Deep Research, Wide Results: Updating Format Descriptions in NARA's Digital Preservation Framework

Leslie Johnston, Amanda May, Mackenzie Beasley

What is the Digital Preservation Framework?

NARA Digital Preservation Framework

The Digital Preservation Framework is both an internal and external resource for assessing risks posed to file formats in NARA's holdings and planning for their long-term preservation. It represents NARA's current thinking and internal capabilities.

The Framework consists of:

- File Format Risk Matrix
- File Format Preservation Action Plans
- Record Category Preservation Action Plans

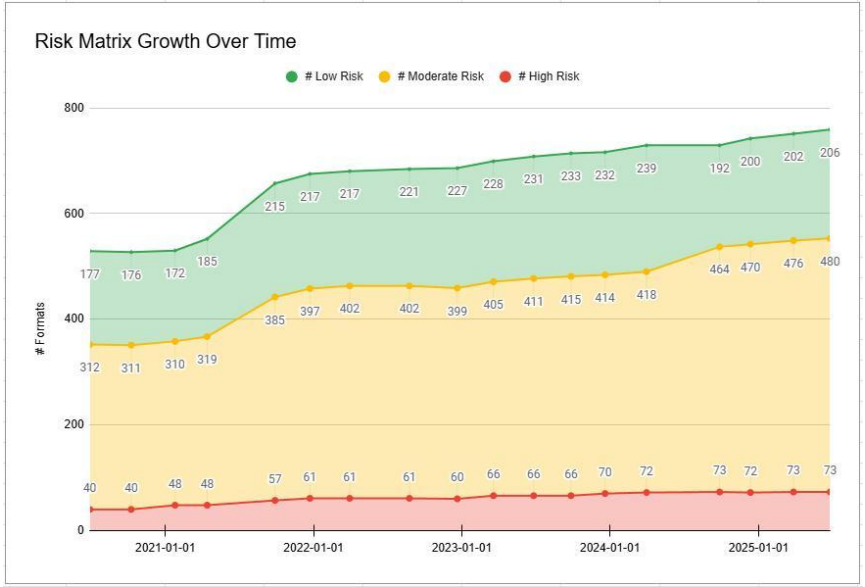
The Framework was first published in its current form in 2020, is updated quarterly, and is currently made available on GitHub and as linked data on archives.gov for reuse and adaptation.

The intended audiences for the Framework are (1) NARA staff and (2) the international community of digital preservation practitioners.

File Format Risk Matrix

- Risk factors for file formats fall into eight categories:
  - Disclosure
  - Adoption
  - Transparency
  - Self-Documentation
  - External Hardware Dependencies
  - External Software Dependencies
  - Impact of Patents
  - Technical Protection Mechanisms

- By answering 27 questions related to the ability to preserve and sustain a file format, the weighted answers generate numeric scores for relative risk levels in the NARA context (Low / Moderate / High).



Preservation Action Plans

There are two types of Preservation Action Plans:

- File format plans: over 750 plans made available in a spreadsheet on GitHub and as linked data on archives.gov
  - Document the outcome of the Risk Matrix assessment (Low / Moderate / High Risk)
  - Collate links to format specifications or documentation
  - Describe the format and briefly justify its risk level
  - Identify whether the format is “Preferred” or “Acceptable” per the NARA Transfer Guidance
  - Recommend NARA preservation outcomes, preferred tool(s), and available viewer(s)

- Record Category Plans: 16 plans made available on GitHub. Documents significant properties to be retained in a format migration. These can be used as metrics to test potential tools for preservation migration.
  - Appearance
  - Structure
  - Behavior
  - Context

2024-2025 Improvements

Risk Matrix Update and Rescoring

In 2024, NARA's Digital Preservation Unit overhauled the Risk Matrix to make it more precise and easier to use.

- Reduced the number of questions from 39 to 27, and 9 categories to 8. Many questions asked nearly the same thing. While redundancies were removed, 8 new questions were added to address additional risk factors that emerged over the years.
- New scoring logic - added N/A options, updated guidance in the SOP, and added "If/Then" language to the questions. Added justifications for each question.
- New weighting - some questions impact the final risk score more than others. Answering "unknown" for questions impacts the score negatively, because it is not possible to plan for and mitigate unknown risks.

Rescoring each of the formats based on the new Matrix took two Senior Digital Preservation staff three months.

Format Description Updates

- A template was put in place for standardizing file format descriptions:
  - One sentence that states the name of the format, its abbreviation (if applicable), and what it is generally used for.
  - One sentence about the history of the format/software.
  - One sentence that describes how this format is different from related formats in its family/domain.
  - One to two sentences summarizing whether it is open/proprietary, its general adoption or use in a community, and whether or not there is software available for rendering.
  - One to two sentences justifying the proposed preservation action plan.
  - One sentence on Transfer Guidance (if it applies)

- Addresses inconsistencies in insufficient data, content, and tone:
  - Example: Graphics Interchange Format (GIF) 87a (NF00202)
  - **Before**: Widely adopted. Many software tools exist for encoding and decoding. Natively supported by web browsers.
  - **After**: Graphics Interchange Format (GIF) is a raster image web graphics format that stores bitmap images in a single file. The 87a version is the original version and was created by CompuServe in 1987. GIF used patent-protected LZW lossless compression; the patent was owned by Unisys but expired in 2003. GIF has open specifications and is well-documented and in wide use with a variety of software to render. It is recommended to retain GIF files as-is since the files can easily be opened and are prevalent in the field.

- Over 750 formats to review!

Research Process

File Format Research Process

Each File Format in the NARA Framework is the subject of an intensive research process:

- The history of the format and related formats
- The history of the original associated applications
- The history and maintenance of applicable format standards and specifications
- Documentation of current state of technology related to the format, including standards and specifications, MIME type, and current application support for the use and/or rendering of the format

This work requires the use of authoritative community resources, technical documentation, and research products produced by other organizations and experts:

- Library of Congress Format Descriptions, the British Library, PRONOM, Wikidata, ArchiveTeam
- Organizational Preservation Plans and Statements on Preferred and Acceptable Formats

These resources can provide contradictory information on the history of a format and its original context and applications, requiring critical analysis and extensive review of specifications in particular.

Framework Research Output: Preservation Plans

Format Name	File Extension(s)	Category/Plan(s)	NARA Format ID	MIME type(s)	Specification/Standard URL
SIARD 2.2	siard	Databases	NF00797	text/xml+zip	https://github.com/DILCISBoard/SIARD/blob/master/specific
Sigma RAW	x3f	Digital Still Image	NF00476	image/x-raw-sigma	
Signature Block	sig	Email	NF00716	text/plain	
Sonic Scenarist Closed Caption Format	scc	Digital Video	NF00709	application/octet-stream	http://www.theneitherworld.com/mcpoodle/SCC_TOOLS/DC
Sony Alpha RAW 1.x	arw	Digital Still Image	NF00809	application/octet-stream	https://github.com/iclevy/sony_raw
Sony RAW	srf	Digital Still Image	NF00477	image/x-raw-sony	
Sound Designer II Audio File	sd2	Digital Audio	NF00407		
SoundFont 2 Sound Bank	sf2	Digital Audio	NF00814	audio/octet-stream	http://www.synthfont.com/sfspec24.pdf
Spatial Data Transfer Standard (SDTS)	ddf	Geospatial	NF00408	application/x-ogc-sdts	https://web.archive.org/web/20161208010908/http://mcmcw
Standard for the Exchange of Product Model Data (STEP)	step|stp|p21	Digital Design and Vector Graphics	NF00409	application/octet-stream	https://www.iso.org/standard/63141.html
Standard Generalized Markup Language (SGML)	sgm|sgml	Structured Data	NF00410	application/sgml|text/sgml	https://tools.ietf.org/html/rfc1874
STATA Data file version 118	dta	Presentation and Publishing|Structured Data	NF00696	application/x-stata-dta	https://www.stata.com/help.cgi?dta
Static Library	a|lib	Software and Code	NF00199	application/octet-stream	
Stationery for Apple Mail	doc	Email	NF00412		
Statistica Report File	str	Presentation and Publishing	NF00413		
Structured Data eXchange Format	sdxf	Structured Data	NF00415		https://www.ietf.org/rfc/rfc3072.txt
Structured Query Language	sql	Software and Code|Databases	NF00416	text/plain	https://webstore.ansi.org/Standards/ISO/ISOIEC90752011
SubRip Video Subtitle	srt	Digital Video	NF00723	text/plain	https://www.matroska.org/technical/subtitles.html#srt-subtitl
Sun Raster Image	sun|ras|sunras|rs|sr|scr|im1|im	Digital Still Image	NF00417	image/x-sun-raster|image/x-cmu-r	http://www.martinreddy.net/gfx/2d/RAS.txt
Synchronized Accessible Media Interchange	smi|sami	Digital Video	NF00689	application/x-sam	
Synchronized Multimedia Integration Language	smi|smil	Structured Data	NF00782	text/smil|applicati	
Tab Separated Values	tab|tsv	Structured Data	NF00418	text/tab-separate	
Tagged Image File Format (TIFF) 1-6	tif|tiff	Digital Still Image	NF00419	image/tiff	

NARA Risk Level	NARA Preservation Action	NARA Proposed Preservation Plan	Description and Justification	Preferred Processing and Transformation
Low Risk	Retain	Retain	SIARD is the emerging XML-based standard in	DBPTK Database Preservation Toolkit
Moderate Risk	Transform	Transform to DNG	Preferable to convert from proprietary	Adobe Photoshop with Camera RAW; Adobe DN
Moderate Risk	Retain	Retain	Signature Block files, or sig files, contain an	Any supported text editor; Microsoft Outlook
Moderate Risk	Retain	Retain	Sonic Scenarist Closed Caption (SCC) is a	FFmpeg; Adobe Premiere
Moderate Risk	Transform	Transform to DNG	Sony Alpha RAW, also known as ARW, is a	Adobe Photoshop with Camera RAW; Adobe DN
Moderate Risk	Transform	Transform to DNG	Preferable to convert from proprietary	Adobe Photoshop with Camera RAW; Adobe DN
High Risk	Transform	Transform to WAV	The file format is not well documented and	Audacity
Moderate Risk	Transform	Transform to MP3 or WAV	The SoundFont 2 format is based on RIFF. The	FFmpeg; VLC
Moderate Risk	Transform	Transform raster data to TIFF and transfo	Transformation is required as the software is	FME; Global Mapper
Low Risk	Retain	Retain	The file format is well documented, is a	FME; Global Mapper; FreeCAD
Low Risk	Retain	Retain	SGML is no longer actively used, having been	Any supported text editor or web browser
Moderate Risk	Transform	Transform to a TBD format	STATA dta is a proprietery binary format for	Procure and/or develop tools. Stata dta files can
Moderate Risk	Retain	Retain	A static library is a set of routines, external	
High Risk	Retain for Future Assessment	Further research is required	An email stationery file created and used by	This can only be viewed in Apple Mail. It is uncle
Moderate Risk	Transform	Transform to PDF	The STATISTICA Report format (.str) is used	Controlled Image Base (CIB); Compressed ARC
Moderate Risk	Retain	Retain	The format is very uncommon though well	Further research is required
Low Risk	Retain	Retain	SQL (Structured Query Language) is a	Any supported text editor
Low Risk	Retain	Retain	The SubRip programs saves subtitles ripped	Any supported text editor
Moderate Risk	Transform	Transform to TIFF	Sun Raster is a proprietary raster image file	IrfanView; GNU Image Manipulation Program
Low Risk	Retain	Retain	SAMI is a plain text caption/subtitle files and	Any supported text editor
Low Risk	Retain	Retain	Synchronized Multimedia Integration Language	Any supported text editor
Moderate Risk	Transform	Transform to CSV	Tab separated values (TSV) is a simple,	LibreOffice; Microsoft Excel
Low Risk	Retain	Retain	Tagged Image File Format (TIFF) version 1-6 is	Adobe Photoshop; IrfanView; GNU Image Manip

Research Applications

- NARA Digital Preservation
  - The primary research goal is to collate documentation on file format technologies and risk to determine any actions that NARA needs to take
- NARA Custodial Units
  - Technical overview of holdings across all areas
  - Provide authoritative research to support:
    - Processing of accessions
    - Preservation of files
- Archival Community Worldwide
  - Complement other format registries
  - Preserve data about legacy formats
  - Share information about best practices, useful tools

Thank you!

Leslie Johnston, Digital Preservation

leslie.johnston@nara.gov

Amanda May, Archives Specialist

amanda.may@nara.gov

Mackenzie Beasley, Archives Specialist

mackenzie.beasley@nara.gov