Journal of Western Archives

Volume 4 | Issue 1 Article 1

2013

Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study

Laura Wilsey
Stanford University, lauraw15@stanford.edu

Rebecca Skirvin
Stanford University, ramhist@gmail.com

Peter Chan
Stanford University, pchan3@stanford.edu

Glynn Edwards
Stanford University, gedwards@stanford.edu

Capturing and Processing Born-Digital Files in the STOP AIDS Project Records: A Case Study

Laura Wilsey
Rebecca Skirvin
Peter Chan
Glynn Edwards

ABSTRACT

In September 2012, the Manuscripts Division of the Stanford University Libraries Department of Special Collections and University Archives completed a one-year National Historical Publications and Records Commission (NHPRC)-funded project to process the records of the STOP AIDS Project, an HIV prevention non-profit organization in San Francisco, California. This project marked the department's first large-scale processing project to capture and process born-digital records. Building upon the nascent framework outlined by the AIMS white paper and the infrastructure developed by Stanford University Libraries, the project team captured born-digital records and implemented new processing strategies using digital forensics tools. This case study will document the strategies and workflows employed by the project team to capture and process the born-digital component of the STOP AIDS Project records. We will describe the successes, challenges and roadblocks encountered while forensically imaging 3.5 inch floppy disks, Zip disks, and CDs using Forensic Toolkit (FTK) Imager software. We will then outline our approach to processing nearly 30,000 unique digital files captured from the computer media using AccessData Forensic Toolkit (FTK) software, discuss our current delivery strategy, and offer some concluding thoughts.

Introduction

In September 2012, the Manuscripts Division of the Stanford University Libraries Department of Special Collections and University Archives completed a one-year National Historical Publications and Records Commission (NHPRC)-funded project to process the records of the STOP AIDS Project, an HIV prevention nonprofit organization in San Francisco, California. Like many late twentieth century archival collections, the STOP AIDS Project's records are hybrid—containing textual, audiovisual, photographic, born-digital, and graphic materials. The project team's objective was to process the entire collection, including born-digital files stored on removable computer media dating from the mid 1980s to the 2000s.

The Stanford University Libraries (SUL) had been involved in developing a born-digital program since 2009; before the grant period, the work had centered on building a forensic lab, selecting software programs, and testing various forensic capture and processing workflows on small sets of born-digital records. In addition, SUL had been participating in Born-Digital Collections: An Inter-Institutional Model for Stewardship (AIMS)[1]—a collaborative, grant-funded project focused on developing good practice for institutions charged with stewarding born-digital content.

The STOP AIDS Project records processing project marked the department's first effort to capture and process born-digital records in production mode. Building upon the nascent framework outlined by the AIMS white paper[2] and the infrastructure developed by SUL, the project team captured born-digital records and implemented new processing strategies using digital forensics tools. This case study will outline our workflows and describe the successes and challenges met while forensically imaging 3.5 inch floppy disks, Zip disks, and CDs.[3] We will also outline our approach to processing nearly 30,000 unique digital files using AccessData Forensic Toolkit (FTK) software,[4] discuss our current delivery strategy, and offer some concluding thoughts.

Project Background

The STOP AIDS Project Records

Founded in 1985, a year in which approximately 8,000 gay and bisexual men became infected with HIV in San Francisco, the STOP AIDS Project works to prevent HIV transmission among gay and bisexual men through multicultural, community-based organizing and outreach. Their programs are based on established public health, community organizing, and volunteer management principles; are free to participants; and are built using input from members of the populations they serve. Throughout its history, the STOP AIDS Project has been successful in reducing HIV transmission rates within the San Francisco gay community through innovative outreach and education programs. The STOP AIDS Project has also served as a model for community-based HIV/AIDS prevention across the nation and around the world.

1. For more information on the AIMS Project, see http://www2.lib.virginia.edu/aims/ (accessed November 6, 2012).

2. AIMS Work Group, "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship," University of Virginia, http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf (accessed November 6, 2012).

3. A handful of other computer media, including 1 mini disc, 1 flash drive and 2 5.25 in floppy disks, were present in the collection, but this paper will focus on 3.5 inch floppy disks, Zip disks, and CDs as they constituted the bulk of the media.

4. FTK is currently the industry standard in computer forensics software and is used by government agencies and law enforcement for digital investigations as well as financial institutions such as banks and credit card companies. For more information, see AccessData's website http://accessdata.com/.

On November 1, 2011, the STOP AIDS Project joined the family of HIV prevention and care programs at the San Francisco AIDS Foundation.

The STOP AIDS Project records were donated to Stanford University Libraries Special Collections in multiple accruals from 2005 to 2012 and include materials related to workshops; events and community forums; outreach activities including anonymous surveys conducted on the streets of predominantly gay neighborhoods in San Francisco; program files; and marketing materials such as posters, flyers and brochures. Also included are records documenting the STOP AIDS Project's administrative activities, including financial reports, Board of Directors meeting minutes, and grants along with administration, personnel and volunteer services files. Audio and video recordings, computer media, photographs and artwork are also present in the collection.[5]

Born-Digital Stewardship Strategy

Grant funding to process the STOP AIDS Project records came at an opportune time—the Born-Digital Program @ Stanford University Libraries[6] had the equipment and infrastructure in place to capture and process born-digital records. In addition, staff from both Special Collections and the Stanford University Libraries Digital Library Systems and Services group (DLSS)[7] was in the midst of working on the AIMS project—an Andrew W. Mellon Foundation-funded partnership between the University of Virginia Libraries, Stanford University Libraries, the University of Hull Library, and Yale University Library. This group developed the AIMS Framework to work towards best practice in stewarding born-digital material throughout its lifecycle. While broken down into four main functions including Collection Development, Accessioning, Arrangement & Description, and Discovery & Access, the Framework is meant to be flexible and customizable. As stated in the AIMS White Paper:

> The Framework is divided into four main functions that should be thought of as sequential steps in a very high-level workflow. However, it is also important to view the process as a whole. Decisions made at the beginning of the process will have a direct impact on later outcomes. Furthermore, with growing legacy collections of data on disks and servers already sitting in our stacks, the process at an individual institution may begin somewhere in the

5. The STOP AIDS Project Records finding aid can be accessed at http://www.oac.cdlib.org/findaid/ark:/13030/c8v125bx.

6. See https://lib.stanford.edu/?q=digital-forensics for more information on Stanford's Born-Digital Program.

7. See http://www-sul.stanford.edu/depts/dlss/about/index.htm for more information on DLSS.

> middle or may require moving through the functions in an order different than what is presented here.[8]

The decision was made at the outset of the grant term that the project team would follow the AIMS Framework while working on the born-digital component of the STOP AIDS Project records. This opportunity allowed project staff to become trained in and test preliminary workflows developed by SUL during the AIMS project. It also provided the team with the chance to draft workflow documentation[9] and learn from the roadblocks and challenges faced during the process. This knowledge and experience will be essential as the department processes subsequent collections containing computer media.

The discussion of our workflow below mirrors the four main functions detailed in the AIMS Framework. Our goal is to provide an outline of the steps we took while working with the born-digital material in the STOP AIDS Project records, as well as to provide insight into the challenges we faced and how we tackled them. We will also discuss some common issues we met and strategies we employed while working with specific computer media types in hopes that they may be useful to others involved in capturing and processing born-digital records.

This case study is offered in an attempt to present one repository's workflow, modeled after the AIMS Framework. It is not meant to be prescriptive, but rather one example of the hardware, software, and methods used to capture and process born-digital records from removable storage media. Our hardware and software choices were based upon extensive testing undertaken during the AIMS project and the process of piloting techniques to process born-digital records in FTK, inspired by Jeremy John's use of FTK Imager at the British Library.[10] Due to the rapid rate of technology development, our repository will likely be using the hardware and software tools as described in this paper for a limited period of time, but we believe that the functions behind the selected tools are widely applicable. Different repositories will choose tools and approaches based on their institutional frameworks, the nature of their collections, and their preferences.[11]

8. AIMS Work Group, "AIMS Born-Digital Collections," 1.

9. Our workflow documentation website is a work-in-progress, and can be found at https://sites.google.com/site/workflowdocumentation/.

10. Jeremy John, *Adapting Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools* (London, England: The British Library, 2008), http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf (accessed January 18, 2013).

11. For examples of tools used by other archival repositories see: Michael Forstrom, "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library," *American Archivist* 72, no. 2 (2009): 460-477, http://archivists.metapress.com/content/b82533tv17713471/ (accessed January 18, 2013); John A. Blythe, "Digital Dixie: Processing Born Digital Materials in the Southern Historical Collection" (Master's paper, University of North Carolina, 2009); Martin Gengenbach, "'The Way We Do It Here': Mapping Digital Forensics Workflows in Collecting

Workflow

Collection Development (Pre-Accessioning)

The first function outlined in the AIMS Framework is Collection Development (Pre-Accessioning), or the tasks that should be completed by a repository before taking in born-digital content. These tasks include conducting a preliminary survey/ assessment of the computer media and born-digital files;[12] analyzing the feasibility of taking in born-digital material (e.g. is the staffing, funding, and infrastructure in place to capture, preserve, and provide access to the files?); determining whether any enhanced curation is desired (e.g. conducting an oral history or photographing the creator's work space); drafting a legal agreement that addresses issues such as restricted material and copyright; and preparing the materials for transfer to the repository.

While performing a detailed assessment of born-digital content before receiving a collection is ideal, it is not always feasible. Like many collections, the STOP AIDS Project records were a legacy collection. A born-digital program was not in place at the time the collection was first donated to Stanford University Libraries in 2005. Therefore, an assessment of the computer media was not done at the point of negotiation, appraisal, or transfer. Only cursory, box-level inventories existed for the multiple accruals of the collection, and a rough estimate of 210 pieces of removable computer media existed at the start of the grant term.

Although the project team did not have the opportunity to perform pre-accessioning tasks on the legacy computer media, during the grant period we worked closely with the donor and their IT contractor to transfer additional born-digital material. In this way, we had the chance to test out the pre-accessioning tasks outlined in the AIMS Framework. By meeting with the donor, we gained insight into the content and extent of the files, and determined that ingesting the data was feasible since we had the equipment, staffing, and technical infrastructure in place. We determined that no additional enhanced curation was necessary, and were able to discuss potentially restricted files as well as delivery options with the donor.

The process of working closely with the donor proved to be integral in acquiring and processing both the analog and digital portions of the collection. SUL project staff met regularly with the creators throughout the grant term to discuss the unique

Institutions" (Master's paper, University of North Carolina, 2012), http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf (accessed January 18, 2013); Catherine Stollar and Thomas Kiehne, "Guarding the Guards: Archiving the Electronic Records of Hypertext Author Michael Joyce," New Skills for the Digital Era, Case Study 4 (paper presented at Society of American Archivists colloquium New Skills for a Digital Era, May 31, 2006–June 2, 2006), http://www.azlibrary.gov/diggovt/documents/pdf/4_Stollar_Kiehne.pdf (accessed January 18, 2013).

12. For guidance on possible questions to ask donors during the Collection Development (Pre-Accessioning) phase, see Appendix F of the AIMS Digital Material Survey at http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appF.pdf.

aspects of the textual, audiovisual, and digital files in the collection. Regardless of format, the privacy and confidentiality of their program participants and volunteers was of utmost importance to the donors. While the collection did not contain medical records, it did include records with the name, contact information, date of birth, and sexual orientation identification of program participants. In addition, because an important aspect of the STOP AIDS Project's programming is targeting specific subsets of the gay and bisexual community, a participant's connection with a certain program may identify them as being associated with certain fetishes or subcultures, or expose their HIV status. [13]

Our work with the donors to identify sensitive information in the textual records gave them the confidence that we would be able to identify and restrict similar records in the born-digital files. We also had the opportunity to discuss in detail the infrastructure and workflows we had in place to capture and process their born-digital files. The fact that the infrastructure and processes were already in place to deal with these records gave them the confidence that we would be able to screen the files for restricted material before delivery. Because of this trust, the donor felt comfortable selecting and transferring an additional 836 gigabytes of data stored on STOP AIDS Project computers to our repository during the grant term.

The benefits of working closely with the donor strengthened our resolve to collaborate with creators while processing legacy collections of born-digital material. The opportunity to communicate with the donor provided us with key information about the STOP AIDS Project's technical infrastructure. We regularly asked questions related to the computer media donated in the early accessions during the imaging process, such as how their disks were typically formatted, whether they used PCs or Macs, and if they conducted routine backups of their data.

This process also strengthened our resolve to work closely with creators and/or donors to gain as much information as possible about born-digital materials when receiving new collections. For various reasons, donors may become unavailable between the time a collection is donated and when it is processed. Performing a detailed assessment of born-digital content upfront will reduce the likelihood of a donor becoming unreachable and unable to answer key questions that may be the difference between born-digital material being captured and preserved or irrevocably lost.

Accessioning

The process of accessioning born-digital content under the AIMS Framework is multifaceted. It ensures that adequate physical control exists over the files and that

13. See slides from Project Archivist Laura Wilsey's presentation at the 2012 Society of California Archivists Annual General Meeting entitled "To Restrict or Not to Restrict: Balancing Access, Privacy and Confidentiality in the STOP AIDS Project Records;" available at http://www.calarchivists.org/Resources/Documents/AGM_Past/2012_AGM_presentation_session-07_Williams.pdf.

virus-free disk images are made of the computer media. At SUL, physical control entails counting, labeling, and assigning a unique ID to the physical media or gaining control through virtual transfer, such as copying files from a network drive. Creating disk images involves undertaking slightly different processes for each type of computer media. In our case the bulk of the media and main differences we discovered were between floppy and Zip disks on one hand and CDs on the other. These steps are discussed in detail below.

Establishing Physical Control

The first action we took towards accessioning the born-digital component of the STOP AIDS Project records was to locate the estimated 210 items of computer media scattered throughout over 300 record storage boxes of collection material. In doing so, we discovered almost twice the number of computer media in the collection than anticipated—an argument for closely appraising born-digital content at the point of transfer. When computer media was housed with textual material, we created a separation sheet linking the two, and separated the computer media for cataloging and imaging.

Next, we labeled each item with a unique ID number using the convention "Collection Call No._CMxxx" (e.g. M1463_CM001). We placed the labels directly on the front of floppy and Zip disks in a way that did not obscure any writing on the media label. For CDs, we placed the labels in the upper left or right corner of the CD sleeve so that when the CD was photographed in the sleeve, any writing on the actual CD was not obscured.

We used an Excel spreadsheet to catalog the computer media to the item level and track it throughout the accessioning process. This log was based on a prototype developed by the SUL Digital Archivist during the AIMS project. It allows the department to track statistics such as loss rates for various types of media and track the progress of an item throughout the accessioning process. Information documented for each piece of computer media includes the unique identification number; media type; manufacturer; folder title (if the item was removed from a collection folder); a transcription of any metadata on the media case and/or label; whether the disk is formatted for a PC, Apple or undetermined; virus scan status; whether disk imaging was successful or unsuccessful; the earliest modification and/or creation year for the files on the disk; the date the disk was imaged; the staff member that imaged the disk; and the type of computer used for capture.

During the project, we determined that instead of using a separate spreadsheet to track computer media for each collection, it would be preferable to build a FileMaker Pro database. Using this database would allow us to track multiple projects and aggregate statistics across collections. It would also allow us to track metrics for capture. We designed the Stanford Computer Media Log database in FileMaker Pro based on the fields in the Excel media log, and imported the data. Future processing projects involving the capture of computer media will use this database.

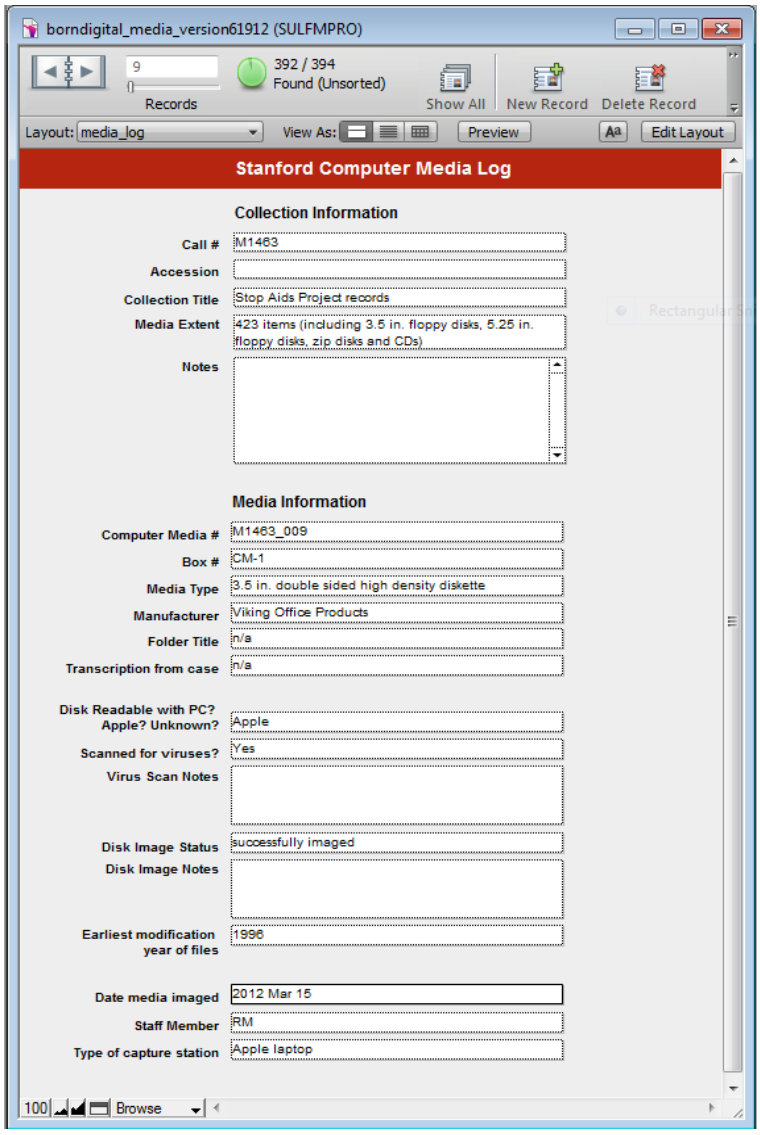

Figure 1. Screen capture of Stanford Computer Media Log

Photographing Computer Media

Once included in the Stanford Computer Media Log, we photographed the computer media. These photos serve not only as an image of the artifact itself but also as a record of any metadata recorded on the original media label. While sometimes cryptic or irrelevant (for example, if media has been overwritten), the media label may be the only external hint of the creator's organization.

We set up a Canon EOS T1i digital camera using a camera stand, added a ruler into the shot for scale, and saved each digital image with the item's unique ID as the filename. We found the process of taking photos of the computer media to be fairly quick—the first image in a shoot took approximately 2-3 minutes because of the need to focus the camera, take a test shot and adjust the media accordingly. After that, it took less than a minute per item. It was most efficient to have a batch of media labeled and ready to photograph in a single session.

Figure 2. Digital photograph of a 3.5 inch floppy disk

Virus Scanning

Screening born-digital files for viruses mitigates the risk of infecting computers used to open the files. Our virus-scanning workflow for this project involved first checking to see if the media was readable. For readable floppy and Zip disks we ran Sophos[14] antivirus software and removed any viruses before continuing to the disk imaging phase. We removed any viruses from the data captured off of the CDs and hard drive after the disk images of these media were created.

Due to inconsistencies in workflows between media types and a number of issues encountered during virus scanning, we will be revising our workflow. From now on, we will image the media, process the captured files, and run the virus software only on the files that will be exported to the Fedora-based Stanford Digital Repository

14. For more information on Sophos software, see http://www.sophos.com/en-us/.

(SDR).[15] If any viruses are found, they will be cleaned before accessioning into the digital repository and making them available for research. Running a virus scan in this way will streamline the workflow and ensure that the files researchers open will not infect their computers.

Imaging

Once any viruses were removed, we began capturing files off of the legacy computer media. To do this, we used FTK Imager, a freely-downloadable software utility by AccessData.[16] After researching several options, Special Collections decided to implement AccessData's software products for capture and processing computer media for a number of reasons. First, the products were originally developed for forensic analysis of digital storage media by law enforcement and therefore have safeguards in place to ensure the authenticity and integrity of born-digital files. FTK Imager provides automatic disk image verification with two checksums and is compatible with the powerful processing capabilities provided by Forensic ToolKit (FTK). In addition, the PC version of FTK Imager employs an easy-to-use Graphical User Interface (GUI) and does not require processors to run the imaging software using a command line interface.

We used a wide variety of computer hardware to create disk images of the STOP AIDS Project computer media. These included FRED (Forensic Recovery of Evidence Device),[17] a powerful Windows computer equipped with an integrated write-blocker that ensures that no data will be written to the source drive; three PCs running Windows 7 and their built-in DVD drives; a Macintosh laptop and its built in DVD drive; a Digital Intelligence Forensic portable USB 3.5 inch floppy disk drive; a Fujitsu portable USB 3.5 inch floppy disk drive; and an iOmega portable USB Zip disk drive.[18]

While we chose the hardware and software tools outlined above for a number of reasons including prior testing during the AIMS project, subsequent forensic lab build out, and existing infrastructure, other repositories may choose other tools to perform the same functions. While a detailed description of available options is outside the scope of the present case study, a number of excellent papers and blogs

15. For more information on the Stanford Digital Repository, see https://lib.stanford.edu/?q=sdr.

16. The free download can be found at http://accessdata.com/support/product-downloads.

17. For more information on FRED, see http://www.digitalintelligence.com/products/fred/.

18. For more information on how and why we chose these tools for imaging, see this blog post http://lib.stanford.edu/digital-forensics-stanford/processing-born-digital-materials-stop-aids-project-records-imaging and the short case study on the born-digital records in the Stephen Jay Gould papers in Appendix E of the AIMS white paper http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appE.pdf.

outline the various options available to repositories embarking on the capture and processing of born-digital files.[19]

Imaging Philosophy

Before imaging, we needed to think carefully about our approach. Should we attempt to capture every bit of every removable storage media? Would we be satisfied with imaging most of the media? In the end we decided to attempt to capture media that was readable by Mac or PC computers with the hardware (such as external floppy disk and Zip drives) that we already had in place. We chose to attempt reading each disk, but did not attempt to diagnose why a particular disk wasn't readable. Under a tight deadline, we chose to follow this philosophy so that we could image as many readable disks as possible without spending too much time on troubleshooting.

We also needed to determine whether we wanted to perform a forensic (i.e. disk) or logical image capture of the computer media. A forensic image is a bit-by-bit copy of a storage medium or device, such as a hard drive, SSD (solid state drive), tape drive, floppy disk, optical disc, or flash memory device. The image can be stored in one or more files. Deleted files, if any, may be copied in this process. A logical disk image is a copy of the files in the directories or folders specified during the imaging process. During this process, the full path of each file is recorded and the files are embedded in one or more files in AD1 format (a file format for logical disk images). Since deleted files and unpartitioned space are not represented in a directory, they are not copied in the process. Thus, the imaging process is quicker, and not as much storage space is required. We decided to create forensic images of the removable storage media but forego recovering any deleted files during processing and to create a logical disk image of the external hard drive received during the grant term.

Imaging Floppy and Zip Disks

We discovered early on in the imaging process that Apple-formatted floppy and Zip disks could not be imaged using FTK Imager in a Windows platform. Therefore, two different processes for these types of disks were required. For PC-formatted disks, we used FTK Imager's GUI, which allowed us to select the type of image to be made (forensic or logical), the format, and add notes to the text file that is generated with each image. For Apple-formatted disks, we used FTK Imager's command line terminal program. The image commands included assigning a name to the image file based on

19. For a comparison of forensic hardware and software, see: Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, DC: Council on Library and Information Resources, 2010), http://www.clir.org/pubs/abstract/reports/pub149 (accessed January 18, 2013). For information on a range of available digital forensics tools, see: Appendix G of the AIMS White Paper, available at http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_appG.pdf, which contains descriptions and evaluations of tools available for accessioning and processing born-digital files including Curator's Workbench, Karen's Directory Printer, a comparison of 5.25 inch floppy disk drive solutions, AccessData FTK Imager 3.0 and AccessData FTK 3.3. Chris Prom's Practical E-Records Blog at http://e

the unique ID, mounting and unmounting the disk, and running two checksums to ensure that the original files and the disk images were identical.

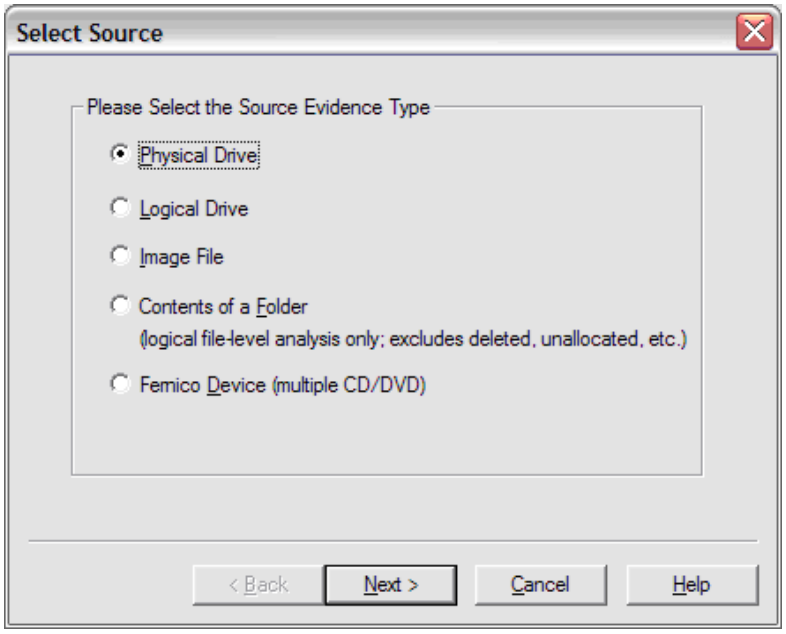

Figure 3. FTK Imager's GUI for PC computers

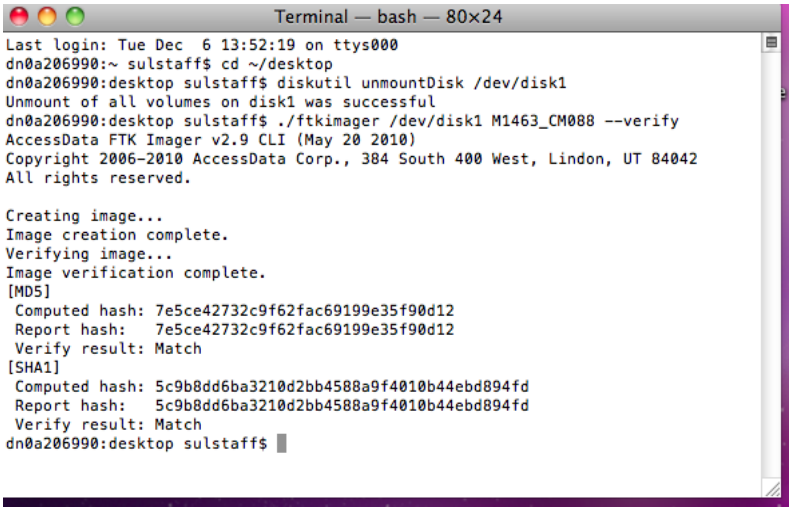

Figure 4. FTK Imager's command line interface for Mac computers

Discounting the amount of time it took to determine whether a disk was readable within a particular drive (and if not, trying it in others), we found that it took

approximately 1.5 minutes to image 3.5 inch floppy disks, and about 6 minutes to image Zip disks. However, it should be noted that imaging time is dependent on how much data is stored on the disk; more data means a longer amount of time to image the disk.

Overall, we found that the imaging success rate for Zip disks was very high—96% or 22 out of 23 total Zip disks imaged successfully. 6 out of 6 PC-formatted Zip disks imaged successfully (100%), and 16 out of 17 Apple-formatted Zip disks imaged successfully (94%).

Formatting	**Total number of disks**	**Number of disks successfully imaged**	**Success rate**
PC	6	6	100%
Apple	17	16	94%

Table 1. Success rates for imaging Apple and PC-formatted Zip disks

The imaging success rate for 3.5 inch floppy disks was lower than that of the Zip disks. 60% or 109 out of 183 disks were imaged successfully. Formatting was known for 123 disks (94 Apple and 29 PC). Out of these, 83 Apple and 26 PC disks were imaged, bringing the imaging success rate for Apple-formatted disks to 88% and for PC-formatted disks to 90%.

Formatting	**Total number of disks**	**Number of disks successfully imaged**	**Success rate**
PC	29	26	90%
Apple	94	83	88%

Table 2. Success rate for imaging Apple and PC-formatted 3.5 inch disks

Considering our experiences, we formulated the following list of common issues and lessons learned through the process of imaging floppy and Zip disks:

- Because of the way floppy and Zip disks are formatted, only Apple computers can read disks formatted for Macs, and only Windows PCs can read disks formatted for Windows.

- Some external floppy disk drives may be more compatible with certain systems. For example, we found the Digital Intelligence Forensic Floppy Drive worked well with both Windows and Mac-formatted disks, while a more generic Fujitsu drive worked best with Windows.

- Write protection can occur in a number of ways, and is an important step in imaging so that files are not unintentionally altered. We used two 3.5 inch floppy disk drives to image disks during the project; one had a write blocker but the other did not. Working with the drive that included the write blocker was straightforward—before imaging disks, we made sure the write blocker was set to "read only." When working with the drive that did not have a write blocker, we had to be sure that the read/write tab (in the upper left hand corner of the back of the disk) was set to the "read only" position.

- Disks are not always labeled as Apple or PC-formatted, and experimentation with different drives to see which can read the disk is often required. In addition, one PC machine may not be able to read what another PC machine can. Thus, it is helpful to have a number of computers available for testing and imaging disks.

- There are times when a disk simply does not read. There will be no error message on Apple computers, the disk icon simply will not appear on the desktop. On PCs, a dialog box will inform you that the disk is either not formatted or is formatted incorrectly, and it will ask if you want to reformat the disk.

- There are instances when a disk is readable but does not image. Unfortunately, no error message appears when the disk does not image properly. With experience, it may be possible to determine by sound whether the drive is trying to read the disk, or is just caught up and spinning.

- An unsuccessful imaging attempt may be directly followed by a successful attempt while using the same disk on the same machine. It is worth attempting to image disks multiple times if not successful on the first try.

Imaging CDs

Unlike floppy and Zip disks, CDs all use the same file formatting system (Compact Disk File System, in accordance with ISO 9660). Thus, we were able to use FTK Imager on PC machines to image all readable CDs in the STOP AIDS Project records. On average, it took about five minutes to image each CD.

While we had a relatively high success rates for imaging 3.5 inch and Zip disks, we faced some unforeseen issues when imaging the CDs. Out of 218 CDs, only 18 were imaged, yielding an 8% success rate. We attempted to read each CD on both PC and Mac machines. The unsuccessful disks would simply not image on Macs, and on PCs the disk image would be empty. We undertook three separate rounds of unsuccessful imaging for the CDs.

Total number of CDs	**Number of CDs successfully imaged**	**Success rate**
218	18	8%

Table 3. Success rate for imaging CDs

Although we are unsure of the exact reason for the low CD capture rate, we were able to posit some explanations. The majority of CDs in the collection appeared to have been burned in-house by STOP AIDS Project staff. Notations on the cases hinted at a large portion of the CDs consisting of backups of computers on different floors of their main office building. It is possible that at the time the disks were burned, the data was not written successfully. It is also possible that the low quality of the discs themselves or poor CD burning hardware may have caused the media to degrade and become unreadable over time. While we may not be certain of the reason for the low capture rate, our experience strengthens the literature asserting that recordable optical discs are an unreliable storage media.[20]

Imaging Hard Drives

An external hard drive containing 836 gigabytes of STOP AIDS Project computer files was transferred to Stanford University Libraries during the grant project. Since the files were selected and transferred by the donor (as opposed to giving us a hard drive from one of their machines used for business), we chose to perform a logical disk image, as a bit-by-bit copy of the entire external drive was not needed.

We found the average capture rate to be approximately 1 gigabyte every 4 minutes. It took just under 24 hours for FTK Imager to create a logical image of the

-records.chrisprom.com/resources/software/ also contains reviews of several software options for this type of work.

20. Kevin Bradley, "Risks Associated with the use of Recordable CDs and DVDs as Reliable Storage Media in Archival Collections—Strategies and Alternatives" (Paper presented at the UNESCO Memory of the World Programme Sub-Committee on Technology, Paris, France, October, 2006), http://unesdoc.unesco.org/images/0014/001477/147782e.pdf (accessed January 18, 2013); Daniel P. Wells,

hard drive, and another 21 hours for FTK Imager to verify the image. While it took approximately 45 hours to complete the entire imaging process, it only took about 5 minutes of staff time to set up the specifications for imaging the drive; the rest of the process was automated.

Similar to floppy and Zip disks, hard drives are format-dependent (e.g. a PC-formatted hard drive cannot be imaged on a Mac computer). We also were confronted with a number of viruses during the Sophos scan. The Sophos software cleaned some of these, but others had to be manually removed by the Digital Archivist. In addition, we were struck by how time-consuming the imaging process was. This was compounded by a number of unsuccessful imaging attempts caused by spontaneous computer restarts and network outages. We ended up imaging the external hard drive to a local drive rather than a network drive to mitigate these pitfalls.

Outcomes

Overall, the project team completed item level cataloging, labeling, and photographing of 423 pieces of computer media (about double the number estimated in the grant proposal). We were able to successfully read and image 22 Zip disks, 109 3.5 inch floppy disks, and 18 CDs, yielding nearly 100,500 files captured off of the computer media. After filtering for duplicates, the number of unique files totaled 29,423. The disk images, along with the photographs of the media, will be ingested into the SDR for long-term preservation.

Media type	**Success rate**
Zip disk	96%
3.5 in. floppy disk	60%
CDs	8%

Table 4. Imaging success rates by media type

Arrangement and Description

At the outset of the project, we had an amorphous idea of what processing the born-digital files in the collection would entail. Processing efforts by the SUL Digital Archivist on much smaller born-digital collections during the AIMS project included filtering out potentially sensitive material and adding tags and bookmarks in FTK to groups of files. In our case, we knew that we would be dealing with a large corpus of digital files under a hard deadline, and ended up modifying our processing strategy as we progressed through the project.[21]

After loading the disk images created during the accessioning process into FTK, we began by searching the files for potentially sensitive information. We chose to do this first so that we could filter out the restricted files and have fewer files to process for delivery afterward. Our methodology for identifying restricted information involved running two powerful FTK searches: Live Searches and Index Searches.

The Live Search function in FTK allows the user to specify terms or patterns, and then conducts a search of the files for those terms or patterns. The patterns that users can search for are called "regular expressions" or common patterns of numbers or text that can be boiled down to computational abbreviations. FTK can then "match" portions of text to a regular expression, which then show up as search results. For example, social security numbers follow a particular pattern of three numbers, a space, two numbers, a space, and four numbers. FTK can run a search for this pattern and return files with the specified pattern highlighted within each document.

We ran two Live Searches on the STOP AIDS Project born-digital files: one to identify social security numbers and the other to identify credit card numbers. Project staff sorted through the returned files to determine which ones actually contained social security and credit card numbers and which contained false positives. Since there is currently no way to redact information within an individual file in FTK, we used the flagging feature to mark the files containing social security numbers and credit card numbers as "privileged." When exporting the files from FTK, the processor can filter out files flagged as "privileged" so that they will not be available for viewing by the public. Each search returned less than 100 files. Upon review, credit card numbers were found in a handful of files and social security numbers were found in fewer than 20.

The Index Search function was also useful in conducting searches for potentially sensitive information. Because participation in certain STOP AIDS Project programs might imply not only sexual orientation but also HIV status, we wanted to restrict all files containing personally identifiable information of program participants, such as

"Predicting the Longevity of DVD-R Media by Periodic Analysis of Parity, Jitter, and ECC Performance Parameters" (Master's paper, Brigham Young University, 2008), http://contentdm.lib.byu.edu/cdm/ref/collection/ETD/id/1534 (accessed January 18, 2013).

name, address, telephone number, date of birth, gender identification and sexual orientation identification. We also wanted to restrict personally identifiable information about STOP AIDS Project volunteers, personnel files, and donor files. We discussed this strategy with the donor and restricted similar information in the textual portion of the records. To find the documents that contained this information, we formulated a list of 18 search terms that might indicate the presence of sensitive content, including "payroll," "timesheet," "evaluation," and "donor." We then ran an Index Search for each of the 18 search terms in FTK. Each search returned between hundreds and thousands of files. Project staff then reviewed the files and flagged the ones containing private information as "privileged."

The process of searching for keywords was time consuming, as search results needed to be manually skimmed to identify files that should be restricted. One of the biggest challenges in this process was the number of false positives. For example, searching for a term such as "evaluation" returned employee evaluations (which we wanted to flag as "privileged"), but also returned blank evaluation forms, grant proposals and employee handbooks that referenced evaluations (which we wanted to leave open). Familiarity with the types of files in the analog portion of the collection decreased the amount of time necessary to skim each file.

Due to time constraints, searching for and filtering out files containing restricted information was the extent of born-digital processing performed in FTK for this project. If we had the opportunity, a proposed strategy would have been to use FTK's bookmark function to arrange the files into folders corresponding to the series structure of the textual component of the collection. Bookmarks in FTK allow the user to identify and organize files into hierarchical groups. For example, a processor might apply the bookmark "Series 1. Programs" to a large number of files and the bookmark "Subseries 1. Workshops" to a subset of those files. We also could have used FTK's tagging feature to apply tags to groups of files.

This type of arrangement would have been particularly interesting to test out on the born-digital files captured during our project. The files were captured off of a large number of computer media with relatively small storage capacities (e.g. floppy disks, Zip disks, CDs). Thus, the files lacked any overarching directory structure such as one commonly found on computer hard drives. To arrange born-digital files lacking any overarching structure, the workflow would involve constructing a list of terms that might indicate the presence of materials associated with a particular series or subseries, and using the Index Search function to call up, sort through, and assign bookmarks or tags to those files. As more born-digital collections are processed, it will be interesting to see if researchers find imposed arrangement useful, or whether navigating the files through access points such as file type (e.g. documents, spreadsheets, graphics), date, or keyword searches better suits their needs.

Outcomes

Searches for credit card numbers, social security numbers and 18 keywords that might indicate restricted content were performed on 29,423 unique digital files. Of

these, 1,816 files were flagged as containing restricted information. In total, 27,607 files (5,925 MB) are now available for research, including 582 database files, 14,459 document files, 2,869 graphics files, 79 presentations, and 1,557 spreadsheets. The set of processed files, along with associated technical metadata, will join the original disk image and photographs of the media for long-term preservation in the SDR.

Delivery and Discovery

Along with building the AIMS Framework, the AIMS project partners were also tasked with developing an open-source software solution for digital archival materials management, preservation, and access. The partners developed Hypatia (Hydra Platform for Access to Information in Archives).[22] This program is "a Hydra-based Ruby-on-Rails application with a Fedora Repository back-end designed to provide digital archivists with a platform for managing, preserving and providing access to born digital archival materials."[23]

While a prototype was built during the AIMS project, Hypatia is still under development and is therefore not yet a viable delivery option for SUL's born-digital files. Thus, we needed to develop a short-term solution for delivering this content to researchers. We decided to deliver born-digital files that have been screened for restricted information through the Special Collections reading room on a secure network server. Files will be exported from FTK based on the file type (documents, spreadsheets, graphics, etc.) and placed into corresponding folders within a main collection folder on the network drive. Users may choose to browse within the folders for a specific collection. If more advanced techniques are desired, such as cross-collection searching or full text searching, researchers can receive additional assistance with accessing the files through FTK by making an appointment with the Digital Archivist.[24]

21. We only had time during the grant term to capture and process the material acquired before the grant period and included in the grant proposal—the hard drive received during the grant period was accessioned and forensically captured, but no processing was performed on these files during the grant term.

22. Additional information about Hypatia can be found at: Hydra site (http://projecthydra.org/), Hypatia Project Wiki (https://wiki.duraspace.org/display/HYPAT/Home), JIRA project site (https://jira.duraspace.org/browse/HYPAT), and a Hypatia demonstration application hosted at Stanford (http://hypatia-demo.stanford.edu).

23. 2011 DLF Forum, *Hypatia: Research & Early Developments on a Platform for Managing Born Digital Archival Materials*, Digital Library Federation, http://www.diglib.org/forums/2011forum/schedule/hypatia-research-early-developments-on-a-platform-for-managing-born-digital-archival-materials/, (accessed January 13, 2013).

24. At the time this article was written, we had not yet received researcher requests to use the born-digital files. This may be a result of the finding aid only recently being posted online and the press release about the collection yet to be written, resulting in limited promotion of the collection's

We chose to describe the born-digital files in aggregate within a scope and content note for the born-digital component of the collection in the EAD finding aid.[25] We chose not to insert born-digital description throughout the guide at the file or series/subseries level for a number of reasons. First, we did not have time to review the born-digital records in detail for subject content. Instead, we decided to prioritize the search and filtering of sensitive files. Second, because nearly 30,000 unique born-digital files were present in the collection, it would not be scalable to read each document or view each image, and integrate description of these into the collection finding aid.[26] The problem of integrating born-digital content description into EAD finding aids will continue to be compounded by collections containing hundreds of thousands, or millions, of files.

Metrics

Developing metrics for accessioning and processing computer media in this collection proved to be somewhat challenging. For example, while imaging the 3.5 inch floppy disks may have only taken 1.5 minutes per disk on average, another 15 minutes of time over several days may have been spent on trying to see if the disk was readable by different drives, running a virus scan, and cleaning up any resulting viruses.

We estimate that total imaging time for the media was 14 hours. This does not include time spent on photographing the media, virus scanning, or troubleshooting. On average, each floppy disk took 1.5 minutes to image. We attempted 183 captures, which resulted in 4.5 hours of straight imaging time for 3.5 inch floppy disks. Zip disks took approximately 4 minutes per disk or 1.5 hours of imaging (23 Zip disks attempted). Unsuccessful CD imaging attempts took about 2 minutes per disc, and successful imaging took approximately 5 minutes. There were 200 unsuccessful attempts at 2 minutes each and 18 successful attempts at 5 minutes, resulting in about 8 hours to image the CDs.

Performing Live and Index Searches for restricted material and reviewing the returned results took approximately 16 hours of staff time. The searches for social security and credit card numbers took approximately 5 minutes of computer processing time to complete. Reviewing the returned files for actual social security or credit card numbers took about an hour per search. Each index search for potentially restricted files took less than a minute of computer processing time (searches were performed for 18 terms). Depending on the number of returned files, reviewing them took between one half to 2 hours per search.

availability to researchers and limited time during which the collection has been available. Once researchers begin to access the files, it will be interesting to evaluate their experiences using the processed files, and receive their input on this method of delivery. As access to born-digital content in archival repositories throughout the world increases, user studies will be an important field of research and will inform our delivery and discovery efforts.

Because imaging, processing, and creating metrics for born-digital records is a relatively new endeavor, there is a gap in the literature regarding benchmarks for capture and processing. This project was the first at SUL to image and process born-digital files in production mode, so we were not even able to compare these rates with other projects within our repository. Our hope is that this paper will not only provide detailed steps on capturing, processing, and delivering born-digital materials, but also contribute statistics including success rates for capture and the amount of time taken to perform various tasks to the literature.

Conclusion

The NHPRC grant-funded project to process the STOP AIDS Project records was an important milestone in processing born-digital content at SUL. We successfully created disk images of the readable computer media in the collection and in the process, tested and refined workflows that were developed during the AIMS project. We have also begun to establish a baseline for estimating time and effort to process future born-digital collections.

We gained valuable insight into working with different computer media types through the challenges we faced during the imaging phase of the project. Chief among these are noting the differences in process required for working with PC and Apple-formatted disks, and the difference between working with floppy and Zip disks and CDs. Above all, we learned that trial and error, patience, and a sense of humor are necessary when imaging computer media.

The project also shaped our conception of what it means to process a large corpus of born-digital files. Our experiences showed us that searching, reviewing, and flagging restricted material is very time-consuming. This is certainly a task that we will allocate ample time for in future projects. We also learned that even with pattern searches, such as those for social security and credit card numbers, human intervention is required. While it is tempting to rely solely on returned results for restricting materials, because of the large number of false positives, a project staff member needed to skim each individual file to see if it warranted restriction. We also struggled with how much time and how many searches constitute "due diligence" in searching for restricted information. We ran what we thought was an appropriate number of pattern and keyword searches, but it is always possible that some files containing private information may have slipped through the cracks.

Early in the project, we had envisioned sorting the screened files into a series structure mirroring that of the textual portion of the collection. Because we did not have the time to do so, the extent of our processing ended with searching for and restricting files containing private or confidential information. We were prompted to ask ourselves: Is searching for restricted information adequate "processing" for born-digital collections? How much time should we spend on imposing structure onto born-digital files? Will researchers find access based on series structure or based on file format, date or full-text searching most useful?

This line of questioning led us to the comparison between processing the textual and born-digital component of the records. The formats and processes were different, but the records were similar in content and type. Similar privacy and confidentiality issues appeared in both. Because the analog portion was processed first, it provided the processing team familiarity with the types of documents likely to contain sensitive information. This aided tremendously in both developing the list of 18 search terms for restricted material unique to this collection, as well as in reducing the amount of time required for manually scanning the returned search results for false positives. The major difference between the analog and digital records we found was the unique characteristics of born-digital records. For example, detailed physical arrangement and categorization of born-digital files may not be necessary as these files are full-text searchable, and can be aggregated using a number of facets or tags. This fungibility allowed the processing team to forego arranging the files before searching and navigating them using FTK's search capabilities.

We did note that processing the files from the removable storage media in this case study will differ slightly from processing the material on the hard drive (or any material imaged off of a computer hard drive). Whereas the files imaged from the disks and CDs in this case study had no overarching file directory structure, the hard drive consisting of folders from STOP AIDS Project staff computers likely will. This leads to the questions: How will the directory structure of the hard drive shape processing strategies? Will this structure aid researchers by exposing additional contextual information?

This project marked the first in the department to use an archival processing team to process both the textual and born-digital records in a collection. Before this project, the SUL Digital Archivist had been the only staff member testing born-digital processing strategies. This processing occurred while other staff processed the corresponding textual materials. The familiarity project staff gained while processing the textual portion of the collection proved to be invaluable in processing the born-digital records. It facilitated the creation of keywords to be searched in FTK for potentially restricted content. If we had more time, this knowledge would also have aided the project team in formulating a list of search terms related to particular series and subseries and sorting the results into folders or applying tags. In-depth knowledge of the collection is a significant argument for training processing staff working on collections containing both analog and born-digital content and processing the two concurrently.

While we were thrilled with the success of the project, there are many questions yet to be answered, and workflow concerns yet to be resolved. We will continue to gain better statistics on capture rates as we image more computer media. We will continue to refine our workflows and shape our processing strategies. Once Hypatia is developed further, we will be able to offer researchers a more dynamic delivery environment for born-digital materials at Stanford. We look forward to continuing the learning process as we work towards preserving and providing access to born-digital materials.