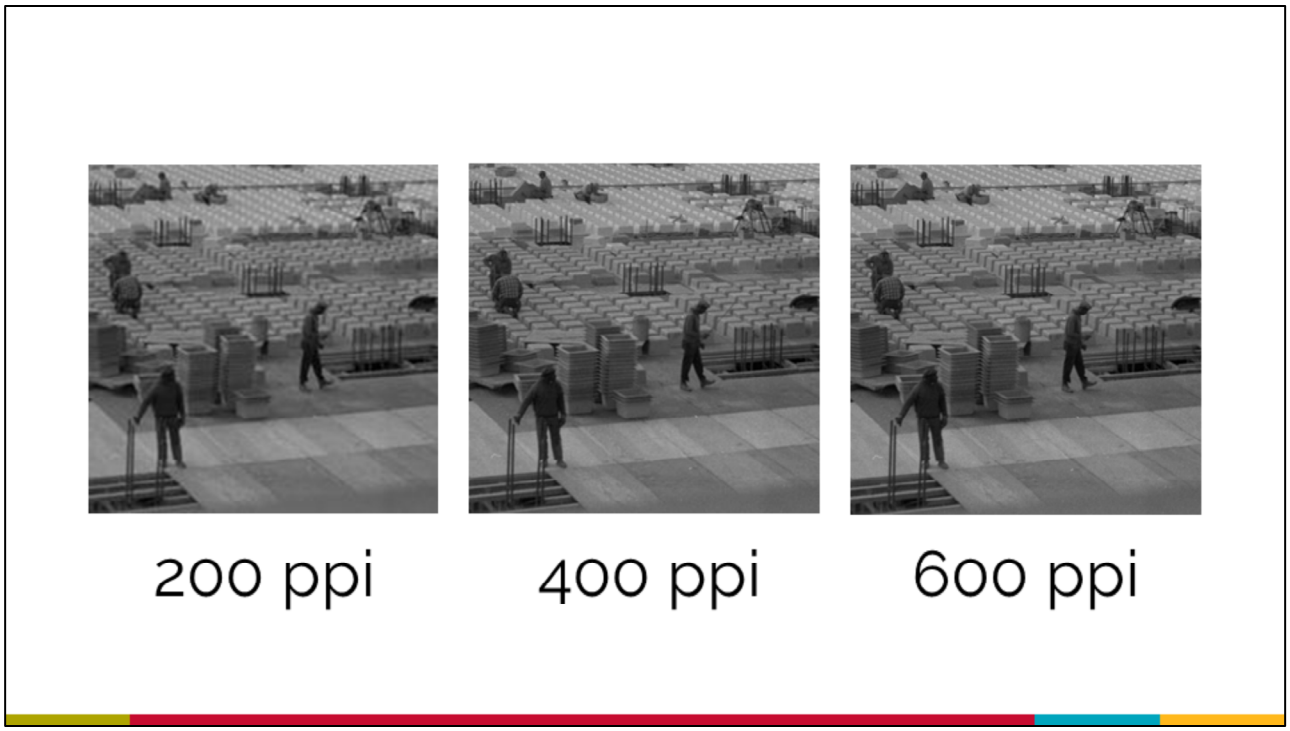
- Adapt or use existing standards and guidelines
  - Standards organizations
  - Format-specific working groups
  - Leaders in the field
- Find examples
- Adapt to your needs and goals



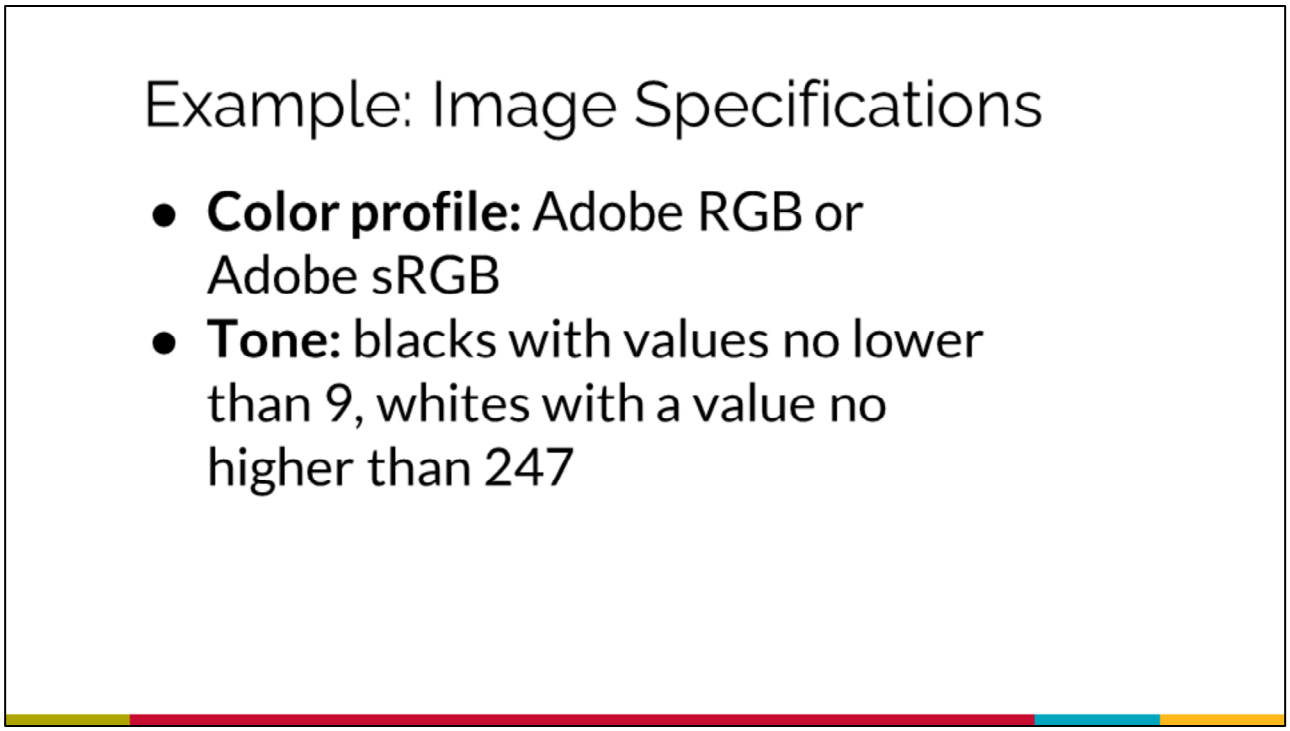
- Please reference the workshop on Image Digitization Projects on a Budget that is on the SHN under workshops - sharing the different types of factors for image quality in depth.
- Some examples of commonly used specifications for image digitization -
    - Format = type of file (save as)
        - TIFF is the UNCOMPRESSED file type, largest and highest quality
        - Some examples of other versions of your files might be access copies, or copies for printing in a newsletter (jpg or pdf might be more appropriate for these versions, depending on your project goals)
    - Bit depth = color information that the file can store (the reason a greyscale newspaper photo that might be 8bit with a few hundred values and shades, up to a 48 bit FULL color photo)
    - Resolution = number of pixels per inch - the density of pixels - this is what makes the difference between a crisp and detailed photo and a pixelated and blurry photo



- BIT DEPTH EXAMPLE -
  - The highest recommendation is at 48 bit (captures the most information)- the archival standard, but 24 bit is used widely for projects that you don't need to print, or preserve in the long term.
    - 48 bit lets you capture 281.5 trillion colors
- The images of the leaf in this slide are the same file, compressed and saved at various bit depths. You can see the differences with the naked eye here.
  - 1 bit (2 colors) - size: 4kb
  - 2 bit (4 colors) - size: 6kb
  - 4 bit (16 colors) - size: 13kb
  - 8 bit (256) - size: 37kb
  - 24 bit (16,777,216) - size: 98kb
- Then to the right, the extremely magnified image of a 24 bit image vs a 48 bit image shows that a photograph of the sky will capture better with the 48 bit, which will show smoother transition and variation between colors.
  - This can be hard to see with just your eyes, and the quality/calibration of computer display monitors also affect whether you would catch these small details.
- Images of leaf: Thegreen / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)
- 24 bit vs 48 bit image: Howtoscan.ca https://howtoscan.ca/scanning-tips/difference-between-24-bit-vs-48-bit-scans.php

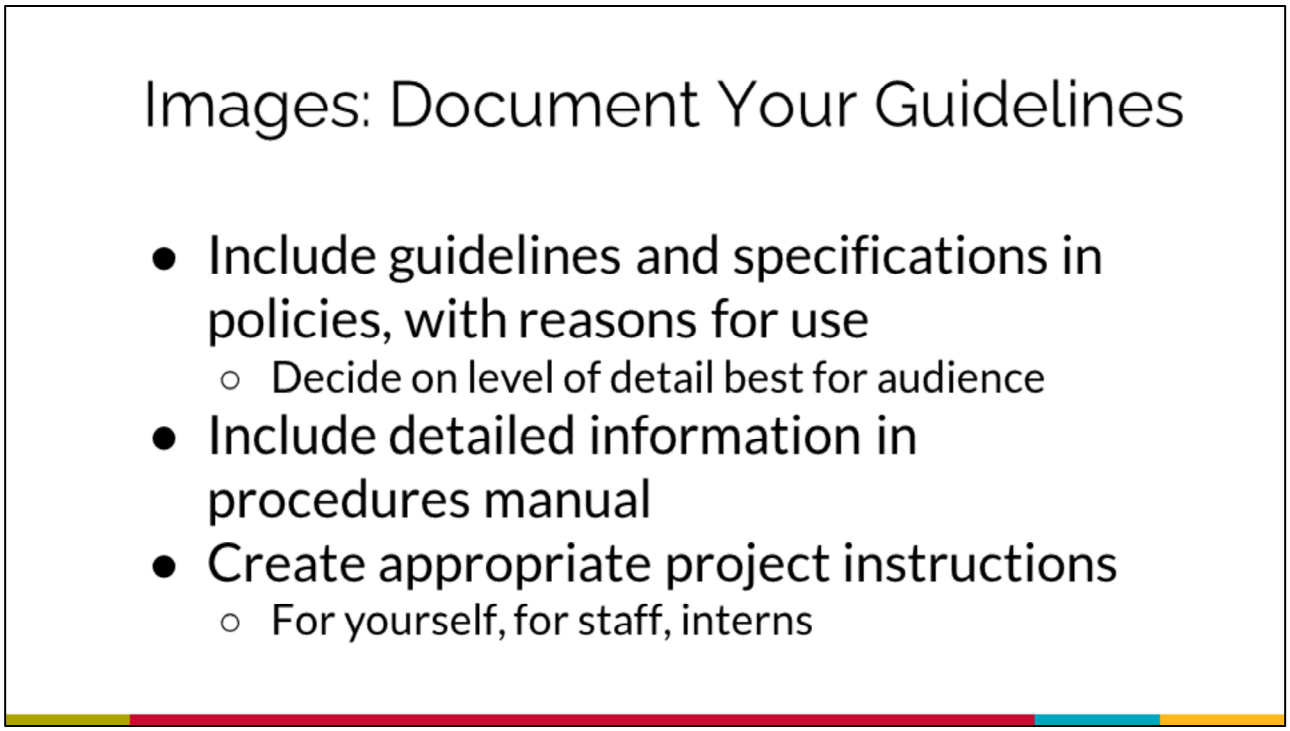


- RESOLUTION EXAMPLE -
  - The image on the slide is the exact same, except for different image resolutions. Lower pixels per inch in the two images to the left, and the highest concentrations of pixels per inch on the right.
  - The 200 ppi is the most blurry, with the most loss of detail.
  - The 400 ppi begins to have more clear details, and sharper definition between shapes and tones.
  - And the 600 ppi is the most clear, you can make out details of the people's faces, and background details.
- Image Credit: Jeanine Nault

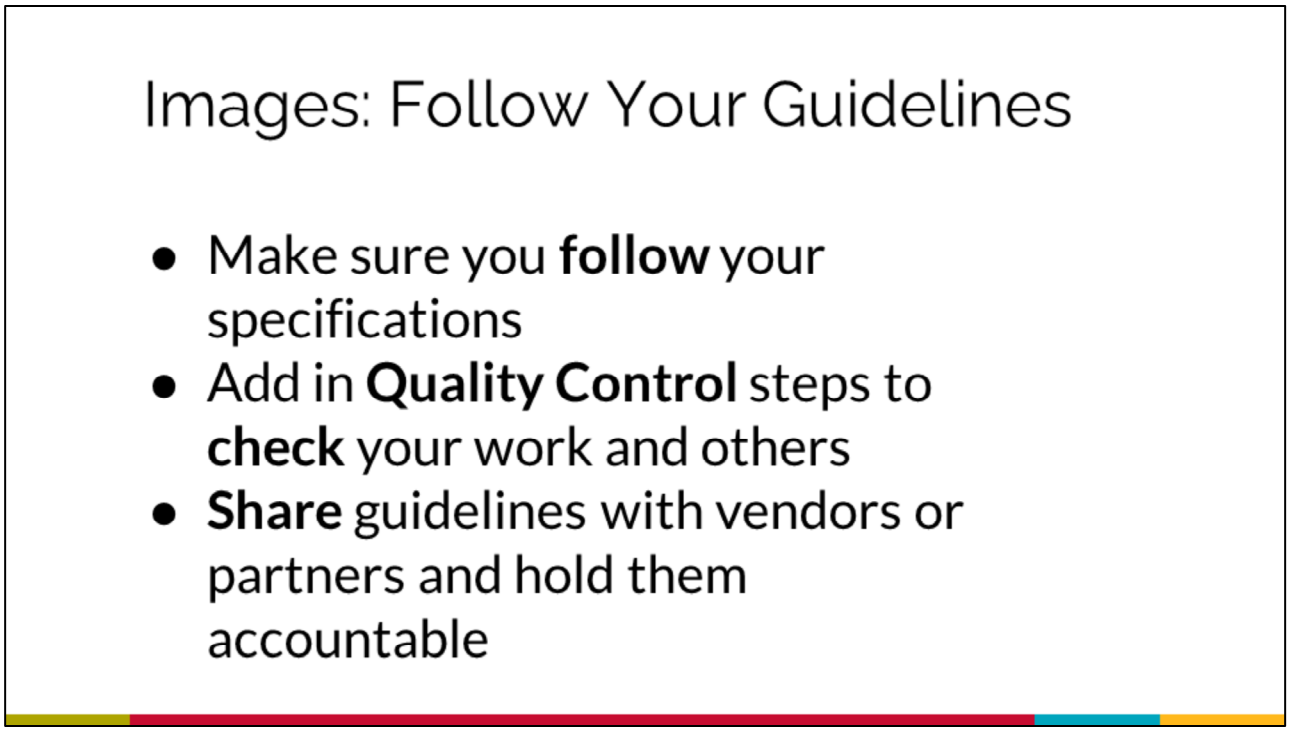


- Use widely accepted standards, these are some.
- If there is a recommended color profile in the computer system you use, software, equipment, go with that as long it is widely accepted.
    - Adobe RGB or sRGB are widely used
- Tone - basic idea is that you don't want your **darkest black to be too dark or your whitest white to be too bright** - too dark or too blown out like when you are taking a photo or capturing the image
    - You can use a color checker to include in a scan or other capture, to be able to accurately calibrate your equipment, and be able to check back that an item was scanned correctly

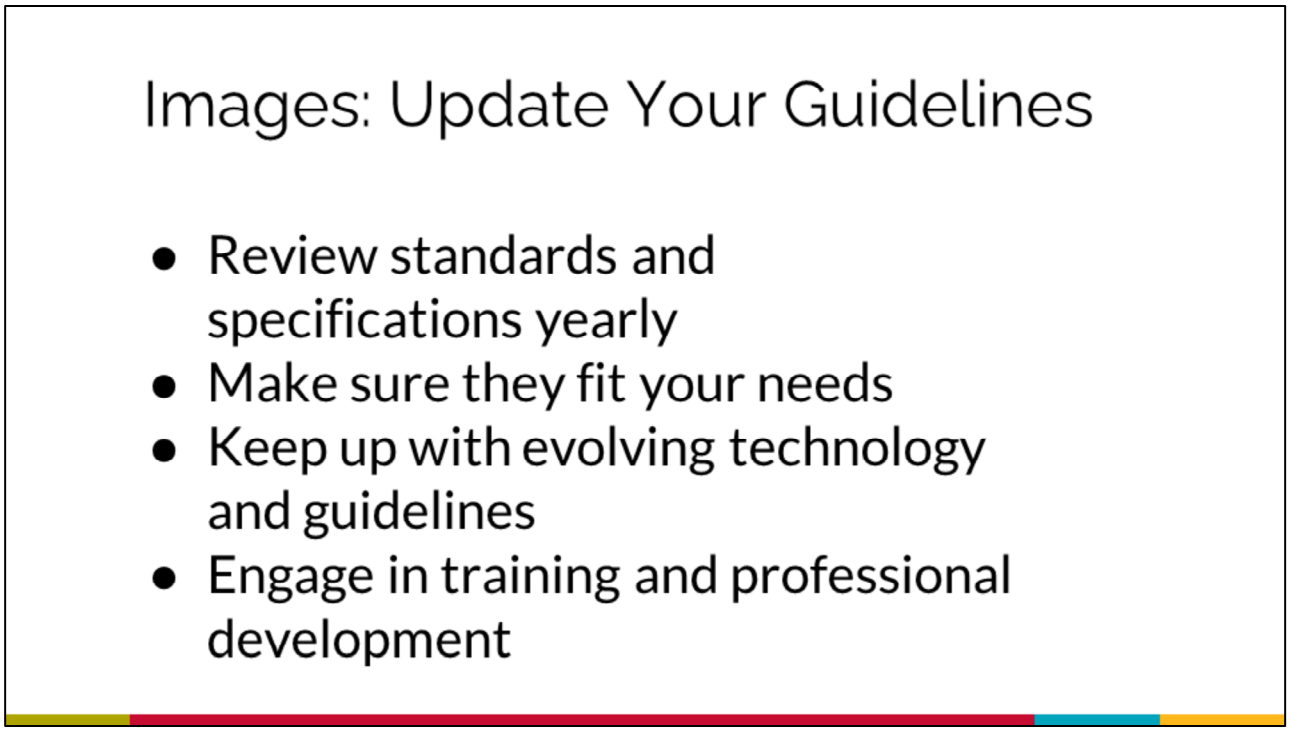
- These are some resources to use, but there are additional resources and examples available to draw from -
  - Still images working group
  - Minimum Digitization Standards
  - Sustainability of Digital Formats



- Document in your policy- explain to your audience at their appropriate level (staff may be different than supervisors or community members).
- Procedures are a good place for specifics and details (may not need the WHOLE FADGI guidelines, but could include a summary of the pieces you use).
- As well as project instructions (simplify).



- Instructions next to scanner, procedures manual?
- Quality control - check work!
- Share guidelines if you send something out



- Or every few years - have timeline built in -- you should have this for your policy already!
- Things change make sure they still work.
- Includes monitoring technology, and other guidelines that change.
- Workshops, webinars, conferences.

- Do you have any files on your computer that you can take a look at?
  - Most helpful for files that you created for digitized
  - Start with image files, since we have the examples of image specifications in slides 12 and 15
- Examine the files
  - In Windows, Right click and select Properties to see information about your files (General and Details tabs)
    - Locate the bit depth field - what is the bit depth of the file?
    - Locate the resolution field - what is the resolution of the file?
    - What other information are you noticing within this embedded metadata?

- If you have scanning software, open it up and examine the settings available
  - Can you locate where to change the file type, resolution, bit depth, and file save location?
  - What other settings do you have the option to change?
  - Do you understand what all of the settings do?
  - Are there any settings you do not use?

Credits

- Slide 4: Image created by the WSU Center for Digital Scholarship and Curation
- Slide 14 : *Images of leaf:* Thegreenj / CC BY-SA (http://creativecommons.org/licenses/by-sa/3.0/)
  24bit vs 48bit image: Howtoscan.ca
  https://howtoscan.ca/scanning-tips/difference-between-24-bit-vs-48-bit-scans.php
- Slide 15: Jeanine Nault
- *This template is free to use under Creative Commons Attribution license*.
- Presentation template by SlidesCarnival.
- Minicons by Webalys
- 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.