Documenting Born-Digital Ingest Workflows

Mike Shallcross
Indiana University Libraries

Best Practices Exchange
May 1, 2019
Indiana University & Born Digital Archives

- Extensive digital collections since early 90s (digitized AV, images, texts, TEI)
- Founding member of HathiTrust; Samvera partner
- Born Digital Archives
  - Custom projects: Virtual CD-ROM / Floppy Disk Library (ca. 2007-08)
  - Institutional repository (IUScholarWorks)
  - First digital preservation librarian: 2015-2017
  - Born Digital Preservation Lab: BitCurator and disk imaging
Ingest Goals

- Create standardized Submission Information Packages (SIPs)
- Reduce human errors/inconsistencies and increase overall efficiency.
- Facilitate content appraisal and identification of sensitive information before moving materials into longer-term storage.
- Capture information about preservation actions to ensure the authenticity and integrity of content.
Challenges

● Backlog

● Finding / retrieving legacy Submission Information Packages (SIPs) for collecting units
  ○ Lack of description
  ○ Disk images of 500 GB - 1 TB external hard drives

● Data management guidelines and storage of ‘critical’ data

● Limited IT support for BitCurator and programming
Opportunities and Considerations

- Digital forensics tools and strategies
- Metadata and format standards (esp. PREMIS and bagit)
- Opportunities for iterative improvements and interoperability
- Walsh, Sampson, Algana, Pendergrass (2018 BC Forum and forthcoming American Archivist article)
  - Emphasis on critical appraisal of content and capture procedures
- Ben Goldman (2016 PASIG and forthcoming SAA publication)
  - Authenticity and “meaningful metadata about the context and provenance of digital objects”
Influences

BitCurator Reports and PREMIS

Brunnhilde (Tim Walsh)

Disk Image Processor (Tim Walsh)
Similar Projects: National Library of the Netherlands

diskimgr

omimgr

tapeimgr
bdpl_ingest: General Approach

- Python; microservice design (includes key elements from Brunnhilde)

- Intended to facilitate the transfer and analysis of content in 4 main job types:
  - **Disk images**: use cases involving digital material stored on physical media, including 5.25" floppies, 3.5" floppies, zip disks, optical media, USB drives, and hard drives.
  - **Copy only**: use cases where disk imaging is not appropriate or where content has arrived via email, network transfer, or download.
  - **DVD**: use cases where moving image content is stored as DVD-Video on optical media.
  - **CDDA**: use cases where sound recordings are stored as Compact Disk Digital Audio on optical media.

- Collecting units:
  - Document media/individual transfers in a spreadsheet (include barcode, collection information, label transcription, notes for technician, etc.)
  - Appraisal decisions (with technical support as needed)
bdpl_ingest Interface

**Image Description:**
- The first image shows the bdpl_ingest Interface, with options for Manifest, Barcode, and Media type.
- The second image displays a command line interface for bdpl_launch.bat, with steps for file replication.

**Text from Image:**
```
STEP 1. TRANSFER CONTENT

FILE REPLICATION: TERACOPY
SOURCE: \bl-1bg-crystal.ads.iu.edu\home\micshall\Avalon
DESTINATION: \files

FILE REPLICAION COMPLETED; PROCEED TO NEXT STEP.
```
Transfer

- **Disk imaging**
  - ddrescue (production of raw images)
  - cdrdao (production of bin and cue files for CDDA use cases)

- **File replication**
  - tsk_rescue (file extraction from disk images with file systems that include ntfs, fat, exfat, hfs+, etc.)
  - unhfs (file extraction from disk images with file systems that include hfs and hfsx)
  - TeraCopy (replication of files in other use cases, including from optical media with ISO9660 or UDF file systems)

- **Normalization**
  - cdparanoia (production of single .wav and cue files for CDDA use cases)
  - ffmpeg (production of one .mpeg per title for DVD-Video use cases, with content information provided by lsdvd)
Analysis

- Virus scan
- Sensitive data scan: bulk_extractor
- Forensic feature analysis:
  - disktype (document disk image file system information)
  - fsstat (document range of metadata values and blocks/clusters)
  - ils (document allocated and unallocated inodes on the disk image)
  - mmls (document the layout of partitions on the disk image)
  - cdrdao disk-info (CDDAs) or lsdvd (DVD-Videos)
- Format identification: Siegfried
- Documentation of file directory structure: tree
- Checksum creation: fiwalk or md5deep (depending on use case)
- Scanned image(s) of physical media and packaging
Resulting Directory Structure (per barcode)

```
[/barcode]/
  __ /disk-image/ (if produced)
  __ /files/ (normalized versions of content for DVD-Video and CDDA use cases)
  __ /metadata/  
      [barcode]-dfxml.xml
      [barcode]-premis.xml
  __ /logs/
  __ /media-images/ (if produced)
  __ /reports/ (including version of Brunnhilde html report)
```
Documenting Ingest

- Log files
- Reports:
  - Siegfried format characterization
  - Brunnhilde HTML (and additional CSV reports generated from Siegfried output)
  - Tree output (directory structure)
  - Reports specific to job type (i.e., cdrdao disk-info, lsdvd, The Sleuth Kit, etc.)
- Scanned images of media
- DFXML
- PREMIS
- Spreadsheet for review/appraisal
  - Descriptive/administrative metadata (from collecting unit)
  - Technical/preservation metadata (from ingest procedures)
PREMIS Event Information

● Create a dictionary of values upon completion of each microservice:
  ○ `eventIdentifier`
    ■ Type: UUID
    ■ Value from Python uuid module
  ○ `eventType`: PREMIS Preservation Events Controlled Vocabulary
  ○ `eventDateTime`: timestamp
  ○ `eventDetail`: command line arguments
  ○ `eventOutcome`: exit code returned by tool
  ○ `eventOutcomeDetailNote`: indication of successful/failed completion
  ○ `linkingAgentIdentifier`
    ■ Implementer: Indiana University Libraries
    ■ Executing software: software and version number

● Save each dictionary to a list and write to XML with Python lxml at conclusion
PREMIS Event Information

<premis:event>
  <premis:eventIdentifier>
    <premis:eventIdentifierType>UUID</premis:eventIdentifierType>
    <premis:eventIdentifierValue>fb3fdde6-be4d-4eed-98e1-8057a84d9321</premis:eventIdentifierValue>
  </premis:eventIdentifier>
  <premis:eventType>disk image creation</premis:eventType>
  <premis:eventDateTime>2019-04-16 10:25:30.767206</premis:eventDateTime>
  <premis:eventDetailInformation>
    <premis:eventDetail>cdrdao read-cd --read-raw --session 1 --datafile
X:\disk-image\UAC2017010081-01.bin --device 0,0,0 --driver generic-mmc-raw -v 1
X:\disk-image\UAC2017010081-01.toc</premis:eventDetail>
  </premis:eventDetailInformation>
  <premis:eventOutcomeInformation>
    <premis:eventOutcome>0</premis:eventOutcome>
  </premis:eventOutcomeInformation>
</premis:event>
<table>
<thead>
<tr>
<th>Identifier</th>
<th>Transfer method</th>
<th>Migration date</th>
<th>Migration outcome</th>
<th>Migration notes</th>
<th>Extent</th>
<th>No. of files</th>
<th>No. of Duplicate Files</th>
<th>No. of Unidentified Files</th>
<th>File Formats</th>
<th>Date Range</th>
<th>Virus Status</th>
<th>PI Status</th>
<th>Appraisal results</th>
<th>Full report</th>
<th>Directory tree</th>
<th>Images of media</th>
</tr>
</thead>
</table>
Ongoing investigations...

● Additional documentation for:
  ○ Manual workarounds
  ○ Work performed by vendors (upcoming: data cartridges and tape)

● Appraisal process
  ○ Documenting separations/deaccessioning and redactions
  ○ Improving information in spreadsheet

● Potential workflow integration with:
  ○ ArchivesSpace (describe and track digital objects...and events?)
  ○ Digital preservation system (Archivematica? Preservica?)

Feedback / suggestions: micshall@iu.edu

https://github.com/IUBLibTech/bdpl_ingest