The Challenges of Archiving Personal, Community, & Event-based Born-Digital material

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

Background & The Problem of Personal, Community, & Event-based Digital Archiving (for both text and image)

The PDA Conferences

My Projects that provided key concepts & Challenges
  - InterPARES
  - Preserving Digital Public Television
  - Activist Archivists & the Occupy Movement

Important Lessons & Approaches Learned (From Activist Archivists & Other projects)

Other projects to Monitor

Further Applied Research needed

Personal born-digital (from PDA Conferences)

- Correspondence/email
- Personal photos/movies and group collections
- Manuscript drafts, camera original footage, rough cuts
- 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 this type of material?

- Archives and 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.)
- Collections documenting an event

Documenting an Event

Woodstock: The Definitive 50th Anniversary Archive (35mm + Blu-Ray)

DISC 1

Woodstock: The Definitive 50th

Anniversary Archive

DISC 2

Woodstock: The Definitive 50th

Anniversary Archive

Documenting an Event
Well-funded “anti-racist” activists, with backing from billionaire Jews and government institutions, prevent racially-aware White people from engaging in online commerce. The National Alliance and Cosmotheist Church, for example, have been effectively banned from online commerce for nearly two years. Now the Jews and their pawns are trying to extend this censorship to totally non-political venues like this Bloomington, Indiana farmers’ market.

https://nationalvanguard.org/2019/06/indiana-pro-white-farmers-threatened-persecuted-by-jewish-funded-leftists/

In the analog world

• Traditionally, we have come to understand the work of writers, scientists, filmmakers by scholars studying their papers and rough-cuts in Special Collections and Archives
• Their correspondence and progressively different drafts of papers and rough-cuts reveal their changing thoughts and craft
• But how do we gather these in the Digital Age?
Correspondence

Where can we find these today?

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

Stages of the problem

• Stage #1: People record on digital media instead of analog
• 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 (Vimeo, YouTube, Instagram)

Stage #1 Issues—digital replaces analog

• This will require
  – new interventions (like changing creators’ workflow, saving EDLs, or intervening in email handling software)
  – New tools (like for analyzing email)
  – new approaches like digital archeology, forensics

Stage #2 Issues—content no longer on hard disk

• Rise of Online Services and Social Media is changing where this content resides (and is imposing TOS restrictions that go beyond the rightsholder)
Much Content is on Facebook

Much Content is on Facebook

Much Content is on Facebook

Much Content is on Facebook

Identitarian Platform only on Web

Identitarian Platform only on Web

Identitarian Platform only on Web

Identitarian Platform only on Web

Core Multi-location Problems

• It’s difficult enough when someone’s photos or movies are spread throughout their hard disk. But today some images there, but others on their phone(s), YouTube, Vimeo, 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 how do we handle donations after an important person dies?

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

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.

And these issues are also true for Community Grps & Assns

And these issues are also true for Community Grps & Assns

And these issues are also true for Community Grps & Assns

And these issues are also true for Community Grps & Assns

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests

Documenting Protests
When aggregated, many different personal collections form an important picture of:

- An event
- 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 and home movies 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

Digital Diaspora Family Reunion

And we know from past works that aggregations create new meanings

But aggregating items from disparate sources causes significant problems

- Vast quantity of user-contributed material
- Rights Issues
- 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, YouTube, Vimeo, Facebook, ...)

THE PDA CONFERENCES

Personal Digital Archiving 2015

http://personaldigitalarchiving.com/

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 & ISU
- 2015 New York University
- 2016 University of Michigan
- 2017 Stanford University
- 2018 University of Houston
- 2019 University of Pittsburgh
Sample PDA Talks-Indianapolis 2014
http://visions.indstate.edu/pda2014/conference-program.html

- The Social Life of Personal Information,
- Defining the Personal in Digital Archiving and Communities
- Personal Archiving as a Gateway to Data Literacy
- Finding Roots, Gems, and Inspiration: Understanding the Ultimate Uses of Digital Materials
- Opportunities and Challenges in Accessing Personal Digital Archives
- Public Libraries and Personal Digital Archiving: Outreach Lessons Learned
- Clear as Glass: Controlling the Data Generated by Wearable Technology
- Personal Archiving Within the Scholarly Workflow: Zotero as Connector for Collecting Faculty Work
- Preserving Your Digital Photos Using Free or Low-Cost Software
- Collaborative Data Management and Provenance as Core Pillars of Personal Digital Curation
- Family Archive as a Narrative Organization
- Designing a Personal & Family Archive for the 21st Century

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-

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 McInroy, Ben Simon Haydn and Ashley Blewer, “Project Amiga On Modern Hardware” (PAOMH). McInroy and Simon Haydn explained that since Project Amiga On Modern Hardware is not commercially viable, they offered to work with other professionals to create an archive of Amiga software.
- Peter Chan, “5.25 inch floppy disks” (PDF).
- Chad Ragone, “Path as an Opportunity for Collaborative Advocacy and Murder Mystery Intrigue” (PDF) and “PDA as an Opportunity for Collaborative Advocacy and Murder Mystery Intrigue” (PDF) and "PDA as an Opportunity for Collaborative Advocacy and Murder Mystery Intrigue” (PDF)
- Cheyenne Sundahl, “Building Digital Collections in Omaha for The Lawvere”
- Ashley Blewer, “I don’t know about you, but I’m feeling like SHA-2!”
- Peter Chan, “Stay Away from Happy Days” (PDF).

Workshop on Do-It-Yourself Personal Digital Archiving
SOME KEY CONCEPTS FROM INTERPARES, PDPTV, ACTIVIST ARCHIVISTS

Key Concepts & Challenges 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.
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)

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.”


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.

Examined Potential Points of Metadata Capture

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

Activist Archivists

- NYU MIAP students and grads originally working on archiving media from the Occupy movement
- Guidelines both activist creators and archives
- Developed newer low-impact methods

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

Activist Archivist Website

AcIvist Archivist Website

- "Toolkit" for Occupy archiving
- Coordinating discussions among various groups archiving different parts of Occupy
- Exploring methods for obscuring identities

IMPORTANT LESSONS FROM ACTIVIST ARCHIVISTS (& OTHERS)
Lessons Learned for Archivists:

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

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.

Develop Cooperative Relationships:

- Try to better understand what their aims are; get involved in their activities
- Develop partnering relationships

Skill-shares for Occupiers

Workshop: UWS Archive Share Day

participated in Self-help activities:

- Skill-shares for Occupiers
- Workshop: UWS Archive Share Day
Self-help activities:
Other Archive Share-Day and Hackathon activities

• Batch download from FLICKR with selected attributes (#OWS, Creative Commons, EXIF metadata, tagged-text metadata)
• Re-mixing of older footage
• Creating a visual timeline
• Mining material for data (eg. number of co-locations of an officer’s name with "pepper spray")

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
• Instructions for redundant metadata collection (to make sure that it is captured)

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

Best Practices for Content Creators

• Security
  – Hidden camera laws, parties’ consent laws
• Capturing Content
  – Highest quality, set date and time-stamps, note location
• Offloading Content
  – Raw files directly onto computer, keep material organized
• Uploading Content
  – Importance of tagging, review of diff services
• Depositing with an Archive
• Copyright

Occupy Archiving Kit

• Why Archive?
• What is an "archive"? How do I create an archive?
• Creating archiving-friendly content
• How can I collect materials for the archive?
• What should I save?
• How should I organize my materials? How do I get it into the archive?
• Description/Metadata
• Media Management
• Storage & Preservation
• Access
• Exhibition and Presentation/Outreach
• Rights and Re-Use
Think Tank metadata redundancies

- Guidelines stipulate that person holding recording device will check to see that time and date stamp are correct before beginning recording (mostly didn’t happen)
- Guidelines stipulate that a script be read verbatim at the beginning of the recording, with date, time, proposed subject, etc. (and would eventually allow voice-recognition software to create appropriate metadata). Script also stated that all participants agreed to Creative Commons licensing of the recording
- Guidelines requested that date/time be embedded in the applied file-name

Find smart ways to deal with Scale-

Collecting Streaming Media
The NYU Mellon Composers Project

- Traditional Web Crawlers (Heritrix) follow links and capture most web content. But they are less successful with streaming video and dynamic content executed in the browser (like JavaScript).
- NYU collaborated with IA to create a combined crawler and browser-
Stream capture relies on Youtube-dl
https://rg3.github.io/youtube-dl/supportedsites.html

Tamiment YouTube collecting

- Tamiment Archive was 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 they didn’t realize that this wouldn’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
Handle Privacy & Security responsibly-

Police in Berlin have raided an anarchist library.

Promoting Privacy Protection
Example from WITNESS
• “ObscuraCam is a visual privacy app for photo and video, that gives you the power to better protect the identity of those captures in your photos, before you post them online”
• Developed by Guardian Project in conjunction w/Human Rights group WITNESS-

Discuss issues around commercial services with Creators/Recorders-
• Disappearance of embedded metadata from YouTube & Vimeo
• More general Rights issues
• Give archives the IP right to download

Study of metadata loss through uploading to services

UCLA Deed of Gift template

“In an effort to protect the privacy and personal safety of contributors to the Iranian Green Movement Collection, DONOR and UCLA Library agree to work together to develop methodology and an approach and will redact the email addresses or other personally identifiable information from broad public presentation.”
Even with good records, Rights Issues remain

1999 WTO Seattle Protest 20th anniversary Video Preservation

Creative Commons Guidance

- Creative Commons lets you mix-and-match four different conditions:
  - Attribution: You let others copy, re-use and distribute your video, but they must credit you.
  - Share-Alike: You let others copy, re-use and distribute your video, only if they do the same with the work they create.
  - Non-Commercial: You let others copy, re-use and distribute your video for non-commercial purposes only.
  - No Derivative Works: You let others copy and distribute your video, but not to create new works using it.
- You can use these conditions in different combinations to share your work in a controlled way. Creative Commons licenses are legal tools that depend on pre-existing copyright laws. Having a Creative Commons license on your work may give you legal recourse, but it may not actually prevent people from downloading and re-using your video illegally.

YouTube User Agreement

- 5B “You shall not download any Content unless you see a ‘download’ or similar link displayed by YouTube on the Service for that Content.”

Marking Creative Commons licenses

- There are a few ways to mark your video with a Creative Commons license. One way is to include a Creative Commons “bumper” or test card in your video. Creative Commons has created some with graphics that you can download from their website. This method is useful if your video is going to be shared offline (e.g. on DVD, live screenings), as the license information is attached to the video itself.
- Another way to mark your video with a Creative Commons license is to publish your video on platforms that are Creative Commons-enabled, such as YouTube, Vimeo, or Internet Archive. These platforms allow you to easily select a license during the upload process. This method is useful because the license is machine-readable. A search engine, for example, can detect the license.
Tips for Archivists on Outreach to Communities

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

Other Projects to monitor

- Documenting the Now
  [https://www.docnow.io/](https://www.docnow.io/)
- International Digital Ephemera Project
  [http://idep.library.ucla.edu/](http://idep.library.ucla.edu/)

Remaining Applied Research Questions

- How do deal with Scale with the tsunami of born-digital content (appraisal, description, discovery, workflow, …)
- How to avoid our collections continuing to reflect the world of the rich, well-known, and powerful
- Better methods for collecting social media
- Tension between preservation & privacy (& handling massive amounts of redaction)
- Improving public policy and TOS in areas like IP, privacy, and what happens upon death

The Challenges of Archiving Personal, Community, & Event-based Born-Digital material

- [besser.tsoa.nyu.edu/howard/Talks](besser.tsoa.nyu.edu/howard/Talks)
- [idep.library.ucla.edu/](idep.library.ucla.edu/)
- [acivist-archivists.org/](acivist-archivists.org/)
  (use Wayback)
- [facebook.com/AcivistArchivists/](facebook.com/AcivistArchivists/)
- [archive.org/details/personaldigitalarchiving](archive.org/details/personaldigitalarchiving)
- [www.docnow.io/](www.docnow.io/)