Digital Preservation

Concepts and Principles

There is no starting and stopping point for digital preservation.

— Kara Van Malssen, “Planning Beyond Digitization”

What are Digital Files?

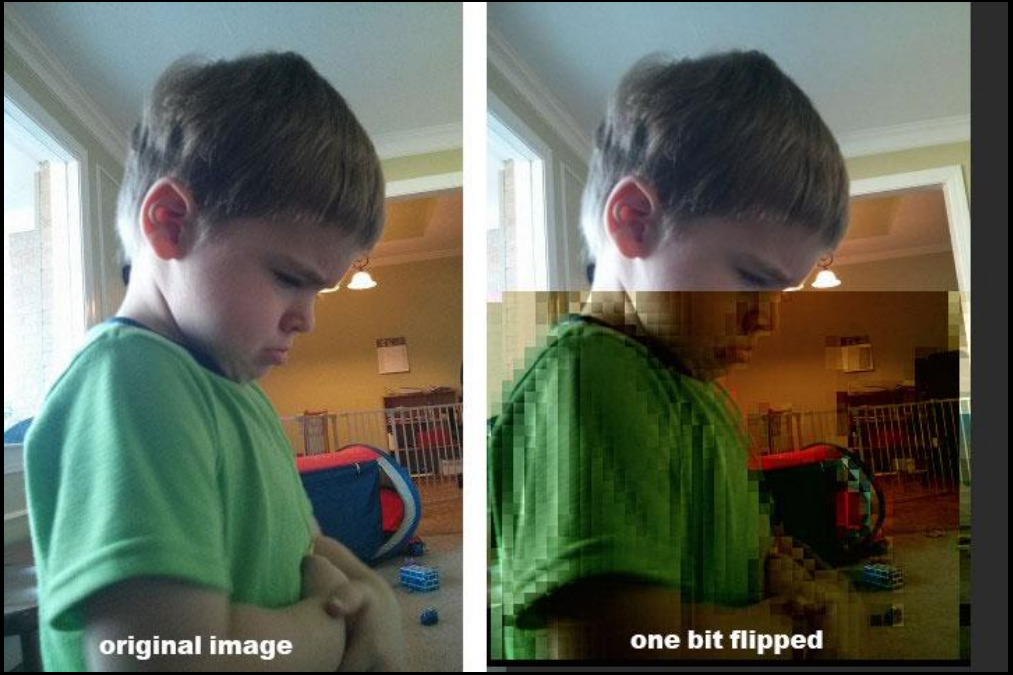
Specific amount of information physically written to some device or media

Part of a filesystem

Examples: Magnetism on a spinning disc HDD or data tape, stored charge in a SSD, pits or dyes on an optical disc (CD or DVD)

Risks to Long-term Preservation of Digital Files

Data Degradation / Data Rot

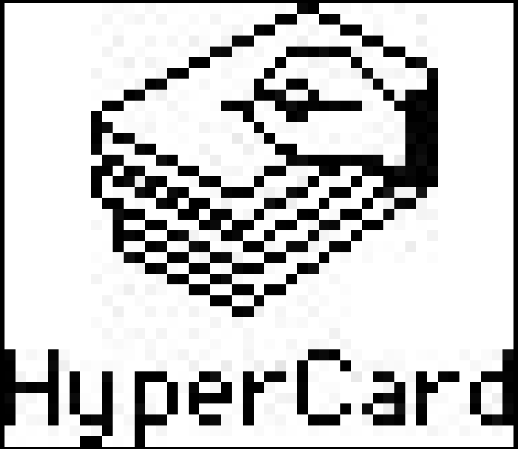


Data rot is a risk to the fixity or integrity of files

Loss of Data

Data can become extinct through a natural disaster

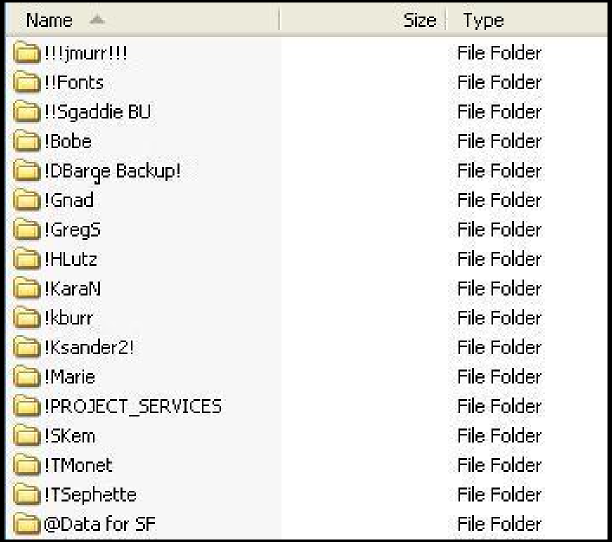
Obsolescence



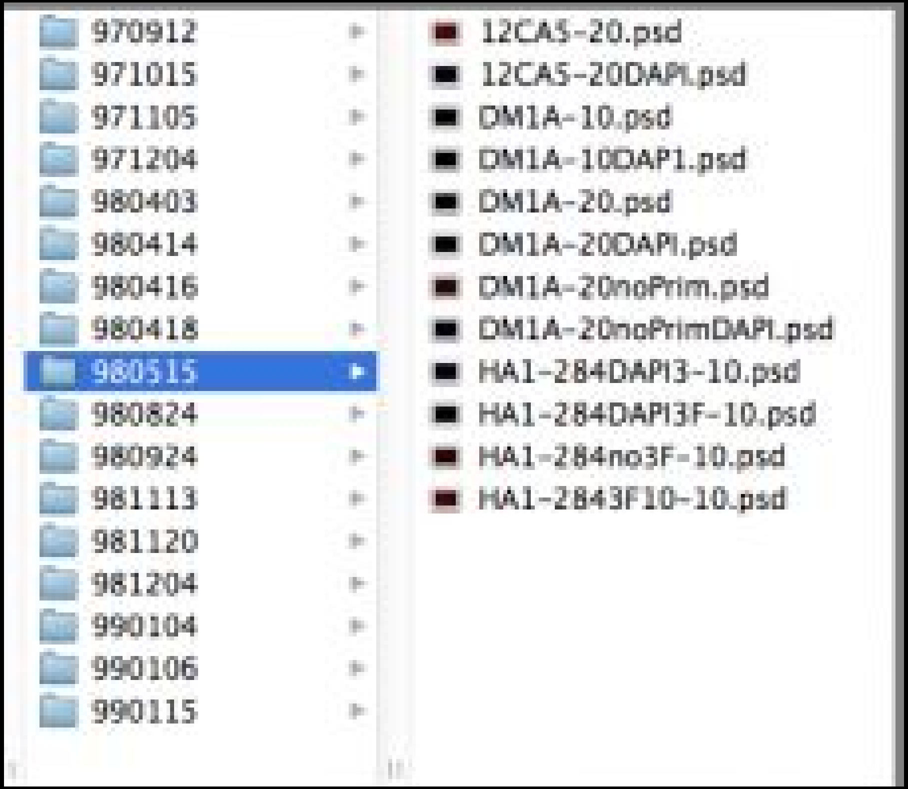
Obsolescence of file format

Obsolescence of storage media

Lack of Metadata



Name	Size	Type
!!!jmurr!!!		File Folder
!!Fonts		File Folder
!!Sgaddie BU		File Folder
!Bobe		File Folder
!DBarge Backup!		File Folder
!Gnad		File Folder
!GregS		File Folder
!HLutz		File Folder
!KaraN		File Folder
!kburr		File Folder
!Ksander2!		File Folder
!Marie		File Folder
!PROJECT_SERVICES		File Folder
!SKem		File Folder
!TMonet		File Folder
!TSephette		File Folder
@Data for SF		File Folder



970912
971015
971105
971204
980403
980414
980416
980418
980515
980824
980924
981113
981120
981204
990104
990106
990115

12CAS-20.psd
12CAS-20DAPI.psd
DM1A-10.psd
DM1A-10DAP1.psd
DM1A-20.psd
DM1A-20DAPI.psd
DM1A-20noPrim.psd
DM1A-20noPrimDAPI.psd
HA1-284DAPI3-10.psd
HA1-284DAPI3F-10.psd
HA1-284no3F-10.psd
HA1-2843F10-10.psd

Questions?

What are the basic rules and strategy of the sport volleyball?

Strategies for Preservation of Digital Files

Fixity / Integrity Checks

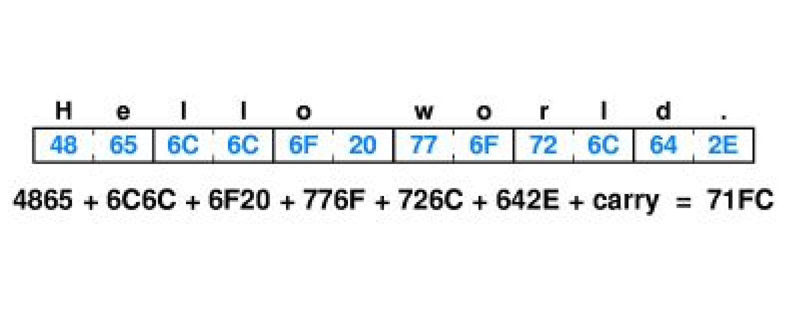
What’s a Checksum ?

A digital “signature” for a file

An algorithm goes through every bit and calculates checksum

Also called a “cryptographic hash,” “hash value,” or simply “hash”

Examples: SHA-1, SHA-2, MD5

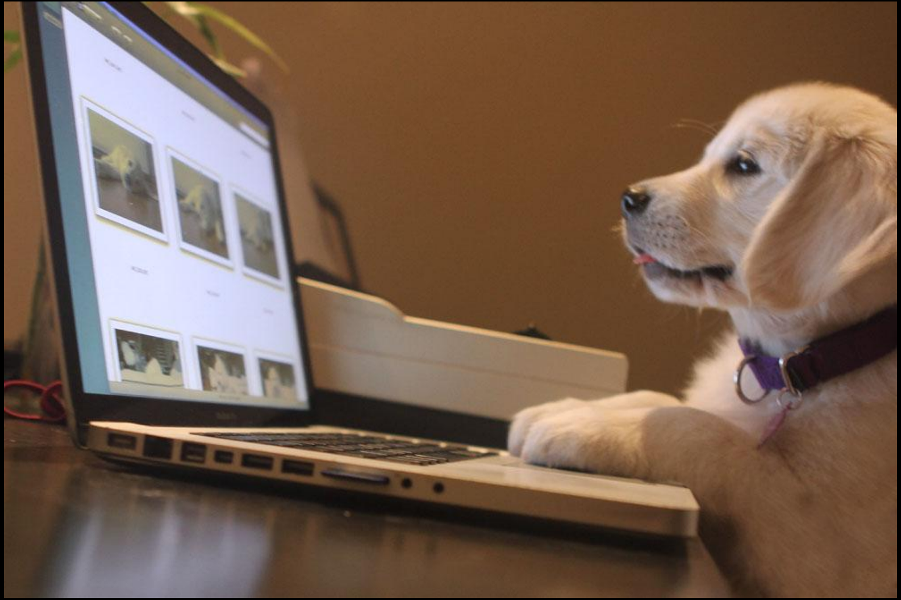


H	e	l	l	o		w	o	r	l	d	.
48	65	6C	6C	6F	20	77	6F	72	6C	64	2E

4865 + 6C6C + 6F20 + 776F + 726C + 642E + carry = 71FC

An example of how a checksum is computed

Demo Time



Scheduling Fixity Checks

Should be automated - open source programs and scripts exist to run fixity checks

Not too often: A fixity check could cause a read/write error or wear of storage media

Duplication of Files

Backup Strategies

Create many copies

Store copies on different media

Store copies in separate locations

Check backups regularly

Securing Data

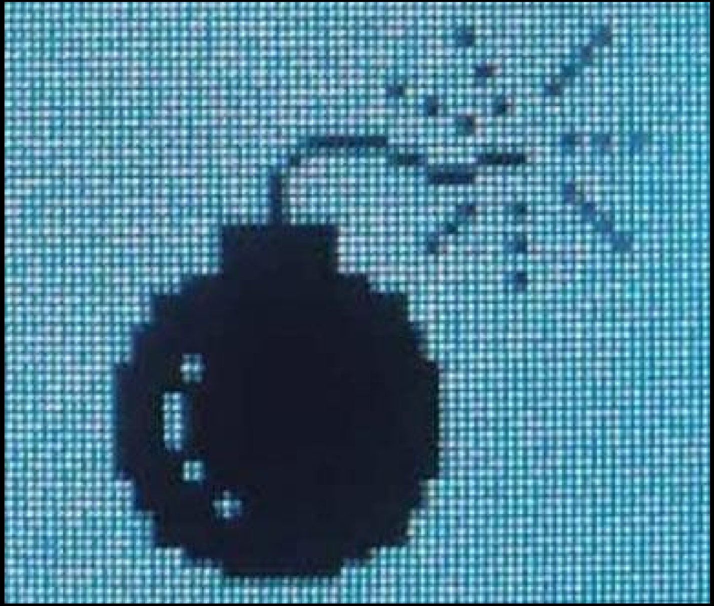
Risks to Consider

Computer viruses

Hackers

Researchers

Employees



Refreshing of Files

Storage Media

Spinning Disc
Hard Drive

Solid State
Hard Drive

Data Tape
(LTO)

Optical Disc

Migration of Files

Selecting a Format

Compression - Does it use lossy compression, lossless compression, or no compression?

Openness - Does one company control the format?

Documentation - Do detailed documents about its specifications exist? Is it easy to get those documents?

Self-descriptiveness - How good is the format's metadata?

Universality/ubiquity - How much is it being used the real world?

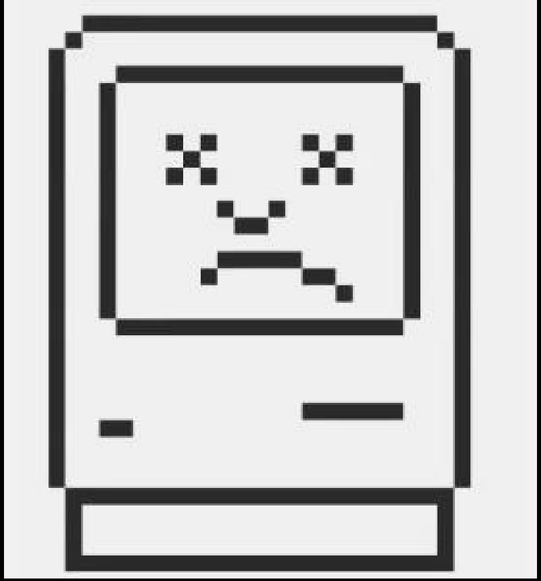
Issues with Compression

Lossy compression discards data forever!

Data rot or errors have a greater effect on compressed files

The codec used for compression may not be supported in the future

The codec may not be well-documented



Demo Time

Monitoring for Obsolescence

Pay attention to the film/video/digital production world

Obsolescence of born-digital formats will happen at a faster pace than analog

File formats widely used on the internet may be more resistant to obsolescence

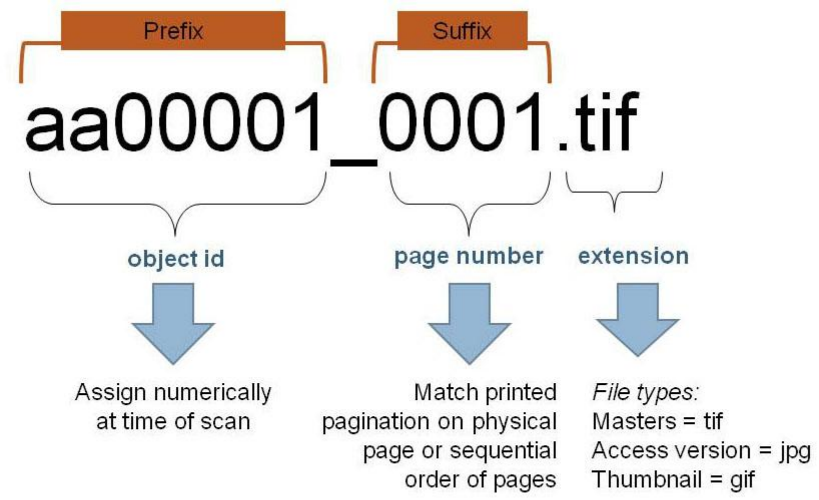
Creating/Maintaining Appropriate File Metadata

Naming Conventions

Conventions are necessary at both the file and the folder levels

Value consistency above almost everything else

Make conventions readable by all computer operating systems



Embedded Metadata and Sidecar Files

Both strategies can be used

As with file format, choose a metadata standard that is open and universally accepted

Be Careful: Any program used to manipulate files at an archive should not strip metadata from files

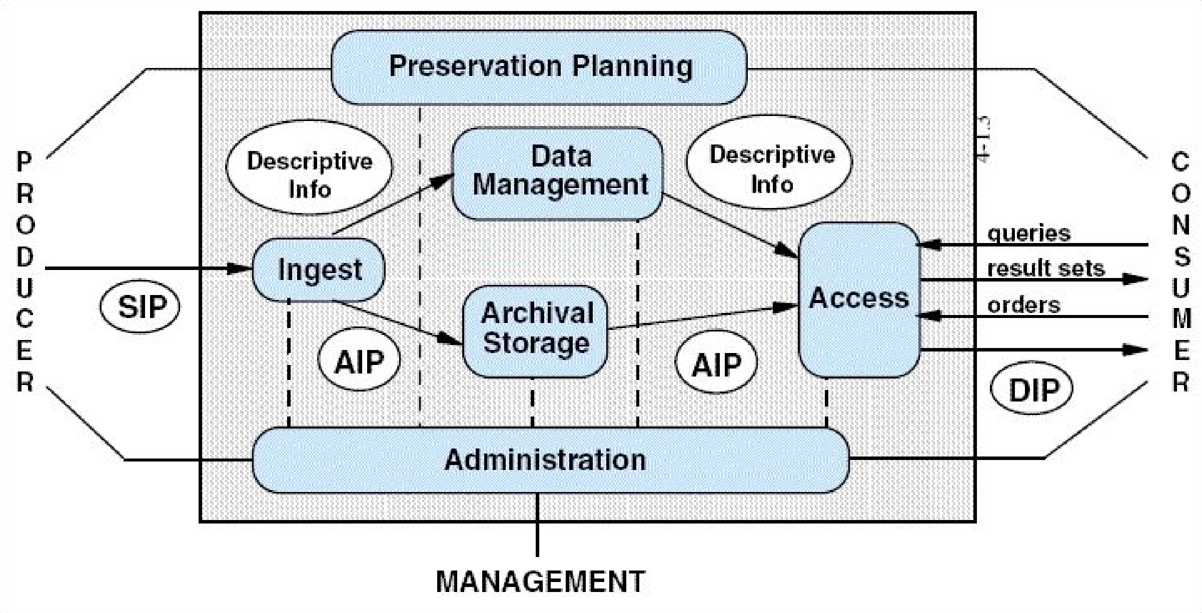
Sustainability of a Digital Preservation Program

If an organization chooses to no longer support the digital preservation environment – either due to bankruptcy, change of mission, or simply a lack of funds – the digital resources risk disappearing.

— Kara Van Malssen, “Planning Beyond Digitization”

Digital Repositories

The OAIS Model



Submission Information Package (SIP)

Archival Information Package (AIP)

Distribution Information Package (DIP)

Submission Information Package (SIP)

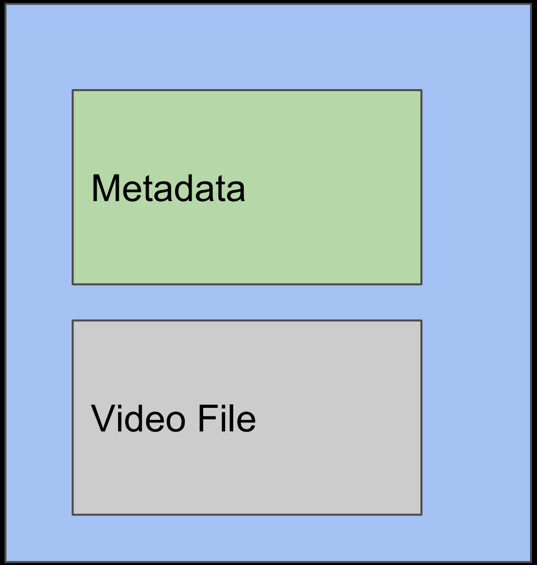
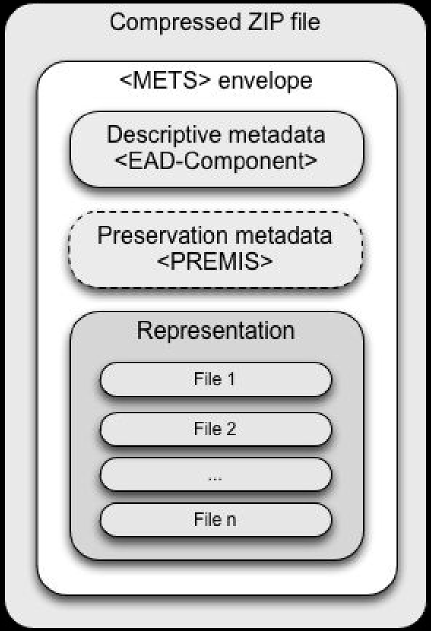


Diagram of a Simple SIP



Typical SIP for a digital repository

NDSA Levels of Digital Preservation

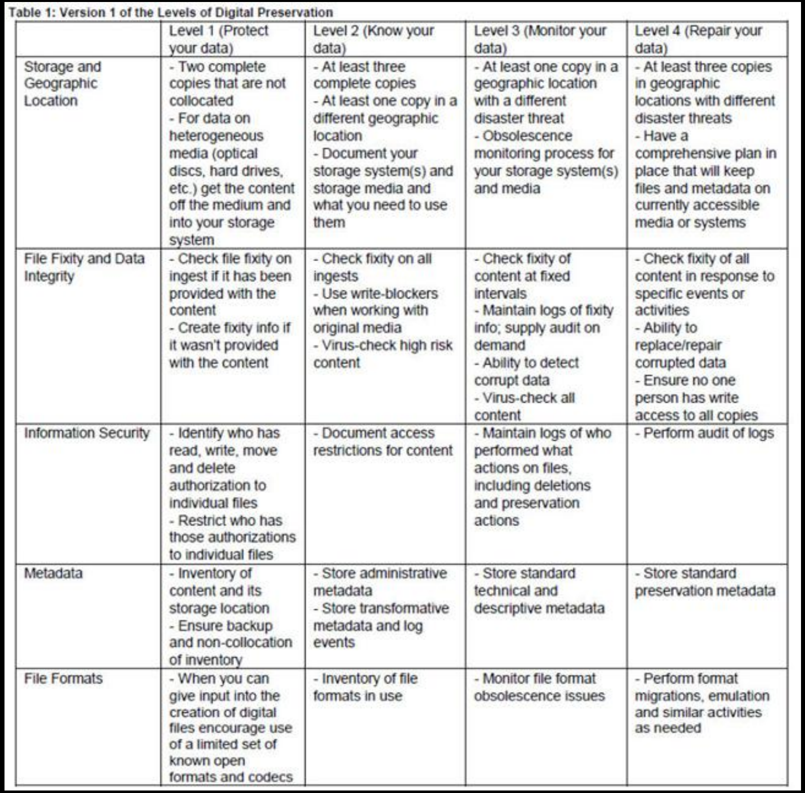


Table 1: Version 1 of the Levels of Digital Preservation

	Level 1 (Protect your data)	Level 2 (Know your data)	Level 3 (Monitor your data)	Level 4 (Repair your data)
Storage and Geographic Location	- Two complete copies that are not collocated - For data on heterogeneous media (optical discs, hard drives, etc.) get the content off the medium and into your storage system	- At least three complete copies - At least one copy in a different geographic location - Document your storage system(s) and storage media and what you need to use them	- At least one copy in a geographic location with a different disaster threat - Obsolescence monitoring process for your storage system(s) and media	- At least three copies in geographic locations with different disaster threats - Have a comprehensive plan in place that will keep files and metadata on currently accessible media or systems
File Fixity and Data Integrity	- Check file fixity on ingest if it has been provided with the content - Create fixity info if it wasn't provided with the content	- Check fixity on all ingests - Use write-blockers when working with original media - Virus-check high risk content	- Check fixity of content at fixed intervals - Maintain logs of fixity info; supply audit on demand - Ability to detect corrupt data - Virus-check all content	- Check fixity of all content in response to specific events or activities - Ability to replace/repair corrupted data - Ensure no one person has write access to all copies
Information Security	- Identify who has read, write, move and delete authorization to individual files - Restrict who has those authorizations to individual files	- Document access restrictions for content	- Maintain logs of who performed what actions on files, including deletions and preservation actions	- Perform audit of logs
Metadata	- Inventory of content and its storage location - Ensure backup and non-collocation of inventory	- Store administrative metadata - Store transformative metadata and log events	- Store standard technical and descriptive metadata	- Store standard preservation metadata
File Formats	- When you can give input into the creation of digital files encourage use of a limited set of known open formats and codecs	- Inventory of file formats in use	- Monitor file format obsolescence issues	- Perform format migrations, emulation and similar activities as needed

Contact

Email: jfarbowitz@gmail.com

Twitter: @jfarbowitz