Electronic Records Modules
Electronic Records Committee
Congressional Papers Roundtable
Society of American Archivists

File Format Migration Pathways

Erin Wolfe
Dole Center, University of Kansas
edw@ku.edu

Date Published: December 15, 2016

Module#: ERCM008

Created 2016-07
CPR Electronic Records Committee

File Format Migration Pathways

Overview and Rationale

Overview

This module is designed to assist with the development of file format migration pathways: the process of transforming digital files from one format (source format) to another (target format) through the use of a migration tool, with the goal of ensuring sustainable access to the files.[1] It is intended for use by archival staff responsible for ensuring that digital files are accessible across environments and/or for long-term preservation.

Rationale

File formats are the standardized ways in which software encodes bits of electronic information for storage in a computer environment. These formats require specific software and/or hardware to interpret and render for human access. Technological changes and developments create the possibility that, over time, formats may be rendered inaccessible, resulting in a loss of intellectual content and the associated value(s) of the data objects. The process of format migration is a crucial step in managing electronic files, both in terms of (a) providing access to and (b) creating preservation copies of electronic files and digital objects. Generally, the target format will be a widely adopted open format, ensuring community support and usage and increasing the likelihood of interoperability and long-term support of the format.

What are the benefits of format migration?

The transformation of files to an open format helps to ensure access to the files over time. Using open formats reduces barriers that may hinder access to digital files, such as software obsolescence, software and/or operating system incompatibility, proprietary software challenges, or other technology issues.

The process also carries the benefits of reducing the number of formats held by the institution, simplifying the ongoing management of these files and decreasing the likelihood of losing content due to obsolescence.

[1] http://www2.archivists.org/glossary/terms/f/format-migration

What are the costs of format migration?

The inherent cost in any format migration process is that of institutional resources. Creating format policies, defining metadata requirements, selecting and testing software, and other related tasks will all require staff time to research, test, troubleshoot, and implement. Once the migration pathway is established, staff time will still be required to perform the migration and quality check the results, in addition to periodically reviewing the pathway to ensure that it is kept current with best practices.

In terms of impact on the files, one of the most common costs of format migration may be the loss of functionality of the original object. For example, macros in a Microsoft Access database may not transfer to a mysql format, or layers in a Photoshop .psd file will be lost on migration to a .tif file. In many cases, as determined by the institution, the original format may be retained as a preservation copy in addition to the migrated copy, mitigating some potential concerns around loss during migration. This approach also leaves open the potential for restored functionality of the original formats that may occur with future technological developments. However, the practice of keeping multiple copies of files will increase storage needs, likewise increasing the monetary costs associated with digital preservation.

Another potential cost is the loss of formatting of the original object. For example, fonts, tables, spacing, or other design elements may display differently in the target file format, or some elements may disappear altogether. Institutions should test samples before migrating, as well as performing quality control following migration.

When does format migration occur?

Format migration primarily occurs at one of three stages of a digital object's lifecycle:

- Migration on ingest (generally referred to as "normalization"). This is the process of transforming file formats as they are ingested into a repository or institution. In most cases, this is the preferred method of format migration for long-term preservation.[2]
- Migration on obsolescence. This refers to the process of migrating files only as their existing format becomes obsolete. An institution may determine that, for example, a widely adopted proprietary format (such as a Photoshop .psd file) is an acceptable format for short-term preservation, with the awareness that the format status will need to be monitored over time. While this approach may have some benefits (e.g., reducing the resources needed to processing complex file types, or retaining functionality of the original), it carries inherent risks around losing content, as well as increased complexity of records management and future migration.

[2] Although the practice will vary between institutions, this approach may best be done in conjunction with some level of appraisal to reduce investing unnecessary resources in migrating files that will ultimately not be retained.

- Migration on access. This refers to the process of migrating files only as they are requested for access. As with migration on obsolesce, this approach may have some benefits in the initial stages of file management, but does carry the associated risks referenced above.

What comprises a format migration pathway?

Components of a format migration pathway may include the following:

- A pre-defined list of formats supported by the institution (preservation and access) [3]
    - This should include notes on limitations of target format, such as potential loss of functionality or formatting.
- A format policy for various file types and formats,[4] including:
    - A crosswalk to map source format to target format
    - Notes on exceptions to migration, including rationale
- Defined software to perform the migration process (migration tools)
- Defined software to provide access to the target format (access tools)
- A defined process for capturing and recording metadata around the migration process, with notes on any that are required.

Procedural Assumptions

The processes described here are largely at a policy level. Any specific actions were developed and tested on Windows computers. While the underlying steps behind them are the same regardless of operating system, some software and/or steps may be different for other users.

This module assumes that:

- The institution has a list of acceptable formats for preservation and/or access
- The institution does not have a digital repository that will automate the process of format mapping and format migration, such as Archivematica or Preservica
- The files have been checked and are virus free. If they are not, any processing of these files should be done on a non-networked workstation.

---

[3] Selection of formats that will be supported by the institution is an important precursor to format migration. A discussion of format selection is outside the scope of this document, but there many resources available to help with this step. For one approach to selection criteria, see the UK's National Archives' *Selecting File Formats for Long-Term Preservation*: https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf or the Digital Preservation Coalition's *File Formats for Preservation*: http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation.pdf.
See also Appendix A of this module.

[4] See Appendices B and C.

Hardware and Software Requirements

There are a variety of software available to help with format migration, including open source and proprietary.

Xena (XML Electronic Normalization for Archives)

Xena is free and open source software that can detect many types file formats of digital objects and convert them into open formats for preservation.[5] It runs in Java on Windows, Mac, or Linux as a stand-alone program or as a part of the National Archives of Australia's Digital Preservation Software Platform. A detailed user manual is included in the software.

Xena works to preserve objects in three different ways:

1. Normalization – supported formats are converted to an open format, encoded as ASCII text, and wrapped (along with metadata) in XML. The resultant .xena file is plain text and can be rendered in the Xena viewer. This process is reversible.
2. Conversion – supported formats are converted to an open format and stored as individual data objects.
3. Binary normalization – non-supported formats are encoded as ASCII text and wrapped (along with metadata) in XML.

Workflow

Preliminary Step: Format Policy creation

1. Identify which formats are preferred for various file types.
    a. Access
    b. Preservation
2. Identify any formats that should not be migrated automatically and note the rationale for this decision.
3. Identify which migration tools are preferred for specific migration tasks.
4. Identify any metadata that will be migrated and confirm that the selected migration tools will handle these in an acceptable manner.

[5] Xeno: Software for Digital Preservation: http://xena.sourceforge.net/, accessed November 2016. This is a good software platform that can handle a number of common file types. It is worth noting, however, that the latest version was released in 2013, and it appears to no longer be in development. Testing will be required to see how well it works for any particular institution. See also Chris Prom's 2010 *Review of XENA Normalization Software* for another perspective of some of the limitations: http://e-records.chrisprom.com/review-of-xena-normalization-software/

Process for migration

Note: the following steps will utilize the sample e-records workstation setup described in Appendix D.

1. Identify original file format(s) of files to be migrated.[6]
   - Ensure that a format policy for each of these formats is defined for your institution. Update as necessary.
2. Create virtual workspace for each group of files (e.g., all files in a specific accession).
   This should include an identified folder specific to that group of files, with separate sub-folders for preservation and working copies. These can be further divided into separate folders for access copies, metadata, logs, etc.
3. Copy original files and any supporting documentation to workstation/Preservation_Repository and to AIP/preservation_locked/originals folders. Confirm checksums. These are your backup copies. No actions should be done to any files in these folders.
   - If not already done, use this step to generate a full manifest of original files, including metadata and checksums.[7]
4. Copy any supporting documentation to the appropriate folder(s) in workstation/Descriptive_System and AIP/pres_desc_info
5. Make another copy of original files to workstation/Working_Space folder. Confirm checksums. These are your working copies. Any processing actions should only be done to files in this folder.
6. Do any pre-processing steps, as determined by institution's workflow (e.g., de-duplication, removal of system files, etc.). Record any actions taken.
7. Run migration processes for preservation formats on selected files and save to workstation/Working_Space/preservation_copies folder.
8. Run migration processes for access formats on selected files and save to workstation/Working_Space/access_copies folder.
9. Perform quality check of file conversions and required metadata.
10. Perform any additional processing actions to converted files (e.g., filename normalization, reorganization). Record any actions taken.
    - *Note*: *Depending on the institution's practice, this step may be done in conjunction with Step 5.*
11. Generate full manifest of migrated files (preservation and access), including metadata and checksums.
12. Move newly migrated access files to appropriate folders and verify checksums. These may include:
    a. Workstation/dissemination_copies/Lan_access_copies

[6] For one method of gathering all format information, see Module# ERCM001: *Generating File Format Identification and Checksums with DROID*: https://cprerc.wordpress.com/electronic-records-modules/

[7] For one method of generating full manifest of files with checksums, see Module# ERCM003: *Create AIP with DataAccessioner and Bagger*: https://cprerc.wordpress.com/electronic-records-modules/

b. Workstation/dissemintation_copies/Online_access_copies
c. AIP/preservation_locked/migrated_originals
d. AIP/processed_access_copies

Things to record as part of the file format migration process:

1. A manifest of original formats and file names, including checksums
2. Type of tools in the pathway, including software version information
3. A manifest of converted formats and file names, including checksums.

Appendix A: Sample file format table for preservation[8]

Media	High probability for full long-term preservation	Medium probability for full long-term preservation	Low probability for full long-term preservation
Text	• Plain text (encoding: USASCII, UTF-8, UTF-16 with BOM) • XML (includes XSD/XSL/XHTML, etc.; with included or accessible schema) • PDF/A-1 (ISO 19005-1) (*.pdf)	• Cascading Style Sheets (*.css) • DTD (*.dtd) • Plain text (ISO 8859-1 encoding • PDF (*.pdf) (embedded fonts) • Rich Text Format 1.x (*.rtf) • HTML (include a DOCTYPE declaration) • SGML (*.sgml) • Open Office (*.sxw/*.odt) • OOXML (ISO/IEC DIS 29500) (*.docx)	• PDF (*.pdf) (encrypted) • Microsoft Word (*.doc) • WordPerfect (*.wpd) • DVI (*.dvi) • All other text formats not listed here
Raster Image	• TIFF (uncompressed) • JPEG2000 (lossless) (*.jp2) • PNG (*.png)	• BMP (*.bmp) • JPEG/JFIF (*.jpg) • JPEG2000 (lossy) (*.jp2) • TIFF (compressed) • GIF (*.gif) • Digital Negative DNG (*.dng)	• MrSID (*.sid) • TIFF (in Planar format) • FlashPix (*.fpx) • PhotoShop (*.psd) • RAW • JPEG 2000 Part 2 (*.jpf, *.jpx) • All other raster image formats not listed here
Vector Graphics	• SVG (no Java script binding) (*.svg)	• Computer Graphic Metafile (CGM, WebCGM) (*.cgm)	• Encapsulated Postscript (EPS) • Macromedia Flash (*.swf) • All other vector image formats not listed here

[8] Excerpted from Cornell University's *Research Data Management Services Group* documentation under a CC BY 4.0 license. To see the full list, visit http://data.research.cornell.edu/content/file-formats. Accessed November 2016.
For a different approach, see NARA's Tables of File Formats: https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

Audio	• AIFF (PCM) (*.aif, *.aiff) • WAV (PCM) (*.wav)	• SUN Audio (uncompressed) (*.au) • Standard MIDI (*.mid, *.midi) • Ogg Vorbis (*.ogg) • Free Lossless Audio Codec (*.flac) • Advance Audio Coding (*.mp4, *.m4a, *.aac) • MP3 (MPEG-1/2, Layer 3) (*.mp3)	• AIFC (compressed) (*.aifc) • NeXT SND (*.snd) • RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) • Windows Media Audio (*.wma) • Protected AAC (*.m4p) • WAV (compressed) (*.wav) • All other audio formats not listed here

Appendix B: Sample format policy table 1[9]

Media type	File formats	Preservation format(s)	Access format(s)	Normalization tool
Audio	AC3, AIFF, MP3, WAV, WMA	WAVE (LPCM)	MP3	FFmpeg
Email	PST	MBOX	MBOX	readpst
Email	Maildir**	Original format	MBOX	md2mb.py
Office Open XML	DOCX, PPTX, XLSX	Original format	Original format	Tool search in progress
Plain text	TXT	Original format	Original format	None
Portable Document Format	PDF	PDF/A	Original format	Ghostscript
Presentation files	PPT	Original format	PDF	Tool search in progress
Raster images	BMP, GIF, JPG, JP2*, PCT, PNG*, PSD, TIFF, TGA	Uncompressed TIFF	JPEG	ImageMagick
Raw camera files/Digital Negative format**	3FR, ARW, CR2, CRW, DCR, DNG, ERF, KDC, MRW, NEF, ORF, PEF, RAF, RAW, X3F	Original format	JPEG	ImageMagick/UFRaw
Spreadsheets	XLS	Original format	Original format	None
Vector images	AI, EPS, SVG	SVG	PDF	Inkscape
Video	AVI, FLV, MOV, MPEG-1, MPEG-2, MPEG-4, SWF, WMV	FFV1/LPCM in MKV	MP4	FFmpeg
Word processing files	DOC, WPD, RTF	Original format	Original format	Tool search in progress***

[9] From Archivematica's documentation under a CC BY SA 3.0 license. https://wiki.archivematica.org/index.php?title=Media_type_preservation_plans. Accessed November 2016.

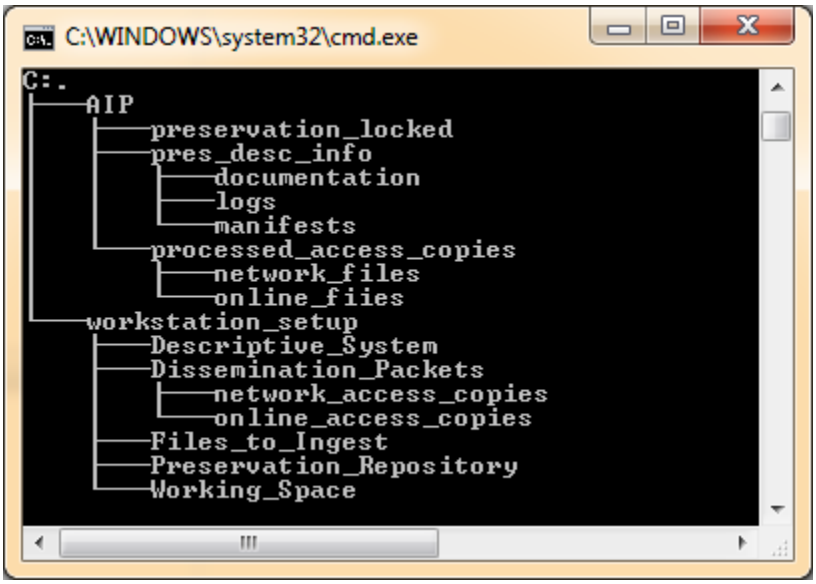
Appendix C: Sample format policy table 2[10]

Genre	Supported Formats	Preservation formats	Supported Access Format	Supported Access Tool	Migration Path/Notes
Text, source code	. txt or files of any extension or no extension containing ASCII or MIME data	Original format	.txt	NotePad++ or jEdit	To facilitate rendering use Renamer or Thunar to append .txt, without removing original extension.
Word Processed Documents	.doc, .wpd, .odt,	.odt	PDF/A	Adobe Reader, Open Office	Use Adobe Acrobat or Open Office to Migrate doc, wpd and odt to PDF/A, where appropriate.[1]
Raster Images	.jpg, .png	.tif	.jpg	GIMP	Use ImageMagick or Adobe to migrate.
Vector Images	.ai	.sgv	.sgv	Inkscape, any modern browser will open .sgv files, but inkscape imports .ai files .	Convert to sgv by importing to Inkscape.
Audio	.mp3	pcm stored as wav file	.mp3	VLC Media Player	Ffmpeg or Audacity
Video	avi, mov, wmv	mpeg2	.ogv	VLC Media Player	Ffmpeg

1. Some significant properties may be lost on migration, particularly if Open Office Writer is used for migration.

[10] Excerpted from *Practical E-Records* blog under a CC BY NC ND 3.0 license. For the full table and additional reading, visit: http://e-records.chrisprom.com/recommendations/supported-formats/migration-pathways-table/. Accessed November 2016.

Appendix D: Sample virtual workstation setup[11]



1. [File group identifier] – top-level folder for a single group of files that are being processed
   1.1. AIP – preservation folder for final, processed files
      1.1.1. Preservation_locked – final folder for all preservation copies, including original and migrated preservation format. Folder could be sub-divided into original and migrated folders.
      1.1.2. Pres_desc_info – top-level folder for preservation descriptive information
         1.1.2.1. Documentation – notes about the files, processing, migration, etc.
         1.1.2.2. Logs – any logs generated during automated processes
         1.1.2.3. Manifests – manifests of (a) originals and (b) processed copies
      1.1.3. Processed_access_copies – final folder for preservation copies of access formats
         1.1.3.1. Network files – files for internal use/storage
         1.1.3.2. Online files – files that will be made available online
   1.2. Workstation_setup – working folder, to be used until all actions are completed
      1.2.1. Descriptive_System – metadata and notes about metadata
      1.2.2. Dissemination_packets – top-level folder for access copies (working)
         1.2.2.1. Network_access_copies – files for internal access
         1.2.2.2. Online_access_copies – files that will be made available online
      1.2.3. Files to ingest – temporary home for files that are ready to be ingested into the digital repository
      1.2.4. Preservation_repository – initial home for preservation copies

[11] Based on a model presented by Chris Prom in a 2014 SAA workshop, *Arrangement and Description of Electronic Records.*

1.2.5. Working_space – temporary home for copies of files that are being processed