LIBRARY OF CONGRESS

March 28, 2023

Digital Collections Repository Development and Storage Infrastructure

Daniel Krech
Information Technology Specialist
Mission Platforms Section

Mark Cooper
Senior Digital Collections Specialist
Digital Content Management Section

Background

- **2018**: Digital Content Management section forms in Library Services
- **2019**: DCM completes analysis of current digital collections inventory and storage systems
  - At that time, inventory system covered approximately **1 billion files, 1 million inventories**
- **Selected key findings:**
  - No practical capability to validate collections at scale and produce actionable data
  - Widespread untracked collections content and untracked changes
    - **15%** of files not tracked in inventory system
    - **12%** of inventory records had errors
  - Significant content duplication
  - Lack of review of storage usage
    - Content not in long-term storage, or that should not be in access storage
  - Lack of derivative management processes or other automation capabilities
  - No ability to assess at scale whether content intended to be available to users actually was

Implications of System Gaps on Storage Usage

- General & International Collections (GICD) presentation storage example:
  - 2020: 230 TB / 77 million files; 67 million of those not managed in inventory
  - 2022: Significant and manual remediation efforts by DCM resulted in **50% reduction in presentation usage by both size and count, 98% reduction in unmanaged files**
  - Anticipate even higher ongoing reductions across all collections
- Very high numbers of unneeded files
  - Very difficult to remediate in current systems
  - Unnecessary, outdated derivatives including ~50 million GIF thumbnail images
  - Large original files in presentation when not needed and not presented
- System workflows resulting in significant duplication
  - Routine processes that include repeated receives of content
  - Ex: collection with 50 TB / 8 million files; de-duping identical files would remove 90%

Next-Gen Digital Collections Repository Goals

- Maintain integrity of all collections content
- All changes to collection content tracked and versioned, with no possibility of untracked changes
- Remove duplication of content
- Increase storage usage efficiency and intentionality
- Independent management of preservation and derivative content
- Capability for users to work at scale and with automated processes
- Ability to track and report on access status of collections

LIBRARY LIBRARY OF CONGRESS

Digital Collections Repository - Paprika

- **2018**: Digital Content Management section formed
- **2019**: DCM completes analysis of current digital collections inventory and storage systems
- **March 2020**: Next-gen system development begins
- **March 2021**: Launch as production system, ingest of 170,000 digitized books
- **March 2022**: Begin automated release of collections to access systems
- **June 2022**: User interface implemented
- **December 2022**: Reach 160,000 books (30+ million pages), previously unmanaged in LC systems, released to loc.gov
- **Ongoing** implementation of new features including expanding ingest streams, enhanced derivative services, reporting integration, user interface enhancements, and more

Digital Collections Repository - Process

- Close collaboration between Mission Platforms section and Digital Content Management section to deeply understand desired product
- Iterative and agile development delivering immediate value
- Cloud-first development
  - Use of well suited high level services enabling rapid delivery of the system and establishing automated processes
- Foundational use of fixed content addressed storage as an integral aspect of the system

Digital Collections Repository - Architecture

- Write-only for collections storage
  - Objects stored by multi-hash
- All actions and versioning tracked in ledger
  - Leveraging AWS Quantum Ledger Database
- Derivative and access storage managed separately from permanent collections
  - Increasing efficiency of usage; presentation storage only used for content actually being made available
  - Ability to update or recreate derivatives as needed
- Storage architecture evolved to meet requirements from assessment of current practice
  - Unmanaged content, untracked changes, content duplication not possible by definition in construction of the architecture
  - Validation capability built in to all steps of storage with content managed by hash

LIBRARY | PAPRIKA
Digital Collections Repository

General Digital Collections