LIBRARY LIBRARY OF CONGRESS

2019-09-09

Data Integrity Review Project

Digital Content Management Section (DCM)
Library Services

Mark Cooper
Digital Collections Specialist

The Challenge

- Over 1 million inventories in the authoritative digital content inventory system
- Inventories have never been systematically validated at scale
- When content enters the system or an operation is performed, content is validated
- Static content has never been completely validated

[Crowded stacks, 1970]
https://www.loc.gov/resource/ds.10193/

Project Unknowns

- Is there systemic data corruption or system failure?
- Can the system practically perform this task?
- What types of issues will we find?
- How will we understand the results and fix the issues?

Library of Congress deposits in basement
https://www.loc.gov/resource/npcc.20063/

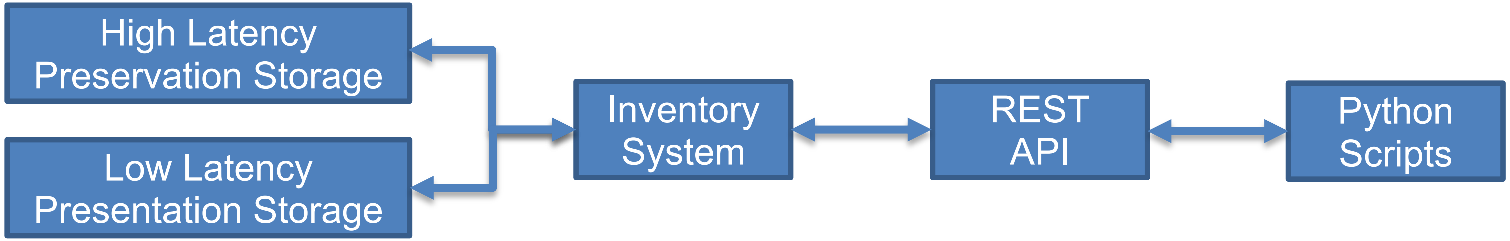
Project Goals

- Develop a method to validate data integrity of all digital collections
- The process must be repeatable
- Results must be understandable and actionable
- Remediation of errors must be possible and practical

[People working in Card Division, Library of Congress, Washington, D.C.]
https://www.loc.gov/resource/cph.3c18631/

Developing a Plan

- Leverage inventory system REST API
    - Designed a suite of Python scripts to utilize the API to perform checks, extract results, and execute remediation
- Select area for first phase of review large enough to be representative
    - Initiated manifest and MD5 hash validation on 100 million files / 240,000 inventories in low latency presentation storage


High Latency Preservation Storage
Low Latency Presentation Storage
Inventory System
REST API
Python Scripts

Results and Analysis

- API enabled utilizing existing system to successfully validate content at scale
- Generated MD5 hash for 100 million files to validate against manifests
- No evidence of systemic corruptions or failures
- Very rare cases of issues with content caused by system errors had previously been reported, but not corrected
- However, 30% of inventories failed validation check – largely presentation derivatives

Results and Analysis

- Content on storage is correct, inventory is not
- Content custodians working around system limitations, resulting in broken inventory records
- Content in the digital storage system needs to be understood as potentially dynamic, in particular for presentation and access
- System needs to facilitate required actions in ways that are logged and versioned

Next Phases

- 99% of first phase inventory issues are resolved
- DCM is systematically expanding scope across all systems
- Clean inventory system that reflects the current state of content, and a corrupted file should prompt immediate action
- Generating future systems recommendations

[Woman at Main Reading Room card catalog in the Library of Congress]
https://www.loc.gov/resource/cph.3c00400/