Julie Seifert. The Time-Cost of Digital Forensics for Archival Collections. A Master's paper for the M.S. in L.S. degree. April, 2014. 84 pages. Advisor: Christopher Lee.

This study describes an experiment in which I performed a series of digital forensics tasks on different forms of digital media. The experiment was conducted in order to determine the time needed to complete these digital forensics tasks. Findings indicate that disk images and forensic reports can be generated quickly, often within seconds, for media of small storage sizes, such as floppy disks. However, for media with large amounts of storage, such as an external hard drive, a single task (generating a disk image) can take several days to complete.

Headings:

Archives

Archives -- Electronic information resources -- Management.

Data recovery (computer science)

Digital preservation

Electronic records -- Management.

Forensic sciences

THE TIME COST OF DIGITAL FORESNICS FOR ARCHIVAL COLLECTIONS

by
Julie Seifert

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April, 2014

Approved by:

____

Christopher A. Lee

Table of Contents

The Time-Cost of Digital Forensics for Archival Collections

In these times of limited budgets, all institutions are concerned about cost. At the same time, as budgets are shrinking, demands on institutions are growing. The digital age has brought on an unprecedented amount of information, and institutions need to help manage it. Additionally, many institutions receive digital data from producers and donors, whether it is stored on hard drives, floppy disks or other media. Institutions want to protect, obtain and preserve this valuable data, but sometimes they are at a loss of how to do so.

These institutions may consider using digital forensics to deal with their acquisitions. "At the most basic level, forensic practices are geared toward establishing the authenticity of files, conducting analysis to discern their characteristics, and generating documentation about what has been done and why."[1] Although originally developed for law enforcement, digital forensics practices are now being adopted by many collecting institutions, which use these tools not only to determine authenticity and the characteristics of the digital data, but also for the purpose of preserving materials of enduring value.

However, while digital forensics can help an institution to achieve the goals discussed above, many are still not implementing these procedures. This could often be because of cost. While it is fairly easy to know the cost of the equipment required to perform these

[1] Matthew G. Kirschenbaum, Richard Ovenden, Gabriela Redwine, "Digital Forensics and Born-Digital Content in Cultural Heritage Collections" (Washington, D.C.: Council on Library and Information Resources, 2010): 39, http://www.clir.org/pubs/reports/pub149/pub149.pdf (accessed January 2014).

tasks, there is another factor that is also important to compute: Time. Many institutions wonder how long various digital forensics tasks will this. Do their staff have enough time to complete these tasks, given their many other responsibilities? Do they have enough staff to complete these tasks, or will they have to hire someone else? Will they be able to finish these tasks before the technology becomes obsolete? How much does this cost, in terms of time?

While institutions wonder about these time costs, often they do not have the opportunity to experiment, and to see whether or not they do have time for digital forensics. However, this is a Catch-22. They want to know how long it will take before they start the project, but in order to determine how long the tasks will take, they must first start the tasks.

In this study, I have been able to experiment with these different tasks and to record how long they take, but without the same demands as a professional. More importantly, I have the opportunity to use the equipment without purchasing it, since it has already been purchased by my parent institution, The University of North Carolina at Chapel Hill.

It is my hope that the results of this paper will prove useful to institutions who are considering implementing digital forensics into their workflows, and that it will help those institutions determine the time needed to complete a certain task or project. This may also help institutions who are considering accessing a collection that contains digital materials, but are not sure if they have the resources to process these materials. Time is a limited resource, perhaps even more limited than funding. The aim of this paper is to see how much time different digital forensics tasks require, so that institutions can better determine how to allocate this valuable resource.

Literature Review

Introduction to Digital Forensics

Before beginning a discussion of digital forensics in cultural heritage institutions, it is important to understand what digital forensics is and what it is used for. It is also important to understand the history of digital forensics, what it can be used for, and its potential to uncover information.

There are many definitions and interpretations of digital forensics, as Jones and Valli note.[2] According to Jones and Valli, the Scientific Working Group for Digital Evidence defines digital forensics as "any information of probative value that is either stored or transmitted in binary form."[3] Simson Garfinkel, a leading expert in the field, writes that digital forensics is "the uncovering and examination of evidence located on all things electronic with digital storage, including computers, cell phones, and networks."[4] "It is concerned with discovering, authenticating, and analyzing data in digital formats to the standard of admissibility in a legal setting."[5] The practices of digital forensics "are geared toward establishing the authenticity of files, conducting analysis to discern their characteristics, and generating documentation about what has been done and why."[6] Another definition, which Jones and Valli call more "usable," is this: "Computer

[2] Andy Jones and Craig Valli, *Building a Digital Forensic Laboratory(* Burlington, MA: Butterworth-Heinemann and Syngress Publishing, Inc., 2009), 7

[3] Jones and Valli, 7

[4] Simson Garfinkel, "Digital Forensics*," American Scientist* 101 (2013): 370-377

[5] Kirschenbaum, Ovenden, and Redwine, 1

[6] Kirschenbaum, Ovenden, and Redwine, 31

forensics is the collection, preservation, analysis, and court presentation of digital-related evidence."[7]

The mention of court presentation in the previous definition demonstrates the history of digital forensics, which, although it is now being adopted by cultural heritage institutions, was first created to investigate alleged crimes. Digital forensics was developed in the United States by federal law enforcement agents in the 1980s, when they "noticed the rise of white-collar crimes that were aided by these new personal computers."[8] As Garfinkel points out, computers can contain evidence for "a crime that took place in the physical world" or "cases…in which the crime was inherently one involving computer systems, such as hacking."[9] Because of these origins in crime investigation, much of the literature surrounding digital forensics discusses concepts that are not obviously relevant to cultural heritage institutions, such as suspects, and, as in the previous definition, court proceedings. However, as will be discussed later in this paper, digital forensics is still highly relevant for cultural heritage institutions, and many of the issues that concern forensics investigators, such as maintaining a chain of custody, also concern archivists.[10]

As Garfinkel says, "Digital forensics is powerful because computer systems are windows into the past."[11] Certainly, digital forensics presents many opportunities for those wishing to learn more about the past, whether it is for legal purposes or historical research. Many computers retain large quantities of information and savvy investigators can recover chat logs, email messages, Google search terms, and other kinds of data that

[7] Jones and Valli, 7
[8] Jones and Valli, 6
[9] Garfinkel, "Digital Forensics," 370
[10] Jones and Valli, 8-9
[11] Garfinkel, "Digital Forensics," 370

were created weeks, months, or even years earlier.[12] These records can provide insight into an "individual's state of mind or intent."[13]

Despite its potential, digital forensics also presents many challenges and limitations. "Electronic data are easily changed, damaged or erased if handled improperly." Garfinkel lists more limitations, such as that "information on a computer system can be changed without a trace, the scale of data that must be analyzed is vast, and the variety of data types is enormous"[14] Data can be purposely "tampered with and manipulated,"[15] and many mistakes can also occur in the digital forensics process, such as inadvertent changes to data.[16] Finally, one key challenge of digital forensics, and the one that will be examined in this paper, is the time required to carry out a forensic investigation. As Valli and Jones point out, "With volumes of storage now in common use," there may not be enough time to examine every bit of information on a device, and decisions must be made about what should be examined.[17] Furthermore, although digital forensics can uncover a great deal of information, some might not be found, or the information may be incomplete.[18]

Digital Forensics in Cultural Heritage Institutions

As more cultural heritage institutions receive born-digital materials and materials stored on removable media, they are beginning to see the value and importance of digital forensics for cultural heritage institutions. "The methods and tools developed by forensics experts represent a novel approach to key issues and challenges in the archives and

[12] Garfinkel, "Digital Forensics*,"* 370
[13] Garfinkel, "Digital Forensics*,"* 370
[14] Garfinkel, "Digital Forensics*,"* 370
[15] Kirschenbaum, Ovenden, and Redwine, 6
[16] Jones and Valli, 13
[17] Jones and Vali, 13
[18] Dan Farmer and Wietse Venema, *Foresnic Discovery* (Upper Saddle River, NJ: Addison-Wesley, 2005): 147-150

curatorial community."[19] It is becoming more common for libraries, special collections and other collecting institutions to receive digital storage media, or even and entire computers, as part of an acquisition.[20] As Lee et al. write:

> The acquisition of digital materials by collecting institutions – libraries, archives and museums (LAMS) – has resulted in the need to incorporate new tools and methods into curatorial practices. LAMS are increasingly called upon to move born-digital materials from removable media into more sustainable preservation environments…[21]

The amount of digital material being received is not likely to decrease. ARMA International estimates that ninety percent of records being created now are born digital.[22] This volume of digital data presents new challenges for archives and other cultural heritage institutions, as well as new opportunities, and digital forensics may be able to help these institutions take advantage of these opportunities.

Harvard University's Houghton Library is one of the many institutions grappling with the question of what to do with digital materials. Leslie Morris, the curator of writer John Updike's papers, received a "steady stream of manuscripts and papers" from Updike.[23] However, in February of 2009, after the writer died, the Library received "approximately 50 three-and-a-half and five-and-a-quarter-inch floppy disks – artifacts from late in the author's career when he, like many of his peers, began using a word processor."[24] At another institution, Emory University, archivists received four laptops, an external hard drive, and a Palm Treo personal digital assistant from writer Salman

[19] Kirschenbaum, Ovenden, and Redwine, 1
[20] Kirschenbaum, Ovenden, and Redwine, 1
[21] Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods, "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions," *D-Lib Magazine* 18: 5/6,(May/June 2012), http://www.dlib.org/dlib/may12/lee/05lee.html (accessed January 2014).
[22] Kirschenbaum, Ovenden, and Redwine, 2
[23] Steve Kolowich, "Archiving Writer's Work in the Age of E-Mail," *The Chronicle of Higher Education* 55:31 (April 10, 2009): 1, http://chronicle.com/article/Archiving-Writers-Work-in/22770 (accessed January 2014)
[24] Kolowich, 1

Rushdie.[25] Additionally, The Harry Ransom Center at the University of Texas Austin holds the computers and disks of authors such as Norman Mailer and Terrance McNally.[26] The Center has reportedly been receiving born-digital items as part of collections for nearly twenty years, and approximately thirty-nine of the Center's holdings contain electronic records, including correspondence and manuscript files on a variety of disks and computers.[27] The digital materials provide valuable information about their creators. For example, the Tom Zigal papers contain a set of proofs created in Microsoft word that Zigal exchanged with his editor at The Toby Press.[28] "Their tracked changes and comments provide valuable insight into the creative process."[29]

Indeed, there are many new insights to be gained from the digital materials of persons and institutions. In his article, "Archiving Writer's Work in the Digital Age," Kolowich writes at length about the possibilities of these materials for literary collections which, evidently, have also been recognized by the archivists at the Harry Ransom Center. "The trappings of the digital age," he writes, such as computers floppy disks "will transform the way libraries preserve and exhibit literary collections."[30] Kolowich argues, "The great American novelists of the digital era – the ones who own BlackBerrys, use Gmail, Facebook, and Twitter, and compose only on computer screens – will soon begin

[25] Kolowich, 2
[26] Matthew G. Kirschenbaum, Erika L. Farr, Kari M. Kraus, Naomi Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside, "Digital Materiality: Preserving Access to Computers as Complete Environments*" Proceedings of the Sixth International Conference on Digital Preservation (iPRES)* (October, 2009): 106-107, http://mkirschenbaum.files.wordpress.com/2009/10/digitalmaterialityipres2009.pdf, (accessed January 2014)
[27] Kirschenbaum, Farr, Kraus, Nelson, Stollar Peters, Redwine, and Reside, 106-107
[28] Kirschenbaum, Farr, Kraus, Nelson, Stollar Peters, Redwine, and Reside, 106-107
[29] Kirschenbaum, Farr, Kraus, Nelson, Stollar Peters, Redwine, and Reside, 106-107
[30] Kolowich, 2

shipping their hard drives off to university libraries."[31] Access to these laptops could potentially allow scholars to gain great insight into the minds of these writers.[32] Although Kolowich focuses on literary collections, the digital materials of politicians, scholars and even organizations, or anyone else using a computer, could prove incredibly informative. "Computers today function as personal environments and extensions of self – we inhabit and customize our computers, and their desktops are the reflecting pool of our digital lives."[33] The question, then, is how to uncover these digital lives, especially when they are stored on seemingly obsolete media, such as the floppy disks that were donated to Harvard. This is one place where digital forensics can help.

As Rogers and John write, "At the most basic level, both digital archivists and digital forensics practitioners are concerned with discovering, understanding, describing and presenting information inscribed on digital media."[34] Using digital forensics techniques, library, archives and museum professionals can work to ensure the authenticity, integrity, and provenance of digital materials.[35] Although the field of forensics might at first seem vastly different from that of archives and museums, "three central requirements of digital forensics match those of archivists: capturing the information without changing it, demonstrating that the information has not been changed or that the changes can be identified, and analyzing and auditing the analysis of the

[31] Kolowich, 3

[32] Kolowich, 3-5

[33] Kolowich, 7

[34] Corinne Rogers and Jeremey Leighton John, "Shared Perspectives, Common Challenges: A History of Digital Forensics and Ancestral Computing for Digital Heritage," in *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation Conference, Vancouver, British Columbia, Canada, September 26-28, 2012* (Vancouver, Canada: UNESCO, 2012): 2. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC__Rogers_John_26_D_1620.pdf (accessed March 2014).

[35] Lee, Kirschenbaum, Chassanoff, Olsen, and Woods, 2

information, again without changing it."[36] Furthermore, the same tools that are sometimes used to solve computer crimes can be adapted for archival purposes, allowing archivists to ensure the integrity of their digital materials, to recover and reconstruct files from source media, and to create a list of electronic files that have been donated, for example.[37] Using forensics techniques, archivists can capture these digital environments and piece together "the relationships of the materials contained within."[38] Forensics tools can also help archivists to make "informed preservation and access decisions" and to search digital media for private, sensitive or personally identifying information.[39] Lee argues that the "incorporation of digital forensics methods will also be essential to the sustainability of archives as stewards of personally identifying information..."[40] The specific tools that will allow an archivist to perform all of these tasks will be discussed in detail in a later section.

However, digital forensics also presents unique challenges for cultural heritage institutions. For example, there is the question of who owns the rights to digital materials. As Kolowich points out, more information is being stored in the cloud and on the Web, and in this environment it is not always clear who owns the information.[41]

In addition to these legal concerns, collecting institutions must also protect the privacy of donors. Digital forensics may allow the archivist to uncover data that the

[36] Rogers and Leighton John, 2
[37] Kirschenbaum, Ovenden, and Redwine, 2
[38] Kirschenbaum, Ovenden, and Redwine, 25
[39] Kam Woods, Christopher Lee, and Sunitha Misra. "Automated Analysis and Visualization of Disk Images and File Systems for Preservation," in *Proceedings of Archiving 2013, Washington, D.C.: April 2-5, 2013* (Springfield, VA: Society for Imaging Science and Technology, 2013): 2, http://ils.unc.edu/callee/p239-woods.pdf (accessed March 2014).
[40] Christopher A. Lee, "Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision," in *Proceedings of the International Council on Archives Congress, Brisbane, Australia, August 20-24, 2012*, http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00290.pdf (accessed March 2014).
[41] Kolowich, 9

donor did not realize he or she was donating.[42] For example, the archivist, in creating a disc image of the media, may also uncover private or personally identifying information.[43] Archivists may also be able to recover files that the user deleted.[44] This private and deleted information could provide the archivists and future researchers with valuable information and insight into the creator.[45] However, before making use of hidden or private data, collecting institutions should first consult with the donor, if possible. Woods and Lee write:

> In order to determine appropriate levels of access to data from an acquired disk, archivists will ideally be able to consult individual producers, representatives of creating organizations, detailed donor agreements, and (when appropriate) applicable laws that dictate who is entitled to access data. However, such information is often not available, and archivists must make their best professional judgments.[46]

John echoes this sentiment, writing that maintaining good relationships with donors and respecting their privacy “ultimately depends on appropriate, effective and open policies and protocol, and astute curatorial decision-making…”[47] Whether or not the collecting institution can make this information available should be addressed in the donor agreement. Lee urges professionals responsible for the care of digital materials to “expand the traditional notion of a donor agreement to address the various forms of

[42] Kirschenbaum, Ovenden, and Redwine, 51-53

[43] Kam Woods and Christopher A. Lee, “Acquisition and Processing of Disk Images to Further Archival Goals," in *Proceedings of Archiving 2012, Copenhagen, June 2012* (Springfield, VA: Society for Imaging Science and Technology, 2012), 147-152. http://ils.unc.edu/callee/archiving-2012-woods-lee.pdf (accessed March 2014).

[44] Woods, Kam, Christopher A. Lee, and Simson Garfinkel, “Extending Digital Repository Architectures to Support Disk Image Preservation and Access,” in *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, (*New York, NY: ACM Press, 2011), 57-63. http://www.ils.unc.edu/callee/p57-woods.pdf (accessed March 2014).

[45] Woods, Lee, and Garfinkel, 57

[46] Woods and Lee, 2012, 3

[47] Jeremy Leighton John, “Digital Forensics and Preservation,” *DPC Technology Watch Report 12-03* (Digital Preservation Coalition, November 2012), http://www.dpconline.org/component/docman/doc_download/810-dpctw12-03.pdf. (accessed March 2014).

representation that are manifested in the digital objects."[48] The person responsible for working with the donor in the donation of these materials should ask, for example, "Exactly what does the donor intend to transfer to the repository?" and "What types or levels of representation are particularly sensitive to the parties represented in the materials?"[49] It is important to balance the interests of researchers with the interests of donors, and to give researchers access to the materials without compromising the privacy of the donors.[50] Archives do not want to risk losing the trust of potential doors.[51] John recommends the following steps:

> (i) Establish open policies and procedures; (ii) inform and seek consent of donors and families; (iii) preview content of personal archives; (iv) discern as far as feasible the interests of third parties; and (v) take actions to comply with policies and expressed wishes.[52]

Finally, forensics tools, such as *fiwalk* and *bulk extractor,* can be used to identify and redact private and personally identifying information.[53] These tools will be discussed in more detail in a later section.

Digital materials and digital forensics also raise new questions about appraisal and selection. "Digital storage is cheap, easy and virtually unlimited,"[54] and therefore donors may accumulate huge volumes of digital information over a lifetime, which they

[48] Christopher Lee, "Donor Agreements," in "Digital Forensics and Born-Digital Content in Cultural Heritage Collections" (Washington, D.C.: Council on Library and Information Resources, 2010), 57 http://www.clir.org/pubs/reports/pub149/pub149.pdf (accessed January 2014).

[49] Lee, "Donor Agreements," 57

[50] Jeremey Leighton John, "The Future of Saving Our Past," *Nature* 459 (June 2009): 775-776. http://www.nature.com/nature/journal/v459/n7248/full/459775a.html (accessed March 2014)

[51] John, "Digital Forensics and Preservation," 33

[52] John, "Digital Forensics and Preservation," 33

[53] Christopher A. Lee and Kam Woods, "Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage," *in Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: An International Conference on Permanent Access to Digital Documentary Heritage, 26-28 September 2012, Vancouver, British Columbia, Canada,* edited by Luciana Duranti and Elizabeth Shaffer, 298-313: United Nations Educational, Scientific and Cultural Organization, 2013. http://ils.unc.edu/callee/p298-lee.pdf (accessed March 2014)

[54] Kolowich, 10

may then try to donate to the institution. At some point, digital media may actually contain too much information. “Mining, sorting, and archiving every bit of data stored on an author’s [or other donor’s] computer could become a chore of paralyzing tedium and diminishing value.”[55] Kolowich adds that, if an institution attempted to save everything, digging through this data could prove frustrating for scholars. He reports the perspective from Matthew Kirschenbaum that “unless scholars are able to find what they want in that sea of data, it is not worth archiving it in the first place.”[56]

Trustworthiness can be another issue when dealing with digital materials, and as Kirshenbaum discusses, born-digital *fonds* are “mobile,” as they pass from the creator to perhaps an intermediary to the staff archival repository and then to storage and ingest into a digital repository.[57] This movement can pose a threat to the trustworthiness of the digital objects especially when it comes to intermediaries, such as manuscript dealers or family members, handling the materials. Redwine et al. encourage repositories to communicate with donors and dealers during the transfer process, and to make donors and dealers aware that simply viewing the files can alter them, among other things.[58] Ideally, digital materials should arrive at the collecting institution with a “documented chain of custody (perhaps even including access history) and authentication information that can be verified upon arrival.”[59] However, this is not always the case. Contemporary recordkeeping is “rarely consistent with recordkeeping ideal.”[60] But an archivist

[55] Kolowich, 10
[56] Kolowich, 10
[57] Kirschenbaum, Ovenden, and Redwine, 27
[58] Gabriela Redwine, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas, “Born Digital: Guidance for Donors, Dealers, and Archival Repositories” (Washington, D.C.: Council on Library and Information Resources, 2013). http://www.clir.org/pubs/reports/pub159/pub159.pdf (accessed March 2013).
[59] Kirschenbaum, Ovenden, and Redwine, 29
[60] Lee, 2012, 3

presented with digital media should try, as best she can, to reconstruct the chain of custody from before she first encountered the media.[61] For example:

> "an archivist acquiring a floppy disk containing records from a donor often will not know with certainty what the states and transitions of the records were before they were last saved onto that disk, but she can use various forms of information (e.g. other records, discussion with the donor) to make inferences about earlier points in the 'life' of the records."[62]

Once the archivist has acquired the media, however, she should implement more detailed record-keeping practices. Lee adds that for purposes of legal compliance and authenticity, archivists need to "document and account for all states of a record and changes of state….from the point of creation to each instance of use and (when appropriate) destruction."[63] Kirshenbaum asserts that in order for these materials to be safeguarded during the transfer process, "dealers and other will need to assume some level of responsibility for the trustworthiness of the digital files."[64] Finally, much responsibility lies with the repository, in addition to the intermediaries. In order to earn the trust of current and future donors, archival repositories should develop a strong technical infrastructure and a sound preservation plan, and demonstrate that the staff and repository are qualified to manage born-digital materials.[65] Digital repositories should also follow agreed-upon models or standards, such as the Reference Model for an Open Archival Information System (OAIS).[66] "Adopting forensic practices geared toward establishing a chain of custody and implementing a series of checks and balances to ensure that when

[61] Lee, 2012, 3
[62] Lee, 2012, 3
[63] Lee, 2012, 3
[64] Kirschenbaum, Ovenden, and Redwine, 28
[65] Kirschenbaum, Ovenden, and Redwine, 29
[66] Kirschenbaum, Ovenden, and Redwine, 29

digital objects arrive at an archival repository they are transferred intact and with appropriate documentation are two other important steps."[67]

Digital objects are also vulnerable to being altered or otherwise damaged if they are not properly handled by the collecting institution. As Kirshenbaum states, "the mere act of opening a file or booting up a computer to alter the archival materials in fundamental ways."[68] Even if one just turns on a computer, new data can be written to the hard drive.[69] Furthermore, removable optical and magnetic media have a limited shelf-life. "Degradation of the media can occur due to incorrect or inadequate storage, damage during handling, and wear on the media during access."[70]

Collecting institutions should also strive to maintain the original order of the digital materials. Lee writes that although the original order of digital materials is often "messy and idiosyncratic," it should be preserved because "it conveys meaningful information about the recordkeeping context, and additional layers of description can be laid on top of that order to facilitate various forms of navigation and access."[71]

Cost proves to be another barrier for cultural heritage institutions considering adopting digital forensics practices and collecting digital materials. "New tools and new training…mean new money."[72] Richard Ovenden, associate director of Oxford's Bodleian Library, says that in order to adopt digital curation, most institutions will have to divert funds from other, "more traditional areas," and they may not be willing to do so,

[67] Kirschenbaum, Ovenden, and Redwine, 31
[68] Kirschenbaum, Ovenden, and Redwine, 28
[69] Kirschenbaum, Ovenden, and Redwine, 28
[70] Kam Woods and Geoffrey Brown, "From Imaging to Access - Effective Preservation of Legacy Removable Media," in *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations: Final Program and Proceedings* (Springfield, VA: Society for Imaging Science and Technology, 2009), 213-218.
[71] Lee, 2012, 3
[72] Kolowich, 11

at least not right away.[73] The speed at which they adopt these practices "could be at a slower pace than the speed of technological invention itself."[74] Cook also supports this statement, and adds that there is a human cost, writing that "unless you can get substantial new financial and human resources, you will need to stop doing something important that you are doing now, and reallocate significant resources to electronic records, period."[75] He suggests that this problem is especially relevant to small institutions.[76] Kirshenbaum adds that the full costs of providing an infrastructure for digital forensics is still unknown. He cites costing models such as the LIFE2 model but says that these methodologies "are probably too generic to provide anything more than broad guidance about the costs of acquiring, capturing, managing, securing, and providing controlled access to sensitive digital information."[77] However, he adds the cost of adopting these practices "is likely to be high for the foreseeable future."[78] The cost of equipment and software may also be a challenge, although some tools, such as the BitCurator environment are open source and freely distributed.[79] Leighton John is more optimistic about the cost of digital forensics: "Smaller institution will be able to do much with a combination of free and inexpensive tools (write blockers, open source software, FTK imager and others); larger institutions may be able to justify greater expenditure in

[73] Koliwich, 11

[74] Koliwich, 11

[75] Terry Cook, "Byte-ing Off What You Can Chew: Electronic Records Strategies for Small Archival Institutions," *Archifacts* (April 2004). http://www.aranz.org.nz/Site/publications/papers_online/terry_cook_paper.aspx (accessed January 2014).

[76] Cook, 1-5

[77] Kirschenbaum, Ovenden, and Redwine, 48

[78] Kirschenbaum, Ovenden, and Redwine, 49

[79] For more information, see: Christopher A. Lee, Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods, "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions," *D-Lib Magazine* 18, No. 5/6 (May/June 2012), http://www.dlib.org/dlib/may12/lee/05lee.html

part so that a wide range of tools can be tried and tested for the benefit of the wider community…."[80]

Finally, for some, there may be a psychological barrier to working with digital materials. The tasks can seem daunting, perhaps overwhelming, and Cook writes that, for archivists, "a starting point is getting past the fear factor, and recognizing that the whole solution is not resting on their shoulder or actions."[81] Archivists may feel that they do not have enough knowledge to work with these materials, but Cook assures us that "no one is qualified to speak about electronic records with full authority."[82] Cook believes that no one has the "one answer" for perfectly capturing and managing electronic records, for ensuring their authenticity, or for preserving them well into the future, even as technologies change.[83] He asserts that there is probably not just one solution to these complex problems.[84] The solution will be different depending on a number of factors, including the size, complexity, and type of digital information and the resources of the collecting institution.[85] Therefore, Cook adds that in order to find these solutions, archivists "will certainly need a vast array of tools…in our professional toolkit."[86]

Despite the risks and drawbacks presented by digital forensics and digital materials, the risks of not adopting these practices, and of not collecting digital materials, could be far greater than the risks of doing so. In his article, "Byte-ing Off What You Can Chew," Terry Cook discusses a statement made in April 2004 by Eduard Mark, a senior

[80] John, 2012, 19
[81] Cook, 1
[82] Cook, 1
[83] Cook, 1
[84] Cook, 1
[85] Cook, 1-3
[86] Cook, 1

historian with the United States Department of Air Force. According to Cook, Mark writes:

> I wrote a history of the invasion of Panama, which remains classified. I began my research within weeks of the operation and found that many electronic records had already been purged from computers…I will mince no words. It will be impossible to write history of recent diplomatic and military history…Too many records are gone…History as we have known it is dying, and with it the public accountability of government and rational public administration.[87]

Unless these and other electronic records are saved, much of the historical record may be lost.

Finally, in order to overcome many of the challenges presented by digital media, such as cost, training, and equipment, many in the field have suggested that cultural heritage institutions might work together or collaborate in order to overcome this seemingly-overwhelming task. For example, in "Swatting the Long Tail of Digital Media: A Call for Collaboration," Ricky Erway proposes that instead of all institutions attempting to deal with all kinds of digital media, a few institutions could each specialize on certain types of media. She writes, "A community-based approach would use SWAT [software and workstations for antiquated technology] sites wherein a few self-selected institutions acquire and maintain the gear and expertise to read data and transfer content from particular types of obsolete media."[88]

Key Concepts in Digital Forensics

To understand the methods of forensic investigation carried out in this paper, it is important to first understand some key concepts about computers and digital storage media, such as how computers write and store data. This section is not meant to be an

[87] Cook, 2

[88] Ricky Erway, "Swatting the Long Tail of Digital Media: A Call for Collaboration." Dublin, OH: OCLC Research, 2012. http://www.oclc.org/content/dam/research/publications/library/2012/2012-08.pdf (accessed January 2014).

exhaustive explanation of how computers work and will focus only on the aspects of computers most relevant to digital forensics and to the tasks carried out in this paper.

Computers organize, store, and retrieve data using the file system, "which means that it is important not only in relation to the files themselves but also to their metadata."[89] The file system allows users to store data "in a hierarchy of files and directories," and it organizes data so that the computer knows where to find it.[90] The file system is "independent from any specific computer."[91] File systems have specific procedures and structures for storing information, whether it is a small amount of data on a floppy disk or thousands of files on a personal computer, and this "underlying structure allows any computer that supports the type of file system to process it."[92] For digital forensics, it is important to understand the file system because this will allow the investigator to understand how computers write data and how deleted data can be recovered. Additionally, in digital forensics, file system analysis "examines data in a volume (i.e. a partition or disk) and interprets them as a file system."[93] File system analysis will allow the investigator to perform many tasks, such as listing the files in a directory, recovering deleted content and viewing the contents of a portion of the disk.[94]

The smallest discrete unit of data that a computer can handle is called a bit. The bits are then grouped into groups of eight called bytes, to store data. When recorded on a hard drive or memory card, these bytes are grouped in blocks called sectors that are typically 516 or 4,096 bytes in length. A sector is the smallest block of data that a drive

[89] Kirschenbaum, Ovenden, and Redwine, 15
[90] Brian Carrier, *File System Forensic Analysis* (Upper Saddle River, NJ: Addison-Wesley, 2005), 174
[91] Carrier, 173
[92] Carrier, 174
[93] Carrier, 174
[94] Carrier, 177-178

can read or write. Each sector on the disk has a unique identifying number, called the sector's logical blog address. Two or more sectors form a cluster or block. The number of bytes in a cluster depends on the disk's size and on the version of the operating system used to format the disk.[95] A cluster is the smallest unit of memory that an operating system will use to store information. Even if a file contains only two bytes, the operating system will still write this information to a cluster, which may be as large as 32 kilobytes. As a result, there is usually extra, unused space in a cluster, called slack space. Slack space is useful in digital forensics because an investigator can sometimes finds remains of deleted files in the slack space.[96]

Computers write data to the disk by allocating files to specific storage sectors. "Allocated files are ones that can be viewed through the file system and whose contents under normal circumstances will not be inadvertently overwritten by the operating system."[97] In other words, these are the files that one can easily see on a computer without using any specialized software, such as through opening "My Documents" on a Windows operating system. The files are called "allocated" because they are assigned to a particular cluster or clusters on the disk, and that space cannot be taken up by another file.

However, if a file is deleted by the user, then those sectors are deallocated. If that happens, then new data can be written to those sectors. They are no longer assigned to the old data. As Kirshenbaum writes, "The 'delete' command simply tells the file system to make the clusters associated with a given file available again for future use."[98] However,

[95] Garfinkel, "Digital Forensics*,"* 372
[96] Kirschenbaum, Ovenden, and Redwine, 44
[97] Garfinkel, "Digital Forensics*,"* 375
[98] Kirschenbaum, Ovenden, and Redwine, 43

sometimes the space is not immediately allocated to a new file. In those instances, the file may continue to be stored in memory, on the hard drive, or on external media, "even though the metadata that could be used to locate it are lost."[99] Even when users delete a file, it remains on the disk, in unallocated data blocks and in unallocated file attribute blocks, until it is overwritten by other data.[100] When a user deletes a file, the computer does not immediately wipe the file from its memory. Instead, it marks the space as available for future data, but the old file will not be erased until something new is written into that memory space.[101] As Farmer and Venema write, "Destroying information turns out to be difficult. Memory chips can be read even after a machine is turned off. Data on a magnetic disk can be recovered even after it has been overwritten multiple times."[102] Information about the files, called MAC (modified, accessed, created/changed) times, also can survive for several months or even years.[103]

Data from deallocated sectors (that is, data that has been deleted by the user but not overwritten by new data) can be recovered using a technique called file carving.[104] A file carver makes use of "characteristic sequences of bytes at the beginning and end of each file…called file headers and footers."[105]

Computers can also use a technique called compression is to "squeeze data" so it uses less storage.[106] However, compressed data can be more difficult to reconstruct through file carving.[107] For example, if the file has been compressed, it may be corrupted

[99] Garfinkel, "Digital Forensics," 375
[100] Farmer and Venema, 146
[101] Garfinkel, "Digital Forensics," 375
[102] Farmer and Venema, 146
[103] Farmer and Venema, 149-150
[104] Garfinkel, "Digital Forensics," 375
[105] Garfinkel, "Digital Forensics," 375
[106] Garfinkel, "Digital Forensics," 375
[107] Garfinkel, "Digital Forensics," 375

or partially missing.[108] However, compression can also be useful in digital forensics, because compressed data can be processed more quickly.[109]

Another key concept is that of hash functions. A hash function generates a unique, fixed-length sequence of characters. As Garfinkel writes, "Hash functions are designed so that changing a single character in the input," that is, in the bits that generate the hash, "results in a completely different output," which is the string.[110] However, occasionally, two different groups of bits will generate the same hash, which is called a hash collision, although this is rare.[111] Hash functions are useful both for ensuring that data has not been changed and for recognizing specific files.[112]

Random Access Memory (RAM) is also of interest to digital forensics investigations. "RAM gets its name because the data in its stores can be accessed in any order."[113] Because RAM can be accessed quickly, it is often used for temporary storage and working space for the computer's operating systems.[114] But RAM can prove challenging for forensic examiners, because its contents change quickly and are gone shortly after a computer is turned off.[115] In order to capture RAM, the examiner must use a dedicated program (a *memory imager*).[116] This information is then stored in its own special kind of file, called a *memory dump*.[117] RAM might contain a lot of useful information, such as bits of programs that have been recently run and closed, but it is also

[108] Garfinkel, "Digital Forensics*,"* 375
[109] Woods, Lee, and Misra, 4
[110] Garfinkel, "Digital Forensics*,"* 374
[111] Garfinkel, "Digital Forensics*,"* 374
[112] Garfinkel, "Digital Forensics*,"* 374
[113] Garfinkel, "Digital Forensics*,"* 372
[114] Garfinkel, "Digital Forensics*,"* 372
[115] Garfinkel, "Digital Forensics*,"* 372
[116] Garfinkel, "Digital Forensics*,"* 372
[117] Garfinkel, "Digital Forensics*,"* 372

very difficult to capture and, when captured, is usually incomplete.[118] However, archivists rarely collect data from the RAM, since archivists most often collect digital data from a donor's computer after they are no longer actively using that computer.

Tools and Procedures

To perform a digital forensics investigation, it is necessary not only to have the correct tools to carry out the investigation, but also to know how and when to use them. In this section, I will briefly describe some basic steps in a digital forensics workflow as well as some key tools. Again, this is not meant to be an exhaustive description of every possible tool and every possible step, but is instead meant to give a general idea. Additionally, the tools and steps used here will be the ones used most in the procedure later described in this paper.

Kirschenbaum et al. discuss pre-accession procedures. They write that "the transfer process for digital materials needs to be managed carefully, and with rigorous adherence to documented procedures incorporating standard elements of archival accessing that have been adapted to the needs of digital objects."[119] When transferring materials, it is crucial to take any steps necessary to minimize risk to the materials.[120] Kirshenbaum also stresses the importance of documentation, completing a transfer list which contains information such as ownership and permissions, and of the "generation of checksums for comparison in future integrity checks."[121]

[118] Garfinkel, "Digital Forensics*,"* 372
[119] Kirschenbaum, Ovenden, and Redwine, 38
[120] Kirschenbaum, Ovenden, and Redwine, 38
[121] Kirschenbaum, Ovenden, and Redwine, 38

Both before and during access, it is important for collecting institutions to communicate with donors.[122] As has been mentioned previously, it is important that donors understand that the institution may be able to recover files that the donor had not intended to donate, such as deleted files.[123] Additionally, the donor should understand the accessioning process and policies so that they can provide "necessary information and guidance."[124] Additionally, the collecting institution should be certain that it will be able to "gather sufficient information to establish an appropriate level of physical, administrative, and intellectual control over the materials being transferred."[125] The institution needs to establish guidelines about what kinds of records, curatorial area, and creators they will focus on.[126] The institution needs to be certain that the scale of the potential donation fits the scale of their institution.[127] That is, do they have the resources to make these digital materials available in a timely manner?[128] Additionally, do they have the technical knowledge and infrastructure to properly transfer and preserve these records?[129] The institution must also be certain that they will be able to gain legal custody of the records.[130] Finally, it is important that institutions not delay in the accessing process, as accessing "benefits from being carried out as soon as possible after selection, to better ensure preservation and integrity of digital content."[131]

[122] AIMS Work Group. "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship." 2012. http://www.digitalcurationservices.org/aims/white-paper/
[123] AIMS Work Group, 18
[124] AIMS Work Group, 18
[125] AIMS Work Group, 19
[126] AIMS Work Group, 19
[127] AIMS Work Group, 19
[128] AIMS Work Group, 19
[129] AIMS Work Group, 19
[130] AIMS Work Group, 19
[131] AIMS Work Group, 19

Glisson and Maxwell provide a description of a typical digital forensics workflow. First, "one must decide where to store the information."[132] Additionally, the target drive, that is, where the information will be stored, should be forensically cleaned, to wipe any data that might be lingering on the target device, which could be confused with the new information.[133] The next step is to record information about the hardware, such as serial numbers and manufacturer information.[134] Then, the next step is to "start the chain of custody and to transport the device to a secure lab for processing."[135]

Once the device has been transported, "a bit stream copy of the removable media should be made by creating either a clone or a forensic image of the device."[136] Garfinkel provides an explanation of the process of creating this copy, also called a disk image:

> To preserve the data on a computer or phone, each of these sectors must be individually copied and stored on another computer in a single file called a disk image or physical image. This file, which contains every byte from the target device, naturally includes every visible file. But the physical image also records invisible files, as well as portions of files that have been deleted but not yet over-written by the operating system.[137]

Additionally, by using a disk image during "triage and analysis tasks," instead of using the source media, the examiner can reduce the risk of erasing or otherwise damaging the source media. [138] Across several collecting institutions, Gengenbach found that the creation of forensic disk image is a "central," important part of the workflow.[139] Write-

[132] Brad Glisson and Rob Maxwell, "A Digital Forensics Workflow," in, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, D.C.: Council on Library and Information Resources, 2010), http://www.clir.org/pubs/reports/pub149/pub149.pdf (accessed January 2014), p. 16

[133] Glisson and Maxwell, 16

[134] Glisson and Maxwell, 16

[135] Glisson and Maxwell, 16

[136] Glisson and Maxwell, 16

[137] Garfinkel, "Digital Forensics*,"* 372

[138] Woods, Lee, and Misra, 240

[139] Martin J. Gengenbach, "'The Way We Do it Here': Mapping Digital Forensics Workflows in Collecting Institutions," A Master's Paper for the M.S. in L.S degree, August, 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf (accessed March 2014)

blocking hardware should be used in this process to avoid accidental changes to the data, and this hardware should be tested before being used on the data.[140] The data then needs to be authenticated using hash functions which, as discussed in the previous section, will ensure that the data are identical.[141] Institutions might make several copies of the disk image, and keep one "isolated" as a master copy, while another is used to create access copies for researchers.[142]

Next, the examiner should identify active files and inactive files. "Active files are readily identifiable and can be access with the appropriate software and, in some cases, the required security information."[143] Inactive files, or deleted files, can be found in allocated space and slack space, as previously discussed. Furthermore, other tools allow the user to extract files from the disk image, to search for a specific word or phrase within the files, or to find encrypted data.[144] The examiner can then extract the relevant data from the disk image "so that they are easier to analyze"[145] As Woods et al. write, "the disk image can be mounted on a host system or in a virtual machines…and any readable filesystems can be explore manually."[146] The examiner can also use commercial tools such as FTK Imager to identify and explore "both the filesystems and unallocated spaces."[147] However, there are issues with these approaches. For example, manually exploring the file system can be "error prone" and time consuming and may not yield useful results.[148] Furthermore, the tool previous mentioned, FTK Imager, is free but has

[140] Glisson and Maxwell, 16
[141] Glisson and Maxwell, 16
[142] Gengenbach, 76
[143] Glisson and Maxwell, 16
[144] Garfinkel, "Digital Forensics*,"* 373
[145] Garfinkel, "Digital Forensics*,"* 372
[146] Woods, Lee, and Misra, 240
[147] Woods, Lee, and Misra, 240
[148] Woods, Lee, and Misra, 240

"a limited set of filesystem analysis utilities," and tools with more capabilities might also cost more.[149]

Once a collecting institution has captured the digital materials, they may then wish to arrange and describe them. The AIMS working group notes that "success within arrangement and description of born-digital materials can be describe in the same way as traditional archival records."[150] That is, the institution should work to preserve the context in which the records were "created, managed, assembled or accumulated."[151] To do so, collecting institutions should strive, throughout the accessing and capture process, to gather evidence of the context and preserve the metadata embedded within the files.[152] The institution should also strive to maintain intellectual control over the materials and to provide some means of discovery of the materials, such as a finding aid.[153] The AIMS working group also stresses the importance of documenting the processing of the materials.[154]

As the AIMS Work group writes, "Discovery and access workflows…are shaped by the needs of user communities, but also need to be carried out with regard to legal and ethical issues relating to the material and the information contained within it."[155] Understanding the user needs and user base of a collecting institution can be especially difficult with digital materials. Since digital materials are often made available online, users often have less interaction with the archivists, and the institution can become less

[149] Woods, Lee, and Misra, 240
[150] AIMS Work Group, 32
[151] AIMS Work Group, 32
[152] AIMS Work Group, 32
[153] AIMS Work Group, 32
[154] AIMS Work Group, 32
[155] AIMS Work Group, 44

familiar with its user base.[156] Additionally, although having materials online allows for wider access, it also "significantly increases the risk of misuse or abuse of copyrighted or sensitive information."[157] It is important to mitigate these risks in order to protect the interests and maintain the trust of the donor. The AIMS working group suggests that institutions provide "clear statements regarding usage rights, clear and effective policies on restriction and data curation and [demonstrate to donors] a working system of access restrictions and long-term preservation."[158]

There are many digital forensics methods which could be used throughout this process and which could support the goals and functions of collecting institutions. Chief among these is the creation of disk images. Disk images can, for example, help collecting institutions ensure the provenance, integrity, authenticity of digital materials.[159] "Both the file system metadata within the disk image and the supplementary metadata within the disk image package can be used to document provenance and chain of custody."[160] Acquiring a disk image would also allow an institution to perform "data triage and data integrity tasks," such as creating cryptographic hashes and "creating maps and hierarchies of allocated and unallocated space on the original device."[161]

Additionally, as previously mentioned, by acquiring disk images, collecting institutions also lessen the risk that they will damage the source media.[162] "Disk images allow researchers to retain and investigate aspects of the systems that could be

[156] AIMS Work Group, 44
[157] AIMS Work Group, 44
[158] AIMS Work Group, 45
[159] Woods and Lee, 2012, 1
[160] Lee and Woods, 2013, 303
[161] Woods and Lee, 2012, 2
[162] Woods and Brown, 1

inadvertently altered during normal operating of a typical operating system."[163] The original media could also degrade, and using the original media limits patron access.[164] If an institution chose to have patrons access the original source media, only one patron could access the media at the same time, and patrons would probably have to use a designated work station.[165] Furthermore, the "speed of access" is usually higher when using disk images as opposed to accessing the original source media.[166]

Disk images can also provide collecting institutions and researchers with contextual information. A disk image "provides the user with valuable information about how the device was organized, who uses, and which users had access to particular contents on the device."[167] Disc images "that contain complete operating systems capture significant information about the 'digital ecosystem' in which the documents and media were created."[168] Furthermore, as previously mentioned, disk images may uncover damaged data, private data, or data that was thought to be lost or deleted.[169] By using disk images, the repository will not need to be concerned about having equipment to mount the source media.[170] Despite these benefits, Woods and Lee reported that "generation and management of disk images remains relatively rare in current repositories."[171]

Disc images are typically saved in either raw (also known as dd) format, in ISO format (for optical media) or as forensically packaged disk image formats, such as the

[163] Woods, Lee, and Garfinkel, 58
[164] Woods and Brown, 1
[165] Woods and Lee, 2012, 2
[166] Woods and Lee, 2012, 2
[167] Woods and Lee, 2012, 1
[168] Woods and Lee, 2012, 2
[169] Woods and Lee, 2012, 1
[170] Woods and Lee, 2012, 1
[171] Woods and Lee, 2012, 2

Advanced Forensic Format (AFF) or Guidance Software's Evidence Witness Format (E01). While raw disc images have the advantage of being widely supported by many software tools, hey also have many limitations. "As sector-by-sector copies of the drive contents, they do not retain additional metadata about the capture process or supporting actions performed during acquisitions."[172] In contrast, AFF and E01 include both the disk images and metadata generated during imaging. This metadata can provide insight into the user who performed the capture, the system that performed the imaging, the physical storage medium as well as cryptographic checksums and timestamps.[173] These formats may also provide information about areas of the source media that might be damaged, as well as manufacturer data associated with the media.[174] "This information may be used to support technically consistent workflows, improve records of provenance, and assess issues associated with authenticity and duplication."[175]

Garfinkel also discusses "file-based approaches" to digital forensics, which are "widely used" and "implemented by popular tools such as guidance Software's EnCase…and AccessData's FTK."[176] File-based approaches are useful because they are easy to understand, since "they mirror the way that users interact with computers."[177] However, "They have the disadvantage of ignoring data not contained within files."[178] Another approach to digital forensics, as discussed by Garfinkel, is bulk data analysis. In this approach, "Digital content is examined without regard to file system metadata.

[172] Woods and Lee, 2012, 1

[173] Woods and Lee, 2012, 2

[174] Woods, Lee, and Misra.1

[175] Woods, Lee, and Misra.1

[176] Simson Garfinkel, "Digital media triage with bulk data analysis and *bulk_extractor*," *Computers and Security* 32: 56-72 (2013) http://simson.net/clips/academic/2013.COSE.bulk_extractor.pdf (accessed January 2014), 57

[177] Simson Garfinkel, "Digital media triage with bulk data analysis and *bulk_extractor*."52

[178] Garfinkel, Simson Simson, "Digital media triage with bulk data analysis and *bulk_extractor*." 52

Instead, data of interested is identified by content and processed, extracted, and reported as necessary."[179] Bulk data approaches also "have the advantage of being applicable to all types of computers systems, file systems and file types."[180]

One set of tools that can be used to analyze a disk image is the BitCurator environment. "BitCurator incorporates software designed to improve coverage and efficiency when analyzing disk images, and reduce the potential for error when handling these materials in archival workflows."[181] The BitCurator environment "use[s] and expand[s] on tools including Simson Garinkel's *fiwalk* and *bulk extractor* and Basis Technology's The Sleuth Kit to produce human readable reports using technical metadata extracted from raw and forensically packaged images"[182] Furthermore, "the data generated by these tools can be used to improve triages of and access to digital collection, and to support a range of preservation decisions."[183]

Bulk extractor is a tool that performs bulk analysis, which was discussed previously. In the BitCurator environment, *bulk extractor* "is employed to identify potentially private and sensitive information, and to search for relevant patterns in the bitstream specified by the user. *Bulk extractor* does not parse filesystems but instead reads the raw contents of the disk image."[184]

Fiwalk is a "disk image parsing tool"[185] which "identifies and interprets the contents of filesystems contained in disk images…"[186] It "can produce both XML (as digital forensics ML) and simple text reports on the processed media: filesystem(s) and

[179] Garfinkel, Simson Simson, "Digital media triage with bulk data analysis and *bulk_extractor*."52
[180] Garfinkel, Simson Simson, "Digital media triage with bulk data analysis and *bulk_extractor*." 52
[181] Woods, Lee, and Misra, 1-2
[182] Woods, Lee, and Misra, 1
[183] Woods, Lee, and Misra, 1
[184] Woods, Lee, and Misra, 3
[185] Lee, Kirschenbaum, Chassanoff, Olsen, and Woods, 3
[186] Woods, Lee, and Misra, 3

volume(s) encountered, file objects and associated metadata within a given filesystem, and information on byte runs associated with file fragments."[187] *Fiwalk* can "generate reports of all files on a drive, along with their associated filesystem metadata and locations within the filesystem hierarchy"[188]

In order to generate reports about which files contain information of interest, such as personally identifying information, another tool, identify_filesnames.py, can also be used.[189] This tool matches this information found by *bulk extractor* to the file from which it came.[190] This is necessary because "*bulk extractor* ignore filesystem structure."[191]

Once the disk image has been processed, the BitCurator environment will produce two sets of data. The first is "a detailed report – based on Digital Forensics XML – on data from the filesystem" and "details the filesystem hierarchy information in a single XML file using the current set of Digital Forensics XML tags." Digital Forensics XML is an XML schema which represents "an initiative to enable to production of interoperable metadata by digital forensics tools."[192] Some of the current DFXML tags include "volume structure, permissions, [and] timestamps…"[193] "With this metadata, one can rapidly produce informative, human-readable reports," which include information such as "timelines of modification," "location and contents of user accounts," and "'hidden data.'"[194]

The second set of data that will be generated by the BitCurator environment, after the disk image has been processed, is "sets of features corresponding to information

[187] Woods, Lee, and Misra, 3
[188] Woods and Lee, 2012, 4
[189] Woods, Lee, and Misra, 3
[190] Woods, Lee, and Misra, 3
[191] Woods, Lee, and Misra, 3
[192] Lee, Kirschenbaum, Chassanoff, Olsen, and Woods, 3
[193] Woods, Lee, and Misra, 5
[194] Woods, Lee, and Misra, 5

within the filesystem that may be private, sensitive, individually identifying, or indicative of specific actions on the part of the user."[195] The BitCurator environment can generate reports from the contents of the disk image and the DFXML outputs.[196] The reports show the distribution of data on the disk and indicate areas likely to contain large amounts of private data.[197] The reports can also show if an external device was used, and can create a timeline of email activity.[198]

Woods et al. provide some insight into the time cost of processing materials using the BitCurator environment. They write, "The time required to process a given disk image with *fiwalk*, *bulk extractor*, the annotation tool, and the BitCurator reporting module is a function not only of the processing and disk speed of the workstation, but also on the composition of the disk images."[199] Furthermore, images with larger amounts of data take longer, and smaller amounts take less time.[200] In terms of time, "The limiting factor…is generally the BitCurator report generation tool, which may have to process extremely large text feature reports and XML file system repots as produced by *bulk extractor* and *fiwalk*.[201]

Another tool that will be used in this paper is the FRED (Forensic Recovery of Evidence Device), a specialized computer produced by Digital Intelligence. According to the company's website, "FRED systems are optimized for stationary laboratory acquisition and analysis."[202] The FRED station is useful for digital forensics because it has extensive memory, which is useful for processing and well as creating disk images.

[195] Woods, Lee, and Misra, 1
[196] Lee and Woods, 2013, 305-306
[197] Lee and Woods, 2013, 305-306
[198] Lee and Woods, 2013, 305-306
[199] Woods, Lee, and Misra, 4
[200] Woods, Lee, and Misra, 4
[201] Woods, Lee, and Misra, 4-5
[202] Digital Intelligence, http://www.digitalintelligence.com/products/fred/ Accessed March 2014.

The FRED contains several memory drive bays as well as a drive cooling system to keep the drives cool during the imaging process.[203] The FRED also has a write-protected imaging bay. Finally, according to the manufacturer's website, the FRED can "acquire data directly from IDE/EIDE/ATA/SATA/ATAPI/SAS/Firewire/USB hard drives and storage devices and save forensic images to Blu-Ray, DVD, CD or hard drives."[204] FRED systems also acquire data from Blu-Ray, CD-ROM, DVD-ROM, Compact Flash, Micro Drives, Smart Media, Memory Stick, Memory Stick Pro, xD Cards, Secure Digital Media and Multimedia Cards.[205]

Kirshenbaum writes that, in order to ensure the integrity of the data after ingest, institutions must follow good archival practice, and use available tools and technology to ensure that the data is not "interfered with or altered" while in the custody of the repository.[206] The archivist should also strive to be active in each stage of the archival life cycle for the digital records.[207] He stresses the importance of maintaining metadata, which can be used to manage "both the use and administration of digital records."[208]

Jones and Valli discuss common mistakes in forensics investigations, and although they write in the context of a legal investigation, much of this discussion is also applicable to a cultural heritage institution. For example, one of the most common mistakes, they say, is the "failure to maintain the proper documentation."[209] Also, examiners may accidentally alter the data by opening a file.[210] Examiners may also fail to "adequately control access to the digital evidence" and thus they will jeopardize the chain

[203] Digital Intelligence.
[204] Digital Intelligence.
[205] Digital Intelligence.
[206] Kirschenbaum, Ovenden, and Redwine, 38-39
[207] Kirschenbaum, Ovenden, and Redwine, 38-39
[208] Kirschenbaum, Ovenden, and Redwine, 38-39
[209] Jones and Valli, 10
[210] Jones and Valli, 10

of custody.[211] Finally, the investigator might fail to realize or to admit when he had reached the limits of his own knowledge, and might not ask for help.[212] Although Jones and Valli are writing about forensic investigators, this statement could also be applied to a staff member at a cultural heritage institution. As they say, “the subject is now so vast and complex it is not possible for one person to have the necessary level of knowledge in all relevant areas.”[213]

Methodology and Results

In this experiment, I sought to answer the question: how long would it take to complete a certain digital forensic task, with a certain set of options within that task, on a piece of digital media of a certain size? To determine this, I performed a series of different forensics tasks that might be performed by a collecting institution. Often I did similar tasks with slight variations. For example, I made a disk image in a raw disk image format, and then imaged the same disk, but using E01 format. I did this in order to provide insight into how long each task takes, and then what within those tasks is the most time-consuming. I hope that, by performing similar tasks with slight variations, it allows others to see what activities and tasks are most time consuming. This, in turn, will hopefully allow others to perform a cost-benefit analysis of whether the time required for a certain task or certain variation will be worthwhile. For most tasks, the software itself timed the task and provided information about how long it took to perform each task. When that was not available, I used a stopwatch function on my iPhone.

I recorded only the time it took for the computer to perform the task itself. I did not record the time for any of the activities leading up to the task, or activities I did in

[211] Jones and Valli, 11
[212] Jones and Valli, 11
[213] Jones and Valli, 11

preparation for another task. For example, I did not record the time it took to start a program or start up the computer.

To complete all of the following tasks, I used a FRED system. The computer I used for these tasks was running on Windows 7 Ultimate with 24 GB of memory. It had a 64 bit operating system. The main C drive of the FRED had a speed of 10,000 RPM. It had a NTFS file system. I used a write-blocked USB connection to move data from the hard drive to the computer. I used an external floppy disk drive for the floppy disks, which I connected to the computer via a write-blocked USB connection .

For the first set of tasks, I used a 3.5 inch floppy disk, which contained approximately 1.5 MB of data. This disk was part of a collection of disks, which in turn was part of a collection in the UNC Southern Historical Collection. I was granted permission to use this disk by Meg Tuomala, the Electronic Records Archivist at the University of North Carolina, and I am extremely grateful that she gave me the opportunity to use these materials. I used the same disk for all of the following experiments, to control for variations in different disks.

The first disk was labeled “Ques2.Dat to Ques231.dat.” It is assumed that this label was created by the producer. The disk contained 32 DAT files, a file format for data. The disk also contained two deleted, or unallocated, files of unknown size. The file system was FAT 12. There were no files on the disk larger than 1 MB. Files ranged in size from 24576 bytes to 51100 bytes. There were no image files on the disk. The files on the disk were all last modified in January of 1995.

I first created a disk image using Guymager, which is a disk imaging tool that is part of the BitCurator environment. Guymager allows the user to choose from several options when imaging a disk. First, the user can choose whether to create a raw disk image or an image in the .e01 format. The user can also choose to create an MD5 hash for the disk, a SHA-1 hash, a SHA-256 hash, two of these, all three, or none. The user can also choose to have the program re-read the source after acquisition and/or to verify the image after acquisition. Verifying the image after acquisition means that the program will use the generated hash to ensure that the acquired image has not been altered during acquisition. In re-reading the source after acquisition, the program will re-scan the source media to ensure that the disk image matches the source. Using the 1.5 MB floppy disk, I imaged the disk using some, all, or none of these options, and the time taken for each variation is recorded below. I also assigned a number to each image, for later use, as seen in Table 1.

For all tables, the time is recorded as hours, minutes, seconds unless otherwise noted.

Table 1: Results of creating disk image in Guymager for Floppy Disk 1

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:33	1
E01	No	No	No	No	No	00:00:32	2
Raw	Yes	No	No	No	No	00:00:33	3
E01	Yes	No	No	No	No	00:00:32	4
Raw	No	Yes	No	No	No	00:00:32	5
E01	No	Yes	No	No	No	00:00:31	6
Raw	No	No	Yes	No	No	00:00:32	7

E01	No	No	Yes	No	No	00:00:33	8
Raw	Yes	No	No	Yes	No	00:01:02	9
E01	Yes	No	No	Yes	No	00:01:03	10
Raw	Yes	No	No	No	Yes	00:00:32	11
E01	Yes	No	No	No	Yes	00:00:31	12
Raw	Yes	Yes	Yes	Yes	Yes	00:01:00	13
E01	Yes	Yes	Yes	Yes	Yes	00:01:02	14
E01	Yes	Yes	No	No	No	00:00:30	15
E01	Yes	Yes	Yes	No	No	00:00:30	16
E01	Yes	No	Yes	No	No	00:00:30	17
E01	No	Yes	No	Yes	No	00:01:00	18
E01	Yes	Yes	No	Yes	No	00:01:00	19
E01	Yes	Yes	Yes	Yes	No	00:01:00	20
E01	No	No	Yes	Yes	No	00:01:00	21
E01	No	Yes	No	No	Yes	00:00:31	22
E01	No	No	Yes	No	Yes	00:00:31	23

The type of image (raw or E01) and the creation of the various hashes had little effect on the time needed to complete the task. The only option that had a significant effect on the time was re-reading the source after acquisition. This took about twice as long, whether or not I chose to add other options, such as MD5 hash. In order to select the option to re-read the source after acquisition, the user must also select at least one of the hash options. The user is not able to select only “Re-read source after acquisition.” Interestingly, although the dialog box in Guymager tells the user that verifying the image after acquisition will take twice as long, the results do not support this.

Table 2: Creating BitCurator reports from Disk Image 14 of Floppy Disk 1

Bulk extractor, with default options	00:00:07
Run All	00:00:04
Fiwalk	00:00:06
Annotate File Names	00:00:02
BitCurator reports	00:00:03
Bulk extractor, with default options and word list scanner	00:00:04
Bulk extractor, with just email scanner	00:00:01

Next, I created BitCurator reports, using several different variations. I used Image 14 from the above task to create these reports.

I first had to run the image through *bulk extractor*. I used the default options, with the default scanners. As discussed previously, *bulk extractor* scans a disk image and extracts information from it without parsing the file system. It is important to run *bulk extractor* first before attempting to create other reports because the other reports use the output from *bulk extractor*. I used the default options, with the default scanners.[214]I then used the “Run All” option to create the reports. Using the run all tab means that the program creates the *fiwalk* report, the annotate fie names reports, and the BitCurator reports. [215]

[214] For a full list of the scanners and their functions, please see: http://wiki.bitcurator.net/index.php?title=Bulk_Extractor_Scanners

[215] BitCurator Wiki, “Using the Run All Tab” http://wiki.bitcurator.net/index.php?title=Using_the_Run_All_Tab. Accessed March 2014.

I created a report using just *fiwalk*. *Fiwalk*, as discussed previously, is "a program that processes a disk image using the SleuthKit library and outputs its results in Digital Forensics XML."[216]

I used the annotate file names option. This step matches the features found by *bulk extractor* to the file they are in on the disk image. This is necessary since *bulk extractor* ignores the file system and instead just scans the raw bit stream.[217]

I then ran the BitCurator reports. The BitCurator report combines the outputs of *bulk extractor, fiwalk* and the annotation tool to "generate both machine and human readable reports that can be read directly or crosswalked to other archival tools." [218]

I ran the image through *bulk extractor* again, this time using the default scanners, plus an addition scanner, word list, which creates a list of all the words found on the disk.[219]

I ran the image through *bulk extractor* again, this time using just one scanner, the email scanner, which "discovers RFC822 email headers, HTTP cookies, hostnames, IP addresses, email addresses, and URLs" and is "useful for recreating email correspondence on a device."[220]

I then loaded the image into a case in FTK. Again, FTK allows for several different options when loading in a case. For example, the user can chose to use MD5

[216] Forensics Wiki, "Fiwalk." http://www.forensicswiki.org/wiki/Fiwalk. Accessed March 2014.
[217] BitCurator Wiki, "Generating An Annotated Features Report." http://wiki.bitcurator.net/index.php?title=Generating_an_Annotated_Features_Report. Accessed March 2014.
[218] BitCurator Wiki, "Generating BitCurator Forensic Reports." http://wiki.bitcurator.net/index.php?title=Generating_BitCurator_Forensic_Reports. Accessed March 2014.
[219] Bit Curator Wiki, "Bulk Extractor Scanners." http://wiki.bitcurator.net/index.php?title=Bulk_Extractor_Scanners Accessed March 2014.
[220] Bit Curator Wiki, "Bulk Extractor Scanners."

hash, SHA-1 Hash, SHA-256 hash, or all three, or none, or two. The user can choose (among other options) whether or not to:

- Flag duplicate files
- Run file signature analysis, which "analyzes files to indicate whether their headers or signatures match their extensions."[221]
- Flag bad extensions, which "identifies files whose types do not match their extensions, based on the file header information"[222]
- Generate a dtSearch Text Index. Doing so will allow you to an index search of acquired images.[223] An index search will allow you to search for discrete words or number strings in the allocated and unallocated space on a disk image.[224] Dtsearch is "one of the leading search tools available" and "can quickly search gigabtyes of text."[225]
- Create thumbnails for graphics
- Optical Character Recognition (OCR) on the files. OCR "scans graphics files for text and converts graphics-text into actual text. That text can then be indexed, searched and treated as any other text in the case."[226] Note that OCR is only used for graphics files.

I loaded the disk image into FTK while selecting and de-selecting these options, in several different combinations. I again used Disk Image 14 which was taken from Floppy Disk 1. The results of these tasks can be seen in Table 3 below.

[221] AccessData, "Forensic ToolKit: User Guide" http://cse.spsu.edu/raustin2/coursefiles/forensics/FTK_UG_3-4-1.pdf. Accessed March 2014.
[222] AccessData.
[223] AccessData.
[224] AccessData.
[225] AccessData.
[226] AccessData.

Table 3: Adding Disk Image 14 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:26
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:39
Yes	Yes	Yes	No	No	No	No	No	No	00:00:27
No	No	No	No	No	No	No	No	Yes	00:00:29
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:29
No	No	No	No	No	No	No	Yes	No	00:00:27
No	No	No	No	Yes	Yes	No	No	No	00:00:27
Yes	No	No	Yes	No	No	No	No	No	00:00:26
No	No	No	No	Yes	No	Yes	No	No	00:00:38

When the user selects to check for bad file extensions or to use the dtSearch text index, the program automatically selects the option for file signature analysis, and the user cannot un-select it. The two must be performed together. Also, when the user selects to flag duplicates, the user must also select to use the MD5 hash option. This is because creating a hash is what allows the system to flag the duplicates.

The second disk was a 3.5 inch floppy disk with 1.5 MB of data. It came from the same collection as the first and was labelled “Modferty + Famferty.” It is assumed that this label was created by the producer. The second disk contained 47 files in all and 1.5 MB of data. The disk contained 23 DOC files. The rest of the files had the file extension .FIG and appeared to be figures. The disk contained 4 directories and one deleted file.

There were no files on the disk bigger than 1 MB. The disk also contained document files and data files. The files ranged in size from 6656 bytes to 49,152 bytes. The documents were all last modified between 1992 and 1997. The disk also used the FAT 12 file system.

I again created a disk image of the floppy disk in Guymager, selecting and de-selecting the many options previously discussed, in a variety of combinations. The time needed to complete these tasks can be seen in Table 4.

Table 4: Results of creating disk image in Guymager for Floppy Disk 2

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:30	24
E01	No	No	No	No	No	00:00:31	25
Raw	Yes	No	No	No	No	00:00:29	26
E01	Yes	No	No	No	No	00:00:30	27
Raw	No	Yes	No	No	No	00:00:30	28
E01	No	Yes	No	No	No	00:00:30	29
Raw	No	No	Yes	No	No	00:00:29	30
E01	No	No	Yes	No	No	00:00:31	31
Raw	Yes	No	No	Yes	No	00:01:00	32
E01	Yes	No	No	Yes	No	00:01:01	33
Raw	Yes	No	No	No	Yes	00:00:30	34
E01	Yes	No	No	No	Yes	00:00:30	35

Raw	Yes	Yes	Yes	Yes	Yes	00:01:00	36
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	37
E01	Yes	Yes	No	No	No	00:00:30	38
E01	Yes	Yes	Yes	No	No	00:00:30	39
E01	Yes	No	Yes	No	No	00:00:30	40
E01	No	Yes	No	Yes	No	00:01:00	41
E01	Yes	Yes	No	Yes	No	00:01:00	42
E01	Yes	Yes	Yes	Yes	No	00:01:00	43
E01	No	No	Yes	Yes	No	00:01:01	44
E01	No	Yes	No	No	Yes	00:00:30	45
E01	No	No	Yes	No	Yes	00:00:30	46

As with the first floppy disk, I then generated reports using BitCurator, and the time needed to complete these tasks can be seen in Table 5.

Table 5: Creating BitCurator reports from Disk Image 37 of Floppy Disk 2

Bulk extractor, with default options	00:00:03
Run All	00:00:06
Fiwalk	00:00:02
Annotate File Names	00:00:02
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:03
Bulk extractor, with just email scanner	00:00:02

I then loaded Disk Image 37 as evidence into a case in FTK, using a variety of available options, as with the previous floppy disk, and the results can be seen in Table 6.

Table 6: Adding Disk Image 37 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:29
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:34
Yes	Yes	Yes	No	No	No	No	No	No	00:00:31
No	No	No	No	No	No	No	No	Yes	00:00:25
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:30
No	No	No	No	No	No	No	Yes	No	00:00:29
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:29
No	No	No	No	Yes	No	Yes	No	No	00:00:33

The third disk was a 3.5 inch floppy disk with 1.5 MB of data. It was labelled and came from the same collection as the previous ones and was labeled “Programs for Demographic Analysis with Compliments of: Population Research Laboratory, The University of Alberta, Edmonton, Alberta, Canada T6G 2H4.” It is assumed that this label was created by the producer. It contained approximately 1.5 MB of data. It contained a variety of file formats and sizes. It also included one deleted file. It did not include any files larger than 1 MB. The disk used the FAT 12 file system. The disk contained 38 data files, 12 files in the lotus 1-2-3 wk1 document data format, 2 Corel WordPerfect files, 4 SysEx files, 2 DOS floppy hard disk boot sector files, 32 MS-DOS executable files, 4 files of ASCII English text and 4 empty files. The files ranged in size from 158 bytes to 58928 bytes. The files were all last modified in 1993.

I created a disk image of the third floppy disk using Guymager, and the results can be seen in Table 7.

Table 7: Results of creating disk image in Guymager for Floppy Disk 3

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:30	47
E01	No	No	No	No	No	00:00:31	48
Raw	Yes	No	No	No	No	00:00:31	49
E01	Yes	No	No	No	No	00:00:31	50
Raw	No	Yes	No	No	No	00:00:30	51
E01	No	Yes	No	No	No	00:00:31	52
Raw	No	No	Yes	No	No	00:00:30	53
E01	No	No	Yes	No	No	00:00:30	54
Raw	Yes	No	No	Yes	No	00:01:00	55
E01	Yes	No	No	Yes	No	00:00:59	56
Raw	Yes	No	No	No	Yes	00:00:30	57
E01	Yes	No	No	No	Yes	00:00:30	58
Raw	Yes	Yes	Yes	Yes	Yes	00:01:01	59
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	60

E01	Yes	Yes	No	No	No	00:00:32	61
E01	Yes	Yes	Yes	No	No	00:00:30	62
E01	Yes	No	Yes	No	No	00:00:31	63
E01	No	Yes	No	Yes	No	00:01:01	64
E01	Yes	Yes	No	Yes	No	00:01:01	65
E01	Yes	Yes	Yes	Yes	No	00:01:00	66
E01	No	No	Yes	Yes	No	00:01:01	67
E01	No	Yes	No	No	Yes	00:00:31	68
E01	No	No	Yes	No	Yes	00:00:30	69

I also generated reports of Disk Image 60 using the BitCurator environment. The time needed to complete these tasks can be seen in Table 8.

Table 8: Creating BitCurator reports from Disk Image 60 of Floppy Disk 3

Bulk extractor, with default options	00:00:02
Run All	00:00:02
Fiwalk	00:00:01
Annotate File Names	00:00:01
BitCurator reports	00:00:01
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 60 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 9.

Table 9: Adding Disk Image 60 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:32
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:37
Yes	Yes	Yes	No	No	No	No	No	No	00:00:39
No	No	No	No	No	No	No	No	Yes	00:00:27
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:34
No	No	No	No	No	No	No	Yes	No	00:00:33
No	No	No	No	Yes	Yes	No	No	No	00:00:33
Yes	No	No	Yes	No	No	No	No	No	00:00:35
No	No	No	No	Yes	No	Yes	No	No	00:00:36

The fourth disk was a 3.5-inch floppy disk with approximately 1.5 MB of data. It was labelled “Gateway 2000 Mach 64 Drivers and Utilities, Disk 2 of 3, Version 1.43, 12/5/94.” This disk was also part of the same collection as the previous one. The disk contained five files with the LZH extension (LZH compressed), one Exe file, two HLP files (help files), one SYS file, one SCR file, one NT file, one INF file, and on DLL file. The files ranged in size from 997 bytes (a LZH file) to 195142 bytes (also an LZH file). There was one partition that contained a deleted or unallocated file. The disk used the FAT 12 file system, and the files were last modified in 1994.

I created a disk image of the fourth floppy disk using Guymager, and the results can be seen in Table 10.

Table 10: Results of creating disk image in Guymager for Floppy Disk 4

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:29	70
E01	No	No	No	No	No	00:00:32	71
Raw	Yes	No	No	No	No	00:00:30	72
E01	Yes	No	No	No	No	00:00:30	73
Raw	No	Yes	No	No	No	00:00:31	74
E01	No	Yes	No	No	No	00:00:31	75
Raw	No	No	Yes	No	No	00:00:31	76
E01	No	No	Yes	No	No	00:00:31	77
Raw	Yes	No	No	Yes	No	00:00:59	78
E01	Yes	No	No	Yes	No	00:01:01	79
Raw	Yes	No	No	No	Yes	00:00:30	80
E01	Yes	No	No	No	Yes	00:00:30	81
Raw	Yes	Yes	Yes	Yes	Yes	00:00:59	82
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	83
E01	Yes	Yes	No	No	No	00:00:31	84
E01	Yes	Yes	Yes	No	No	00:00:30	85
E01	Yes	No	Yes	No	No	00:00:30	86
E01	No	Yes	No	Yes	No	00:01:00	87
E01	Yes	Yes	No	Yes	No	00:01:00	88

E01	Yes	Yes	Yes	Yes	No	00:01:00	89
E01	No	No	Yes	Yes	No	00:01:01	90
E01	No	Yes	No	No	Yes	00:00:31	91
E01	No	No	Yes	No	Yes	00:00:30	92

I also generated reports of Disk Image 83 using the BitCurator environment. The time needed to complete these tasks can be seen in Table 11

Table 11: Creating BitCurator Reports from Disk Image 83 of Floppy Disk 4

Bulk extractor, with default options	00:00:01
Run All	00:00:03
Fiwalk	00:00:01
Annotate File Names	00:00:01
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 83 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 12.

Table 12: Adding Disk Image 83 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:29
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:33
Yes	Yes	Yes	No	No	No	No	No	No	00:00:29
No	No	No	No	No	No	No	No	Yes	00:00:27

Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:30
No	No	No	No	No	No	No	Yes	No	00:00:28
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:29
No	No	No	No	Yes	No	Yes	No	No	00:00:33

The fifth disk was not part of the same collection as the previous disks and was instead part of a collection of disk used by the digital forensics lab at the University of North Carolina School of Information and Library Science. I used this disk to create variety in my sample. It was a 3.5 inch floppy disk and contained 737.3 KB of data. It was labelled “Harry S. Truman Library, oral histories.” The disk contained 18 Corel Word Perfect files, four files of data, tow DOS_tor DOS floppy hard disk booter files and two deleted files. The files ranged in size from 56,279 bytes to 117,428 bytes. The files were last modified in 1994 and used the FAT 12 file system.

I created a disk image of this floppy disk using Guymager and the many options allowed by Guymager. The results of these tasks can be seen in Table 13.

Table 13: Results of creating disk image in Guymager for Floppy Disk 5

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:28	93
E01	No	No	No	No	No	00:00:28	94
Raw	Yes	No	No	No	No	00:00:28	95
E01	Yes	No	No	No	No	00:00:29	96

Raw	No	Yes	No	No	No	00:00:28	97
E01	No	Yes	No	No	No	00:00:28	98
Raw	No	No	Yes	No	No	00:00:28	99
E01	No	No	Yes	No	No	00:00:29	100
Raw	Yes	No	No	Yes	No	00:00:56	101
E01	Yes	No	No	Yes	No	00:00:56	102
Raw	Yes	No	No	No	Yes	00:00:28	103
E01	Yes	No	No	No	Yes	00:00:28	104
Raw	Yes	Yes	Yes	Yes	Yes	00:00:55	105
E01	Yes	Yes	Yes	Yes	Yes	00:00:56	106
E01	Yes	Yes	No	No	No	00:00:29	107
E01	Yes	Yes	Yes	No	No	00:00:28	108
E01	Yes	No	Yes	No	No	00:00:28	109
E01	No	Yes	No	Yes	No	00:00:56	110
E01	Yes	Yes	No	Yes	No	00:00:57	111
E01	Yes	Yes	Yes	Yes	No	00:00:57	112
E01	No	No	Yes	Yes	No	00:00:56	113
E01	No	Yes	No	No	Yes	00:00:28	114
E01	No	No	Yes	No	Yes	00:00:27	115

I also generated reports of Disk Image 106 using BitCurator. The time needed to complete these tasks can be seen in Table 14.

Table 14: Creating BitCurator Reports from Disk Image 106 of Floppy Disk 5

Bulk extractor, with default options	00:00:01
Run All	00:00:02
Fiwalk	00:00:01
Annotate File Names	00:00:01
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 106 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 15.

Table 15: Adding Disk Image 106 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:29
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:34
Yes	Yes	Yes	No	No	No	No	No	No	00:00:29
No	No	No	No	No	No	No	No	Yes	00:00:29
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:29
No	No	No	No	No	No	No	Yes	No	00:00:29
No	No	No	No	Yes	Yes	No	No	No	00:00:29
Yes	No	No	Yes	No	No	No	No	No	00:00:29
No	No	No	No	Yes	No	Yes	No	No	00:00:33

The sixth disk was also part of the materials available for use in the digital forensics laboratory at the University of North Carolina School of Information and Library Science. The floppy disk was labelled “Stever Papers Finding Aid. STEVER.PAP. Ford Library.” It is assumed that the label was created by the producer of the item. The floppy disk contained 737.3 KB of data. The disk contained 15 allocated files and 15 deleted files. It used the FAT 12 file system and contained no files larger than 1 MB. The largest file on the disk was called STEVER.PAP and contained 180331 bytes. The smallest file on the disk was called CTOOLS.BAT and contained 51 bytes. The files were mostly last modified in 1991, with one file being last modified in 1994. The deleted files included EXE file formats, as walls as .BAT, .CFG, .TXT, .COM, and .OVL.

I created a disk image for this floppy disk using Guymager, selecting and deselecting the various options. The time needed to complete these tasks can be seen in Table 16.

Table 16: Results of creating disk image in Guymager for Floppy Disk 6

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:28	116
E01	No	No	No	No	No	00:00:29	117
Raw	Yes	No	No	No	No	00:00:27	118
E01	Yes	No	No	No	No	00:00:28	119
Raw	No	Yes	No	No	No	00:00:28	120
E01	No	Yes	No	No	No	00:00:29	121

Raw	No	No	Yes	No	No	00:00:28	122
E01	No	No	Yes	No	No	00:00:28	123
Raw	Yes	No	No	Yes	No	00:00:56	124
E01	Yes	No	No	Yes	No	00:00:57	125
Raw	Yes	No	No	No	Yes	00:00:56	126
E01	Yes	No	No	No	Yes	00:00:29	127
Raw	Yes	Yes	Yes	Yes	Yes	00:00:54	128
E01	Yes	Yes	Yes	Yes	Yes	00:00:56	129
E01	Yes	Yes	No	No	No	00:00:28	130
E01	Yes	Yes	Yes	No	No	00:00:28	131
E01	Yes	No	Yes	No	No	00:00:28	132
E01	No	Yes	No	Yes	No	00:00:56	133
E01	Yes	Yes	No	Yes	No	00:00:57	134
E01	Yes	Yes	Yes	Yes	No	00:00:55	135
E01	No	No	Yes	Yes	No	00:00:56	136
E01	No	Yes	No	No	Yes	00:00:29	137
E01	No	No	Yes	No	Yes	00:00:28	138

I also generated reports of Disk Image 129 using BitCurator. The time needed to complete these tasks can be seen in Table 17.

Table 17: Creating BitCurator Reports from Disk Image 129 of Floppy Disk 6

Bulk extractor, with default options	00:00:01

Run All	00:00:03
Fiwalk	00:00:01
Annotate File Names	00:00:01
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 129 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 18.

Table 18: Adding Disk Image 129 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:29
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:33
Yes	Yes	Yes	No	No	No	No	No	No	00:00:31
No	No	No	No	No	No	No	No	Yes	00:00:24
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:30
No	No	No	No	No	No	No	Yes	No	00:00:30
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:29
No	No	No	No	Yes	No	Yes	No	No	00:00:32

The seventh disk was labeled “Program for 1993 Survey” and contained 1.5 MB of data. This disk came from the same collection as Disks 1-4 and was part of a collection from the Southern Historical Collection at the University of North Carolina at Chapel

Hill. It contained four PRG files, one DAT file, three PGM files, and one DAT file. The files ranged in size from 322 bytes to about 480,422 bytes.

Table 19: Results of creating disk image in Guymager for Floppy Disk 7

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:30	139
E01	No	No	No	No	No	00:00:31	140
Raw	Yes	No	No	No	No	00:00:31	141
E01	Yes	No	No	No	No	00:00:31	142
Raw	No	Yes	No	No	No	00:00:30	143
E01	No	Yes	No	No	No	00:00:31	144
Raw	No	No	Yes	No	No	00:00:31	145
E01	No	No	Yes	No	No	00:00:30	146
Raw	Yes	No	No	Yes	No	00:01:00	147
E01	Yes	No	No	Yes	No	00:01:00	148
Raw	Yes	No	No	No	Yes	00:00:30	149
E01	Yes	No	No	No	Yes	00:00:30	150
Raw	Yes	Yes	Yes	Yes	Yes	00:01:01	151
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	152
E01	Yes	Yes	No	No	No	00:00:30	153
E01	Yes	Yes	Yes	No	No	00:00:31	154

E01	Yes	No	Yes	No	No	00:00:31	155
E01	No	Yes	No	Yes	No	00:01:00	156
E01	Yes	Yes	No	Yes	No	00:01:00	157
E01	Yes	Yes	Yes	Yes	No	00:01:01	158
E01	No	No	Yes	Yes	No	00:01:00	159
E01	No	Yes	No	No	Yes	00:00:30	160
E01	No	No	Yes	No	Yes	00:00:30	161

I also generated reports of Disk Image 152 using BitCurator. The time needed to complete these tasks can be seen in Table 20.

Table 20: Creating BitCurator reports from Disk Image 154 of Floppy Disk 7

Bulk extractor, with default options	00:00:01
Run All	00:00:03
Fiwalk	00:00:02
Annotate File Names	00:00:02
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:01
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 152 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 21.

Table 21: Adding Disk Image 152 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:30

Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:34
Yes	Yes	Yes	No	No	No	No	No	No	00:00:29
No	No	No	No	No	No	No	No	Yes	00:00:23
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:29
No	No	No	No	No	No	No	Yes	No	00:00:30
No	No	No	No	Yes	Yes	No	No	No	00:00:29
Yes	No	No	Yes	No	No	No	No	No	00:00:30
No	No	No	No	Yes	No	Yes	No	No	00:00:33

The eighth disk was labeled simply "1" and contained approximately 1.5 MB of data. It came from the same collection as the previous disk. It contained only two files, both with the extension .001. One file contained 1,456,128 bytes and the other contained 1,271 bytes.

Table 22: Results of creating disk image in Guymager for Floppy Disk 8

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:30	162
E01	No	No	No	No	No	00:00:30	163
Raw	Yes	No	No	No	No	00:00:31	164
E01	Yes	No	No	No	No	00:00:29	165
Raw	No	Yes	No	No	No	00:00:31	166
E01	No	Yes	No	No	No	00:00:30	167

Raw	No	No	Yes	No	No	00:00:31	168
E01	No	No	Yes	No	No	00:00:29	169
Raw	Yes	No	No	Yes	No	00:01:00	170
E01	Yes	No	No	Yes	No	00:00:59	171
Raw	Yes	No	No	No	Yes	00:00:31	172
E01	Yes	No	No	No	Yes	00:00:31	173
Raw	Yes	Yes	Yes	Yes	Yes	00:01:00	174
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	175
E01	Yes	Yes	No	No	No	00:00:31	176
E01	Yes	Yes	Yes	No	No	00:00:30	177
E01	Yes	No	Yes	No	No	00:00:31	178
E01	No	Yes	No	Yes	No	00:01:01	179
E01	Yes	Yes	No	Yes	No	00:01:00	180
E01	Yes	Yes	Yes	Yes	No	00:01:00	181
E01	No	No	Yes	Yes	No	00:01:01	182
E01	No	Yes	No	No	Yes	00:00:31	183
E01	No	No	Yes	No	Yes	00:00:31	184

I also generated reports of Disk Image 175 using BitCurator. The time needed to complete these tasks can be seen in Table 23.

Table 23: Creating BitCurator reports from Disk Image 175 of Floppy Disk 8

Bulk extractor, with default options	00:00:01
Run All	00:00:03

Fiwalk	00:00:02
Annotate File Names	00:00:02
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I loaded Disk Image 175 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 24.

Table 24: Adding Disk Image 175 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:30
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:34
Yes	Yes	Yes	No	No	No	No	No	No	00:00:29
No	No	No	No	No	No	No	No	Yes	00:00:23
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:29
No	No	No	No	No	No	No	Yes	No	00:00:29
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:30
No	No	No	No	Yes	No	Yes	No	No	00:00:34

The ninth disk was not labeled and was part of the same collection as the previous disk. It contained eight files with the extension SYS, four SSD files, two EXE files, and a variety of other file formats, including LQ, NLQ, BAK, HLP, and CMD. The files ranged in size from 1,063,372 bytes to 1 byte.

Table 25: Results of creating disk image in Guymager for Floppy Disk 9

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:31	185
E01	No	No	No	No	No	00:00:30	186
Raw	Yes	No	No	No	No	00:00:30	187
E01	Yes	No	No	No	No	00:00:30	188
Raw	No	Yes	No	No	No	00:00:30	189
E01	No	Yes	No	No	No	00:00:31	190
Raw	No	No	Yes	No	No	00:00:30	191
E01	No	No	Yes	No	No	00:00:30	192
Raw	Yes	No	No	Yes	No	00:01:01	193
E01	Yes	No	No	Yes	No	00:01:00	194
Raw	Yes	No	No	No	Yes	00:00:30	195
E01	Yes	No	No	No	Yes	00:00:30	196
Raw	Yes	Yes	Yes	Yes	Yes	00:00:59	197
E01	Yes	Yes	Yes	Yes	Yes	00:00:59	198
E01	Yes	Yes	No	No	No	00:00:30	199
E01	Yes	Yes	Yes	No	No	00:00:31	200

E01	Yes	No	Yes	No	No	00:00:31	201
E01	No	Yes	No	Yes	No	00:01:00	202
E01	Yes	Yes	No	Yes	No	00:01:00	203
E01	Yes	Yes	Yes	Yes	No	00:01:00	204
E01	No	No	Yes	Yes	No	00:01:00	205
E01	No	Yes	No	No	Yes	00:00:31	206
E01	No	No	Yes	No	Yes	00:00:31	207

I also generated reports of Disk Image 198 using the BitCurator environment. The time needed to complete these tasks can be seen in Table 26.

Table 26: Creating BitCurator reports from Disk Image 198 from Floppy Disk 9

Bulk extractor, with default options	00:00:01
Run All	00:00:03
Fiwalk	00:00:02
Annotate File Names	00:00:01
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:01
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 198 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 27.

Table 27: Adding Disk Image 198 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:31

Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:33
Yes	Yes	Yes	No	No	No	No	No	No	00:00:30
No	No	No	No	No	No	No	No	Yes	00:00:30
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:30
No	No	No	No	No	No	No	Yes	No	00:00:30
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:30
No	No	No	No	Yes	No	Yes	No	No	00:00:35

The tenth disk was labeled ““Ques 137. DAT, 138, 138” and was part of the same collection as the previous disk. The file contained three large .DAT files which were each about 470,000 bytes in size.

Table 28: Results of creating disk image in Guymager for Floppy Disk 10

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number
Raw	No	No	No	No	No	00:00:30	208
E01	No	No	No	No	No	00:00:31	209
Raw	Yes	No	No	No	No	00:00:30	210
E01	Yes	No	No	No	No	00:00:30	211
Raw	No	Yes	No	No	No	00:00:30	212
E01	No	Yes	No	No	No	00:00:30	213

Raw	No	No	Yes	No	No	00:00:31	214
E01	No	No	Yes	No	No	00:00:32	215
Raw	Yes	No	No	Yes	No	00:01:01	216
E01	Yes	No	No	Yes	No	00:01:01	217
Raw	Yes	No	No	No	Yes	00:00:30	218
E01	Yes	No	No	No	Yes	00:00:31	219
Raw	Yes	Yes	Yes	Yes	Yes	00:01:00	220
E01	Yes	Yes	Yes	Yes	Yes	00:01:00	221
E01	Yes	Yes	No	No	No	00:00:31	222
E01	Yes	Yes	Yes	No	No	00:00:30	223
E01	Yes	No	Yes	No	No	00:00:31	224
E01	No	Yes	No	Yes	No	00:01:00	225
E01	Yes	Yes	No	Yes	No	00:01:00	226
E01	Yes	Yes	Yes	Yes	No	00:01:00	227
E01	No	No	Yes	Yes	No	00:01:00	228
E01	No	Yes	No	No	Yes	00:00:30	229
E01	No	No	Yes	No	Yes	00:00:31	230

I generated reports of Disk Image 221 using the BitCurator environment. The time needed to complete these tasks can be seen in Table 29.

Table 29: Creating BitCurator reports from Disk Image 221 of Floppy Disk 10

Bulk extractor, with default options	00:00:02

Run All	00:00:03
Fiwalk	00:00:01
Annotate File Names	00:00:01
BitCurator reports	00:00:02
Bulk extractor, with default options and word list scanner	00:00:02
Bulk extractor, with just email scanner	00:00:01

I also loaded Disk Image 221 as an evidence item in a case in FTK, using a variety of options. The time needed to complete these tasks can be seen in Table 30.

Table 30: Adding Disk Image 221 to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature Analysis	Flag Bad Ext.	dtSearch Text Index	Create Thumbnails for graphics	OCR	Time
No	No	No	No	No	No	No	No	No	00:00:30
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:34
Yes	Yes	Yes	No	No	No	No	No	No	00:00:30
No	No	No	No	No	No	No	No	Yes	00:00:30
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:30
No	No	No	No	No	No	No	Yes	No	00:00:30
No	No	No	No	Yes	Yes	No	No	No	00:00:30
Yes	No	No	Yes	No	No	No	No	No	00:00:30
No	No	No	No	Yes	No	Yes	No	No	00:00:33

After completing these tasks using the floppy disk, I had intended to complete the same set of tasks using an external 2-terabyte hard drive, which also came from the Southern Historical Collection at the University of North Carolina. This disk contained a variety of file formats, including JPEG and TIFF image files, Microsoft Word 97 files, WordPerfect files, Adobe Acrobat files, HTML files, Access 2000 files, Excel 97 files, some unknown file types. It also contained unallocated space as well as slack space. There were 47,742 items on the disk. There were approximately 600 JPEG files on the disk, which were each around 100 KB. There were also approximately five hundred TIFF image files, which ranged in size from 4096 bytes to 134.1 MB. The drive also contained about two hundred Adobe Acrobat files, which ranged in size from about 98 KB to about 4 MB. The large number of images on the hard drive, as well as these large Adobe Acrobat files, may help to explain why it took so long for the program to create a disk image of the drive.

When I created a disk image of the hard drive, it took 97 hours, 22 minutes, and 33 seconds. It was created as an E01 file with the default options for FTK Imager. By default, the program does not select other options, such as to verify the image after it is created, or to create a directory listing of all the files in the image after the image is created. As such, I did not have any of these options selected when I created this disk image. After that, it became clear that I would not have time to complete the same series of tasks, and so I decided that I would not be able to do so many variations.

I also loaded the image as evidence for a case in FTK. I did not select any of the possible options for this task, such as created an MD5 hash. It took 40 minutes and 12 seconds to complete this task.

In Table 31, I have compiled all the fastest times (that is, requiring the least amount of time) in which the different disk image tasks were completed. If multiple disk images required the same amount of time, I have included them both.

Table 31: Fastest Time For Disk Image Creation

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number(s)
Raw	No	No	No	No	No	00:00:28	93, 116
E01	No	No	No	No	No	00:00:28	94
Raw	Yes	No	No	No	No	00:00:27	118
E01	Yes	No	No	No	No	00:00:28	119
Raw	No	Yes	No	No	No	00:00:28	97, 120
E01	No	Yes	No	No	No	00:00:28	98
Raw	No	No	Yes	No	No	00:00:28	99, 122
E01	No	No	Yes	No	No	00:00:28	123
Raw	Yes	No	No	Yes	No	00:00:56	101, 124
E01	Yes	No	No	Yes	No	00:00:56	102
Raw	Yes	No	No	No	Yes	00:00:28	103
E01	Yes	No	No	No	Yes	00:00:28	104, 127
Raw	Yes	Yes	Yes	Yes	Yes	00:00:54	128
E01	Yes	Yes	Yes	Yes	Yes	00:00:56	106, 129
E01	Yes	Yes	No	No	No	00:00:28	130
E01	Yes	Yes	Yes	No	No	00:00:28	108, 131
E01	Yes	No	Yes	No	No	00:00:28	109, 132
E01	No	Yes	No	Yes	No	00:00:56	110, 133

E01	Yes	Yes	No	Yes	No	00:00:57	111. 134
E01	Yes	Yes	Yes	Yes	No	00:00:55	135
E01	No	No	Yes	Yes	No	00:00:56	113, 136
E01	No	Yes	No	No	Yes	00:00:28	114
E01	No	No	Yes	No	Yes	00:00:27	115

In table 32, I have listed all the slowest times for disk image completion (that is, the tasks that required the most time to complete).

Table 32: Slowest Time for Disk Image Creation

Image format	MD5	SHA-1	SHA-256	Re-read source	Verify image	Time	Image number(s)
Raw	No	No	No	No	No	00:00:33	1
E01	No	No	No	No	No	00:00:32	2
Raw	Yes	No	No	No	No	00:00:33	3
E01	Yes	No	No	No	No	00:00:32	4
Raw	No	Yes	No	No	No	00:00:32	5
E01	No	Yes	No	No	No	00:00:31	6, 52, 75, 144, 190
Raw	No	No	Yes	No	No	00:00:32	7
E01	No	No	Yes	No	No	00:00:33	8
Raw	Yes	No	No	Yes	No	00:01:02	10
E01	Yes	No	No	Yes	No	00:01:03	11

Raw	Yes	No	No	No	Yes	00:00:32	12
E01	Yes	No	No	No	Yes	00:00:31	13, 173, 219
Raw	Yes	Yes	Yes	Yes	Yes	00:01:01	59, 151
E01	Yes	Yes	Yes	Yes	Yes	00:01:02	15
E01	Yes	Yes	No	No	No	00:00:32	61
E01	Yes	Yes	Yes	No	No	00:00:31	154, 200
E01	Yes	No	Yes	No	No	00:00:31	63, 155, 178, 201, 224
E01	No	Yes	No	Yes	No	00:01:01	64, 179
E01	Yes	Yes	No	Yes	No	00:01:01	65
E01	Yes	Yes	Yes	Yes	No	00:01:01	158
E01	No	No	Yes	Yes	No	00:01:01	44, 67, 182
E01	No	Yes	No	No	Yes	00:00:31	23, 68, 91, 183, 206
E01	No	No	Yes	No	Yes	00:00:31	23, 184, 207, 230

In Table 33, I have indicated the fastest times in which reports were generated in the BitCurator environment, and from what image or images they were generated .

Table 33: Fastest Time for Creating Reports in BitCurator

Task	*Time*	*Image Number(s)*
Bulk extractor, with default options	00:00:01	83, 106, 129, 152, 175, 198,
Run All	00:00:02	60, 106
Fiwalk	00:00:01	60, 83, 106, 129, 221
Annotate File Names	00:00:01	60, 83, 106, 129, 198, 221
BitCurator reports	00:00:01	60
Bulk extractor, with default options and word list scanner	00:00:01	152, 198
Bulk extractor, with just email scanner	00:00:01	15, 60, 83, 106, 129, 152, 175, 198, 221

In Table 34, I have indicated the slowest times in which reports were generated in the BitCurator environment, and from what image or images they were generated .

Table 34: Slowest Time for Creating Report in BitCurator

Task	*Time*	*Image Number(s)*
Bulk extractor, with default options	00:00:07	15
Run All	00:00:06	37
Fiwalk	00:00:06	15
Annotate File Names	00:00:02	15, 37, 152, 175
BitCurator reports	00:00:03	15
Bulk extractor, with default options and word list scanner	00:00:04	15
Bulk extractor, with just email scanner	00:00:02	37

Table 35 shows the fastest times in which a disk image was loaded into FTK, and indicates what image was used in the case, as well as what options were applied.

Table 35: Fastest Time for adding Disk Image to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate Files	File Signature	Flag Bad Ext.	dtSearch Text	Create Thumbnails	OCR	Time	Image Number(s)

				Analysis		Index	for graphics			
No	No	No	No	No	No	No	No	No	00:00:26	15
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:33	83, 129, 198
Yes	Yes	Yes	No	No	No	No	No	No	00:00:27	15
No	No	No	No	No	No	No	No	Yes	00:00:23	152, 175
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:29	15, 106, 152, 175
No	No	No	No	No	No	No	Yes	No	00:00:27	15
No	No	No	No	Yes	Yes	No	No	No	00:00:27	15
Yes	No	No	Yes	No	No	No	No	No	00:00:26	15
No	No	No	No	Yes	No	Yes	No	No	00:00:32	129

Table 36 shows the fastest times in which a disk image was loaded into FTK, and indicates what image was used in the case, as well as what options were applied.

Table 36: Slowest Time for adding Disk Image to a case in FTK

MD5	SHA-1	SHA-256	Flag Duplicate	File Signatu	Flag Bad	dtSearch	Create Thumb	OCR	Time	Image Numb

			Files	re Analysis	Ext.	Text Index	nails for graphics			er(s)
No	No	No	No	No	No	No	No	No	00:00:32	60
Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	00:00:39	15
Yes	Yes	Yes	No	No	No	No	No	No	00:00:31	37, 129
No	No	No	No	No	No	No	No	Yes	00:00:30	198, 221
Yes	Yes	Yes	No	Yes	No	No	No	Yes	00:00:34	60
No	No	No	No	No	No	No	Yes	No	00:00:33	60
No	No	No	No	Yes	Yes	No	No	No	00:00:33	60
Yes	No	No	Yes	No	No	No	No	No	00:00:35	60
No	No	No	No	Yes	No	Yes	No	No	00:00:38	15

Discussion

One of the most surprising findings of this experiment was that it took so little time to perform the different tasks on the floppy disks. Many of the tasks took only half a minute, or even one second, to complete. It was also interesting that the time needed for a task often did not change even when I added different options to the task. For example, when I created a disk image of the floppy disk using Guymager, and selected the options to create an E01 image with a SHA-1 hash, it took 31 seconds to complete the task. When

I selected the same options, plus the option to verify the image, it still took 31 seconds to complete the task.

On the other hand, it was also surprising how long it took to complete the forensic tasks on the external hard drive. As I stated previously, it took almost a hundred hours to create a disk image of a two terabyte hard drive using FTK. Also, the size of the disk made completing the tasks more complicated. For example, because of the size of the hard drive, I had to make sure I saved the image of the hard drive on a completely empty drive on the computer. Initially, I tried to save it on a drive that contained a few other small files, and the program would not complete the task because there was not enough room on that computer drive to save the disk image.

With these results in mind, it would be useful to consider when it would be "worth the time" for an institution to complete a certain task, and when they would want to select certain options for a task. For example, when would it be worth the time to create an MD5 hash when also creating a disk image? When would it be worth the time to use OCR? There are many factors that might affect this decision. For example, how important are the materials to the institution? How do they fit into the collecting scope? Additionally, how will they be used? What kind of resources does the institution have? Does it have the staff and the time to devote to a lengthy project? Do they have the storage capacity to store digital materials? The answers to these questions will be different for each institution, and could also vary across collections. However, with the results of this experiment in mind, it may be possible to begin to answer some of these questions. For example, in some cases, it would be valuable, and worth the time, to create a hash for a disk image. Doing so will help the institution to ensure the authenticity of the materials. It will help them to know that the materials have not been changed because, as previously discussed, the hash will change if the data is changed even a little bit. It might also be valuable to flag duplicate materials. The institution might consider deleting these materials, which would save them storage space. Again, when I selected this option in creating the disk image of the floppy disk, it did not significantly increase the time needed to complete the task.

Additionally, I was able to generate reports in the BitCurator environment about the disk images a few seconds, sometimes in just one second. These reports provide information that could be valuable to a collecting institution, such as metadata about the contents of the disk. Therefore, an institution could access this valuable information with a relatively low time commitment. However, if the institution had not already created a disk image, they would also have to factor in the time needed to do so, as they would need to first have a disc image before generating the reports.

Institutions which are considering acquiring digital materials which have a large volume of storage, like the two terabyte hard drive, might also want to consider the time cost and other complications of these materials. As I mentioned previously, when creating the disk image for the two terabyte hard drive, I had to save the disk to a completely empty drive. I was able to do this because I was using a FRED device, which has many drives available. However, it would be difficult to find that empty space if one was trying to do this task on a more standard desktop computer. Additionally, the process of creating the disk image was very time-consuming, and this may be frustrating to the institution. However, a hard drive such as the one in this experiment might contain valuable materials, and it may be worth the time for an institution to create a disk image of the hard drive.

Limitations

One limitation of this study is that I did not record how long I spent setting up the tasks, including troubleshooting them when they went wrong, or figuring out exactly how to do a certain task. Although I had had previous exposure to the tools I used in this experiment, I still needed some time to familiarize myself with them before completing the tasks, and I did not record the time I spent on these activities. However, I devoted more time to these activities than to many of the tasks I recorded, so it may have been helpful to record them. However, it is also true that time spent on setting up and trouble-shooting would be different for every individual, and every situation – it would depend on the software being used, and how familiar the person is with the particular task they need to do. Therefore, even if I had recorded the time I spend on these activities, it may not have been a reflection of a universal experience.

It was also unfortunate that I was not able to perform more tasks on the external hard drive. However, I would not have been able to complete all those tasks in the allotted time for this study. Just creating the one disc image took several days.

Another limitation of this study is that I did not have much variety of media to use in the testing, just the floppy discs and the external hard drive. It would be beneficial to perform these tests on different types of materials, and to see how long it took to complete the tasks using those materials.

Additionally, there was also not much variety in the contents of the floppy disks I used. Most of the floppy disks were similar in that the contained WordPerfect documents, data files, and other files of ASCII text. I would have liked to use a floppy disk that contained image files, such as JPEG. This would have likely created some variety in the time it took to complete some of the tasks. For example, if I had used OCR on a disk with images, it would have likely taken longer to create these tasks, as OCR is used with graphics.

Additionally, this experiment was not meant to be an exhaustive study of every possible digital forensics task that could be performed, on every possible type of data. There are many data types and forensics tasks that were overlooked. Institutions may find that their data does not perfectly match that used in this experiment. For example, institutions may have data that has a larger file size than the data used in this experiment, or a small file size, or a different type of file format. There are many forms that data can take. There are also a great variety of forensic tasks that can be performed. However, it is still possible that institutions could use these experiments as a guide. These experiments should give institutions a better idea of how time-consuming these tasks are. With these guides, hopefully institutions will be able to estimate the time needed to perform the tasks, while also considering the size of their data or their collections.

Conclusion

Digital forensics, although originally created for law enforcement officials, is now making its way into the world of archives and other cultural heritage institutions. Digital forensic tasks have the potential to help cultural heritage institutions manage, preserve and curate their digital materials. Despite this potential, institutions may be slow to adopt these practices, for a variety of reasons. For example, institutions could be concerned about how these tasks might drain their resources and how much time will be required to complete a certain digital forensics task. However, the only way to know how much time a task will take is to complete the task. Thus, I undertook this experiment, in an effort to determine how long it would take to complete certain tasks.

I performed a series of tasks on several floppy disks and an external hard drive. Often, I performed similar tasks with slight variations, in order to determine how much time would be needed for these variations. Whether an institution would want to apply these same variations would depend on their individual needs, and how much time they wanted to devote to the task. For example, some institutions might find it useful to also flag duplicate materials when creating a disk image, while others might not. However, it is my hope that the results of this experiment will help institutions make these decisions. The results of this experiment will give others an idea of how long it would take to complete a certain task. With this in mind, the institution can then begin to consider whether this task would indeed be worth their time.

Bibliography

AccessData. “Forensic ToolKit: User Guide” http://cse.spsu.edu/raustin2/coursefiles/forensics/FTK_UG_3-4-1.pdf. Accessed March 2014.

AIMS Work Group. “AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.” 2012. http://www.digitalcurationservices.org/aims/white-paper/ Accessed March 2014.

BitCurator Wiki. http://wiki.bitcurator.net. Accessed March 2014.

Carrier, Brian. *File System Forensic Analysis,* Upper Saddle River, NJ: Addison-Wesley, 2005.

Cook, Terry. “Byte-ing Off What You Can Chew: Electronic Records Strategies for Small Archival Institutions,” *Archifacts* (April 2004). http://www.aranz.org.nz/Site/publications/papers_online/terry_cook_paper.aspx (accessed January 2014).

Digital Intelligence. http://www.digitalintelligence.com/products/fred/ Accessed March 2014

Erway, Ricky. "Swatting the Long Tail of Digital Media: A Call for Collaboration." Dublin, OH: OCLC Research, 2012. http://www.oclc.org/content/dam/research/publications/library/2012/2012-08.pdf (accessed January 2014).

Farmer, Dan and Wietse Venema. *Forensic Discovery* Upper Saddle River, NJ: Addison-Wesley, 2005.

Forensics Wiki. http://www.forensicswiki.org. Accessed March 2014.

Garfinkel, Simson. “Digital Forensics*,” American Scientist* 101 (2013): 370-377

Garfinkel, Simson. “Digital media triage with bulk data analysis and *bulk_extractor*.” *Computers and Security 32*: 56-72 (2013) http://simson.net/clips/academic/2013.COSE.bulk_extractor.pdf (accessed January 2014)

Gengenbach, Martin J. “The Way We Do it Here': Mapping Digital Forensics Workflows in Collecting Institutions” A Master’s Paper for the M.S. in L.S degree, August, 2012. http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf (accessed March 2014)

Glisson, Brad, and Rob Maxwell.“A Digital Forensics Workflow” In , “Digital Forensics and Born-Digital Content in Cultural Heritage Collections,” Washington, D.C.: Council on Library and Information Resources, 2010, http://www.clir.org/pubs/reports/pub149/pub149.pdf. Accessed January 2014), p. 16

John, Jeremy Leighton. “Digital Forensics and Preservation,” *DPC Technology Watch Report 12-03*, Digital Preservation Coalition, November 2012. http://www.dpconline.org/component/docman/doc_download/810-dpctw12-03.pdf. Accessed March 2014.

John, Jeremey Leighton. “The Future of Saving Our Past,” *Nature* 459 (June 2009): 775-776. http://www.nature.com/nature/journal/v459/n7248/full/459775a.html. Aaccessed March 2014.

Jones, Andy and Craig Valli. *Building a Digital Forensic Laboratory,* Burlington, MA: Butterworth-Heinemann and Syngress Publishing, Inc., 2009

Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine. “Digital Forensics and Born-Digital Content in Cultural Heritage Collections,” Washington, D.C.: Council on Library and Information Resources, 2010. http://www.clir.org/pubs/reports/pub149/pub149.pdf Accessed January 2014.

Kirschenbaum, Matthew G., Erika L. Farr, Kari M. Kraus, Naomi Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. “Digital Materiality: Preserving Access to Computers as Complete Environments*” Proceedings of the Sixth International Conference on Digital Preservation (iPRES) (October, 2009)*: 106-107, http://mkirschenbaum.files.wordpress.com/2009/10/digitalmaterialityipres2009.pdf, Accessed January 2014

Kolowich, Steve. “Archiving Writer’s Work in the Age of E-Mail,” *The Chronicle of Higher Education* 55:31, April 10, 2009, http://chronicle.com/article/Archiving-Writers-Work-in/22770 Accessed January 2014.

Lee, Christopher A. "Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision." In *Proceedings of the International Council on Archives Congress, Brisbane, Australia, August 20-24, 2012.* http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00290.pdf Accessed March 2014.

Lee, Christopher. “Donor Agreements.” In *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington, D.C.: Council on Library and Information

Resources, 2010, p. 57 http://www.clir.org/pubs/reports/pub149/pub149.pdf. Accessed January 2014.

Lee, Christopher A., Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods. "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions." *D-Lib Magazine* 18: 5/6, May/June 2012, http://www.dlib.org/dlib/may12/lee/05lee.html (accessed January 2014).

Lee, Christopher A. and Kam Woods. "Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage." *In Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: An International Conference on Permanent Access to Digital Documentary Heritage, 26-28 September 2012, Vancouver, British Columbia, Canada,* edited by Luciana Duranti and Elizabeth Shaffer, 298-313: United Nations Educational, Scientific and Cultural Organization, 2013. http://ils.unc.edu/callee/p298-lee.pdf. Accessed March 2014

Redwine, Gabriela, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas. “Born Digital: Guidance for Donors, Dealers, and Archival Repositories” Washington, D.C.: Council on Library and Information Resources, 2013. http://www.clir.org/pubs/reports/pub159/pub159.pdf Accessed March 2014.

Rogers, Corinne and Jeremey Leighton John. “Shared Perspectives, Common Challenges: A History of Digital Forensics and Ancestral Computing for Digital Heritage.” *Proceedings of the Memory of the World in the Digital Age: Digitization and Preservation Conference, Vancouver, British Columbia, Canada, September 26-28, 2012,* 314-337, Vancouver, Canada: UNESCO. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC__Rogers_John_26_D_1620.pdf Accessed March 2014.

Woods, Kam and Christopher A. Lee. “Acquisition and Processing of Disk Images to Further Archival Goals." In *Proceedings of Archiving 2012, Copenhagen, June 2012,* Springfield, VA: Society for Imaging Science and Technology, 2012, 147-152. http://ils.unc.edu/callee/archiving-2012-woods-lee.pdf Accessed March 2014.

Woods, Kam, Christopher A. Lee, and Simson Garfinkel. “Extending Digital Repository Architectures to Support Disk Image Preservation and Access.” In *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, 57-66.* New York, NY: ACM Press, 2011. http://www.ils.unc.edu/callee/p57-woods.pdf Accessed March 2014.

Woods, Kam, Christopher Lee, and Sunitha Misra. “Automated Analysis and Visualization of Disk Images and File Systems for Preservation.” In *Proceedings of Archiving 2013, Washington, D.C.: April 2-5, 2013.* 239-244. Springfield, VA: Society for Imaging Science and Technology, http://ils.unc.edu/callee/p239-woods.pdf Accessed March 2014.

Woods, Kam, and Geoffrey Brown. “From Imaging to Access - Effective Preservation of Legacy Removable Media.” In *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations: Final Program and Proceedings*, 213-218. Springfield, VA: Society for Imaging Science and Technology, 2009.