Handling a Digital Backlog and Analyzing Content in Archivematica

August 7, 2012
Research Forum
Society of American Archivists Annual Conference, Beyond Borders
San Diego, California

Courtney C. Mumma, MAS/MLIS, Systems Analyst and Archivematica Community Manager
Overview

- Action research in the Archivematica project
- Digital backlog
- Processing digital objects
- Managing transfer backlogs
- Creating SIPs in Archivematica
Action research

- To quote action research's instigator Kurt Lewin: "if you want truly to understand something, try to change it". This kind of work is not simply about changing, but also improving an environment. As John Elliott says, action research is “the study of a social situation with a view to improving the quality of action within it”

Goals of action research

- The improvement of professional practice through continual learning and progressive problem solving;

- A deep understanding of practice and the development of a well specified theory of action;

- An improvement in the community in which one's practice is embedded through participatory research.

Agile development

Individuals and interactions over processes and tools

Working software over comprehensive documentation

Customer collaboration over contract negotiation

Responding to change over following a plan

Six criteria for action research

Roles: Clarify the roles of researchers (Artefactual staff) and practitioners (our clients)

Documentation: Explain the data collection approach and how data quality is managed (Observing and documenting local workflows and requirements)

Control: Explain the control measures (standards and best practices)

Usefulness: Establish the usefulness of the findings in the problem situation (assess community response)

Frameworks: Relate the action taken as well as the findings to frameworks to support the study (build up Archivematica tools and processes)

Transferability: Explicate conditions for transfer of findings to other situations (commonalities make up default workflows in Archivematica)
Backlogs are inevitable

- Most of us have backlogs
- Most of us lack resources (tools, staff, funding) to prioritize and process those backlogs
Minimum data about the backlog

- Verify transfer compliance
- Create transfer UUID and assign file UUIDs to objects
- Verify metadata directory checksums (verifies any checksums included with the transfer)
- Assign checksums to objects
- Generate METS.xml document
- Extract packages
- Sanitize object's file and directory names
- Scan for viruses and malware
- Characterize and extract metadata
- 0.9 added transfer indexing, 1.0 accessioning as PREMIS event
Accession metadata

* PREMIS Event = Accession

  <event>
    <eventIdentifier>
      <eventIdentifierType>UUID</eventIdentifierType>
      <eventIdentifierValue>35cbe00d-d661-4174-b11a-e203f5608008</eventIdentifierValue>
    </eventIdentifier>
    <eventType>accession</eventType>
    <eventDateTime>2012-03-14</eventDateTime>
    <eventDetail>accession#2012-029</eventDetail>
    <eventOutcomeInformation>
      <eventOutcome>
        <eventOutcomeDetail>
          <eventOutcomeDetailNote></eventOutcomeDetailNote>
        </eventOutcomeDetail>
      </eventOutcome>
    </eventOutcomeInformation>
    <linkingAgentIdentifier>
      <linkingAgentIdentifierType>archivist</linkingAgentIdentifierType>
      <linkingAgentIdentifierValue>Courtney Mumma</linkingAgentIdentifierValue>
    </linkingAgentIdentifier>
  </event>
0.9 Transfer workflow

1. Allow MCP access to media or storage where transfer is located
2. Select preconfigured transfer type (generic, dspace, bagit, etc.)
3. Enter transfer name
4. Enter accession number
5. Browse to transfer and select
6. Browse to all submission documentation and select
7. Start transfer
8. Create structured transfer folder
9. Assign transfer UUID
10. Assign file UUIDs and checksums
11. Quarantine
12. Log directory structure
13. Extract packages
14. Assign file UUIDs and checksums to extracted files
15. Verify checksums included in transfer
16. Scan for viruses (generate log)
17. Reject transfer or remove infected files and continue processing

Go to next page
Transfer indexing

- Full text content
- File embedded metadata
- Formats - by folder, by transfer
- Keyword & pattern matching for privacy/security sensitive information (e.g. social insurance numbers/social security numbers, credit card numbers, email addresses security keywords like 'private', 'confidential' - find or generate domain-specific taxonomies)
- Possible reports: PDFs that have not been OCR'ed, password protected / encrypted files, duplicates with their file paths
What to do with that data

- Examine transfer(s)
- Use visualization tools and index to access transfer content
- Assign transfer(s) to appropriate intellectual levels of arrangement
- Create SIP(s) from transfer(s)
0.9 Ingest – Create SIP workflow

1. Select transfer(s) for arrangement into SIPs
2. Load selected transfer(s) into Original Order pane of Create SIP window
3. Generate reports (see transfer indexing requirements)
4. Generate file visualizations (see transfer indexing requirements)
5. Remove transfer(s) from OO pane
6. Remove selected transfer(s) from OO pane
7. Delete cached reports
8. Delete cached file visualizations
9. Select any file to view in new tab if viewer available
10. Highlight selected file
11. Grey out viewer icon if no viewer available for selected file
12. Open selected file in viewer
13. Select transfer(s) or contents
14. Highlight selected transfer(s) or contents
15. Examine transfer(s) or contents (See Examine contents mockups)
16. Delete and move files as needed
17. Generate log
Examine contents splash page

- Information about the transfer or selected file group (number of files, size, name, uuid, accession #, and?)
- Pie graph showing file type distribution overall and bargraph showing file type by folder and ordered by size)
- Clickable links: file type opens into new tab with file browser interface of all of specified format), folders opens into new tab with file browser interface of entire folder in context of rest of transfer)
- Search box to search index (opens in new tab)
- Report options (each opens in new tab): duplicates with full path locations, security keywords, CC numbers, SIN/SS#s, email addresses (with distribution graph), see password protected files with distribution (graph?)
File extension breakdown (by number of files) in transferObjects/:

- pdf: 28.2 MB
- tiff: 1.2 MB
- Jobs/Applications: 7.6 MB
- Proofs: 87 MB
- Requirements: 6 MB

File extensions by folder in transferObjects/:
Total size of objects per folder.
File viewers

- This will allow the user to see individual documents in the transfer to get a better idea of their contents and technical metadata before assigning them to SIPs.
- Viewers are browser-dependent; viewer option is greyed out if viewer is not supported in browser.
- Examine Contents window allows for viewing technical MD and other metadata available after Transfer microservices as well as indexing MD.
@archivematica (official Twitter)
@snarkivist (me)

archivematica.org (wiki with links to documentation, downloads, user group and issues list)

courtney@artefactual.com (email)