The Challenges of Future Special Collections: Handling diverse media types coming from dispersed locations

Howard Besser, NYU
http://besser.tsoa.nyu.edu/howard/Talks/

WHAT MY TALK IS NOT ABOUT

The Nat’l Library is already mixing born-digital with analog…

And doing interesting things with Audio

40+ years ago, I did my UG Thesis on a Serbian filmmaker
And it would have been great to have digital access to documents & papers

But today I’m not talking about remote digital access or about Душан Макавејев

And I won’t have time to talk about the complex relationships btwn Galleries, Libraries, Archives, & Museums (GLAMs)

The Met Museum has several Libraries and an Archive

A Library can have Museum-like Exhibits (Smells at Latvian Nat’l Library, May 2017)

A Library can have Museum-like Exhibits (Smells at Latvian Nat’l Library, May 2017)
A Library can have Museum-like Exhibits (Winged books at Latvian Nat'l Library, May 2017)

 Libraries can attract school kids like Museums do (Latvian Nat'l Library, May 2017)

 An Archive can have Museum-like Exhibits (Latvian Nat'l Archive, May 2017)

GENERAL PROBLEMS OF BORN-DIGITAL SPECIAL COLLECTIONS CONTENT

What is the digital-world equivalent to this type of Spl Collectns material from Младен Секуловић?

In the analog world

- Traditionally, we have come to understand the work of writers and scientists by scholars studying their papers in Special Collections and Archives
- Their correspondence and progressively different drafts of papers reveal their changing thoughts and craft
- But how do we gather these in the Digital Age?
Where can we find this type of material created today?

- Do people write letters on paper? Can we see the iterations of changes on manuscripts?
- Where can we find today’s equivalent of these?

This will require:
- New interventions (like changing authors’ workflow, or intervening in authoring or handling software)
- New tools (like for analyzing email)
- New approaches like digital archeology, forensics

Stages of the problem

- Stage #1: People write on computers instead of paper
- Stage #2: People no longer store their digital works in places over which they have absolute control
  - Email services (gmail, yahoo)
  - Cloud storage for documents (google docs)
  - Social network services (FLICKR, YouTube)

Our Changing Environment

If we’re lucky...
Our Changing Environment

- Rise of Online Services and Social Media is changing where this content resides (and is imposing restrictions that go beyond the rightsholder)

Core Multi-location Problems

- It’s difficult enough when someone’s photos are spread throughout their hard disk. But today some photos are there, but others are on their phone(s), Instagram, Flickr, Facebook, in Tweets, etc.
- Similar problems plague email
- Most Social Network TOS policies prohibit the owner from giving their password to anyone else (even Library)

And how do we handle donations after an important person dies?

And these issues are also true for Community Grps & Assns
Imagine collecting from the Chicago Serbian Film Festival…

In addition to the “official” records, you probably also want to collect images from other sources

In addition to the “official” records, you probably also want to collect images from other sources

The issues posed by multiple sources are also true for Community Grps & Assns

- w/Social Media, group activity is more important than ever
- But each person in the group is an individual collector. And frequently a set of individual collections forms the group collection.

Audience photo of famous visitor to UCLA
Documenting Protests

Every Image Collector has a Different Approach
• Different file-naming conventions
• Different file formats
• Different compression schemes
• Different metadata
• Stored in different arrangements/hierarchies
• Stored in different places (cellphone, personal hard disk, Instagram/Facebook, ...)

Where are people struggling with these questions?

THE PDA CONFERENCES

PDA Covers born-digital
• Correspondence/email
• Personal photos and group photo albums
• Manuscript drafts
• Personal documents
• Home movies
And has been extended to encompass:
• Family history
• Community/Ethnic history & Movements
• Genealogy
• Digital humanities.

PDA: Who Attends & Presents
• Professional librarians & archivists (and their programming support staff)
• Citizen Archivists
  — People who want to step in and rescue content in peril
  — People who like to create software/Apps/Guidelines to help others facing similar problems
• Community or Ethnic groups and activists wanting to save portions of their heritage
• Regular software developers
• Researchers (both academic and computer industry)
PDA Goals—Sharing knowledge

• What worked and what didn’t; what parts turned out to be more difficult than anticipated
• New and different types of content to collect
• Guidelines, procedures, workflows, methodologies
• Software

PDA History

Initially started by Internet Archive with co-sponsorship from Netherlands Sound & Vision, LC/NDIIPP and CNI

• 2010 Internet Archive
• 2011 Internet Archive
• 2012 Internet Archive
• 2013 Univ of Maryland
• 2014 Indiana State Library
• 2015 New York University
• 2016 University of Michigan
• 2017 Stanford University

Personal Digital Archiving 2015

http://personaldigitalarchiving.com/

PDA 2015

• Preserving Digital Photos
• Digital Preservation of Art-
• Creative use of Open Source tools-
• College Library sponsored Personal Digital Archiving Days
• Community Engagement
• Video Preservation-
• Digital Humanities & Social Science
• Workshops-
Preserving Digital Photos

- Todd Wemmer, "Methods and Projects for Recording Audio Related to Family Photography" (PDF)
- Meredith Powers, "Instagram as Curation, Participation, and Documentation" (PDF)
- Melissa Condron, "Digitization & Curation of Personal Photo Collections: Recommendations, Methods & Tools from a Recent Project"
- Sarah Severson, "How I Used Flickr, Wordpress and Embedded Metadata to Build My Family Archive" (PDF)
- Rita Nix, "Methods for Organizing, Cataloging, and Developing Metadata for Family History Photos" (PDF)
- Joel Neville Anderson, "The Circulation of Domestic Portraiture Following Japan's Triple Disaster of 2011." (PDF)

Creative use of Open Source tools

- Jason Scott, "When the Emulators Broke Free." Scott's basic message was that the Internet Archive has made significant advances with emulation of old games online.
- Justin McKinney, Mark Simms-Hodges and Ashley Brewer, "FBI Analytics said BitTorrent is not inherently evil, it's just a tool, and people — especially professional archivists and librarians — should be more open minded about using BitTorrent as a viable, effective tool to transfer large (illegal) files quickly online.
- Cal Lee and Kam Woods, "Does BitTorrent + Private Trackers = The New Film Archive?" McKinney said BitTorrent is not inherently evil, it's just a tool, and people — especially professional archivists and librarians — should be more open minded about using BitTorrent as a viable, effective tool to transfer large (illegal) files quickly online.
- Cal Lee and Kam Woods, "Does BitTorrent + Private Trackers = The New Film Archive?"
- Ashley Brewer, "Don't know about you, but I'm feeling like SHA-2!"
- Peter Chan, "5.25 inch floppy disks" (PDF).

Video Preservation

- Paul Dougherty, "40 Years of Video Husbandry" (PDF)
- Zack Lischer-Katz, "Reflections on Digitizing a Personal Audiovisual Collection" (PDA)
- Rebecca Fraimow, "Digital Video Remix: A Risk Assessment." (PDF)

Workshop on Do-It-Yourself Personal Digital Archiving

Workshop on ArchiveMatica

Workshop on Email Archiving
SOME KEY IDEAS FROM ACTIVIST ARCHIVISTS

Activist Archivists

• MIAP students and grads originally working on archiving media (photos, video, audio) from the Occupy movement

How Occupy material resembles what we’ll be facing in the future

• Vast quantity of user-contributed material
• No easy way to control for quality, file format, metadata
  – no enforcing guidelines as with organizational records
  – no semi-consistency as in a single individual’s personal records
• Much of the material can most easily be found on Social Networks
• ...we need to find smart ways to harvest metadata and analyze files, as well as to influence behavior of potential contributors

Make it easy for future contributors to create “archival-friendly” works-

• Low-hanging fruit
• Easy instructional material that appeals to what they think is important

Low-Hanging fruit

• Turn GPS on
• Develop strategies for automating a profile and uploads (our ideal App)

7 Tips to Ensure Your Video Is Usable in the Long Term

• Collect details while filming
• Keep your original raw footage, unaltered
• Make your video discoverable
• Contextualize it
• Make it verifiable
• Allow others to collect and archive
• Or archive it yourself
Communicate well with your future Contributors-

- Learn to speak their language
- Help them to realize the importance of archiving

“Why Archive” video

“Why Archive” postcard

- ACCOUNTABILITY. Archives collect evidence that can hold those in power accountable.
- SELF-DETERMINATION. We define our own movement. We need to create and maintain our own historical record.
- SHARE. Archives are a point of entry to our movement’s rich record. We can use them to ensure transparency, generate discussion, and enable direct action.
- EDUCATE. Today’s videos, flyers, web-pages, and signs are material for tomorrow’s skill-shares, classes, and mobilizations.
- CONTINUITY. Just as past movements inspire us, new activists will learn from the experiences we document.
- RECORD & COLLECT what's happening around you.
- PRESERVE the record.

Find smart ways to deal with Scale-
Tamiment YouTube collecting

- Tamiment has been selectively browsing through YouTube Occupy videos, trying to choose which ones to keep, then cataloging them with
  - Title, Creator, Creation Date, Upload Date, Description, URL, Youtube Username, License, Format, Codec, Source Media, On Internet Archive, CC License type

But this won’t scale!

March 24, 2012 YouTube stats (just 6 months after start of movement)

- “#Occupy” 169,000
- “Occupy Wall Street” 98,400
- “Occupy Protest” 70,500
- “Occupy Movement” 54,800
- “#OWS” 50,300
- “Occupy Oakland” 13,400
- “Zucotti Park” 6,690

Alternative approach to YouTube Selection process

- Develop categories of important YouTube videos
  - Celebrity visits, Internal workings (library, kitchen, media), Confrontations with police, Labor, Housing, etc.
- Have Occupiers fill in an online form listing the 5 most important videos in each category

Advantages of YouTube Collaborative Filtering Selection Process

- Scalable and manageable
- Consistent with Occupy ideas of inclusiveness and of managing own story
- Tamiment can still choose to be selective in collecting only a portion of what is voted in, but the total set for review is a manageable scale
Some key ideas from Activist Archivists

- Guidelines for recorders to make their works more easily preservable: make notes, turn on GPS, upload to a service that doesn’t strip out metadata, keep raw footage, don’t compress
- For meeting recordings, have them read a script at start of recording

Tips on Outreach to Communities

- Build trust
- Speak in their language (not MARC-speak)
- Identify ways you can meet needs they already perceive
- Approach projects as collaboration whenever possible
- Don’t only focus on content and metadata, but also rights that can be an impediment to preservation

What I know from my prior work with other types of Digital Content

- InterPARES II—if we hope to preserve electronic records, archivists need to be involved early in the life-cycle of that record, long before the record enters the archive
- Preserving Digital Public Television—Pushing metadata gathering upstream into the production cycle

Some key ideas from other projects

Preserving Digital Public Television Workflow in Production Process-

- Site Visits to productions
- Interview Production staff
- Diagrams of Workflow-
It also needs to be here!

Besser, PDPTV, Austin, 3/10/11

Pushing Metadata Gathering Upstream: The Problem

TRADITIONALLY...
• Very little metadata required for preservation accompanies an object to a repository.
• Archives, libraries and other repositories must create (or re-create) most of the necessary metadata.
• This requires many manual hours, and significant resources - both time and money.

IN THE DIGITAL WORLD...
• This doesn’t scale up. Repositories will be unable to continue in this manner, as more metadata than ever is required.

Examined Potential Points of Metadata Capture

Besser, PDPTV, Austin, 3/10/11

But much of the necessary metadata has already been gathered during production
• For each element/clip, production team usually notes source, date, place, people, and other descriptive info
• But this is treated as internal information, and often various parts of the info are distributed among the personal notebooks of different production assistants
• There is seldom a central location for this info, and the info is seldom turned over to the archive (which later tries to recreate much of it)

Besser, PDPTV, Austin, 3/10/11

When the Archive tries to re-create this info, it is seldom successful

Producers know much more about the content of their productions than the archivists do. Archivists wanting accurate info must go back to the production staff (often years later) to start brainstorming over the info.

"Once the (television) program is finished, it is passed on to the archive or library for safe keeping. Librarians will catalog and classify the content, possibly using a proxy copy, and enter the resulting informative metadata in their database so they can retrieve it in the future. However, rarely if ever is the metadata from the rest of the process passed onto them, except, perhaps, for the title, tape number, and basic technical information about recording formats. It has to be re-created, with all the associated risk of errors and lack of accuracy—not to mention the work and time involved."

- Cox, Tadic, and Mulder, Descriptive Metadata for Television (2006)

We need to find ways to push metadata access upstream
• Digital requires even more metadata than Analog
  • As the workflow becomes file-based, the need for robust and accurate metadata will become critical. File relationships, video codecs, bit rates, and rights information must be explicit, accurate, and immediately accessible. This will require a much deeper level of metadata than is currently captured in tape-based archives.
  • We can’t continue to supply this metadata at ingest; that won’t scale
    • Obtaining the necessary metadata at the end of production and broadcast life cycle is not feasible. Metadata will need to be systematically gathered during the production lifecycle and submitted with the programs to the preservation repository.
Examined Potential Points for Metadata Capture

- Much of the necessary metadata for preservation is already generated by the production unit, but discarded after their internal use. This needs to be captured throughout the workflow.

- “Those in the production unit are the creators and have first hand knowledge of who, what, where, when, and why the content was created.” — Mary Ide and Leah Weisse, WGBH Archivists.

Proposed Solutions…?

- Preservation becoming a shared responsibility between content creators, distributors, curators, and preservationists.
- Partnerships are needed to come to unified solutions.
- Preservationists seek reliable metadata back upstream in the production workflow…

WorldFocus

- Nightly news program began Oct 2008
- We began working with Workflows six months before program began
- Had ability to engineer metadata gathering into the creation/production process

Lessons Learned

- Communicate well with your future Contributors
- Develop Cooperative Relationships
- Make it easy for future contributors to create “archival-friendly” works
- Try to “push metadata gathering upstream” into the creation process
- For Cooperative Projects, allow for instructions not being followed
- Find smart ways to deal with Scale
- Handle Privacy & Security responsibly

The Challenges of Future Special Collections: Handling diverse media types coming from dispersed locations

- [http://besser.tsoa.nyu.edu/howard/Talks/](http://besser.tsoa.nyu.edu/howard/Talks/)
- [https://archive.org/details/personaldigitalarchiving](https://archive.org/details/personaldigitalarchiving) or Google personaldigitalarchiving
- [https://activistarchivists.wordpress.com/](https://activistarchivists.wordpress.com/)