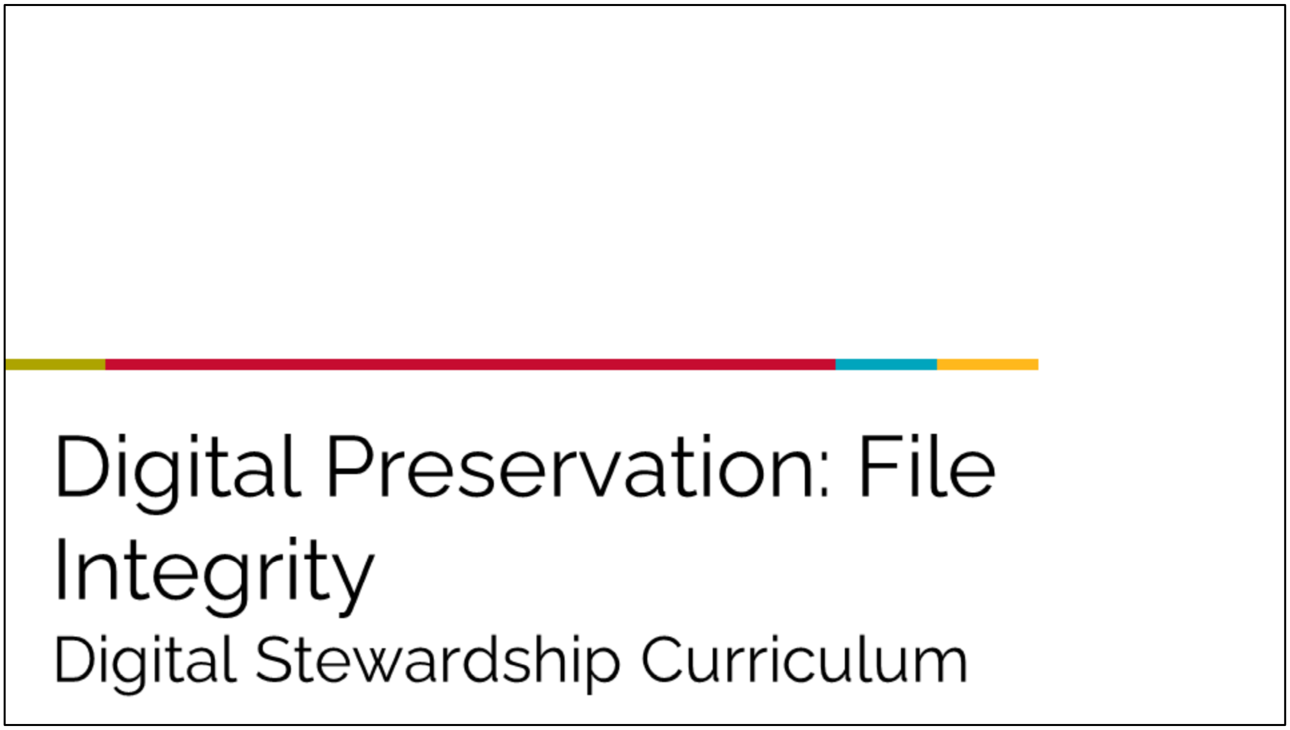
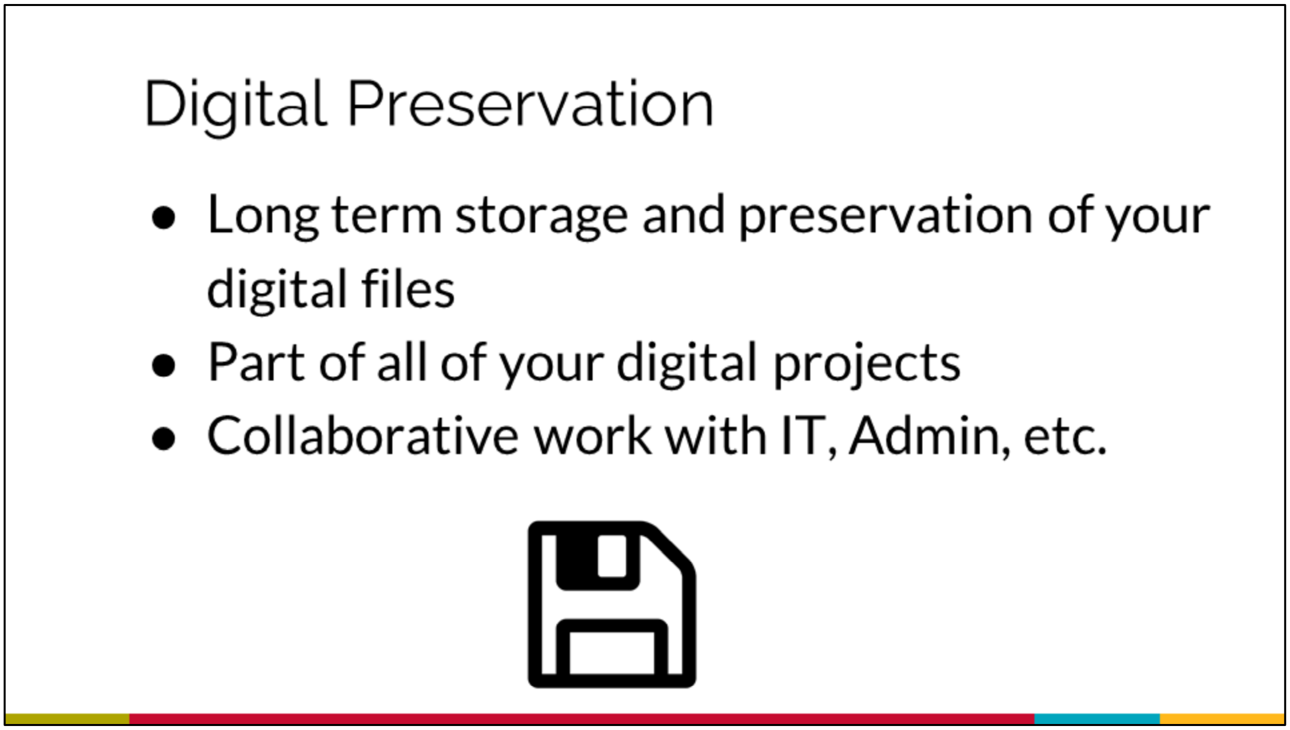


Digital Preservation: File Integrity

Digital Stewardship Curriculum

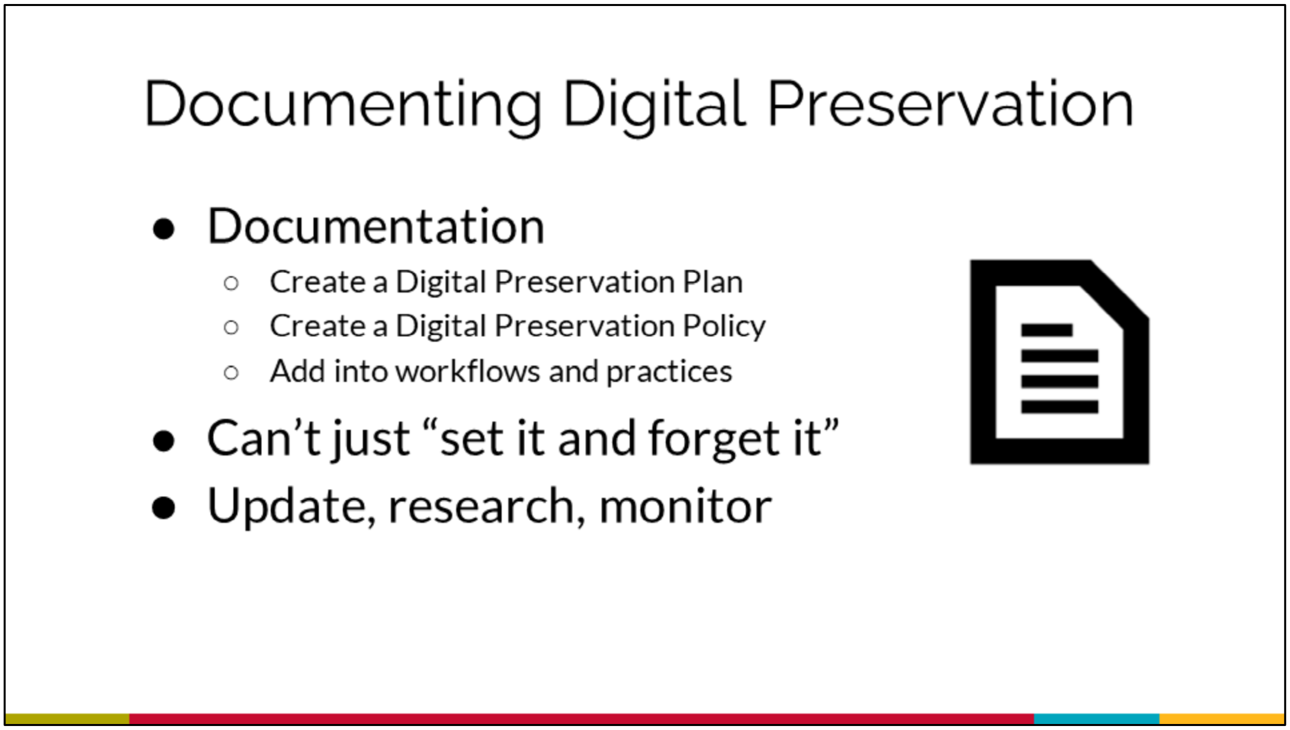
- This Digital Preservation: File Integrity presentation builds on the SHN resource Introduction to Digital Preservation and Storage



Digital Preservation

- Long term storage and preservation of your digital files
- Part of all of your digital projects
- Collaborative work with IT, Admin, etc.

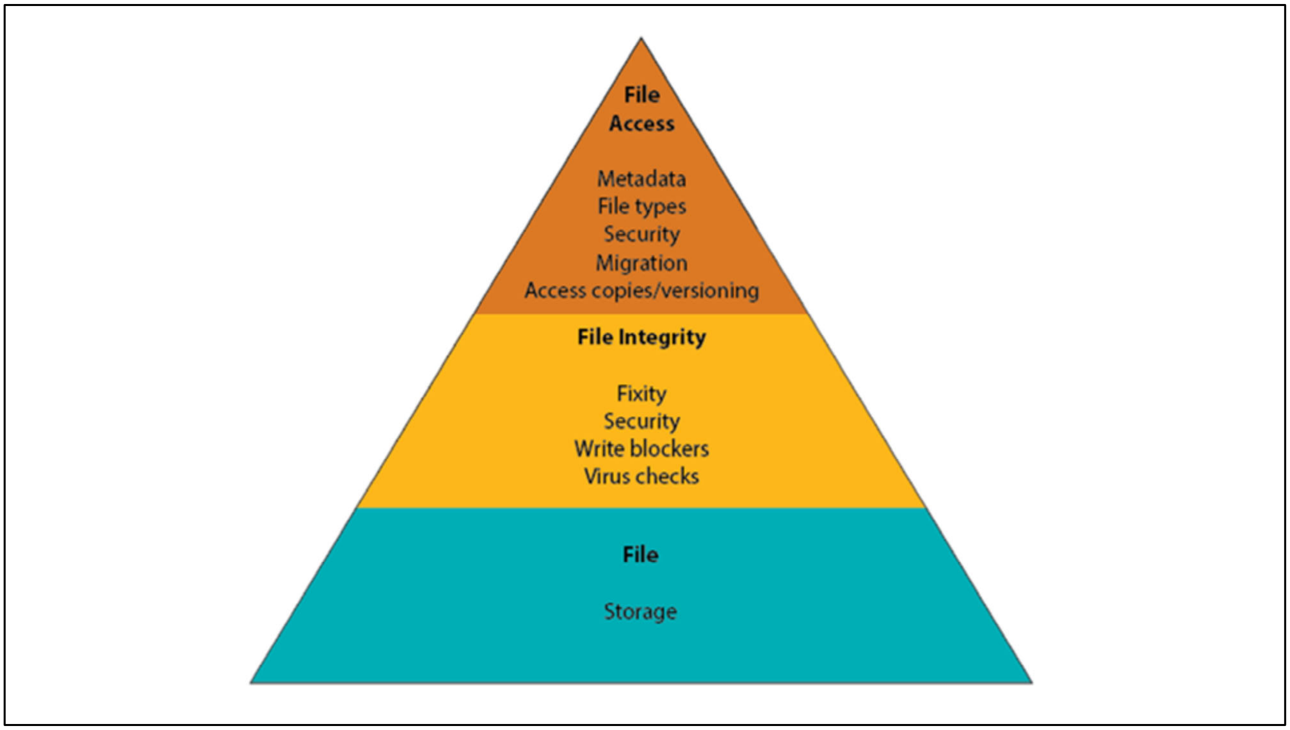
- Review on general digital preservation concepts
- Ensure that all the work you put into digitizing will be saved in the long term!
    - Digital preservation should be a conversation throughout your department/institution - if not, you will have to start small and keep at it
- Should be considering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



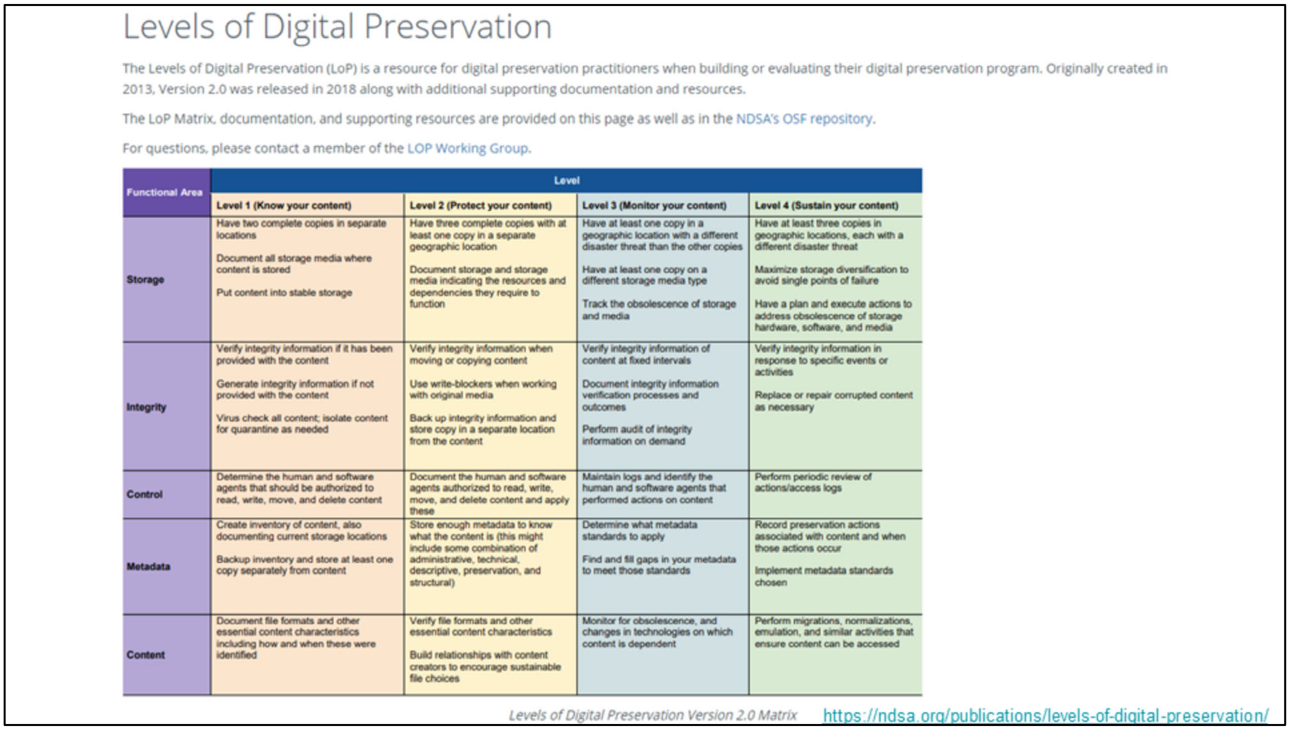
Documenting Digital Preservation

- Documentation
  - Create a Digital Preservation Plan
  - Create a Digital Preservation Policy
  - Add into workflows and practices
- Can't just "set it and forget it"
- Update, research, monitor

- 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



- Three Essentials of Digital Preservation Pyramid
- Broke down digital preservation into 3 essential parts that any Digital Preservation Plan will have to involve
- 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.

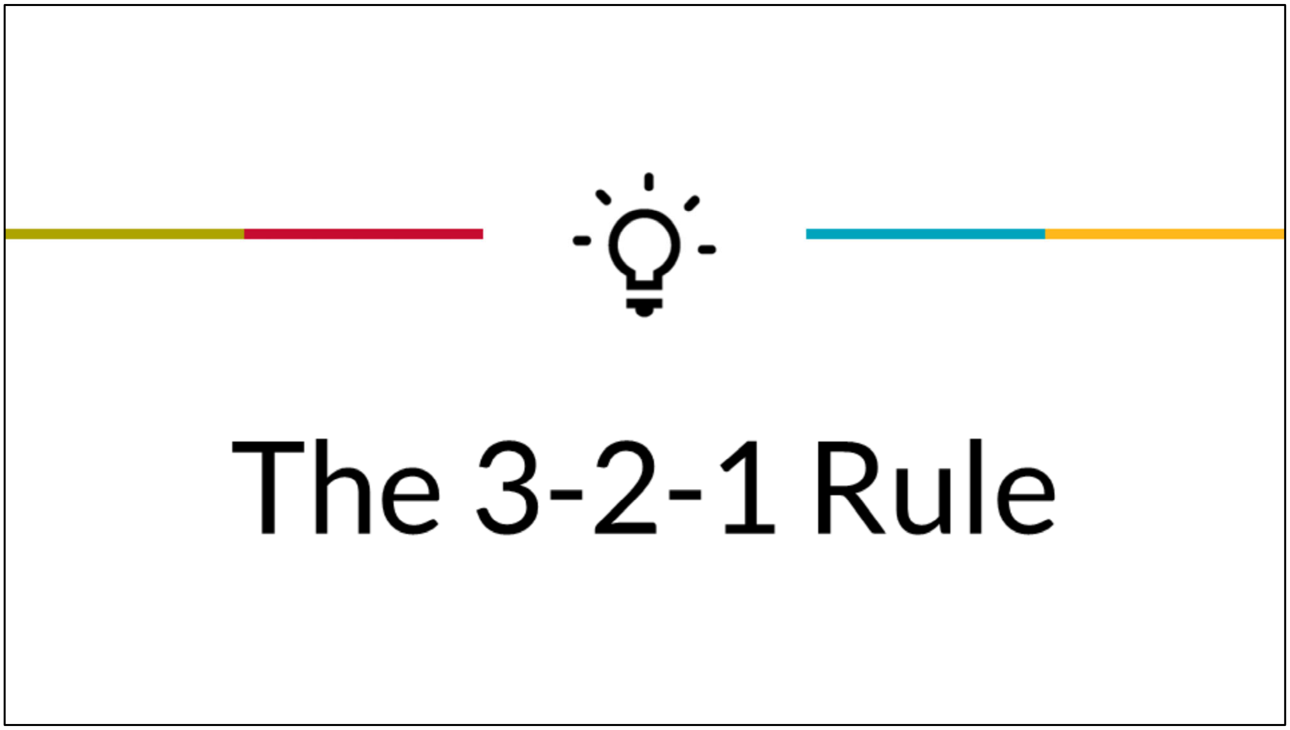
Functional Area	Level			
	Level 1 (Know your content)	Level 2 (Protect your content)	Level 3 (Monitor your content)	Level 4 (Sustain your content)
Storage	Have two complete copies in separate locations Document all storage media where content is stored Put content into stable storage	Have three complete 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 https://ndsa.org/publications/levels-of-digital-preservation/

- 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

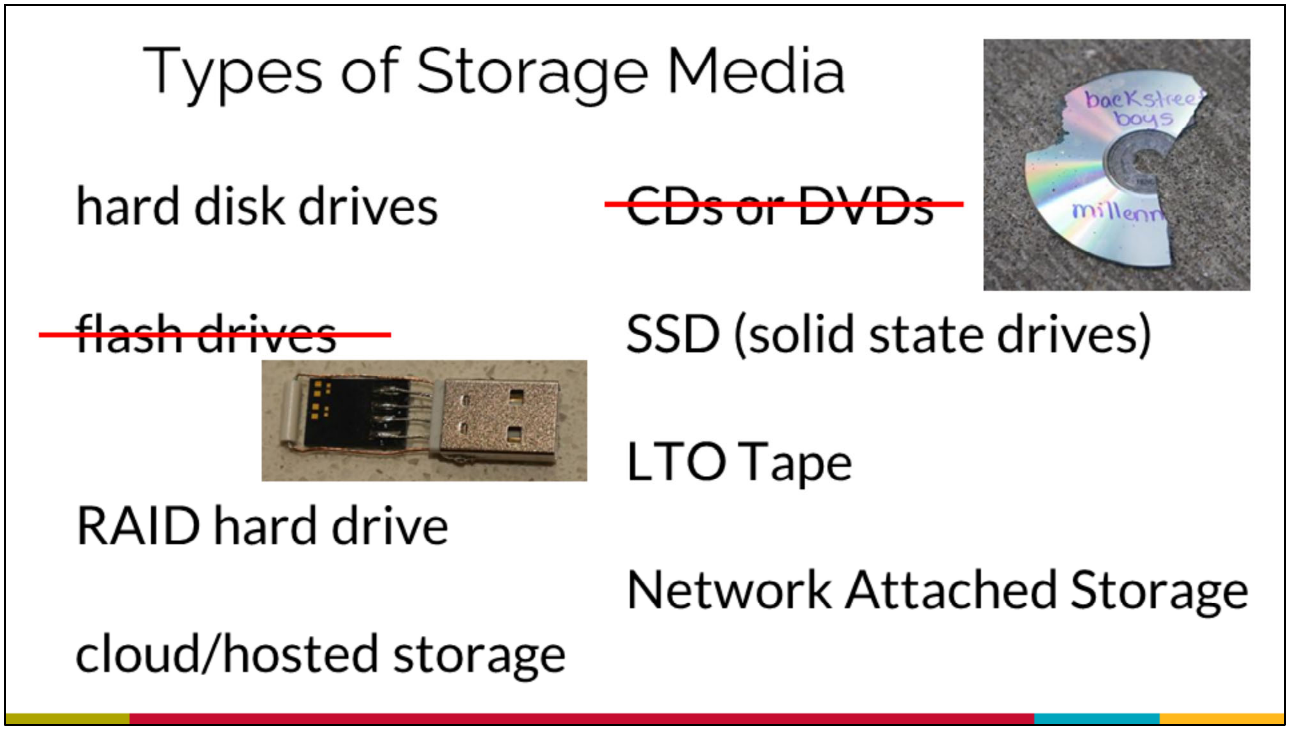
- See also, the SHN resource Levels of Digital Preservation Preparedness



The 3-2-1 Rule

- Review of 3 -2 - 1 rule
  - 3 copies of content to be preserved in the long time
  - Stored on at least 2 types of digital storage media
  - 1 copy should be located in a different disaster area (geographic region)

- We call it a rule, but it is really a guideline, your IT departments may have another way that they describe their system of storage and backups, but those concepts are important to include. Making sure you have that minimum in place, or something else that covers it.



Types of Storage Media

hard disk drives

~~flash drives~~

RAID hard drive

cloud/hosted storage

~~CDs or DVDs~~

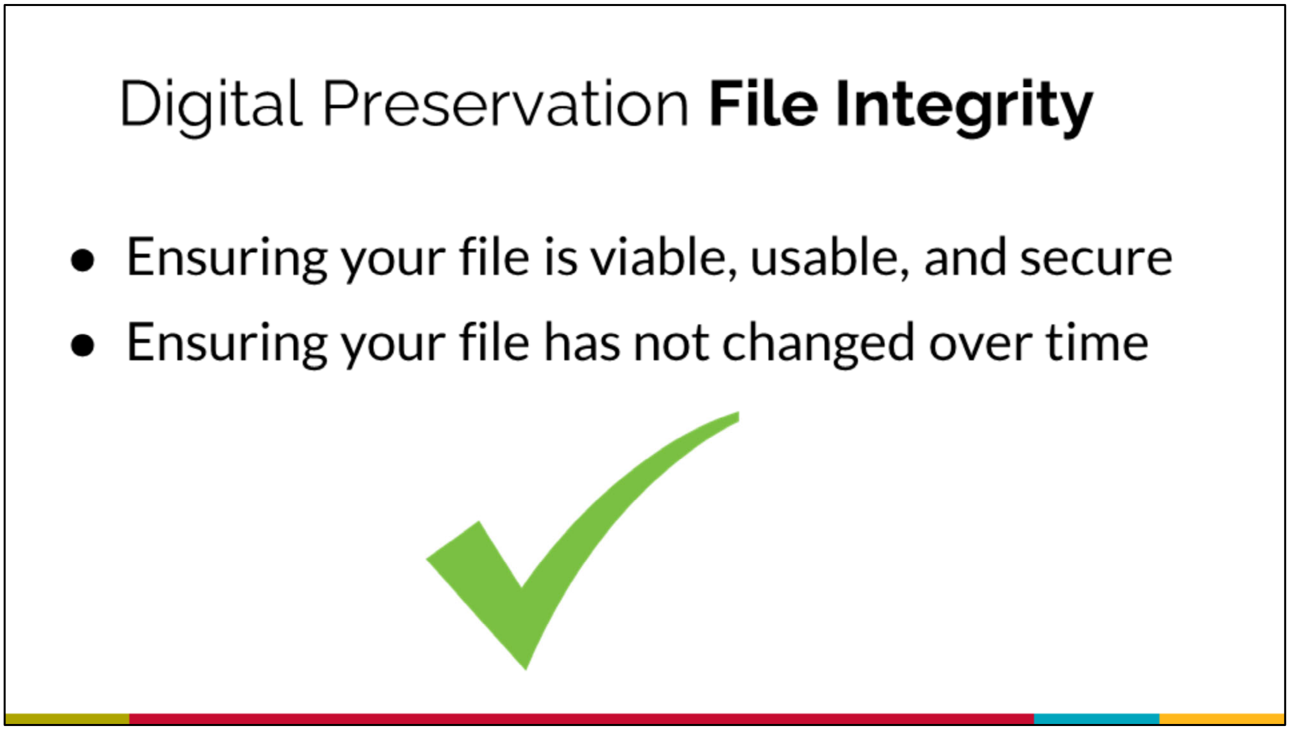
SSD (solid state drives)

LTO Tape

Network Attached Storage

backstree
boys
millenn

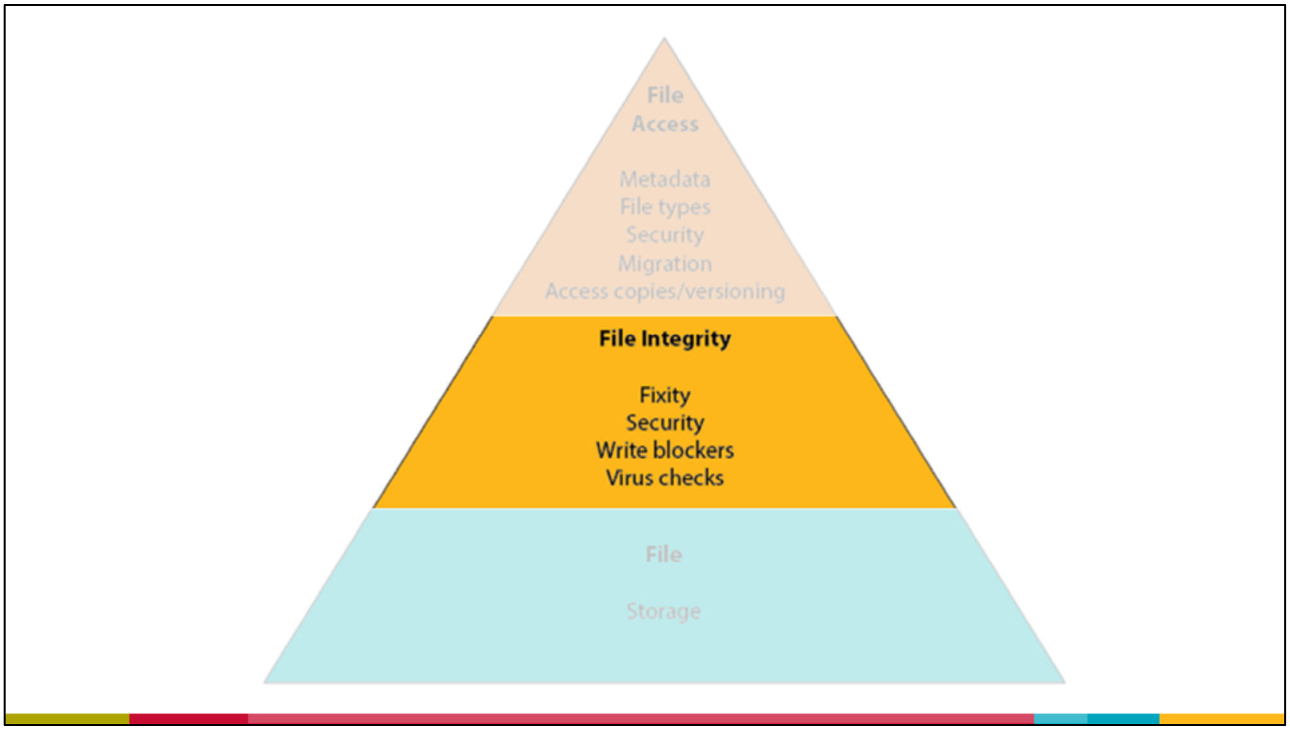
- Here are some different types of storage as a review
    - Remember:
    - Flash drives - fail easily, carry viruses easily
    - CDs or DVDs- obsolete, not used anymore - many new laptops no longer have an optical drive built in, media AND DATA can degrade without warning
- Flash drive image: Miles Goodhew Attribution 2.0 Generic (CC BY 2.0) https://www.flickr.com/photos/m0les/18582870504
- CD image: Brian Teutsch Backstreet Broken Attribution 2.0 Generic (CC BY 2.0) https://www.flickr.com/photos/brianteutsch/129913509



Digital Preservation **File Integrity**

- Ensuring your file is viable, usable, and secure
- Ensuring your file has not changed over time

- 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



File Access

Metadata
File types
Security
Migration
Access copies/versioning

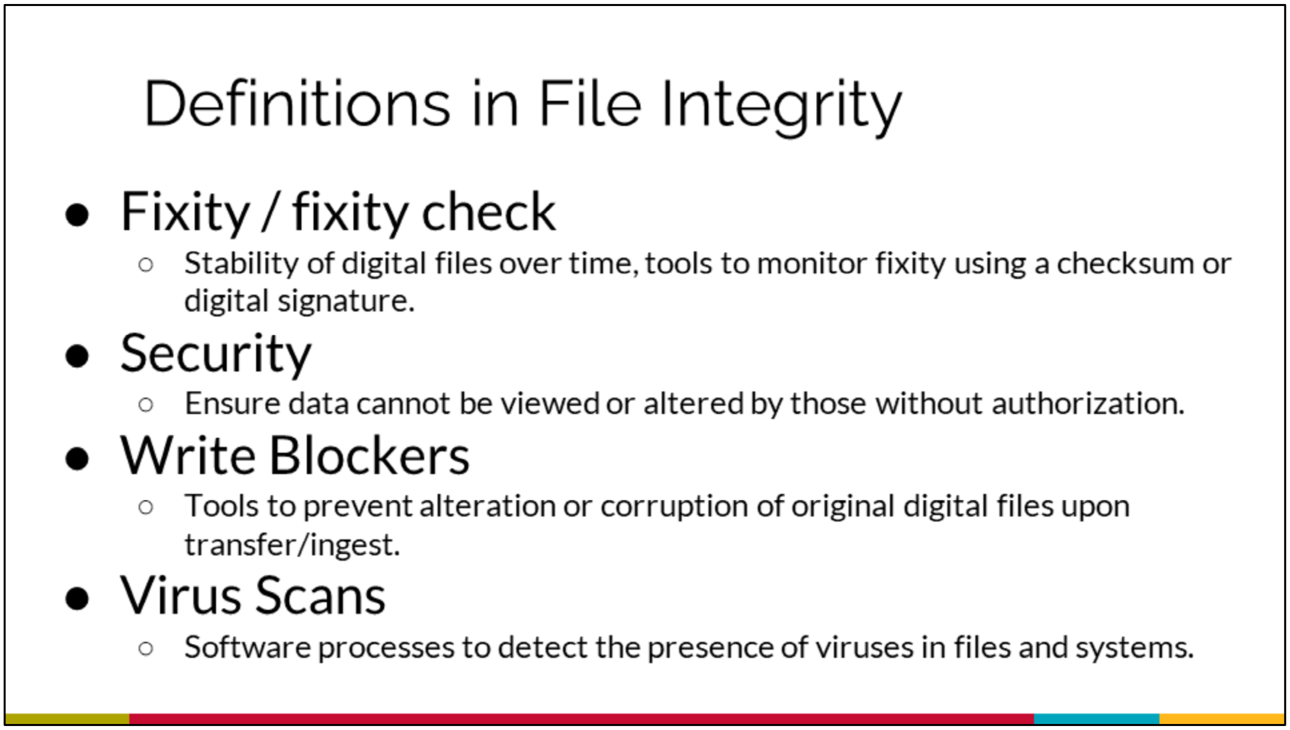
File Integrity

Fixity
Security
Write blockers
Virus checks

File

Storage

- File integrity - making sure the file you save and manage stay unchanged, secure, and authentic.

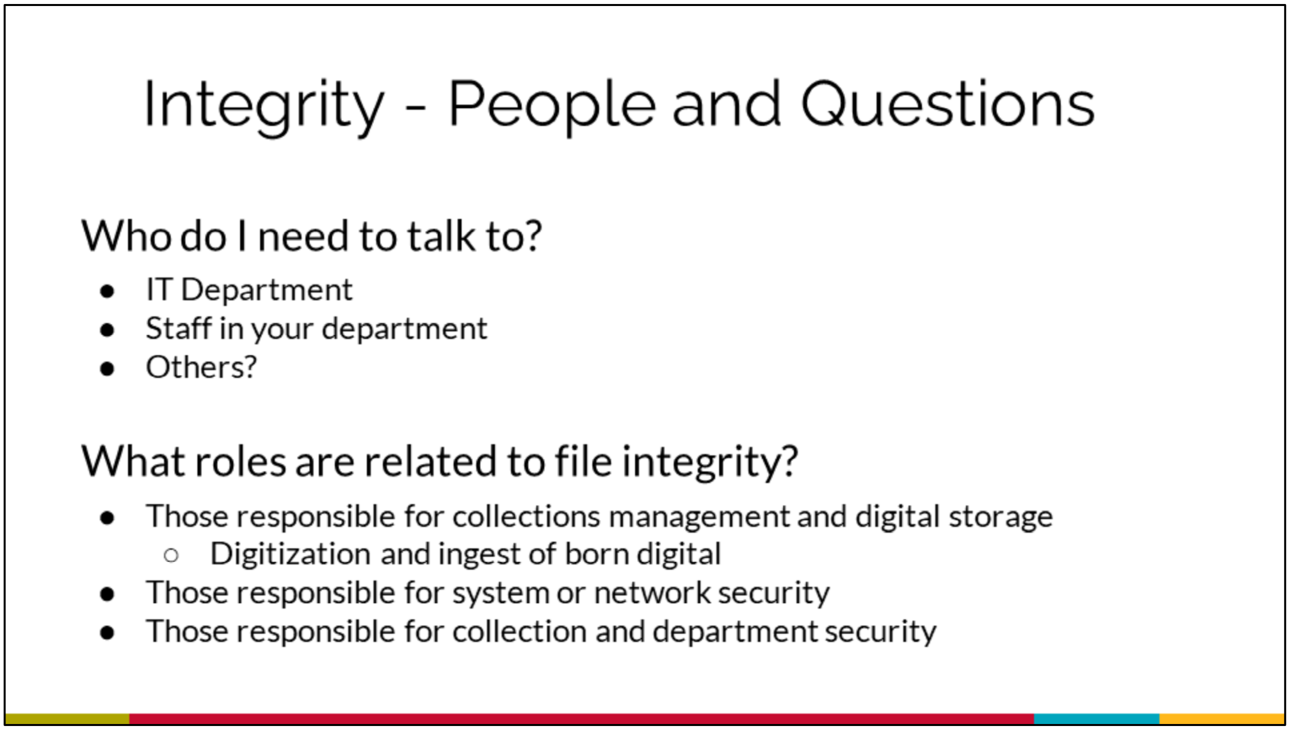


Definitions in File Integrity

- Fixity / fixity check
  - Stability of digital files over time, tools to monitor fixity using a checksum or digital signature.
- Security
  - Ensure data cannot be viewed or altered by those without authorization.
- Write Blockers
  - Tools to prevent alteration or corruption of original digital files upon transfer/ingest.
- Virus Scans
  - Software processes to detect the presence of viruses in files and systems.

- Files can degrade WITHOUT warning, on any storage media (some are more dangerous than others) (BIT ROT)
- Fixity - fixity checks are how you monitor this issue

- Write Blockers allow you to transfer files from another storage media onto your computer without being able to write to the original storage media. There are both hardware and software write blockers.



Integrity - People and Questions

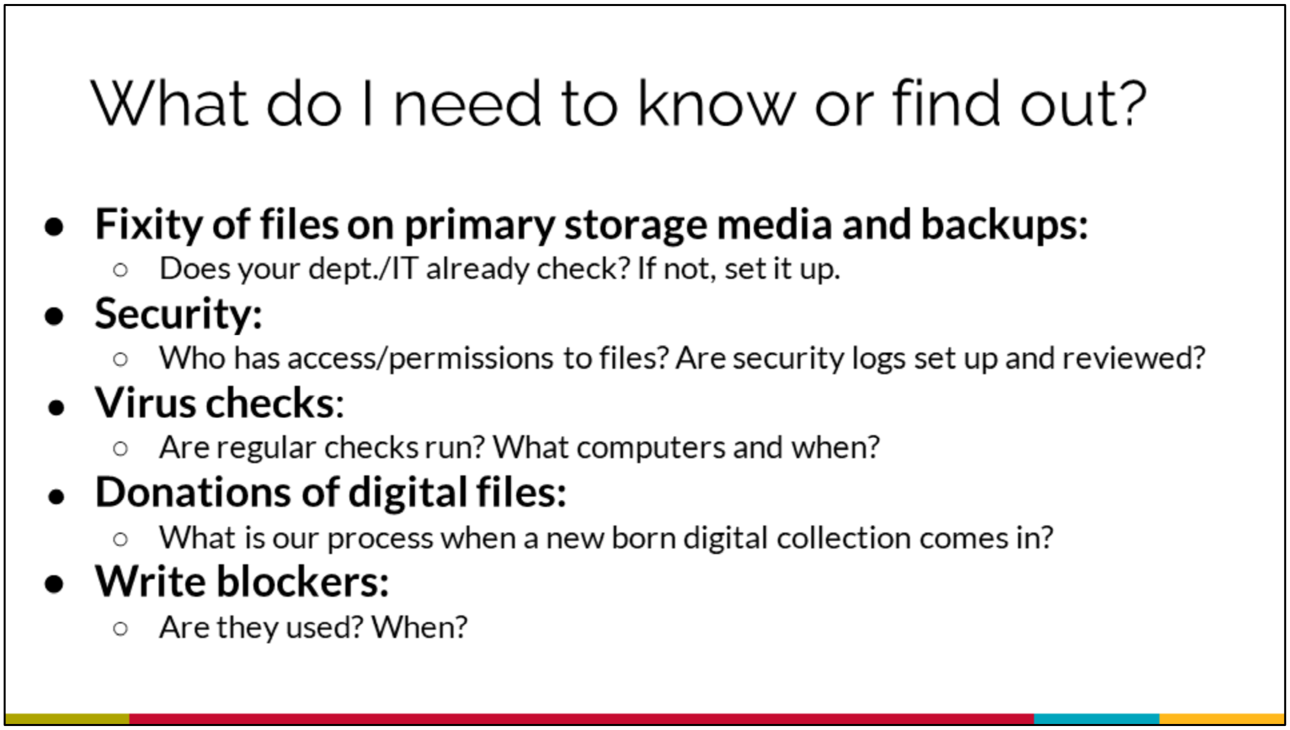
Who do I need to talk to?

- IT Department
- Staff in your department
- Others?

What roles are related to file integrity?

- Those responsible for collections management and digital storage
  - Digitization and ingest of born digital
- Those responsible for system or network security
- Those responsible for collection and department security

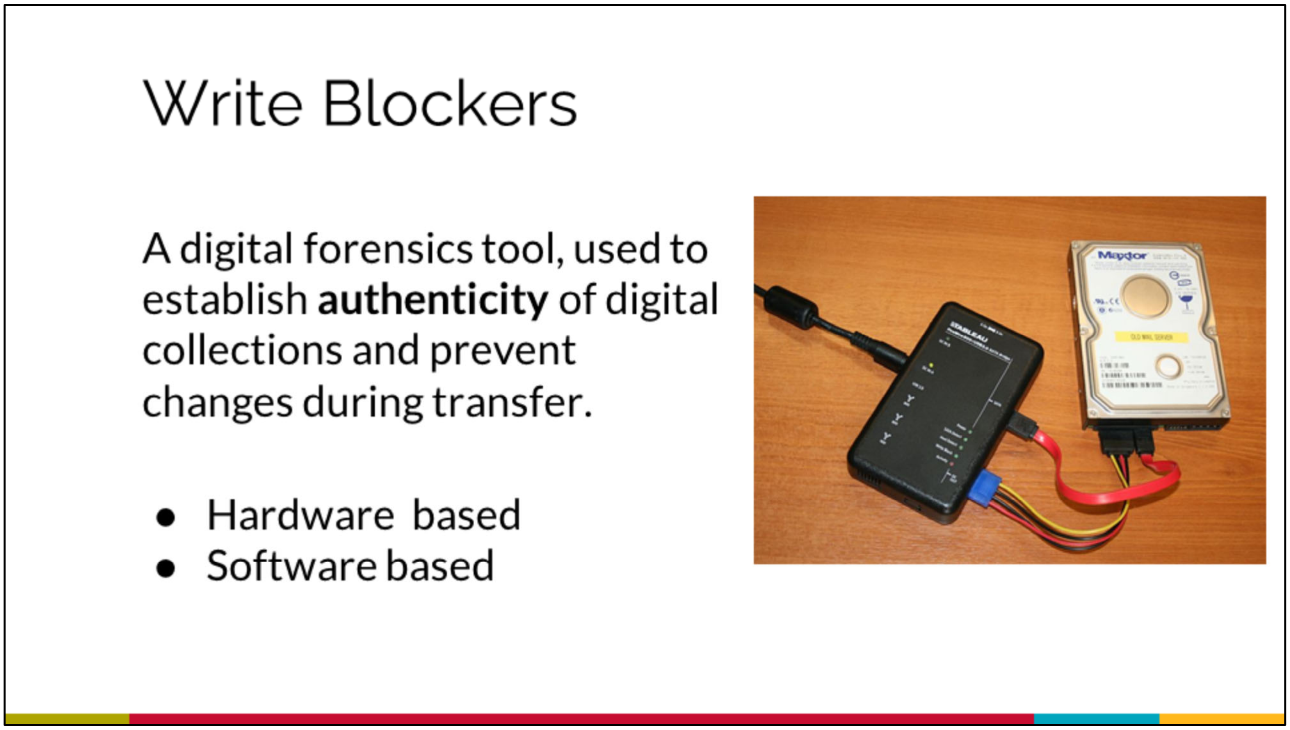
- Staff in your department need to know problems to look out for, and the correct procedures for preventing and dealing with issues
- Admin, staff that are responsible for digital collections management - eg: department head or collections manager
- IT, administration, or if you have someone who handles technology in your department



What do I need to know or find out?

- **Fixity of files on primary storage media and backups:**
  - Does your dept./IT already check? If not, set it up.
- **Security:**
  - Who has access/permissions to files? Are security logs set up and reviewed?
- **Virus checks:**
  - Are regular checks run? What computers and when?
- **Donations of digital files:**
  - What is our process when a new born digital collection comes in?
- **Write blockers:**
  - Are they used? When?

- These are topics that you should add into your vocabulary and knowledge base, and then start adding into your digital preservation plan and workflows



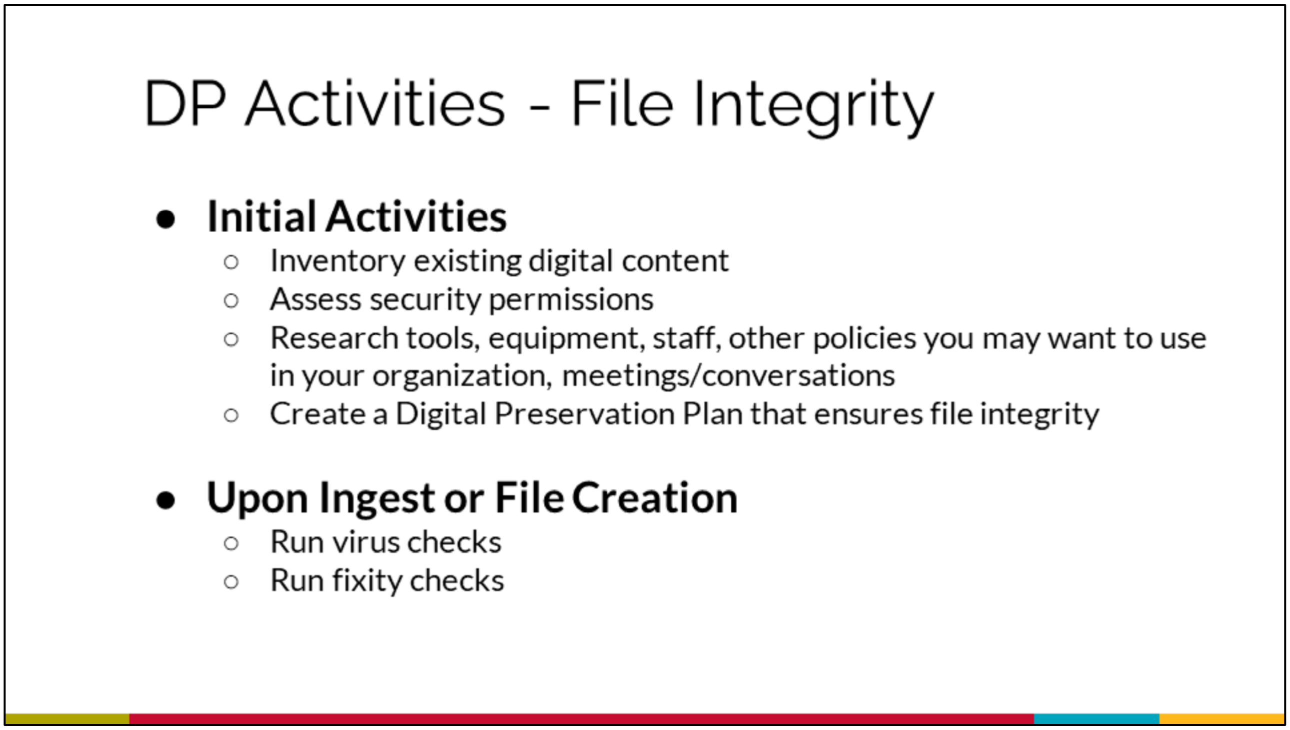
Write Blockers

A digital forensics tool, used to establish **authenticity** of digital collections and prevent changes during transfer.

- Hardware based
- Software based

- Write blockers are a digital forensics tool, used to establish authenticity of digital collections and prevent changes during transfer.
- They would fit into a digital files/media accessioning plan - to establish authenticity
- Give you more certainty that you or your computer will not accidentally change the file when plugging in a drive or copying over
- Hardware - need appropriate connectors (Hardware seems preferred if this is a regular part of your workflow)
    - Weibetech $350 http://www.cru-inc.com/products/wiebetech/usb-3-0-writeblocker/
    - Digital Intelligence - Tableau
- Software - more variables, more complicated
    - BitCurator
- Let you transfer data from a drive without creating the possibility of accidentally damaging the drive contents.
    - Can only read and copy, can't delete or modify
- Write Blockers were developed in law enforcement and criminal investigation - where it was really important that authenticity can be proven

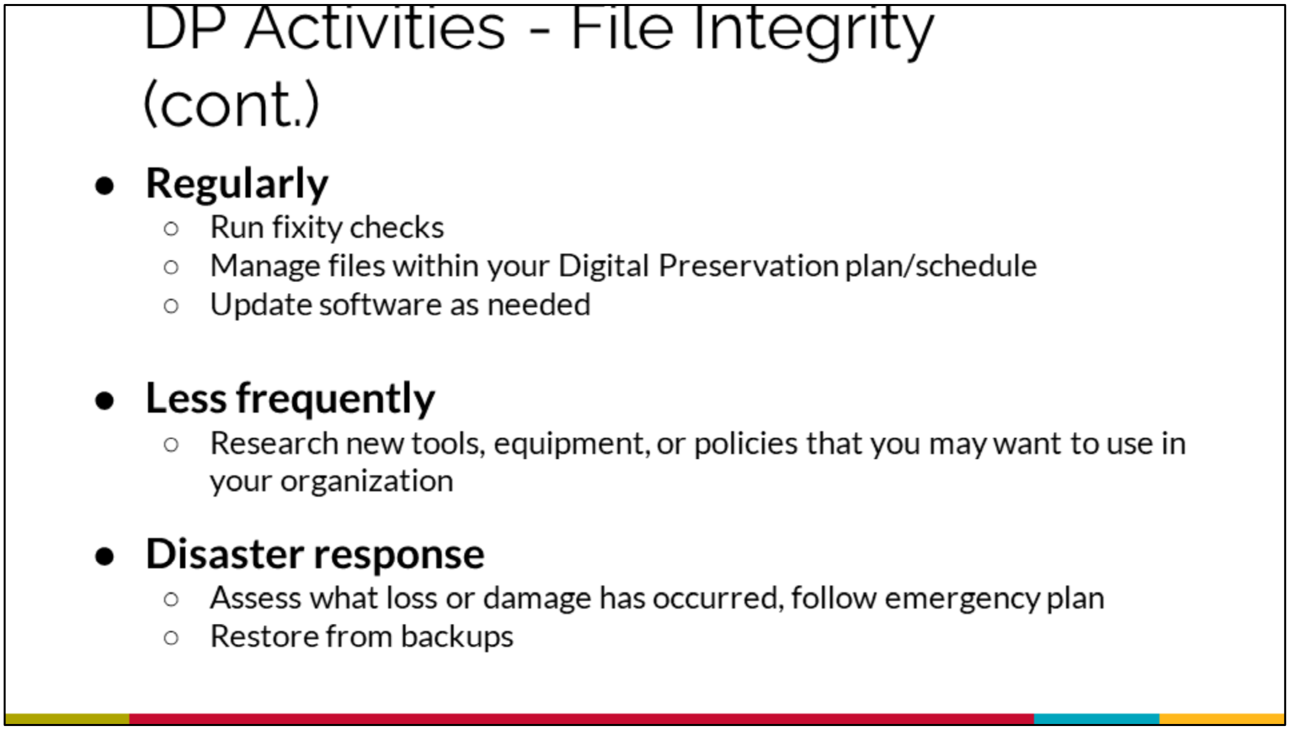
- ErrantX / Public domain https://commons.wikimedia.org/wiki/File:Portable_forensic_tableau.JPG



DP Activities - File Integrity

- **Initial Activities**
  - Inventory existing digital content
  - Assess security permissions
  - Research tools, equipment, staff, other policies you may want to use in your organization, meetings/conversations
  - Create a Digital Preservation Plan that ensures file integrity
- **Upon Ingest or File Creation**
  - Run virus checks
  - Run fixity checks

- For file integrity, there are
- For more information, see the SHN resource Activities to Include in a Digital Preservation Plan



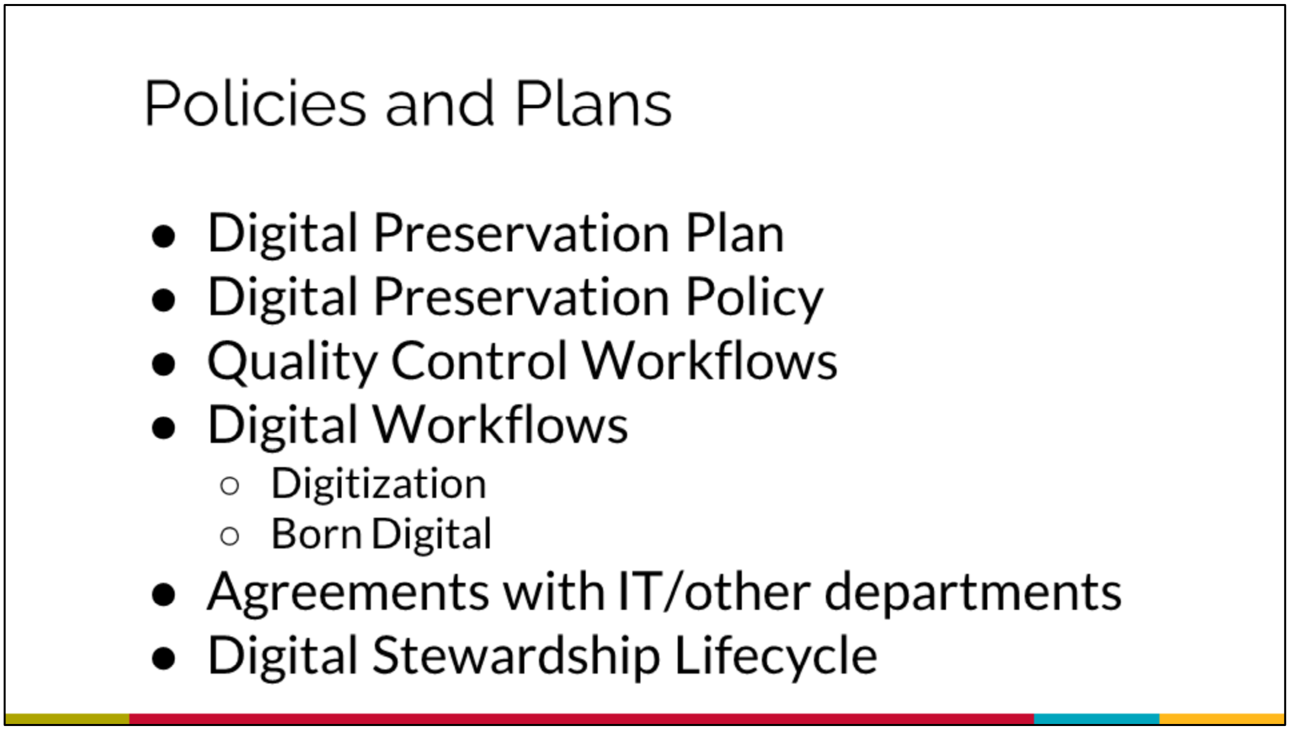
DP Activities - File Integrity (cont.)

- **Regularly**
  - Run fixity checks
  - Manage files within your Digital Preservation plan/schedule
  - Update software as needed

- **Less frequently**
  - Research new tools, equipment, or policies that you may want to use in your organization

- **Disaster response**
  - Assess what loss or damage has occurred, follow emergency plan
  - Restore from backups

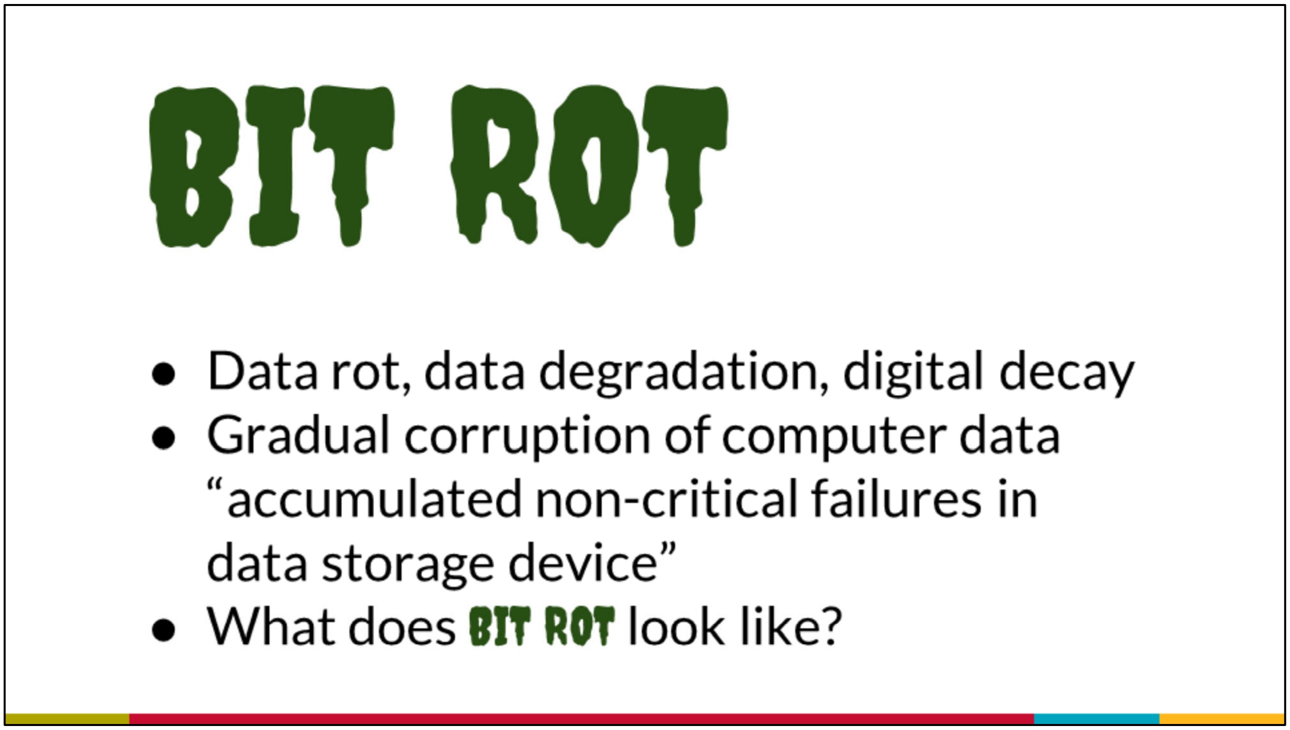
- For more information, see the SHN resource Activities to Include in a Digital Preservation Plan



Policies and Plans

- Digital Preservation Plan
- Digital Preservation Policy
- Quality Control Workflows
- Digital Workflows
    - Digitization
    - Born Digital
- Agreements with IT/other departments
- Digital Stewardship Lifecycle

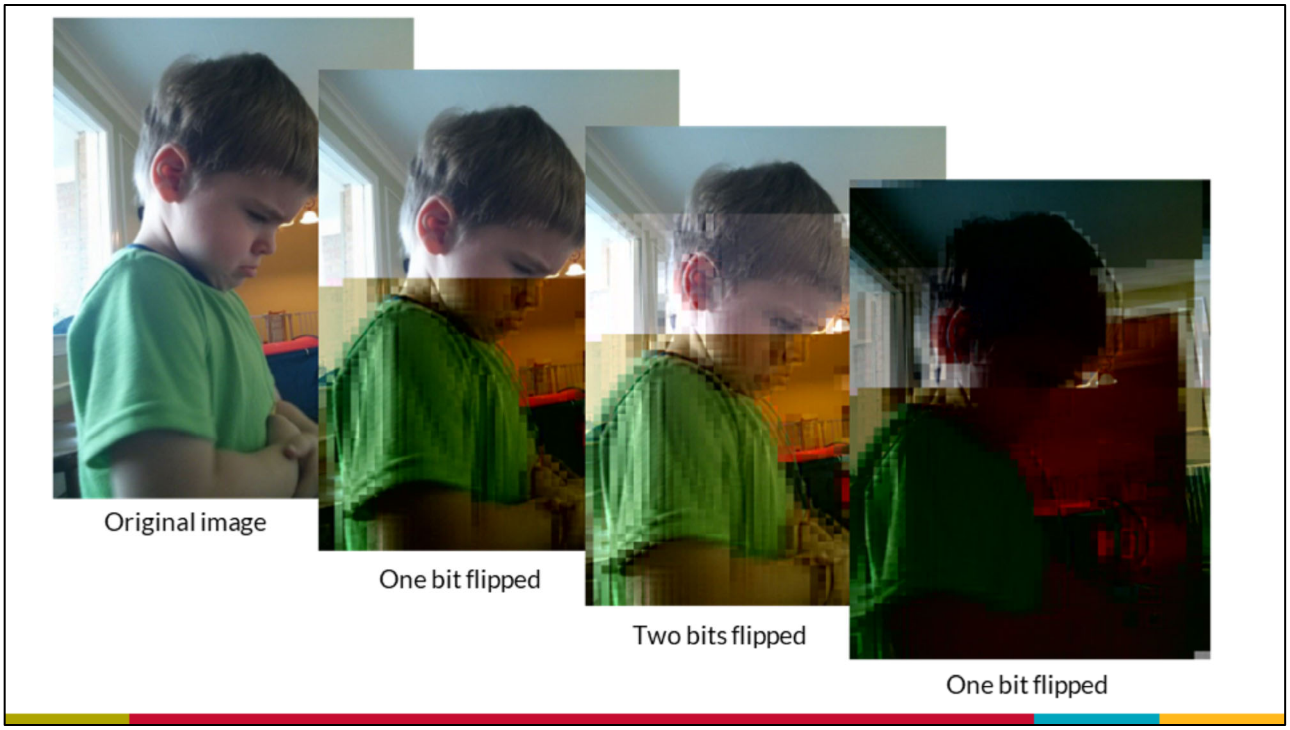
- Where does digital preservation and file integrity information appear in policies and plans?
- You may not have ALL of these policies, plans, and agreements in place right away
    - A good place to start is a digital preservation plan and workflows
- Expand from a section in your policy to a full document/larger section
- Important for staff working with digital content



BIT ROT

- Data rot, data degradation, digital decay
- Gradual corruption of computer data “accumulated non-critical failures in data storage device”
- What does BIT ROT look like?

- Bit Rot - a risk to your digital collections
- Other names: Data rot, data degradation, digital decay
- Gradual corruption of computer data “accumulated non-critical failures in data storage device”, things that might not make the storage device inaccessible, but tiny failures and issues that build up and cause corruption in your data
- Also affects software at the code level

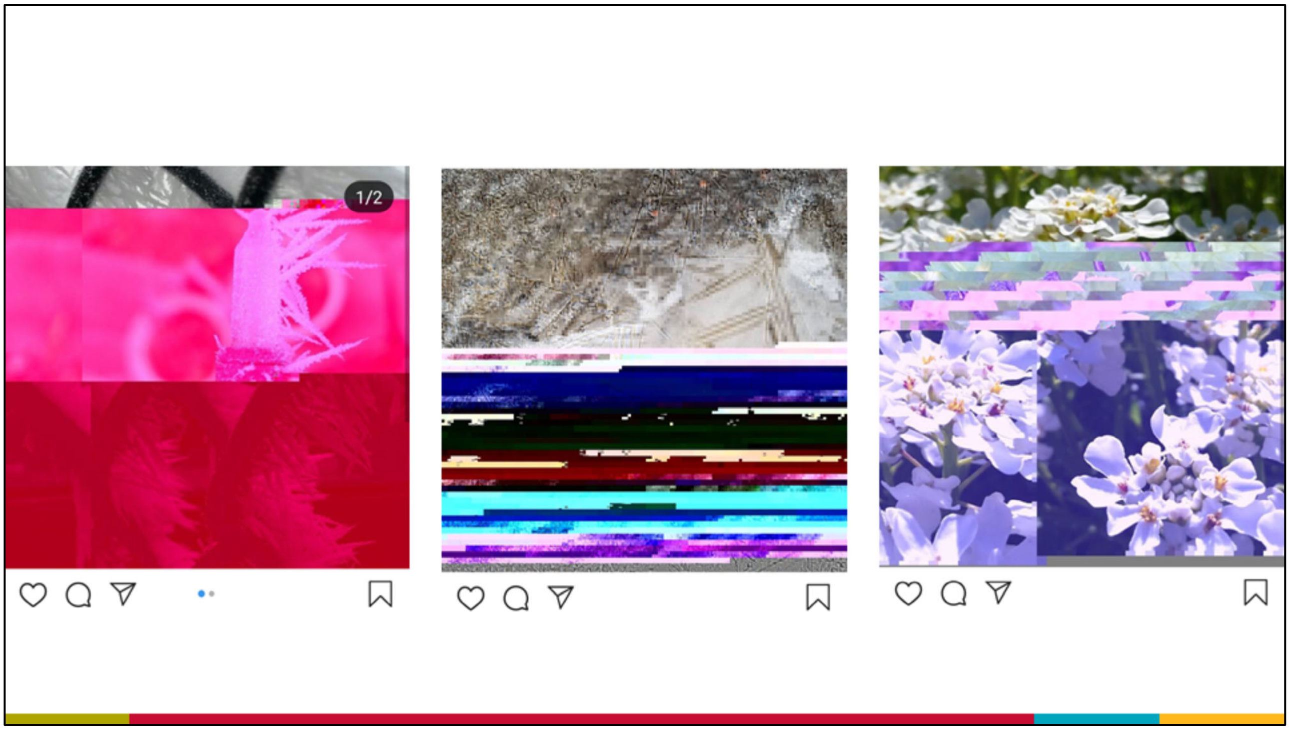


- Below are several digital images illustrating data degradation, all consisting of 326,272 bits. The original photo is displayed on the left. In the next image to the right, a single bit was changed from 0 to 1. In the next two images, two and three bits were flipped. On Linux systems, the binary difference between files can be revealed using cmp command (e.g. cmp -b bitrot-original.jpg bitrot-1bit-changed.jpg).
  - https://en.wikipedia.org/wiki/Data_degradation

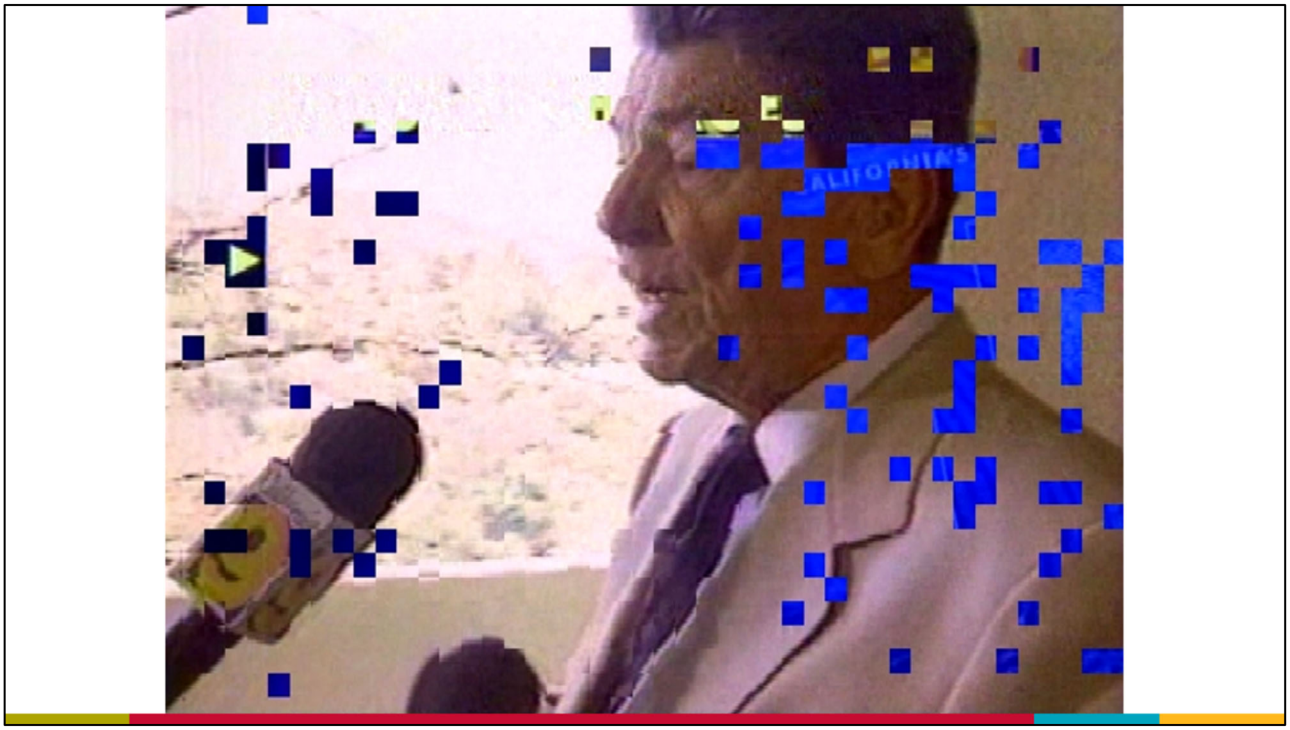
- Jim Salter / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0) https://commons.wikimedia.org/wiki/File:Bitrot_cascade.png

- From left to right, we see the continuing degradation of a JPEG image as bits are flipped. The left image is the original; each succeeding image has one bit flipped from the last.
- This serves as an example of how badly relatively small amounts of damage can degrade a JPEG file's visual presentation.

- The full original files, which can be diffed to see the changes between them, are also available beginning at https://commons.wikimedia.org/w/index.php?title=File:Bitrot_in_JPEG_files,_0_bits_flipped.jpg.“”



- Corruption of digital files
- Example: instagram copies of posts saved on smartphone, appear pixelated, distorted, colors off
- Lotus Norton-Wisla / CC BY-NC-SA (https://creativecommons.org/licenses/by-nc-sa/4.0/)

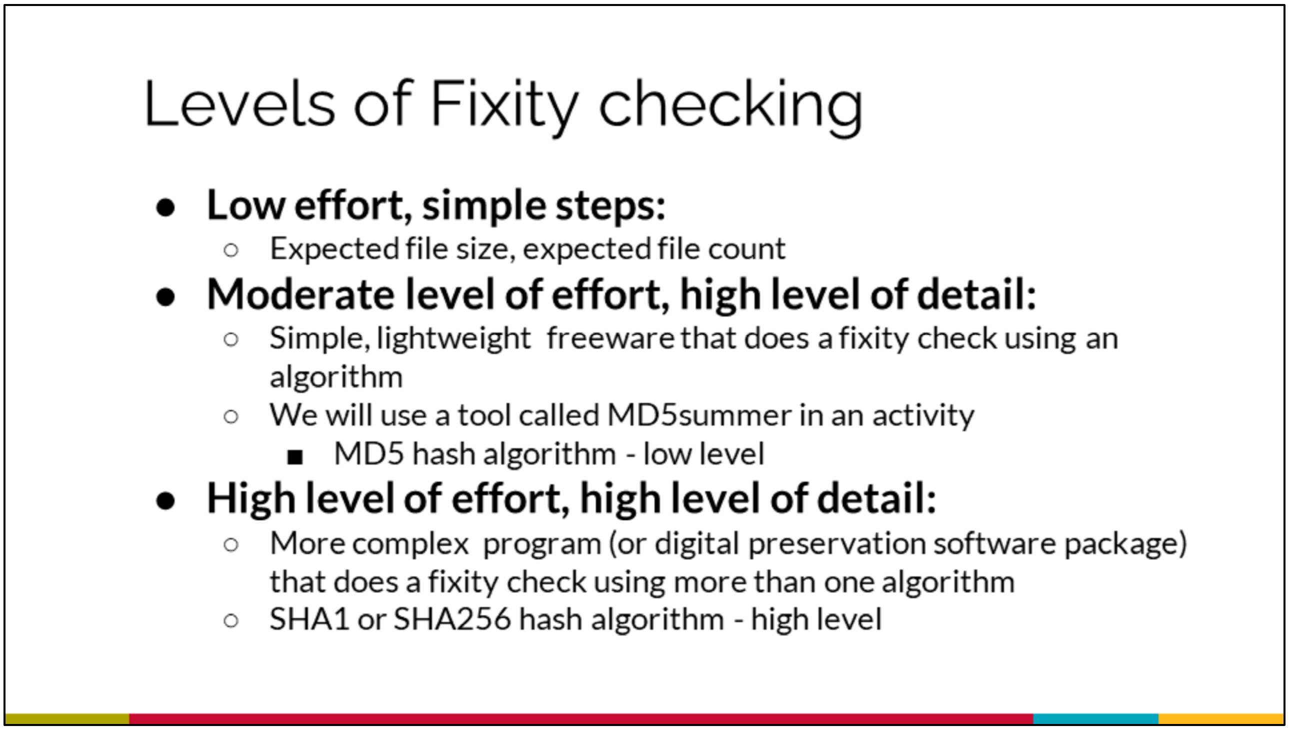


- Video - pixelation is a type of digital artifact also found in video
- Flickr user Wlef70: “Transfer from Betacam SP analog video tape to MPEG-2 digital video created this digital artifact.”

- Audio - digital artifacts - clicks and pops
- Loss of structure - unopenable
- Metadata - corruption of metadata

- Wlef70 Reagan pixelated https://creativecommons.org/licenses/by-nc-sa/2.0/ https://www.flickr.com/photos/wlef70/8097509137/

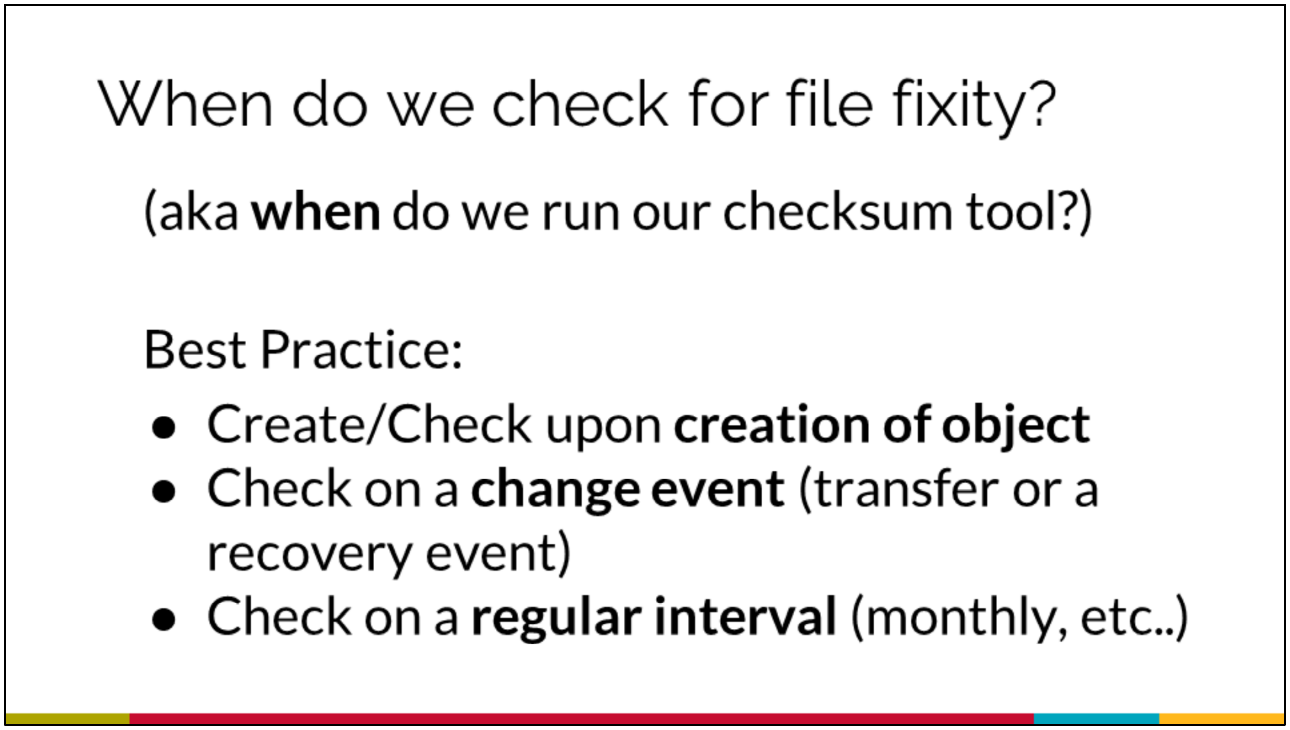
- More: Glitch Art
  - https://www.flickr.com/photos/r00s/3293434558



Levels of Fixity checking

- **Low effort, simple steps:**
  - Expected file size, expected file count
- **Moderate level of effort, high level of detail:**
  - Simple, lightweight freeware that does a fixity check using an algorithm
  - We will use a tool called MD5summer in an activity
    - MD5 hash algorithm - low level
- **High level of effort, high level of detail:**
  - More complex program (or digital preservation software package) that does a fixity check using more than one algorithm
  - SHA1 or SHA256 hash algorithm - high level

- Low level of effort, low level of detail
- Just gloss over Low level and SHA1, focus on what we will be learning (MD5Summer).
- File fixity or integrity can be checked using one or a combination of several mechanisms.
  - The simplest forms of file fixity are expected file size and expected file count.
    - These checks are used on a regular basis when large amounts of files are being transferred between file systems.
    - These fixity tools are the easiest to perform, as these checks can be easily read without specialty software, but also generate the least amount detail about the integrity of the files being checked.
  - As you move up the food chain of fixity checking you see that with a little more effort you can get a high level of detail using a cryptographic hash algorithm.
    - The lowest level of hash algorithm is MD5, it has the ability to, at least in all of our cases, uniquely identify a file with a string of letters and numbers and is relatively lightweight when it comes to computer overhead.
    - The highest level of detail available is by using an SHA1 or SHA256 has algorithm…… these two algorithms have a higher level of detail because each file is represented by a longer string of letters and numbers…..
    - The differences between the MD5 and the SHA is the difference between a 300DPI scan and a 600 or 700 DPI scan. To illustrate what a HASh algorithm does we will use a few normal english phrases…..and push them against a hash algorithm



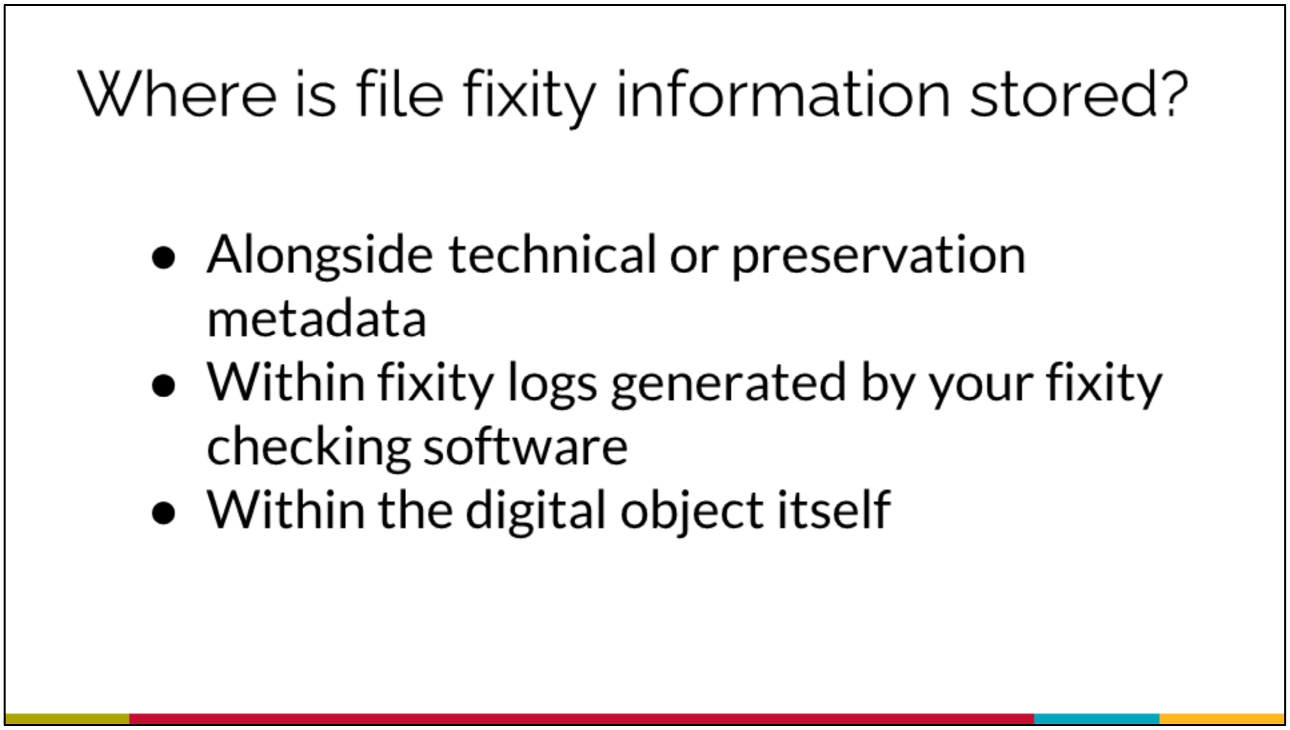
When do we check for file fixity?

(aka **when** do we run our checksum tool?)

Best Practice:

- Create/Check upon **creation of object**
- Check on a **change event** (transfer or a recovery event)
- Check on a **regular interval** (monthly, etc..)

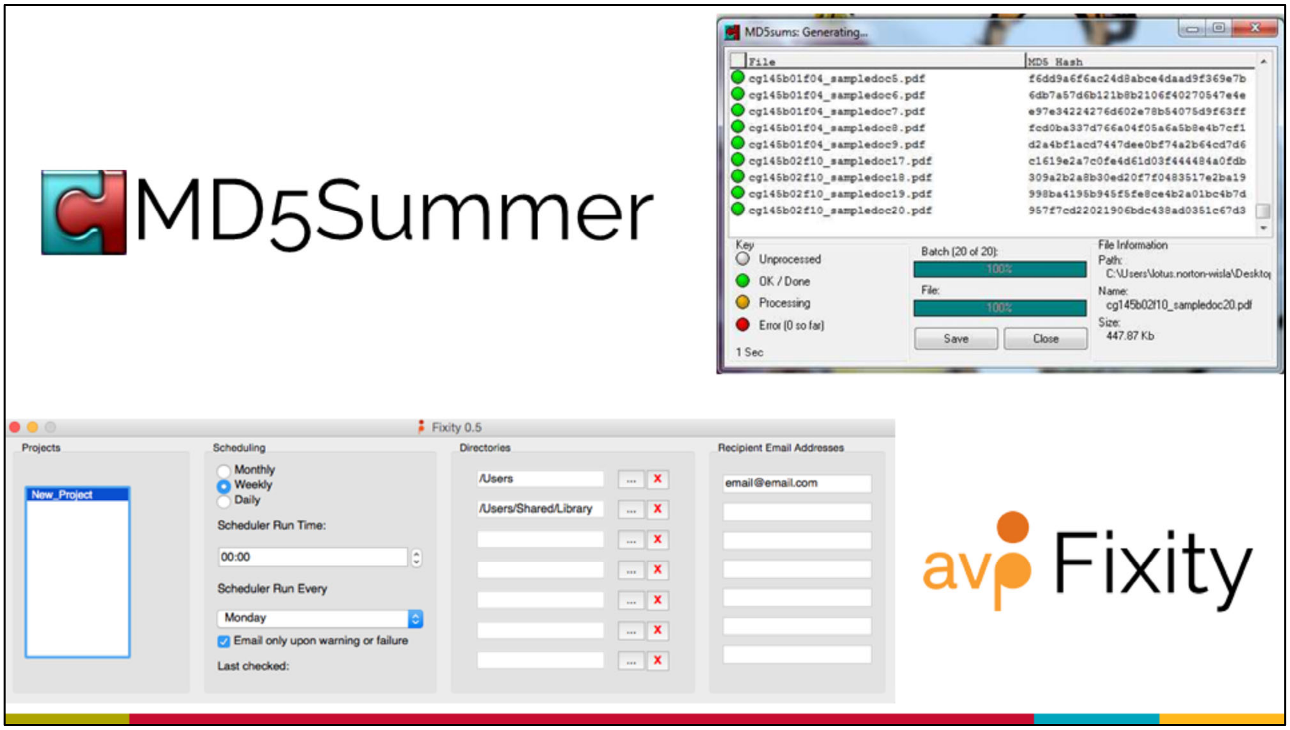
- Need fixity info integrated into a larger Digital Preservation Plan! If your files degrade, you need to find out and act BEFORE that file is backed up incorrectly.



Where is file fixity information stored?

- Alongside technical or preservation metadata
- Within fixity logs generated by your fixity checking software
- Within the digital object itself

- It can be treated as a part of the technical or preservation metadata and stored within the metadata record in the repository or content management system where the digital object resides.External fixity logs can be generated by software packages that include all component parts of a fixity check, the location of the file on the filesystem, the resultant hash of that file, the last time fixity was checked for that file and whether it passed. The screenshot below is from a Fixity report generated by the Fixity software distributed by the AVPreserve.
- Some folks try to store fixity information within the object itself. This is a complicated endeavor as it means the fixity check must only select certain parts of the file to feck file integrity or a wrapper based file format like MXF could be used. This is not for the faint of heart.



MD5sums: Generating...

File	MD5 Hash
🟢 cg145b01f04_sampledoc5.pdf	f6dd9a6f6ac24d8abce4daad9f369e7b
🟢 cg145b01f04_sampledoc6.pdf	6db7a57d6b121b8b2106f40270547e4e
🟢 cg145b01f04_sampledoc7.pdf	e97e34224276d602e78b54075d9f63ff
🟢 cg145b01f04_sampledoc8.pdf	fcd0ba337d766a04f05a6a5b8e4b7cf1
🟢 cg145b01f04_sampledoc9.pdf	d2a4bf1acd7447dee0bf74a2b64cd7d6
🟢 cg145b02f10_sampledoc17.pdf	c1619e2a7c0fe4d61d03f444484a0fdb
🟢 cg145b02f10_sampledoc18.pdf	309a2b2a8b30ed20f7f0483517e2ba19
🟢 cg145b02f10_sampledoc19.pdf	998ba4195b945f5fe8ce4b2a01bc4b7d
🟢 cg145b02f10_sampledoc20.pdf	957f7cd2202190 6bdc438ad0351c67d3

Key
○ Unprocessed
🟢 OK / Done
🟠 Processing
🔴 Error (0 so far)
1 Sec

Batch (20 of 20): 100%
File: 100%

File Information
Path: C:\Users\lotus.norton-wisla\Desktop
Name: cg145b02f10_sampledoc20.pdf
Size: 447.87 Kb

Fixity 0.5

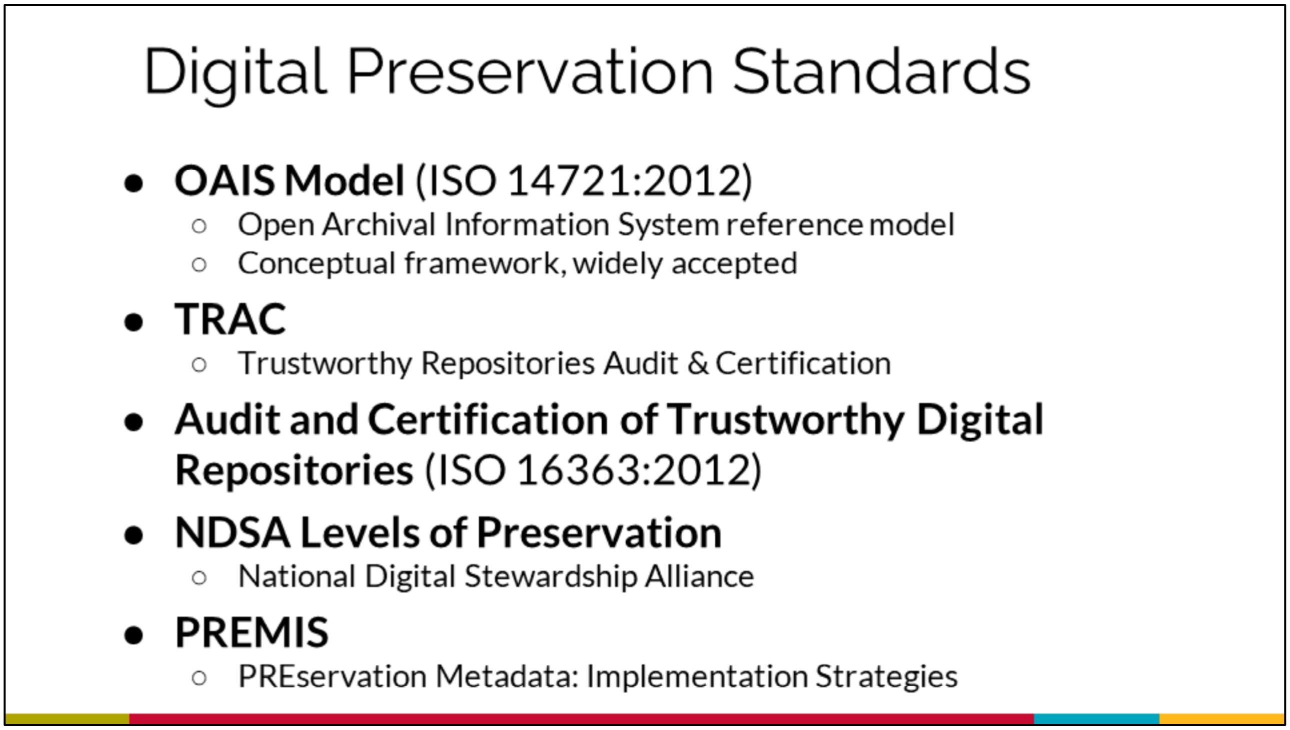
Projects
New_Project

Scheduling
○ Monthly
● Weekly
○ Daily
Scheduler Run Time: 00:00
Scheduler Run Every: Monday
☑ Email only upon warning or failure
Last checked:

Directories
/Users
/Users/Shared/Library

Recipient Email Addresses
email@email.com

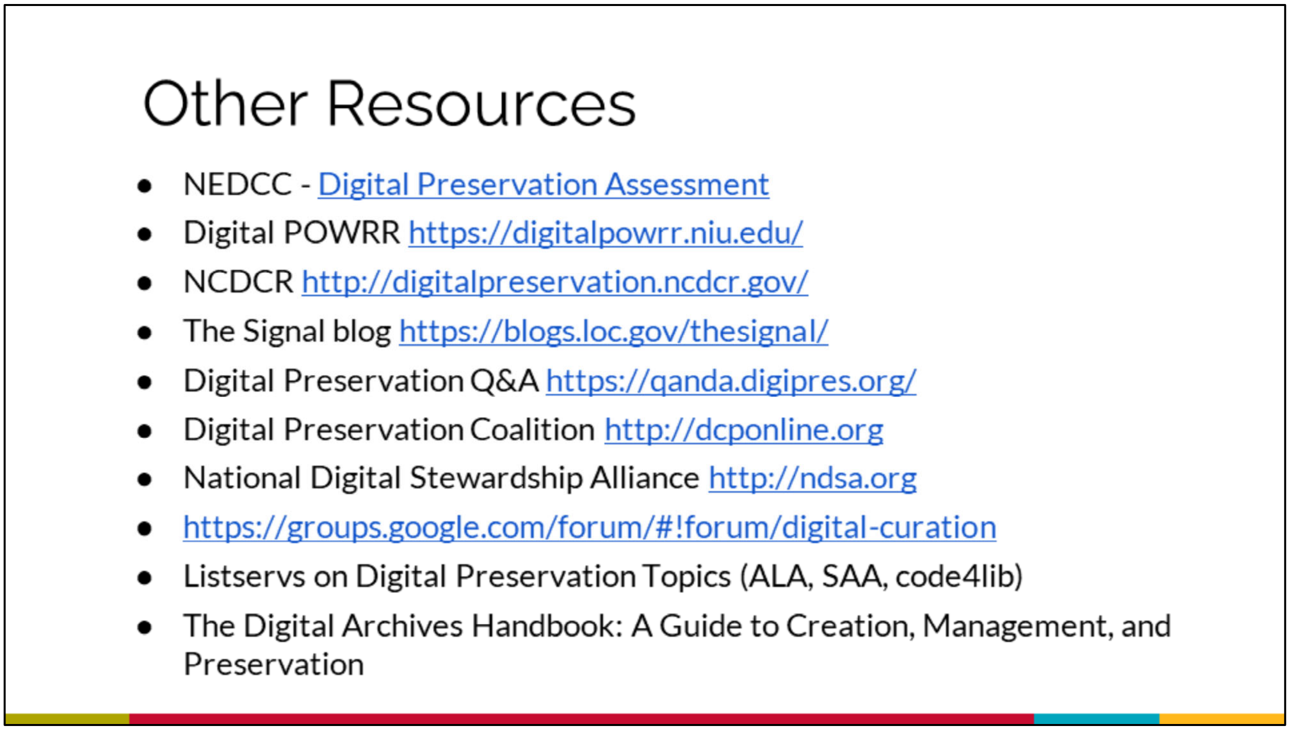
- These are come free tools to start testing out fixity software and understanding how you might create and verify checksums in your own institution.
- Available on the SHN - demonstration of MD5Summer http://www.md5summer.org/
- Available on weareavp.com - demonstration and webinar on Fixity



Digital Preservation Standards

- **OAIS Model** (ISO 14721:2012)
  - Open Archival Information System reference model
  - Conceptual framework, widely accepted
- **TRAC**
  - Trustworthy Repositories Audit & Certification
- **Audit and Certification of Trustworthy Digital Repositories** (ISO 16363:2012)
- **NDSA Levels of Preservation**
  - National Digital Stewardship Alliance
- **PREMIS**
  - PREservation Metadata: Implementation Strategies

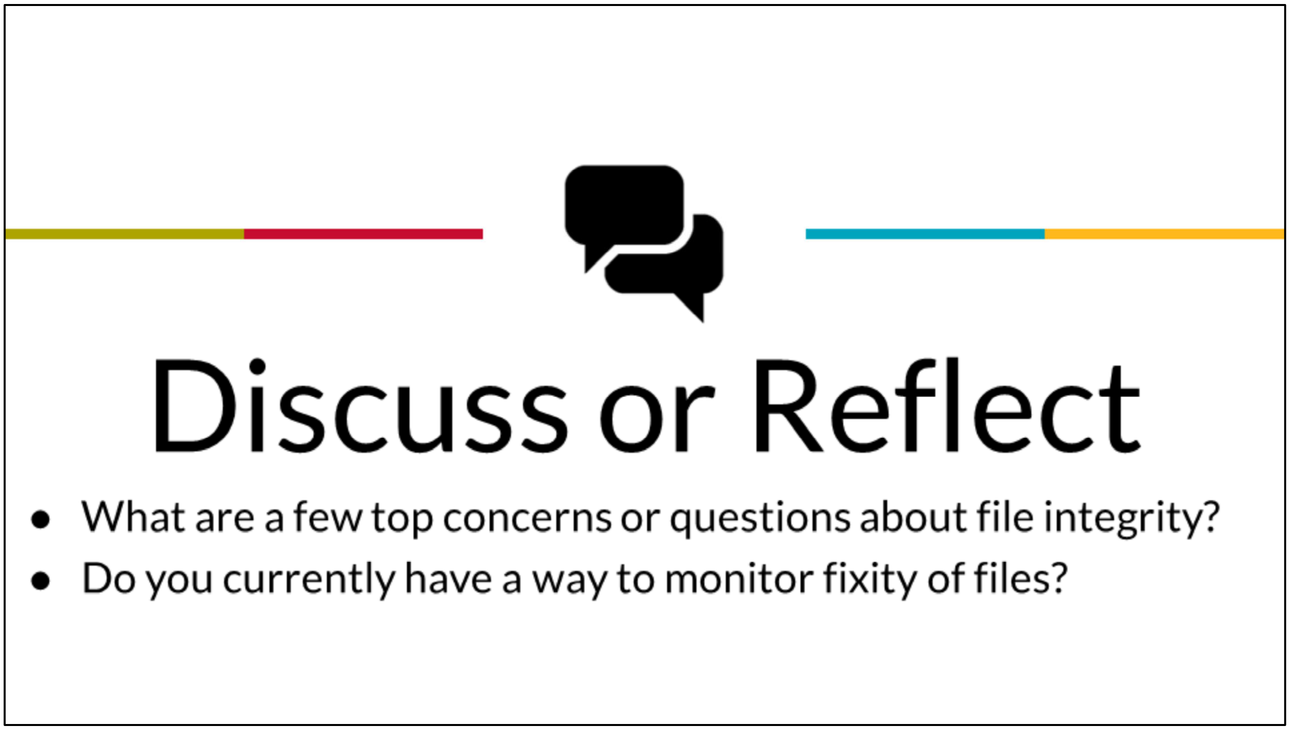
- These are some important concepts to learn more about, reference in your policies and plans, and base your own actions and thinking on - though you will have some things that are specific to your own institution and needs



Other Resources

- NEDCC - Digital Preservation Assessment
- 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
- https://groups.google.com/forum/#!forum/digital-curation
- Listservs on Digital Preservation Topics (ALA, SAA, code4lib)
- The Digital Archives Handbook: A Guide to Creation, Management, and Preservation

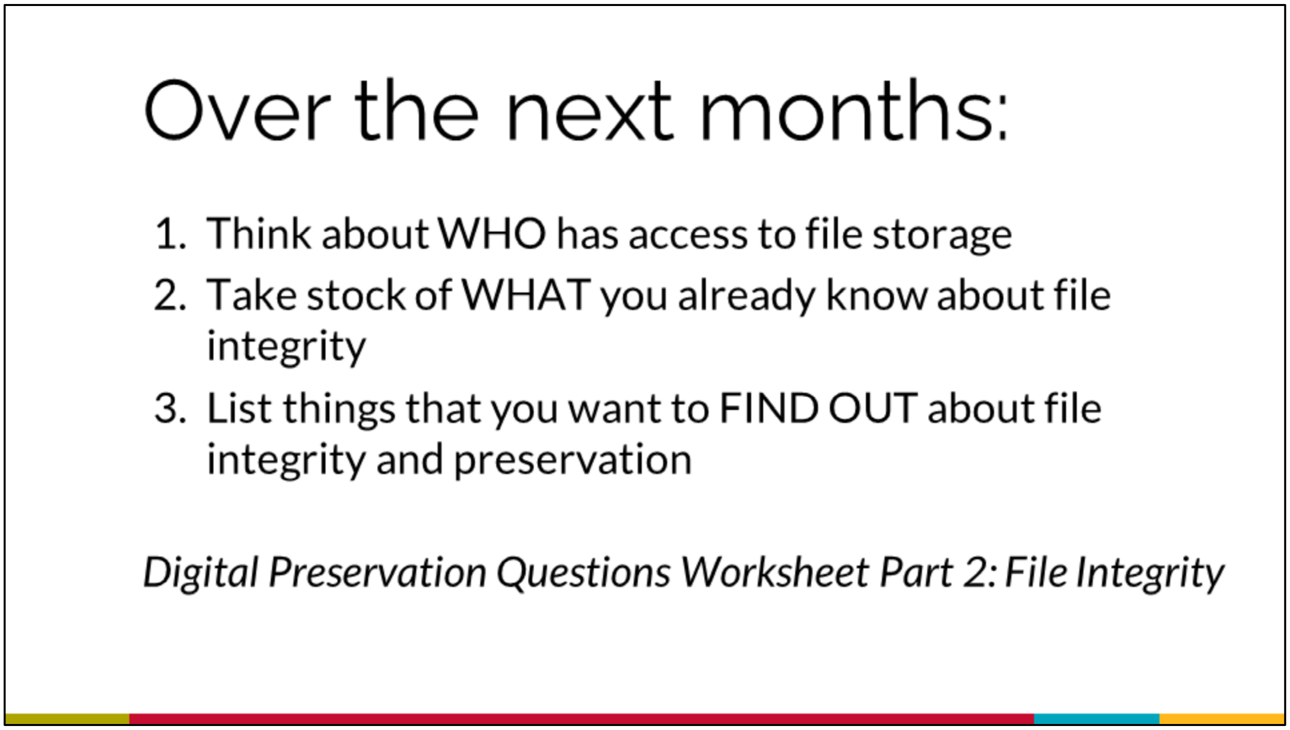
- Here are some places to learn and stay informed. The digital preservation community is growing, and it is a good idea to keep updated on this information.
- UK resource: https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=digital-preservation



Discuss or Reflect

- What are a few top concerns or questions about file integrity?
- Do you currently have a way to monitor fixity of files?

- Take 20-30 minutes and discuss with others, or reflect by yourself and take notes
- What are a few top concerns you have about digital file integrity and preservation? Based on what you have learned so far.
    - Is there anything that you already do in your institution to maintain integrity of files?
    - Is there anything you are particularly worried about?
    - Have you ever seen examples of “bit rot” or data degradation in your work or personal files?
    - Do you think that using a write blocker is a low or high priority for your department?
- Do you currently have a way to monitor fixity of files?
    - If yes, how is this done?
    - Would it be helpful to add on anything to your process?
    - If not, how might you start monitoring file fixity?

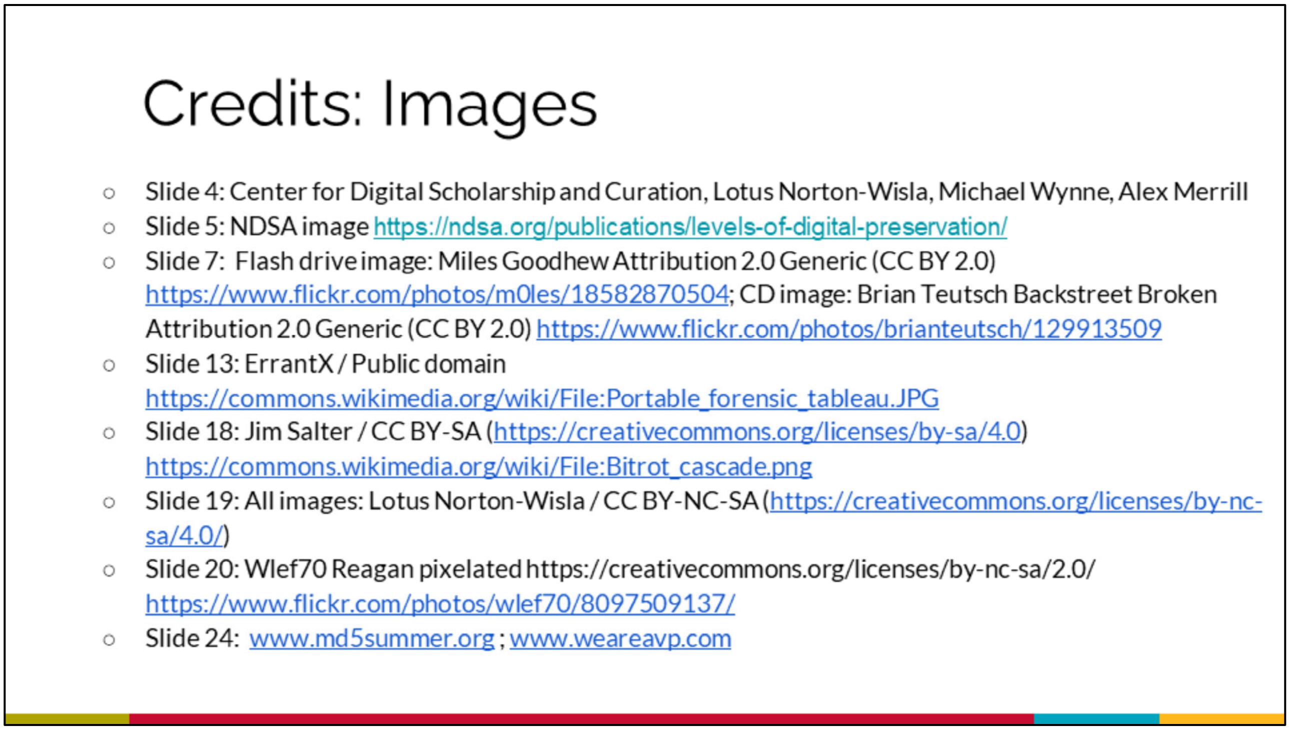


Over the next months:

1. Think about WHO has access to file storage
2. Take stock of WHAT you already know about file integrity
3. List things that you want to FIND OUT about file integrity and preservation

Digital Preservation Questions Worksheet Part 2: File Integrity

- Complete the SHN resource Digital Preservation Questions Worksheet Part 2: File Integrity to get some helpful questions and discussion started around this topic. Bring others into the conversation



Credits: Images

- Slide 4: 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 7: Flash drive image: Miles Goodhew Attribution 2.0 Generic (CC BY 2.0) https://www.flickr.com/photos/m0les/18582870504; CD image: Brian Teutsch Backstreet Broken Attribution 2.0 Generic (CC BY 2.0) https://www.flickr.com/photos/brianteutsch/129913509
- Slide 13: ErrantX / Public domain https://commons.wikimedia.org/wiki/File:Portable_forensic_tableau.JPG
- Slide 18: Jim Salter / CC BY-SA (https://creativecommons.org/licenses/by-sa/4.0) https://commons.wikimedia.org/wiki/File:Bitrot_cascade.png
- Slide 19: All images: Lotus Norton-Wisla / CC BY-NC-SA (https://creativecommons.org/licenses/by-nc-sa/4.0/)
- Slide 20: Wlef70 Reagan pixelated https://creativecommons.org/licenses/by-nc-sa/2.0/ https://www.flickr.com/photos/wlef70/8097509137/
- Slide 24: www.md5summer.org ; www.weareavp.com

Credits: Presentation

- 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.

All presentations and resources created by the CDSC are licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 license (CC BY-NC-SA). Please share, reuse, and adapt the resources and provide attribution to the Center for Digital Scholarship and Curation, Washington State University.