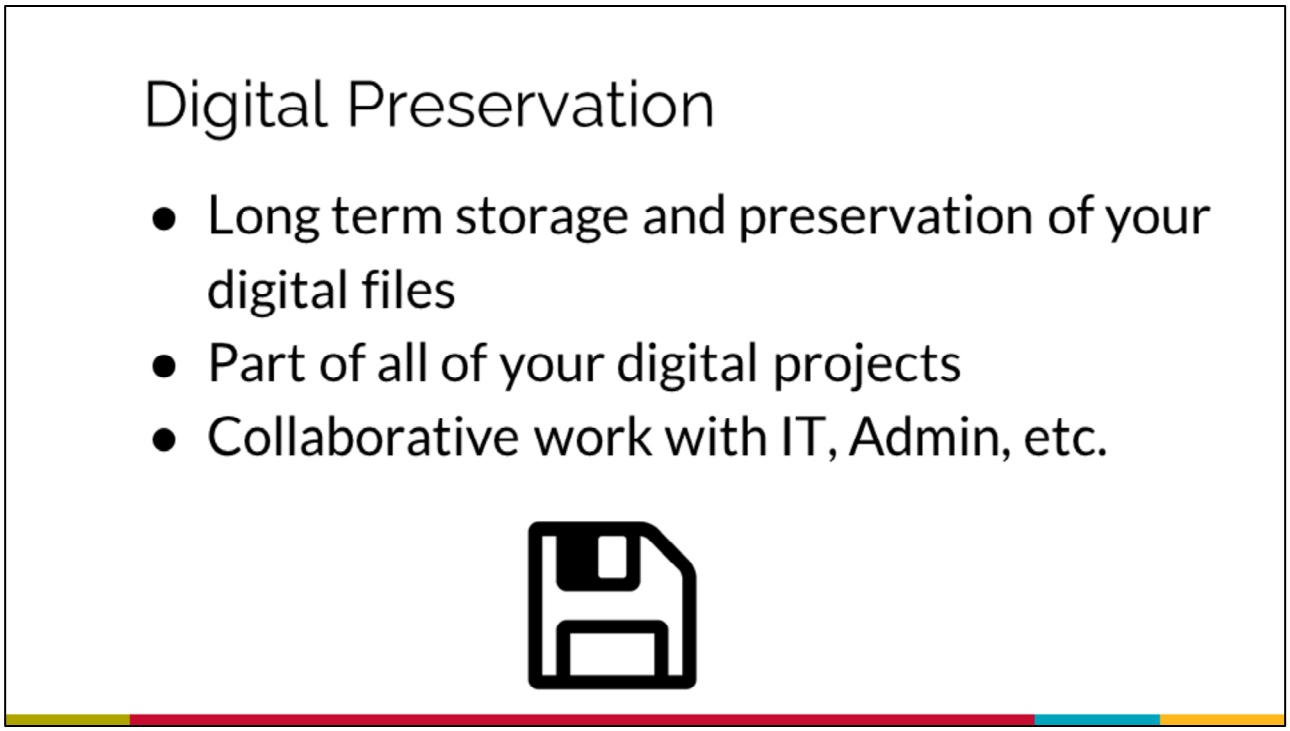
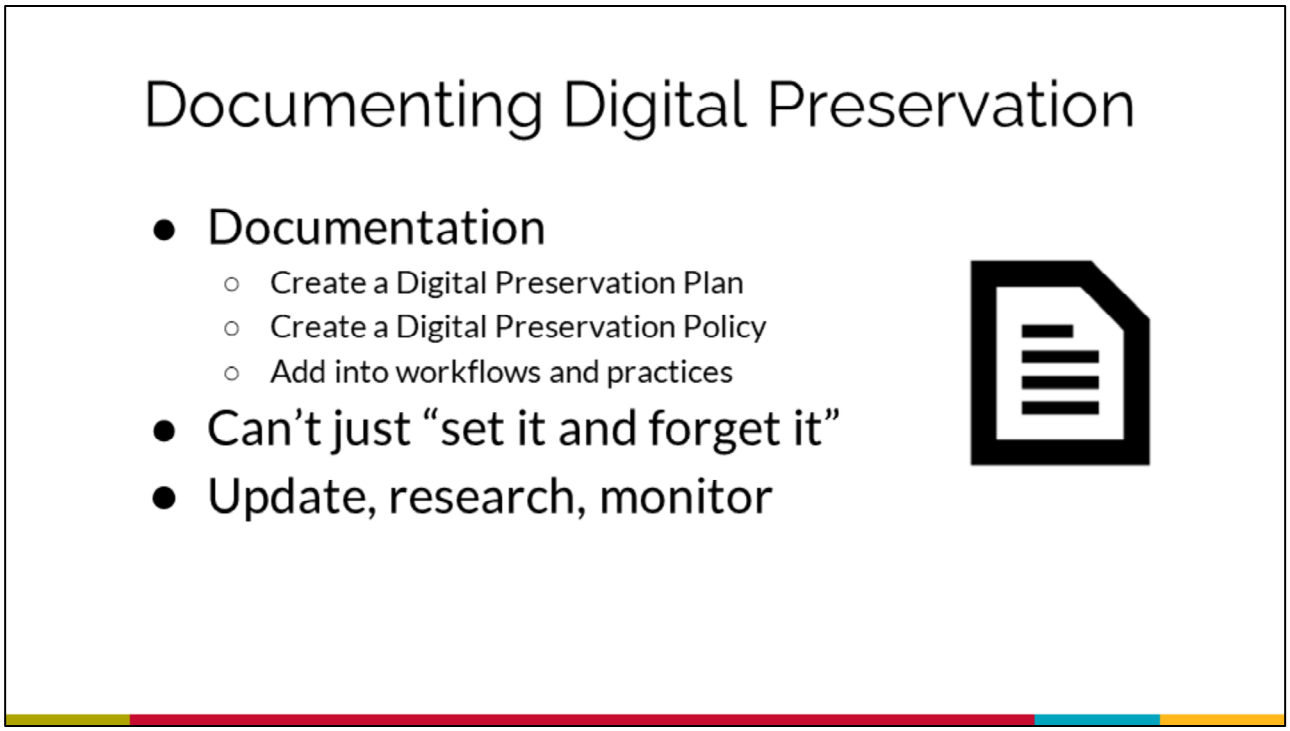
Digital Preservation: File Access

Digital Stewardship Curriculum

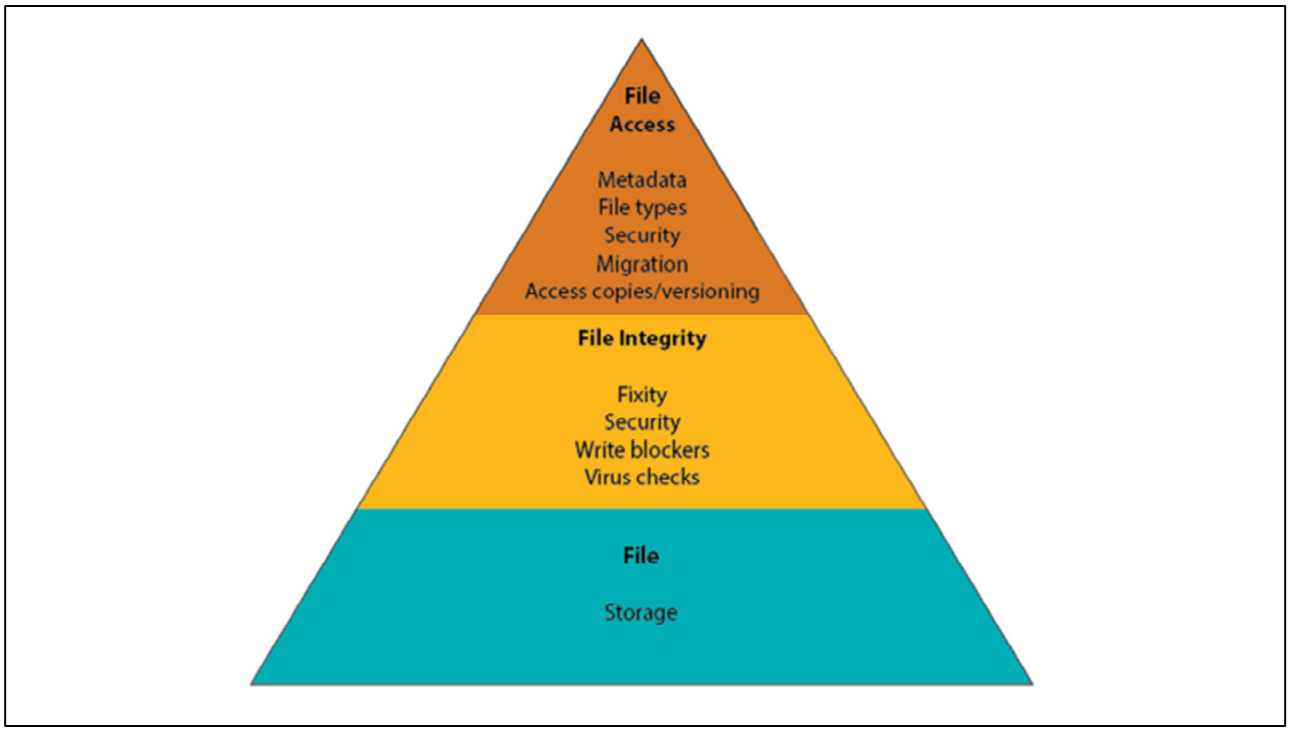
- This Digital Preservation: File Access presentation builds on the SHN resource Introduction to Digital Preservation and Storage and Digital Preservation: File Access, we recommend viewing those resources



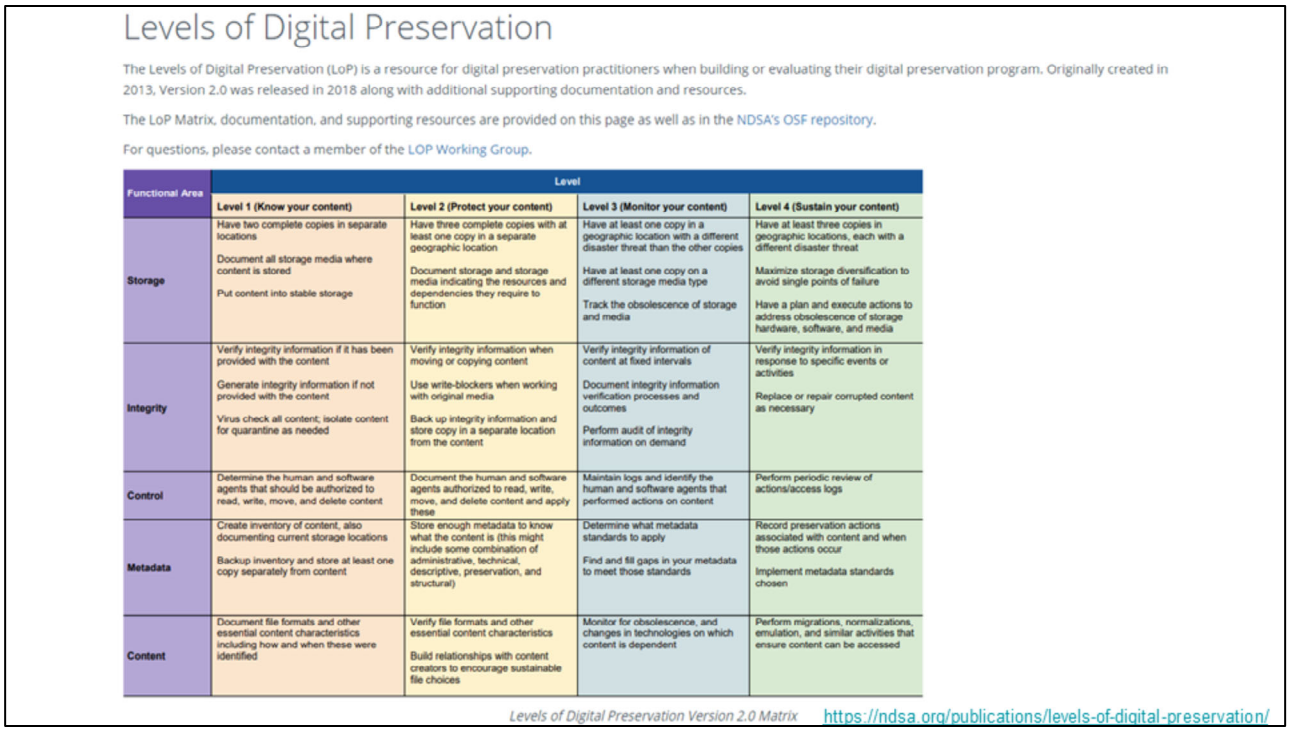
- Ensure that all the work you put into digitizing will be saved in the long term!
  - Digital preservation should be a conversation throughout your department/institution - if not, you will have to start small and keep at it
- Should be considering digital preservation with every digital project that you start



- Documentation
  - A Digital preservation plan that includes all parts of saving and preserving files, managing files, checking files - making sure it all works together and is carried out
    - See the SHN Resources *Activities to Include in a Digital Preservation Plan* and *Digital Preservation Plan Worksheet*
  - A policy is a written version of this information, that ties into institutional and departmental goals
    - See the SHN resource *Developing a Digital Preservation Policy*
  - Your workflows and practices are what gets carried out day-to-day, the information from your plan and policy must be applicable to daily/weekly/monthly/yearly actions to implement effective digital preservation
  - All of this documentation and implementation must be updated as technology and approaches change and evolve
- Similar to digitization projects - the most time goes into the planning (this planning requires time up front, but will help sustain the project)
- Your plan - not just created once and complete...like with your other policies, it must be revisited (especially with changing technology- updates)
- As hardware, software, security changes, your plan must also stay up to date



- Broke down digital preservation into 3 essential parts that any Digital Preservation Plan will have to involve
    - We have probably talked the most about File storage
    - We did start with some content on File Integrity over the last few months, starting to look into fixity and checksums, as well as security
    - File access builds on these first two sections, I will speak to these topics later in the video
- You can look at this pyramid broadly - as an outline of topics that you should plan to learn more about, address in meetings and conversations within your organization, and include in policies.
- You can also look at it more in detail, to plan for a workflow for each subtask.



Levels of Digital Preservation

The Levels of Digital Preservation (LoP) is a resource for digital preservation practitioners when building or evaluating their digital preservation program. Originally created in 2013, Version 2.0 was released in 2018 along with additional supporting documentation and resources.

The LoP Matrix, documentation, and supporting resources are provided on this page as well as in the NDSA's OSF repository.

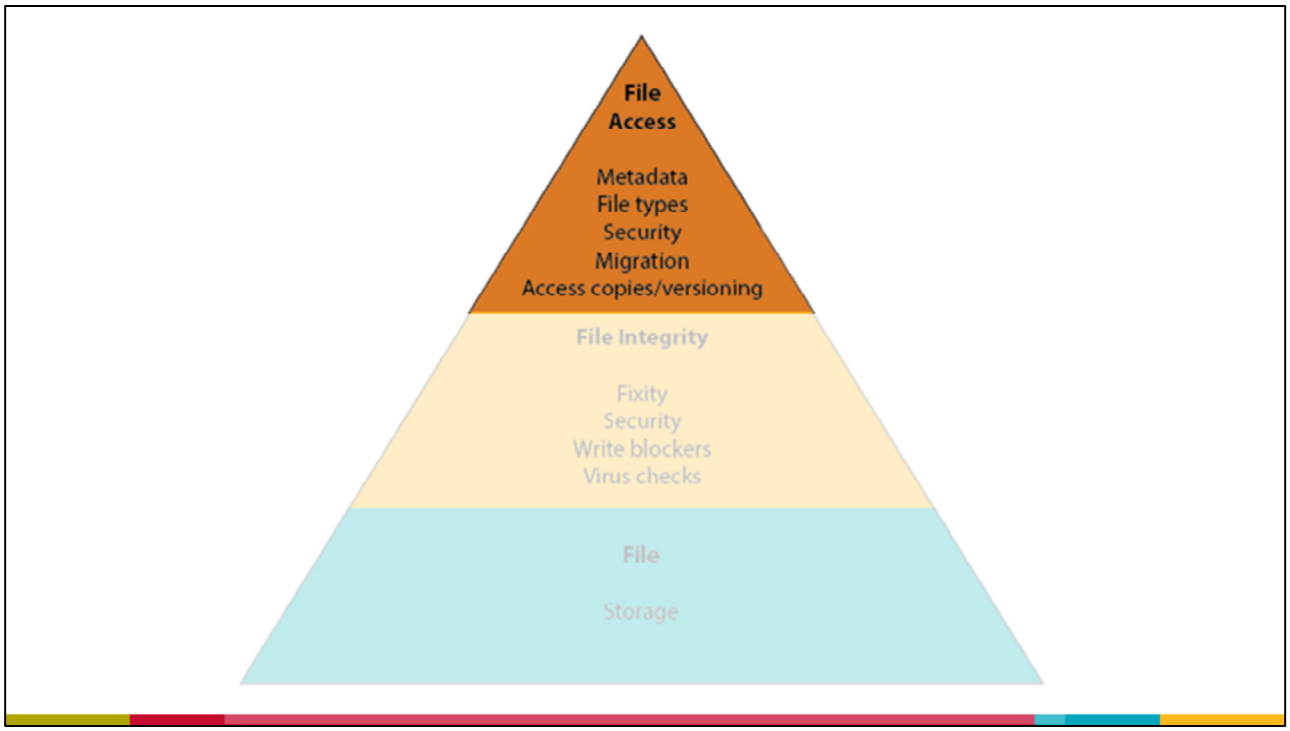
For questions, please contact a member of the LOP Working Group.

Functional Area	Level			
	Level 1 (Know your content)	**Level 2 (Protect your content)**	**Level 3 (Monitor your content)**	**Level 4 (Sustain your content)**
Storage	Have two complete copies in separate locations Document all storage media where content is stored Put content into stable storage	Have three complete copies with at least one copy in a separate geographic location Document storage and storage media indicating the resources and dependencies they require to function	Have at least one copy in a geographic location with a different disaster threat than the other copies Have at least one copy on a different storage media type Track the obsolescence of storage and media	Have at least three copies in geographic locations, each with a different disaster threat Maximize storage diversification to avoid single points of failure Have a plan and execute actions to address obsolescence of storage hardware, software, and media
Integrity	Verify integrity information if it has been provided with the content Generate integrity information if not provided with the content Virus check all content; isolate content for quarantine as needed	Verify integrity information when moving or copying content Use write-blockers when working with original media Back up integrity information and store copy in a separate location from the content	Verify integrity information of content at fixed intervals Document integrity information verification processes and outcomes Perform audit of integrity information on demand	Verify integrity information in response to specific events or activities Replace or repair corrupted content as necessary
Control	Determine the human and software agents that should be authorized to read, write, move, and delete content	Document the human and software agents authorized to read, write, move, and delete content and apply these	Maintain logs and identify the human and software agents that performed actions on content	Perform periodic review of actions/access logs
Metadata	Create inventory of content, also documenting current storage locations Backup inventory and store at least one copy separately from content	Store enough metadata to know what the content is (this might include some combination of administrative, technical, descriptive, preservation, and structural)	Determine what metadata standards to apply Find and fill gaps in your metadata to meet those standards	Record preservation actions associated with content and when those actions occur Implement metadata standards chosen
Content	Document file formats and other essential content characteristics including how and when these were identified	Verify file formats and other essential content characteristics Build relationships with content creators to encourage sustainable file choices	Monitor for obsolescence, and changes in technologies on which content is dependent	Perform migrations, normalizations, emulation, and similar activities that ensure content can be accessed

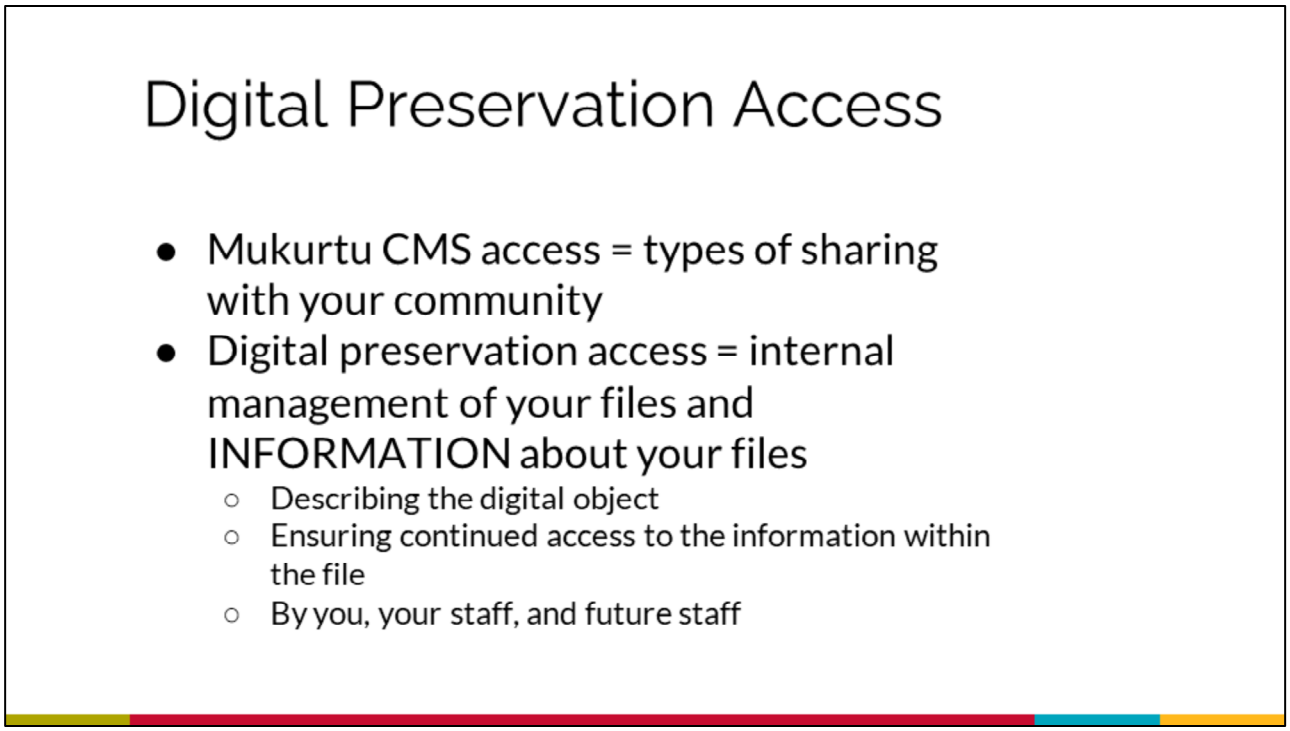
Levels of Digital Preservation Version 2.0 Matrix https://ndsa.org/publications/levels-of-digital-preservation/

- Our pyramid was based (in part) on the National Digital Stewardship Alliances Levels of Digital Preservation - here you can actually see suggested steps to add to a workflow

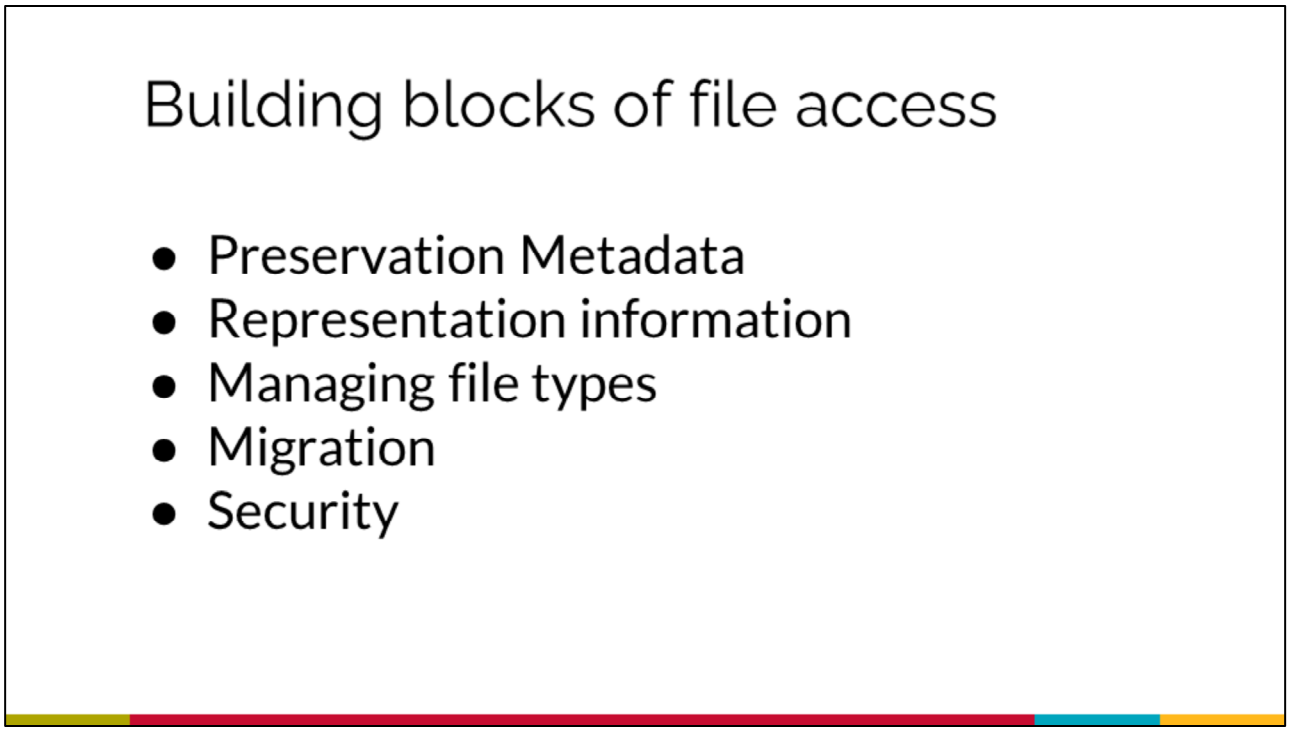
- They have this grid divided into four levels (columns)
  - Know your content
  - Protect your content
  - Monitor your content
  - Sustain your content
- You can start small and build up more and more management and preparedness
- Then on the left you can see the rows are these different “functional areas”
  - Storage
  - Integrity
  - Control
  - Metadata
  - And Content



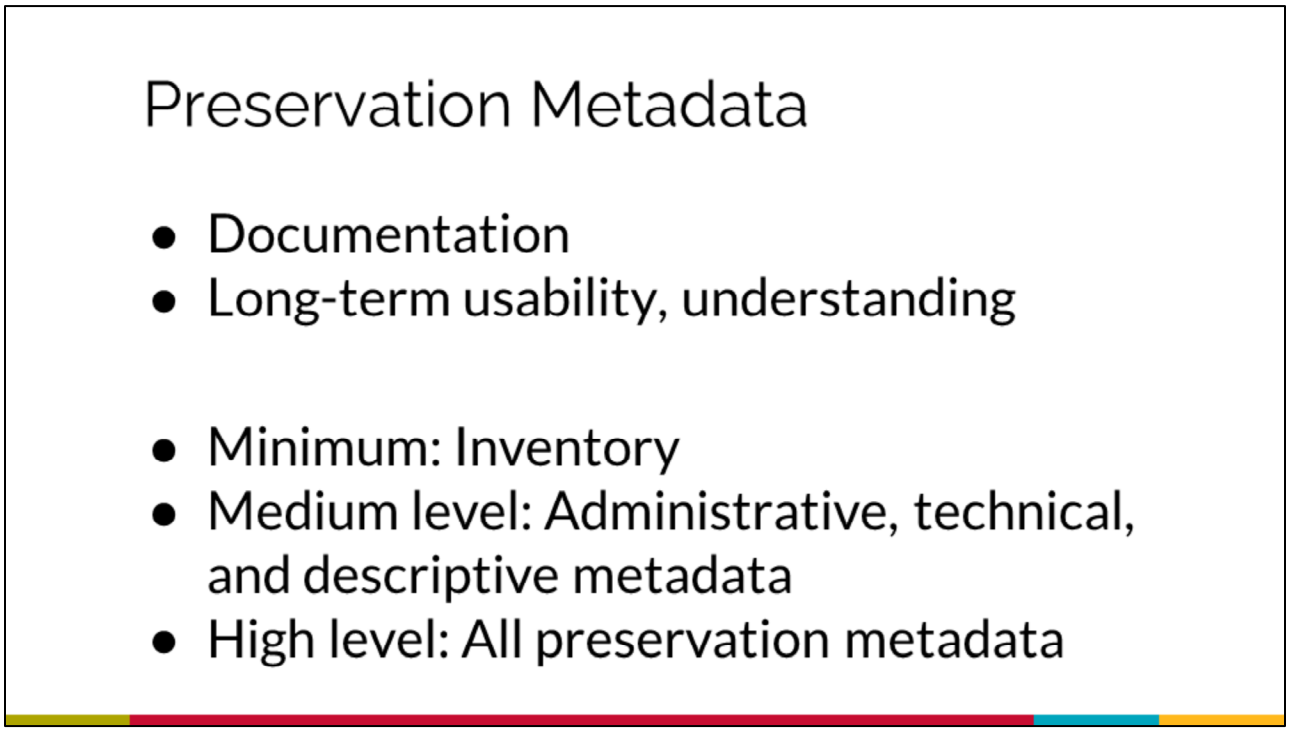
- File Access is the focus on this presentation - the top of the pyramid, all important pieces to have in place to complete your system of digital preservation and make it adaptable for what comes in the future



- First, we want to differentiate “digital preservation file access” from “access” for your community or public audiences
    - Access on a sharing platform like Mukurtu CMS for example, is to bring collections and knowledge to your stakeholders, users, and audiences to share, learn, and enjoy
    - Access for digital preservation is INTERNALLY focused - keeping constant access to the file itself, and the surrounding information or metadata
- Making sure we describe the file enough to find it again and making sure we can get at the information stored within -- whether that information is audio, visual, textual….. etc.



- These are some important pieces of files access that we will summarize in this presentation
  - Preservation metadata
  - Representation information
  - Managing file types
  - Migration
  - Security

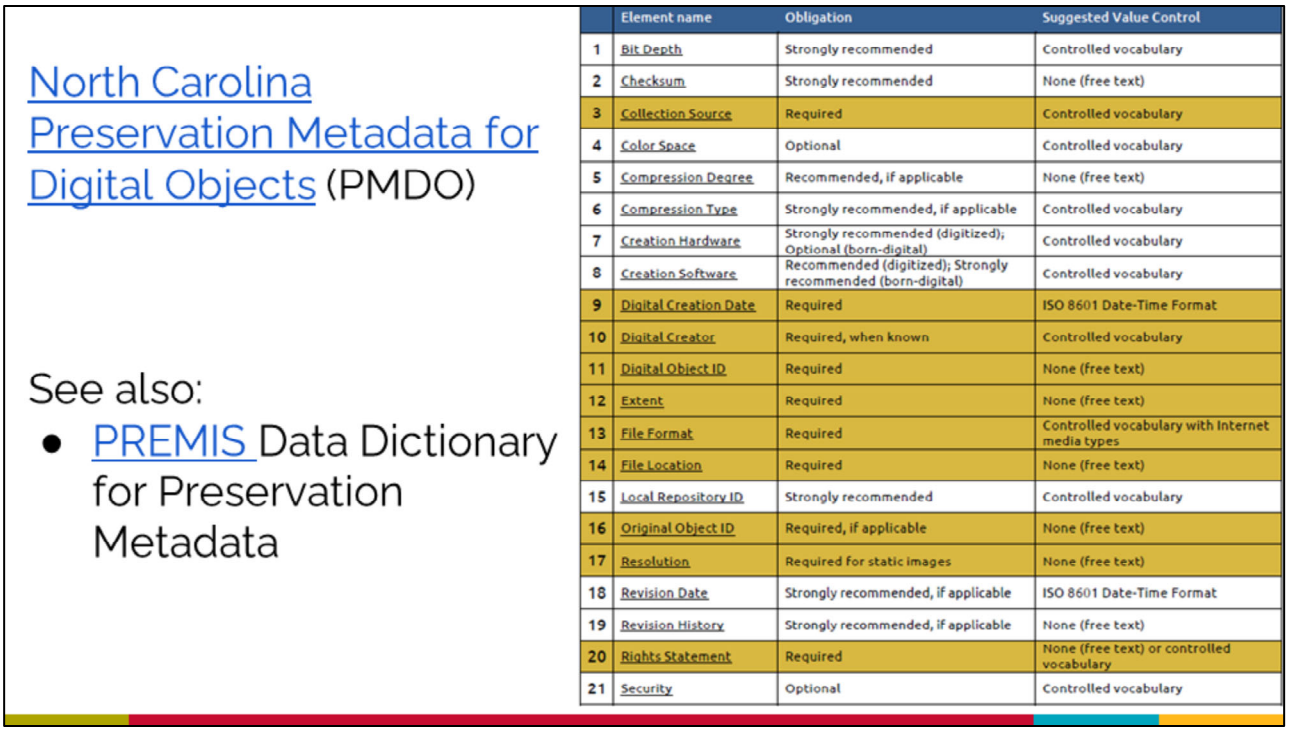


- Metadata is structured information that describes, explains, or facilitates the retrieval, use, or management of an object (either digital or physical). Wherever possible, employ embedded metadata, minimizing the chance that critical information may be separated from the files.
- Having an inventory of everything that you are preserving and where those files are located is just as important in digital preservation as it is with physical items. In most cases, if you are preserving everything that is being digitized, you can use the metadata recorded within your content management system for access as an “inventory” of digital files. This does assume that the metadata record is structured so that some piece of metadata (in many cases, the identifier) can be used to navigate the file system structure down to the unique file in the preservation filesystem or management system.
- For example, this item in the WSU Libraries content management system: (http://content.libraries.wsu.edu/cdm/compoundobject/collection/p16866coll2/id/103)
- was given an identifier, (pc156_f01), and is packed with useful information. This allows us to know not only exactly where the physical item is (pc156 indicates “Photo Collections number 156”) but that it is in folder number one (f01), and also allows us to quickly traverse to the correct file in our reservation storage system since our directory structure is a digital mirror of our physical arrangement(Archival Storage > Photo Collections > 156 > f01).
- Administrative metadata: Provides management information (eg: file type, access permissions, rights). This information may not always be displayed to the end user, but is important to effectively and appropriately manage your digital collections.
- Preservation metadata: Provides information needed to archive, preserve, and access an object.
- Technical metadata: Sometimes considered a subset of administrative and/or preservation metadata. Provides information about the technical attributes of an object, including creation and digitization (eg: hardware and software used, resolution, file formats). One common example is the EXIF metadata automatically embedded in images taken with digital cameras.
- Structural metadata: Describes arrangement of compound objects (eg: book page and chapter ordering)

North Carolina Preservation Metadata for Digital Objects (PMDO)

See also:

- PREMIS Data Dictionary for Preservation Metadata

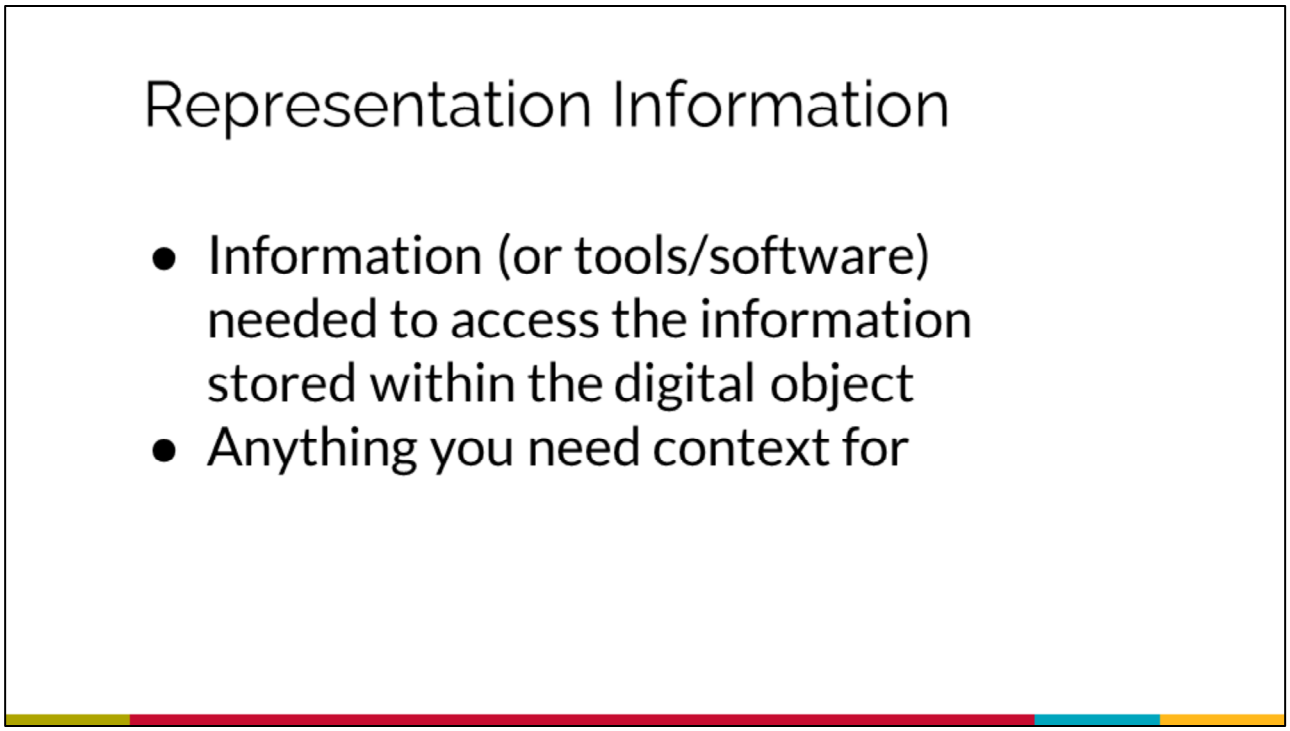


	Element name	Obligation	Suggested Value Control
1	Bit Depth	Strongly recommended	Controlled vocabulary
2	Checksum	Strongly recommended	None (free text)
3	Collection Source	Required	Controlled vocabulary
4	Color Space	Optional	Controlled vocabulary
5	Compression Degree	Recommended, if applicable	None (free text)
6	Compression Type	Strongly recommended, if applicable	Controlled vocabulary
7	Creation Hardware	Strongly recommended (digitized), Optional (born-digital)	Controlled vocabulary
8	Creation Software	Recommended (digitized); Strongly recommended (born-digital)	Controlled vocabulary
9	Digital Creation Date	Required	ISO 8601 Date-Time Format
10	Digital Creator	Required, when known	Controlled vocabulary
11	Digital Object ID	Required	None (free text)
12	Extent	Required	None (free text)
13	File Format	Required	Controlled vocabulary with Internet media types
14	File Location	Required	None (free text)
15	Local Repository ID	Strongly recommended	Controlled vocabulary
16	Original Object ID	Required, if applicable	None (free text)
17	Resolution	Required for static images	None (free text)
18	Revision Date	Strongly recommended, if applicable	ISO 8601 Date-Time Format
19	Revision History	Strongly recommended, if applicable	None (free text)
20	Rights Statement	Required	None (free text) or controlled vocabulary
21	Security	Optional	Controlled vocabulary

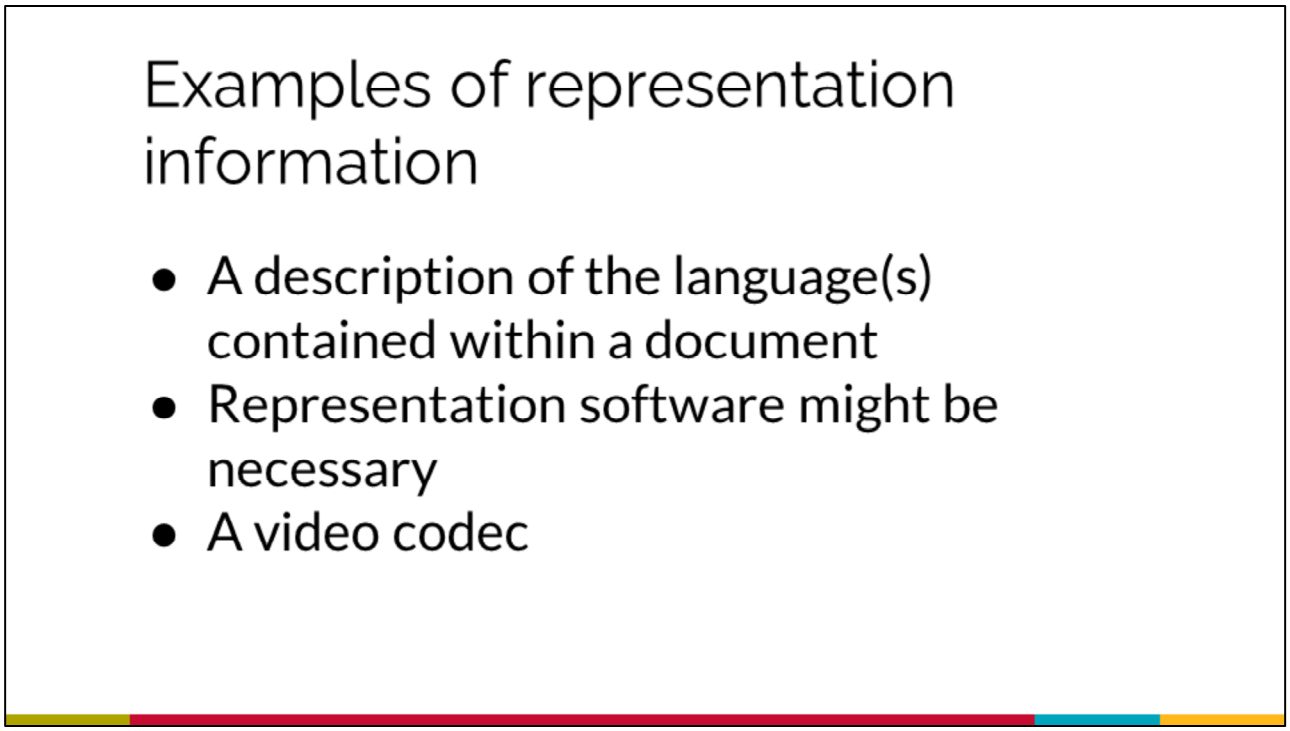
- This example contains some required and suggested fields for preservation metadata from the North Carolina Preservation Metadata for Digital Objects (PMDO), where this State Library created guidelines for mid-sized institutions

- From the NCDCR site: http://digitalpreservation.ncdcr.gov/PMDO.html
  - **Purpose**
  - The Preservation Metadata for Digital Objects (PMDO) contains a list of preservation metadata fields
  - that institutions creating or caring for digital objects might consider recording to help in the longterm access and management of their collections. Elements are mapped to the broader and more
  - comprehensive PREMIS Data Dictionary for Preservation Metadata, to ensure interoperability and
  - Shareability

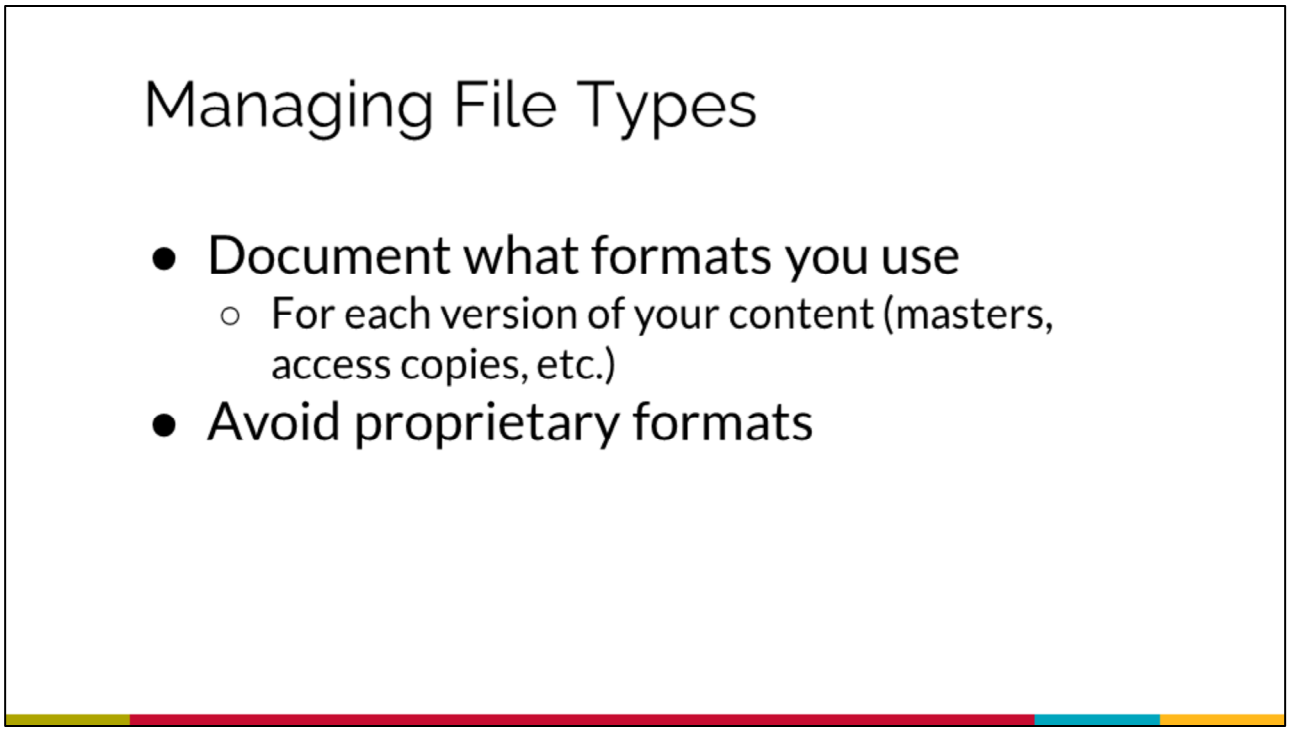
- See also:
  - PREMIS Data Dictionary for Preservation Metadata http://www.loc.gov/standards/premis/
  - North Carolina Digital Preservation Program Overview https://files.nc.gov/dncr-archives/documents/files/20190422_DNCR_preservationguidelines_final.pdf



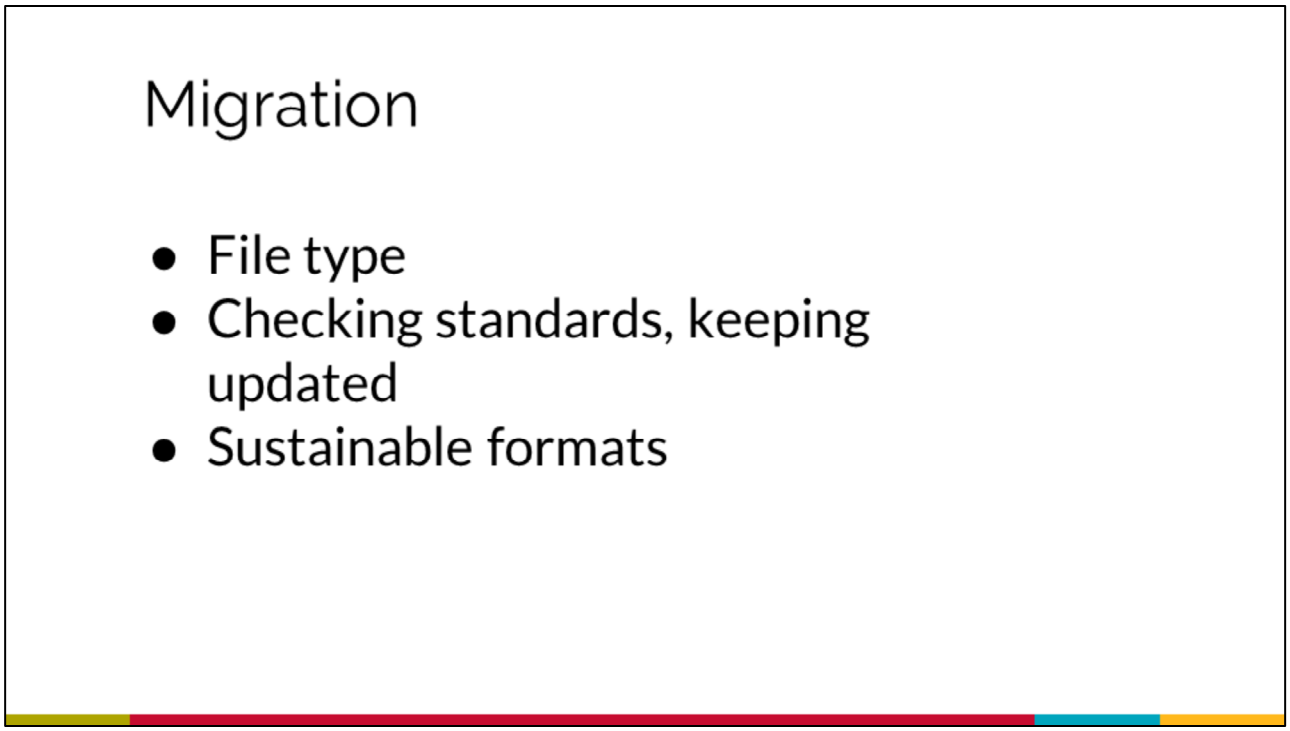
- This can be the most esoteric topics in Digital Preservation. The idea of representation information is ANY type of information that a future person might need to access the semantic content held within the digital object. In some ways it could be described as a decoder ring of sorts….. Representation information could be as small/simple as noting a specific dialect of a language that is spoken on the tape to describing an elaborate technical setup needed to accurately represent the information in a particular file
- We do this already with Physical Media -- for example an example of representation information for physical media is type -- most specifically for things like VHS tapes if a tape is PAL or NTSC. Someone trying to actually access the semantic content on that VHS tape (i.e watch it) will need to know if the tape is PAL or NTSC to actually be able to successfully watch the tape.
- Another good example of representation information, although more complicated, is emulation. There are times where a specific software/hardware combination is needed to run a program (this usually happens when a vendor goes out of business and the data format is proprietary) and the only way to get see/interact with a file is to re-create the hardware/software environment . You see this alot with video game platforms -- if you look online you find emulators for the original nintendo games, Sega Saturns, Intellivision https://www.lifewire.com/video-game-emulators-need-to-know-4687006



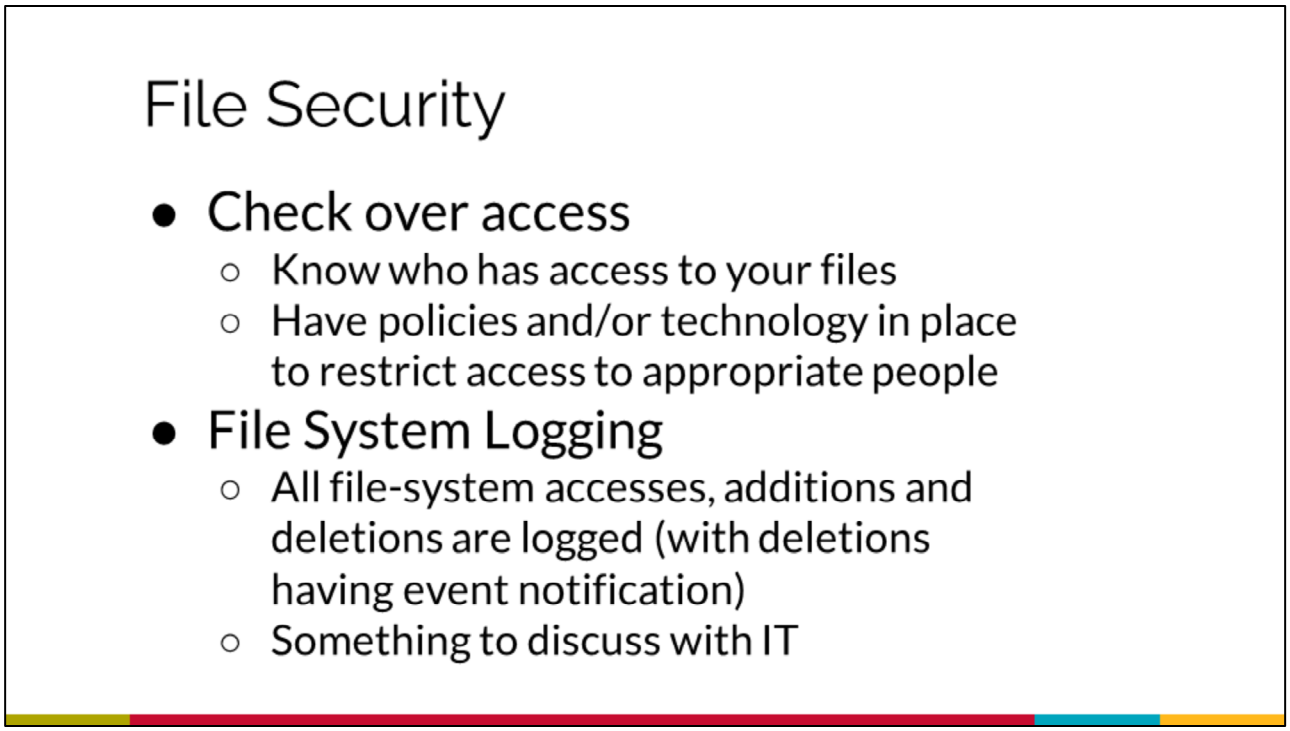
- Here a some more concrete examples of rep information. As a part of the metadata created (whether it is in the descriptive or preservation metadata) for an object you might (should add) a description of the languages that are contained within a document or languages spoken in an oral history -- this might be very important if there are a number of dialects of a language (that are very close but also might share words and sounds but not necessarily the same meaning)
- If you have a specific character-set that a text document is in: like ASCII or UTF-8. Recording the character set and the software and version used can really help folks down the road. For example -- have you ever open up an .xls in another program -- like wordpad? You probably would not be able to g
- Or in some cases it might just be as simple as:
    - “This file needs to be opened by ESRI Arcgis ArcView 7.x and above” to read all the map information correctly. Sure maybe Google Earth or another program could open that file but it may not translate all the prorietary ESRI data which could really throw what the viewer sees in a complex GIS file.
- The last one could even be just as easy as --- please use a video player that can read mp4 or an audio player that can read ogg vorbis.



- All this talk of differing codecs and character sets flows in very nicely to the next topic which is managing file types. As we have talked about before there are multitudes of filetypes that are used across media types for different purposes from jpeg to FLAC to Ogg Vorbis to MotionJpeg.
- In general it is **good** to start out with a short list of files to stick to for both Access and Master files --- hopefully those types will not be proprietary and will have the backing of the community so there will be support. Make a list and try to stick to it. If we stick to open and/or standard file types which are used by most everyone else (JPEG,TIFF, MP4, MP3,etc…) We will be in pretty good shape moving forward.
- hopefully those types will not be proprietary and will have the backing of the community so there will be support. Make a list and try to stick to it.

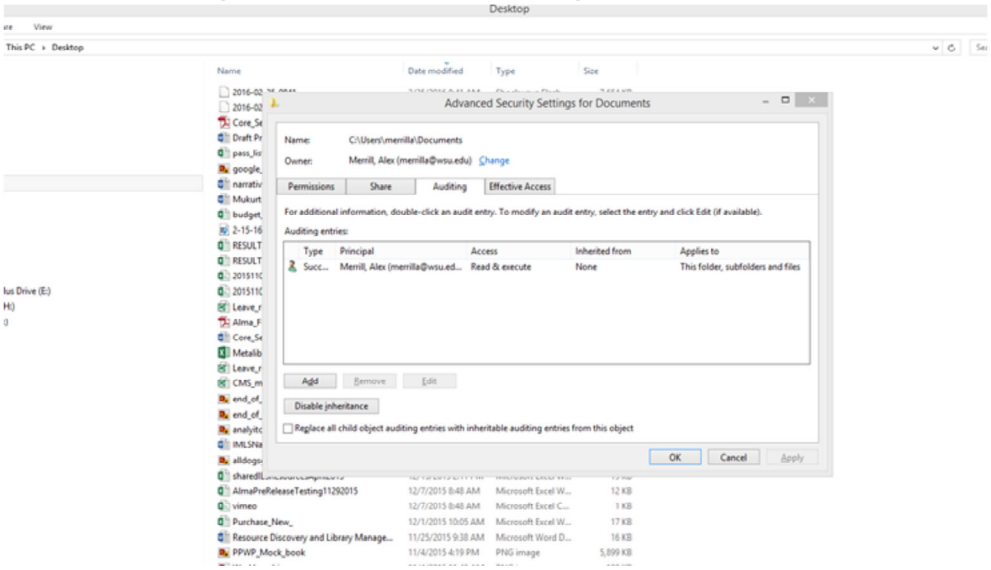


- Here we are talking about two things here really -- file types and the storage media you use. We already talked a bit about migrating file types -but to reiterate we need to choose open formats or migrate content to open formats when they become available. As new standard formats come out for example in the 3D model world there are all kinds of file formats right now (.obj, .ply, .3ds) and not really any real front runner for which is going to be the gold standard for preserving and archiving that 3D model. So looking for open, heavily used file formats (hopefully based on a standard or adopted by the community as a standard) will help you as you try to migrate file formats as there will probably already be pathways to follow
- But we also need to talk a bit about migrating your storage media or platform. Making sure that you keep up with technology, and don't try to keep using storage media for longer than its lifespan
- Why would we ever need to migrate our storage media?
    - In the early 2000's vendors and technologists alike were touting Gold CDs as an archival standard. The WSU Libraries had rows and rows of them as "archival" storage. We now know that even the gold CDs degrade erratically and all digital files should be pulled into a "live" storage medium…a storage arrangement like we discussed last time with "spinning" disks --- like Network Attached storage drives in Raid 5 or 6 arrangement. Migrating of storage media / platform can be one thing that can be offloaded to a cloud service easily -- if your IT and institutional policies allow for it.

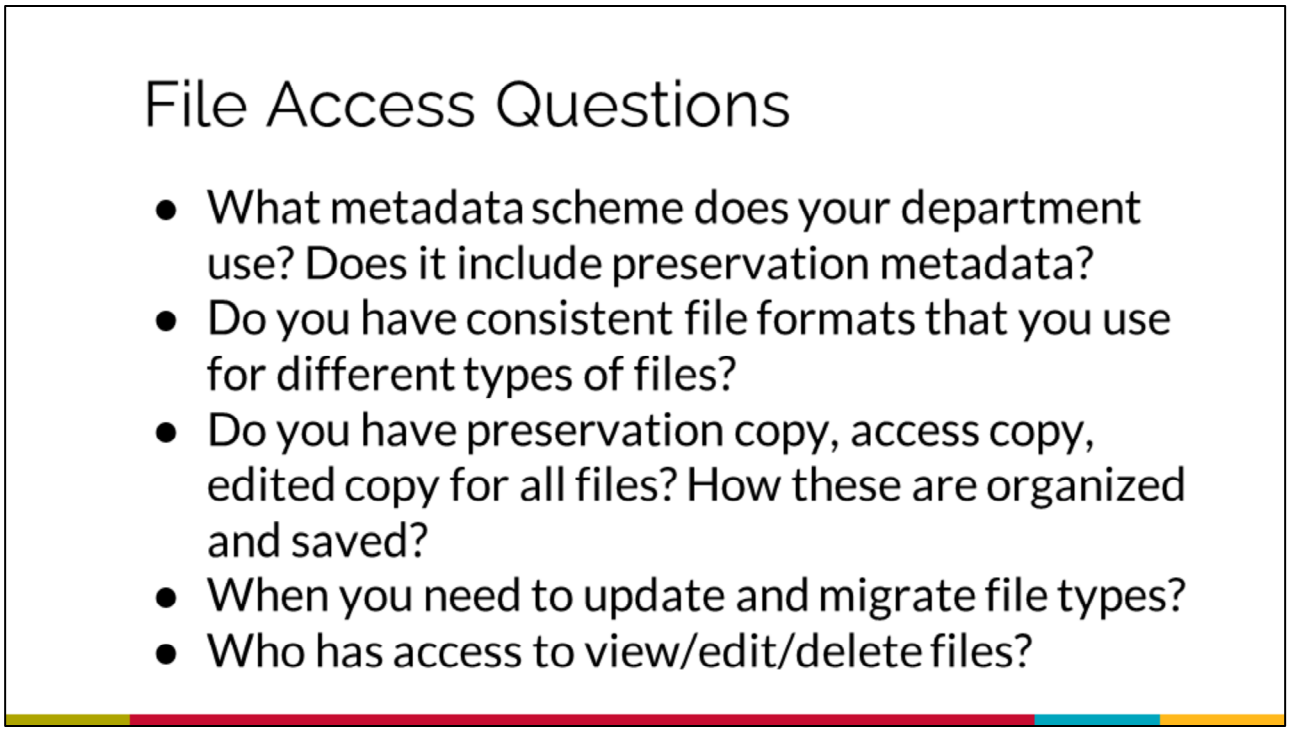


- Security - this was a topic in file integrity presentation, but it is just as important in files access
- Check over your access
    - Who has access?
    - Try to have both policies and technology in place to restrict access to only the people who need it
- Example --- WSU Archives utilizes a staging and a preservation storage area. This can be done using Operating system logging and Network level credentialing or standalone software…...
- While this is ideal… it is also pretty involved and creates a fair amount of overhead
- File system logging
    - This is a way to have a record of who is accessing files and what actions they take
- In the next slides we show how to turn on Security Auditing in Windows

Turning on Auditing in Windows

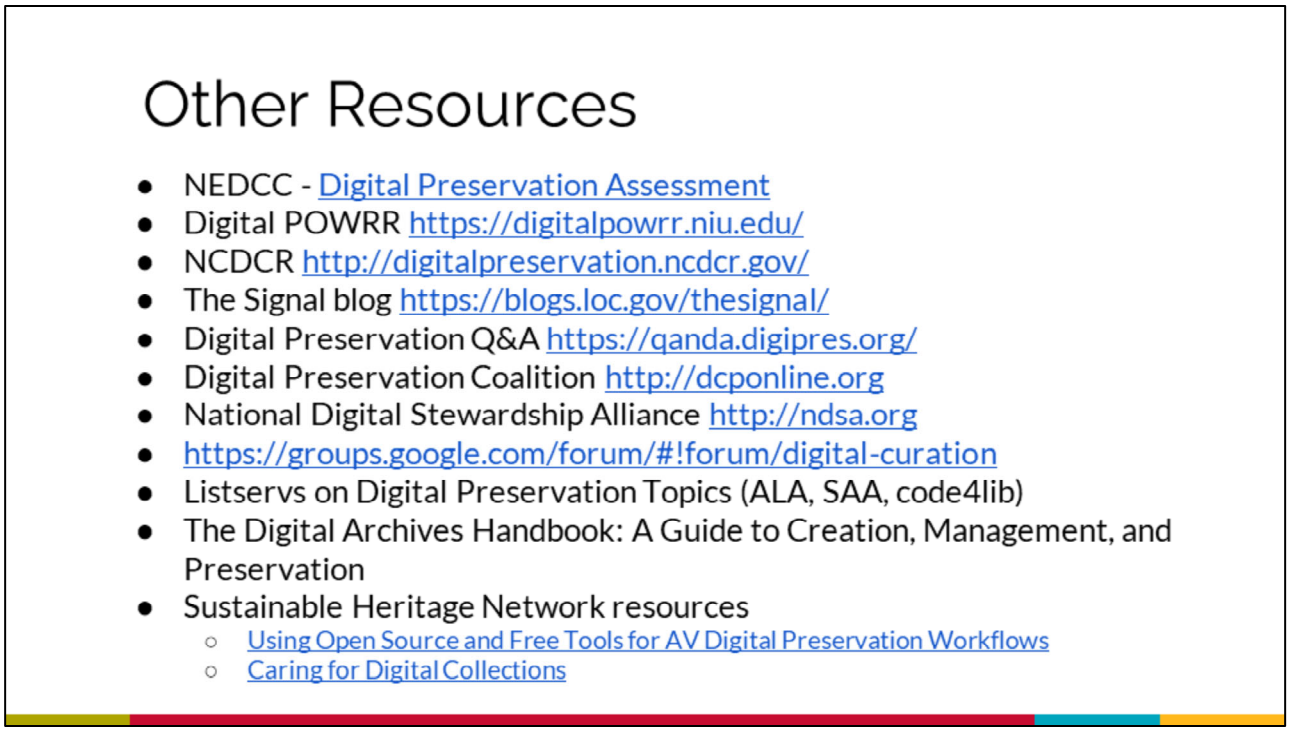


- Many different ways to do it. In windows go to the directory > Properties > Security > Advanced > Auditing> Add people or groups and that to
- If this seems important to your work, talk with IT or others to understand how you can implement Auditing

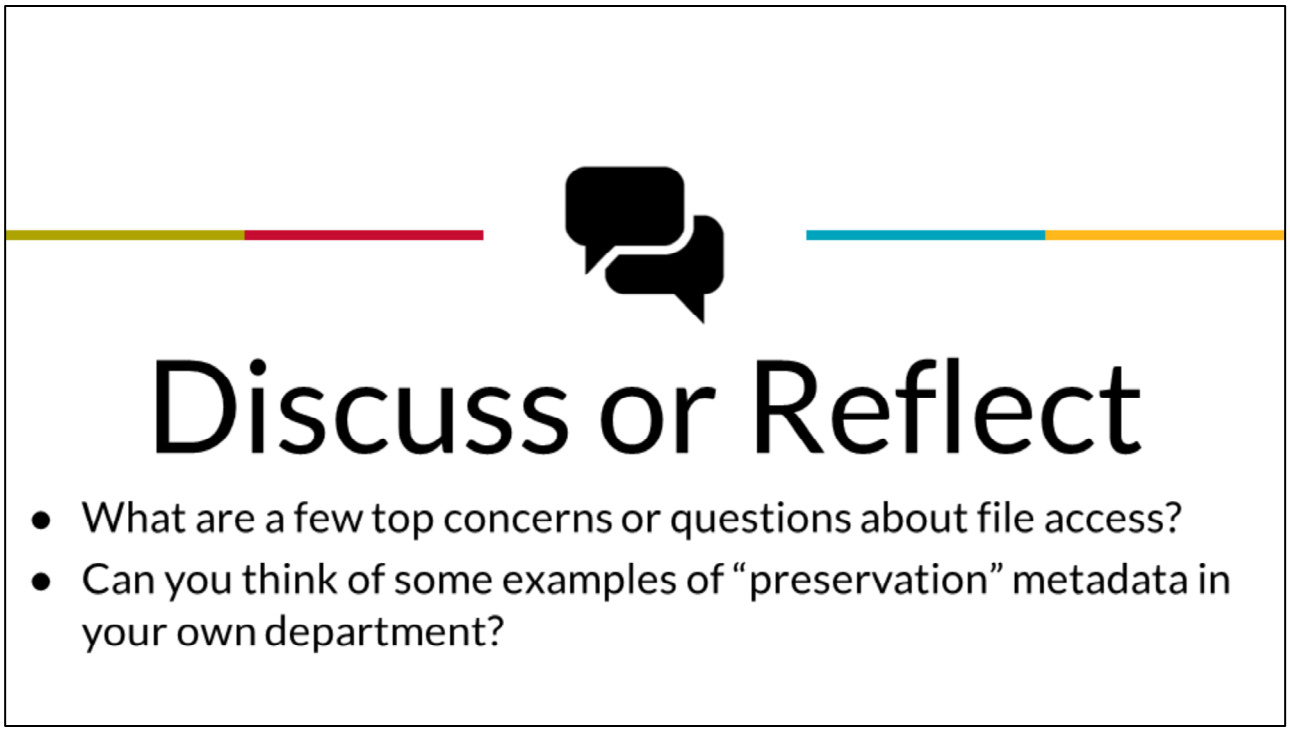


- These are some questions to start asking yourself and others about the file access aspects of digital preservation

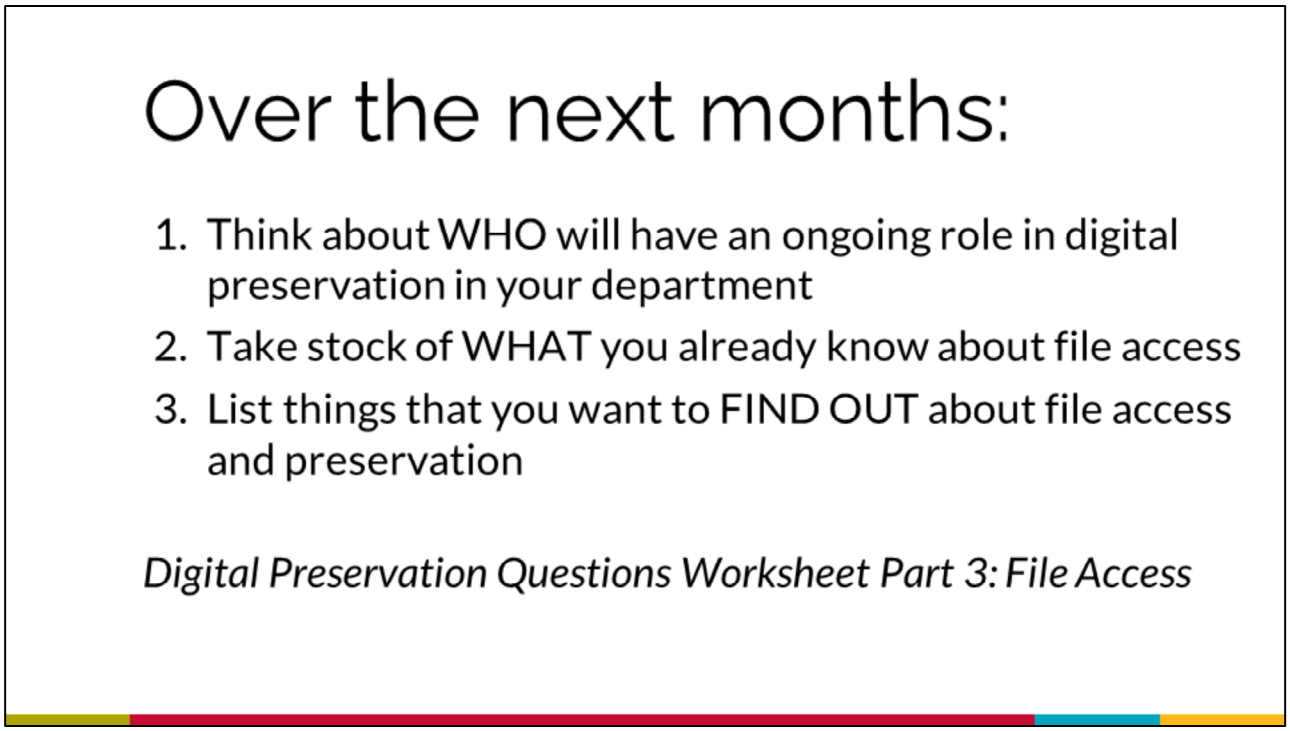
- These are some important concepts to learn more about, reference in your policies and plans, and base your own actions and thinking on - though you will have some things that are specific to your own institution and needs



- Here are some places to learn and stay informed. The digital preservation community is growing, and it is a good idea to keep updated on this information.
- UK resource: https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=digital-preservation



- Take 20-30 minutes and discuss with others, or reflect by yourself and take notes
- What are a few top concerns/questions you have about digital file access and preservation? Based on what you have learned so far.
    - What is the difference between internal digital preservation access and providing access to the community?
    - Had you heard of “preservation metadata” before?
    - Do you have any access/security concerns?
    - When you think of “representation information,” what comes to mind for your own collections?
    - What file types do you use for different formats? Are you consistent, or is there a variety of file types for any given format? How might you make things more consistent?
    - What about migration as hardware and software changes? Think about file types and storage media too.
- Can you think of some examples of “preservation” metadata in your own department?
    - Do you already keep track of anything like this?
    - If yes, how is this done?
    - If not, what are some examples of information that would be important for future people in your department in charge of preserving and maintaining digital files?



- Complete the SHN resource Digital Preservation Questions Worksheet Part 3: File Access to get some helpful questions and discussion started around this topic. Bring others into the conversation

Credits

- Images:
  - Slide 4, 6: Center for Digital Scholarship and Curation, Lotus Norton-Wisla, Michael Wynne, Alex Merrill
  - Slide 5: NDSA image https://ndsa.org/publications/levels-of-digital-preservation/
  - Slide 10: North Carolina Preservation Metadata for Digital Objects (PMDO)
- 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.