CCA Format Policy Registry

Version 2
Approved: April 11, 2018

Introduction

General Introduction

The CCA Format Policy Registry (FPR) defines the file format policies implemented by CCA's digital preservation program. CCA always keeps each digital object accessioned into the Collection in its original file format. For files in some formats, derivative versions in **Preservation Formats** are also created and stored alongside original files in the Dark Archive. The FPR defines which file formats receive this type of treatment.

Preservation Formats have qualities that increase their likelihood of being usable over a longer period of time. These qualities include:

- Wide use and support
- Stable and preferably open specifications
- Uncompressed or lossless compression

When Preservation Formats are created, they are stored alongside the original files in the Dark Archive in Archival Information Packages (AIPs).

For consultation by users, files are delivered in Dissemination Information Packages (DIPs), which by default contain copies of files with their original filenames, last modified dates, and file formats preserved. This is an intentional choice to ensure that complex digital objects with links between files (CAD files, HTML/CSS/JS web assets, source code, et al) continue to work in the packages we provide for end-user access. When original formats prove difficult for researchers, CCA can provide normalized preservation files on demand.

Levels of Support

Levels of support indicate CCA's ability to maintain the usability of a digital object over time. These assessments are made on the basis of file formats. CCA has three levels of support: **Bit-level**, **Watch**, and **Full.**
All files in the CCA Dark Archive are stored on secure and backed-up servers and are regularly audited using checksums to ensure that no files have corrupted or changed in any way.

Bit-level

The minimum level of treatment that all digital objects accessioned into the CCA Collection receive.

At this level, CCA preserves the bitstream (i.e. the 1s and 0s that comprise the code) of a file exactly as-is. No format migration is performed. This practice ensures that CCA is able to provide an exact copy of original files over time. It does not necessarily ensure that files will be usable by software available at a future point in time.

Any formats not explicitly mentioned in the Format Policy Registry are preserved by CCA at the Bit-level.

Watch

File formats at this level of support are those for which CCA is currently only able to offer Bit-level support, but for which we hope to provide Full support in the future. This may be because the formats are common or highly valued at CCA (such as computer-aided design models) or because there is reason to believe that developments in the software industry and digital preservation community will make it easier to perform high-quality batch file format migrations or other means of access (e.g. emulation) in the future.

Full

File formats at this level of support are those for which CCA has high confidence in their long-term usability, either because the original format is already a preferred Preservation Format, or because CCA consistently and reliably normalizes files in this format to a documented Preservation Format.

Format Policy Registry

Text and Word Processing

Type	File Formats	Preservation Format	Level of support	Notes
Plain text	txt, csv, tsv	Original	Full	Preferred format
Rich text	rtf	Original	Full	
Microsoft Office Open XML Word	docx, docm	Original	Full	Preferred format
Microsoft Word and Works (legacy)	doc, wps	docx	Full	
OpenDocument Text Document	odt	Original	Full	Preferred format
Clarisworks Word Processor	cwk	odt	Full	
WordPerfect	wpd	odt	Full	
Markup languages and scripts	py, sh, rb, md, xml, et al.	Original	Full	
WordStar	ws	Original	Watch	

Presentation

Type	File Formats	Preservation Format	Level of support	Notes
Microsoft Office Open XML Powerpoint	pptx	Original	Full	Preferred format
Microsoft Powerpoint (legacy)	ppt	pptx	Full	
OpenDocument Presentation Document	odp	Original	Full	Preferred format
Keynote	key	Original	Watch	

Portable Document Format (PDF)

Type	File Formats	Preservation Format	Level of support	Notes
PDF/A	pdf/a	Original	Full	Preferred format
Standard PDF	pdf	pdf/a	Full	

Desktop Publishing

Type	File Formats	Preservation Format	Level of support	Notes
Quark XPress	qxd, qxp, qxl, qxt	Original	Watch	Files may be selectively manually normalized to PDF/A during processing on a case-by-case basis.
Adobe InDesign	indd	Original	Watch	Files may be selectively manually normalized to PDF/A during processing on a case-by-case basis.
Adobe Pagemaker	dmd, pmt	Original	Watch	Files may be selectively manually normalized to PDF/A during processing on a case-by-case basis.

Data sets

Spreadsheets

Type	File Formats	Preservation Format	Level of support	Notes
Microsoft Office Open XML Excel	xlsx	Original	Full	Preferred format
Microsoft Excel (legacy)	xls	xlsx	Full	
OpenDocument Spreadsheet	ods	Original	Full	Preferred format
Clarisworks Spreadsheet	cwkx	ods	Full	
Lotus 1-2-3	wk1, wks	Original	Watch	

Databases

Type	File Formats	Preservation Format	Level of Support	Notes
SQL-based	various	Original	Watch	Original formats preserved for now. Investigate SIARD and related projects.
NoSQL	various	Original	Watch	

Architecture and Design Models/Drawings

Type	File Formats	Preservation Format	Level of support	Notes
Computer-aided design (CAD) and 3D modeling	dwg, dxf, dwf, 3ds, max, 3dm, dgn, ma, mb, wire, fmz, stl, igs, step	Original	Watch	
Building Information Modeling (BIM)	rvt, ifc	Original	Watch	IFC is ideal format, though scalable migration pathways remain unclear.

Images

Raster images

Type	File Formats	Preservation Format	Level of support	Notes
TIF	tif	Original	Full	Preferred format
GIF (animated or still)	gif	Original	Full	Preferred format
Windows Bitmap	bmp, jpg, jp2, pct, png, psd, tga	uncompressed tif	Full	
JPEG	jpg	uncompressed tif	Full	
JP2 (JPEG 2000 part 1)	jp2	uncompressed tif	Full	
PICT	pct	uncompressed tif	Full	
PNG	png	uncompressed tif	Full	
Adobe Photoshop	psd	uncompressed tif	Full	
Truevision TGA Bitmap	tga	uncompressed tif	Full	
Silicon Graphics Image	sgi	uncompressed tif	Full	

Vector images

Type	File Formats	Preservation Format	Level of support	Notes
Scalable Vector Graphics	svg	Original	Full	Preferred format
Illustrator files	ai, eps	svg	Full	

Camera Raw

Type	File Formats	Preservation Format	Level of support	Notes
Digital Negative	dng	Original	Full	Preferred format
Camera raw formats	3frm, arw, cr2, crw, dcr, dng, erf, kdc, mrw, nef, orf, pef, raf, raw, x3f, ari, bay, cap, data, dcs, drf, eip, fff, iiq, k25, mdc, mef, mos, nrw, obm, ptx, pxn, r3d, rwl, rw2, rwz, sr2, srf, srw	Original	Watch	Camera raw files may be normalized to Adobe DNG prior to ingest if they belong to the Photography collection or are particularly important to the work of an archive's creator and CCA does not have corresponding raster images.

Video

Type	File Formats	Preservation Format	Level of support	Notes
Matroska	mkv	Original	Full	Preferred format (ffv1 encoding)
MOV	mov	ffv1/matroska	Full	
AVI	avi	ffv1/matroska	Full	
MPEG-1	mpg, mpeg	ffv1/matroska	Full	
MPEG-2	mpg, mpeg, mp2	ffv1/matroska	Full	
MPEG-4	mp4	ffv1/matroska	Full	
Macromedia FLV	flv	ffv1/matroska	Full	
Material Exchange Format	mxf	ffv1/matroska	Full	
Windows Media	wmv	ffv1/matroska	Full	

Video				
Digital video from media such as MiniDV, DVCAM, DVCPro, Digital Betacam, Digital-8	(DV in various container formats)	Original	Full	Wrapped in AVI container. DV often contains key metadata that would be lost in transcoding to other formats such as ffv1/matroska.

Note: CCA's preferred lightweight access format for streaming and internal use is H264 (video)/AAC (audio)/MPEG-4 (container)

Audio

Type	File Formats	Preservation Format	Level of support	Notes
WAVE	wav, bwf	Original	Full	Preferred format
Audio Interchange File Format	aiff	wav	Full	
MPEG-3	mp3	wav	Full	
Advanced Audio Coding	aac	wav	Full	
Windows Media Audio	wma	wav	Full	

Websites

Type	File Formats	Preservation Format	Level of support	Notes
Web ARChive	warc	Original	Full	Preferred format
HTML	html	Original	Full	

Email

Type	File Formats	Preservation Format	Level of support	Notes
MBOX	mbox	Original	Full	Preferred format
Outlook inbox	pst	Original	Watch	
Maildir	maildir	Original	Full	

E-publications

Type	File Formats	Preservation Format	Level of support	Notes
EPUB	Epub	Original	Full	Preferred format

Geospatial (GIS) data

Type	File Formats	Preservation Format	Level of support	Notes
ArcGIS geodatabase file	gdb	Original	Watch	
Shapefile	shp	Original	Watch	

Video games

Type	File Formats	Preservation Format	Level of support	Notes
All formats		Original	Bit-level	

Executables, installation files, and binaries

Type	File Formats	Preservation Format	Level of support	Notes
All formats		Original	Bit-level	