Release of the Digital Preservation Framework by the National Archives (NARA)

Leslie Johnston, Director of Digital Preservation
U.S. National Archives and Records Administration (NARA)
CNI Fall 2019

The Context for Formats 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.

It Started with a Collection 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.
- The reporting didn't match in terms of granularity for the various systems, given different tooling for format analysis and reporting.
  - One system uses DROID 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 use DROID.
  - One system could provide a report of all the file names including extensions but no format identifiers.
  - One system provided a 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. It is not complete or perfect.
- 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 as Adobe Acrobat PDF 1.4. This required some normalization when aggregating the data together to compare across the holdings.

Assessing Risk

- In 2018 NARA created a 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 category is weighted using relative weightings that map to the level of risk for each question and, to the extent that it can be defined, cost. The Matrix is designed to help us understand two basic aspects of the formats in our holdings:
  - How much do we know about a format?
  - What impacts our ability to process, render, and preserve that format?
- Risk factors are then added to a Prioritization Matrix to assess the preservation actions that could be taken vis-à-vis our current environment and capabilities, which generates a final numeric scores. These scores are mapped to High, Moderate, and Low Risk.
  - The risk mapping thresholds are open to review and revision over time.

Translating Risk Assessment Results into Preservation Plans

- NARA developed draft Preservation Actions Plans for file formats, applying the Matrix factors to the file formats in our custody to create a prioritized list of formats for preservation actions and to identify processing tool needs. The plans cover over 350 format variations for 15 record types:

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

What Does Each Plan Contain?

- Current NARA Transfer Guidance formats for the record type.
- Significant preservation properties/essential characteristics for the record type:
  - Appearance
  - Structure
  - Behavior
  - Context
- An appendix for each related file format currently identified in the holdings:
  - Current assigned level of preservation risk and priority for preservation actions
  - Links to specifications/documentation
  - Recommended preservation migration actions, including no action if appropriate
  - Recommended tools for processing and preservation
- Information, from preliminary review, on format types currently provided to researchers through reference requests and through the National Archives Catalog.

Public Release for Comment/Discussion

- The Matrix and the Draft Plans were made available on NARA's Github account, ready for public comment and discussion:

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

- We solicited feedback on the essential characteristics and recommended preservation actions, and help identifying tools for both processing and access.

- Public comments were collected through November 15, 2019.

Reviewing the Feedback

- We received 26 comments on GitHub and over 2 dozen in-person comments and questions when the framework release was presented at iPres and DLF.
- No negative feedback was received, only praise and suggestions.
- The comments fell into several broad categories:
  - Add Formats (3)
  - Digitized versus Born-Digital (2)
  - Links to Other Resources (6)
  - Machine Readability (5)
  - Plan Formatting (2)
  - Policy Recommendation (9)
  - Processing Practice Recommendation (5)
  - Tool Recommendation (2)

Feedback Details

- Add Formats (3)
  - Suggestions for additional formats that we should include.
- Digitized versus Born-Digital (2)
  - Asked for clarification about format standards in digitized versus born-digital records.
- Links to Other Resources (6)
  - Suggested specific links in plans that were not included, but more generally requested as complete a set of links to other related online resources as possible for each format.
- Machine Readability (5)
  - Requested a machine-readable version of the plans.
- Plan Formatting (2)
  - Suggestions about plan layout or structure.
- Policy Recommendation (9)
  - Recommendations for additional information to be tracked and included in the plans, requests to release detailed format specifications, a request to release specific settings for tools used in the transformations, and a suggestion for plan update triggers.
- Processing Practice Recommendation (5)
  - Suggestions for the use of different formats.
- Tool Recommendation (2)
  - Suggestions for tools to be considered in processing and preservation migration.

Outcomes from the Release

- It became clear that there was not enough explicit clarification that these plans represented NARA's current practices based on its current infrastructure and capabilities to support our digital preservation strategy to create normalized versions of files, and that these were NOT a recommendation for the entire archival field. This will be made more explicit in the next release.
- We will add suggested formats to the plans that were deemed important by the community. These include eBook formats, research data formats, and noSQL databases.
- We will review all practice and policy-related feedback and update the plan content and our practices/policies accordingly. Appropriate expert NARA processing staff will respond to the feedback on Github.
- We must review and ensure that applicable URIs are consistently included for the Library of Congress Format Sustainability site, PRONOM, any formal or reverse-engineered specifications that can be found, and to Wikidata, because it was not consistent across all the formats in all the plans.
- Eleven comments focused on links/relationships to other resources and machine-readability in the increasingly important linked resource community. The community absolutely requires that this resource be released as Linked Data with URIs that link to as many authoritative online resources as applicable. Wikidata is becoming the hub for digital preservation format information across the community. This will happen in 2020.
- We will continue the discussion on Github. It is clear that this level of transparency is welcomed and valued by the community.

Thank You

Leslie Johnston
leslie.johnston@nara.gov