Digital Preservation at NARA
Policy, Records, Technology

Leslie Johnston

Director of Digital Preservation

US National Archives and Records Administration (NARA)

ARMA, April 18, 2018

Policy

- Managing Government Records Directive (2012)
  - All email managed electronically by end of 2016
  - All permanent electronic records managed electronically by end of 2019
- 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
- 2014-04: Revised Format Guidance for the Transfer of Permanent Electronic Records
- 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

Digital Preservation Strategy

- NARA published its first Digital Preservation Strategy in June 2017 to guide its 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

  - Data Integrity

  - Format and Media Sustainability

  - Information Security

Record Content

- **Domain**: Records of US Federal Government

- **History**:
  - Electronic records covered by our records law from 1943 ("other documentary materials, regardless of physical form or characteristics")
  - NARA started collecting electronic records in 1970

- **Process**:
  - Records selected for preservation either through one of several federal regulations or through the records scheduling process. Permanent records:
    - Document individual rights;
    - Document actions of officials: provide government accountability; and/or
    - Document the national experience.

Record Formats

- **Domain**: Records of US Federal Government

- **History**:
  - Electronic records covered by our records law from 1943 ("other documentary materials, regardless of physical form or characteristics")
  - NARA started collecting electronic records in 1970

- **Process**:
  - Records selected for preservation either through one of several federal regulations or through the records scheduling process. Permanent records:
    - Document individual rights;
    - Document actions of officials: provide government accountability; and/or
    - Document the national experience.

Technology

- Phase 1: Tape, 1970- present
  - Classified records still managed on tape

- Phase 2: Electronic Records Archives (ERA), 2008 - present
  - Approximately 550 TB
  - Three separate environments aligned with record regulations, plus the National Archives Catalog
  - Online support for records management transactions
  - Preservation repository, PREMIS metadata catalog

- Phase 3: ERA 2.0, 2018 - future
  - Cloud-based scalability for processing and storage
  - Flexible processing environment for running diverse tools
  - Improved search and access in preservation repository

How NARA Organized its Digital Preservation Planning

Policy

Goal: Policies and processes are in place to preserve all electronic records transferred to NARA for permanent retention and preservation as well as for digitized surrogate files.

Example Policies:
- Digital Preservation policy for NARA.
- Policy on the preservation of digital surrogate master files.

Example Gaps:
- Policy and SOPs for the assessment of file formats in the holdings and the triggers for format transformations.

Infrastructure

Goal: NARA has adequate storage, network capacity, systems, and tools for the ingest, processing, active file management, and preservation of its digital holdings.

Infrastructure Examples:
- Managed, replicated preservation storage.
- Tools for the characterization and validation of file formats.

Example Gaps:
- Hardware and software to read an increasing variety of legacy storage media upon which records are transferred.
- Tools for file format preservation transformations.

Staffing

Goal: NARA has sufficient staff with appropriate training to do the work of digital preservation, so that electronic records are properly and actively managed through their lifecycle.

Examples:
- Archivists responsible for the ingest, processing, and description of electronic records.
- IT specialists who research technologies and develop or modify open source tools for the processing of electronic records.

Example Gaps:
- Ongoing training in community best practices and technologies.
- Archivists who analyze file formats in the holdings and run system preservation operations.

Digital Preservation Assumptions

- Electronic records received should conform to the NARA Transfer Guidance for file formats.

- Master preservation files are retained.

- All files must have recorded fixities

- All actions taken on files must be recorded and tracked.

- Public use copies of files are created.

- At this time, preservation format transformations are not performed but are planned.

- Regular audits must be performed.

Current Digital Preservation Activities

- File Format Transfer Guidance for agencies to ensure that records are transferred to NARA as sustainable formats.

- Digital Preservation Strategy and Digital Preservation advisory group to guide internal operations, including SOPs and File Format Preservation Action Plans.

- Ingest of files from agency media (drives, optical media) and network transfers of files directly from agencies, including:
  - 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

Case Study: Email

- Email is an obvious area of records management emphasis, but we lack a systematic approach to receive, process, and make emails available for public access

- Growing scale of email as a record type

- Need to determine best technical approach
  - Similar but not identical business/functional needs across Presidential, Congressional, and Federal records and the systems that currently hold them
  - How to ensure the availability of messages (including attached files)

- Need to assess staff resources for reviewing emails
  - How to identify and deploy capabilities in ERA 2.0 to eventually consolidate all systems
  - How to reduce dependence on human reviewers

What Recent Preservation Technology Trends Should We be Watching?

- Web Archiving is not new, but there is greater public awareness of web content as potentially valuable but often transitory, and its use in Federal records management and preservation is increasing. There are exciting new tools for indexing and playback of web archives.

- Format obsolescence is always an issue for digital preservation. There have long been tools to identify and characterize file formats, such as JHOVE and DROID, but more organizations internationally are starting to actively share their risk assessments that they use to gauge the sustainability of formats and what preservation actions to take.

- There is new recognition that digital items have a context -- they're not always files sitting in directories on desktops and servers -- they live in systems and in complex web applications and our interactions with them are based on algorithms that control how search results are ranked (or shown to us at all) and in some cases so personalized that those algorithms are actually mediating what information comes our way.

- There is increased interest in the preservation of software. The Library of Congress has a symposium on architectural design files and software preservation. The Software Preservation Network and Software Heritage are testing the boundaries of public policy and legal framework to preserve code.

- The issues of working with vintage hardware and media are getting more attention: The National Archives of Australia was interviewed about its practices. And video preservation continues to be a critical topic. Born digital media files only 4 years old can be at risk without stewardship. As is email.

Thoughts and Questions?

Leslie Johnston

Director of Digital Preservation

National Archives and Records Administration

leslie.Johnston@nara.gov