Projects to Collect Born-Digital Archival Material from Individuals and Community Groups

Howard Besser
Moving Image Archiving & Preservation
New York University
http://besser.tsoa.nyu.edu/howard/Talks/
http://www.nyu.edu/tsch/preservation/

Projects to Collect Born-Digital Archival Material from Individuals and Community Groups

• Background & The Problem of Personal Digital Archiving
• The PDA Conferences
• Interesting solutions and approaches to these problems; Lessons learned
  – InterPARES
  – Preserving Digital Public Television
  – Activist Archivists & the Occupy Movement

PDA Covers born-digital

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

In what environments do we find PDA material?

• Library Special Collections
• Collections documenting a community
• Collections documenting an ethnic group
• Collections documenting a social movement
• Collections documenting the work of any other type of group (a group of Architects, a set of law-makers, etc.)

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?
Alasdair Gray's Lanark (Glasgow U Library)

Correspondence

Where can we find these 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 email 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 there, but others 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

- 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.
When aggregated, many different personal collections form an important picture of:

- An ethnic group
- A community
- A social movement
- A set of architects
- A set of law-makers
- What is important to them, how they go about their business, ...

And we know from past works that aggregations create new meanings

- Aggregating all the photos of the Digital Diaspora is hugely more meaningful than a single photo
- One tweet says very little, but thousands of tweets can show trends or depict a particular event or day

But in the PDA world, aggregating items causes significant problems

- Vast quantity of user-contributed material
- No easy way to control for quality, file format, metadata (not even any consistency for any of these)-

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, ...)
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
- 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
- Professional librarians & archivists (and their programming support staff)
- 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

SOME KEY IDEAS FROM INTERPARES, PDPTV, ACTIVIST ARCHIVISTS (HTTP://ACTIVIST-ARCHIVISTS.ORG/)
What I know from my prior work with other types of Digital Content

• InterPARES—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

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.

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 into a 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 loss of accuracy—nor to mention the work and time involved.*

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

Preserving Digital Public Television Workflow in Production Process

• Site Visits to productions
• Interview Production staff
• Diagrams of Workflow

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)

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

Activist Archivists

- MIAP students and grads originally working on archiving media from the Occupy movement
- Guidelines for recorders to make their works more easily preservable: make notes, turn on GPS, upload to service that doesn’t strip out metadata, keep raw footage, don’t compress
- For meeting recordings, have them read a script at start of the recording

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

WITNESS: Activists’ Guide to Archiving Video, Yvonne Ng

http://archivangelog.witness.org/

Find smart ways to deal with Scale:

- plug-in for FireFox (downloadhelper.net)
- As of April, ca. 250 items, policy: large events
- Fair Use: 2012 ARL Code of Best Practices
  - “transformative” collecting with context
- 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
Tamiment YouTube Cataloging

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

Tips on Archiving the Born-Digital

- Monitor the work of the PDA community
- Find smart ways of dealing with Scale, as we’ll be facing a tsunami of born-digital records in the future
- Try to “push metadata gathering upstream”; get cooperation from creators to create more archivable records
Tips on Outreach to Communities

- Build trust
- Speak in their language (not DACS-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

Projects to Collect Born-Digital Archival Material from Individuals and Community Groups

- [https://archive.org/details/personaldigitalarchiving](https://archive.org/details/personaldigitalarchiving)
- [http://www.docnow.io/](http://www.docnow.io/)
- [http://besser.tsoa.nyu.edu/howard/Talks/](http://besser.tsoa.nyu.edu/howard/Talks/)