Tools for digital preservation / Code4Lib Toronto / March 28th, 2018

Your host today is Grant Hurley, Digital Preservation Librarian, Scholars Portal

Agenda

1. What’s this “digital preservation” thing?
2. Go with the flow
3. Need more?
4. Questions?

What's this "digital preservation" thing?

What's this "digital preservation" thing?

- Digital objects (both born digital and digitized) need active management to ensure ongoing access
- Quickly-changing technological norms create risks that must be managed from an object's creation forward
- Digital preservation is a set of theories and practices that work to keep digital objects **authentic**, **available** and **reliable** over time.

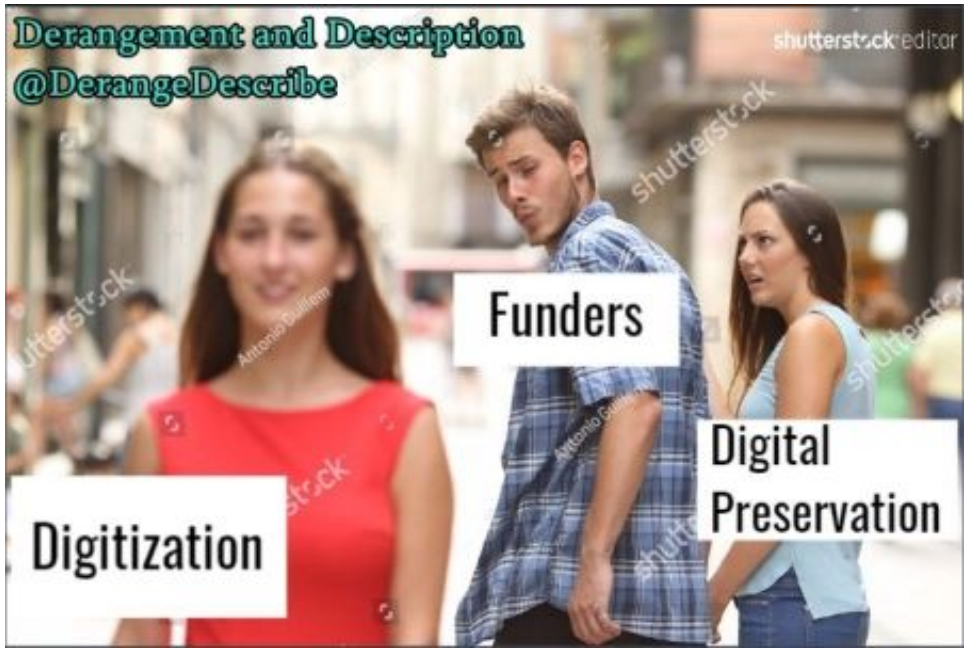


Image: Derangement and Description

Authenticity boils down to:

Identity: what it is; format identification, descriptive information, provenance information, etc.

Integrity: establishing that a file remains unaltered over time

Availability: ensures that records are accessible into the future by periodically migrating copies of digital objects to new formats.

Reliability: is a combination of authenticity and availability - a reliable digital object can be trusted when proof of authenticity and availability are transparent.

"Doing" digital preservation needs:

- **Tools** and **infrastructure** to process digital objects and then keep them alive into the future

- **People** and **money** to implement these tools and infrastructure

- Sets of **standards**, **policies**, and **practices** that mature and develop over time in response to technological changes

Key steps in a preservation workflow:

- Transfer/Capture
- Establish fixity
- Identify
- Characterize
- Validate
- Normalize
- Store
- Manage

Go with the flow: Transfer/ Capture

Bringing materials under the custody of the stewarding organization so they can be processed for preservation and access

City of Toronto Archives, Fonds 1266, Item 15419

Transfer/Capture

- If already on an accessible system (connected computer, server) can be transferred the network - FTP, cloud-based transfer point, shared network etc.

- If on external media (CDs, USBs, floppy disks!) a whole set of digital forensics processes kick in to ensure content is retrieved safely without modifying it
  - This subject is deeply out of scope for today.
  - **But** check out Jess Whyte's upcoming workshop:
    Preservation in a Historical Computing Environment - How to Recover Information from Vintage Tech

    April 27, 2018 - hosted by the Toronto Area Archives Group (TAAG):
    https://aao-archivists.ca/event-2863782

Go with the flow: Establish fixity

Establishing checksums as soon as possible so that a chain of integrity can be demonstrated and validated as required into the future

City of Toronto Archives, Fonds 1568, f1568_it0240

Integrity: The almighty checksum

md5 checksum =
cf8d829ca657ee3860c3434294d1bae6

md5 checksum =
cc1fae67e8a61f6fcb4b38cf8f72af5f

Corrupted with [Image Glitch Tool]

Establish Fixity

- Easy way to implement fixity processes is by using tools that conform to the **BagIt specification**
- Bags are files wrapped with a simple text files that describe them in a predictable way
  - Original files get put in a folder called ‘data’
  - An inventory with checksums is created
  - Metadata about the bag can be added

- Bagit is intended for transfer workflows - to send files from one place to another with integrity
  - But it’s also great for internal use

Establish Fixity

- Various tools implement Bagit:
  - Exactly by AV Preserve (Nice GUI, but changes file modified dates)
  - Bagger by Library of Congress (Gnarlier GUI, does not change dates)
  - Bag It Python by Library of Congress (No GUI, lightweight & easy)

- Can use Bagit tools as part of capture and transfer workflows by validating bags on the other end

Go with the flow: Identify

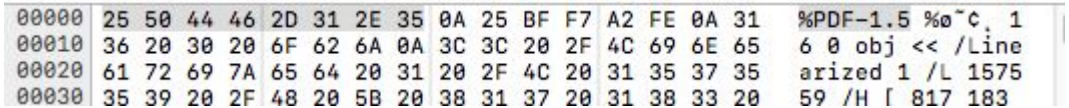
Determine what file format and version a particular file is

City of Toronto Archives, Fonds 1266, Item 4469

Identity: File formats

- The key is to identify the file's signature rather than its extension or MIME type
  - A signature is a series of bytes that occur in a predictable manner at the beginning (usually) of a file

- PDF 1.5 file in hex editor



```
00000  25 50 44 46 2D 31 2E 35  0A 25 BF F7 A2 FE 0A 31   %PDF-1.5 %¿÷¢þ 1
00010  36 20 30 20 6F 62 6A 0A  3C 3C 20 2F 4C 69 6E 65   6 0 obj << /Line
00020  61 72 69 7A 65 64 20 31  20 2F 4C 20 31 35 37 35   arized 1 /L 1575
00030  35 39 20 2F 48 20 5B 20  38 31 37 20 31 38 33 20   59 /H [ 817 183
```

- File format signature in PRONOM

File extension: pdf

Name	PDF 1.5	
Description	BOF: %PDF-1.5 EOF (offset up to 1024 bytes): %%EOF	
Byte sequences	**Position type**	Absolute from BOF
	Offset	0
	Byte order	
	Value	255044462D312E35
	Position type	Absolute from EOF
	Offset	0

Identify

- Various tools for this function.

- Good for spot IDs:
  - FIDO
  - Siegfried
- Batch processing & error reporting:
  - DROID

Go with the flow: Characterize files

The process of extracting metadata related to a file's intrinsic properties

City of Toronto Archives, Fonds 200, Series 376, File 3, Item 1b

Characterize

- Useful to get to know the components of an individual file better
- Can provide detailed information on quality characteristics for audiovisual materials, photographs, etc.
- Provides reliable information on created and modified dates

- Common tools used:
  - ExifTool (lots of formats, but especially good at images)
  - Apache Tika (nearly everything)

Go with the flow: Validate

The process of determining if a file is well-formed and valid according to its specification

City of Toronto Archives, Fonds 16, Series 71, Item 9910

Validate

- File formats have specifications that dictate how files are structured and interpreted
- Some file formats have these specifications published
- Validating a file means confirming that it is well-formed according to these specifications
    - Purpose is to ensure that files being stored have not been corrupted/are of necessary quality for long-term storage

- Tools:
    - JHOVE - images, documents
    - VeraPDF - PDF/A
    - MediaConch - video

Identify, Characterize, Validate

- Bundles many tools together:
  - FITS (File Information Tool Set)

Go with the flow: Normalize

The process of converting files from source formats to designated preservation or access formats/specifications

City of Toronto Archives, Fonds 200, Series 376, File 5, Item 10

Normalize

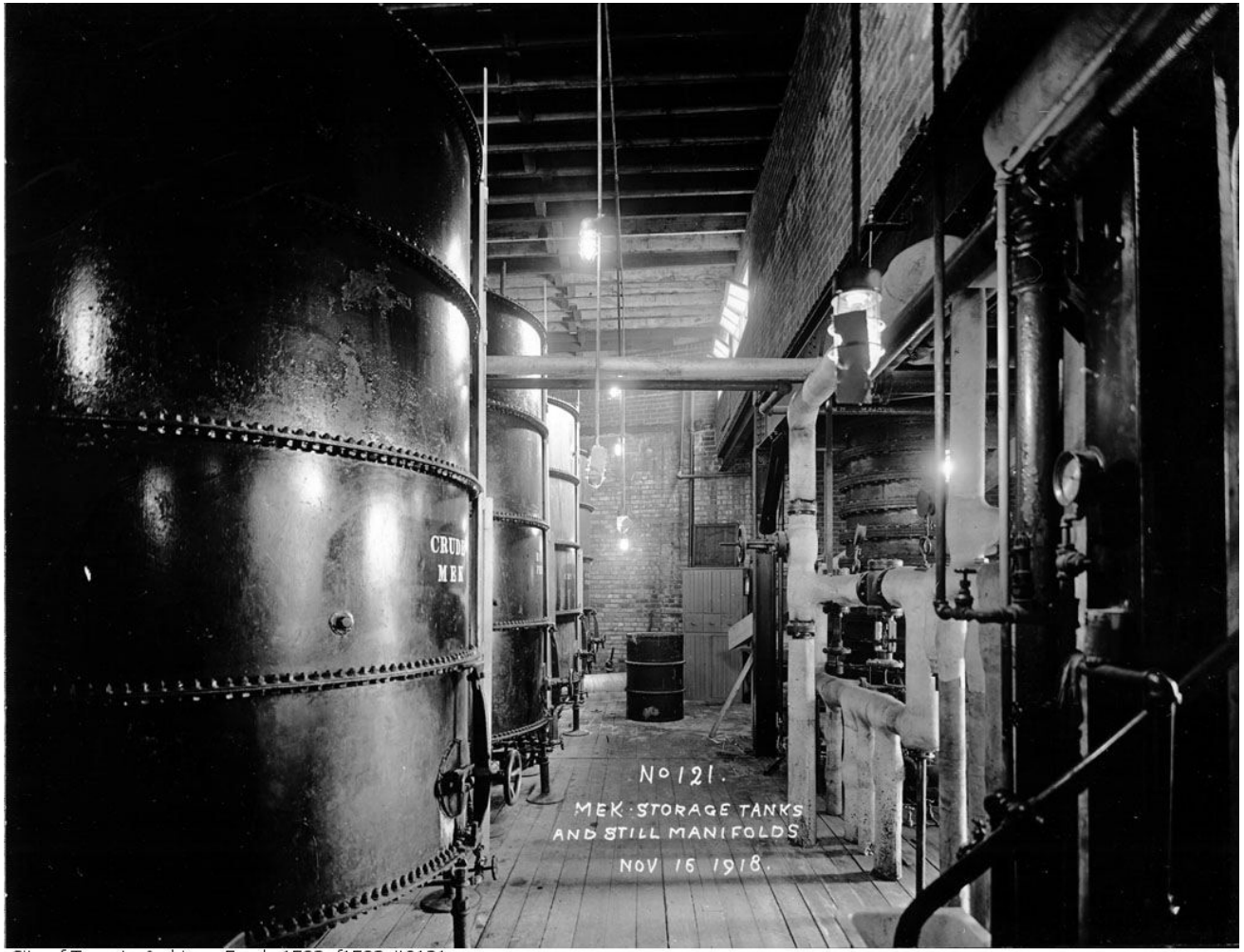
- Two uses: access and preservation
- Access derivatives are usually smaller file sizes in common formats
- Preservation copies are normalized to a standard set of files based on institutional policies

Various tools support normalization:

- Convert (ImageMagick): images
- FFMPEG: audio/video
- Libreoffice (Headless): office documents
- Ghostscript: PDF/A
- Inkscape: other PDF and SVG

Go with the flow: Store

Put the files and associated metadata somewhere safe



City of Toronto Archives, Fonds 1583, f1583_it0121

Store

- No specific tools - you need access to infrastructure!
  - Ideally: replicated in multiple places
  - Available when you need it
  - Not a removable hard drive or other risky media
  - Can enable oversight and management

- Draft digital preservation storage criteria specification available
- InterPARES checklist for cloud service contracts

Go with the flow: Manage

Keep an eye on the files and metadata into the future

City of Toronto Archives, Fonds 1244, Item 31

Manage

- Digital preservation doesn't stop once you've stored a thing
- Keep an eye on file formats over time and migrate as needed
- Check fixity on a regular basis
    - AVPreserve's Fixity can be used for this purpose

- Current lack of comprehensive, lightweight management tools in this space

Looking for more?

- Look into Archivematica!
- Think about PREMIS-in-METS files!
- Get excited about digital forensics!
- Dig into standards!
  - Open Archival Information System Standard (OAIS): ISO 14721
  - Trustworthy Digital Repository certification standard: ISO 16363
- Read up on digital preservation!

Questions? Feelings? Ideas?

Get in touch:
grant@scholarsportal.info

www.granthurley.ca

Thanks to the City of Toronto Archives for making beautiful images from their collections available online!