Developing, Implementing, and Maintaining the NARA Digital Preservation Framework and the move to Linked Data

Leslie Johnston
National Archives and Records Administration

Starting with 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 into a spreadsheet.
- 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.

Assessing Risk in the Digital Preservation Framework

- In 2018 NARA created an extensive Risk Matrix spreadsheet, 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.

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								

Preservation Action Plans in the Framework

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 Category Plans**: Word Documents which comprise 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**: A single spreadsheet containing close to 700 file formats across all record types, containing the specifications, resource links, format information, and preservation actions for all formats across all record types to ensure the Framework's actionability and extensibility to other institutions.

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.	

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

Public Release and Maintenance

- The Framework is on GitHub as a CSV file (the format Preservation Action Plan) and PDFs (from the Record Category Plan Word documents), with quarterly updates.

  https://github.com/usnationalarchives/digital-preservation
- Maintaining the data
  - Links break!
  - New Formats need to be added
  - Incorporation of newly available resources
  - Revising available tools
  - New Record Categories are identified
    - Added Calendars and Navigational Charts
  - Risk Matrix
    - Changes due to factors such as format age and availability of specifications or tools
- Managing 1 dataset, 2 spreadsheets and 16 Word files.

Linked Data Goals

- Document the decision-making process and decisions made by NARA for the long-term preservation of specific file formats in its electronic record holdings in a Digital Preservation Framework.
- Share the Digital Preservation Framework in both human readable and machine-readable formats.
- Publish the Linked Data in a format that can be incorporated into other Linked Data resources, expanding the scope of file format preservation information available to the community, and that allows NARA to easily reference those resources in the Digital Preservation Framework.
- Wikidata for Digital Preservation is rapidly becoming the hub for all resources that document file formats and related information, so the goal is to ensure as much integration there as possible.

Transitioning to a Semantic Resource

- There are many other authoritative online resources on file formats, official specifications, and reverse-engineered documentation.
    - The Preservation Action Plans already contain URIs which connect to and add to the graph of resources, as well as NARA-researched content to be shared.
- Given the increasing use of Linked Data for descriptive metadata and preservation metadata such as PREMIS, there are several existing sets of elements and properties NARA can use from WikiData and Dublin Core.
- We mapped out the current spreadsheet structure to a data model to clarify the vocabularies, internal data relationships, and relationships to external resources.
- NARA needs its own unique identifier, as well as elements and brief ontologies because we did not find anything in use to document risk.
- We had to review and normalize our data, since it was prepared by a multi-person team over time and had some inconsistencies.

2021-2022 Pilot and Plans

- Mapped the Preservation Action Plan spreadsheet to existing Linked Data elements/properties, and identified NARA-specific elements and controlled vocabularies.
- Converted the Plan from a CSV file to RDF-Turtle
- Shared the draft document with community leaders for feedback on the draft implementation. The feedback included:
  - Express the RDF schema using the ShEx (Shape Expressions) language
  - Identify the language for the linked resources
  - Suggestions about the use of some specific elements
  - Support the use cases to filter on target formats and preservation actions
  - Work toward more granularity and break out individual tools, and software publishers and versions
- The plan is to incorporate a first round of changes and release the linked data version later in 2022, consulting closely with the Wikidata for Digital Preservation team.

Thank You

Leslie Johnston
leslie.johnston@nara.gov