Applied Digital Preservation and Risk Assessment

Leslie Johnston

NISO Plus 2022

What is the Discipline of Digital Preservation?

- Any digital object is in scope for digital preservation, whether born-digital or digitized.
- Digital preservation encompasses all format types: texts and images, databases and spreadsheets, vector or raster images, software, email and social media, games, movies, music and sound, and the web.
- With every IT innovation, digital preservation managers must respond by devising effective strategies for ensuring the durability and ongoing accessibility and usability of new digital materials, so digital preservation will remain an always-emerging challenge.

There are Several Digital Preservation Standards & Models

- **Open Archival Information System (OAIS) for trusted digital repository systems--ISO 14721:2012—applied and measured with ISO 16363:2012**
  - A reference model for a trusted digital repository (both the organization and the system(s) with metrics covering the organization, digital object management, and IT infrastructure and security.
  - https://www.iso.org/standard/57284.html and https://www.iso.org/standard/56510.html
- **Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)**
  - An assessment that includes identification of an organization's mandate, constraints, objects, activities, and assets, in addition to several risk categories: physical environment; personnel, management, and administrative procedures; operations and service delivery; and hardware, software, communication equipment, and facilities.
  - http://www.repositoryaudit.eu/
- **Simple Property-Oriented Threat (SPOT) Model for Risk Assessment**
  - Focused on safeguarding against threats to six properties of digital objects: availability, identity, persistence, renderability, understandability, and authenticity.
  - https://www.oclc.org/research/publications/2012/threats-successful-digital-preservation.html
- **National Digital Stewardship Alliance (NDSA) Levels of Digital Preservation**
  - A lightweight, overall digital preservation program maturity measure with categories of metrics that cover storage, integrity, control, metadata, and content.
  - https://ndsa.org/publications/levels-of-digital-preservation/
- **Digital Preservation Coalition Rapid Assessment Model (DPC RAM)**
  - A program maturity modelling tool that helps an organization benchmark their progress in digital preservation in organizational and service capabilities.
  - https://www.dpconline.org/digipres/dpc-ram
- **CoreTrustSeal**
  - Aimed at data repositories that want a core level certification based on the DSA–WDS Core Trustworthy Data Repositories Requirements catalogue and procedures.
  - https://www.coretrustseal.org/
- **Digital Archiving Graphical Risk Assessment Model (DIAGRAM)**
  - Focused on intellectual control and renderability, as measured by more operational metrics: operating environment, replication and refreshment, physical disaster, storage medium, technical skills, digital objects, information management, system security, and checksums.
  - https://nationalarchives.shinyapps.io/diagram-jr/

Which one is “The Best?”

- None of Them.
- All of Them.
- It Depends. What are your goals for assessing your digital preservation program and the risks related to your collections?

Applied Work at NARA

The First Step: A Guiding Policy/Strategy

- NARA published its first Digital Preservation Strategy in June 2017 to guide its internal operations.

https://www.archives.gov/preservation/electronic-records.html

- This outlines the specific strategies that NARA will use in its digital preservation efforts, and specifically addresses:
  - Infrastructure
  - Format & Media Sustainability and Standards
  - Data Integrity
  - Information Security

- It applies to born-digital agency electronic records, digitized records from agencies, and NARA digitization for access and preservation reformatting.

The File Format Risks at NARA

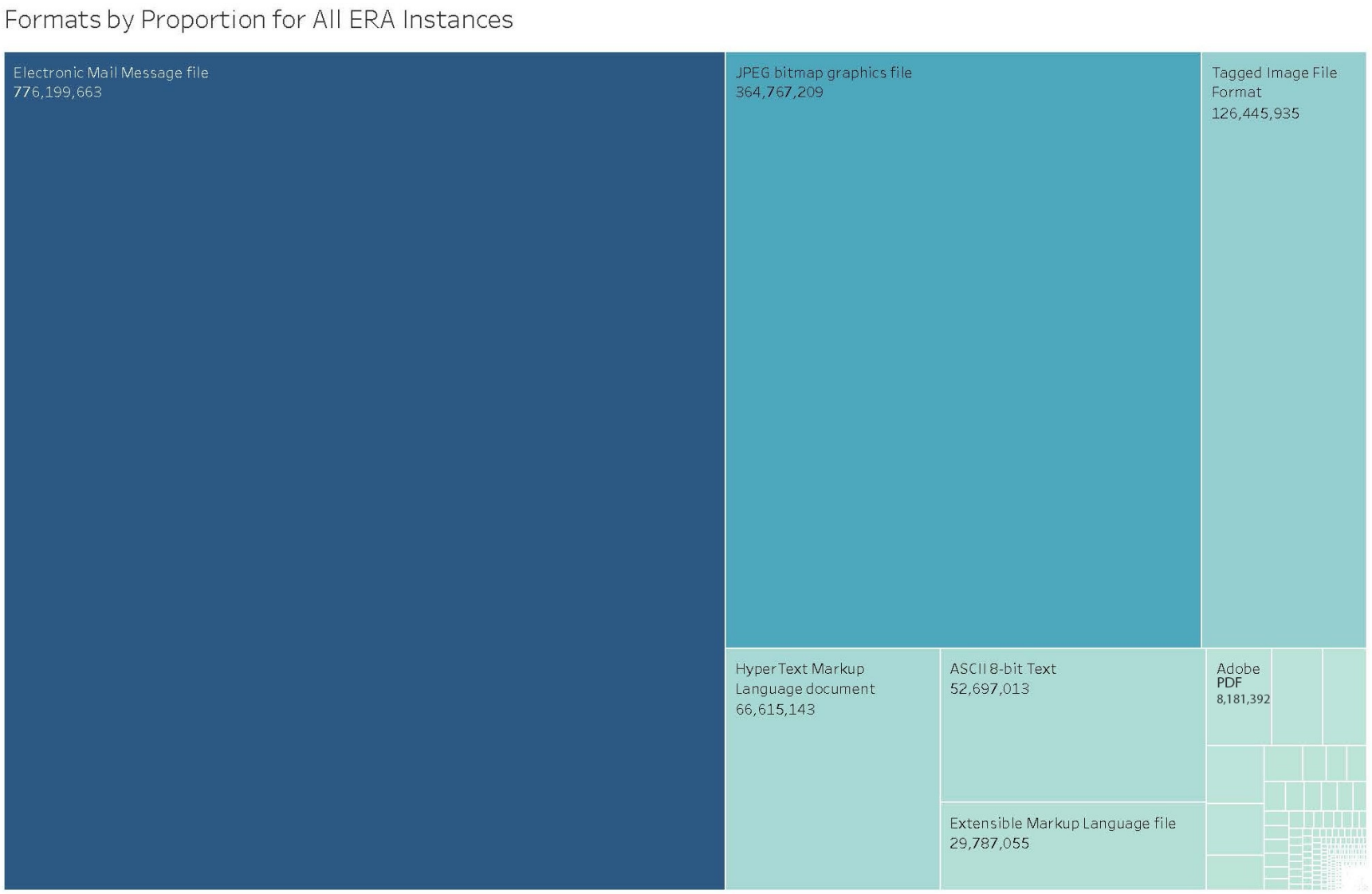
- An integral part of NARA's work is the issuance of guidance on all aspects of Federal electronic records management and transfer to NARA, including media types, file formats, and metadata.

https://www.archives.gov/records-mgmt/policy/transfer-guidance.html

- By regulation, NARA cannot be 100% proscriptive in the formats it accepts. When records are transferred, they are validated to ensure that they are uncorrupted, and, if possible meet NARA's format guidance. There are "Preferred" and "Acceptable" formats, but sometimes has to take in records in the format the agencies have because those are the tools and formats they use to do their jobs, and there must always be exceptions.

Creating a Collections Format Profile

- NARA has several electronic records systems: Federal Records, Congressional Records, Census, and two different systems for Presidential Records. This meant we had no single profile or measure of what NARA has in its holdings.
- A manual process was used to combine reports from all the systems to create a list of the formats in the holdings since the reporting didn't match in terms of granularity for the various systems, given different tooling for format analysis and reporting.
- There were different granularity levels reported for file formats, e.g., files identified as Adobe Acrobat PDF vs. files identified as Adobe Acrobat PDF 1.4. This required normalization when aggregating the data together to compare across the holdings.
- Not every file could be characterized and mapped to documented formats with certainty.
- Several hundred file formats are present in the holdings if one counts all the variations of PDF or Microsoft Word, for example.
- There were discoveries, such as decisions made in the past about format normalization in one portion of the holdings meant to improve access that had to be taken into account.



NARA File Format Analysis by Proportion in the Holdings

Assessing Risk

- The Holdings Format Data was vital to assess risk in the electronic records holdings.
- In 2018 NARA created an extensive Risk Matrix, designed to apply a series of weighted factors related to the preservation sustainability of the file formats in the Collection Format Profile to generate a numeric score.
- Each question has a relative weighting that maps to the level of risk for each question and, to the extent that it can be defined, resource costs (staff time or budget).
- The Matrix also includes high level factors that assess the preservation actions that could be taken vis-à-vis our current environment and capabilities.
- The Matrix calculates numeric scores, which are mapped to High, Moderate, and Low Risk. The risk thresholds are open to review and revision over time.

What are Some of the Assumptions?

- The openness of a format and availability of full documentation--which enables the development of tools to process that format and/or perform preservation format transformations--provides a higher positive effect than the lack of openness and documentation.
- The level of adoption of a format translates to a higher likelihood of the availability of tools that read, display, or transform the format. A low level of adoption provides an equal negative effect on format sustainability.
- The ability to represent and analyze formats directly adds to the sustainability rating, and the inability to do has an equal negative impact.
- The presence of self-documentation, where a file describes its own characteristics provides a higher positive impact than negative. All files can provide some basic technical information, but not all can have descriptive metadata embedded, so all file formats are self-documenting to some degree.
- The requirement to maintain specific software (or, in some cases, hardware) for processing or access to formats has a higher negative impact on sustainability that the lack of required software does on positive impact. Requiring such software or operating systems has cost and expertise implications.
- The presence or absence of licenses or patents and open source licensing status have limited and equal positive or negative impacts on the sustainability.
- The age of a format is an additive risk factor; all formats have inherent risk, especially the lack of tools to read, render, or transform the format, so there are no potential positive impacts; risk increases based on the age of the format and the currency of its versions.

Preferred	Acceptable	Risk Level	NARA Forr	Format Name	File Extension(s)	Record Type/Plan(s)	Is the format proprietary?	Does the format have a published open specification?	Are there available tools that can validate the technical integrity of a file encoded in this format against the published specification?	Has the specification been approved and published by an internationally recognized standards body?	Is the available specification complete and accurate?	Total Disclosure Score. Highest possible score = 10; Lowest possible score = -6
		Low Risk	NF00139	CALS Compressed Bitmap	cal; ct1; ct2	Digital Still Image	2	2	2	0	2	8
		High Risk	NF00468	Canon RAW 1.0	crw	Digital Still Image	-1	-1	-1	-1	-1	-5
		High Risk	NF00469	Canon RAW 2.0	cr2	Digital Still Image	-1	-1	-1	-1	-1	-5
		Low Risk	NF00141	Cascading Style Sheets 1.0	css	Web Records; Software	2	2	0	2	2	8
		Low Risk	NF00543	Cascading Style Sheets 2.0	css	Web Records; Software	2	2	0	2	2	8
		Low Risk	NF00544	Cascading Style Sheets 2.1	css	Web Records; Software	2	2	0	2	2	8
		Moderate Risk	NF00191	CCITT T.4 Group 3 compression (Fax image file)	tiff; others	Digital Still Image	0	0	0	-2	0	-2
		Moderate Risk	NF00488	CCITT T.6 Group 4 compression (Fax image file)	tiff; others	Digital Still Image	0	0	0	-2	0	-2
		Moderate Risk	NF00411	Checksum File	sum	Software and Code	-1	-1	-1	-2	0	-5
		Moderate Risk	NF00545	Cold Fusion Component File	cfc	Software and Code	-1	-1	0	-2	0	-4
		Moderate Risk	NF00142	Cold Fusion Markup Language	cfm	Software and Code	-1	-1	0	-2	0	-4
X		Low Risk	NF00143	Comma Separated Values	csv	Structured Data	2	2	2	2	2	10
		Low Risk	NF00144	Common Data Format Toolkit	cdf	Software and Code	-1	2	2	-1	2	4
		Moderate Risk	NF00111	Compressed Archive File	arc	Software and Code	2	2	2	-2	2	6
		Moderate Risk	NF00145	Computer Graphics Metafile Binary 1	cgm	Multimedia	2	2	0	2	2	8
		Moderate Risk	NF00546	Computer Graphics Metafile Binary 2	cgm	Multimedia	2	2	0	2	2	8
		Moderate Risk	NF00547	Computer Graphics Metafile Binary 3	cgm	Multimedia	2	2	0	2	2	8
		Moderate Risk	NF00548	Computer Graphics Metafile Binary 4	cgm	Multimedia	2	2	0	2	2	8
		Moderate Risk	NF00146	Configuration File	ini; conf; cnf; cfg; cf	Software and Code	2	-1	-1	-2	0	-2
		High Risk	NF00148	Corel CMX Compressed	cpx	Digital Design	-1	-1	0	-2	-1	-5
				CorelDraw Compressed								

NARA Risk Matrix

Preservation Action Plans

- Risk Assessment is not enough – it must be translated into actionable plans to mitigate the risks.
- The Plans identify essential characteristics for electronic records held by NARA, document file format risk, and collate links to specifications and other digital preservation resources. The recommended preservation tools and actions for formats included in the Plans are based on current NARA decisions and capabilities.
- The Plans consist of two sets of documents:
  - Record Type Plans which document the characteristics of different categories of records:

Calendars	GIS
Databases	Multimedia/Publishing/Presentations
Digital Audio	Navigational Charts
Digital Cinema	Software Code
Digital Design/CAD	Spreadsheets
Digital Still Image	Structured Data
Digital Video	Web Records
Email	Word Processing

  - Preservation Action Plans: A single spreadsheet containing over 500 file formats across all record types.

What Do the Plans Include?

- Record Type Plans
  - Documentation of Essential Characteristics for record types (Email, GIS, Databases, Still Images, etc.)
    - Appearance, Structure, Behavior, and Context.
  - These are the properties that should, if possible, be retained in any format migration, and are used as metrics to test potential tools for the preservation migrations.

- Preservation Action Plans
  - Current NARA assigned level of preservation risk and priority for preservation actions.
  - Links to specifications, both official and reverse-engineered by the community.
  - Links to community resources.
  - Format description.
  - Recommended preservation migration actions, including no action if appropriate.
  - Recommended tools for processing and preservation.

Preservation Action Plan: Digital Still Image AKA Raster Images
National Archives and Records Administration (NARA)
Plan Date: 20200629
Template: 201907

Electronic Record or Digital Surrogate Types and Associated Formats

Digital still images are digitally encoded representation of the tonal and brightness information of a subject into a bitmap. Data from digital cameras and scanning devices record light characteristics as numerical values into a grid or raster of picture elements (pixels). The term raster data is often contrasted with vector data, in which geometrical points, lines, curves, and shapes are based upon mathematical equations, thus creating an image without specific mapping of data to pixel. Bit-depth, spatial resolution, and color encoding, for example, are all important characteristics of still images.

There are two types of raster file digital image record types: Digital still photographs of natural, real-world scenes or subjects produced by digital cameras, and scanned images of textual documents, illustrations, posters, graphics, cartographic records, photographic prints, slides, and negatives. Image file formats are standardized means of organizing and storing rasterized data that can be used on a computer display or printer.

Essential Characteristics of Digital Still Image/Raster Images

To render an authentic digital photograph one must preserve the structural, technical, and descriptive metadata that allow certain appearance characteristics to persist. Many of the characteristics native to raster image file formats are the result of industry efforts to develop common standards and interoperability. Many file formats for digital photography and scanning are the same except that most digital cameras create native camera raw proprietary formats, JPEG, and DNG, and rarely TIFF, JPEG2000, PNG, or GIF.

Appearance is a critical characteristic for this record type due to the common purpose or use of this type of record is to depict scenic information or to render the informational and artifactual aspects of a scanned original. Tone fidelity, resolution, bit depth, color encoding as well as compression algorithms all contribute to the preservation of the file. There is widespread adoption of most formats and rendering and display platforms. A unique characteristic for scanned multi-page documents is descriptive and administrative metadata that may be held external to the electronic record and could be a risk for long term identification of the context.

Appearance

Name	Definition	Function Description
Size	Determined by bit-depth, spatial resolution, compression, and color encoding.	
Color	Color mode, color space.	Mathematical representations of color information needed to encode and decode color information such as Hue, Chroma, lightness, white point.
Bit-depth	The number of bits used to indicate color and tone information of a pixel.	High or low bit depth contributes to the pleasing transformation of color accuracy, gradients, and tonal information. Also greatly affects issues such as signal clipping and transformative image editing.
Orientation	Portrait versus Landscape.	

Structure

Name	Definition	Function Description
Layout Structure	Embedded technical metadata captured at the time of creation describing, among other things: File format/encoding; Compression; Resolution; Bit depth; and EXIF (Exchangeable Image File Format) information.	

Example NARA Record Type Preservation Plan

Format Name	Extension(s)	Record Type/Plan(s)	NARA Format ID	MIME type(s)	Specification/ Standard URL	PRONOM URL	LOC URL	British Library URL	WikiData URL	ArchiveTeam URL	ForensicsWiki URL	Wikipedia URL
Canon RAW 1.0	crw	Digital Still Image	NF00468	image/x-canon-crw; image/x-raw-canon	https://exiftool.org/canon_raw.html	https://www.nationalarchives.gov.uk/PRONOM/fmt/593	https://www.loc.gov/preservation/digital/formats/fdd/fdd000241.shtml		https://www.wikidata.org/wiki/Q3651247	http://fileformats.archiveteam.org/wiki/Camera_Image_File_Format		https://en.wikipedia.org/wiki/Camera_Image_File_Format
Canon RAW 2.0	cr2	Digital Still Image	NF00469	image/x-raw-canon	http://lclevy.free.fr/cr2/	https://www.nationalarchives.gov.uk/PRONOM/fmt/592	https://www.loc.gov/preservation/digital/formats/fdd/fdd000241.shtml		https://www.wikidata.org/wiki/Q27866048	http://fileformats.archiveteam.org/wiki/Canon_RAW_2		
Cascading Style Sheets 1.0	css	Web Records; Software and Code	NF00141	text/css	https://www.w3.org/TR/CSS1	https://www.nationalarchives.gov.uk/PRONOM/x-fmt/224	http://www.digitalpreservation.gov/formats/fdd/fdd000482.shtml		https://www.wikidata.org/wiki/Q19942840	http://fileformats.archiveteam.org/wiki/Cascading_Style_Sheets		http://en.wikipedia.org/wiki/Cascading_Style_Sheets
Cascading Style Sheets 2.0	css	Web Records; Software and Code	NF00543	text/css	https://www.w3.org/TR/1998/REC-CSS2-19980512/	https://www.nationalarchives.gov.uk/PRONOM/x-fmt/224	http://www.digitalpreservation.gov/formats/fdd/fdd000482.shtml		https://www.wikidata.org/wiki/Q27826458	http://fileformats.archiveteam.org/wiki/Cascading_Style_Sheets		http://en.wikipedia.org/wiki/Cascading_Style_Sheets
Cascading Style Sheets 2.1	css	Web Records; Software and Code	NF00544	text/css	https://www.w3.org/TR/CSS2/	https://www.nationalarchives.gov.uk/PRONOM/x-fmt/224	http://www.digitalpreservation.gov/formats/fdd/fdd000482.shtml		https://www.wikidata.org/wiki/Q27826457	http://fileformats.archiveteam.org/wiki/Cascading_Style_Sheets		http://en.wikipedia.org/wiki/Cascading_Style_Sheets
CCITT T.4 Group 3 compression (Fax image file)	tiff; others	Digital Still Image	NF00191	image/g3fax		https://www.nationalarchives.gov.uk/PRONOM/x-cmp/14			https://www.wikidata.org/wiki/Q28234649	http://fileformats.archiveteam.org/wiki/CCITT_Group_3		https://en.wikipedia.org/wiki/Fax#Compression
CCITT T.6 Group 4						https://www.nationalarchives.g	https://www.loc.gov/preservation/digital/forma		https://www.wi	http://fileformats.archiveteam.		https://en.wikipedia.org/wiki/G

NARA File Format Preservation Plans

Format Name	Preservation Action	Proposed Preservation Plan	Description and Justification	Preferred Processing and Transformation Tool(s)
Canon RAW 1.0	Retain; Transform	Retain; Transform to Digital Negative Format (DNG)	Preferable to convert from proprietary manufacturer file to Adobe DNG. However, there is some support for camera raw files from major camera manufacturers and there is some desire to retain the raw sensor data.	Adobe Camera RAW; Adobe Photoshop; ACDSee Pro; darktable; dcraw; ExifTool; ImageMagick; LightZone
Canon RAW 2.0	Retain; Transform	Retain; Transform to Digital Negative Format (DNG)	Preferable to convert from proprietary manufacturer file to Adobe DNG. However, there is some support for camera raw files from major camera manufacturers and there is some desire to retain the raw sensor data.	Adobe Camera RAW; Adobe Photoshop; ACDSee Pro; darktable; dcraw; ExifTool; ImageMagick; LightZone
Cascading Style Sheets 1.0	Retain	Retain. Web archives should never be migrated, as that would change the fundamental linkages and functionality of web objects. Uncompiled code that is plain text ASCII or Unicode does not require format migration.	Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS files are plain text.	Any supported text editor.
Cascading Style Sheets 2.0	Retain	Retain. Web archives should never be migrated, as that would change the fundamental linkages and functionality of web objects. Uncompiled code that is plain text ASCII or Unicode does not require format migration.	Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS files are plain text.	Any supported text editor.
Cascading Style Sheets 2.1	Retain	Retain. Web archives should never be migrated, as that would change the fundamental linkages and functionality of web objects. Uncompiled code that is plain text ASCII or Unicode does not require format migration.	Cascading Style Sheets (CSS) is a style sheet language used for describing the presentation of a document written in a markup language like HTML. CSS files are plain text.	Any supported text editor.
CCITT T.4 Group 3 compression (Fax image file)	Transform	Transform to TIFF or JPEG2000	FAX files are generally a variant of the TIFF format. Experimentation is needed to determine the best tool for opening and converting these files.	Procure and/or develop tools; possible tools include Adobe Photoshop; Adobe Elements; GraphicsMagick; GNU Image Manipulation Program; ImageMagick; Farbfeld Utilities

NARA File Format Preservation Plans

Digital Preservation Framework 2020 Release

- The NARA Digital Preservation Framework(the Risk Matrix, the Record Type Plans, the Format Preservation Actions Plans, and guidelines for their use) were formally released on GitHub in June 2020. All forms of feedback and community use and adaptive re-use is welcome.

  https://github.com/usnationalarchives/digital-preservation

- The Framework can be applied across the electronic records lifecycle
- The Plans are iterative and are updated quarterly.
- A Linked Data release is in pilot.

NARA Digital Preservation Program Assessment

- In 2021 and 2019, NARA completed a self-assessment of its programs and systems using the PTAB (Primary Trustworthy Digital Repository Authorisation Body) instrument based on ISO 16363:2012.

  http://www.iso16363.org/iso-certification/preparation/

  https://public.ccsds.org/Pubs/652x0m1.pdf

- We readily acknowledge that there are gaps in our processes, documentation, and systems.

NARA Self-Assessment Outcomes

Fiscal Year 109 Metrics	**Metrics Met**	**Metrics Partially Met**	**Metrics Not Met**
2021	52	54	3
2019	32	64	13

How will NARA Proceed?

- NARA addressed 16363 self-assessment gaps in stages, starting with documentation. This included a definition of Designated Community, a key part of the metrics. 93 documents—SOPs, Sys Admin and User Guides, and IT Security documents—were created, updated, or added to the justifications.
- Some gaps are related to repository functionality, so priorities have shifted for system development.
- Some gaps come from different capabilities across different systems at NARA, so migration and consolidation is a priority.
- There is an ongoing risk assessment of new file formats, updates to guidance for records creators, assessment of the formats in the holdings, and updates to preservation plans as new tools are identified and acquired.
- Update the self-assessment every 2 years.

Thank You.

Leslie Johnston

Leslie.Johnston@nara.gov