NDSA NE Regional Meeting, UMass Amherst, MA, Oct. 30, 2014

VIDEO ARCHIVING BREAKOUT DISCUSSION SESSION NOTES

Participants: Jennifer Mullins (Dartmouth), Jennifer Betts (Brown), Shaun Trujillo (Mt. Holyoke), Casey Davis (WGBH), Andrea Goethals (Harvard), Angelike Contis (Vermont Access Network).

We took on the DPOE model to center our discussion: *Identify, Select, Store, Protect, Manage, Provide*

IDENTIFY

Casey discussed the need to create an inventory to start with.
Several of the universities present pointed to a lack of inventory of audiovisual content. Often the catalogue of them is "in people's heads" and when they retire, the info is lost.
Increasingly, university groups produce various videos in different formats.
"We take it all," is a policy for many archivists. At Harvard, for instance, there are a lot of classroom captures. How to process and save different video formats is a challenge. There is a new video reformatting lab at Harvard, as an example of a solution.

WGBH keeps a Filemaker database, for each video file created. There is a template. Rights clearance follows after. For much older material, there is no legal documentation, but it has to be cleared to distribute archival footage.

Within some institutions, there are challenges with regards to how much storage space is available and how much can be asked of, for instance, "the analogue guy," who digitizes yet doesn't systematically store his material. There are different challenges and ways of dealing with server space and cloud storage options.

SELECT

At Harvard, it's a matter of what curators select and how much funding is available.
In Vermont public access, usually only final edited programsis preserved, but some special projects' original footage is kept, depending on staff choices.
WGBH keeps all original footage, with example of finding Obama in footage shot at Harvard, while he was a student, that didn't make the final cut of that program.

Digitization is often funding- or project- dependent. At some institutions, A/V priorities are not always very clear. Some materials are often neglected and found in basements, scattered around campus.

WGBH has successfully produced a Boston local TV news digitization-on-demand project, a paid service operating under fair use principles.

STORE

One option is LT04 tapes, with Iron Mountain backup. WGBH has shifted to LT06 tapes in a vault, with a new LT06 work station.

There are a lot of challenges with IT departments in handling uncompressed video and archival requirements.

There is the issue of how access-focused archives of video will be. What are the high-access storage systems and migrating plans?

Some companies: Crawford Media Services (www.crawford.com), Amazon Glacier, Duracloud (used for just proxy files in some cases).

PROTECT

How to built integrity checks and emergency plans into systems.

At WGBH, the MD5 checks take time; four hours for each tape. It there are 1,000 tapes…

There are some new solutions, like checks on new drives (tape T100?) that don't require removing materials. Discussion of scripts tool & frame-level checksums on RAID systems.

When it comes to storage drives, Lacie seems more stable than Seagate.

MANAGE

There is a discussion of crowd-sourced metadata, being used at some schools, with students watching material. See this Dartmouth alum-identifying link:
www.tiltfactor.org/tiltfactor-dartmouth-college-library-launch-game-to-identify-alumni

Also there is this cool site: www.metadatagames.org

Some organizations opt to hire AV Preserve (www.avpreserve.com) for this work.

EBUCore is used for technical metadata by some.

PROVIDE

Ran out of time, had to present. To be continued!

NOTES FOR DIGITAL PRESERVATION SYSTEMS AND TOOLS BREAKOUT DISCUSSION GROUP

- Impetus for the topic
    - BC just received a hard drive
        - Need to get a write blocker, integrity checking, look for idiosyncratic file names
        - Wants to put together the workflow
            - What to do in what order
            - And can you find a tool that's widely adopted, open source, and free
        - Looking for viruses – use ClamAV
            - Then what??
                - IT says erase the files
                    - So they deleted the 23 infected files out of 30,000 files
            - Transferred the files on the hard drive
            - 3 copies
                - original hard drive, inventory w/ write blocker, check for viruses using Clam AV, deleted on external hard drive prior to transfer
                - Copied to a designated Digital Preservation computer, labeled ORIGINAL DO NOT TOUCH
                - 2nd copy, labeled WORKING COPY, used for processing and isolation
- Archivematica Virus issues – do not pass go
- FRED FTK
    - Microsoft Essentials detected the Virus
        - Kept 2 copies of the files
            - 1 as the original infected
            - 1 as the use copy – extracted and cleaned
- Keeping Analog copies
    - BC keeps external hard drive
    - HBS does not keep external hard drive
        - Discussing whether or not to keep analog copy of disk images
        - Gray area with deleted files dependent on Creator's needs
- Duplicate Files
    - "Fast duplicate file finder"
        - finds each group of identical files. And will delete all but the last one
- File Naming
    - "**Renamer**": develop sophisticated rules for looking at what's in the text, removing a space, replacing an extra character, and either eliminate or replace with something else
- Access Copy

    - HBS
        - Trained archivist → deleted files not for researcher access, noted which files deleted due to restrictions, flagged for review in 50-80 years.
    - BC
        - Turn over to archivist to make deliberations
    - Bitcurator
        - You can do virus check prior to creating your disk images
        - Cool reporting tools
        - Working on a redaction script
            - It would go through and look for strings – SSNs as example – and create a clean version
    - Integrity Checking
        - Fixity: AV Preserve
            - You can schedule it to revisit your checksums and send reports
            - Can pick which algorithm to chose: MD5, SHA256
        - GitHub – Fixity Checker
            - Run natively on the Linux server that's storing
            - (BC = MetaArchive Cooperative)
                - Keep multiple copies on internal server, but also through the MetaArchive or Hathi Trust
- Resources:
    - Marty Gengenbach paper
        - http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf
    - AIMS White Paper
    - Stanford's Forensic Plan

NOTES FROM THE PRESERVING BORN-DIGITAL VS. DIGITIZED CONTENT BREAKOUT DISCUSSION SESSION

In our allotted thirty minutes, we covered a lot of ideas and personal experiences that fit into this general topic. We discussed differences in digitization for print materials (books) and analog AV material—noting that the value of the digitized product varies greatly across content type. We wondered about how to prioritize content for various levels of long-term preservation action. We considered differences between licensed born-digital library materials, research data, and various born-digital content found in archives. In the end we didn't reach specific conclusions, but we posed three questions to the overall group.

1. How do we prioritize preservation actions (selection, reformatting, processing, long-term storage, reappraisal/deselection) for digital content

(born-digital, born-digital legacy media, digitized analog AV media, digitized print/photographs)?

2. If it’s born digital, is it more valuable?
3. How do we highlight the importance of long-term digital preservation at the outset of research, object creation or digitization — rather than pushing quick for access and leaving digital preservation as an after-thought?

NOTES FROM THE LEVERAGING INTELLECTUAL DATA USING TAXONOMIES AND OTHER TOOLS BREAKOUT DISCUSSION SESSION

Most of what we talked about were the problems that each of us were having with the concept. Some of us were trying to design ontologies to use for our work with very specialized groups or sets of information, but we had problems defining the things we were using. For example we discussed the possibility of using SKOS as a way to try and join disparate taxonomies together, but that ran into the problems of the "true" definitions of words and the ability for other users to take that SKOS template and use it for themselves. We ultimately came away with more questions than answers.