National Heritage Digitization Strategy – Digital Preservation File Format Recommendations

Ern Bieman, Canadian Heritage Information Network
William Vinh-Doyle, Provincial Archives of New Brunswick

List of abbreviations

AVI	Audio Video Interleave
BWAV	Broadcast Wave Format
DPX	Digital Picture Exchange
DV-NTSC	digital video – National Television Standards Committee
IEC	International Electrotechnical Commission
ITU	International Telecommunication Union
LAC	Library and Archives Canada
LC	Library of Congress
LPCM	linear pulse-code modulation
MBps	megabyte per second
MRC	mixed raster content
MXF	Material Exchange Format
NAA	National Archives of Australia
NARA	National Archives and Records Administration
NHDS	National Heritage Digitization Strategy
NPTA	National, Provincial and Territorial Archivists Conference
OCR	optical character recognition
PCM	pulse-code modulation
PDF/A	Portable Document Format/Archive
RGB	red-green-blue
RIFF	Resource Interchange File Format
SMPTE	Society of Motion Picture and Television Engineers
TPM	technological protection measure
XMP	Extensible Metadata Platform

Abstract

This document was produced by the Canadian Heritage Information Network (CHIN) in collaboration with members of the Digitization and Digital Preservation Discussion Group. It was submitted to the National Heritage Digitization Strategy (NHDS) steering committee, which had listed it as an activity in their 2018–2019 Business Plan. The document is meant to help Canadian cultural heritage institutions (art galleries, libraries, archives and museums) in selecting file formats for long-term preservation of their digitized content.

Scope

This document recommends file formats to be used in the process of preserving digitized content. Considerations are limited to formats already recommended by authoritative sources (including Library and Archives Canada [LAC], Library of Congress [LC], National Archives and Record Administration [NARA] and Harvard University). Since a rigorous evaluation of these formats for use in digital preservation has already been carried out by a number of organizations, that process is not repeated here. Instead, this document summarizes commonly considered evaluation criteria and limits the evaluation of each file format to a brief discussion of weaknesses and strengths, based on these criteria.

In keeping with the scope of the NHDS, this document is limited to formats relating to digitized content. It does not consider born-digital materials. Likewise, this document does not discuss other matters related to digital preservation, such as best practices or workflows, nor does it address any component of the digitization process.

For additional information on guidelines and best practices for digitizing documents, please consult the following online resources:

- Digitization Standards for the CMCC: Scan and Artifact Photography
- *Recueil de règles de numérisation* (in French only)

For additional information on guidelines and best practices for digitizing audio, video and motion picture films, please consult the following online resource:

- Recommendations on Preservation Files for Use in the Digitization of Analog Audio and Video Recordings and Motion Picture Films (PDF format), produced by the National, Provincial and Territorial Archivists Conference (NPTAC) Audiovisual Preservation Working Group in collaboration with the NHDS Steering Committee (henceforth referred to as the "NPTAC recommendations").

Introduction

Preservation file formats differ from access formats, web publication formats and, in some cases, preferred digitization formats. The criteria that make a format ideal for preservation have been identified by several organizations; some of these organizations have gone on to establish a rigorous evaluation and selection process, and many of them have made recommendations of their own (consult Evaluation criteria for examples).

This document is broken into three main sections:

1. a summary of evaluation criteria,

2. a summary of institutions adopting or recommending formats for preservation and
3. a discussion of formats recommended for each digitized media type.

Evaluation criteria

There are several existing documents that cite criteria for the selection of digital preservation file formats. A few examples include:

- The Digital Preservation Coalition's Digital Preservation Handbook compares open source and proprietary formats, as well as lossless and lossy formats. It also considers the availability of file specifications, its adoption rate, support for metadata, and the file format's ability to preserve the content properties that are deemed most significant.
- LC outlines digital preservation file format Sustainability Factors, which include disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents and technological protection measures (TPMs).
- The National Library of the Netherlands uses criteria for Evaluating File Formats for Long-term Preservation (PDF format), which include openness, adoption, complexity, TPMs, self-documentation, robustness and external dependencies.
- LAC's *Local Digital Format Registry – File Format Guidelines for Preservation and Long-term Access* (PDF format) (now archived) identifies criteria that include openness or transparency, adoption as a preservation standard, stability and compatibility, and standardization.

Some criteria are common among all of these institutions. Others are a topic of debate. Common to all institutions listed above are the following two fundamental criteria:

1. The format must be widely adopted. The adoption criterion is sometimes considered in the context of memory institutions using it for digital preservation, but the adoption of the format by the wider community is more important. This helps ensure that there is, and will continue to be, support for the format, that software and tools will be developed so as to use the format and work with it, and that the format will continue to be used in the long term.
2. The format must be well documented and open to inspection. In that regard, a number of criteria are often cited; each has a slightly different meaning, but all refer to this general goal. Openness, transparency and disclosure are all examples, and these can be assessed by the degree to which: documentation of the format is freely available, a file's detailed contents are easily inspected and software tools for inspecting and editing are readily available.

Further to these two general criteria are a number of others which are common to many selection criteria models. These include the following:

- The format should be subject to minimal external dependencies. For instance, formats relying on proprietary hardware, such as various digital audio tape formats from the 1970s and 1980s, or that are exclusively accessible on a single operating system or through specific software are

all reliant on the ongoing availability of that external factor. This is also true of text files that rely on external font definitions, as well as audio and video files that rely on external codecs.

- The format should, ideally, allow for the inclusion of internal metadata. Embedding documentation about content in the same file as the content itself helps to ensure that the two are never separated.
- There must be no legal barrier to using the format for the purposes of digital preservation. Selection criteria models often identify this risk by specifying the impact of a patent related to the encoding of content within the format or by indicating whether the format is open source or proprietary. Note, however, that some proprietary formats have become de facto standards, whose widespread use is encouraged by the format creators.

  In Canada, a legal barrier may also present itself in the form of a TPM; namely, any technology designed to prevent the copying of content. The *Copyright Modernization Act* of 2012 (PDF format) prevents anyone in Canada (including memory institutions) from circumnavigating TPMs in order to make copies of content for any purpose.

- A format's versions should be backward and forward compatible with their supporting environment, ensuring that content remains accessible regardless of the version. This is sometimes referred to as compatibility, and it is also sometimes included in the definition of robustness.

The following are selection criteria which may further improve a format's candidacy for use in preservation, but which are contentious if considered as mandatory requirements:

- A requirement that files be lossless upon edit or compression. This is unquestionably a desirable quality in any preservation format, but there may be exceptions where near lossless compression is acceptable. It may also be the case that a digitization output format is lossy but represents, by all other accounts, a better candidate than other formats produced by that hardware.
- A requirement that files be robust in the sense that the format is more likely to remain accessible in spite of (minimal) read or write failures. While this is useful for data recovery, such recovery is generally not required if good practices (such as keeping multiple copies and the use of fixity checking) are followed.
- A requirement that files have minimal complexity: this criterion focuses on the ability to manually decipher and read the contents of files, suggesting that human readable formats are preferable. However, the matter of accessing the file's contents has already been addressed (namely, that formats be well documented and open to inspection) so long as the format is clearly documented and the file is migrated to newer formats while the necessary software tools remain available.
- A requirement that a file be subject to a rigorous standardization process, ideally by an authoritative and widely recognized standards body. While this consideration would clearly improve a format's candidacy for use in preservation, it should not be considered a

requirement, as it overlooks the value of previously mentioned open and transparent proprietary formats which have become de facto standards.

Institutions adopting or recommending formats for digital preservation

Appendix A contains tables outlining these formats. The tables are based, in part, on a literature review previously conducted by the authors of this document and its contributors (consult the Digital preservation format literature review). The tables summarize file formats that are known to have been recommended or accepted as preservation formats by a subset of the previously reviewed institutions. An effort has been made to distinguish preservation formats from transfer formats (that is, formats that may be accepted by a memory institution, but which may not be the final format that is preserved). Only the preservation formats adopted by at least two of the institutions cited here have been considered for recommendation by NHDS.

Recommended file formats

Text – standardized typeface fonts

The following text formats are recommended for long-term preservation.

PDF/A (both versions 1 and 2)

Description: Portable Document Format (PDF) files allow the inclusion of text, graphics, formatting information and other features to control the layout and appearance of a document. Portable Document Format/Archive (PDF/A) has been developed specifically for archiving by preserving a document's "visual appearance over time, independent of the tools and systems used for creating, storing or rendering the files" (ISO 2005). As such, PDF/A prohibits features unsuitable for long-term archiving, including relying on external font definitions, using encryption, launching executable code from within the document and similar elements. Some scanning equipment is capable of scanning directly to PDF/A, and optical character recognition (OCR) will generally convert text components of a scanned object to machine-readable text in the resulting PDF document.

Considerations: Among the memory institutions reviewed, this is the most widely adopted preservation format for formatted text. Originally controlled by Adobe as proprietary formats, PDF formats were released to the public under royalty-free licence in 2008 and are published as open standards. PDF/A is published under ISO 19005. There are multiple versions of the PDF/A standard; versions 1 and 2 are both heavily adopted. Version 3 (PFD/A-3) allows the embedding of arbitrary file formats, which may not be suitable for archiving. It should also be noted that some embedded metadata in a PDF/A may need to be manually corrected (depending on the digitization process). In

such cases, edits may need to be done to the document firstly as a PDF, prior to the creation of the PDF/A version.

In addition to PDF/A, a number of image formats are also recommended for circumstances in which text in a scanned document cannot be converted through OCR to machine-searchable text, or where a detailed (and possibly lossless) image of the document is important. These formats, which are described in greater detail in the Still image section, include:

- TIFF
- JPEG 2000
- PNG

Still image

The following still image formats are recommended for long-term preservation.

TIFF

Description: The Tagged Image File Format (known as TIFF or TIF and having these terms as file name extensions) is a de facto standard widely adopted in the print and imaging industry for full colour, greyscale and bi-level images. Although the TIFF standard can accommodate image data in other forms, TIFF files are most commonly recognized as lossless compressed bitmap (raster) images. Header tags help identify, among other things, the image's size and definition as well as the image data structure and compression. The most recent version of the format is TIFF 6.0, released in 1992.

Considerations: A well-established and commonly known format for more than two decades, TIFF's ubiquity, lossless compression and image stability (while being edited) has made the format popular for both imaging hardware and software as well as image editing and rendering software, and archiving. As a result, the format is likely to continue to benefit from widespread support in the long-term.

Copyright for TIFF format is privately held by Adobe Systems. However, this copyright has never been exploited, and the full TIFF 6.0 Specification (PDF format) is freely available on the International Telecommunication Union (ITU) website.

Adobe does provide a process (using the Extensible Metadata Platform [XMP]) for embedding preservation metadata within an image's header. However, these are not native fields to the format.

The TIFF format has no external dependencies, and there are no concerns regarding TPMs.

JPEG 2000

Description: Joint Photographic Experts Group (JPEG) 2000 is an image compression and coding system developed by the Joint Photographic Experts Group in 2000 to replace the older JPEG standard. Among other improvements over traditional JPEG, this format supports transparency, optional lossless compression and variable levels of definition within the same image. It is also generally more robust, should file information be damaged or partially lost. File size tends to be smaller than with JPEG, and unlike JPEG, the newer format allows the embedding of metadata using XMP in the header information. Files associated with this format are recognized by the .jp2 and .jpx filename extensions.

Considerations: While the format has been taken up by larger memory storage institutions, the degree to which JPEG 2000 will be supported in the long term is not clear. As of 2018, support within imaging hardware and image viewing and editing software exists, but it is less common than with other popular formats.[1] There are no known cameras that export to JPEG 2000. The format is not backward compatible with JPEG, and while a free licence has been offered to coders to encourage development, concern remains over the multiple patents that cover it. All components of the JPEG 2000 specification are listed on the JPEG website. Some of these documents are available for purchase as ISO/IEC standards. Others are available free of charge from the ITU website.

If you are using JPEG 2000 as a preservation format, be sure that any compression is lossless (as both lossy and lossless compression are possible) or that the degree to which lossy compression is used is acceptable. Note that it is possible to set the degree of lossy compression in this format.

PNG

Description: Portable Network Graphics (PNG) is a patent-free ISO/IEC-recognized image format published in 2004 to replace the Graphics Interchange Format (GIF), which was covered by a patent until 2005. The raster-based format, which was designed with web publishing in mind, supports up to 48-bits per pixel of colour information (in red-green-blue [RGB] format) and an additional 16 bits, which can be used to render transparency. Lossless compression and interlaced rendering are supported. Unlike GIF, PNG does not formally support animation. PNG files are recognized by their .png filename extension.

Considerations: This format has been widely adopted in the web publishing industry and is recognized as a preservation format by some large memory institutions. The format is also widely supported by image editing and image management software. Care should be taken with regards to compression, as both lossless and lossy forms are supported, and with respect to embedded metadata, which the format does provide for, but which most software does not support. The format does not support TPMs and has no external dependencies.

PDF/A (versions 1 and 2)

Description: This description expands on the information provided for PDF/A (versions 1 and 2) under the formatted text section. All PDF formats incorporating images use a mixed raster content (MRC)

mechanism, meaning portions of a scan interpreted as an image are raster-based and the remaining portions are not. PDF image data is compressed in a number of ways, depending on image type (bi-tonal versus grayscale or colour) and settings. The bi-tonal compression is lossless, whereas greyscale and colour algorithms can either be lossy or lossless (Bärfuss 2014).

Considerations: While scanning equipment has traditionally scanned to TIFF or JPEG, some newer scanners are scanning to PDF or PDF/A. Lossless grayscale and colour image compression for PDF documents is comparable to that for TIFF. However, PDF/A has the advantage of also allowing for the embedding of colour profiles, optically recognized text and metadata.

Note: While other formats for still images adopted for preservation by memory institutions have been reviewed, the preceding represent the formats deemed most suitable for long-term preservation.

Audio

Note that digitized audio formats may refer to an encoding format (a bitstream of audio data, usually compressed, that may be stored or transmitted). The format of this bitstream is often referred to by the process (or algorithm) that is used to encode and decode it, which is known as a "codec." Strictly speaking, the codec (executable software) and the format it accesses or produces (a bitstream of audio/visual information) are two different things, but both are often referred to by the codec's name.

Audio formats may also refer to a file into which an encoded bitstream is stored. If the file into which the audio is stored holds nothing else (save, possibly, metadata about the audio), this file is often referred to as a "wrapper." If the file contains additional content (such as moving images, text or additional audio tracks), it is generally referred to as a "container." The distinction between containers and codecs is made in all of the recommendations below.

For additional information on the audio digitization process, please consult the NPTAC recommendations (PDF format), produced in collaboration with the NHDS Steering Committee.

The following audio format is recommended for long-term preservation.

BWAV container, with linear pulse-code modulation codec

Description: Waveform Audio (WAV) is a proprietary file format developed by Microsoft and IBM to store audio information. Broadcast Wave Format (BWAV) is identical to the WAV audio format, and the same player may be able to decode both file types. However, BWAV contains additional information in its header, most notably for the purpose of preservation: optional descriptive metadata.

The codec most commonly used within a BWAV container, and the one recommended for digital preservation, is the non-compressed linear pulse-code modulation (LPCM). For both WAV and BWAV formats, file size is limited to 2 GB, which is about three hours of play time at CD-level quality.

However, information in the BWAV header allows multiple files to be linked for longer playtimes. File extensions for BWAV are .bwf and .wav (note that the latter is also used by WAV).

Considerations: This is the only audio format currently recommended by NHDS, as it is the only one that meets the criteria of being adopted by many of the large memory institutions in our literature review. There may exist other formats that are suitable for preservation within the context of your institution. Specifications for WAV and BWAV (PDF format) are readily available online, and their use is licence-free. Some file editors may recognize BWAV metadata in different ways; thus, care should be taken to ensure it is not lost or corrupted upon edit or migration. The recommended sampling rate for high quality recordings is 96 kHz with a bit depth of 24 bits.

Video and motion picture

All recommended formats in the audio, video and motion picture sections are in keeping with the NPTAC recommendations (PDF format), produced in collaboration with the NHDS Steering Committee. This document should be consulted for further information on the digitization process for these forms of media.

Video

The formats in this section refer to content that was digitized from analogue video. As with audio, digitized video formats may refer to one of two things: the format of a bitstream that is either transmitted or stored (often referred to by the codec that can produce or read this bitstream), or the container file into which this bitstream (and often other forms of multimedia) is stored. A publicly maintained summary of common video containers and their supported video formats is found on Wikipedia.

The following video formats are recommended for long-term preservation.

MXF container, using JPEG 2000 lossless image format

Description: Developed by the Professional MPEG Forum in the late 1990s and early 2000s, and standardized through the Society of Motion Picture and Television Engineers (SMPTE), Material Exchange Format (MXF) is designed to address most components of the digital video lifecycle, including content capture, editing, distribution and archiving. Because of MXF's weight (large file size), which is due to its typically lossless content, it is not designed to be a consumer-playable format. Instead, content is kept in discrete tracks (including an unlimited number of audio tracks, generally stored in WAV-LPCM BWF format) for simple editing. This is the recommended format for preservation of associated audio. While MXF is capable of holding video in any format, it is commonly used to hold discrete image frames in lossless JPEG 2000 format (also recommended), further facilitating the editing process.

Considerations: This is one of two video preservation formats recommended by the NPTAC A/V Preservation Working Group, which recommends recordings at a 10-bit variable bitrate with a lossless minimum average of 50 MBps. Further details are found in the NPTAC recommendations (PDF format).

Incorporating a subset of the Advanced Authoring Format (AAF) standards developed by the networked (broadcast) media community, MXF is designed to help standardize the management of video across various hardware and software platforms. Taken up by the digital video and digital cinema production and management sectors, it is widely recognized by professional applications supporting this field on Windows, Apple and Linux operating systems. It has also been implemented in video equipment, since a large number of the manufacturers are members of the standards bodies from which MXF arose. All standards related to the container and the recommended video format are documented, and use of the standard is royalty-free. Patents exist for components of the JPEG 2000 standard, although rights have been waived for their application in this format.

MXF recognizes video archiving as part of the multimedia workflow it is intended to address, and the container format holds the appropriate metadata for this purpose. Two main components exist: system (or structural) metadata describes internal media formats, their relationships, etc.; user (or descriptive) metadata describes information generated at the production level (including content capture and editing), as well as that required for end-user and archival stages. Among the memory institutions reviewed, these container and media formats have been adopted for preservation by LAC, NARA and Harvard University.

MXF is suitable for the lossless preservation of video. However, accessible (playable) video formats will also be required.

QuickTime container with an uncompressed 4:2:2 chroma subsample

Description: QuickTime (QTFF), sometimes referred to by its file extension MOV, is a proprietary container format developed by Apple Computer, Inc. It is a de facto industry standard and is the basis for the ISO/IEC MPEG-4 standard, which is similar, but not identical. Many video recording devices export to the QuickTime container format, which is recognized by its ".mov" or ".qt" file extension. It is often the only format to which these devices will record. QuickTime can store video, multichannel audio and text (for subtitles) in separate tracks, each capable of being edited individually.

Considerations: This is the second of two video preservation formats recommended by the NPTAC A/V Preservation Working Group, which recommends a recording using 10-bit uncompressed v210 codec and approximately 36 MBps (consult the NPTAC recommendations [PDF format] for details). Because this recommended recording requires an uncompressed and, thus, lossless video format, it is too heavy to play in real time by typical hardware and software, but it is ideal for preserving all image information. However, it is recognized that resource limitations may not permit this in some heritage institutions. Lossless, uncompressed is ideal and recommended, but lossy or near lossless codecs such as JPEG 2000, digital video – National Television Standards Committee (DV-NTSC,

particularly for migration from digital tape) and Apple ProRes 422 are acceptable. All audio should be recorded in pulse-code modulation (PCM) format. The Quicktime format is widely supported on both Apple and Windows products. The format does include a feature for digital rights management (apparently for use with Apple's online iTunes Store), but this is not a concern for those preserving content which does not take advantage of that feature.

AVI container with JPEG 2000 or DV-NTSC codec

Description: The Audio Video Interleave (AVI) container format was first released by Microsoft in 1992 as a sub-format to the Resource Interchange File Format (RIFF). The container specification describes a file format that holds content in "chunks." The audiovisual content can be of any kind, and any codec (or combination of codecs) can be used to play both audio and video synchronously. These files are recognizable by their ".avi" filename extension.

Considerations: This format is not recommended but is deemed acceptable by the NPTAC A/V Preservation Working Group (which recommends more than one possible bit depth and sample rate; consult the NPTAC recommendations [PDF format] for details). Acceptable codecs include JPEG 2000 (ideally lossless) and DV-NTSC (acceptable for migrating from digital tape formats). Regardless of the video codec chosen, audio should be preserved in PCM format.

The AVI container is a proprietary format developed by Microsoft and IBM, although licensing appears not to be an issue and format documentation is available online. Being a native multimedia file format for the Windows operating system, AVI has a long history of widespread use and support. Due to the format's ubiquity, it is also supported in software on MAC, Linux and Unix operating systems. Memory institutions that have adopted this standard include LAC and NARA (which both prefer the format), as well as LC (which accepts the container with either the H.264 or H.262 video codecs).

There are no issues regarding TPMs. The file header contains technical metadata about the video and may also have metadata (as part of a RIFF info chunk) pertaining to provenance and copyright. AVI files may have further metadata using the XMP, but this is not standardized.

Digitized motion picture

This section refers specifically to the digital capture and preservation of motion picture film. Unlike digitized audio or video, digitized motion picture involves multiple files. It is computationally too complex to play in real time with consumer-grade hardware, and it is generally stored as a structured collection of files (not in containers). For additional information on the digitization process of motion pictures, please consult the NPTAC recommendations (PDF format), produced in collaboration with the NHDS Steering Committee.

DPX file format

Description: Developed in 1994 as a format for the Kodak Cineon film scanner, released in 2003 as an SMPTE standard (SMPTE 268M-2003) and revised through SMPTE on numerous occasions since then, the Digital Picture Exchange (DPX) file format is used to store high-quality detailed image information (in uncompressed raster format) and supporting metadata of a single motion picture frame. DPX is used as an output format by film scanning equipment, as a digital intermediate format for colour management during the production of a motion picture and as a digital format to print to film for distribution.

There are three commonly used blocks of metadata contained in the header of each file. The first type contains fields relating to the image itself, such as the file format identification "magic number," the image resolution, colour space information, the creation date/time, the creator's name, the project name, copyright information and other technical details. Some of these fields are core fields required by DPX "core compliant" software.

The second block of metadata is industry-specific. DPX also has roots in the telecine industry (film-to-TV broadcast equipment), and this section has metadata specific to broadcasting (typically with only the relevant grouping of these elements being filled out). Adoption of DPX in the TV broadcast sector has not been as successful as it is with motion picture capture.

The third type is optional user-defined metadata, the structure and length of which is not defined in the SMPTE standard. However, the Federal Agencies Digital Guidelines Initiative Audio-Visual Working Group has released Guidelines: Embedding Metadata in DPX Files, which takes advantage of this block to document, among other things, the digitization process history.

A fourth block, described as "image data," is not well-defined in version 2.0 of the SMPTE standard. It may be better defined in more recent revisions.

As per the NPTAC recommendations (PDF format), images should be stored in discrete uncompressed 4K resolution raster image files with a 10-bit RBG colour bit depth for 35mm film, and 2K resolution image files with a 10-bit RGB colour bit depth for 89mm, Super 8 and 16mm film. Audio should be stored as uncompressed PCM in a BWAV wrapper, with a 24-bit depth sample and a 48 kHz sampling rate. Because each image file contains only a single frame, metadata for the entire film is also often stored externally. The naming convention for each frame's file is also important, and it is commonly saved using a "name.n.dpx" format, where "n" is an 8-digit numeric value representing the frame's sequence within the film and ".dpx" is the file name extension. All DPX frame files for a given film are generally stored within a single directory.

Other information, such as the BWAV file for audio, is commonly saved separately. It is possible to scan an area wider than the original image frame and to embed audio, or other information, as visual content outside the image, but doing this is not part of the standard.

Considerations: This is a heavy format to save to disk. It takes upwards of 4 TB per hour of recording for larger format film. This format is an industry standard for film scanning, film colour management

and film printing. Among the memory institutions reviewed, it has been adopted as a preservation format by LAC and NARA.

Specifications of the format are fully disclosed as an SMPTE standard, and there are no known licensing or patent issues, nor are there any TPMs built into the standard. While access to the content is reliant on industry-specific software (and is not possible with consumer-level applications), a number of professional applications exist to support the format. The format also allows for ample self-documentation through its numerous metadata blocks and fields. Apart from the necessary software to access the content, there are no external dependencies. As such, DPX is a reasonable preservation format to which film can be scanned and archived. Access formats, however, will also be required.

For a more complete list of preservation formats adopted by large memory institutions, please consult the Digital preservation format literature review.

Acknowledgements

The authors would like to acknowledge the following people for their contributions to this document:

Paul Durand, Canadian Museum of History
Émilie Fortin, Bibliothèque de l'Université Laval

Appendix A: Summary of preservation formats adopted by reviewed memory institutions

The following tables summarize content from the Digital preservation format literature review, conducted by the authors of the present document as well as Paul Durand. Each table focuses on a specific type of digital media, or format discipline, and includes the format names, file name extensions for each format, if applicable, and their PRONOM Unique Identifier (a unique code assigned to that file format by the UK National Archives' PRONOM Registry) as well as a column for each cited institution, indicating whether that format was adopted or recommended for preservation by the institution.

All formats with recommendations or adoptions from two or more of the cited institutions are discussed in the present document, and most (with a few exceptions) are recommended by NHDS as a viable preservation format.

Text formats

Table 1: summary of formats adopted by cited institutions for formatted and unformatted textual data

Format Discipline	Extension	PRONOM Unique Identifer	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University
American Standard Code for Information Interchange (ASCII Text)	.txt .asc	x-fmt/111 x-fmt/22 x-fmt/283	yes	yes	yes	no	no
Unicode	.txt	x-fmt/111 x-fmt/22 x-fmt/283	yes	yes	yes	no	no
EPUB 3	.epub	fmt/483	yes	yes	no	yes	no
OpenDocument Text 1.2	.odt .ott	fmt/136 fmt/290 fmt/291	yes	no	yes	yes	no
PDF/A-1	.pdf	fmt/95 fmt/354	yes	yes	yes	yes	yes
PDF/A-2	.pdf	fmt/476 fmt/477 fmt/478	yes	no	yes	yes	yes
PDF/UA	.pdf	-	no	yes	no	no	no
Plain Text (encoding) (UTF-8, UTF-16)	n/a	x-fmt/111	no	yes	no	no	no
Word XML	.docx	fmt/412	no	no	no	no	yes

Format Discipline	Extension	PRONOM Unique Identifer	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University
WordPerfect (various versions)	.wpd	x-fmt/44 x-fmt/203 x-fmt/393 fmt/949 x-fmt/394 fmt/892	no	no	no	no	yes
Rich Text	.rtf	fmt/969 fmt/45 fmt/50 fmt/52 fmt/53 fmt/355	no	no	no	no	yes
Word (binary)	.doc	x-fmt/329 fmt/473 fmt/609	no	no	no	no	yes

Still image formats

Table 2: summary of formats adopted by cited institutions for still images

Format Discipline	Extension	PRONOM Unique Identifier	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University
TIFF (raster)	.tiff .tif	fmt/353 (many other versions)	yes	yes	yes	no	yes
JPEG 2000 (raster)	.jp2	xmt/392	yes	yes	yes	no	yes
JPEG (raster)	.jpeg .jpg	fmt/42 fmt/43 fmt/44	yes	yes	no	yes	yes
PNG (raster)	.png	fmt/11 fmt/12 fmt/13	yes	yes	yes	yes	no
GIF (raster)	.gif	fmt/3 fmt4	no	yes	no	no	yes
PDF (various versions, PDF/A preferred)	.pdf	-	no	yes	yes	yes	no
Scalable Vector Graphics (vector)	.svg	fmt/92 fmt/413	no	yes	no	yes	no

Format Discipline	Extension	PRONOM Unique Identifier	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University
Digital Negative (DNG) (raster)	.dng	fmt/436 fmt/152 fmt/437 fmt/438 fmt/730	no	yes	no	no	no
BMP (raster)	.bmp	fmt/114 fmt/115 fmt/116 fmt/117 fmt/118 fmt/119	no	yes	no	no	no
OpenDocument Graphics (raster)	.odg .otg	fmt/139 fmt/296 fmt/297	no	no	no	yes	no
Encapsulated PostScript (*.eps)	.eps	fmt/122 fmt/123 fmt/124	no	yes	no	no	no
JPEG File Interchange Format	.jfif	-	no	yes	no	yes	yes

Audio formats

Table 3: summary of formats adopted by cited institutions for audio files

Format Discipline	Extension	PRONOM Unique Identifier	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
Broadcast Wave Format	.wav	Version 1: fmt/2 Version 2: fmt/527	yes	yes	yes	no	no	yes
WAV	.wav .wave	fmt/6	no	yes	no	no	yes	no
MPEG-4	.mp4 .m4v .m4a .f4v .f4a	fmt/199	no	yes	no	no	yes	no
QTA_AAC, QuickTime Audio, AAC Codec	.aac	-	no	yes	no	no	no	no
AAC_ADIF, Advanced Audio Coding (MPEG-2), Audio Data Exchange Format	.m4p .m4b	-	no	yes	no	no	no	no
QTA_AAC, QuickTime Audio, AAC Codec	.m4p .m4b	-	no	yes	no	no	no	no

Format Discipline	Extension	PRONOM Unique Identifier	Library and Archives Canada	Library of Congress	National Archives and Record Administration	National Archives of Australia	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
WMA_WMA9_PRO, Windows Media Audio File with WMA9 Professional Codec	.wma	-	no	yes	no	no	no	no
Audio Interchange File Format (AIFF)	.aif .aiff	x-fmt/135	no	yes	no	no	no	no
MP3	.mp3	fmt/134	no	yes	no	no	no	no
Free Lossless Audio Codec (FLAC)	.flac	fmt/279	no	no	no	yes	no	no

Motion picture and video formats

The following table summarizes motion picture file formats, including digital cinema formats as well as end-user/consumer-level formats. Note that an additional column has been added to indicate what digital bitstream formats (the content that is encoded or decoded by a codec) are recommended for use within a given file format.

Table 4: summary of formats adopted by cited institutions for motion picture files

Discipline / Format	Extension	PRONOM Unique Identifier	Bitstream Format or Codec (if applicable)	Library and Archives Canada	Library of Congress	National Archives and Record Administration	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
Digital Cinema Distribution	File set (may be	N/A	N/A: Generally WAV and	no	yes	no	no	no

Discipline / Format	Extension	PRONOM Unique Identifier	Bitstream Format or Codec (if applicable)	Library and Archives Canada	Library of Congress	National Archives and Record Administration	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
Master (DCDM)	wrapped in MXF)		lossless frames					
Digital Picture Exchange (DPX)	.dpx	fmt/193 (1.0) fmt/541 (2.0)	N/A: Sequence of still raster images	no	yes	no	yes	no

Table 5: summary of formats adopted by cited institutions for video files

Discipline / Format	Extension	PRONOM Unique Identifier	Bitstream Format or Codec (if applicable)	Library and Archives Canada	Library of Congress	National Archives and Record Administration	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
Audio Video Interleave (AVI)	.avi	fmt/5	(LAC: uncompressed 4:2:2; LC: MPEG4 codec, MPEG-1 codec; NARA uncompressed 4:2:2.)	no	yes	yes	yes	no
MPEG-4	.mp4	fmt/596 fmt/199	(LC: MPEG4 codec; NARA: MPEG4 codec.)	no	no	yes	yes	no

Discipline / Format	Extension	PRONOM Unique Identifier	Bitstream Format or Codec (if applicable)	Library and Archives Canada	Library of Congress	National Archives and Record Administration	Harvard University	National, Provincial and Territorial Archivists Conference A/V Preservation Working Group
MOV	.mov .qtm	x-fmt/384	(Harvard: jpeg2000/MPEG-2; LAC: uncompressed 4:2:2; LC: MPEG4 codec; NARA: Uncompressed 4:2:2.)	no	yes	yes	yes	yes
MPEG-2	.mpg .mpeg .mp2	x-fmt/386	(LC: MPEG2 codec with AAC audio; NARA: MPEG2 codec.)	no	no	yes	yes	no
WMV	.wmv .asf	fmt/133	(NARA: VC-1.)	no	no	no	yes	no
MXF	.mxf	fmt/200 fmt/783 to 791	(LAC: JPG2000 lossless; Harvard: JPG2000; NARA: JPEG2000 lossless.)	no	yes	no	yes	yes

Appendix B: Glossary of technological terms

bit
: The most basic unit of information managed by a computer (set either to 1 or 0).

bit depth
: The number of bits assigned to one unit of information, such as colour information in a picture or the fidelity of a waveform in audio.

byte
: Eight bits.

codec
: An algorithm, expressed as software or firmware, for encoding and decoding audio or video. The format of the information stored by the codec is often referred to by the codec that is used to produce the file or access it.

compressed
: Digital information that is stored or transmitted in fewer bytes than what was produced or presented.

lossy
: An adjective describing the loss of some information through an action, such as when a file is compressed for storage or transmission.

lossless
: An adjective describing the complete retention of information in a file in spite of an action, such as when a file is compressed for storage or transmission.

pixel
: The smallest visible component of an image. In a black and white raster image, it is represented by a single bit.

raster image
: An image stored as pixels within a rectangular space (such as a Cartesian coordinate system).

uncompressed
: Digital information that was not compressed.

Websites consulted

Federal Agencies Digital Guidelines Initiative, Guidelines

National Heritage Digitization Strategy

The Technical Registry PRONOM

Wikipedia (for the following terms: Audio Video Interleave, Broadcast Wave Format, Digital Picture Exchange, JPEG 2000, Material Exchange Format, PDF/A, Portable Network Graphics, QuickTime, TIFF and Comparison of video container formats)

Bibliography

Advanced Media Workflow Association. *Structure of an MXF File*. N.p: Advanced Media Workflow Association, n.d.

Aldus Developers Desk. *TIFF Revision 6.0* (PDF format). Seattle, WA: Aldus Corporation, 1992.

Bärfuss, H. Scan to PDF/A: Some Insights. PDF Tools AG, 2014.

Digital Preservation Coalition. *Digital Preservation Handbook*, 2nd ed. Glasgow, Scotland: Digital Preservation Coalition, 2019.

Digitization and Digital Preservation Discussion Group (Canada). Digital preservation format literature review. Ottawa, ON: Canadian Heritage Information Network, 2019.

Duce, D., ed. *Portable Network Graphics (PNG) Specification*, 2nd ed. N.p.: World Wide Web Consortium, 2003.

European Broadcasting Union. *Specification of the Broadcast Wave Format; A Format for Audio Data Files, Supplement 6* (PDF format). Geneva, Switzerland: European Broadcasting Union, 2009.

Harvard Library. Data Management: File Formats and Naming, Formats. Cambridge, MA: Harvard Library, n.d.

International Organization for Standardization. ISO 19005-1:2005, *Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)*. Geneva, Switzerland: International Organization for Standardization, 2005.

Joint Photographic Experts Group. Workplan & Specs of JPEG 2000. N.p.: Joint Photographic Experts Group, n.d.

Library and Archives Canada. *File Format Guidelines for Preservation and Long-term Access, Version 1.0* (PDF format). Ottawa, ON: Library and Archives of Canada Local Digital Format Registry, n.d.

Library and Archives Canada. *Guidelines on File Formats for Transferring Information Resources of Enduring Value*. Ottawa, ON: Library and Archives Canada, 2015.

Library of Congress. Sustainability of Digital Formats: Planning for Library of Congress Collections. Washington, D.C.: Library of Congress, n.d.

Microsoft. AVI RIFF File Reference. N.p.: Microsoft, 2008.

National Archives and Records Administration. "Appendix A: Tables of File Formats." *Records Management Regulations, Policy, and Guidance*. College Park, MD: National Archives and Records Administration, 2017.

National Archives of Australia. Long-term File Formats. Canberra, Australia: National Archives of Australia, 2019.

National, Provincial and Territorial Archivists Conference Audiovisual Preservation Working Group and the National Heritage Digitization Strategy Steering Committee. *Recommendations on Preservation Files for Use in the Digitization of Analog Audio and Video Recordings and Motion Picture Films* (PDF format). N.p.: January 2018.

Rog, J., and C. van Wijk. *Evaluating File Formats for Long-term Preservation* (PDF format). The Hague, The Netherlands: National Library of the Netherlands, n.d.

Endnote

1 Consult https://en.wikipedia.org/wiki/JPEG_2000

Published by:
Canadian Heritage Information Network
Department of Canadian Heritage
1030 Innes Road
Ottawa ON K1B 4S7
Canada

Cat. No.: CH57-4/9-2019E-PDF
ISBN 978-0-660-33840-8

Également publié en français.