UNIVERSITÀ DEGLI STUDI DI PADOVA

SISTEMA BIBLIOTECARIO DI ATENEO

Recommended file formats for long-term archiving and for web dissemination in Phaidra

Edited by Gianluca Drago
May 2019

Premise

This document is intended to provide an overview of the file formats to be used depending on two possible destinations of the digital document:

- long-term archiving
- uploading to Phaidra and subsequent web dissemination

When the document uploaded to Phaidra is also the only saved file, the two destinations end up coinciding, but in general one will probably want to produce two different files, in two different formats, so as to meet the differences in requirements and use in the final destinations.

In the following tables, the recommendations for long-term archiving are distinct from those for dissemination in Phaidra.

There are no absolute criteria for choosing the file format. The choice is always dependent on different evaluations that the person who is carrying out the archiving will have to make on a case by case basis and will often result in a compromise between the best achievable quality and the limits imposed by the costs of production, processing and storage of files, as well as, for the preceding, by the opportunity of a conversion to a new format.

This choice is particularly significant from the perspective of long-term archiving, for which a quality that respects the authenticity and integrity of the original document and a format that guarantees long-term access to data are desirable.

This document should be seen more as an aid to the reasoned choice of the person carrying out the archiving than as a list of guidelines to be followed to the letter.

In the tables that follow, the "Recommended Web" column includes only those formats that allow direct viewing in the browser, without the aid of external applications. The column headed "Type of object in Phaidra" has the practical function of specifying which "type of object" must be selected by those who are archiving in Phaidra at the time of loading.

For an explanation of the abbreviations used here, see the "Abbreviations used" section at the end of the document.

Format selection criteria

There are some general criteria that can be followed when choosing the most suitable format for archiving. Although there are some discrepancies, archives and institutions involved in the preservation of digital heritage identify the most important as summarised below.

- **Openness**: an open format is defined as conforming to public specifications, i.e. available to anyone interested in using that format. The availability of the format's specifications should make it possible to decode it, even in the absence of products that perform this operation automatically. Some archives specify more strictly that the format must not be proprietary[1].
- **Portability**: means the ease with which formats can be used on different platforms, both hardware and software. When defining the portability of a format, account is also taken of the availability of tools that make it accessible both when creating files and when accessing data. The presence of external dependencies, technical protection mechanisms or patents are in contrast to portability.
- **Quality and functionality**: the ability of a format to provide features that ensure quality and richness of data, good performance in terms of speed (and possibly compression), inclusion of metadata and data of different nature.
- **Development support**: this refers to the resources needed to maintain and develop the format and the IT products that manage it.

1 Generally, "proprietary" format means a format that is not open, is covered by patents or licenses, or whose specifications are not fully available.

- **Diffusion**: measures the degree of diffusion of the format in the world, and in particular the level of adoption by the most important international archives. The diffusion of a format has implications on the probability that it will be supported over time, through the availability of computer products suitable for its management and visualization.
- **Transparency**: refers to the degree to which the digital representation is open to direct analysis with basic tools. Transparency is best if the content is encoded in standard encodings. Encryption is incompatible with transparency; compression inhibits transparency (however, for practical reasons, digital audio and video are rarely stored in an uncompressed form).
- **Self-documentation**: digital objects that are self-documenting are more sustainable and less vulnerable in the long run than objects that are stored separately from the metadata needed to make them usable. A digital object that contains basic descriptive metadata (the analog to the title page of a book) and incorporates technical and administrative metadata related to its creation and early stages of its life cycle will be easier to manage, to monitor for integrity and usability, and to transfer from one archiving system to the next.

For more information on format selection criteria, compare:

- DPCM 3 dicembre 2013, alleg. 2 *"Regole tecniche in materia di sistema di conservazione"*, ai sensi del *"Codice dell'amministrazione digitale"* https://www.agid.gov.it/sites/default/files/repository_files/leggi_decreti_direttive/dpcm_3-12-2013_conservazione.pdf pag. 17
- LOC, *Sustainability of Digital Formats: Planning for Library of Congress Collections,* 2017 *https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml#disclosure*
- Evelyn Peters McLellan, *Selecting Formats for Digital Preservation: Lessons Learned during the Archivematica Project*, 2010 https://groups.niso.org/apps/group_public/download.php/4237/IP_McLellan_Selecting_Formats_isqv22no2.pdf
- Universität Wien, *Formats for long-term preservation* https://datamanagement.univie.ac.at/en/about-phaidra/formats/formats-for-longterm-perservation/
- NARA, *Frequently asked questions about Digital Audio and Video*, 2016 https://www.archives.gov/records-mgmt/initiatives/dav-faq.html
- PACKED, *A short guide to choosing a digital format for video archiving masters*, 2014 https://www.scart.be/?q=en/content/short-guide-choosing-digital-format-video-archiving-masters

Text

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Object type in Phaidra	Notes
PDF/A		PDF/A		Document (PDF, TeX)	Widely used archiving format, which prohibits the use of some functions of PDF that may be difficult to render in the future. Well supported; fully available specifications
	PDF	PDF		Document (PDF, TeX)	Generally, lighter than PDF/A, but not suitable for long term archiving
	TeX, LaTeX		TeX, LaTeX	Document (PDF, TeX)	Markup format particularly used for the representation of mathematical formulas. Without its own accessory files (images, table of contents, bibliographies...) there can be loss of content, formatting, functionality. For Phaidra it is preferable to export to PDF/A
	HTML, XHTML		HTML, XHTML	Unknown	Without its own accessory files (images, CSS, JavaScript...) there can be loss of content, formatting, functionality. For Phaidra it is preferable to export to a compressed folder (e.g. ZIP) that also contains the accessory files. There is also the possibility of archiving in WARC format

					(WebARChive), which is however complex both in creation and in reading.
	XML		XML	Unknown	Includes XML-based markup formats including DTD/Schema and XSD/XSL stylesheets (TEI, DocBook...). Character encoding may vary (UTF-8, UTF-16, ASCII...) and must be explicitly stated
	EPUB		EPUB	Unknown	XML-based format. Not widely used, but well documented. Must not be encrypted or contain access restrictions
	TXT		TXT	Unknown	The simplest and most supported text format. Character encoding may vary (UTF-8, UTF-16, ASCII...). UTF-8 is recommended
	ODT (ODF)		ODT (ODF)	Unknown	Open format. Fully documented specifications. PDF/A export preferable
	ODT (ODP)		ODP (ODF)	Unknown	Open format. Fully documented specifications. PDF/A export preferable
	DOCX (OOXML)		DOCX (OOXML)	Unknown	Open format. Fully documented specifications. PDF/A export preferable. To exclude: macros, binary files, cross-references with external files
	PPTX (OOXML)		PPTX (OOXML)	Unknown	Open format. Fully documented specifications. PDF/A export preferable

Clarifications

For PDFs, Phaidra allows double uploading: downloadable PDF and "lightweight" PDF for quick viewing in the browser (see Guida all'archiviazione).

The PDF generated by the *Phaidra Importer* is not PDF/A.

Image

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Object type in Phaidra	Notes
	JPEG	JPEG		Image	Widely used, particularly for web usage (compressed files, with loss). Not suitable for long-term archiving of quality images
	JPEG 2000 (JP2)		JPEG 2000 (JP2)	Unknown	Standard covered by patents; only Part 1 specifications are fully available. There are several implementations that are not necessarily compatible with each other. Not yet widely used, but with increasing diffusion. Allows lossless compression. For long-term archiving must include descriptive and technical metadata
TIFF			TIFF	Image	Fully documented specifications. This is the *de facto* standard for archiving images. It must be uncompressed TIFF 6.0, with Intel byte order (PC), and inclusion of descriptive and technical metadata. Compared to lossy images, e.g. JPEG, these are heavy files
	PNG	PNG		Image	Lossless compressed format, fully open. Widely used in particular contexts
	PDF, PDF/A	PDF, PDF/A		Document (PDF, TeX)	Widely used archiving format. Well supported; specifications fully

					available. PDF/A prohibits the use of certain PDF functions that may be difficult to render in the future
SVG		SVG		Unknown	Based on XML. Open and fully documented format, suitable for archiving vector images
	DNG			Unknown	It is the open RAW (digital negative) format; it is fully documented
	RAW proprietari (NEF, CRW, 3FR...)			Unknown	The RAW formats (digital negatives) produced by cameras (CAM_RAW), are not suitable for long-term archiving, as they are proprietary (with the exception of DNG). However, provided that a copy in TIFF format is also archived, storing RAW files is useful since they contain raw data.

Clarifications

In general, for the distribution of raster images over the Web, it is preferable to use the compressed JPEG format; however, for PNG images that one wants to distribute without loss of quality use the native format.

The PDF format has been included here - in addition to the text formats - because it can be composed only of images, for example in the case of scanned books. Note that there is no unanimous view on whether PDF/A should be used for PDFs that contain only images, for which a conversion to PDF/A does not improve their long-term preservation (see, for example, that written by *The Open Preservation Foundation*)[2].

For PDF, Phaidra allows double uploading: downloadable PDF and lightweight PDF for quick viewing in the browse (see the Archiving guide).

2 https://openpreservation.org/blog/2014/08/27/when-not-migrate-pdf-pdfa/

The PDFs created by the *Phaidra Importer* are not PDF/A.
In image PDFs created by *Phaidra Importer* the images are partially modified (if nothing else, EXIF data is lost).

For a comparison of image formats, see also: FADGI, *Summary Table: Raster Still Images for Digitization: A Comparison of File Formats,* 2014, page 6 http://www.digitizationguidelines.gov/guidelines/FADGI_RasterFormatCompare_p3_20140417_r.pdf

Audio

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Object type in Phaidra	Notes
WAVE, Broadcast Wave (BWF)			WAVE, Broadcast Wave (BWF)	Audio	It's a *de facto* standard[3], well documented. For archiving, it should only contain uncompressed, Linear PCM bitstream (LPCM) encoded audio. WAVE files are relatively heavy compared to lossy formats. Broadcast Wave (BWF) files are WAVE files with metadata included, and are the preferred archiving format by IC, LOC, NAA, SIA, IASA, LAC and NARA (for NARA, along with FLAC)
	MP3	MP3		Audio	Format originally covered by patents (now expired), public specifications. Widely used, also thanks to the high compressibility. Lossy compression, therefore not suitable for long-term storage
FLAC			FLAC	Audio	Open format, developed as an open source project. It allows lossless compression, with production of files that are about a third the size of uncompressed files. When converting from BWF, it allows inclusion of the metadata present[4]. For NARA it is one

3 https://www.webarchive.org.uk/wayback/archive/20160101152346/http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/audio-wrappers
4 http://dericed.com/2013/flac-in-the-archives/

					of the two preferred formats (along with BWF). Does not use the Phaidra player
	MPEG-4 AAC		MPEG-4 AAC	Unknown	Widely used, also for the high compressibility. Lossy compression and covered by patents, therefore not suitable for long-term storage. Does not use the Phaidra player
	AIFF		AIFF	Unknown	For archiving, it should only contain uncompressed, Linear PCM bitstream (LPCM) encoded audio. Does not use the Phaidra player

Clarifications

Usually, for archiving purposes, uncompressed formats are preferred (although discussion is open, particularly for the FLAC format), and at the native resolution rather than through resampling.
See also: CAVPP, *Target Audio and Video Specifications*, 2017 https://calpreservation.org/wp-content/uploads/2017/03/CAVPP-File-Specs-2017.03.08.pdf

Video

A video format is a complex object, which can be seen as a container of multiple files of a different nature: audio, video, subtitles and others. These files can be compressed with different encodings, some more suitable for long-term preservation and others for web use. Therefore, here it is necessary to detail both the container formats and the encodings of the video and audio contents (in general for long preservation the audio will be uncompressed WAVE, or alternatively FLAC); consequently, the following two tables should be used together[5].

Differently from what occurs in audio digitization, where WAVE (normally in the BWAV variant) is widely considered the *de facto* standard for archiving, in the video field there is no consensus among archiving institutions.

For video archiving, historically there are two opposing trends: television broadcasting agencies, possessing enormous amounts of video footage in which content is often more important than image quality, tend to archive in compressed and lossy format, so as to speed up procedures and reduce storage costs[6]. Conversely, the institutions (archives, museums, libraries) that must preserve the cultural and video heritage of which they are the repositories, prefer to archive the material in the best possible quality, uncompressed or, more frequently, with lossless compression, that is, without loss of data[7].

For long-term archiving, large international archives suggest several containers - AVI, MOV, MXF, Matroska - and several encodings - JPEG2000, FFV1 or V210.

According to an analysis of FADGI's comparative study of major digitization projects[8], there are two different approaches, or "communities", that of large national archives and libraries more likely to use standards with a capital "S" (MXF or MOV with JPEG2000 encoding) and those of specialists, mainly located in Europe, active in the adoption of formats created in open source projects even when not widely established or not well documented[9] (Matroska with FFV1 encoding) [10].

5 For a correspondence between containers and supported encoding formats, see Wikipedia *"Comparison of video container formats"* https://en.wikipedia.org/wiki/Comparison_of_video_container_formats

6 The Digital Learning and Multimedia Office of the University of Padova, for example, recommends archiving in the original format, but, if the format is obsolete, accompanied by a copy in a widely used format, even compressed and lossy (e.g., MP4 or Matroska, with H.264 video encoding, and FLAC audio) (Marco Toffanin, voice communication).

7 http://download.das-werkstatt.com/pb/mthk/info/video/comparison_video_codecs_containers.html#lossy_vs_lossless

8 http://www.digitizationguidelines.gov/guidelines/video_reformatting_compare.html?loclr=blogsig

9 The limit of not being well documented present in FFV1 is destined to disappear shortly since the specifications for FFV1 are now at the status of "last call" thanks to the IETF cellar working group. (https://datatracker.ietf.org/doc/draft-ietf-cellar-ffv1/ February 2019)

10 https://blogs.loc.gov/thesignal/2014/12/comparing-formats-for-video-digitization/

For the video quality of files intended for long-term preservation, the resolution[11], bitrate[12] and all other parameters that best preserve the authenticity and integrity of the original video are also taken into consideration. In particular, in the case of conversion to a new format, the resolution and bitrate of the original must be maintained.

In general, archiving in the original format is recommended for preservation, but, if the format is obsolete, it is also recommended to archive a copy in a format currently widely used and suitable for the purpose (i.e. well documented, uncompressed or lossless compressed, etc.).

As for the display in Phaidra, the player currently used (Video.js) plays only the MP4 format with H.264 video encoding and MP3 or AAC audio encoding. In the future, Phaidra is expected to adopt more efficient streaming solutions, compatible with a wider range of formats and encodings.

Containers

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Object type in Phaidra	Notes
AVI			AVI	Video	Very popular format, well supported and with fully documented specifications. For archiving, IC recommends uncompressed AVI (encoding not specified). LAC also prefers this format (AVI with 4:2:2:2 chroma subsampling; uncompressed), along with MXF and MOV
	MP4 (MPEG-4 part 14)	MP4 (MPEG-4 part 14)		Video	Very popular and well documented container. Usually used with MPEG-H Part 2 (H.265/HEVC), MPEG-4 Part 10 (H.264/AVC) and MPEG-4 Part 2 video encoding; MPEG-4 AAC is the

11 In ascending order, from the worst to the best resolution: VHS, PAL, DVD, Blu-Ray (720p, 1080p), 4K, 8K (https://datamanagement.univie.ac.at/en/about-phaidra/formats/background-knowledge/).

12 In ascending order, from the worst to the best *bitrate*: 1 Mbps (480p), 2-5 Mbps(720p), 4.5 Mbps (1080p), 9.8 Mbps (DVD), 40 Mbps (HB Blu-Ray) (https://datamanagement.univie.ac.at/en/about-phaidra/formats/background-knowledge/).

					only allowed audio encoding. For access, IC and CAVPP recommend MPEG-4
Motion JPEG 2000 (MJP2 or MJ2)			Motion JPEG 2000 (MJP2 or MJ2)	Unknown	Diffuse format, with fully documented specifications. Intraframe compression. Heavy files. NAA uses Motion JPEG 2000 for its projects.
MXF			MXF	Unknown	Popular format more professional than desktop, with fully documented specifications. The LOC National Audio-Visual Conservation Center uses MXF (JPEG 2000 included in MXF) when reformatting videotapes for preservation. LAC also prefers this format (always in combination with lossless compressed JPEG2000), along with AVI and MOV (4:2:2 chroma subsampling; uncompressed).
Matroska (MKV)			Matroska (MKV)	Video	Developed as an open source project. Open, widely used and rapidly expanding format. Supports a large number of audio and video encodings. Can contain complex objects. Format recommended by PREFORMA. Archivematica and SNA use Matroska for preservation (with FFV1 video and LPCM audio encoding)
Quicktime (MOV)			Quicktime (MOV)	Video	Very common format. Fully available specifications. Can be considered as a variant of MP4 to which it has given

					rise and from which it incorporates many updates[3]. May contain complex objects. Used by CAVPP for masters. LAC also prefers this format (MOV with uncompressed 4:2:2 encoding), along with MXF and AVI
	OGG		OGG	Unknown	Open format, developed by the open source project Xiph. Limited diffusion. Can incorporate different audio and video encodings
DPX			DPX	Unknown	Open format. It is considered a standard for high quality digital conversion from cinematic films. Normally it does not include audio, which is saved separately (usually in WAVE). Used by Motion Picture, Broadcasting, and Recorded Sound Division (LOC), NARA and LAC

Video coding formats [13]

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Notes
	H.264 (also AVC) (MPEG-4 part 10)	H.264 (also AVC) (MPEG-4 part 10)		Compressed format, lossy or lossless[14], widely used especially for web use. Protected by various patents. For web access LOC and CAVPP use MPEG-4_AVC
	H.265 (also HEVC) (MPEG-H Part 2)	H.265 (also HEVC) (MPEG-H Part 2)		Successor format of H.264, more efficient, but extremely demanding in terms of CPU resources. Covered by numerous patents. Growing circulation
JPEG 2000 (JP2)			JPEG 2000 (JP2)	Standard covered by patents, lossy or lossless compression (intraframe). Demanding in terms of CPU resources. Supported by few software. Adopted by important audio-video archives. There are various implementations that are not necessarily compatible with each other. Used by NAA for its projects (in Motion JPEG 2000 container), by the National Audio-Visual Conservation Center of the LOC in the reformatting of *videotapes* for preservation (in MXF container) and by LAC
	Theora		Theora	Open format, developed by the open source project Xiph. Compression with loss. Limited diffusion
FFV1			FFV1	Open format, developed as an open source project. Created for digital preservation. Intraframe compression like Motion JPEG2000, but much more efficient. Good performance; well supported by software. Still little used, but of increasing popularity

13 Erroneously, sometimes we refer to encoding formats with the English term *codec* (which stands for "coder-decoder"). For a terminological distinction between "*encoding format*" and "*codec*" see Wikipedia:: https://en.wikipedia.org/wiki/Video_coding_format#Distinction_between_"format"_and_"codec"

14 In reality the encoding format is not optimized for lossless compression and most players are not able to read compressed H.264 wihtout loss (personal communication by Peter Bubestinger-Steindl).

				in archiving[15]. For storage, Archivematica uses FFV1 in Matroska container, OM uses FFV1 in AVI container. On FFV1 in Matroska, see also FIAF[16]
	Dirac		Dirac	Compressed encoding format, lossless or lossy, not covered by patents. Limited distribution. Demanding in terms of CPU resources. Developed by the BBC
V210, YUY2, UYVY e altri (tutti con 4:2:2 chroma subsampling)			*V210, YUY2, UYVY e altri (tutti con 4:2:2 chroma subsampling)*	Well documented and well supported formats, uncompressed, used by many archives. They result in large files. V210 is used by CAVPP for the masters. YUY2 is used by NARA for reformatting videotapes

Clarifications

Figure 1 graphically summarizes the diffusion of video containers and encodings recommended for long-term preservation by some important archives and projects in the world.

It can be noted that some archives leave a wide spectrum of choice, even including formats considered "at risk" (e.g. WMV), while others restrict the choice to practically a single format (e.g. PREFORMA with FFV1 encoding format included in MKV container, or LOC with JPEG2000 encoding format included in MXF container). NARA has also been included in the figure of the recommended formats, although these archives, for video formats, provide only "accepted" formats and no "preferred" ones.

Among the containers, besides the "classic" MOV and AVI, also MXF and Motion JPEG2000 are widely used, and also the open source MKV is beginning to have a significant basis of use for professional preservation.
Among the encoding formats, besides the uncompressed ones (4:2:2 and others) the use of JPEG 2000 and the open source format FFV1 stand out.
It should also be noted that some "traditional" containers - in particular AVI and MOV - are used in combination with a wide variety of encodings, while the more recent containers are used in more defined configurations (MXF with JPEG2000 and MKV with FFV1).

15 https://www.webarchive.org.uk/wayback/archive/20160101152356/http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/video-codecs
16 https://www.fiafnet.org/images/tinyUpload/E-Resources/Commission-And-PIP-Resources/TC_resources/FFV1_and_Matroska_reading_list.pdf

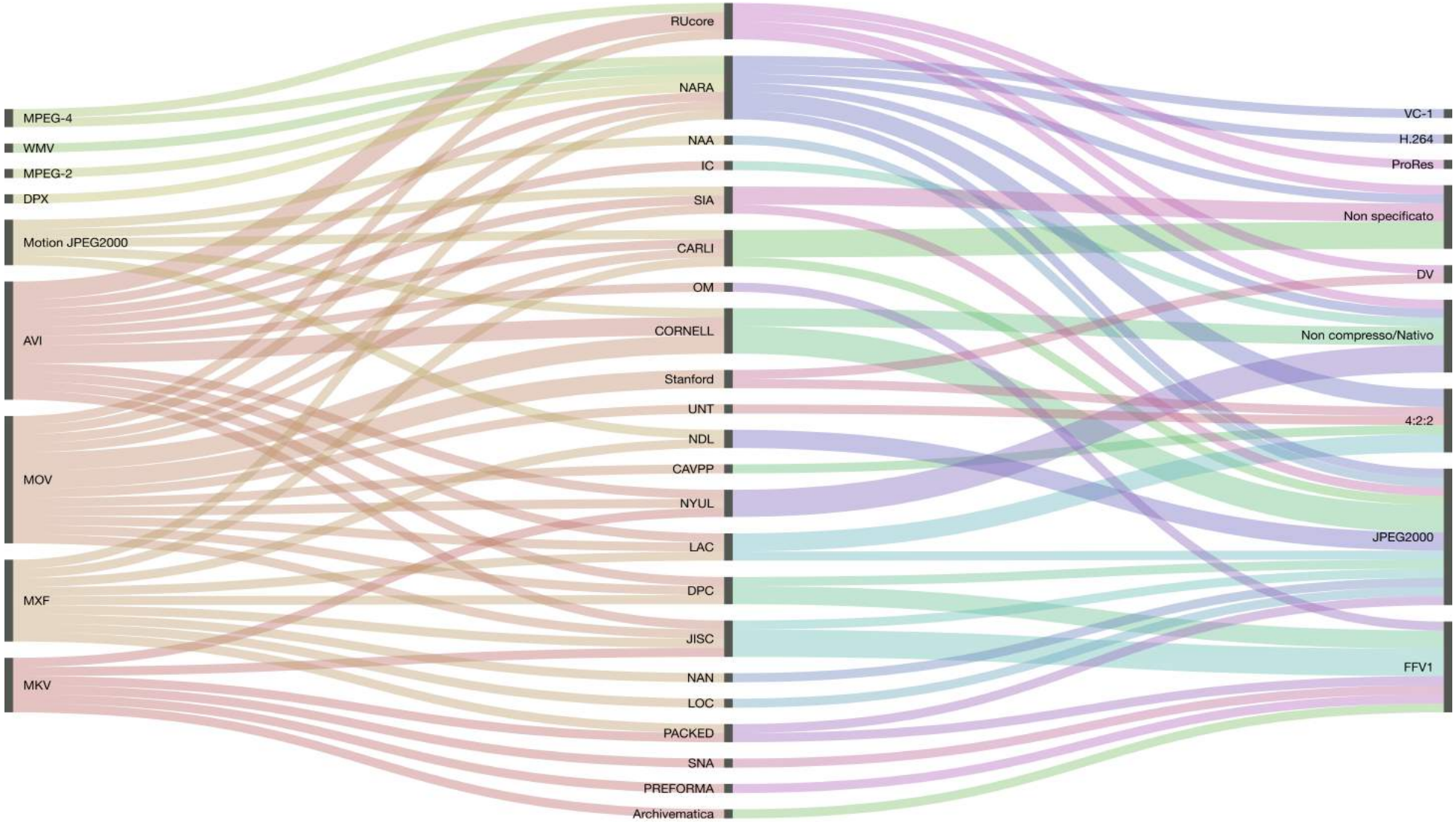

Figure 1: Diffusion of video containers (columned on the left) and codings (on the right) recommended for long preservation by some important archives and projects in the world (in the center)

A useful tool to determine the technical data of a video file (container, encoding, bitrate, frame rate, color space, bit depth and more) is MediaInfo, available for Windows, MacOS and Linux: https://mediaarea.net/en/MediaInfo. The MediaInfo developers themselves are working on a more in-depth analysis tool, MediaTrace (https://mediaarea.net/MediaTrace), available, for example, in MediaConch (https://mediaarea.net/MediaConch).

Due to the complexity of the topic, for more information on the formats of video files intended for preservation, please refer to the extensive documentation on the network:

- FADGI, *Digital File Formats for Videotape Reformatting – Part 5. Narrative and Summary Tables*, 2014. At pages 16-17 two tables, one for the containers and one for the coding formats http://www.digitizationguidelines.gov/guidelines/FADGI_VideoReFormatCompare_p5_20140908.pdf
- FADGI, *Digital File Formats for Videotape Reformatting*, 2014 http://www.digitizationguidelines.gov/guidelines/video_reformatting_compare.html
- FADGI, *Guidelines: MXF Application Specification*, 2017 http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html
- FADGI, *Creating and Archiving Born Digital Video*, 2014 http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p1_20141202.pdf http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p2_20141202.pdf http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p3_20141202.pdf http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p4_20141202.pdf
- NDL, *File formats*, page 13 and following http://digitalpreservation.fi/files/File-Formats-1.6.1-en.pdf
- Wikipedia, *Comparison of video container formats* https://en.wikipedia.org/wiki/Comparison_of_video_container_formats
- Harvard Library Digital Preservation Program, *Video Format Matrix*, 2016 https://wiki.harvard.edu/confluence/display/digitalpreservation/Video+Formats
- IASA, *Guidelines for the Preservation of Video Recordings IASA-TC 06,* 2018 https://www.iasa-web.org/sites/default/files/publications/IASA-TC_06-B_20180518.pdf https://www.iasa-web.org/sites/default/files/publications/IASA-TC_06-B-app_20180520.pdf
- CAVPP, *Target Audio and Video Specifications*, 2017 https://calpreservation.org/wp-content/uploads/2017/03/CAVPP-File-Specs-2017.03.08.pdf
- AMIA, *Digital Formats - Video Part One*, 2017 https://www.youtube.com/watch?v=BkYLEsjIdK4

- AMIA, *Digital Formats - Video Part Two*, 2017 https://www.youtube.com/watch?v=-TOJp-AL8Z4
- Stanford Media Preservation Lab, *Capture specs*, https://library.stanford.edu/research/digitization-services/labs/stanford-media-preservation-lab/capture-specs
- RUcore, *Video and Moving Image Objects: Recommended Minimum Standards For Archival and Presentation Datastreams*, 2015 http://odin.rutgers.edu/standards/2015/Video%20Object%20Standards%20Analysis-2015.pdf
- ALCTS, *Minimum Digitization Capture Recommendations*, 2013 http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations
- George Blood [for the LOC], *Refining Conversion Contract Specifications: Determining Suitable Digital Video Formats for Medium-term Storage*, 2011 http://www.digitizationguidelines.gov/audio-visual/documents/IntrmMastVidFormatRecs_20111001.pdf
- New York University Libraries, *Digitizing Video for Long-Term Preservation: An RFP Guide and Template*, 2013 http://memoriav.ch/wp-content/uploads/2014/07/VARRFP.pdf
- CARLI (Consortium of Academic and Research Libraries in Illinois), *Guidelines for the Creation of Digital Collections, Digitization Best Practices for Moving Images,* 2017 https://www.carli.illinois.edu/sites/files/digital_collections/documentation/guidelines_for_video.pdf
- FIAF, FFV1 and Matroska reading list https://www.fiafnet.org/images/tinyUpload/E-Resources/Commission-And-PIP-Resources/TC_resources/FFV1_and_Matroska_reading_list.pdf
- LAC, Guidelines on File Formats for Transferring Information Resources of Enduring Value, 2014 https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Pages/guidelines-file-formats-transferring-information-resources-enduring-value.aspx
- LAC shows a list of “container/coding format” pairs, preferred or acceptable, at page 14 of https://www.bac-lac.gc.ca/eng/services/government-information-resources/guidelines/Documents/file-formats-irev.pdf
- P. Bubestinger-Steindl, H. Lewetz, M. Jaks, *Comparing Video Codecs and Containers for Archives*, 2015 http://www.av-rd.com/knowhow/video/comparison_video_codecs_containers.html
- P. Bubestinger-Steindl, *Risk Assessment Considerations: Using FFV1 for Preservation*, 2016 http://www.av-rd.com/knowhow/video/risk_assessment.html

- eCommons: Cornell's Digital Repository, *Recommended File Formats*, 2018 https://guides.library.cornell.edu/ecommons/formats
- Smithsonian Institution Archives, *Recommended Preservation Formats for Electronic Records* https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records
- Reto Kromer, *Matroska and FFV1: One File Format for Film and Video Archiving?*, 2017 https://www.fiafnet.org/images/tinyUpload/Publications/Journal-Of-Film-Preservation/Matroska-and-FFV1_Kromer_JFP96.pdf
- National Archives of the Netherlands, *Preferred formats National Archives of the Netherlands: in view of sustainable accessibility*, 2016 https://www.nationaalarchief.nl/sites/default/files/field-file/National%20Archives%20of%20the%20Netherlands%20preferred%20and%20acceptable%20formats.pdf
- Reto Kromer, *Matroska and FFV1: One File Format for Film and Video Archiving?*, in Journal of Film Preservation, n. 96 (April 2017) https://retokromer.ch/publications/JFP_96.html

Structured data

Archiving Recommended	Archiving Possible	Web Recommended	Web Possible	Object type in Phaidra	Notes
CSV		CSV		Unknown	Open format. *De facto* standard. One of the formats preferred by LOC
TSV		TSV		Unknown	Open format. One of the formats preferred byLOC
	ODS (ODF)		ODS (ODF)	Unknown	Open format. Fully documented specifications. Prefer to export to PDF/A, or CSV. However, the following should be excluded: macros, binary files, cross-references with external files
	XLXS (OOXML)		XLXS (OOXML)	Unknown	Open format. Prefer to export to PDF/A or CSV. However, the following should be excluded: macros, binary files, cross-references with external files
JSON		JSON		Unknown	Open format. Typically an interchange format for data. One of the formats preferred by LOC
XML		XML		Unknown	Open format. In Phaidra, one can declare the encoding in the Metadata Editor (Technical Data → Requirements for using the object). E.g. UTF-8 and UTF-16 (with BOM), US-ASCII, ISO 8859...
	PDF/A	PDF/A		Document (PDF, TeX)	Open format. Widely used storage format, which prohibits the use of

					certain functions of the PDF that may be difficult to render in the future. Well supported; fully available specifications

Clarifications

Data files and databases must be transferred as flat files or as rectangular tables, i.e. as two-dimensional arrays, lists or tables.
Structured data must be transferred together with the associated files needed to verify the validity of the data, e.g. DTDs, schemas and data dictionaries.

Acknowledgements

Thanks to Marco Toffanin[17] and Antonio Zanonato[18] for their advice on video formats.
Thanks to Jérôme Martinez[19] and Peter Bubestinger-Steindl[20] for the revision of the sections on audio and video.

17 Digital learning e multimedia Office of the University of Padova
18 Dipartimento dei Beni Culturali of the University of Padova
19 https://archive.fosdem.org/2018/schedule/speaker/jerome_martinez/
20 http://www.preforma-project.eu/advisory-board.html

Abbreviations used

AGID = Agenzia per l'Italia digitale https://www.agid.gov.it/

ALCTS = Association for Library Collections & Technical Services http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations

AMIA = Association of Moving Image Archivists https://amianet.org/

Archivematica https://wiki.archivematica.org/Media_type_preservation_plans

CAVPP = California Audiovisual Preservation Project https://calpreservation.org/

CORNELL = eCommons: Cornell's Digital Repository https://guides.library.cornell.edu/ecommons

CPP = California Preservation Program https://calpreservation.org

DCC = Digital Curation Centre http://www.dcc.ac.uk/

DPC = Digital Preservation Coalition https://www.dpconline.org/

FADGI = Federal Agencies Digitization Guidelines Initiative http://www.digitizationguidelines.gov/

FIAF = International Federation of Film Archives https://www.fiafnet.org

FIAT = Fédération Internationale des Archives de Télévision / The International Federation of Television Archives (FIAT/IFTA) http://fiatifta.org/index.php/about/

IASA = International Association of Sound and Audiovisual Archives (IASA) https://www.iasa-web.org/

IC = Internet Culturale http://www.internetculturale.it/it/1131/linee-guida-e-standard

JISC = Joint Information Systems Committee Digital Media https://www.webarchive.org.uk/wayback/archive/20160101151358/http://www.jiscdigitalmedia.ac.uk/infokit/file_formats/digital-file-formats

LAC = Library and Archives Canada http://www.bac-lac.gc.ca/eng/Pages/home.aspx

LOC = Library of Congress https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml

NAA = National Archives of Australia http://www.naa.gov.au/information-management/managing-information-and-records/preserving/long-term-file-formats.aspx

NAN = National Archives of the Netherlands https://www.nationaalarchief.nl

NARA = The U.S. National Archives and Records Administration https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html

NDL = National Digital Library's Digital preservation service (Finland) http://digitalpreservation.fi/en/specifications

OM = Österreichische Mediathek https://www.mediathek.at/

PREFORMA – PREservation FORMAts for culture information and e-archives http://www.preforma-project.eu/media-type-and-standards.html

PRONOM – Digital Preservation Department of the UK National Archives http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

SIA = Smithsonian Institution Archives https://siarchives.si.edu/what-we-do/digital-curation/recommended-preservation-formats-electronic-records

SMPTE = Society of Motion Picture and Television Engineers https://www.smpte.org/

SNA = Swedish National Archives https://riksarkivet.se/startpage

UNT = University of North Texas Libraries, Digital projects unit https://library.unt.edu/digital-projects-unit/standards/

UW = Universität Wien https://datamanagement.univie.ac.at/en/about-phaidra/formats/