From Scratch: Building a Sustainable Born-digital Workflow in LOC's Manuscript Division

DPC's Digital Preservation Workflows Series

February 4, 2025

Kathleen O'Neill
Senior Archives Specialist
Manuscript Division

LIBRARY OF CONGRESS

Overview

- 270+ collections with born digital materials
- 160+ collections are processed, described, and available to researchers onsite-only in the Manuscript Division reading room (3.56 TB, 750,450 files)
- Files formats and media formats span the late 1970s to the present
- Collections range in size from 1 file to over 2 million files
- Largely hybrid collections. Beginning to receive large acquisitions from cloud-based storage and platforms (e.g., email, Dropbox, etc.)
- All collecting areas, particularly strong in science, political, judicial, African-American, and women's history

Taking Stock

“Great cooking favors the prepared hands.”
Jacques Pépin

Identify Your Resources

- Budget
- Supervisors, Staff, Colleagues
- Time
- Tools, Technology, and Infrastructure
- Wider Library Community

Understand Your Holdings

- Survey
- Co-located media
- Scale
- Prioritize
- Advocate for resources

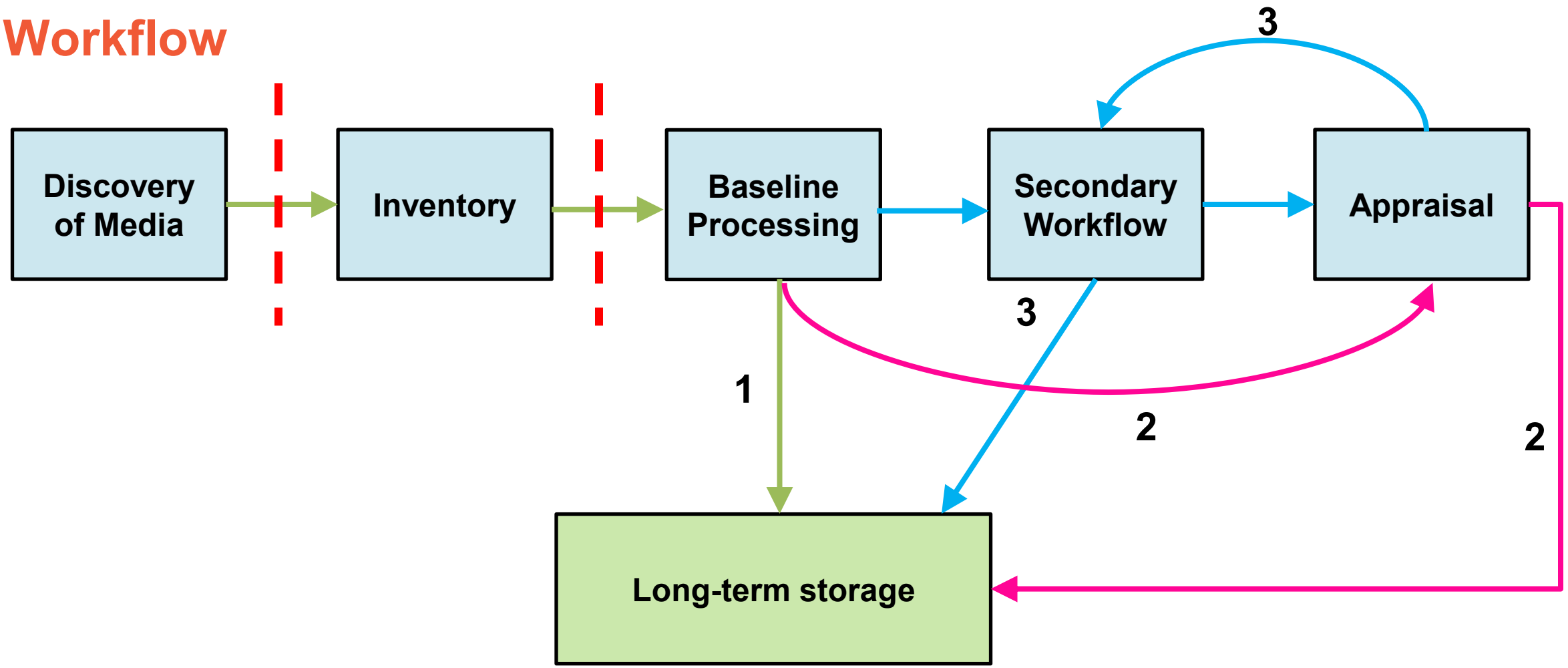
Make an Honest Assessment

- What are the minimum steps sufficient to responsibly curate the materials?
- What can we realistically and responsibly do in the present?
- What can we realistically and responsibly sustain into the future?

Building the Workflow

“Do what you can, with what you've got, where you are.”
Squire Bill Widener of Widener’s Valley, Virginia

Workflow


Discovery of Media
Inventory
Baseline Processing
Secondary Workflow
Appraisal
Long-term storage
3
3
1
2
2

Discovery of Media and Registration

- **Tasks**
  - Survey the collection
  - Remove media when feasible
  - Document context & location w/removal forms
  - Rough count of media and formats
  - Register
  - Estimate file and byte count
  - Shelve
  - Publish initial catalog record
- **Tools**
  - Removal form (Word Doc)
  - Access database tracks all registrations
  - Spreadsheet (tracking processing status)
  - Spreadsheet (metadata & processing)

Inventory

- **Tasks**
  - Physical appraisal
    - Identify duplicates & content inappropriate for retention
    - Physical condition
    - Group by format, content, subject
    - Update media formats counts
  - Photograph media
    - Assign digital IDs
    - Photograph media, save images to project files
  - Inventory informs processing plan
- **Tools**
  - Archival permanent pens
  - Image form
  - Camera and camera stand
  - WorldCat & LOC catalog

Baseline Processing

- **Tasks**
  - Metadata
    - Directory listings
    - Media and process info logs
  - Copying content to local storage
    - Bagging
    - Checksums and inventory
    - Virus scan
    - Document any errors
    - Initial reports
      - PII, file formats, duplicates, keyword
- **Tools**
  - Cmd line
  - Spreadsheet (metadata & processing)
  - Three (3) digital workstations
    - 2 uFREDs
    - 1 LOC standard workstation
    - Various drive readers
  - Bagger tool (GUI)
  - Microsoft defender virus scan
  - DROID
  - FTK & Sleuthkit

Secondary workflow

- **Tasks**
  - Media/files needing additional processing
    - Rip to .wav
    - Unzip
    - Disk images
    - Data recovery
    - Emulation
    - Obsolete media and file formats
    - Redaction
    - Migration
- **Tools**
  - Spreadsheet (metadata & processing)
  - Bagger tool (GUI)
  - Exact Audio Copy, Handbrake, fre:ac
  - Stuffit Expander
  - FTK Imager
  - FRED w/full FTK license
  - Emulation – QEMU, DosBox
  - DROID, Pronom, Sustainability of Formats
  - HxD
  - Kryoflux
  - Sleuthkit

Appraisal

- **Tasks**
  - Review content
  - Record technical & descriptive metadata
  - Delete temporary/empty/system files
  - Identify & manage restricted/PII materials
  - Minimal arrangement of files
  - De-duping at media or folder level
  - Favor emulation over migration
  - Selective retention of disk images
  - Consider unique research use cases
- **Tools**
  - File viewers
    - QuickView Plus
    - File Viewer Pro 4
  - Beyond Compare
  - File Format reports
    - DROID
    - FTK
  - Emulation
    - QEMU
    - DosBox

From Scratch

“Cooking requires confident guesswork and improvisation— experimentation and substitution, dealing with failure and uncertainty in a creative way”

Paul Theroux