An Introduction to Archivematica

For INF 2122H

October 31, 2019

Grant Hurley, Digital Preservation Librarian, Scholars Portal

Agenda

- Where it came from
- What it does - the short version
- What it isn’t
- Who uses it
- What its functions are
- Preparing transfers
- Processing transfers
- Looking at the outputs
- Final thoughts

Where it came from

- Standards for digital preservation developed in late 1990s and early 2000s, but no easy way of applying them
- UNESCO released 2007 report advocating for open source digital preservation system
- Artefactual Systems started up by creating Access to Memory (AtoM) platform for archival descriptions
- Various small open source tools were also being developed by others for particular tasks
- Artefactual developed Archivematica beginning in 2008
- Beta release in 2012; current release is 1.10 (2019)

What it does

- **The goal**: create well-formed data packages of digital objects, including metadata about those objects, for long-term preservation and access

- Takes a pre-structured transfer from a data source
- Makes a Submission Information Package (SIP)
- Transforms the SIP into an Archival Information Package (AIP) for preservation storage
- Also can create a dissemination information Package (DIP) for access

What it does con’t

- Stores and applies format policies for preservation normalization and access copies
- Allows access to, and deletion of, AIPs
- Assists in ingest of descriptive metadata, rights information
- Manages data flows in and out of system through separate Storage Service module
- Can connect to access systems for DIP deposit (mostly just AtoM)
- Can be fully automated

What it isn’t

- A storage system
- An access system
- Easy to install or maintain in production
- User friendly
- A complete digital archives workflow

Who uses it

Largely, memory institutions (libraries, archives, galleries, museums) with digital collections that need preserving

- Libraries:
    - Digitized/born-digital content in institutional repositories
    - Research data
    - Digital collections (books, journals, maps, etc.)
- Archives
    - Digitized collections (photographs, audio-visual materials, etc.)
    - Born digital donations (all sorts of stuff)
        - Private papers/collections
        - Records from corporate bodies, institutions, etc.

What it is

“Archivematica is a **web- and standards-based**, **open-source** application which allows your institution to preserve long-term access to **trustworthy, authentic and reliable** digital content.” - Archivematica website

“Web- and Standards-based”

Web-based part:

- Accessed through a web-based dashboard
- This does not mean it is publicly accessible
- Typically installed as a virtual machine on a server and deployed to a local network
- This VM needs adequate resources: CPUs, RAM and disk space

“Web- and Standards-based”

Standards-based part:

- Explicitly modeled on OAIS
  - Uses concepts of SIPs, AIPs, and DIPs directly in workflow

- Adopts metadata standards
  - Simple Dublin Core (for descriptive metadata)
  - METS (XML wrapper for information about digital objects)
  - PREMIS (preservation metadata standard)

OAIS - SIPs, AIPs, DIPs

Conceptual - and actual - data packages managed by an OAIS archive

Include both digital objects and metadata

SIP = Submission information package

- The version of the information package when it is ready to be ingested in the archive.

AIP = Archival information package

- The version of the information package when it is stored and maintained by the archive.

DIP = Dissemination information package

- The version of the information package made available to consumers.

Metadata:
METS (Metadata Encoding and Transmission Standard)

- XML-based metadata standard
- A container for metadata about digital objects
- Intended for transferring data about digital objects between systems
- Can contain PREMIS metadata

- Main sections:
    - **Descriptive metadata** (identifies objects; DublinCore often used)
        - This gets ingested into Archivematica via CSV or information entered via the interface

    - **Administrative metadata** (records of events, agents and outcomes e.g. fixity check, file identification - PREMIS used here)
        - The bulk of the METS file created by Archivematica is here

    - **File section** (files in AIP and relationships between them)
    - **Structural map** (links all elements together)

Metadata: PREMIS

- Came from the **Pre**servation **M**etadata: **I**mplementation **S**trategies working group
- Initially released in 2005
- Sets out core terms for preservation metadata organized around:

Intellectual entities - objects that can be described
Objects - the digital objects themselves
Events - actions that involve an object
Agents - people, organizations or software that perform events (and otherwise)
Rights - asserts what actions can be taken and by whom
Environments - hardware or software required to process or interpret objects

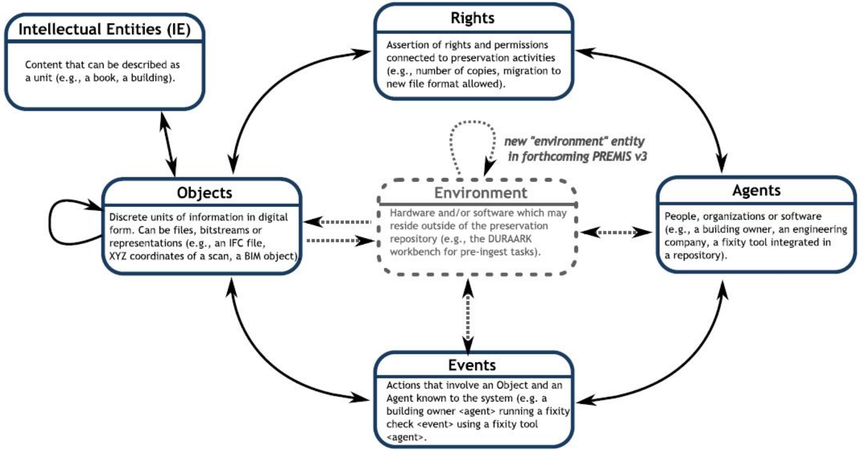
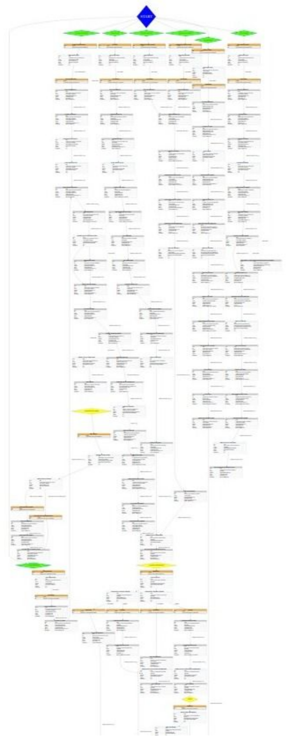

Intellectual Entities (IE)
Content that can be described as a unit (e.g., a book, a building).
Rights
Assertion of rights and permissions connected to preservation activities (e.g., number of copies, migration to new file format allowed).
new "environment" entity in forthcoming PREMIS v3
Objects
Discrete units of information in digital form. Can be files, bitstreams or representations (e.g., an IFC file, XYZ coordinates of a scan, a BIM object).
Environment
Hardware and/or software which may reside outside of the preservation repository (e.g., the DURAARK workbench for pre-ingest tasks).
Agents
People, organizations or software (e.g., a building owner, an engineering company, a fixity tool integrated in a repository).
Events
Actions that involve an Object and an Agent known to the system (e.g. a building owner <agent> running a fixity check <event> using a fixity tool <agent>.

Image source: DURAARK

“Open source”

- Free, code open on GitHub
- Integrates a large number of open-source tools in a “micro-services” architecture
- Developed primarily by Artefactual Systems Inc. via “bounty model” of development



Micro-services chainlinks in Archivematica from @archivalistic

“Trustworthy, Authentic and Reliable”

A definition of digital preservation I like to use:

Digital preservation is a set of theories and practices that work to keep digital objects **authentic**, **available** and **reliable** over time.

Authenticity

Breaks down into:

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

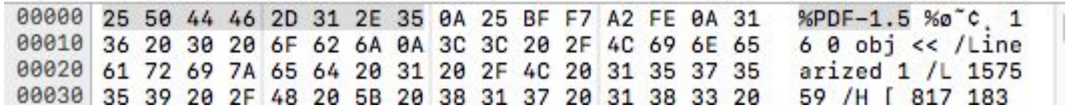
Integrity: establishing that a file remains complete and unaltered over time

Identity: File formats

- Determine what file format and version a particular file is

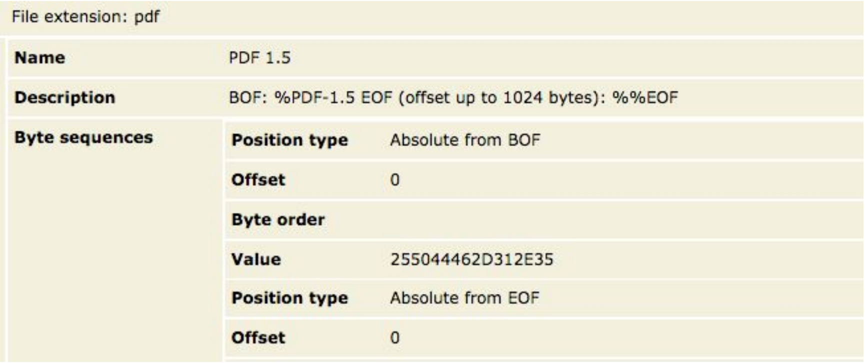
- The key is to identify the file's signature rather than its extension
  - A signature is a series of bytes that occur in a predictable manner at the beginning (usually) of a file
  - Many old file types do not have them

- 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 %ø˜¢¸ 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

- Tools for file format identification:
  - Siegfried
  - FIDO

Identity: Characterization

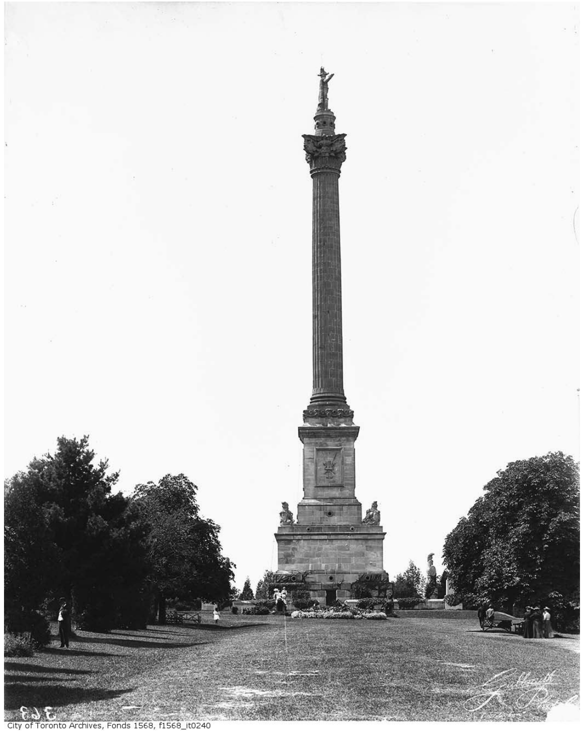
- The process of extracting metadata related to a file’s intrinsic properties

- 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](#) (images)
  - [MediaInfo](#) (audio and video)
  - [Ffprobe](#) (video)

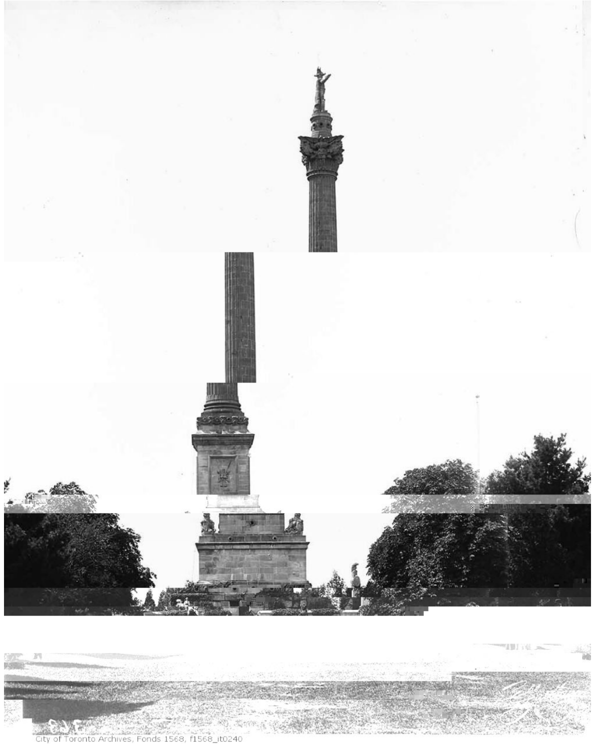
City of Toronto Archives, Fonds 200, Series 376, File 3, Item 1b

Integrity: The almighty checksum



City of Toronto Archives, Fonds 1568, f1568_it0240

md5 checksum =
cf8d829ca657ee3860c3434294d1bae6



City of Toronto Archives, Fonds 1568, f1568_it0240

md5 checksum =
cc1fae67e8a61f6fcb4b38cf8f72af5f

- Archivematica creates and validates checksums throughout its workflow
- It can accept and validate Bags at the front end, and stores AIPs as [Bags]
  - Bag-making programs create checksums for files in a particular package in a predictable way
  - They can be validated over time

Corrupted with [Image Glitch Tool]

Integrity: Validation

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

- 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

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

Integrity: Validation

Test: is a file well-formed and valid?

A **well-formed file** obeys the syntactic rules of its file format. That is, it follows the structural rules as set out by its file format standard.

A **valid file** is first well-formed. Secondly, it meets higher-level semantically defined rules. That is, it meets certain quality standards defined for that file format, such as minimum bit depth, for example.

- Some tools:
  - JHOVE - images, documents - used in Archivematica
  - MediaConch - video - Archivematica can be used to validate derivatives

Availability

Ensures that objects are accessible into the future by periodically migrating copies to new formats and concerns other access-related issues in general

The best test of a preservation program is that content is accessible to users

Availability: Normalization

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

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

Various tools support normalization:

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

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

Reliability

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

- This is part of the “trustworthy” claim an archives can make - but only part
- Archivematica can help you with establishing authenticity, availability and reliability, but it does not by itself enable trustworthiness
- Trust is also about your relationship with your user community, your ability to recover from a disaster, and much more

The Workflow

Pre-Transfer*	Transfer	Backlog	Appraisal	Ingest	Storage & Access*
Selection of objects to preserve Metadata preparation Packaging for transfer	Generates METS file to be written to Virus scan File ID, characterization, validation	You can send something here if you don't want to continue processing it	File format view/analysis Selection for retention ID sensitive data	Normalize files Create & store AIP/DIP	Store in location Send access copies to other systems

*Not in Archivematica

*Linked to by Archivematica

Preparing transfers

Steps

- Determining content and structure (1 SIP = 1 AIP = fonds, series, item? Or section of one of these?)
- Gather and structure metadata (next slide)
- Gather submission documentation (not in demo)
- Package and structure for ingest
    - All data needs to be in a directory, at minimum

Metadata

Descriptive metadata

- Uses simple Dublin Core as key standard, other information is recorded as ‘Custom’
- Transfer level can be added through interface or imported
- Item level must be imported via CSV file

Rights metadata

- Mapped to PREMIS
- Same import structure as above

Demo

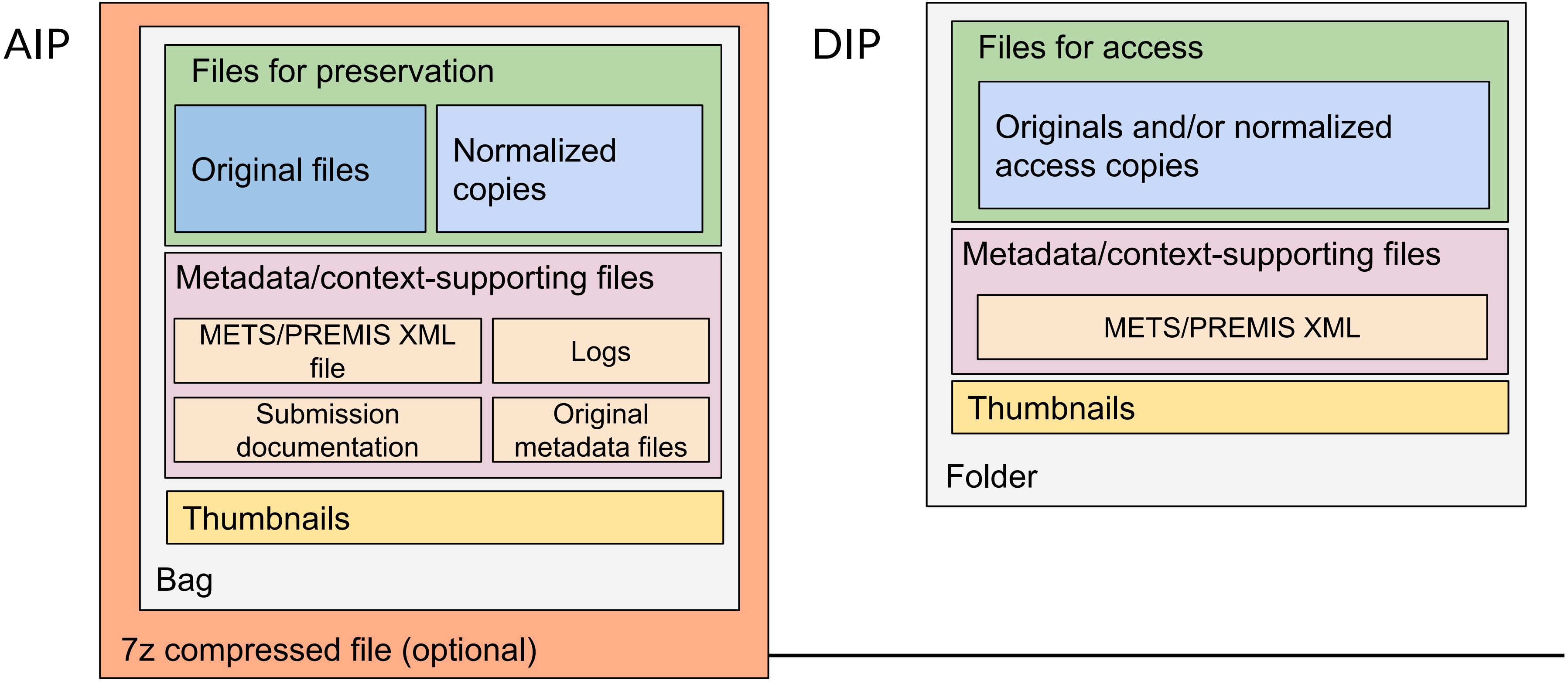
- Photos, PDF, WordPerfect files + metadata csv file
- Bagging using library of Congress Python tool

Processing transfers

Demo

- Same materials as before
- Uploaded to transfer source on Ontario Library Research Cloud
- Process using standard workflow and settings
- Briefly demo backlog/appraisal tabs
- Store AIP and DIP on OLRC

Outputs



AIP

Files for preservation

Original files

Normalized copies

Metadata/context-supporting files

METS/PREMIS XML file

Logs

Submission documentation

Original metadata files

Thumbnails

Bag

7z compressed file (optional)

DIP

Files for access

Originals and/or normalized access copies

Metadata/context-supporting files

METS/PREMIS XML

Thumbnails

Folder

Format Policy Registry (FPR)

- Accessed under “Preservation planning” tab
- Consists of a **format index** and **tools** paired with **rules** and **commands**
- If a file format is unidentified or there is no tool/rule/command, an action will fail

  - Format index: list of known formats and versions with yes/no if suitable for preservation and/or access
  - Tools: open source tools that perform certain functions
    - e.g. the tool ffmpeg normalizes audio and video
  - Rules: pair a format with a command to perform a policy-based action
    - e.g. for an AVI file, normalize to MKV
  - Commands: pair a tool with an output to fulfill a rule
    - e.g. normalize to MKV with ffmpeg

Thoughts about Archivematica

Pros:

- Connects functions/tools for preservation processing in a workflow you can start using right away
  - Does not require much setup/development
  - Actually implements METS/PREMIS
  - Reasonably scalable, if you have the computing resources and know your content
- Free, open source
  - Though not something you can really run on your personal computer
- Active and supportive user community

Thoughts about Archivematica

Cons:

- Overly prescriptive/conservative about normalization
    - Preservation normalization not as necessary as initially thought
- Not granular enough when it comes to file characterization/validation/normalization
    - Not all files need this metadata
- Fairly steep learning curve
    - Sometimes gives the impression that it will take care of all digital preservation work for you - spoiler alert - it does not
- Software development model means some features are permanently in beta

That’s all for now!

Questions now or to grant@scholarsportal.info !