The Digital-Age Challenges of Preserving “User-generated” content: Manuscript Drafts, Correspondence, & Social Movements

Howard Besser, NYU
http://besser.tsoa.nyu.edu/howard
http://activist-archivists.org/

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?

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)

Stage #1 Issues-

• Obsolescence Problem
• Making Sense of the Donor’s desktop
• new approaches like
  – Digital archeology
  – Digital forensics

Old Digital Formats

Obsolescence—the Viewing Problem

💡 Digital Info requires a whole infrastructure to view it
💡 Each piece of that infrastructure is changing at an incredibly rapid rate
💡 How can we ever hope to deal with all the permutations and combinations

Hardware—Issues & steps

• Reading the storage device
  – Finding a device that will play it (8", 5.25", 3.5"; double/single density, zip, jazz, …)
  – Finding an interface plug to hook it to a current computer (SCSI, Firewire 1,2; USB 1,2,3)
  – Making sure that a recent operating system can recognize the device (driver)
HRC-Collections of British writer Bernard Kops, donated 2000

Dear Mr. Kops,

My name is Gabriela Rodvine, and I’m the archivist responsible for preserving your computer files at the Ransom Center. I’ve found several 3-inch compact disks among your papers. This is a format we haven’t encountered before—several collections have 3.5-inch disks, but yours is the only one with the 3-inch variety—and I’m wondering whether you remember which brand and model of computer you might have used to create these disks. I’m hoping to use a similar machine to preserve your computer files. I’d really appreciate any information you can offer.

7th March 2008

Dear Gabriela,

I’m afraid I can’t help you much. I believe the compact disks you refer to were out of usage years ago. As far as I can recall they may have been part of a now redundant system called Amstrad. A very early sort of computer that is no longer in circulation. You might be lucky to have one knocking around the junk room at Austin.

Sorry I can’t be more helpful!

Making sense of a donor’s desktop

Software—Issues & initial steps

• Create Disk Image of the storage media (so you don’t lose “hidden files”, edit histories, profiles and settings, etc.)
• Begin to explore the Directory Structure of the disk image
Making sense of a donor’s desktop

Appraisal & Preservation/Access

- Emulation/Migration just to read the files
- Digital forensics (BitCurator) to decide whether or not to save edit histories, possibly contribute to a Finding Aid, etc.
- Evaluate privacy issues to decide about redaction (more later)

Digital Forensics for Scholarly analysis

Doug Reside research on “Rent” lyrics

Word 5.1 emulator
Text Wrangler
http://www.nypl.org/blog/2011/04/22/no-day-today-look-jonathan-larsons-word-files

Stage #2 Issues

- New tools (like for analyzing email)
- What happens when a service shuts down?
- What happens when the account owner dies?

Analyzing Email
Analyzing Email

Issue of 3rd Parties & Personal records

- Emails, documents, attachments, photos with someone else’s personal information (social security number, tax returns, libelous accusations)
- Sensitive works involving this individual and others (meetings with Battered Women or victims of oppression in foreign countries, sex photos/videos)

3rd Parties & Personal records

- As number of records increases w/digital, so does the amount of personal (as opposed to corporate or governmental) and 3rd party material
- A similar increase in personal and 3rd party material comes from collecting email, chats, tweets, etc.

Protecting Privacy of Donors & 3rd Parties

- The only way to try to protect this material is:
  - De-accessioning (but sometimes this involves important records where only small parts are sensitive)
  - Embargoing until the sensitive becomes far less sensitive (ie. Named people die)
  - Redacting the sensitive parts
    - Within the archival records
    - Upon disclosing the records
  - Coding names to protect privacy, human rights, prevent defamation

Underlying Sensitive Issues—Defamation (libel & slander)

- In some jurisdictions, even non-digital oral histories constitute “published” material in terms of defamation (Calif Supreme Court ruling on Bancroft
- So, web-published oral histories (and likely other types of records) should certainly be considered “published”, and subject to defamation suits

Other Underlying Sensitive Issues—Examples (1/2)

- Human Rights
  - Genocide testimony (can result in revenge on family members)—Rwanda
  - Documentation of oppression of women (Witness in Congo, RAWA in Afghanistan)
- Laws of different countries
  - France, Germany—Nazi advocacy or revisionism
  - Turkey—Armenian records about post-WWI era using word “genocide”
Other Underlying Sensitive Issues--Examples (2/2)

• Pornography
• Right of Publicity
  - “use of a person’s name, likeness, or other personal attribute”
  - Each state has a different law on this
• Traditional Knowledge (new WIPO protections proposed)
  - tradition-based literary, artistic or scientific works; performances; inventions; scientific discoveries; designs; marks, names and symbols; undisclosed information; and all other tradition-based innovations and creations resulting from intellectual activity in the industrial, scientific, literary or artistic fields.
  - Knowledge systems, creations, innovations and cultural expressions which have generally been transmitted from generation to generation; are generally regarded as pertaining to a particular people or its territory; and, are constantly evolving in response to a changing environment.

Consider this

• The more widely-available you make a record, the more likely it is that someone will find offense in it, and the more likely that someone will challenge you in a foreign jurisdiction (with diff laws)

Privacy is part of our Ethical Standards

• “Archivists protect the privacy rights of donors and individuals or groups who are the subject of records. They respect all users’ right to privacy by maintaining the confidentiality of their research and protecting any personal information collected about them in accordance with the institution’s security procedures.”
  - http://www2.archivists.org/standards/code-of-ethics-for-archivists

What I know from my prior work

• 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.

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
Archiving Media from the “Occupy” Movement: Methods for Archives trying to manage large amounts of user-generated audiovisual media

- Why look at Occupy?
- Projects from Activist Archivists
- Projects from NYU’s Tamiment Library/Archive

What’s interesting/important about Occupy?

Occupy Wall Street (10/17/11)

Occupy Times Square (10/15/11)

Occupy Wall Street (10/11/11)

Occupy Wall Street (10/11/11)
Occupy St Paul Church (12/9/11)

Occupy Bristol (12/6/11)

Occupy Rio (11/14/11)

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

• “#Occupy” 632,089
• “Occupy Wall Street” 164,304
• “Occupy Protest” 179,454
• “Occupy Movement” 40,572
• “#OWS” 113,904
• “Occupy Oakland” 27,202
• “Zucotti Park” 9,164

What’s interesting/important about Occupy?

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
How this material may (or may not) be different

- Huge suspicion of conventional organizations (incl universities and libraries)
- DIY mentality—want to control own story
- Reluctance to sanction turning over material to an institution
- Consensus Process becomes more difficult when each meeting attracts a different set of people

Activist Archivists
http://activist-archivists.org/

- 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

OWS Archives Working Group

Archives Working Group
Posted on January 5, 2012

The Occupy Wall Street Archive is an institutionally-independent collection of historical material created and distributed in and around Liberty Plaza of ephemera, signs, and audiovisual materials.

Occupy Wall Street Archives Working Group is created to ensure that the Occupy Wall Street movement will own its past. Its mission is to keep OWS histrocally self-conscious, and guarantee that our history will be accessible to the public.

Activist Archivist Website
Activist Archivists Projects-

• “Why Archive” postcard & video
• 7 Tips to Ensure Your Video Is Usable in the Long Term
• Study of metadata loss through uploading to services
• Best Practices for Creators/Collectors
• “Toolkit” for Occupy archiving
• Coordinating discussions among various groups archiving different parts of Occupy
• Exploring methods for obscuring identities

“Why Archive” video

• 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.

“Why Archive” postcard

Why Archive Postcard

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
Study of metadata loss through uploading to services

Best Practices for Content Collectors
- Security
  - Sensitive material
  - Scraping for content
- Content Search
  - Internet Archive, Archive-It, YouTube, Vimeo, Bit-Torrent
- Receiving Content
- Metadata Extraction
- Copyright

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

Self-help activities:
Skill-shares for Occupiers
Workshop: OWS Archive Share Day

Self-help activities:
OWS Archive Share Day used variety of tools
Self-help activities:

**Bulkr**

- Batch download from FLICKR
- Set selection filters (#OWS, Creative Commons)
- Set what elements you want to grab (essence, EXIF metadata, tagged-text metadata)

Self-help activities:

**Downloaded FLICKR image**

Self-help activities:

**Downloaded EXIF metadata**

Self-help activities:

**Other Archive Share-Day and Hackathon activities**

- 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”)

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.

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.
Promoting ObscuraCam

- “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-

ObscuraCam

Activist Archivist Collaborations-

- With Occupy Wall Street groups
  - Helping store and manage Media Working Group’s media streams
- With Tamiment Collection
  - Methods for extracting metadata from recordings
  - Methods for collaborative selection of YouTube videos

Occupy’s Media Working Group—Streaming Services (1/3)

- livestream, ustream—almost impossible to hack into downloads
- One of Occupy Oakland’s streamers uses the streaming service as his fileserver.
  - does no recording onto local media (to avoid confiscation of equipment)
  - when he wants to edit, he downloads from streaming service, edits, then uploads it back
  - trusts the streaming service, and thinks that it will be around forever

Occupy’s Media Working Group—Streaming Services (2/3)

- OWS keeps masters for their streams and other media locally, and as numbers grow, they’re having trouble with managing the files and storage and want a DAM, but...
  - Activist Archivists (AA) brokered a deal to put copies of their content on a reliable archival service
  - AA helped them select open source tool to use for cataloging (Omeka), and AA volunteer is currently cataloging their mini DV collection of footage from all General Assemblies and Spokes Councils from September 17th to date—
Cataloging GA/Spokescouncil mini-DVs

Metadata for OWS Core google doc

Omeka for Cataloging/Managing

Occupy’s Media Working Group—Streaming Services (3/3)
• Worked on plan to turn over the Global Revolution streaming collection to NYU’s Tamiment
  — OWS would NOT sign a donor agreement, as they see that as conferring exclusivity to a bureaucratic organization
  — Instead, many in OWS were willing to execute a Creative Commons license letting anyone else use the material, and AA volunteers would make sure that the corpus is transferred to NYU-

AA Collaboration with Tamiment on Collecting OWS Content-
• Think Tank meeting recordings
• YouTube videos
• Other Tamiment OWS activities
  — Websites
  — Meeting notes
  — Ephemera
  — Oral/video history
  — Google, Facebook groups

Collecting – Think Tank
Collecting – Think Tank

• Daily, 2 hours
• Audio capture hardware provided by NYU library (Zoom-H2n)
• Bi-weekly digital file transfers

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

Think Tank Guidelines

Collecting – Think Tank

Tamiment YouTube collecting

• 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

Social Networks starting to police their “Terms of Use”

Facebook privacy chief fires warning shot at bosses who demand access
Social networking site warns against employers who ask for the passwords of job applicants, saying legal action is possible
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.”

But even 1916 Public Domain material on Creative Commons YouTube Channel don’t have “download” button

But using YouTube’s video editor to view and “Re-Mix” exact copies is probably legal

YouTube sees Creative Commons as only for Re-Mix, not archiving

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

“Preservation Week” event

THE ACTIVIST ARCHIVISTS

COMMUNITY ARCHIVING and the OCCUPY WALL STREET MOVEMENT

What will happen to all of the audio, video and physical material we’ve created? What will we do with it? What will it become? JOIN THE DISCUSSION

Join the Activist Archivists as we embark on a retracement of the Occupy Wall Street movement, discussing with a variety of experts the tangled challenges of archiving the political uprisings that began in New York City.

FREE AND OPEN TO THE PUBLIC
APRIL 30, 6 PM
NYU MICHIELSON THEATER
221 BROADWAY, 6TH FLOOR, ROOM 648
The Digital-Age Challenges of Preserving “Individual User-generated” content:
Manuscript Drafts, Correspondence, & Social Movements

- http://activist-archivists.org/
- http://www.digitalpreservation.gov/personalarchiving/
- http://visions.indstate.edu/pda2014/ [past PDA]

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