CurateGear 2016 - Notes

Authored by DigCCur 2015-16 attendees: Mark Ames, Moriah Neils Caruso, Conor Casey, Elizabeth Charlton, Blake Graham, Jeremy Heil, and Margy Jessup

Program http://ils.unc.edu/digccurr/curategear2016.html

Professional Needs and Strategies - Talks

Carolyn Hank - The Impact of Digital Curation Tools and Projects in the Classroom

- 11 syllabi compared for a core of readings
- 729 unique citations!!!!
- Data curation profiles datacurationprofiles.org

Alex Nelson - Navigating Unmountable Media with the Digital Forensics XML File System

- File systems age out of usability - operating systems can't support every files system
- Proposal:
    - treat a file listing like a file system - want to have a files system you can interact with/navigate
    - with DFXML and the disk image get file contents from disk (DFXMLFS)
- Benefits:
    - non-standard disks can be read with normal interfaces
    - no special software needed on reading machine

- Demo of DFXMLFS...
- Summary: if an operating system won't read disk, it can read the dfxml instead
- available on Github https://github.com/ajnelson/dfxmlfs

Matthew Kirschenbaum - Researching the Literary History of Word Processing: A Bucket List--Things I would have loved to have as a researcher:

- multi-institutional Collections census of digital materials
- [online] Finding aids can be v. useful, can't go everywhere, e.g. a f.a. with actual text of disk labels provided, though not technical info
- the original hardware
- the original media-- for the label, as an artifact
- disk images - many will only need to see contents but some want complete surrogate
- analytical tools; similar to a bibliographer analysing the materiality of books
- Emulation
- Lateral content - related physical materials (e.g. notebook of programmer where digital files are unavailable)

- Identification aids within collection (photograph of creator at computer; receipt for computer
- typescript vs. printout - not always easily identified
- Book: *Track Changes* (Harvard U Press) - due out soon

Susan Malsbury and Don Mennerich- Documentation of Born Digital Collections in a Centralized Processing Environment

- Why centralize? Efficient coll mgt; consistent accessioning, arrgt & desc; project planning & resource allocation; curatorial staff free to focus on other work
- NYPL - 7 curatorial units with central Archives Processing Unit
- Importance of education and outreach with 7 curators, ref staff, processing archivists re: best practices, tools avail.
- Project planning important: tracking processing, resource allocation, increased R&D esp with digital projects
- Born digital challenges/solutions
    - understanding collections; provenance
    - Donor relations- transparency and trust
        - Transfer/submission agreements (based on PAIMAS)
        - Can be a lengthy document
    - High level documentation of processes & policies
- Collection summaries - importance of collecting info and documenting while processing.
    - tracking formats, media, problems,
- Summary
    - Documentation needs to be clear
    - efficiency and brevity
    - Repurpusability

Dorothy Waugh - Teaching Files: Incorporating Born-Digital Materials into Instruction

- Currently providing access to born digital in reading room @ Emory
- second objective: teaching researchers & donors re: awesomeness of born-digital (email not so romantic)
- Will be teaching undergrad poetry class using digital materials,
- using Voyant - analytical and visualization tools
- goals
    - establish sustainable model for integration of born digital & digital humanities into instruction program
    - expose fac and students to born digital
    - explore logistical requirements and policies needed for incorporating born digital in classroom & RR
    - train staff to talk confidently with patrons re born digital; enhance support for research methods with bd in RR

Bradley Glisson - Global Positioning System Evidence: Its Impact and Implications for Digital Curation

- research is in residual data -- difficulties, risks, legal implications, tech issues
- impact on society (individuals and technology) - digital footprint
- Why interested in GPS?
    - It's everywhere
    - decisions are based on it
    - becoming more relevant in court cases
- Research findings: criminal and civil cases; focusing on transportation
- this data will eventually make it into archives, if not already

Mark Evans - Integration of Collections Management and Digital Preservation

- Current project with client: Trustees of Reservations
- Team: History Associates and Preservica
- Objectives: enhanced online and internal access; use technology to preserve
- Goals: Implement DAMS- integration of PastPerfect with other systems; Internal mgt of collections based on OAIS; ultimately online access
- Preservica core, enhanced with PastPerfect data - online access
- Approach: exporting from PP data as XML; creating PP package -> Preservica package
- Lessons: data quality; data consistency - object naming is key; buy-in, keep it simple

Data and Repository Management - Talks

Jon Crabtree and Don Sizemore – Using iRODS Policies to Support Preservation Actions

- Odum Institute, UNC
- Project: leveraging iRODS to automate preservation actions; manage rules; scalable storage; enhance security...
- configured to utilyze BitCurator tools -- e.g. testing sensitive info i.d.
- Rules in GitHub
- Next steps include stress testing sensitive info

Nancy McGovern – Digital Preservation Management Tools

- Self assessment tools and updated/new tools
- New tools include:
    - Roles & responsibilities
    - Preservation storage mgt
    - DPM action plans for each leg of 3 leg stool
- Tools on DPM Workshop website - http://www.dpworkshop.org/
- Foundations:
    - 3 Legged Stool

    - Five Stages - tools map to each of 5 stages

Erin Clary - Human Subjects Data in an Open Access Repository: Considerations and Challenges

- Dryad Digital Repository for scientific research data
- Human Subjects data may be open access after removing identifiers (direct and indirect-- SS#, age, gender...)

Data and Repository Management - Breakout Demos/Discussion

Jon Crabtree and Don Sizemore (Odum Institute Dataverse Network)– Using iRODS Policies to Support Preservation Actions

- iRODS scripts/rules/policies created to automate cleanup of data, including:
    - Move data to local server for processing
    - Verify checksums
    - documentation of preservation actions
    - utilyze BitCurator tools to identify PII
    - (etc.)...
- next steps:
    - collect/identify more extensive GPS data
    - identifying connections in PII (direct and indirect)
    - integrating Dataverse Network with larger iRods consortium database(s?)
    - flagging sensitive PII that may need restricted access, and associate with data files
    - ...
- Challenges
    - getting data from behind firewalls
    - dealing with huge amounts of output from BitCurator (in some cases)
- still testing, not yet in production mode
- Output can be viewed in user interface of Dataverse Network
- Rules in GitHub https://github.com/donsizemore/odum/tree/master/irods_rules
- Discussion
    - Risks of inadvertently publishing sensitive data
    - Importance of bringing our tools together, leveraging existing tech & community efforts so we can get on with our work

Erin Clary - Human Subjects Data in an Open Access Repository: Considerations and Challenges

- As funders, journals, publishers (ex. PLOSone) require more open data sharing, the implications of human subjects data are growing. Though some journals may provide exceptions for this kind of data given the sensitive nature.
- History of human subjects protection (human rights abuses committed in the name of science) -- INternational: Nuremberg Code, Declaration of Geneva, Helskini declaration. USA national: Institutional review board (IRB), Tuskegee study etc etc. Publishers: have ethical guidelines as well. Just because human subjects consent to participate, they don't consent to sharing their personal data.
- Then why share?
    - transparency
    - reduce duplication
    - exploration beyond original research question
    - educational
    - credit to researchers
    - scientific process!
- How to safely share
    - access control (ex ICPSR) so data can be more complete. extreme version is a data enclave where data is blocked from any manipulation (view only).
- Identifiers
    - no direct identifiers are allowed
    - indirect identifiers can span many types.
- Repository staff reviews data sets and flag problems, and return data sets to the researcher to redact
- Potential of linking out to full data set somewhere else.
- Federal "Guidance" on open data
    - https://project-open-data.cio.gov/
    - http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
- Refer to slides for great analysis of the review of data sets.

Nancy McGovern - Management Tools for Digital Curation and Preservation (DPM)

- emphasises her "good enough practice"
- to create material, she puts herself in the client's seat, what do I want/need?
- create a record of practice of digital preservation at your institution
- change in language from "archival storage" to "preservation storage" - more processes and action required
- digital preservation is not the exclusive domain of archivists

- digital archivists do not always have digital preservation responsibilities
- use gap analysis to move through the 5 stages
- Skill levels: executive / managerial / operational
- Shout-out to AVPreserve cost of inaction model - should be widened for all types of material - cost of inaction for preservation is huge
- Need to get executives to invest and fund the preservation storage managerial - strategy, planning
- operational - make the best use of what tools we have; give enough time for new skills to become embedded after training
- Nancy mentioned collaboration a lot; collaboration includes trust

New Processes and Workflows - talks

Matthew Farrell - Providing Remote Access with Docker: A Proof of Concept

- Rubenstein Library, Duke U.
- Currently providing local access on locked down machines; Aeon for online access via finding aids
- Needs read write access for all users to run properly
- Build base images - 2 required
- need command line knowledge - adding or pulling files from repository
- Example running on Ubuntu
    - needed dosbox and xpdf
    - He did go through step by step on how to create but it was a bit too detailed to capture via notes
- He also questions as to how much training should he be providing to users to access the images
- IMO - it is too technical for general use but worth watching - Mark
- No audio, currently text materials only.
- Only available in Duke reading room
- Wants to add time limits to how long the virtual machine exists - currently no limit
- Very concerned about user authentication, this would allow use outside of reading room.

Josh Schneider - Using ePADD to Process and Provide Access to Email Archives
Stanford U. Archives

- Email as potentially rich resource, not yet tapped
- java-based natural language processor for email
- Project in phase 2. next release exp'd June
- java-based, Mac/Win
- 4 modules: appraisal, processing, discovery, delivery
- query generator
- allows users to create lexicons
- https://library.stanford.edu/projects/epadd

Terrell Russell - iRODS Pluggable Rule Engine

- overview of iRODS
- Data management middleware, including plugin for rules
- rules in C++
- Coming in vsn. 4.2 Spring 2016:
    - will allow rules written in any language
    - multiple rule engines can be run simultaneously
    - rules can call on rules in other languages

Breakout Demos/Discussion: New Processes and Workflows

Josh Schneider - Using ePADD

- Appraisal Module (capabilities)
    - concatenates email address
    - visualize correspondence over time
    - identifying and processing of names/organizations (merging entities is a feature that is currently in development)
    - less flexible for batch processing
    - browse attachments, and selecting an attachments will bring the user directly to the email message
    - lexicon search, or categories of keywords within an email archive. This feature allows users to browse sentiments within the archive, such as "congratulations," "humor," "superlative," "personal," "job," etc.
    - query generator
    - export to .csv
- Processing Module
    - edit descriptive metadata about the email archive
    - restrict emails discussing "health" before exporting to delivery module
- 5-6 releases every 6 months
- looking forward to linked data, and preservation metadata

Collection Management and Description - talks

Kari Smith - Organizational Considerations for Implementing Archivematica

- MIT evaluating (sandbox and production) since 2012
- Resources: Archivematica Google Group; relationship with Artefactual Systems
- Lessons learned
    - system/storage requirements
    - Developing pipelines
    - Dependencies
    - workflows
    - Understanding content
    - Managing AIPS and DIPS - location pointers, fixity, integrating metadata, managing over time

Carl Wilson - VeraPDF and JHOVE

- VeraPDF - 9 months into the development of the project - EU funder
- partners: Open Preservation Foundation, PDF Association, and Digital Preservation Coalition
- PDF/A validation
- reports to ISO committee
- working on basic metadata repair and specific characteristics in PDF/A
- GitHub - https://github.com/verapdf
- http://verapdf.org/
- JHOVE release will have new installer and bug fixes

Brad Westbrook - Enhancing the ArchivesSpace Public Interface

- Project begun April 2015 to enhance public interface
- Team of 15 volunteers, ArchiveSpace Team members and developer
- Phase 1: User stories and work with design consultant to est new design
- Recommendations included
    - to improve searching functionality and search results
    - to improve browsing design and usability; and highlight collections
- Phase 2….

Collection Management and Description -- Demos and Discussion

Kari Smith - Organizational Considerations for Implementing Archivematica

- MIT has been an evaluation partner with Archivematica since 2012
- functions: storage service, format policy repository, base application, pipelines, dependencies on other software, and workflows

- Preservation planning - Identification, FPR, characterization, event detail, extraction (disk images) normalization, transcription, validation, verification: rules on which you want certain actions to happen.
    - Identification tools within the PP tab include FIDO and FITS
    - for issues encountered during ingest or normalization, users can discuss the issue in Google Groups (via screenshot or copying error messages)
- One pipeline (storage service) per instance of Archivematica.
- Each pipeline has its own workflow
- Dependencies: GhostScript seems to cause most errors (see discussion list)
- Institutions may modify workflows by configuring processing activities (Admin settings)
- Mary vary depending on project/materials
- MIT found that Archivematica rules and results need to be double-checked by institution to ensure that they are appropriate for preservation strategies.
- Archivists using Archivematica need to understand decisions being made (imp. of training)
- Understanding your content: MIT selected top 20 file types to focus on, ensuring appropriate actions (rules/tools/commands) for each
- Less common formats will go through unprocessed, or wait until rules are created
- MIT will share their work
- storage considerations
    - transfer - content needs to be cleared out periodically
    - backlog - accumulates until run as SIPs
    - processing - 3 or 4 times the space of the transfers
    - AIP storage - you choose! (note, if storage location changes, application must be reconfigured to point to new location)

  (can see your Processing storage usage under Admin tab)

- Ways to use Backlog in workflow: batch process multiple accessions, do preliminary SIP preparation and move to Backlog, then break them up to process further as separate AIPS/DIPS
- See Bentley Library's work with Archivematica, DSpace, ArchivesSpace integration http://archival-integration.blogspot.com/
    - Will increase appraisal/processing capabilities or archivematica, among other cool stuff.
    - Metadata goes directly to ArchiveSpace
- Next step at MIT includes exploring packaging options of AIPS/DIPS
- Google Group listserv - great resource for troubleshooting

Hosted and Distributed Services - Talks

Sam Meister and Katherine Skinner - Preservation as a Process: The MetaArchive Cooperative and Distributed Digital Preservation http://metaarchive.org/

- Distributed digital preservation (geographic distribution)
- Institutions maintain control over content

    - control curation processes, where and how content gets into the network
- Pricing - 3 membership levels
- Cooperative not a vendor (governance etc)
- Compatible with any repository system
- 12 year track record

Jack O'Sullivan - Preservica: New Access Options, Workflows and Ways to Eliminate Duplicate Work

- Suite of OAIS tools, but working now on Access improvements
- Who: Custodian (curators and archivists?), Researchers, General Public (non-expert users), other applications

Klaus Rechert - Local Usage of Emulation as a Service - Docker, Live-System and More

- EaaS/bwFLA http://bw-fla.uni-freiburg.de/
- Using Docker
- and Boot2Emulator for public displays (local emulator - boot from USB)

Hosted and Distributed Services - Breakout Demos/Discussion

Klaus Rechert - Local Usage of Emulation as a Service - Docker, Live-System and More

- Docker setup and preparation
- http://bw-fla.uni-freiburg.de/image-archive.tgz (for download)

(see slides for details)

- Live demo of bwFLA user interface
- Download USB example: http://bw-fla.uni-freiburg.de/usb-demo.img

(see slides for instructions)

- Here is a link to a blog post about install: http://bw-fla.uni-freiburg.de/wordpress/?p=844

Sam Meister - Preservation as a Process (MetaArchive)

- Stage Collection
    - content on a server
    - prepped to push into the network
- Create Manifest Page
    - simple html page with basic collection description info and links to collection content for LOCKSS
    - LOCKSS crawlers MUST find permission statement ot be able to harvest content

- Develop Collection Plugin (within Conspectus)
    - plugins tell member caches where to find a designated Manifest page and how far to follow the links to harvest
    - every plugin must have a unique name, which will define where a collection is
    - then, you can test a collection plugin. Further review and test is done by MetaArchive staff
- Make collection available
    - collection has now moved from test to the production network
    - auditing control - provides report details on when reports are made, and other integrity checking details
    - there is also functions that occur with damage and repair for archival units being replicated across the network.

Jack O'Sullivan - Preservica: New Access Options, Workflows and Ways to Eliminate Duplicate Work

- access - can set permissions in its "specialised" viewer
- creates DIP as zip file - option for audit history, though the default is not to complete
- metadata can be transferred into METS 1.9 for better readability
- contains in-browser technology
- universal access system - based on Wordpress (non-specialised access)

Media and Disk Images - talks

Doug White - Game Cartridge ROM Capture
NIST

- HowTheyGotGame.stanford.edu
- Sega Genesis and Super Nintendo cartridges
- Integrated into BitCurator Imaging tools (Guymager)
- Value added Retrode scripts - metadata locations (title, versions etc). extracted from cartridges); checksums; added speed; extensible; error/fault descriptions
- See Live Demos

Dianne Dietrich - Vintage Forensics: Wrangling HFS Metadata into DFXML
Cornell

- New Media Art preservation project 2013-15
- Metadata strategy - requirements 1st; then find tools
    - Basic metadata - DFXML
    - Additional namespace elements

- HFS not supported by SleuthKit, therefor combo of tools used (hfsutils script developed - on GitHub hfs2dfxml
- (set debug=true) ??

Kam Woods - BCA-Webtools: Accessing and Visualizing Disk Images in a Web Browser

- Web access to disk images: User friendly alternative to viewing files in Hex viewer
- No pre-processing of image - immediately displays file system
- Optional: can extract DFXML and will index file contents in background (can take awhile)

Media and Disk Images - Breakout Demos/Discussion

Dianne Dietrich - Vintage Forensics: Wrangling HFS Metadata into DFXML
Cornell

- Live demos are given regarding how to run the script (via command line interface), and the resulting DFXML
- HFS2DFXML is on Github: https://github.com/cul-it/hfs2dfxml

Doug White - Game Cartridge ROM Capture
NIST

- No live demo for Sega Genesis :(
- Just showing retrode example
  - Retrode is a commercial usb device
  - Creates immediately playable disk image
- Working from the Sega Genesis requires an extra board to be attached to motherboard for usb connection

Kam Woods - BCA-Webtools: Accessing and Visualizing Disk Images in a Web Browser

- Overview of architecture
- Note: disks are not mounted, contents not added to machine, simply displays file system in web browser
- Can select individual file to download to local machine
- Dynamically generates HTML view of file system in web browser
- HTML display is customizable (css). Selected DFXML metadata fields used by default, can add additional ones
- Option to index text files in disk image (various formats) - works in background
- Can be run on server or standalone; internet or internet
- task manager will parse out tasks depending on processing power
- Software (beta) download from Github: BitCurator/bca-webtools
- Runs on Vagrant in VirtualBox (ultimately will run on Docker (?)

- Additional instructions on access.bitcurator.net
- After startup, you work in web browser interface --pretty basic (assumption is that in most cases users will not use this interface but be sent a link to disk image)
- Disk file system loads immediately, can browse file system, open directories, and select files to download
- Optional indexing will happen in background - may take a while to complete
- Future work includes:
    - Emulation add-on for executable files
    - support for Raw image files
    - Development of interface

Observations and Implications

panel: Elizabeth Charlton, Cal Lee (moderator), and Helen Tibbo

What are your main impressions from today?

- Elizabeth, Lone Arranger at Marist Archive, New Zealand
    - themes of the day that resonated:
        - collaboration
        - open source tools (and scaling down for smaller institutions
        - online community
- Helen: Great to see not just variety, but maturing of projects and tools (since first CuratorGear)
- …

Thoughts about whether coming just to today's event is enough? (in context of other activities this week --DigCCurr, BC Users Group...)?

- Emphasis seems to be on collecting institutions (manuscript repositories) as opposed to institutional or government records with a whole set of different requirements
- CurateGear is about tools; while DigCCurr is primarily focused on big picture management, collaborating, making your case, working with people within organizations
- Cal: CurateGear grew out of expressed desire of DigCCurr attendees to see specific tools being developed; but has also matured over years as people applied their growing knowledge of digital preservation to the tools

What else would you like to see happen?

- More about access
- Policies and decision making regarding access for sensitive materials
- Integration of processing and management systems and workflows for born digital from all sources, not just disk images (ex., electronic transfers). Workflows must differ but parts of process can be integrated

- Importance of High level workflows, procedures, policies that can be applied to any formats into the future (not having to reinvent tools and workflows for every new format/media)