The Digital-Age Challenges of Preserving "Personal" Content: Manuscript Drafts, Correspondence, & Social Movements

Howard Besser; New York University; New York, NY, USA

Abstract

This paper outlines the crisis facing Archives in an age when the material they traditionally acquire is mostly available only in digital form. It discusses how the first stage (writing on computers instead of paper) was exacerbated by the 2nd stage (messages and files hosted on social networks and external services in the Cloud). Placing this in the context of previous studies advocating archivist intervention within the workflow of the creator, it discusses strategies for nudging creators to alter practices so that their works will be more preservable. The presentation will be couched within a case study of efforts to archive user-generated media related to the "Occupy" Movement.

Introduction

Content created by individuals has always composed a critical part of library Special Collections and Archives. Scholars have spent countless hours pouring over our paper collections, analyzing correspondence and manuscript version differences in order to reveal how the thoughts and practice of a scientist or an author has changed and evolved. But today's scientists and authors neither write drafts on paper, nor send paper letters to comment on each others' work. Likewise, today's social movements leave few artifacts in the form of leaflets, posters, photographic prints, and newspaper ads -- but instead leave a multitude of digital remnants in other locations.

When the Occupy Movement arose at the end of 2011, a group of Archivists affiliated with New York University’s Moving Image Archiving & Preservation masters degree program formed the group Activist Archivists in order to support the preservation of records associated with that movement. Many of their efforts followed on the findings of both the InterPARES 2 project [1] and the Preserving Digital Public Television project [2] that both found that digital preservation was most successful if archivists and librarians could slightly alter the workflow during the creation of a record’s lifecycle.

Activist Archivists explored a variety of methods for altering the workflow and metadata preservation of user-generated content from the Occupy Movement, some of which were more successful than others [3] [4]. They also developed innovative ideas for gathering and appraising relevant user-generated content from social networks and similar services.

Unlike previous reports that discussed the Activist Archivist efforts within the context of social movements, this presentation revisits those efforts and places them within the important context of Archives needing to manage user-generated content. It discusses the utility of digital archeology and digital forensics approaches. And it discusses how tools from the Personal Digital Archiving world [5] can prove useful to Archives trying to manage user-generated content

Stage One: Problems posed by writing/storing on computers instead of paper

When individuals began to widely use word processors and email to replace paper and letters sent through the mail, this posed a number of significant problems for the library Special Collections and Archives that traditionally collected the manuscripts and correspondence of scientists, authors, politicians, and other prominent individuals.

Repositories usually only receive these types of analog collections near the end of their life-cycle, often after the individual dies and almost always after the individual has ceased contributing significantly to their field. It is not unusual for a repository to receive these records more than 40 years after they were created (or even 40 years after they were last viewed).

We know that software formats are upgraded to new versions several times per decade. In the 30 years that Microsoft Word has existed, there have been more than a dozen different formats, and most versions of Microsoft Word are only able to view the two previous versions of the format. When our digital repositories begin to take in 40-year-old digital records, they will not be able to use today’s off-the-shelf software versions to view these older records. They will face significant challenges trying to view records stored in obsolete versions of software packages that are still in widespread use, and they will be even more challenged by records created by software packages that have long been abandoned (such as WordStar). Similar problems are posed by changing formats and obsolete versions of both still and moving images, and some of these pose additional problems because they also employ a variety of compression schemes.

Another problem with born-digital writings is that authors often make edits and other changes on their word processors, and save the changed document to the same file. With analog manuscripts, authors would create (and give to others to read) different numbered versions of revised manuscripts. Major changes were reflected in different version numbers, and minor changes were visible through editorial markings on the manuscript pages. Most of this evidence of version and editorial changes disappears when authors convert to all-digital workflows. Potential solutions to this problem include both training authors to save significantly changed documents to different files (document-version1.2, document-version 4.1, etc.), and the repository’s employment of digital forensics tools to view the most recent minor editing changes.

Still another problem facing both image and text records is the physical organization of folders and files. Again, this is a more acute problem in the digital world because folders can be nested into an almost infinite hierarchy, whereas most analog filing systems make it difficult to have hierarchies more than a few levels deep. And folder-naming in the digital world is usually shorter and much less explicit than file-naming in the analog world, partially due to earlier technological limits of 8 characters, and partially due to warnings about special characters in folder/file names causing problems when a file is moved to a different operating system.

All these problems vary from individual to individual. So, when a repository is ingesting records from multiple individuals, they have to worry about each one being from a different version of different software, each having different file/folder-naming conventions, each having different folder organizational schemes, … . The wide variety of different file formats and different forms of intellectual organization that a repository will need to support poses great challenges for the future.

Repositories face parallel digital-age issues in handling correspondence to what they faced handling writings, photographs, and moving images. In the analog age, repositories ingested correspondence that arrived on paper, organized in ways that the donor felt comfortable with. As paper correspondence was replaced by email, our repositories have had to develop methods for locating where the email and attachments were stored on the donated disk, and ingesting that email which might have come from a wide variety of POP-based client software applications (Eudora, Mozilla Thunderbird, Mac Mail, Windows Mail, Microsoft Outlook, …). And researchers have begun building tools to browse [7] and manage [8] collections of email. But it was often much more difficult for a repository to ingest email collections from donors who used server-based email services (like that provided by a university or other employer).

The stage-one problems provoked by potential archive donors transitioning from creating works in analog for to creating digital works in digital form took some time for special collections and archives to adapt to. But with model projects and education leading to changes in repository workflows, more and more collections began to feel like they could handle the ingestion and management of born-digital works. But as the comfort level spread through the community of collections, further technological and social adaption changes threatened to again disrupt a collection's ability to ingest the personal collections of writers, scientists, politicians, and others that have traditionally been the lifeblood of our special collections and libraries.

Stage Two: Problems posed by Social Networks and the Cloud

Stage one was characterized by common creation practices migrating from analog to digital form. Writing was done on computers instead of paper, images were captured as bits instead of as electronic pulses on video or crystals on film, correspondence and memos were sent via email instead of in envelopes or with routing slips. But for the most part, in stage one, personal collections were still under the complete control of their creator (or, in the case of received emails, under the complete control of the recipient), and after they died, under the control of the heirs.

With the rise of Social Networks and Cloud-based applications and storage, a creator and his/her heirs no longer have complete control over his or her blogs, photos, drafts, comments, etc. Under stage one, an aging donor (or their heir) could simply hand over one or several hard disks to a repository, and the repository would have acquired very similar content to what they had acquired in the analog age in the form of file folders inside boxes.

Since the rise of Social Networks and the Cloud, a significant portion of what used to be seen as personal collections are now only partially under the control of the creator, and are fully technologically under the control of companies such as Facebook, Instagram, YouTube, LinkedIn, etc. Terms of Service agreements that these companies force all users to "sign" usually forbid the user from allowing anyone else to use their login and password to access their content. In many cases, the owner of content housed on a social media or cloud storage site cannot legally give a repository their password so that the repository can download and ingest that content. And in most cases, if a donor executes a will giving their content stored on a social media or cloud storage site to a repository, the repository would still be violating a Terms of Service agreement if they tried to download the dead donor's content.

Beyond potentially violating Terms of Service agreements, repositories also face technological challenges in downloading content from social networks and cloud services. Many of the most highly used web mail services do not allow a user to download groups of emails. Other challenges range from the often closed nature of social networks, to problems of extracting the donor's content without extracting inappropriate content belonging to others, to finding appropriate storage and interfaces in order to display the extracted content within an appropriate context, etc.

Still another stage two problem is that, increasingly, what was formerly expressed in memos, annotations and comments, emails, etc. is now sent in the form of short text messages either through telephone services or social networks (both as messaging within an application such as Skype, and as full messaging social networks such as Twitter). Even if one had the cooperation of the corporate entities that own and manage these communications services, it would be very difficult to save even a portion of these; in general, the architecture and services do not support saving these messages, as they're considered ephemeral. In the future, repositories wishing to ingest selections of their donors' contributions to messaging systems may find it fruitful to execute "Freedom of Information" requests to spy agencies, who likely have done the best job of saving this type of communication.

Background on the Occupy Movement

The Occupy movement began in September 2011 in New York City, and quickly spread to cities and towns across the US (and eventually to other parts of the world). [1] The movement's key slogan – "We are the 99%" – reflects that the movement was fueled by a moral outrage at the control exerted on society by a small minority of the populace. The movement's name – "Occupy" – points to its tactic of "occupying" public physical spaces for 24 hours per day 7 days per week both to highlight the importance of those spaces to society's discourses, and to maintain a constant presence where people who pass by cannot help but notice the movement. This 24/7 presence in physical space also

[1] According to Wikipedia, by October 9, 2011 (3 weeks after its beginning in NYC), Occupy protests had taken place in over 95 cities across 82 countries, and over 600 communites in the US. (http://en.m.wikipedia.org/wiki/Occupy_movement accessed August 19, 2012)

led to the development of self-organizing and community-building within the movement itself, and is reflected in the communal feeding of large numbers of participants (numbering in the hundreds, or in the case of NYC sometimes numbering in the thousands), and in collective providing of services for all participants (in the form of lending libraries, electrical power, wireless internet services, etc.).

In addition to physically occupying key public spaces, the movement engaged in extensive large-scale demonstrations involving thousands or tens of thousands of participants. Often these demonstrations highlighted what the movement saw as particular examples of systemic problems in society – the government's bail-out of financial firms (while not rescuing the worst-off individuals), the seizure of peoples' houses via foreclosure, etc. A major characteristic of the movement was the broad creativity shown in signs carried in protest marches, and in creative street-theater, where protestors would dress as bankers or governmental officials and act out satiric scenarios.

Like the Arab Spring movement that preceded it (and inspired it), the growth of the movement was fueled by communication mediated on the Internet. But, partially because of the high level of broadband Internet access and the ubiquity of smart-phones in the US, the number of digital photographs and video and audio recordings that movement participants posted online was astounding. Statistics from the photographic posting service Flickr show that 6 months after the first Occupy demonstration, more than half a million individual photos had been posted to this service with the tag of "#Occupy".[2] Tens of thousands of individual videos were posted to YouTube during the first few months of the movement. By 6 months after the first Occupy action, 169,000 individual postings to YouTube had been tagged with "#Occupy".[3] The vast amount of content created and the dissemination through commercial websites posed interesting problems for libraries and archives interested in preserving this material.

Archiving the media related to the Occupy Movement

The material generated by the Occupy movement looks very much like the type of material that will be entering the archives and library special collections of the future. It is a vast quantity of born-digital user-generated everyday material, created by a multitude of different users [6]. There is no easy way to control for quality, file format, or metadata. Unlike most organizational collections that try to enforce standards for metadata and file formats, there are not even guidelines suggesting what schemes should be followed. And because the content comes from so many individuals, it lacks even the semi-consistency that a single individual would apply to the items that he or she creates. And what might logically constitute a future "Occupy" media collection is actually found today spread over a multitude of commercial social networks (such as Flickr, YouTube, and Facebook) that each add their own organizational idiosyncrasies, and offer no guarantee that the material will remain posted for any length of time.

So, in order to preserve this type of material, we need to find smart ways to harvest metadata and analyze files, as well as to influence the behavior of potential contributors. A number of the methods that might be useful for future user-contributed collections were explored in the projects of the Activist Archivists, which are outlined later in this paper. Many of these methods are based upon the findings in prior projects on preserving born-digital material that Activist Archivists had participated in. From the InterPARES II project (2002-2006) [1] we learned that 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. From the Preserving Digital Public Television project (http://www.thirteen.org/ptvdigitalarchive/) (2004-2010) we learned the effectiveness of automating metadata collection from the moment of first recording [2].

In response to the Occupy movement, in October 2011 students and recent graduates of NYU's Moving Image Archive and Preservation Program – MIAP (see other paper for this conference explaining MIAP in more detail) began efforts to explore the archiving and preservation of the media being generated by the Occupy movement. They felt that much of the spirit, decentralization, self-organization, playfulness, and whimsy of this protest movement would be lost to history if the media that documented this did not survive. Joined by MIAP Director Howard Besser, the group took on the name Activist Archivists, and began work on about a dozen different projects to archive the born-digital media content related to this movement (http://www.activist-archivists.org/), with most of the projects having potential impact on the archiving and preservation of all types of material that might be collected by cultural repositories in the future.

Many of the sub-projects involved collaboration with various partners. These included both collecting institutions (such as the NYU Library's Tamiment Collection) and "working groups" from the Occupy Wall Street movement (including both the "Archives" working group, which mainly dealt with collecting non-digital artifacts such as posters and signs, and the "Media" working group).

It is important to note that certain predispositions of the Occupy movement may not be relevant to libraries and archives building collections from other sources. Those in the Occupy movement were very suspicious of conventional organizations, including universities and libraries, and often needed convincing that a conventional cultural institution might be a good repository for the artifacts that they created. Occupiers could also be characterized as having a "do-it-yourself" (DIY) mentality, not wanting to rely on professionals outside their community to organize and provide access to the material. This was part of a critique of conventional media dissemination outlets which appeared to not do a good job of explaining the movement, and appeared to manipulate news coverage. The Occupiers wanted to

[2] March 24, 2012 Flickr statistics show 632,089 items tagged with "#Occupy", 164,304 tagged with "Occupy Wall Street", 179,454 tagged with "Occupy Protest", 113,904 tagged with "#OWS", 40,572 tagged with "Occupy Movement", 27,202 tagged with "Occupy Oakland", and 9,164 tagged with "Zucotti Park" (location of the first NYC occupation).

[3] March 24, 2012 YouTube statistics show 169,000 items tagged with "#Occupy", 98,400 tagged with "Occupy Wall Street", 70,500 tagged with "Occupy Protest", 50,300 tagged with "#OWS", 54,800 tagged with "Occupy Movement", 13,400 tagged with "Occupy Oakland", and 6,690 tagged with "Zucotti Park" (location of the first NYC occupation).

control their own story. Their ideology also made them suspicious of any type of exclusive arrangement, including giving their material only to a particular repository. And their consensus decision-making process made it difficult for a repository to try to come to an agreement with the group, as a group discussion on a topic such as this might range over several meetings, and each meeting might be composed of a slightly different group of participants (and discussion from previous meetings had to be repeated to and accepted by the newcomers).

Case Study: Activist Archivist projects on the Occupy Movement

A variety of the projects undertaken by Activist Archivists have been outlined elsewhere [3] [4], and the details of these will not be repeated here. Most of these projects were part of a larger goal of encouraging more consistent and systematic practices among content creators. If content contributors would limit their file format and compression choices to a few highly preservable codecs, if they were more consistent in their use of metadata, and if they saved their content within a limited set of services and structures – then the job of the repository ingesting material from multiple contributors would be more achievable. And if repositories can find ways for donors/contributors to do so with only a very minimal amount of regular and ongoing efforts, then this is much more likely to be successful.

Perhaps the most successful idea that Activist Archivists came up with to improve preservability and discoverability was to have potential contributors turn on and check both their time/date and GPS functions, and make sure that this metadata was embedded within all the images and sounds that they recorded. This is an easy, light-weight approach that only requires a minimal amount of initial set-up time on the part of the content creator (recorder), and yields critically important metadata for the repository.

Other Activist Archivist approaches that were moderately successful included guidelines that steered content creators away from compression and from proprietary file formats, and both an empirical study and guidelines that explained why placing content on the Internet Archive made it more preservable than placing it only on commercial social networks.

One idea that was discussed by Activist Archivists but never implemented was the creation of a telephone App that would allow someone recording events to pre-populate metadata to accompany their recordings. Such an App could allow the recordist to execute a Creative Commons license, embed their own name or a pseudonym as metadata into the file header, and add date/time and GPS location metadata to the digital object. It would also allow the recordist to choose which (multiple) sites on which they want to post their recording. Using this App, a person could just record something, pull up the App, either check “OK” or change some parameters, then push a button, and the result would be that the recording with extensive metadata would be sent to all relevant online services. This would also solve the problem that many recordists want to post a video on YouTube because of its wide dissemination, but a copy could also go to the Internet Archive where the metadata would not be stripped out and where it would much more likely to survive over time. Some pieces of this idea have been incorporated into the InformaCam plugin for ObsuraCam as a collaboration between the human rights group Witness and the Guardian Project. An App like this would greatly simplify a repository’s job of identifying, downloading, and ingesting material from multiple donors because the App would both insure consistency in metadata and file formats amongst a wide number of donors, and would make sure that the content was placed on a site where it was easily downloadable by the ingesting repository.

Abstract

This paper has briefly examined the movement from creating analog works to creating works in digital form. It has developed and outlined two key stages of this progression that have had a significant impact on how library special collections and archives collect and manage the types of personal collections they acquire from individual writers, scientists, politicians, as well as from collections of individuals that may constitute a community organization or movement. It has then pointed to the difficulties these transitions have posed for repositories trying to collect this type of material that is now in digital form. And it has pointed to strategies developed around preservation of media related to the Occupy Movement which could prove useful in pressing creators to alter practices so that their works will be more preservable.

References

[1] Duranti, L., and R. Preston, eds. International Research on Permanent Authentic Records in Electronic Systems (InterPARES) 2: Experiential, Interactive and Dynamic Records. Padova, Italy: Associazione Nazionale Archivistica Italiana (2008) http://www.interpares.org/ip2/ip2_index.cfm

[2] Besser, H. and K. van Malssen. “Pushing Metadata Capture Upstream into the Content Production Process: Preliminary Studies of Public Television”. DigCCurr 2007: An International Symposium in Digital Curation, 18-20 April. Chapel Hill, NC (2007)

[3] Besser, H. *Imatge I Recerca, 12es Jornades Antoni Varés 2012, Ponencies, Experiencies I Communicacions*, proceedings of conference in Girona, Catalunya, 20-23 Novembre, 2012: 106-110 (2012) (in Catalan)

[4] Besser, H. "Archiving Aggregates of Individually Created Digital Content: Lessons from Archiving the Occupy Movement", Preservation, Digital Technology & Culture. 42:1, 31-37, (2013)

[5] Personal Digital Archiving 2012 (video documentation) https://archive.org/details/personaldigitalarchiving2012pt1

[6] Besser, H. *Amateur Collections and Scholarship: Lessons from the impact of amateur collections on Cinema Studies research and on Film Archives’ practice.* Photo Archives and the Photographic Memory of Art History, Part III, March 25-26, 2011, Institute of Fine Arts, New York University, NY, NY, USA.

[7] Stanford. Project Muse (Memories Using Email), http://mobisocial.stanford.edu/muse/

[8] Hangal, S. et. al. “Providing Access to Email Archives for Historical Research,” Personal Digital Archiving 2013 http://mith.umd.edu/pda2013/schedule/

[9] Witness. http://blog.witness.org/2012/02/introducing-informacam-the-next-release-of-the-securesmartcam-project/

Author Biography

Howard Besser has been involved with digital preservation since the 1990s, has taught classes and dozens of workshops on the subject, and has published numerous articles on it. In 2009 he was named to the Library of Congress' select list of "Pioneers of Digital Preservation". He has also been involved in the creation of several library metadata standards

(PREMIS, Dublin Core, METS), and has published more than 50 articles dealing with technology and cultural institutions.