Introduction to Digital Preservation at NARA

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

Two Preservation Units at NARA

- **Preservation**, with the mandate to care for the physical holdings and perform preservation reformatting digitization of physical items, and to undertake preservation research.

- **Digital Preservation**, with the mandate to support electronic records processing archivists, perform audits of the holdings, and assess the need to perform preservation actions.

What are Digital Preservation Risks?

- Safeguarding the data: Preservation Planning, Active Object Management, and Bit-level Preservation.
- Ensuring that the content stays accessible and usable: File Format Migration.
- Maintaining the context of the data: Descriptive, Structural, Administrative, and Preservation Metadata.
- Ensuring there is ongoing trust in the data: Auditing and IT Security.
- The keys to all of this are good planning, thorough documentation, capturing and creating as much metadata as is feasible, and ongoing monitoring of the object and the infrastructure they are stored in.

Guiding Assumptions

- Electronic record files should conform to the NARA Transfer and Metadata Guidance whenever possible.
- All files must have recorded fixities to support auditing.
- Actions taken on files must be recorded and tracked.
- Separate public use copies of files are created.
- At this time, ongoing preservation format transformations are not performed but are planned. Format transformations DO happen when formats are received that we cannot process/preserve.
- Regular audits must be performed.

Guidance to Agencies plays a major role at the Beginning of the Digital Preservation Lifecycle

- 2020-01: Guidance on OMB/NARA Memorandum Transition to Electronic Records (M-19-21)
- 2018-01: Revised Format Guidance for the Transfer of Permanent Electronic Records (formerly 2014-04)
- 2015-04: Metadata Guidance for the Transfer of Permanent Electronic Records
- 2015-03: Guidance on Managing Digital Identity Authentication Records
- 2015-02: Guidance on Managing Electronic Messages
- 2014-06: Guidance on Managing Email
- 2014-02: Guidance on Managing Social Media Records
- 2013-03: Guidance for Agency Employees on the Management of Federal Records, Including Email Accounts, and the Protection of Federal Records from Unauthorized Removal
- 2012-02: Guidance on Managing Content on Shared Drives
- 2010-05: Guidance on Managing Records in Cloud Computing Environments

- https://www.archives.gov/records-mgmt/bulletins

Formats for Records Transfers

- An integral part of NARA's digital preservation 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

- 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, and NARA negotiates with each agency about what it can provide.

Digital Preservation 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.

- A new version will be issued in 2022.

Core Digital Preservation Infrastructure

- *ERA Business Objects*, to initiate and record approvals for federal agency electronic records scheduling and transfers
- *Processing Tools:*
  - *AISS*, which runs the automated workflow to read media received from federal agencies to prepare for ingest.
  - *AMIS*, to log ingest and processing workflows for federal records.
  - *AERIC*, to validate structured data received as federal records.
  - *Commercial and Open Source Software*, to automate processing, review files, and transform formats.
- *Preservation Systems:*
  - *ERA 2.0*, to process and preserve federal records.
  - *ERABase*, legacy mechanism for authorizing transfer of federal records to NARA.
  - *ERA Title 13*, to preserve Census records.
  - *ERA EOP*, to preserve Presidential records.
  - *ERA CRI*, to preserve Legislative records.
- *DAS*, to create descriptive metadata for records
- *The National Archives Catalog*, to deliver records to the public

Current Digital Preservation Activities

- File Format Transfer Guidance for agencies to ensure that records are transferred to NARA as sustainable formats.
- Ingest of files from agency media (drives, optical media) and network transfers of files directly from agencies, including:
  - Assuring, if possible, that the formats align with our Transfer Guidance
  - Capturing descriptive, structural, and preservation metadata
  - Transforming file formats where appropriate and/or possible
  - Checking for fixities and assigning fixities if none came with the records
  - Running file format validation checks
  - Creating manifests and logs of all ingest actions
- Audits of media in the collection:
  - Annual sample of media
  - 10 year migration of media
- Ongoing monitoring of the systems and infrastructure:
  - Monitoring of system and storage status
  - Monitoring of the holdings files preserved using those systems
  - Regular emergency system backup restoration tests
- Documentation of Standard Operating Procedures related to digital preservation and file management across all units and locations, made available in a centralized internal location; creation and maintenance of the Digital Preservation Framework.

Digital Preservation Framework: Internal Holdings Format Profile

- One aspect of having several systems is not having a single measure of what we have in our holdings, and not having a single approach to 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.
  - 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.
  - There were also different granularity levels reported for format versions, 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.
- Through a manual process we now have a Holdings Format Profile to document what formats we have across all systems.
- For some files 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.

Digital Preservation Framework: Risk Matrix

- NARA has 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.

Digital Preservation Framework: Preservation Plans for Record Types

- NARA developed its Digital Preservation Framework to document and share recommended preservation actions based on its electronic record holdings and current capabilities. There are Preservation Plans for 16 categories of electronic records (or “record types”) which identify “Significant Properties,” the properties that should, if possible, be retained in any format migration:

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

Digital Preservation Framework: Preservation Plans for File Formats

- NARA developed Preservation Action Plans for over 500 file formats based on an analysis using the Risk Matrix and the Essential Characteristics for each associated record type. The Format Plans include links to standards and specifications and proposed preservation actions and tools based on current NARA thinking and capabilities.
- The Digital Preservation Framework Matrix and Preservation Plans are publicly available on the NARA Github account for reuse and adaptation, as well as discussion.

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

- The Framework can be applied across the lifecycle: its documented format sustainability metrics provide critical context for records creator and management decisions, support records selection and appraisal, and guide the selection of formats available for public access.

Updating the NARA Infrastructure

- ERA 2.0 went into production in October 2018 in the AWS GovCloud: it introduces a more flexible and extensible framework for processing and preservation.
- Updated infrastructure and tools to support more efficient transfer, processing, and delivery of born-digital and digitized records.
- Increased capacity for the transfer and processing of records, and the ability to support additional and every-changing file formats more readily.
- A Digital Processing Environment (DPE) system for the records transfer and ingest process that also incorporates virtual machines which run commercial and open source tools needed by archivists to process electronic records.
- A cloud-based Digital Object Repository preservation system.
- Support for ingest, processing, and preservation of digitized records created in-house and by partners.
- An updated Disposition Scheduling and Transfer administration workflow infrastructure capable of supporting both current ERA and the DPE environment, which will go into production in 2020.

Roadmap for Future ERA 2.0 Functionality

- New tools for processing and preservation actions are reviewed in an ongoing process.
- Planning and starting the migration of the electronic records holdings from several other current production systems, and a consolidation of their functionality.
- An effort is needed to put enterprise-level tools in place to support for the processes and tools used for FOIA, special access requests, and redaction for pubic release.
- Preservation risk auditing and reporting will become easier as NARA consolidates its files into the new environment with more robust format characterization tools, auditing capabilities, and reporting tools. This will allows us to monitor and report on the entirety of the holdings in one place for the first time.
- A parallel environment for Classified electronic records is planned.
- When will this all be in place? It's a gradual process over at least 3-5 years to move from a fully manual to a partly (hopefully fully) automated process, given the record and system migrations and cloud-to-cloud record ingest functions that come first in the development priorities.

How do we assess our progress?

- In 2019 and 2021, NARA completed internal self-assessments of its programs and systems using the PTAB (Primary Trustworthy Digital Repository Authorisation Body) instrument based on ISO 16363.
  - 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. This is the necessary first step in a gap analysis to prioritize investment in the infrastructure, policy revisions, and the normalization and documentation of processes.

NARA Self-Assessment Outcomes

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

Who is working on Digital Preservation?

- The Digital Preservation unit guides internal operations:
  - Advised by an agency-wide Digital Preservation Group
  - Maintains the Holdings Profile
  - Responsible for the Risk Analysis and File Format Preservation Action Plans
  - SOPs for digital preservation and management of files during digitization are submitted to the digital preservation unit
  - Transparency and replicability are key: processes and documentation are made available to all NARA staff
- This is distributed work across the agency, not just the digital preservation unit
  - Agency Services for external guidance, creation of Schedules, and records appraisal
  - Processing archivists in the Legislative Archives, Presidential Libraries, and Research Services for Federal records
  - IT Operations and IT Security for monitoring of systems and storage status, and regular emergency system backup restoration tests
  - The Innovation unit provides public access through the National Archives Catalog
  - Agency leadership supports the work through policies and resources

For More Information, Feel Free to Contact Us

Leslie Johnston

leslie.johnston@nara.gov