SOP for NARA Digital Preservation Framework

Table of Contents

SOP Revision and Review History

Version	Date Created/Updated	Who	Revision Description
0.1	10/22/2018	Leslie Johnston	Initial SOP Draft
1.0	2/15/2019	Criss Austin, Meg Guthorn, Leslie Johnston, Jana Leighton, Andrea Riley	Revised after review
1.1	6/25/2020	Elizabeth England	Revised in response to updating formatting and content for the Preservation Action Plans and File Format Matrix
1.2	4/26/2021	Elizabeth England	Revised threshold number of files in the holdings from 1,000 to 2,000 for inclusion in the Framework; expanded file management; updated terminology, template, SOP version numbering
1.3	10/7/2021	Elizabeth England	Revised SOP title; added information about dates for file formats in risk matrix; changed separator between multiple values from being a semicolon to a pipe
1.4	1/5/2022	Elizabeth England	Revised Proposed Preservation Actions Section to reflect updated controlled lists.

SOP Purpose Statement and Scope

Having documented decisions that guide digital preservation operations is a vital activity. A Preservation Action Plan is a document used to assess risks and lay out the steps to mitigate the risks to best preserve the holdings in question. A Plan must identify risks, prioritization, proposed preservation actions, necessary resources, and the steps required. NARA has developed a Digital Preservation Framework for documents to serve this need, documenting the risk associated with file formats in its File Format Matrix, and recording the decisions made about the preservation of file formats in its File Format Preservation Action Plans. This SOP covers the creation, review, and updating of both the Matrix and the Preservation Action Plans.

When does this SOP take effect?

- This guidance takes effect as of **June 29, 2020**.
- The SOP consists of **13** pages.
- This SOP about digital preservation supersedes and replaces version 1.3.

Terms Used

NARA Acronyms and Terms

Acronym	**Definition**
R	Research Services. NARA's Office of Research Services is the custodian for federal archival records that are stored, managed, and made available at several locations across the country, which provide services for the public to discover, locate, and use the records.
L	Legislative Archives, Presidential Libraries and Museum Services

Non-NARA Acronyms and Terms

Acronym	**Definition**
Preservation Action Plan	A Preservation Action Plan is a document used to assess risks and lay out the steps to mitigate the risks to best preserve the holdings in question. A Plan must identify risks, prioritization, proposed preservation actions, necessary resources, and the steps required.

Infrastructure/Equipment

Computer Hardware, Software

Tool	**Purpose**	**NARA IT Supported**	**Approved for NARAnet**	**Purchase Yr.**
Google Sheets	File Format Matrix; Preservation Action Plans for File Formats; Control list for NARA IDs; Matrix Weights	Yes	Yes	

Google Docs	Preservation Action Plans for Categories	Yes	Yes	

Other Equipment and Supplies

Tool	**Purpose**	**NARA Supported**	**Purchase Yr.**
N/A			

Methodology

File Format Matrix

Risk

In the File Format Matrix you will answer questions about sustainability factors that impact the ability to preserve those formats. The sustainability factors fall into nine categories, each of which is weighted differently as it relates to the level of risk/sustainability and, to the extent that it can be identified to be taken into account, cost. Categories and questions have different weights related to their impact (positive or negative) on sustainability of a format and therefore its risk level. The weight assigned to each factor is documented in a weights spreadsheet, available in Drive.

- Highest Impact Categories
    - Positive: Disclosure, Adoption
    - Negative: Hardware Dependency, Software Dependency, Age of Format
- Highest Impact Questions
    - Positive: +2 for several questions in Disclosure, Adoption, Self-documentation, Software Dependency, and Technical Protection Mechanism categories.
    - Negative: -4 for Format Age and Hardware Dependency questions.

Instructions:

1. Create a Row for the file format. Each File Format variation, such as a unique release/version, must have its own row in the sheet.
    a. The general guideline is to create an entry for any file format variation for which the count is 2,000 or higher. There are many formats for which the variant/version information is currently unknown. In these cases, if the count is more than 20,000 files identified in the holdings, create an “unspecified version” entry (see JPEG for example). The count information is gathered from the Holdings Profile.
2. *Guidance Rating*: Put an X in the appropriate column if the format is on the NARA Preferred or Acceptable Transfer Guidance list
3. *Overall Risk Rating*: This will be auto-completed at the end of data entry.
4. File Format Identifiers:
    - Provide the file format name, including applicable Publisher and version/date information, e.g., Microsoft Access 97, or Adobe InDesign Document CS.

- List all applicable file extensions. A minimum of one is required although several file extensions may be associated if it is a compound object or container files such as a GIS object. The same file extension may be repeated on different rows for multiple versions across time. Separate multiple extensions with pipes | without spaces around the pipe. *Example: xls|xlsx*
- Enter the Record Type/Plan(s) for which the file format belongs. Separate multiple record types with pipes | without spaces around the pipe.

5. *Sustainability Factors*: For each of the following categories, research the format and enter the value that best addresses the sustainability of the format. Do not overthink or over-research this; the Matrix is not meant to be exhaustive and perfect. This is a living document that will be updated over time.
    - **Disclosure** *The degree to which complete specifications and tools for validating technical integrity exist and are accessible to those creating and sustaining digital content.*
    - **Adoption** *The degree to which the format is already used by the primary creators, disseminators, or users of information resources. This includes use as a primary or "master" format, for delivery to end users, and as a means of interchange between systems.*
    - **Transparency** *The degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor.*
    - **Self-Documentation** *Self-documenting digital objects contain basic descriptive, technical, and other administrative metadata.*
    - **External Dependencies** *External dependencies refers to the degree to which a particular format depends on particular hardware, operating system, or software for rendering or use and the predicted complexity of dealing with those dependencies in future technical environments.*
    - **Licensing and Patents** *Degree to which the ability of archival institutions to sustain content in a format will be inhibited by licenses or patents.*
    - **Encryption and Rights Management** *Implementation of mechanisms such as encryption that negatively impact and prevent the preservation of content by a trusted repository.*
    - **Format Age** *As formats age, the potential risk increases. Risk factors related to the age of the formats and the time since their specifications were reviewed or updated.*

    - **Additional guidance:** When answering Risk questions, the following guidance may be of use. If a topic is not mentioned here, use your best judgement when filling out the matrix. It is acceptable and unavoidable to have factors for which the value is "0" because there is no applicable answer or the answer cannot be determined. This is especially true for proprietary formats for which there is no open specification or other resource to confirm information, specially the status of patent/licensing issues. The Licensing/Patents and Encryption/Rights Management sections are the most likely to have "0" values. You may have more "0" values for unspecified version entries; however

the final risk scores of unspecified versions should always be equal to or greater than the highest risk variant of the format. Ensure the scoring reflects this.

- *What is a "published open specification"?* An open specification is a complete technical format specification created and controlled, in an open and fair process, by an association or a standardization body intending to achieve interoperability and interchangeability; rights, licenses, and patents associated with the specification must be clearly communicated.
- *How do you know if there are available tools that can validate the technical integrity of a file encoded in this format against the published specification?* Check the PRONOM database, JHOVE, and JHOVE2 to determine if the format is listed and includes a format signature, which is necessary for validation.
- *How do you know whether the specification has been approved and published by an internationally recognized standards body?* If the standard comes from NISO, ISO, ANSI, WC3, IANA, another national or international standards organization, or an organization dedicated to a specific format, such as the PDF Association, it meets the criteria.
- *How do you determine whether the specification is complete and accurate?* Complete specifications should include information about file extensions, header structure, the magic number, data structure including byte patterns, extended attributes, and required or optional external linkages. File format Identifiers (FFIDs) are relatively new and may not be assigned/included. Accuracy is difficult to judge, but if you find other format-related resources citing the specification, it is likely to be complete and authoritative.
- *What constitutes "common use" in the federal government?* The answer is yes if NARA has added more than 2,000 files of this type to its holdings. This is a somewhat arbitrary threshold, the same that we have used to identify statistically significant formats for risk assessment. Prior to April 2021, NARA used 1,000 files as the threshold, however due to the rate of growth of the holdings, this was revised to 2,000. Other formats not meeting that threshold can still be added on an as needed, or requested, basis.
- *What constitutes "multiple renderers"?* There should be 3 or more commercial, freeware, or open source tools that can render files of this type.
- *How do you know if the archives or library communities identify the format as one they prefer for creation and transfer of permanent materials?* Confirm that the format is on the NARA Transfer Guidance, the Library of Congress Recommended Formats Statement, or guidance from Libraries and Archives Canada, The National Archives UK, or the National Archives of Australia.
- *How can you determine whether the format relies on standard character or other encoding methods such as IEEE notation?* Hopefully any technical specification for a format or preservation description of the format will include this information. This may not always be discoverable.
- *How do you determine whether the software used to create the format is supported by current computing environments?* Documentation of software generally includes information about the environment(s) in which it runs, e.g. Win95, MacOS7, Windows 7/10, etc. An environment may be considered current, e.g. still commercially supported by the original company or by a 3rd party, but a separate issue is whether it is available at NARA for use in processing. Determining the availability of an environment will often require

consultation with NARA IT on an individual facility and unit basis to determine whether we have matching computing capabilities and support, and/or whether we can obtain appropriate software/hardware and support.

- *What is considered "descriptive metadata"?* Metadata that describes a record for purposes such as discovery and identification, including agency name, record series, names, places and dates. Many formats allow some level of descriptive metadata to be embedded in the file headers.
- *What is considered "technical metadata"?* Technical metadata documents the technical environment in which a file was created and used, including hardware, software, operating systems, and specifications for the file (pixel height/width, encoding codec, etc). This is generally automatically written into the file header when it is created and when it is updated.
- *What is considered "administrative metadata"?* Administrative metadata has some commonalities with descriptive metadata--such as authority for the transfer/accessioning and the provenance of the record--that is used to manage the files, but also includes metadata related to rights and restrictions.
- *How robust should the metadata be for an accurate file analysis?* This varies by record type. The definition of robustness is the ability to determine if the file is the format it purports to be, that it is well-formed, and can be rendered or transformed into a renderable format, and that there is sufficient descriptive metadata to make the records accessible.
- *Does the format require a specific hardware environment, such as a specific graphics card, chipset, or memory requirements, to process or interact with it?* This applies most often to special media and sometimes digital still images, as some formats are specific to the carrier media they are stored on.
- *Does the format rely on specific computing operating system(s) to render or view files?* Some formats are limited to the MacOS, Windows, DOS, or Linux/Unix environments because there is no software that supports the format available for other environments.
- *How do you know about patent claims, their status, and whether there are any associated fees?* Unless a format has its own web site or page documenting its history, the best resources to research this are the Library of Congress Format Sustainability site, Wikipedia, Wikidata, and the ArchiveTeam file format site. Answers to this question may not be easily discovered.
- *What is considered "robust encryption"?* Robust encryption involves mathematically calculated cipher/key pairs that are used to lock and unlock files, as opposed to simple password locking of files. Encryption is most commonly used for sensitive or classified records, to protect PII (such as user accounts or personnel data), or intellectual property (software code). Some formats natively support encryption of their file contents. Some encrypted formats are container formats that can hold other types of files.
- The equivalent for "human readable" for non-text files is that we must be able to run commonly available tools (ubiquitous?) that identify the format and recognize it as a format that can be rendered/played.

- **Resources for researching formats:**

- NARA Table of Preferred and Acceptable File Formats: https://www.archives.gov/records-mgmt/policy/transfer-guidance-tables.html
- Library of Congress (LOC) Sustainability of Formats site: https://www.loc.gov/preservation/digital/formats/
- British National Archives PRONOM database: http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=new
- ArchiveTeam File Format site: http://fileformats.archiveteam.org/wiki/File_Formats
- Wikidata File Format project: https://www.wikidata.org/wiki/Wikidata:WikiProject_Informatics/File_formats/Lists/File_formats
- ForensicsWiki file formats: https://forensicswiki.xyz/wiki/index.php?title=Category:File_Formats
- Fileextensions.org (be cautious, this has lots of links to downloads/shareware, etc): https://www.file-extensions.org / https://www.file-extensions.org/extensions/common-file-extension-list
- dotwhat (same caveat as fileextensions.org): http://dotwhat.net/
- Wikipedia: http://wikipedia.org/
- British Library File Formats Assessments hosted by the Digital Preservation Coalition: https://wiki.dpconline.org/index.php?title=File_Formats_Assessments
- File Format Wiki wiki.fileformat.com

5. Summary Risk/Sustainability Factor Numeric Score: This will be auto-completed after all numeric values are assigned.

Prioritization

Next you will answer questions about NARA-specific factors that impact the prioritization for preservation actions.

Instructions:

1. Prevalence: Enter the count for this format currently identified in the NARA holdings across all NARA preservation systems.
    - If you cannot find the count, ask the Director of Digital Preservation to track it down for you.
    - The percentage in the holdings will be automatically calculated from the count.
    - The Prevalence column is automatically calculated from the percentage.
    - Note that the raw counts are removed before making the Matrix publicly available. To do this, after finalizing the internal Matrix and saving a copy indicating it's the internal version, create an external-facing version. Copy the percentages column and **paste values only** into a new column, then delete the raw count and original percentages columns.
2. Feasibility: The answer to this question comes from the applied experiences of processing archivists in R and L, who are the best resources. Digital Preservation can assist in research on tools in the marketplace that have been identified or are in use at NARA.
    - No acceptable tools available in the marketplace = -5
    - Acceptable tools exist but NARA does not have them = -3

- Transformation already performed at NARA = 3
- Preferred/Acceptable Format as per Transfer Guidance or no transformation is needed = 5

Internal NARA Risk and Prioritization Total: The numeric rating will be automatically calculated.

Preservation Action Plans: Categories

The granularity for Plans is at the Category level, not the file format level, because significant properties of electronic records are shared by a category of records, such as word processing files or email, not to specific formats such as Microsoft Word 97 or EML.

Document Title

Identify the Category in the document Title, with the document date and the date for the version of the Template that was used. Example:

Preservation Action Plan: Web Records

Approved: 09122018
Using Template 201808

Electronic Record/Surrogate Type Section

Provide a brief description of the Category, no more than two brief paragraphs.

Significant Properties Section

Provide brief context--no more than 3-4 short paragraphs--for the Significant Properties, including an explanation of which Properties are the most significant for the Category (Record Type).

Identify the Significant Properties for each category. It is expected that some record type categories will have no significant properties in one or more categories.

- **Appearance** *Properties related to the visible appearance of this record type, which contributes to the re-creation of the record content and are required to convey meaning, as well as properties of the content of this record type which must be conveyed in any migration. Example: font type, appearance features, color and size, and bit depth.*
- **Structure** *Properties that are required to retain the structure of the information contained in this record type, such as information that describes the relationship between two or more types of content, as required to reconstruct its performance. It may be applied to the intrinsic or extrinsic relationships contained in the performance.*
  Ex: logical properties, duration, character count, external file relationships (such as email with attachment or threaded messages, etc.
- **Behavior** *Properties related to user interaction with this record type, which may include the interaction of the user with the software, or interaction with other sources of information, such as an external resource that affects the content, context, structure, or appearance of the resources. Behavior is a difficult property to preserve, as it is often tied to the capabilities of a particular software application and may be difficult to translate.*

- **Context** *Properties required to provide context about this record type or its relationship to other records of the same or different types, such as the environment in which the Context was created or that affects its intended meaning. Ex: Creator name, date of creation, description of the intellectual work, computer environment in which the record was created.*

NARA Transfer Guidance Section

Document the current Preferred and Acceptable file format(s) for this Category for transfer from agencies to NARA.

NARA Public Access and Reference Section

Formats for Public Access are those made available online through the National Archives Catalog. Formats for Reference are defined as those made available to researchers upon direct requests for digital copies.

Comments and Notes Section

Enter any appropriate notes and comments about sources for significant properties and findings.

Preservation Action Plans: File Formats

The columns are as follows:

File Format Identifiers Section

- **File Format Name** *The name of the specific individual format being analyzed. This should be at the same level of granularity as the Matrix.*
- **Extension(s)** *Previously described in the Matrix section.*
- **Category/Plan(s)** *Previously described in the Matrix section.*
- **NARA Format ID** *Previously described in the Matrix section.*
- **MIME type(s)** *Enter the relevant MIME type(s) for the format. You can include unofficial ones if no registered MIME types exist.*

Links Section

It is not expected that each of these fields can be filled out; in many cases a relevant link doesn't currently exist.

- **Specification/Standard URL** *Include a link to the file format specification/standard. If one cannot be located, you can include a reverse-engineered specification, if that can be located.*
- **PRONOM URL** *Include a link to the relevant PRONOM entry. Note this link should end in the PRONOM Persistent Unique Identifier (PUID), and may be different than what appears in your address bar, depending on how you navigated to the entry. If a PUID has been deprecated, include the link to the referenced PUID.*
- **Library of Congress URL** *Include a link to the relevant LC Sustainability of Digital Formats entry.*

- **British Library URL** *Include a link to the relevant British Library PDF, linked from the Digital Preservation Coalition's File Formats Assessments.*
- **WikiData URL** *Include a link to the relevant WikiData entry. If one cannot be found for the individual format, you can include a link to the format family entry if that exists.*
- **ArchiveTeam URL** *Include a link to the relevant ArchiveTeam entry. These are often for the format family, and not individual versions.*
- **ForensicsWiki URL** *Include a link to the relevant ForensicsWiki entry.*
- **Wikipedia URL** *Include a link to the relevant Wikipedia entry. In some cases, this may be for software associated with the file format.*
- **File Format Wiki (wiki.fileformat.com)** *This is optional, and may be more useful for some Record Types over others.*
- **Other URL** *This is optional, if there is another URL that fills a gap in understanding the format being described. This is useful for formats that do not have many other links.*

Proposed Preservation Actions Section

- **Notes** *This is optional, if there is additional information you want to provide that isn't captured elsewhere. For any unspecified version entry, include an explanatory note. See JPEG for example.*
- **NARA Transfer Guidance** *Indicate Preferred or Acceptable here, if applicable. If the format is in multiple Categories and the Transfer Guidance varies based on the Category, indicate that. Such as, "Preferred for Textual and Word Processing; Acceptable for Presentation and Publishing"*
- **Risk** *Include the overall Risk (Low Risk, Moderate Risk, or High Risk) from the Matrix.*
- **Preservation Action** *Controlled List. The options are:*
    - **Identify Version** *Use when the most imminently needed preservation action is to identify the version(s) of a format, as version-based identification has not been performed and the longer-term preservation action may be version dependent (e.g., some versions of Microsoft Word are retained, and some versions are transformed. Therefore the preservation action for NF00659, Microsoft Word unspecified version, is Identify Version).*
    - **Transform**
    - **Retain**
    - **Retain for Future Assessment** *Use when further research is required, unknown if it will be retained in current format or transformed.*
- **Proposed Preservation Plan** *Semi-controlled list. Use the following options, with additional description as necessary:*
    - **Depends on Version** *Use when the Preservation Action is Identify Version.*
    - **Further research is required**
    - **Transform to a TBD format**
    - **Transform to [insert format name]** *Use when target format is known, and preferably transformation tool(s) are already available to NARA.*
    - **Retain** *Use when the Preservation Action is Retain.*
    - **Retain but extract files from the container** *Use when the format is a container such as TAR or ZIP archive.*
- **Description and Justification** *You must supply a Justification for the Preservation Action(s) selected, which should be no more than 2 brief paragraphs. This can be as brief as a single sentence. Provide additional, brief description of the format as necessary.*

- **Preferred Processing and Transformation Tool(s)** *Provide the name of the preferred tool(s) used to review the records in this format and prepare them for ingest. Provide the name of the preferred tool(s) used to transform records in this format into new formats for preservation. If there is a known, preferred software package that should be used or recommended for working with records in this format, document the software here. Separate tools with semicolons.*
    - *If transformation tools are not currently available to NARA, include "Procure and/or develop tools"*

Metadata

All metadata is embedded in the structure of the Matrix and Plans.

Plans must include the date that the current version of the Plan is complete and supersedes any previous versions.

File Management

A **copy** of the most current version of each file should be stored in Drive. Include the date in the filename. We want to keep versioned copies to trace the history of our preservation decisions and justifications.

- NARA_File_Format_Risk_Matrix_Weights_YYYYMMDD (spreadsheet)
- NARA_Deprecated_IDs_YYYYMMDD (spreadsheet)
- control_list_NARA_IDs (spreadsheet used for assigning new IDs, not published to GitHub)
- NARA_PreservationActionPlan_FileFormats_YYYYMMDD (spreadsheet)
- NARA_File_Format_Risk_Matrix_YYYYMMDD (spreadsheet)
- NARA_PreservationActionPlan_WebRecords_YYYYMMDD (text document, follow this naming convention for all category plans)

Preservation Actions Plans must be reviewed and updated as needed on a quarterly basis at a minimum. Trigger events for the creation of a new plan include, but are not limited to, the addition of new Categories to the NARA Transfer Guidance, or updates to the Holdings Profile. When updating the Plans, create a new, dated version (do not directly edit the existing Plans).

Publishing to GitHub

When ready for release, create a new folder in *Preservation Action Plans / Completed Releases* and name the folder with the month and year of the release, such as 2021-04. Any files which should not be uploaded to GitHub, such as working files and the internal-only version of the Matrix with raw counts present, should be placed into a sub-folder that is clearly labeled DO NOT UPLOAD.

The Matrix and File Formats Preservation Action Plans spreadsheets should be released as XLSX and CSV files. The Category Preservation Action Plans text documents should be released as PDFs.

Include an external-facing version of this SOP for GitHub. Links and file paths to NARA-internal systems (ICN, Google Drive, NARA at Work) should be removed for the GitHub copy, and it should be released as a PDF.

SOP Update Strategy

This SOP will be updated every odd numbered fiscal year during the first quarter ***or*** if there is a trigger event such as changes to business processes, to staffing, and/or to systems.

When the SOP is updated, include the most up-to-date version in the next round of publishing to GitHub.