The University of British Columbia Library		Document No.	###
		Approval Date	*May 6, 2016*
		Last Revision	*May 5, 2016*
Title	Digital Preservation File Format Policy		

Since UBC uses Archivematica for digital preservation, normalization of file formats in digital collections is carried out according to the Format Policy Registry (FPR), Artefactual's database defining rules for the treatment of file formats during processing in Archivematica. The FPR rules determine the access and preservation formats for normalization in Archivematica. The FPR is configurable for local needs. Current and future file formats included in the FPR will be assessed according to factors listed below in the section titled "Considerations for File Formats".

Preservation File Formats

During ingest, digital preservation workflow users utilizing Archivematica can make choices about whether to normalize for preservation and/or access or not to normalize.

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***

* PNG and JPEG2000 are not normalized to a preservation format

** in development

*** In early versions of Archivematica, normalization of word processing formats (Microsoft Word, Word Perfect, etc) were normalized to PDF or open office formats using Libre Office. However, testing showed that the results were too inconsistent with significant losses in formatting information to continue using this normalization path. Currently, the FPR does not have any normalization paths for word processing formats [https://wiki.archivematica.org/Format_policies#Format_Policy_Registry_.28FPR.29]

Access File Formats

During the Archivematica ingest process, a DIP with metadata for the access system is created, as well access copies. Depending on the workflow, creation of access copies may not be necessary or projects may require manual creation of derivative access copies.

File name	File format	Preservation normalization attempted	Preservation normalization failed	Already in preservation format	Access normalization attempted	Access normalization failed	Already in access format
edited_tif/OSC_ARC_01_002_016_003.tif	TIFF	Yes	No	No	Yes	No	No
edited_tif/OSC_ARC_01_002_016_005.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_016_007.tif	TIFF	Yes	No	No	Yes	No	No
edited_tif/OSC_ARC_01_002_016_002.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_017_009.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_016_023.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_017_036.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_017_007.tif	TIFF	Yes	No	No	Yes	No	No
unedited_tif/OSC_ARC_01_002_017_011.tif	TIFF	Yes	No	No	Yes	No	No
edited_tif/OSC_ARC_01_002_017_029.tif	TIFF	Yes	No	No	Yes	No	No

Figure 1: Example of a normalization report for a package normalized for access but not for preservation.

Considerations for File Formats

Assessment of and decisions about file formats for preservation copies consider the following six factors. These factors were derived from the Digital Preservation Coalition's (DPC) *Digital Preservation Handbook* (2nd ed.), *Sustainability of Digital Formats: Planning for Library of Congress Collections*, and InterPARES report *Selecting Digital File Formats for Long-Term Preservation*. [1]

1. General Adoption

If formats are widely adopted and used, they are likely to be supported for a longer period of time, as well as have options for migration and emulation in the future.

2. Proprietary vs. Open Source

Potential problems exist for both proprietary and open source options. While proprietary formats are more subject to upgrades (forcing users to move to the latest version) and to vulnerability if the owner goes out of business, open source formats are dependent on the support of the development community. Proprietary formats might also be subject to license or royalty fees. Even if fees are reasonable or nonexistent now, there is no guarantee that owners will not levy fees in the future.

3. Availability of Documentation

It is very important that specifications, standards or other documentation about the format are available. DPC suggests that the format should be listed in the PRONOM file format registry. Generally, documentation is more readily available for non-proprietary formats.

4. Interoperability

File formats have a better chance of being accessible in the future if they are not dependent on a specific hardware, operating system or software.

5. Compression

Lossy compression (file formats using a compression algorithm which looses data) is inappropriate for long-term preservation. Preservation copies of files should be stored using file formats that involve lossless compression or that are uncompressed.

6. Embedded Metadata

File formats that store metadata in the digital object support future understandability and provide easier management than when metadata is stored separately.

Sources

Digital Preservation Coalition (2015-2016). "File formats and standards." In *Digital Preservation Handbook* (2nd edition). Retrieved from http://www.dpconline.org/advice/preservationhandbook/technical-solutions-and-tools/file-formats-and-standards

Library of Congress. (2013) "Sustainability Factors." In *Sustainability of Digital Formats: Planning for Library of Congress Collections.* Retrieved from http://www.digitalpreservation.gov/formats/sustain/sustain.shtml

McLellan, Evelyn Peters, InterPARES 2 Project. (2007). *General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation.* Retrieved from http://www.interpares.org/display_file.cfm?doc= ip2_file_formats(complete).pdf

---

[1] In general, these three documents agree on the same set of considerations. Factors were included where they were listed in at least two of the sources. The Library of Congress Document also adds two further considerations: transparency (extent to which it is possible to analyze the file using basic tools) and technological protection measures.