Creating a Holdings Format Profile and Format Risk and Digital Preservation Prioritization Matrix at the National Archives and Records Administration

Leslie Johnston
Digital Preservation
U.S. National Archive (NARA)
College Park, MS, USA
leslie.johnston@nara.gov

Abstract— **One of the greatest challenges for any archive is the multiplicity of file formats, some of which may be decades old. For the United States National Archives and Records Administration (NARA), with several decades of history accessioning and managing electronic records, this is compounded. NARA required a methodology to analyze and visualize what it has in its holdings in order to understand risk, and elected to undertake a file format profile and risk analysis of its born-digital electronic records. This paper will review the process for the creation of an electronic record holdings format profile and the identification of risk assumptions that lead to the development of a format risk analysis and preservation prioritization instrument.**

***Categories and Subject Descriptors*—• Information systems~Digital libraries and archives • Information systems~Integrity checking** *• General and reference~Computing standards, RFCs and guidelines*

***Keywords*— Digital Preservation, File Formats, File Format Characterization, Risk Analysis, Preservation Repository**

I. A MULTIPLICITY OF FILE FORMATS

One of the greatest challenges for any archive is the multiplicity of file formats. For the United States National Archives and Records Administration (NARA), with several decades of history accessioning and managing electronic records, this is compounded. We received our first transfer of electronic records in 1968 and now have over 1.4 billion files. While there is formal guidance for federal agencies in the transfer of their electronic records appraised as being of permanent value as per records schedules under the Federal Records Act, the Presidential Records Act declares all files created by an administration as permanent with few format restrictions. There are no format restrictions for those records that we hold as a courtesy for the Legislative branch of the U.S. government. And even for Federal records, while there are guidelines for, say, email messages, the attachments come over in their original formats.

NARA operates under several different regulatory mandates, each with different restrictions on collection schedules and scope, as well as access controls. This led to the implementation of multiple systems--developed over more than 20 years with different technologies--which meant a real challenge in understanding the scope of the holdings. NARA required a methodology to analyze and visualize what it has in its holdings, and elected to undertake a file format profile of its born-digital electronic records.

II. WHAT IS A COLLECTION PROFILE?

A collection profile is meant to document key information about collections, the preservation requirements, and document the preservation intent for each individual sub-component of the collections as appropriate. As it has been applied in digital preservation, collection profiling is about documenting what content an organization has and what value each collection component has. This goes hand in hand with assessments of file formats, and documenting processing workflows and the required infrastructure inform preservation planning. The aim of collection profiling is to document preservation commitments—also referred to as “Preservation Intent” (Webb, Pearson, and Koerbin, 2013) —given the requirements of the collection and its component file formats and the capabilities of the organization.

To better understand risk, NARA created a File Format Profile of its holdings, an overview of the file formats in the collection to enable the assessment of the sustainability factors and long-term preservation issues through a quantifiable matrix framework, and recommend and document mitigation options for those formats. The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. CREATING THE FILE FORMAT PROFILE

An integral part of NARA's work is the issuance of extensive guidance[1] on all aspects of Federal electronic records management and transfer to NARA, including media types, file formats, and metadata. In developing and maintaining these types of electronic records guidance, the Policy and Standards Team works with NARA custodial units to identify additions or other changes to the specified file formats or metadata elements. The suitability of formats, file transfer problems, and emerging formats or record types are examined to determine where to take further action. When guidance is first drafted or revised, it is made available to all of NARA to ensure that all internal stakeholders can review and react before it is issued to agencies.

The development of this guidance is now complemented by a greater focus inside NARA on digital preservation; NARA issued its first agency-wide digital preservation strategy in 2017.[2] NARA works with agencies as records creators to develop practicable technical guidance informed by internal agency digital preservation needs, and in turn and is developing internal file format preservation plans that align with the guidance issued to agencies, based in part on the work outlined in this paper. Digital preservation is the most successful when it's considered from the very beginning of the lifecycle, at the creation of the records.

Having strong transfer format guidance is essential, but of NARA is not 100% proscriptive in the formats it accepts, hence the word "guidance." 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 NARA negotiates with each agency about what it can provide, but in the end sometimes has to take in the 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. Since NARA has been accepting permanent electronic records since 1968 from over 200 federal agencies, the White House, and Congressional commissions and committees, there is a wide range of versions of formats that have come into the collection over time.

NARA began work on its Holdings Format Profile by compiling data on the file formats in the unclassified electronic records holdings to get a sense of what exists in all of its systems: ERABase for Federal Records, ERA CRI (Congressional Records Instance), the ERA Title 13 (Census) instance, the Presidential Electronic Records Library systems (PERL) for Reagan, Bush 41, and Clinton, and ERA EOP (Executive Office of the President) for the Bush 43 and Obama administrations). Why does NARA have so many systems? The diversity in part comes from the different regulations covering each area of the collections, which stipulate different access levels and requirements for segregation from other holdings, which led to the development of different systems. In part it is the simple passage of time: NARA received its first electronic Presidential records with the end of the Reagan administration. Presidential electronic records are more often than not under strict access controls and are not be default publicly accessible. The business requirements for review for release and response to Freedom of Information Act access requests required the creation of the PERL system. Other systems have come along over time to meet the regulatory variations of business needs, resulting in nine instances of four systems, not including the Classified record instances. These legacy systems will be gradually subsumed into the Electronic Records Archive 2.0 after it goes into production in August 2018.

The Digital Preservation unit worked with the system owners and IT operations to get the most granular reporting possible from each system: federal, legislative, and individual presidential administrations. The reporting didn't always match in terms of granularity, given different tooling for the format analysis and report generation in the different systems. One system employs DROID[3] to characterize formats and reports were provided that listed the formats identified and the level of certainty, but did not include file names or extensions. None of the other legacy systems used DROID or JHOVE.[4] One system could provide a report of all the file names including extensions but no format identifiers. One system provided report that listed only counts per formats with no file names. For one small subset the report supplied an approximate name of a format but did not receive counts. For the reports that included only extensions, the extensions were mapped to a matrix of formats/applications from Wikidata.[5] It is not complete or perfect since without the infrastructure to run a single authoritative tool on all files, but it provides the first overview. For some files in the holdings the extensions are NOT what a program would create, such as .doc versus .2016report. These cannot be mapped via extension without a scanning tool so are temporarily "unknown" in the profile. There were also different granularity levels reported for file formats, e.g., files identified as Adobe Acrobat PDF vs. files identified specifically as Adobe Acrobat PDF 1.4. This required some normalization when aggregating the data together to compare across the holdings.

The data was then loaded into Tableau,[6] a Business Intelligence system, for review and analysis. Tableau was selected because it had the capacity to load data representing 1.5 billion files, supported SQL queries to create multiple alternate views of the data, and could be used to create clean visualizations of all or part of the holdings.

[1] https://www.archives.gov/records-mgmt/bulletins

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

[3] http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

[4] http://jhove.openpreservation.org/

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

[6] https://www.tableau.com/

The first outcome of this work was the identification of the file formats make up the bulk of the holdings (Table I) and the proportion of the formats in the holdings (Figure 1).

TABLE I: 10 Most Common File Formats in NARA Holdings

Electronic Mail Message file
JPEG bitmap graphics file
Tagged Image File Format
HyperText Markup Language document
ASCII 8-bit Text
Extensible Markup Language file
Adobe Acrobat PDF file
Document (.doc) file format
RAW Image file
Microsoft Word Open XML Document

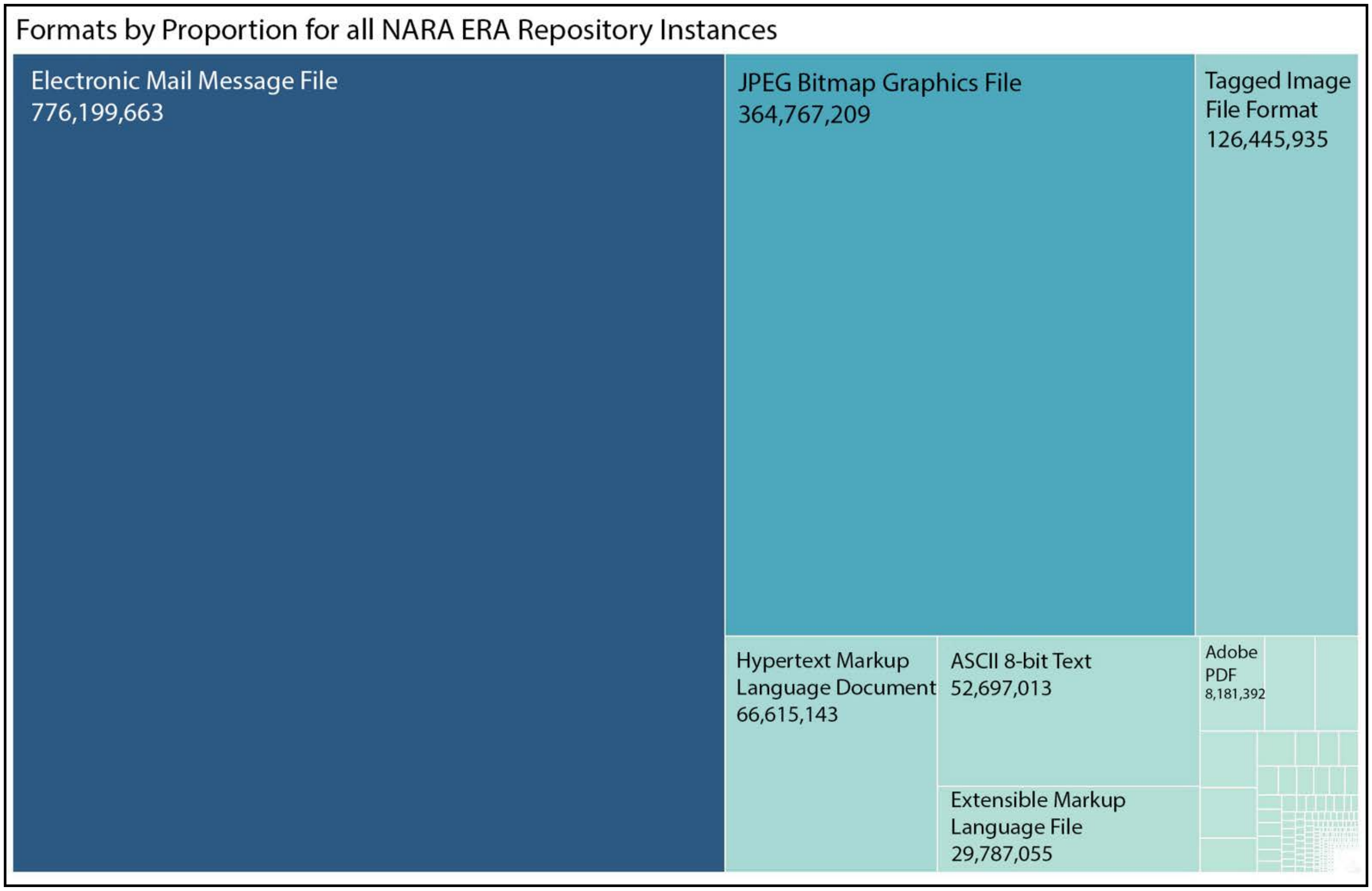

FIGURE 1: Visualization of the Relative Percentages of Formats in NARA Electronic Records Holdings

A percentage of the holdings could not be characterized and mapped to documented formats with certainty by DROID or JHOVE, which is expected. A surprising number of software companies have used the same file extensions over time, which was important for the subset of the holdings where DROID or JHOVE are not in place for format characterization and there are 6 or more possible formats that could be identified based on file extension alone. There are many variations of certain formats that NARA has acquired over the decades, PDF being the most diverse: 8.2 million files in 18 different version variants. NARA has over 700 million email MSG files -- more than anything else -- due to the large numbers of emails received from Presidential administrations. NARA has over 500 million image files - JPEG, TIFF, RAW, GIF, and BMP - across every set of Federal, Presidential, and Legislative records. NARA has tens of millions of ASCII textual records in almost every system. There are databases, GIS data, video, audio, XML, and HTML. And of course documents and spreadsheets and slide decks. And there is a lot more: in some cases there are only a handful of files per format versus the tens or hundreds of millions for other formats.

Several hundred file formats are present in the holdings if one counts all the variations of PDF or Microsoft Word, for example. NARA will have a more granular picture as holdings are migrated into a single environment and ingest processes will run the same format characterization tools across all the files. The NARA Digital Preservation Group is actively working on approaches for identifying format risks and mitigation strategies.

Not every file format could be identified with complete certainty. 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. Developing this complete understanding of the formats, including what we do NOT know about then is informing a plan for the preservation program and the necessary priorities for technology updates.

IV. CREATING THE RISK AND PRIORITIZATION MATRIX

Creating a quantified framework for calculating risk factors has its proponents (Rog and van Wijk; Graf. and Gordea, 2013; Becker, Faria, and Duretec, 2014) and its skeptics (van der Knijff, 2013). That said, NARA has both a long internal history of generating longitudinal statistics to measure its growth and capabilities and a culture of risk identification and measurement, so chose to develop a Risk Matrix as part of this work.

In preparation for the issuance of the 2014 bulletin describing preferred and acceptable file formats for permanent records (NARA, 2014), NARA created a quantified Transfer Format Suitability Matrix with 37 data points on the sustainability of possible formats, arranged in several categories: Disclosure, Adoption Level/Viability, Transparency, Self-Documentation, External Dependencies, Licensing and Patents, and Use of Encryption/Rights Management with varied weighting for each question and category. This work can be directly compared to that of the Library of Congress in its extensive format sustainability analysis,[7] with the addition of the weighted rankings. This matrix assisted the guidance development team in identifying and ranking formats as "Preferred" or "Acceptable" for permanent electronic record transfer to NARA.

While this matrix measured the suitability of file formats for transfer to NARA, it did not measure ongoing risk for those formats in the holdings going forward. To accomplish this to the best of its abilities, the matrix was extended with risk data points on the percentage that a format makes up the holdings, the age of the format (when it was introduced), and the currency of the format (when it was most recently updated). There was extensive discussion of other potential data points – the age of the file(s) in the repository, some sort of measure of risk inherent in format transformations (how much loss will there be in a format transformation). The former was not reported in a consistent way across the sets of files in the holdings so as to be usable in such an analysis, and it was determined that the latter was extraordinarily difficult to quantify given the sheer number of variables, from format version to operating environment to the tool used for the transformation. The identification of appropriate weighting for the final factors in the matrix also required a lengthy period of revision and review. Categories and questions have different weights related to their impact (positive or negative) on sustainability of a format, and therefore its risk level.

Several assumptions were identified to inform the appropriate weightings:

- The openness of a format and availability of full documentation--which enables the development of tools to work with that format and/or to perform preservation format transformations--provides a higher positive effect than the lack of openness and absence of 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 technical characteristics which can be mined for preservation purposes and can have descriptive metadata embedded in its file header, provides a higher positive impact than negative. All files can provide some basics technical information, but not all can have descriptive metadata embedded through its creation process or added by an external tool, so all file formats are self-documenting to some degree.
- The requirement to maintain specific software (or, in some cases, hardware) for the ingest, processing,

[7] http://www.loc.gov/preservation/digital/formats/

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.

After identifying these assumptions, the weights for the factors were adjusted across all the categories to take into account factors where the highest level of effort or, to the ability we could determine it, cost, such as needs for software or hardware. A proof of concept was applied to all formats in the 2014 Transfer Guidance and all formats in the holdings with one million files or more, or seventy formats. This identified 2 high risk, 26 moderate risk, and 42 low risk formats (Table II).

TABLE II: 10 Highest Risk Formats Based on an Analysis of a Subset of 70 Formats from NARA Holdings

High Risk	Camera RAW file	Digital Still Image
High Risk	WordPerfect versions 6-12	Textual Data
Moderate Risk	Advanced Systems Format	Digital Video
Moderate Risk	ESRI Shapefile (Compound)	Geospatial
Moderate Risk	ESRI ESRI ARC/INFO Interchange File Format	Geospatial
Moderate Risk	Microsoft Word Office	Textual Data
Moderate Risk	Vector Product Format	Geospatial
Moderate Risk	Windows Media Video 9 File Format	Digital Video
Moderate Risk	TerraGo Geospatial PDF	Geospatial
Moderate Risk	QTA AAC, QuickTime file with AAC Encoding	Digital Audio

After completion of the proof-of-concept analysis, the framework was further adjusted to split concerns into Risk versus Prioritization. The traditional Need-Use-Value metric that is often used in analog preservation was considered, but none of the usage statistics collected for _items_ could be translated to formats, and the same was true for the relative value of sets of records in the holdings which could comprise multiple formats. Instead we chose to use Need (the Risk value identified in the Risk Matrix); Prevalence (Use as defined by the prevalence of the format in the records created by agencies and transferred to NARA); and Feasibility (the current capabilities at NARA to perform transformations, and if none, the availability of tools to begin doing so). This instrument (see Appendix A) is now in use and work began in Summer 2018 to review the remainder of the formats in the holdings to generate a more complete and informed picture of both risk and prioritization for preservation actions.

IV. NEXT STEPS

The next phase in this work is the creation of File Format Action Plans, documented format preservation recommendations based on a utility analysis of the risk factors identified through the risk matrix. (Becker et.al, 2009; Becker and Rauber, 2011; Stanescu, 2005) Each plan links to format documentation, identifies essential characteristics, aka significant properties to be preserved from the format (Wilson, 2007; Brown 2008), identifies the relevant preferred and acceptable formats from the NARA Transfer Guidance, the relevant internal NARA reference format, the outcome of the risk matrix analysis, the recommended preservation outcomes (transform records to new formats, procure/develop tools that enhance or extend NARA’s capability to manage records in that format, or to explore additional options), preferred normalization/transformation tool(s), and available viewer(s).

What is still missing from this picture is automation. Each of these processes, from the analysis to the documentation and the monitoring, are almost entirely manual at this time. There are reports and scripts to aid in the maintenance of the Holdings profile. While the Matrix is a weighted instrument to produce an informed risk score based on several dozen factors, the answers to the questions about the formats and holdings must be researched and input by expert archivists. The same archivists are responsible for the identification of recommended preservation decisions and documentation of those decisions. And at this point, there is not yet a way to automate the monitoring of the holdings to identify triggers or the tooling to take wide scale preservation actions on hundreds of format variants.

NARA is working toward those goals, preparing to put a major new release of its Electronic Records Archives (ERA) repository into production in August 2018 (Johnston, 2017), a complete update of its original preservation repository (Thibodeau, 2009). Data compilation and reporting will become easier as NARA consolidates its files into the new environment with more robust format characterization tooling that will for the first time be able to report on and monitor the entirety of the holdings. Even the partial documentation created so far on our format risks and plan for format migrations has informed and changed the prioritization for acquiring tools to view, process, and migrate those types of records in the updated system framework. The full holdings will not all be in the system on day one of production, nor will every potential tool be in place, but every step closer to such a consolidated environment will aid in the replicable, more reliable automation of preservation analysis and actions.

This approach is also changing ways in which NARA plans for access. There has always been an explicit link between the preferred formats for preservation and those for access. NARA has a roadmap for the evolution of its systems, including the National Archives Catalog; analysis of the trends in the growth and formats in the holdings will inform the planning and prioritization for new formats to be displayed and delivered to the public.

IV. CONCLUSIONS

Digital preservation is inherently about risk mitigation, which cannot succeed without extensive and transparent documentation of both risks and decisions. The National Archives is extending its well-developed practices of risk-based preservation decision making (Kaplan and Banks, 1990) to its digital preservation operations with these new modes of documentation. The goal is to not only create more reliable processes, but to make the data, processes, and decisions more transparent internally and externally, making digital preservation at NARA more concrete for the staff, the rest of the federal government, and the public.

In implementing the mechanism to create and maintain a full holdings format profile, a risk matrix to quantify format sustainability risks, and documenting preservation action recommendations in file format action plans, NARA is putting in place a framework that enables scalable preservation and monitoring and supports agency trust in its preservation decisions because the evidence for that decision making is documented in a consistent, quantifiable manner.

REFERENCES

[1] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International journal on digital libraries, 10(4), 133-157.

[2] Becker, C., & Rauber, A. (2011). Decision criteria in digital preservation: What to measure and how. Journal of the Association for Information Science and Technology, 62(6), 1009-1028.

[3] Christoph Becker, Luis Faria, Kresimir Duretec, (2014) "Scalable decision support for digital preservation", OCLC Systems & Services: International digital library perspectives, Vol. 30 Issue: 4, pp.249-284, https://doi.org/10.1108/OCLC-06-2014-0025

[4] Brown, A. (2008). Developing practical approaches to active preservation. International Journal of Digital Curation, 2(1).

[5] Graf, R., & Gordea, S. (2013, September). A risk analysis of file formats for preservation planning. In Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres2013) (pp. 177-186).

[6] Johnston, L. (2017). ERA 2.0: The national archives new framework for electronic records preservation. Proceedings of the Association for Information Science and Technology, 54(1), 197-202.

[7] Kaplan, H., & Banks, B. (1990). Archival preservation: the teaming of the crew. The American Archivist, 53(2), 266-273.

[8] National Archives and Records Administration (2014). Bulletin 2014-04: Revised Format Guidance for the Transfer of Permanent Electronic Records. https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html

[9] Rog, J., & Van Wijk, C. (2008). Evaluating file formats for long-term preservation. Data Analysis and Knowledge Discovery, 24(1), 83-90.

[10] Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. OCLC Systems & Services: International digital library perspectives, 21(1), 61-81.

[11] Thibodeau, K. (2009, November). Preserving digital memory at the National Archives and Records Administration of the US. In Workshop on conservation of digital memories. Second national conference on archives, Bologna, Italy (pp. 1-9).

[12] van der Knijff, J. (2013). Assessing file format risks: searching for Bigfoot?. Open Planets Foundation blog, September 30, 2013. http://openpreservation.org/blog/2013/09/30/assessing-file-format-risks-searching-bigfoot/

[13] Webb, C., Pearson, D., & Koerbin, P. (2013). 'Oh, you wanted us to preserve that?!': Statements of Preservation Intent for the National Library of Australia's Digital Collections. D-Lib Magazine, 19(1), 2.

[14] Wilson, A. (2007). Significant properties report. InSPECT Work Package, 2.

Appendix A: NARA File Format Risk and Preservation Prioritization Matrix

Format Name				Disclosure						Adoption				
File Format Identifiers				Disclosure refers to the degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content.						Adoption refers to the degree to which the format is already used by the primar disseminators, or users of information resources. This includes use as a master form to end users, and as a means of interchange between systems.				
Short Name	Long Name	File Extension(s)	Content Category	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	Is the file format commonly used to create or maintain permanent records within the federal government?	Is the file format commonly used outside the federal government?	Is the format actively maintained and updated by an organization, individual, or community?	Are multiple renderers available?	Have the archives or library communities identified the format as one they prefer for creation and transfer of permanent materials?
				-1 = Yes, 0 = N/A or Unknown, 2 = No	2 = Yes, 0 = N/A or Unknown, -1 = No	2 = Yes, 0 = N/A or Unknown, -1 = No	2 = Yes, 0 = N/A or Unknown, -2 = No	2 = Yes, 0 = N/A or Unknown, -1 = No	Sum of Columns I-M					
									0.00					

	Transparency								Self-Documentation						External Hardware Dependencies		
y creators, at, for delivery	Transparency refers to the degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor.								Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.						External dependencies refers to the degree to which a format depends on particular hardware for processing and the predicted complexity of dealing with those dependencies in future technical environments.		
Total Adoption Score: Highest possible score = 10; Lowest possible score = -6	Is the format human readable and can it be opened with a text editor?	Does the available specification provide enough detail to allow basic analysis of the file format with hex editors or other inspection tools?	Does the format rely on standard character or other encoding methods such as IEEE notations?	Is the source code of the software used to create the format available for little or no cost?	Is the software used to create the format supported by current computing environments?	Does the format require the use of compression?	Does the format support user-definable compression levels or other quality settings that affect essential characteristics?	Total Transparency Score. Highest possible score = 7; Lowest possible score = -7	Does the format support self-contained (embedded) descriptive metadata?	Does the format support self-contained technical metadata?	Does the format support self-contained administrative metadata?	Does the format metadata adhere to an international standard?	Is the format metadata robust enough for an accurate file analysis?	Total Self-Documentation Score. Highest possible score = 7; Lowest possible score = -5	Does the format require a specific hardware environment, such as a specific graphics card, chipset, or memory requirements, to process or interact with it?	Does the format require specific playback hardware (e.g., Blu-Ray, Audio CD, etc) to transfer the format to the NARA environment?	Total External Software Dependencies Score. Highest possible score = 4; Lowest possible score = -8
Sum of Columns O-S	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	-1 = Yes, 0 = N/A or Unknown, 1 = No	-1 = Yes, 0 = N/A or Unknown, 1 = No	Sum of Columns U-AA	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	2 = Yes, 0 = N/A or Unknown, -1 = No	2 = Yes, 0 = N/A or Unknown, -1 = No	Sum of Columns AC-AG	-4 = Yes, 0 = N/A or Unknown, 2 = No	-4 = Yes, 0 = N/A or Unknown, 2 = No	Sum of Columns AI-AJ
0								0						0			0

External Software Dependencies					Impact of Patents and Licenses					Technical Protection Mechanisms							
External dependencies refers to the degree to which a format depends on a particular operating system or software for processing or rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.					Degree to which the ability of archival institutions to sustain content in a format will be inhibited by licenses or patents.					Implementation of mechanisms such as encryption that negatively impact and prevent the preservation of content by a trusted repository.						Age of the Format and Curre	
Does the format rely on proprietary software to render or view files?	Does the format rely on plug-ins, scripts, etc. to render or view files?	Does the format rely on specific computing operating system(s) to render or view files?	Does the format rely on special tools to render or view files?	Total External Software Dependencies Score. Highest possible score = 4; Lowest possible score = -8	Is the format subject to patent claims that may impede the development of open source tools for opening and managing the files?	Have the patent claims expired?	Are there fees associated with the format as a result of patent claims?	Does the format have open source license terms?	Total Impact of Patents Score. Highest possible score = 4; Lowest possible score = -4	Does the format have capability to encrypt all or part of the resulting file?	Does the format require the use of encryption?	Does the format support robust encryption?	Can technical protection measures (e.g. digital rights management) be applied?	Does the format allow for embedded information such as watermarking?	Total Technical Protection Mechanisms Score. Highest possible score = 10; Lowest possible score = -6	When was the Format Specification First Created?	Risk Factor: Highest Possible score = 0; Lowest possible score = -4
-2 = Yes, 0 = N/A or Unknown, 1 = No	-2 = Yes, 0 = N/A or Unknown, 1 = No	-2 = Yes, 0 = N/A or Unknown, 1 = No	-2 = Yes, 0 = N/A or Unknown, 1 = No	Sum of Columns AL-AO	-1 = Yes, 0 = N/A or Unknown, 1 = No	-1 = Yes, 0 = Unknown, 1 = No or N/A	-1 = Yes, 0 = N/A or Unknown, 1 = No	1 = Yes, 0 = N/A or Unknown, -1 = No	Sum of Columns AQ-AT	-1= Yes, 0 = N/A or Unknown, 2 = No	-2=Yes, 0=N/A or Unknown, 2 = No	-1= Yes, 0 = N/A or Unknown, 2 = No	-1= Yes, 0 = N/A or Unknown, 2 = No	-1= Yes, 0 = N/A or Unknown, 2 = No	Sum of Columns AV-AZ	Year	0=5 years old or less, -2 for 6-15 years, -4 for 16+ years
				0					0						0	2018	0

Format Age			Summary		
ncy of the Format Specification as Risk Factors			Risk/Sustainability at a Glance		Notes
When was the Format Specification last updated?	Risk Factor: Highest Possible score = 0; Lowest possible score = -4	Total Format Age Score. Highest possible score = 0; Lowest possible score = -8	Total Format Risk/Sustainability Factor Score	Risk/Sustainability Factor Status	
Year	0=5 years old or less, -2 for 6-15 years, -4 for 16+years	Sum of Columns BB-BE		Low Risk = 26-52; Moderate Risk = 1-25; High Risk = -53-0	
2018	0	0	0.00	High Risk	

FILE FORMAT PRESERVATION ACTION PRIORITIZATION MATRIX												
Format Name				Need: Need for Preservation Actions as measured by the Risk Matrix Value		Prevalence: Format Adoption Level as measured by Proportion of File Format in the Overall NARA Holdings			Feasibility: Ability to Convert (tools exist for conversion that does not alter content in unacceptable ways; NARA can perform acceptable transformations)	Transfer Guidance Status		TOTAL
File Format Identifiers				Assigned Risk Level	Risk Matrix Value	Count	Percent of holdings	Highest possible score=-5, Lowest possible score=-15	Highest possible score=5, Lowest possible score =-5			
Short Name	Long Name	File Extension(s)	Content Category		Number calculated in the Risk Matrix	Number of identified files in the combined NARA holdings	Percentage of 1.5 billion total files in the NARA holdings	-5 for 0-2%, -6 for 3-4%, -7 for 5-6%, -8 for 7-8%, -9 for 9-10%, -10 for 11-12%, -11 for 13-14%, -12 for 15-16%, -13 for 17-18%, -14 for 18-20%; -15 for over 20%	No acceptable tools available in the marketplace=-5; Acceptable tools exist but NARA does not have them=-3; Migration already performed at NARA=3; Preferred/Acceptable Format as per Guidance or practice where no migration is needed=5;	Preferred	Acceptable	The lower the number, the higher priority for preservation actions based on risk level, percentage in the holdings, and the requirement/ability to perform format migrations.
				High Risk				-5				-5