U.S. National Archives and Records Administration

Digitization Tools for Archives

By Steven Puglia
U.S. National Archives and Records Administration
8601 Adelphi Road, College Park, MD 20740, USA
Phone: 301-837-3616
Email: steven.puglia@nara.gov

November 2009

Preservation is a long-term management process addressing the identification and mitigation of the risks for loss of information, and for appropriate records includes protection of the original physical form.

Archives are about managing and making information available - maintaining the authenticity and integrity of the records.

New focus for archives and other cultural heritage organizations-

- Get more records online
- Digitize large, complete groups of records

A transition from stand-alone projects to digitization being a core archival activity.

Given the scale of holdings-

Need the most cost-effective and efficient complement of approaches and tools to achieve the long-term preservation of and access to records.

For traditional reformatting, we developed infrastructure and defined workflows.

Need the same for digital reformatting.

Need to take advantage of the benefits of computers and information technology.

Need to develop comprehensive workflow tools for all aspects of digitization and automate as much as possible to gain efficiencies on a large scale.

From a records management perspective, reformatting should be done in a manner that follows ARMA International's *Generally Accepted Recordkeeping Principles* (http://www.arma.org/garp/):

- Principle of accountability
- Principle of integrity
- Principle of protection
- Principle of compliance
- Principle of availability
- Principle of retention
- Principle of disposition
- Principle of transparency

To Ensure Authenticity and Integrity:

The Archives New Zealand S-6 Digitisation Standard
http://continuum.archives.govt.nz/files/file/standards/s6.pdf

- All digitization and digitization processes must be planned, scoped and documented
- An appropriate digitization approach must be selected, documented and implemented
- Technical specifications aligned to the digitization requirements must be selected, documented and implemented
- Equipment and software aligned to the digitization requirements must be implemented

- Systems to support management of the digital output of digitization must be in place
- Guidelines for the preparation of original collections/records must be documented and implemented
- All digital objects created must be assigned metadata to document digitizing processes and to support ongoing business processes
- Quality assurance and quality control procedures must be defined, documented and implemented

- Digital storage and disaster recovery procedures for digital objects and metadata must be defined, documented and implemented
- Systems for the long-term management of digital objects and metadata must be documented and implemented
- Preservation strategies and processes for digital objects and metadata must be defined, documented and implemented

Digitization is a managed process that involves the entire institution and broadly includes:

- Selection
- Assessment
- Prioritization
- Project management and tracking
- Preparation of originals for digitization
- Metadata collection and creation
- Digitizing
- Quality management
- Data collection and management
- Submission of digital resources to delivery systems and into a repository environment
- Assessment and evaluation of efforts, with continual process improvement

Digitization workflow can be divided into four main phases:

- Project planning
- Processes occurring prior to digitization
- Digital conversion
- Post-digitization work

Digitization Activities: Project Planning and Management Outline November 2009

Federal Agencies Digitization Guidelines Initiative

For many institutions, we feel the “managed environment” needs to be extended beyond the digital repository and forward in time to include the entire digitization process.

The managed environment should include support for:

- Collection of information on intellectual and physical organization of holdings/collections
- Definition of approaches to digitization and management of the process
- Creation of additional metadata to document the originals, the digital objects, the process of creating the digital objects, and any changes to the digital objects

The managed environment should include support for:

- Validation and/or authentication of the digital copies in comparison to the originals and to the institutional definitions for the digital objects
- Ability to package digital objects and related metadata for use in access systems and for submission to digital repositories

Build appropriate information technology infrastructure, including well designed and robust IT systems.

Range of IT issues:

- Data validation and integrity
- Records integrity and provenance
- Data and system security
- Metadata (all types)
- Hardware and software
- Policies
- Staff

Effective IT procedures exist for the short-term management of electronic records and digital information-

- Not always followed.
- Not always as easy or as inexpensive as advertised.
- We have been sold on the promise of the technology, but rarely acknowledge the downsides.

Emphasize data security, including back-ups, distributed storage of multiple copies, etc., to prevent catastrophic loss of digital resources.

Goals:

- Produce consistent, high-quality digital objects and related metadata.
- Facilitate the long-term management and preservation of the digital resources.

Use appropriate standards whenever possible for:

- Metadata of all types
- File formats
- Approaches to digitizing physical collections
- Etc.

Collect and create as much metadata as possible:

- Descriptive / Discovery
- Administrative
- Technical
- Preservation
- Behavior / Structural

Our perspective has been-

- Standardize your digital objects, just like your metadata.
- Better to define consistent approaches.
- Treat large batches of images, or other digital objects, in the same way.

Standardization will promote ease of management and lower costs to maintain and preserve digital data/objects/records.

Conceptually, the scope of the digital repository can be pushed forward in time to the point of creation of the digital objects and metadata.

Digitization activities should be geared towards integrating appropriate functionality across all phases and activities of digitization to ensure the relatively automated submission of appropriate and high-quality digital objects and related metadata to digital repositories.

Need to bring digital resources into a managed environment as soon as possible, to facilitate management, access, and long-term preservation.

Needs do not match a specific model or market segment, includes the following functionality-

- Records management application (RMA)
- Content management (CM)
- Document capture and management
- Digital asset management (DAM)
- Media asset management (MAM)
- Digital repository
- Digital preservation

Functionality should include-

- Digitization/image acquisition
- Integration with image/signal processing applications (including batch processing)
- Project management
- Digital storage
- Automated characterization of resources
- Organization
- Description
- Search
- Access

Conclusion:

The creation of integrated tools and IT environment to coordinate and support all phases of digitization and related activities, up to and including submission of digital objects and metadata to the digital repository, will serve the long-term goals of archives best.