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. 131 pages. Advisor: Christopher A. Lee

This paper presents the findings of semi-structured interviews with archivists and curators applying digital forensics tools and practices to the management of born-digital content. The interviews were designed to explore which digital forensic tools are in use, how they are implemented within a digital forensics workflow, and what further challenges and opportunities such use may present. Findings indicate that among interview participants these tools are beneficial in the capture and preservation of born-digital content, particularly with digital media such as external hard drives, and optical or floppy disks. However, interviews reveal that metadata generated from the use of such tools is not easily translated into the arrangement, description, and provision of access to born-digital content.

Headings:

Digital preservation

Data recovery (Computer science)

Personal archives—digitization

Workflow

“THE WAY WE DO IT HERE”: MAPPING DIGITAL FORENSICS WORKFLOWS IN COLLECTING INSTITUTIONS

by
Martin J. Gengenbach

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

August 2012

Approved by

_____

Christopher A. Lee

Table of Contents

Figures

Introduction

A large number of collecting institutions are now acquiring born-digital materials.[1] A 2010 survey of archives and special collections in research institutions notes that some 79% of respondents affirmed that their institution has acquired born-digital content in some format.[2] In this same work, authors Jackie Dooley and Katherine Luce suggest that these collecting activities are not well controlled or monitored; only 35% of respondents could provide a size to their born-digital holdings.[3] Thus, while it is clear that institutions are collecting digital content in some capacity, it is less clear how they are managing the content that comes under their authority, and whether there is any consistency across institutions in their handling of that content. In a 2009 report, a task force formed by the Society of American Archivists (SAA) on best practices for

---

[1]For the purpose of this study, "collecting institutions" refers to "archives that acquire collections from outside donors," as described in Susan E. Davis, "Electronic Records Planning in 'Collecting' Repositories," *American Archivist* 71 (Spring/Summer 2008), 169, http://archivists.metapress.com/content/024q2020828t7332/fulltext.pdf (accessed July 2012). "Born-digital" is taken to mean any materials that come into the control of the collecting institution in digital form, whether on a digital media carrier (CD, DVD, hard drive, etc.) or via online transfer. Thus, scanned photographs may be considered "born-digital," if they were accepted by the archive as digital objects. This work will also alternate between the use of terms such as "digital content," "electronic records," "born-digital material," and "digital objects," depending on the context in which they are being used, as my emphasis is on the actions taken upon born-digital content, and not how it is defined for the purposes of a particular study or project.

[2]Jackie M. Dooley and Katherine Luce, *Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives,* http://www.oclc.org/research/publications/library/2010/2010-11.pdf (Dublin, OH: OCLC Research, 2010), 59. The report, sponsored by the Online Computer Library Center (OCLC), is based on the survey responses of 169 institutions from one of five overlapping membership organizations: Association of Research Libraries (ARL), Canadian Academic and Research Libraries (CARL), Independent Research Libraries Association (IRLA), Oberlin Group, Research Library Group (RLG) Partnership, U.S. and Canadian Members.

[3]Ibid. Ben Goldman points out that this is a frightening statistic in Ben Goldman, "Bridging the Gap: Taking Practical Steps Towards Managing Born-Digital Collections in Manuscript Repositories," *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 12 no. 1 (2011), 12-13, http://rbm.acrl.org/content/12/1/11.full.pdf+html (accessed August 3, 2012).

managing digital content notes that "[i]n most cases, organizations have not made their practices publicly available."[4]

The challenges of digital preservation are well established.[5] At its fundamental level, digital information exists as electronic impulses written onto a physical medium, representing bits— "on" or "off," read as ones and zeroes.[6] These bits serve as instructions, abstracted through the layers of a computer's hardware and software systems,that render a digital object in a way that is human-understandable. Thus digital objects "are recreated each time they are used, based on interactions of numerous technological components."[7] Both the physical media and the digital objects on the media can pose challenges to preservation—hardware obsolescence, bit rot, and corruption of the physical media itself, legacy software that is no longer supported or available, and unreadable, encrypted, and/or proprietary file formats, just to name a few.[8] The sheer magnitude of these challenges has prompted worries about a coming "digital dark age,"

---

[4]Society of American Archivists Technology Best Practices Task Force, "Managing Electronic Records and Assets: A Pilot Study on Identifying Best Practices," Naomi Nelson, chair. (2009), 2, http://www.archivists.org/governance/taskforces/MERA-PilotStudy.pdf (accessed July 2012).

[5]Some of these challenges are first outlined in the seminal report of the 1996 Task Force on Archiving of Digital Information. See Donald Waters and John Garrett, *Preserving Digital Information: Report of the Task Force on Archiving Digital Information.* (Washington, D.C.: Committee on Preservation and Access, 1996), http://www.clir.org/pubs/reports/pub63watersgarrett.pdf (accessed July 2012).

[6]Elizabeth Dow, *Electronic Records in the Manuscript Repository.* (Lanham, Maryland: The Scarecrow Press, 2009), 23.

[7]Christopher A. Lee, *I Digital: Personal Collections in the Digital Era*. (Chicago, IL: Society of American Archivists, 2011), 5. This is a brief explanation of an otherwise complex process that is explained in more detail in Jeff Rothenberg, *Ensuring the Longevity of Digital Information* (Washington, D.C.: Council on Library and Information Resources 1999), http://www.clir.org/pubs/archives/ensuring.pdf (accessed July 2012).

[8]For more detailed information on challenges to digital preservation, see Margaret Hedstrom, "Digital Preservation: A Time Bomb for Digital Libraries," *Computers and the Humanities* 31, no. 3 (1997), 189-202, http://hdl.handle.net/2027.42/42573 (accessed July 2012); Rothenberg, *Ensuring the Longevity of Digital Information*; and Matthew G. Kirschenbaum, Richard Ovenden, and Gabriella Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, D.C.: Council on Library and Information Resources, 2010), 14-21, http://www.clir.org/pubs/abstract/reports/pub149 (accessed July 2012).

and the loss of information critical to understanding the developing born-digital culture.[9]

As the number of institutions trying to better manage their digital content increases, it is important for institutions with existing born-digital preservation programs to document internal practices through case studies, conference presentations, and collaborative research projects.[10] A particularly promising avenue of study focuses on the intersection of digital preservation and digital forensics.[11] Digital forensics has been defined as "a process encompassing the identification, preservation, analysis and presentation of digital evidence in a legally acceptable manner."[12] Originally developed for the law enforcement community to collect and preserve evidence of computer-based criminal activity, several recent research projects have demonstrated the applicability of digital forensics tools and practices to the acquisition, preservation, and provision of

---

[9]See Terry Kuny, "A Digital Dark Ages? Challenges in the Preservation of Electronic Information," *63rd International Federation of Library Associations (IFLA) Council and General Conference*, (September 4, 1997), http://archive.ifla.org/IV/ifla63/63kuny1.pdf (accessed June 2012). Kuny writes: "we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever." Others feel this fear is overblown, and point out that the most common anecdotes about data *loss* are actually about data *recovery*. See Ross Harvey, "So Where's the Black Hole in Our Collective Memory? A Provocative Position Paper (PPP)," Digital Preservation Europe (DPE) (18 January 2008), http://www.digitalpreservationeurope.eu/publications/position/Ross_Harvey_black_hole_PPP.pdf (accessed June 2012).

[10]Examples that will be discussed in more depth in this work include: Matthew G. Kirschenbaum, Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside, "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use," (College Park, MD: University of Maryland, 2009), http://mith.umd.edu/wp-content/uploads/whitepaper_HD-50346.Kirschenbaum.WP.pdf (accessed July 2012); Jeremy Leighton John, Ian Rowlands, Peter Williams, and Katrina Dean, "Digital Lives: Personal digital archives for the 21st century >> an initial synthesis," A Digital Lives Research Paper. Beta Version 0.2 (March 3, 2010) http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf (accessed June 2012); and AIMS Work Group, "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship," (2012) http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf (accessed July 2012).

[11]Digital forensics is also sometimes referred to as computer forensics or forensic computing. For the purposes of this work, the term digital forensics will be used. For a discussion of the variants of the term, see Kirschenbaum, Ovenden, and Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, 3.

[12]Rodney McKemmish, "When Is Digital Evidence Forensically Sound?" *Advances in Digital Forensics IV*, Indrajit Ray and Sujeet Shenoi, eds., *IFIP International Federation for Information Processing* 285 (Boston: Springer, 2008), 4, http://www.springerlink.com/content/048j747850234355/ (accessed August 2012).

access to born-digital content in archival settings.[13]

The study presented in this paper used semi-structured interviews with archivists and curators to investigate the implementation of digital forensics practices for managing born-digital content in collecting institutions. My research objective has been exploratory; my original intent was to examine how collecting institutions integrate digital forensics tools and processes into their workflows for managing born-digital content from acquisition to the provision of access. High-level workflow models based on the information gathered through those interviews provide additional documentation and context for archives and special collections seeking to develop their own processes for managing born-digital content.

Literature Review

Defining Digital Forensics

Before delving into the relevant literature, it will be helpful to provide a brief discussion of the term *digital forensics* in the context of this work. While the definition provided above is accurate, it is useful to consider others. In 2001, at the first meeting of the Digital Forensics Research Workshop (DFRWS), the following was adopted as a definition for digital forensics:

> The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations.[14]

---

[13]This will be further explored in the next section of the present work.

[14]Gary Palmer, *A Road Map for Digital Forensic Research.Technical Report DTR- T0010-01, DFRWS, November 2001. Report from the First Digital Forensic Research Workshop (DFRWS)*: 16, http://www.dfrws.org/2001/dfrws-rm-final.pdf (accessed August 2012).

This definition articulates the specific processes that make up the practice of digital forensics, while making clear that digital forensics developed as an applied field, rather than a theoretical one: indeed, "many experts concede that the scientific method did not underlie much of early digital forensic research."[15] Kirschenbaum notes the practical implications of this definition, which include "working with hard drives and other storage media […] and creating the conditions necessary to ensure that the data has not been tampered with in the process of its recovery or analysis."[16] Digital forensics focuses on the use of hardware and software tools to collect, analyze, interpret, and present information from digital sources, and ensuring that the collected information has not been altered in the process.[17] This work will focus on the use of those tools and processes that are of particular benefit to practitioners in collecting institutions working with born-digital content, beginning with a review of the literature contributing to the development of the fields of digital forensics and digital preservation in collecting institutions.

Digital Forensics Investigation: Background and Methods

Garfinkel has referred to the years 1999-2007 as "the Golden Age of digital forensics."[18] Prior to that period, the practice of digital forensics analysis was ad hoc, limited by the diversity of hardware, software, applications, and the storage capabilities

---

[15]Nicole Beebe, "Digital Forensic Research: The Good, The Bad, and the Unaddressed," *Advances in Digital Forensics V*, Gilbert Peterson and Sujeet Shenoi, eds., *International Federation for Information Processing* 306 (2009), 19, http://dx.doi.org/10.1007/978-3-642-04155-6_2 (accessed August 2012).

[16]Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination.* (Cambridge, Mass: The MIT Press, 2008), 46.

[17]Additional information on the different types and applications of digital forensic practices to digital preservation are outlined in Kirschenbaum, Ovenden, and Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections,* 3-4.

[18]Simson Garfinkel, "Digital forensics research: The next 10 years," *Digital Investigation* 7 (2010): S66, http://dfrws.org/2010/proceedings/2010-308.pdf (accessed July 2012).

of early computers.[19] With the standardization of file types and computing environments running versions of the Microsoft Windows operating system, digital forensics investigators were able to hone their tools and practices on single machines working in generally predictable ways.[20] Owing to this "Golden Age," Beebe notes, "there is now a relatively solid understanding of what digital artifacts exist, where they exist, why they exist and how to recover them."[21] Along with the knowledge of forensics experts, awareness of the field of digital forensics itself has grown prodigiously, as well. Digital forensics investigation is considered a common practice in the law enforcement and information security communities: "it is now mainstream knowledge that the digital footprints that remain after interactions with computers and networks are significant and probative."[22]

With the development of digital forensics as a field, practitioners have made significant progress in standardizing and formalizing forensics practices.[23] Part of this standardization has been achieved through the publication of introductory instructional texts on forensics investigation.[24] Other work has focused on creating broad theoretical frameworks for maintaining the evidence collected in established forensics practices,[25] developing common metadata fields across forensics tools,[26] and creating a standard body

---

[19]Ibid.
[20]Ibid.
[21]Beebe, "Digital Forensic Research,"19-20.
[22]Ibid., 18.
[23]Garfinkel, "Digital forensics research: The next 10 years," S66. Garfinkel notes that "The Golden Age was also marked by a rapid growth in digital forensics research and professionalization."
[24]Introductory texts include: Dan Farmer and Wietse Venema, *Forensic Discovery*. (Upper Saddle River, NJ: Addison-Wesley, 2005) http://www.porcupine.org/forensics/forensic-discovery/ (accessed July 2012); and Brian Carrier, *File System Forensic Analysis*. (Boston, MA: Addison-Wesley, 2005).
[25]Sarah Mocas, "Building theoretical underpinnings for digital forensics research," *Digital Investigation* 1 no. 1 (2004): 61-68, http://www.dblp.org/db/journals/di/di1.html (accessed August 2012).
[26]Simson Garfinkel, "Digital Forensics XML and the DFXML Toolset," *Digital Investigation* 8 (2012), 161-74, http://simson.net/clips/academic/2012.DI.dfxml.pdf (accessed June 2012).

of forensics corpora upon which to test forensics tools in a consistent environment.[27]

Digital forensics authors and practitioners have also characterized the different types of tools used within particular steps of the investigation process. Carrier describes how interpreting data through layers of abstractions can provide solutions to the *Complexity Problem,* the lack of human-readability of data at the bit-level; and the *Quantity Problem,* how to deal with the sheer quantities of digital information to be processed in modern computing environments.[28] Thus manipulating data through layers of abstraction can apply both to the digital information that is the target of forensic processing, and to the software tools used to extract and analyze that information.

Finally, digital forensics authors have characterized the forensic investigation process by its individual components and the relationships between components.[29] Models of the forensic investigation process have attracted particular interest from researchers, and have existed since the 1980s.[30] However, many of these early models were focused on specific technological environments or particular use cases, rather than an abstracted model designed for wider adoption.[31] Researchers have also developed novel forms of modeling digital forensics investigation, such as treating the computer as a separate "digital crime scene," and approaching digital evidence collection with a

---

[27]Simson Garfinkel, Paul Farrell, Vassil Roussev, and George Dinolt, "Bringing Science to Digital Forensics with Standardized Forensic Corpora," *Digital Investigation* 6 (2009), S2-S11, http://www.dfrws.org/2009/proceedings/p2-garfinkel.pdf (accessed June 2012).

[28]Brian Carrier, "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers," *International Journal of Digital Evidence* 1 no. 4 (Winter 2003): 1-12, http://www.informatik.uni-trier.de/~ley/db/journals/ijde/ijde1.html (accessed July 2012).

[29]Yunus Yusoff, Roslan Ismail and Zainuddin Hassan, "Common Phases of Computer Forensics Investigation Models," *International Journal of Computer Science and Information Technology (IJCSIT)* 3 no. 3 (June 2011), 17-31, http://airccse.org/journal/jcsit/0611csit02.pdf (accessed August 2012).

[30]Yusoff, Ismail, and Hassan, "Common Phases of Computer Forensics Investigation Models," 17. Their work provides an extensive bibliography of existing digital forensics process models.

[31]Mark Reith, Clint Carr, and Gregg Gunsch, "An Examination of Digital Forensic Models," *International Journal of Digital Evidence* 1 no. 3 (Fall 2002), 3-4, http://people.emich.edu/pstephen/other_papers/Digital_Forensic_Models.pdf (accessed August 2012).

physical forensics methodology.[32] Others have compiled models chronologically to document their increasing complexity.[33] More recently, researchers have synthesized existing models, isolating common investigative processes in order to create a basic high-level abstract model, usable for the development of new tools and technologies, but also helpful for the application of digital forensics processes outside of criminal investigation and computer security.[34]

Digital Preservation and Digital Forensics in Collecting Institutions

It is only in the recent past that practitioners and researchers in digital forensics and digital preservation have recognized the overlap in their respective fields. Through the 1980s and 1990s, the rapid pace of technology and the development of new hardware and software systems had caused some archivists to entirely reconsider how to approach electronic records, and to question whether traditional conceptions of archival practice were still valid for this new era.[35] Articles by Adrian Cunningham, Tom Hyry and Rachel Onuf during this period note the lack of attention to electronic records in collecting

[32]Brian Carrier and Eugene H. Spafford, "Getting Physical with the Digital Investigation Process," *International Journal of Digital Evidence* 2 no. 2 (Fall 2003), 1-20, https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2003-29.pdf (accessed August 2012).

[33]Mark M. Pollitt, "An Ad Hoc Review of Digital Forensic Models," *Proceeding of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'07)*, Washington, D.C., (2007), http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4155349 (accessed August 2012).

[34]Yusoff, Ismail, and Hassan, "Common Phases of Computer Forensics Investigation Models," 29-30.

[35]There is a sizable body of scholarship available on this topic; for a classic articulation of the "new paradigm" model of electronic records management, see David Bearman, "An Indefensible Bastion: Archives as a Repository in the Digital Age," in D. Bearman (ed.) *Archival Management of Electronic Records, Archives and Museum Informatics Technical Report* 13 (1991), 14-24. An alternative perspective is provided by Linda J. Henry, "Schellenberg in Cyberspace," *American Archivist* 61 no. 2 (Fall 1998): 310-311, http://www.jstor.org/stable/40294090 (accessed July 2012). For summaries of the differences between the projects, see Philip C. Bantin, "Strategies for Managing Electronic Records: A New Archival Paradigm? An Affirmation of our Archival Traditions?" *Archival Issues* 23, no. 1 (1998), 17-34; and Peter B. Hirtle, "Archival Authenticity in a Digital Age," in *Authenticity in a Digital Environment* (Washington, D.C.: Council on Library and Information Resources, May 2000), 8-23, http://www.clir.org/pubs/reports/pub92/pub92.pdf (accessed July 2012).

institutions, calling for established best practices for capturing and preserving digital content, and advocating for additional research and technical expertise specific to the concerns of collectors of personal digital collections.[36] One of the first case studies in personal electronic records is Lucie Pacquet's "Appraisal, Acquisition, and Control of Personal Electronic Records: From Myth to Reality."[37] Pacquet's work highlights the need to collect information about the technological environments in which electronic records are created, the software and hardware platforms encountered, and other technical metadata.[38] She demonstrates that a deep understanding of the technologies involved in digital preservation is needed for collecting institutions to manage born-digital content.

Archivists working with digital collections have thus conducted research into the technical complexities involved in preserving digital content over the long term, including the physical medium upon which digital information is written. In *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*, Seamus Ross and Ann Gow systematically document the ways in which a computer's hard drive, now one of the most commonplace physical carriers of digital content, can degrade and break down.[39]

---

[36]Adrian Cunningham has written several articles articulating these issues; see Adrian Cunningham, "The Archival Management of Personal Records in Electronic Form: Some Suggestions," *Archives and Management* vol. 22, no. 1 (1994), 94-105; and Adrian Cunningham, "Waiting for the Ghost Train: Strategies for managing electronic personal records before it is too late," pre-publication version of a paper delivered to the Society of American Archivists Annual Meeting, August 23-29, 1999, Pittsburgh, PA. Published in *Archival Issues: Journal of the Midwest Archives Conference* 24, no. 1 (1999), 55-64, http://www.mybestdocs.com/cunningham-waiting2.htm (accessed June 2012); Tom Hyry and Rachel Onuf, "The personality of electronic records: the impact of new information technology on personal papers," *Archival Issues* 22 no. 1 (1997): 39-41, http://minds.wisconsin.edu/handle/1793/45828 (accessed June 2012). Cunningham, Hyry, and Onuf all revisited these early works for chapters in Christopher A. Lee, *I, Digital*; see Adrian Cunningham, "Ghosts in the Machine: Towards a Principles-Based Approach to Making and Keeping Digital Personal Records," 78-89; and Rachel Onuf and Tom Hyry, "Take It Personally: The Implications of Personal Records in Electronic Form," 241-256, in *I, Digital*, Christopher A. Lee, ed., (Chicago, IL: Society of American Archivists, 2011).

[37] Lucie Paquet, "Appraisal, Acquisition and Control of Personal Electronic Records: From Myth to Reality." *Archives and Manuscripts* 28, no. 2 (2000): 71-91.

[38]Paquet, "Appraisal, Acquisition, and Control of Personal Electronic Records," 81-83.

[39]Seamus Ross and Ann Gow, *Digital Archaeology: Rescuing Neglected and Damaged Data Resources. London: British Library, 1999,* http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf

However, their work also demonstrates the stubborn resilience of data that is written to a hard drive, outlining the various strategies implemented in recovering "lost" data.[40] The paradoxical fragility and durability of digital information is highlighted by Simson Garfinkel and Abhi Shelat in "Remembrance of Data Passed: A Study of Disk Sanitation Practices," in which the authors reveal that an astonishing amount of information may still be available on hard drives improperly sanitized by their previous owners.[41]

Jeff Rothenberg also explores the nature of digital information in "Ensuring the Longevity of Digital Information," another 1999 study. Rothenberg lays out obstacles to digital preservation through a hypothetical scenario involving a compact disc encountered by his grandchildren: "They must not only be able to extract the content on the disc—they must also interpret it correctly."[42] Rothenberg argues that digital preservation efforts must be based around the bitstream—literally, the linear bit-by-bit stream of information on a storage medium[43]—but also notes that any preservation activity must support the processes and programs necessary to interpret those ones and zeros properly as a digital object.[44] This is because "a document file is not a document in its own right: it merely *describes* a document that comes into existence only when the file is 'run' by the program that created it."[45] The interactions between digital objects described by Rothenberg are

---

(accessed June 2012). The components and operation of a "spinning disk" hard drive are also explained in great detail in Kirschenbaum, *Mechanisms,* 86-96; and Ron White, *How Computers Work.* Timothy Edward Downs, illus. 9th edition (Indianapolis, IN: Que Publishing, 2008), 158-171.

[40]*Ross and Gow, Digital Archaeology,* 17-26.

[41]Simson Garfinkel and Abhi Shelat, "Remembrance of Data Passed: A Study of Disk Sanitation Practices," *IEEE Security and Privacy* (January/February 2003), 17-27, http://cdn.computerscience1.net/2005/fall/lectures/8/articles8.pdf (accessed August 2012).

[42]Rothenberg, "Ensuring the Longevity of Digital Information," 2.

[43]The writing of bits onto a storage media to create the bitstream is described in depth in White, *How Computers Work,* 158-171.

[44]Rothenberg, 2.

[45]Ibid., 10. For more information on the representational levels of digital objects, see David Levy, *Scrolling Forward: Making Sense of Documents in the Digital World* (New York, NY: Arcade Publishing, 2001), especially Chapter 2, "What Are Documents?" pp. 21-38; and Kenneth Thibodeau, "Overview of

examples of what Kenneth Thibodeau has termed the "multiple inheritance" of digital objects: "Every digital object is a physical object, a logical object, and a conceptual object, and its properties at each of those levels can be significantly different."[46] In order to correctly interpret the bits that make up the digital object, then, the programs, files, and underlying systems that support the digital object must also be retained.[47]

The interpretation of digital objects at alternate levels of representation demonstrates one parallel in digital preservation and digital forensics, and as does archivists' increasing use of forensic *disk images*. Commonly used in digital forensics evidence acquisition, Woods, Lee, and Garfinkel describe the disk image as "a 'snapshot' of the medium's content, including all allocated files, filenames, and other metadata information associated with the disk volume," stored as a single file or set of files.[48] The disk image provides a complete copy of the data from the storage medium, including bit-level information from deleted files and other data that would be otherwise lost in standard copying practices.[49] Woods, Lee, and Garfinkel discuss how the disk image can be treated as any other digital object, interacting with an image's content files through the

---

Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation: An International Perspective*. (Washington, D.C.: Council on Library and Information Resources, 2002), http://www.clir.org/pubs/reports/pub107/thibodeau.html/ (accessed July 2012).

[46]Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," in *The State of Digital Preservation*; Garrett and Waters, *Preserving Digital Information*, 12. Garrett and Waters write: "The notion of content, however, is itself a complex idea that operates at several different levels of abstraction."

[47]Rothenberg, 14.

[48]Kam Woods, Christopher A. Lee, and Simson Garfinkel, "Extending Digital Repository Architectures to Support Disk Image Preservation and Access," *JCDL 11, Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, June 13-17, 2011, Ottawa, Ontario, Canada (2011), 58, http://ils.unc.edu/callee/p57-woods.pdf (accessed July 2012).

[49]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," in *Proceedings, iPRES 2009: the Sixth International Conference on Preservation of Digital Objects* (2009), 112, http://escholarship.org/uc/item/7d3465vg (accessed July 2012).

operating system, and facilitating preservation activities.[50]

Archivists have also researched how to ensure the integrity of digital information, and how to demonstrate that the information they preserve can still be considered trustworthy. The 1996 report *Preserving Digital Information: Report on the Task Force on Archiving of Digital Information*[51] highlights the importance of maintaining the integrity of digital objects, and identifies the use of checksum algorithms as a method of ensuring that integrity.[52] The report also recognizes that making content accessible in the future will necessitate its migration—its transfer from one format into another in a manner that may change its "look and feel"—as operating environments change.[53] Archivists have thus invested great effort in establishing the necessary requirements to ensure both the integrity and trustworthiness of digital objects over time, as well as technical processes that can ensure that information is collected in a manner to ensure its authenticity.

An important research program on the trustworthiness of digital information began in 1994 at the University of British Columbia's Master of Archival Sciences (UBC-

---

[50]Woods, Lee, and Garfinkel, "Extending Digital Repository Architectures," 58. See also Kam Woods and Geoffrey Brown, "From Imaging to Access – Effective Preservation of Legacy Removable Media," In *Proceedings of Archiving 2009* (Springfield, VA: Society for Imaging Science and Technology, 2009), 213-218, http://www.digpres.com/publications/woodsbrownarch09.pdf (accessed August 2012); and Kam Woods and Christopher A. Lee, "Acquisition and Processing of Disk Images to Further Archival Goals," In *Proceedings of Archiving 2012* (Springfield, VA: Society for Imaging Science and Technology, 2012), 147-152, http://ils.unc.edu/callee/archiving-2012-woods-lee.pdf (accessed August 2012).

[51]John Garrett and Donald Waters, 12-13.

[52]Garrett and Waters cite Clifford Lynch, "The Integrity of Digital Information: Mechanics and Definitional Issues," *Journal of the American Society for Information Science* 45 no. 10 (1994), 739-740. A checksum is the product of a cryptographic hash algorithm applied to a file. The resulting checksum is a string of characters that is unique to that file. If even one bit is altered and the file is re-saved, running the hash algorithm on that file will produce a different checksum. A further discussion of hash algorithms (also known as *hashing*) is provided in Kirschenbaum, *Mechanisms*, 55-56. Kirschenbaum writes: "A hash algorithm generates a numeric value that is the mathematical surrogate for a particular bitstream." Examples of hash algorithms include MD5 and SHA-1. More information on forensic hashing can be found at: "Forensic Hashing," *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/Hashing (accessed July 2012).

[53]Garrett and Waters, 27.

MAS) program in Vancouver, Canada.[54] The UBC-MAS project, led by Luciana Duranti, defined requirements for the authenticity and reliability of information in an electronic records context, and was theoretically grounded in Duranti's study of diplomatics, the seventeenth-century study of the integrity and authenticity of legal documents.[55] In "Reliability and Authenticity: The Concepts and Their Implications," Duranti describes records in terms of *trustworthiness*, where trustworthy records are *reliable* and *authentic*. Reliability indicates a record's "ability to stand for the facts [it] is about," as "the entity of which it is evidence."[56] Authenticity indicates that a document is what it purports to be, "protected and guaranteed through the adoption of methods that ensure that the record is not manipulated, altered, or otherwise falsified after its creation[.]"[57] Building on this work, in "From Digital Diplomatics to Digital Records Forensics," Duranti explores the theoretical concept of the "digital record" as an entity with particular characteristics that can be supported by digital forensics tools, and how systems may be designed with forensics capabilities in mind to meet the requirements of authentic digital record-

[54]The University of British Columbia-Master of Archival Science (UBC-MAS) project led directly to the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Projects 1-3, expanding upon its findings and applying them in a variety of institutional and organizational contexts. For more information on these projects, see the *InterPARES project* website at: http://www.interpares.org/. Also see Luciana Duranti and Heather MacNeil, "The Protection of the Integrity of Electronic Records: An Overview of the Findings of the UBC-MAS Research Project," *Archivaria* 42 (Fall 1996), 46-67 http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12153 (accessed July 2012); and Hirtle, "Archival Authenticity in a Digital Age," in *Authenticity in a Digital Environment,* 8-23.

[55]Duranti, and MacNeil, "The Protection of the Integrity of Electronic Records," 47; also see Luciana Duranti, "Diplomatics: New Uses for an Old Science," *Archivaria* 28 (1989): 7-27, http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11567/12513 (accessed August 2012); the first of a series of articles by Duranti that were subsequently republished as Luciana Duranti, *Diplomatics: New Uses for an Old Science.* (*Chicago, Ill.: Society of American Archivists, Association of Canadian Archivists, and Scarecrow Press, 1998).*

[56]Luciana Duranti, "Reliability and Authenticity: The Concepts and their Implications," *Archivaria* 39 (1995), 6, http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12063/13035 (accessed August 2012).

[57]Luciana Duranti and Heather MacNeil, "The Protection of the Integrity of Electronic Records," 56.

keeping.[58]

Digital Forensics in Born-Digital Collecting Institutions

The use of digital forensics tools in digital preservation processes developed through archivists' recognition that digital objects are complex entities with multiple inheritance; that digital content is both ephemeral and fragile, as well as persistent and enduring; and that archivists must be able to verify the authenticity of records whether they are analog or digital. Previous work has focused on identifying potential uses for digital forensics tools in cultural heritage settings, and created opportunities for further study in this area.

The Personal ARchives Accessible In Digital Media (Paradigm) project, a collaboration between the Universities of Oxford and Manchester from 2005-2007, developed and documented a workbook containing methods for collecting institutions to acquire, preserve, and provide access to born-digital and hybrid personal collections.[59] Informed by the experiences of project participants at the Bodleian Library at Oxford and the John Rylands University Library at Manchester, the workbook is divided into sections focusing on archival functions, as well as special topics of consideration to those working

[58]Luciana Duranti, "From Digital Diplomatics to Digital Records Forensics." *Archivaria* 68 (2009), 39-66, http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13229 (accessed July 2012). Other notable contributions include Charles T. Cullen, Peter B. Hirtle, David Levy, Clifford A. Lynch, and Jeff Rothenberg, *Authenticity in a Digital Environment* (Washington, D.C.: Council on Library and Information Resources, 2000), http://www.clir.org/pubs/abstract//reports/pub92 (accessed August 2012); and Michael Forstrom, "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library," *American Archivist* 72 (Fall/Winter 2009), 460-477, http://archivists.metapress.com/content/b82533tvr7713471/fulltext.pdf (accessed July 2012).

[59]Susan Thomas and Janette Martin, "Using the papers of contemporary British politicians as a testbed for the preservation of digital personal archives," *Journal of the Society of Archivists* 27, no. 1 (April 2006), 29, http://eprints.hud.ac.uk/7893/ (accessed July 2012).

with digital collections, such as legal issues and digital preservation strategies.[60] While it did not specifically highlight the use of digital forensics practices, it did recognize the possibility of data recovery through the accession of materials on obsolete media types and formats, establishing the necessity of a research program for future work.[61]

There have since been several significant grant projects that have either focused on or incorporated digital forensics into their objectives. The Digital Lives Project, led by Jeremy Leighton John and funded by the British Arts and Humanities Research Council (AHRC), published an *Initial Synthesis* that is an overview of a great many interrelated research activities around the creation and management of personal digital archives.[62] The report explores how to facilitate the capture and preservation of born-digital personal collections (or eMANUSCRIPT collections, to adopt John's term); how people use technology to create digital collections in the twenty-first century; the legal and ethical issues of capturing personal digital archives; and a variety of other topics.[63] A related

---

[60]"paradigm | workbook on digital private papers," *Paradigm Project* (2007) http://www.paradigm.ac.uk/workbook/ (accessed July 2012). Also see Susan Thomas, "Curating the I, Digital: Experiences at the Bodleian Library," in Christopher A. Lee, ed. *I, Digital: Personal Collections in the Digital Era* (Chicago, IL: Society of American Archivists, 2011), 280-305. The final project report, "Paradigm: A practical approach to the preservation of personal digital archives," is available at http://www.paradigm.ac.uk/projectdocs/jiscreports/index.html.

[61]"paradigm | workbook on digital private papers | collection development | transfer via retired media," *Paradigm Workbook.* http://www.paradigm.ac.uk/workbook/collection-development/via-retired-media.html (accessed July 2012). Additional research projects developed from the Paradigm project include cairo (Complex Archive Ingest for Repository Objects), which ended in 2008 and focused on exploring tools to improve ingest processes for digital repositories; and the futureArch project (2008-2011) which was designed to establish policies and systems to facilitate the implementation of an enterprise trusted digital repository service at the Bodleian Library at Oxford University. The system, BEAM (Bodleian Electronic Archives and Manuscripts) is also outlined in this project. For more information on cairo, see: "cairo || Complex Archives Ingest for Repository Objects," *cairo*. http://cairo.paradigm.ac.uk/about/index.html (accessed July 2012); for more information on futureArch and BEAM, see: "futureArch – BEAM," *Bodleian Electronic Archives and Manuscripts.* http://www.bodleian.ox.ac.uk/beam/projects/futurearch (accessed July 2012).

[62]Jeremy Leighton John, Ian Rowlands, Peter Williams, and Katrina Dean, "Digital Lives: Personal digital archives for the 21st century >> an initial synthesis," A Digital Lives Research Paper. Beta Version 0.2 (March 3, 2010) http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf (accessed June 2012).

[63]Ibid, vi. See John et al., "Some key findings," *Digital Lives*, xi-xviii, for a good overview of the range of

article by John explores the use of digital forensics tools and technologies for digital capture and preservation in the British Library.[64]

Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use, a 2009 report funded by the National Endowment for the Humanities (NEH), explored the use of digital forensics techniques in work with the born-digital collections of authors in several different archives, special collections, and research centers.[65] In the NEH-sponsored white paper and subsequent article, Kirschenbaum et al. demonstrate the benefits of preserving original computing environments, using forensic disk images for preservation, and working with donors to acquire their materials, while advocating steps to further integrate digital forensics processes into digital preservation and research into access mechanisms.[66]

The work of this NEH grant-funded group directly fed into a second project, "Computer Forensics and Born-Digital Content in Cultural Heritage Collections," which produced one of the most comprehensive treatments of the topic to date.[67] The report,

---

topics covered. The *Digital Lives* project is not alone in exploring personal information management for common threads within the digital preservation community. See Neil Beagrie, "Plenty of Room at the Bottom? Personal Digital Libraries and Collections," *D-Lib Magazine* vol. 11, no. 6 (June 2005), http://www.dlib.org/dlib/june05/beagrie/06beagrie.html (accessed August 2012); Cathy Marshall, "Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field," *D-Lib Magazine* 14, no. 3/4 (March/April 2008), http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html (accessed August 2012); and Christopher A. Lee and Robert Capra, "And Now the Twain Shall Meet: Exploring the Connections between PIM and Archives," in *I, Digital: Personal Collections in the Digital Era* Christopher A. Lee, ed., 29-77.

[64]Jeremy Leighton John, "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools," Paper presented at the iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK. (2008), http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf (accessed August 2012).

[65]Matthew G. Kirschenbaum, Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside, *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use* (College Park, MD: University of Maryland, 2009), http://mith.umd.edu/wp-content/uploads/whitepaper_HD-50346.Kirschenbaum.WP.pdf (accessed July 2012).

[66]Kirschenbaum et al., "Digital Materiality," 111-112.

[67]Kirschenbaum, Redwine, and Ovenden, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*.

published by the Council on Library and Information Resources, breaks down the challenges of working with digital content: dealing with legacy formats, maintaining the authenticity and trustworthiness of digital materials, and how digital forensics tools and practices can address those challenges. The *Digital Forensics* CLIR report also includes articles on digital forensics workflows and working with donors to address possible privacy concerns in light of the discovery made possible by digital forensics tools.[68]

Another recent project to garner significant attention is *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship,* involving archivists from Yale University, the University of Virginia, the University of Hull, and Stanford University. The goal of the AIMS project has been to "define good practice in terms of archival tasks and objectives necessary for success" in the stewardship of born-digital content.[69] The project's recently released final report is divided according to four primary *functions of stewardship*: collection development, accessioning, arrangement and description, and discovery and access.[70] Each function of stewardship includes generalized keys to successfully implement the function; and a number of outcome-based objectives designed around decisions and practical tasks that break out the components of each objective.[71]

The BitCurator Project in some ways is an extension of these previous efforts, as it involves members of each of the previously mentioned grant-funded projects. The effort is a collaboration of the University of Maryland's Maryland Institute for Technology in the Humanities (MITH) and the School of Information and Library

---

[68]See Kirschenbaum, Ovenden, and Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections,* iv.

[69]AIMS Project Group, *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* (January 2012): ii, http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf (accessed June 2012).

[70]Ibid., 2.

[71]For example, see AIMS Project Group, "Accessioning," *AIMS Born-Digital Collections,* 17-30.

Science at the University of North Carolina at Chapel Hill (SILS), funded by the Andrew W. Mellon Foundation.[72] The project's primary focus is on developing and integrating digital forensics tools into the workflows of institutions collecting born-digital content.[73] The BitCurator Project is using a Professional Expert Panel (PEP) and Development Advisory Group (DAG) to shape the development of digital forensics tools that are geared toward an archival audience.[74] The project is in the first year of a two-year grant, is disseminating software, and has already contributed to the existing literature.[75]

The above projects that implement digital forensics tools and processes demonstrate the willingness of the field to embrace technically innovative approaches to digital preservation. Many of these projects include some form of workflow description or illustration in the products of their research.[76] Glisson's "Use of Computer Forensics in the Digital Curation of Removable Media" is perhaps the only work to focus specifically on workflow development, though visual workflows have also been the focus of a recent post on the Library of Congress Digital Preservation blog, *The Signal*.[77] The present work thus contributes to the body of literature documenting digital preservation workflows in

---

[72]*BitCurator*. http://www.bitcurator.net/ (accessed August 2012).

[73]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 (accessed August 2012).

[74]Ibid.

[75]See "Publications," *BitCurator* http://www.bitcurator.net/publications/ (accessed August 2012).

[76]See AIMS Project Group, "Appendix F: Policies, Templates, Documents, Etc.," *AIMS Born-Digital Collections*, 108-124; John et al., "Forensics," *Digital Lives*, 71-75, and "11.2. Seven Steps in Contemporary Archiving, with Exemplar Tools," *Digital Lives*, 189-198; and Brad Glisson and Rob Maxwell, "A Digital Forensics Workflow," in Kirschenbaum, Ovenden, and Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, 16.

[77]William Bradley Glisson, "Use of Computer Forensics in the Digital Curation of Removable Media," In: Tibbo, H.R. (ed.) *Digital Curation: Practice, Promise and Prospects: Proceedings of DigCCurr 2009*, April 1-3, 2009, Chapel Hill, NC, USA. School of Information and Library Science, University of North Carolina at Chapel Hill. (2009): 110-111, http://eprints.gla.ac.uk/33687/1/33687.pdf (accessed August 2012). Library of Congress blog posts include Bill Lefurgy, "Visualizing Digital Preservation Workflows," March 8, 2012. *The Signal: Digital Preservation* http://blogs.loc.gov/digitalpreservation/2012/03/visualizing-digital-preservation-workflows/ (accessed July 2012).

institutions using digital forensic tools and processes to acquire, preserve, and provide access to born-digital content.

Research Design

The purpose of this study was to explore the ways in which collecting institutions have implemented digital forensic tools and practices in workflows for acquiring and preserving born-digital content. My primary research questions have been:

- What digital forensic tools and practices are implemented in collecting institutions that are acquiring born-digital content?
- What challenges and successes have archivists encountered in their use of digital forensics tools to manage born-digital content?
- Do any commonalities exist in the ways in which collecting institutions implement digital forensics into their born-digital content workflows?

My approach was to collect qualitative data through semi-structured interviews with digital archivists, curators, and electronic records professionals who have demonstrated an engaged interest in digital forensics applications and who, through scholarship and other professional activities, are recognized leaders in the field of digital preservation. These interviews focused on the specific steps involved in the workflows for acquiring, preserving, and providing access to born-digital content in participant institutions, and how the implementation of digital forensics tools and processes has been challenging or beneficial in the execution of their responsibilities. Participant responses were then used to generate workflow maps and accompanying detailed narratives of the archival processes they described.

Research Design: Participants

A total of nine participants were interviewed for this study. Participant recruitment was based on several factors: active work in digital preservation efforts in collecting institutions, involvement in previous and ongoing projects emphasizing digital forensics, and overall demonstrated leadership in the field of digital preservation through publications, presentations, and other professional activities. I began with an initial list of 14 possible participants, whom I contacted to request interviews. All of the those initially contacted are involved in the BitCurator Project, which is developing digital forensic tools for use in libraries, archives, and museums.[78] Through this initial contact three additional individuals were suggested for participation, bringing the total number of archivists, curators, and electronic records professionals invited to participate to seventeen. Of the seventeen professionals invited to participate, nine were ultimately interviewed, and of those nine there were eight workflow maps produced.[79]

The participants hold a variety of positions and titles within their respective institutions. Several hold more than one title, reflecting the variety of responsibilities held by those working in collecting institutions. Similarly, participants work in a variety of collecting institutions, including university archives, special collections and manuscript collections, research centers, city archives, national libraries, and private organizations, in three different countries over two continents. The number of archivists and records professionals actively presenting and publishing in the area of digital forensics and digital

[78]BitCurator is funded by the Andrew W. Mellon Foundation. More information about the project is available on page 21 of this work, as well as in Appendix D: Digital Preservation Tools and Technologies, and on the BitCurator Project website at http://www.bitcurator.net/.

[79] Due to the time constraints of this project, I was unable to confirm all the details of the Library of Congress workflow map, and for this reason I elected not to use the Library of Congress workflow for this project.

preservation is not large; recognizing this small population size, participants agreed to be identified for this project and are listed below with their institution and current position titles.

Participant	Position title(s)	Institution/Organization
Bradley Daigle	Director of Digital Curation Services/Digital Strategist for Special Collections	University of Virginia Libraries
Michael Forstrom	Archivist	Beinecke Rare Book and Manuscript Library, Yale University
Leslie Johnston	Chief of Repository Development/Manager of Technical Initiatives	Library of Congress/National Digital Information Infrastructure Preservation Program (NDIIPP)
Matthew G. Kirschenbaum	Associate Professor/ Associate Director	University of Maryland English Department/Maryland Institute for Technology in the Humanities
Mark Matienzo	Digital Archivist	Manuscripts and Archives, Yale University Library
Courtney Mumma	Digital Archivist/Archivematica Community Manager	City of Vancouver Archives/Artefactual Inc.
Erin O'Meara	Electronic Records Archivist,/Manager of Campus Records	University Archives and Records Management Services, University of North Carolina at Chapel Hill[80]
David Pearson	Manager, Digital Preservation Section	National Library of Australia
Seth Shaw	Electronic Records Archivist	Duke University Archives

[80]Erin O'Meara is now employed as an archivist at the Gates Family Archive, but at the time of the interview had intimate knowledge of the born-digital content workflow for University Archives at the University of North Carolina at Chapel Hill.

Research Design: Methodology

Initial contact with participants was established via email.[81] This email solicitation provided a brief summary of my research topic, and was individualized from a template to provide context relevant to the participant and her/his scholarly activities. The initial email also included a copy of the statement of informed consent.[82] This document outlined the expectations for participation, and the ways in which the data generated through interviews would be used and protected. Participants were not obligated to sign the statement of informed consent; language included in the email solicitation stated that continuing participation indicated agreement with the statement and its terms. As previously noted, of the seventeen total solicitation emails sent, there were nine respondents who agreed to be interviewed. After the initial email and response, follow-up emails established a suitable date and time for the interview to be conducted, and provided answers to any questions that the participants had. Follow-up emails were also used to provide participants with the interview protocol, containing the questions that would be asked in the interview.[83]

Because of the exploratory nature of the study, I chose to conduct semi-structured interviews. Each interview lasted approximately 40-60 minutes. The first question, which was fairly consistent across all interviews, was for the participant to identify her/his position and the institution in which she/he is employed, in order to provide context for the participant's role in the management of born-digital content. The order and precise wording of subsequent questions varied, in order to accommodate the flow of discussion and nature of participant answers. The majority of each interview comprised a detailed

[81]The initial solicitation can be found in Appendix A: Participant Solicitation Email.
[82]The statement of informed consent can be found in Appendix B: Statement of Informed Consent.
[83]The interview protocol can be found in Appendix C: Interview Protocol.

discussion of the individual steps taken within participant institutions upon acquiring born-digital content. Participants also discussed access mechanisms, and methods for arranging and describing born-digital content in their institutions.

For the purpose of my study, I chose to give participants open interpretation of what constitutes born-digital content, as I wanted to allow participants to speak freely and avoid pinning them down to a specific definition. The interviews also provided participants the opportunity to reflect upon the successes and challenges experienced in implementing their current workflows, changes they would like to make, and what they have learned through the creation and evolution of their current processes.

Interviews were conducted in May 2012 via Skype or telephone, and recorded for transcription. In order to allow participants to speak candidly about the problems and challenges they faced in their respective institutions, interview audio files and their subsequent transcription text files were de-identified and assigned a random number, which was associated with a letter for participant labeling in this work. All audio recordings and related materials have been maintained in accordance with the University of North Carolina at Chapel Hill's recommendations for level 2 data security.[84]

Following the interview, an audio transcription was created and used to map institutional workflows for the management of born-digital content. These workflow maps were created using LucidChart, an online browser-based application for process modeling.[85] Upon completion of an institution's workflow map, I presented it to the participant, along with any questions that might have come up as a result of the

[84]Information on level 2 data security requirements can be found at: http://research.unc.edu/ccm/groups/public/@research/@hre/documents/content/ccm3_035154.pdf (accessed June 9, 2012).

[85]Information on LucidChart is available at https://www.lucidchart.com/.

transcription or workflow mapping process. In most cases, the process went through several iterations and refinements before the workflow map was deemed by the participant to be an accurate representation of institutional processes. The completed versions of each institutional workflow map for managing born-digital content can be found in the following section. Transcriptions were analyzed using open coding to identify parallels and common threads throughout the interviews.

A Note on Workflow Mapping

At its most fundamental, process mapping "makes work visible."[86] Process mapping is a way of visualizing and documenting the activities of an organization such that it: "can be performed effectively and efficiently," "[c]an be managed effectively," and it "[o]ffers the potential for a competitive advantage."[87] A flowchart, one of the most basic methods of process mapping, is "a graphic representation of the sequence of steps that make up a process."[88] Flowcharts are frequently used to represent the steps in a high-level workflow, where a workflow is defined as a sequence through which a piece of work passes from start to finish.[89] I have chosen the term *workflow map* for the diagrams I have created, as they *map* the current practices and processes that make up the *workflows* for managing born-digital content within each of the institutions represented by participants. The workflow maps are accompanied by a legend that defines the different processes that are visually depicted, and are based upon workflow symbols

[86]Robert Damelio, *The Basics of Process Mapping*. (Portland, OR: Productivity, Inc., 1996), 1.
[87]Alan J. Raimas and Richard Rummler, "The Evolution of the Effective Process Framework: A Model for Redesigning Business processes." *Performance Improvement* 48 no. 10 (November/December 2009), 26, http://onlinelibrary.wiley.com/doi/10.1002/pfi.20112/abstract (accessed May 2012).
[88]Damelio, *The Basics of Process Mapping*, 10.
[89]"Workflow," *Wikipedia – The Free Encyclopedia* http://en.wikipedia.org/wiki/Workflow (accessed June 2012).

provided in Robert Damelio's *The Basics of Process Mapping*.[90]

Findings

Beinecke Rare Book and Manuscript Library, Yale University

Since its 1963 opening, the Beinecke Rare Book and Manuscript Library at Yale University has been the "principle repository for literary papers and for early manuscripts and rare books in the fields of literature, theology, history, and the natural sciences."[91] One significant part of these collections has been the Yale Collection of American Literature, where the Beinecke has developed a particular strength in manuscript collections of early 20th century authors.[92] In addition to this manuscript material, which has been mostly paper-based, the Beinecke has also begun to acquire born-digital content along with more contemporary collections. The acquisition of these materials, and the relative lack of case studies with respect to born-digital content in manuscript collections, led Archivist Michael Forstrom to publish a case study on the Beinecke's management of personal electronic records which, along with the interview I conducted for this project, form the basis of the Beinecke born-digital content workflow.[93]

In what will be a continuing theme in many of the interviews I conducted, the work of acquiring born-digital content from donors begins long before the material arrives at the archives.[94] Initial conversations involving curators, collection development,

---

[90]Damelio, *The Basics of Process Mapping*, 10. The workflow legend is available following each of the workflow maps in the Findings section.

[91]"Beinecke Rare Book and Manuscript Library: About the Collections," *Yale University: Beineicke Rare Book and Manuscript Library* http://www.library.yale.edu/beinecke/brblinfo/brblguide.html (accessed June 2012).

[92]Michael Forstrom, "Managing Electronic Records in Manuscript Collections," 461.

[93]Interview with Michael Forstrom, conducted on May 29, 2012.

[94]I determined that this pre-custodial work was outside the scope of this work. However, the importance of developing relationships with donors early on to facilitate the transfer of born-digital content to

and archivists ultimately brings digital content across the archival threshold, where it is identified during the accession process and assigned an identifier associated with the collection's accession number, which is retained in the archival collection management system, Archivists' Toolkit.[95] In the case of digital media carriers such as hard drives, floppy disks, and optical disks—which form the majority of the Beinecke's ingested digital content at this time—additional metadata is collected about the carrier and maintained in a separate database before processing of the digital content begins.[96]

One important thing that was difficult to convey in this visual workflow is the Beinecke's active collaboration in recent years with other collections within the Yale University Libraries system to develop the workflow that is currently used. This is due to their participation in the recently completed *AIMS Born Digital Collections* project, which produced the report, *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*.[97] Interview participants representing both the Beinecke and Yale University Library Manuscripts and Archives noted this collaboration as one of the most

---

collecting institutions is a theme that has been widely discussed in archival literature. An excellent and relevant recent example is: AIMS Project Group, "Collection Development," *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*, 4-16. The ways in which to structure donor agreements to document "curatorial intent," regarding the level of content representation to be preserved for future access, are still developing. In addition to the *AIMS* project, see: Christopher A. Lee, "Donor Agreements," in Kirschenbaum, Ovendon, and Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, 57; and Nick del Pozo, Andrew Stawowczyk Long and David Pearson, "'Land of the lost:' a discussion of what can be preserved through digital preservation," *Library Hi Tech* 28 no. 2 (2010): 290-300, doi: 10.1108/07378831011047686 (accessed June 2012).

[95]Acquisition of material may also occur online via file transfer, or via a site visit to the donor's home computing environment. In those cases, a disk image is created, either of the transferred files on a separate drive within the processing station, or of the home computing environment where the target content resides. For more information, see "Archivists' Toolkit," Appendix D: Digital Preservation Tools and Technologies.

[96]This "preliminary processing" captures technical information about the carrier such as brand, type and format, serial number, and if there is any exterior labeling on the media, and also tracks the progress of the media through the process of extracting its digital content.

[97]AIMS Project Group, *AIMS Born-Digital Collections*, v-viii.

important factors in successfully implementing their born-digital content workflows.[98] As a result, the core of each workflow is similar, with differences reflected in the methods of acquisition and the provision of access.

After the digital content has been transferred using a write-blocker to a processing station, the next step is capturing and stabilizing the digital content on the media by creating a disk image. Yale University Libraries has been active in testing different tools used to create disk images, and in particular tools used to image floppy disks—3.5 inch and 5.25 inch diskettes that were standard carriers of digital content in the late 20th century.[99] Tools such as FTK Imager, CatWeasel, KryoFlux, and the FC5025 universal controller have different strengths and weaknesses, and thus one major goal for the Yale University Libraries has been to determine in what circumstances, and for what types of material, each tool is best suited.[100] Following the creation of the disk image, a checksum of the disk image is generated and separately maintained to verify in the future that no data has been altered or corrupted through preservation activities.[101]

The disk image is next analyzed using fiwalk, a Python-based automated forensic analysis program that compiles information on the individual file objects within the disk image. Plug-ins used in the Yale University workflows also conduct a virus check and file format identification. fiwalk outputs an Extensible Markup Language (XML) file that

---

[98]Interviews, Michael Forstrom and Mark Matienzo. My interview with Mark Matienzo took place on May 10, 2012.

[99]White, *How Computers Work*, 167. White states: "For all its deficiencies, the floppy drive was an under-appreciated wonder of its time. Think of it: An entire book full of information can be contained on a cheap disk that you can slip into your pocket. Until the advent of small, cheap USB-based flash drives, floppy drives were still found on nearly every PC, making them a sure an convenient way to get small amounts of data from one PC to another."

[100]For more information on these tools, see Appendix D: Digital Preservation Tools and Technologies.

[101]The exception to this is FTK Imager, which generates a checksum as part of its imaging process. See "FTK Imager" in Appendix D.

conforms to the Digital Forensics XML (DFXML) metadata schema.[102] The DFXML file and a copy of the disk image are then packaged using the BagIt object packaging specification developed by the Library of Congress and California Digital Library.[103] The BagIt package allows the disk image and any associated metadata to be packaged as part of the same "bag," enabling higher-level description and a manifest of the bag contents to be archived in Yale University Libraries digital storage, the Rescue Repository.

The Rescue Repository is a library-wide managed storage environment specifically for the digital materials acquired and ingested by each collection in Yale University Libraries.[104] Within the Rescue Repository, files are verified and validated by JSTOR/Harvard Object Validation Environment (JHOVE), and ingests are monitored through ingest activity logs by a nightly reporting script.[105] The Rescue Repository is a "dark archive," in that it does not allow public access to materials, but only internally to those with proper permissions—mostly the administrators of that material from each of the collections with the ability to ingest into the Repository.

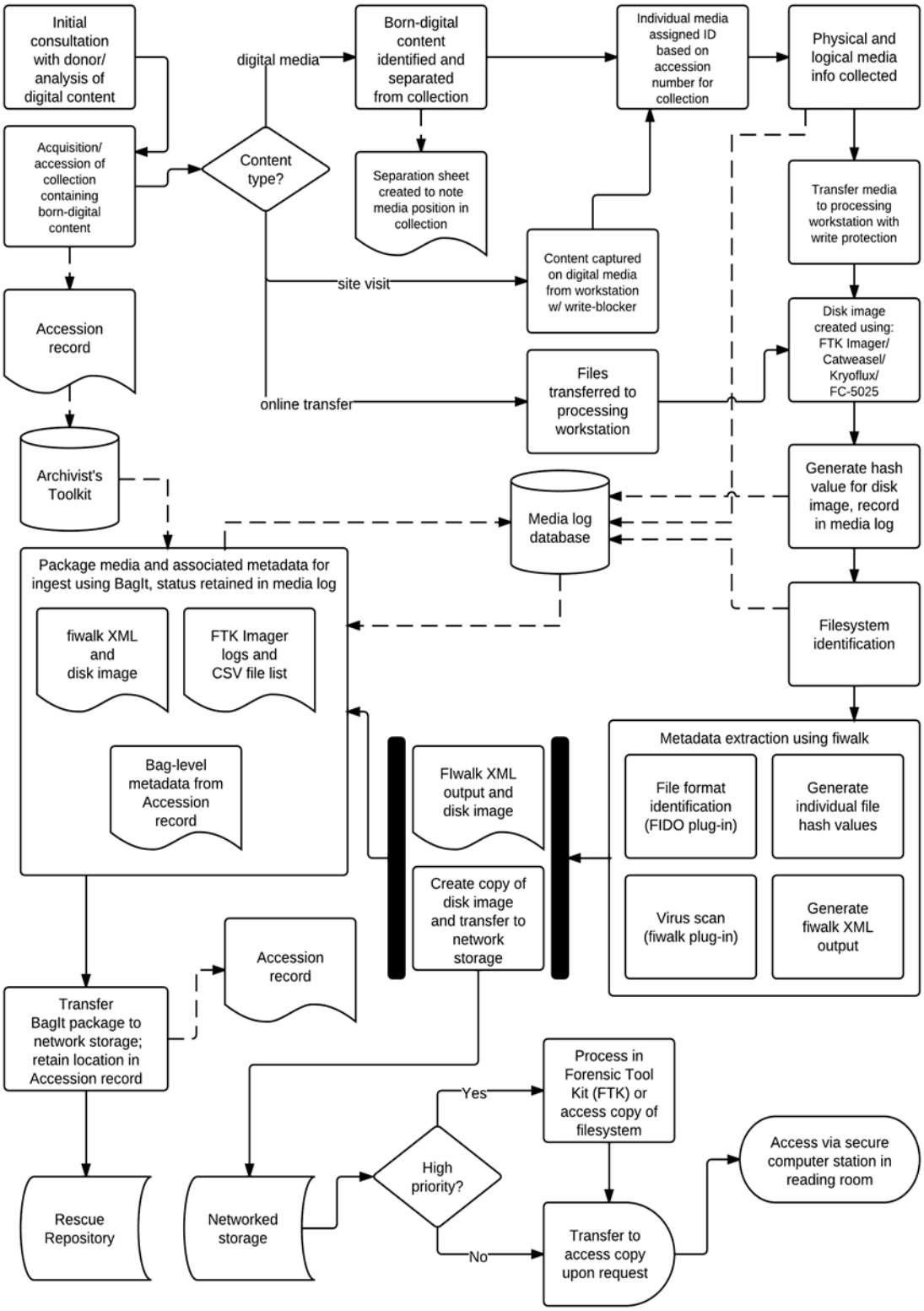
For this reason, the Beinecke has provided an alternative method of access by copying the disk image before it is added to a BagIt package. This access disk image copy is maintained on separate network storage, and may be subject to further processing with a robust forensic analysis tool, the Forensic Toolkit (FTK).[106] This tool is used for arrangement and description, and to define the levels of access available to the user. Accessible material is then provided to the user on a secure access station in the Beinecke

---

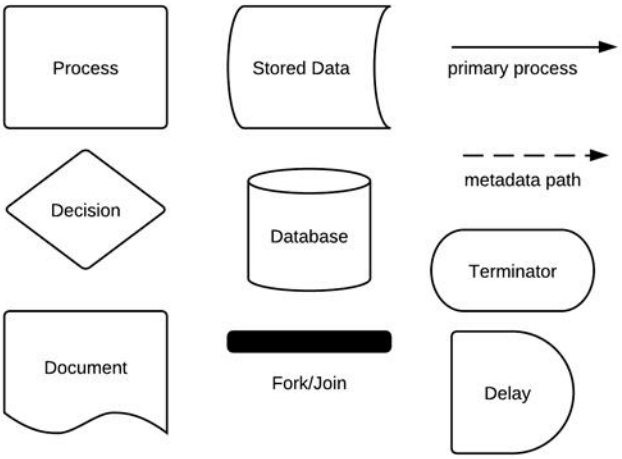
[102]See Appendix D, "fiwalk" and "Digital Forensics XML."
[103]See Appendix D, "BagIt."
[104]"Yale University Rescue Repository: About the Rescue Repository," *Yale University Library: Integrated Systems and Programming Group* http://www.library.yale.edu/ito/RRweb/AboutRescueRepository.html (accessed July 2012).
[105]Ibid.
[106]Note that this is a different program from FTK Imager. For more information, see Appendix D, "Forensic Toolkit (FTK)."

Reading room. The Beinecke Rare Book and Manuscript Library born-digital content workflow map is provided below.

Figure 1. Beinecke Rare Book and Manuscript Library, Yale University



Workflow Map Legend



Manuscripts and Archives, Yale University Library

The Yale University Library, Manuscripts and Archives developed from collections assembled by Yale faculty over the 1950s and 1960s, becoming a fully established unit of the Library in the late 1960s.[107] Like many other institutions, Manuscripts and Archives has encountered incoming born-digital content primarily in the form of "disks in boxes."[108] To better manage this content, collections within the umbrella of Yale University Libraries have begun to collaborate on the development of processes for managing born-digital content in their respective collections. As a result of this collaboration, the Manuscripts and Archives workflow bears similarities to the Beinecke Rare Book and Special Collections Library. Differences in ingest and provision of access will be highlighted through this description, and in the resulting workflows, generated as a product of my interview with Digital Archivist Mark Matienzo.

The majority of the content with which Manuscripts and Archives has been working comes into the collection in the form of removable media such as floppy and optical disks, and there is a fairly straightforward workflow developed around acquiring digital content from these media. Accessioned digital media carriers, when identified by processing archivists, are individually labeled and associated with the accessioned collection. Logical and physical metadata about the physical carrier are then collected prior to the capture of digital objects. This metadata is kept in a separate database called the Media Log, while the physical carrier is connected to a processing station, using a write-blocker to prevent any unintentional alteration of the media carrier's content.

As previously mentioned, there are several tools that are being tested in the

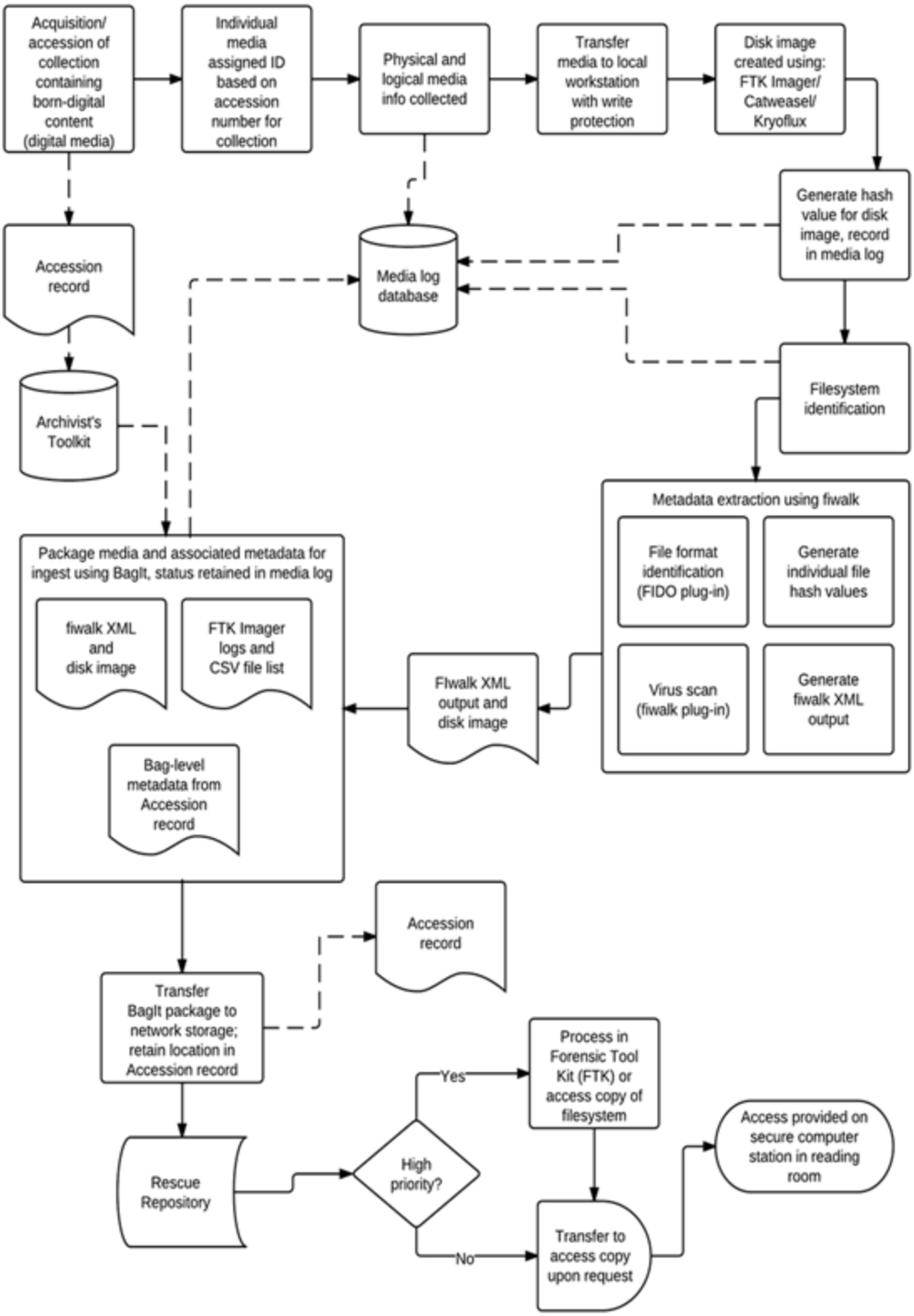
[107]"About Manuscripts and Archives: Introduction," *Manuscripts and Archives, Yale University Library* http://www.library.yale.edu/mssa/about_intro.html (accessed July 2012).
[108]Interview with Mark Matienzo, conducted on May 10, 2012.

processing workflow for the generation of a floppy disk image, including the CatWeasel, KryoFlux, and FC5025.[109] Following the creation of a disk image, a hash of the image file is generated and there is an attempt to determine the file system in use on the target digital media. This is confirmed later in the process by fiwalk, the automated forensic analysis tool, though they also attempt to make an "educated guess" to facilitate the process.

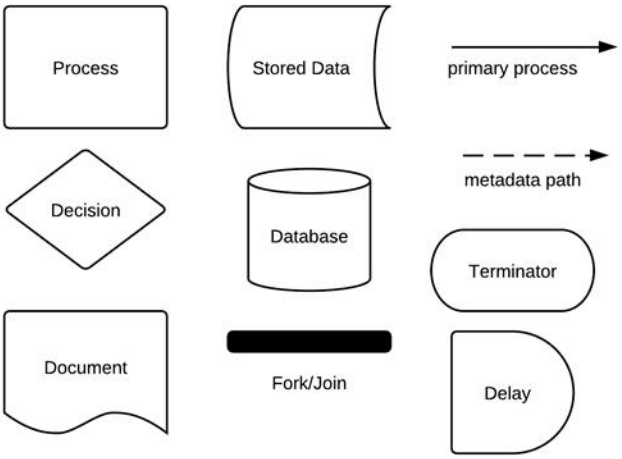
Whereas the Beinecke generated a second copy of the disk image to facilitate access, Manuscripts and Archives has focused on the stabilization and preservation of the born-digital content found on external media such as floppy and optical disks. The fiwalk output and disk image are packaged in conformance to the BagIt specification, and ingested into the Rescue Repository. In cases when additional processing has been required, the disk image is mounted for processing with FTK; otherwise, an access copy of the ingested content is transferred to a user upon request. Similar to the access method at the Beinecke, Manuscripts and Archives uses a secure computer station within the reading room to facilitate access to born-digital materials. The Yale University Library Manuscripts and Archives born-digital content workflow model is provided below.

[109]This is discussed in the previous workflow description, for the Beinecke Rare Book and Manuscript Library; also see Appendix D, "CatWeasel," "KryoFlux," and "FC5025."

Figure 2. Yale Manuscripts and Archives



Workflow Map Legend



City of Vancouver Archives

The City of Vancouver Archives (CVA) in Vancouver, British Columbia, Canada, has the mandate "to acquire, organize, and preserve Vancouver's historical records and make them available to the widest possible audience."[110] The CVA has been actively collecting since 1933, and primary holdings include City of Vancouver government records, and records of non-government organizations, businesses, and individuals, as well as photographs, maps, and other materials.[111] Prior to 2008, the CVA had no preservation program specific to born-digital material, though there has been an active digitization unit scanning historic photographs since 1996.[112] With the 2010 Winter Olympics to be held in Vancouver, and the recognition that a large portion of associated records would be created in digital form, the CVA partnered with Artefactual Systems Inc., to develop a digital preservation system. Courtney Mumma was hired in 2009 to work on the project to acquire and process the incoming collection.[113] The process developed for that collection has become the de facto workflow for subsequent born-digital content entering the CVA.

The acquisition of born-digital content begins with an analysis of the recordkeeping systems in use by the donating party. This allows CVA staff to determine what the most important records are to retain, and how they are being used in their

---

[110]"City of Vancouver Archives – Home Page" *City of Vancouver Archives* http://vancouver.ca/ctyclerk/archives/ (accessed July 2012).

[111]Ibid., and Courtney Mumma, Glen Dingwall, and Sue Bigelow, "A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value="true";," *Archivaria* 72 (Fall 2011): 93, http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13361 (accessed July 2012).

[112]Interview, Courtney Mumma, conducted on May 17, 2012.

[113]Interview, Courtney Mumma. The work undertaken by Mumma and others in processing the records of the Organizing Committee for the 2010 Olympic and Paralympic Winter Games (VANOC) is documented in great detail in Mumma, Dingwall, and Bigelow, "A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds," 2011.

primary context, and how to best proceed with the transfer of physical control of the digital content. Born-digital material is transferred to the CVA, where individual media receive identifiers associated with the collection and accession. This information is maintained in a metadata spreadsheet.

The next step is to create a disk image of the content; the process for this activity depends upon the type of media carrier in question. Imaging external hard drives requires the use of a physical write-blocker to guard against unintended transfer of information from the processing workstation to the media carrier. The CVA uses Linux processing stations with blank external drives to accept the disk image files, formatted for Linux.[114] Disk images are generated in several formats, including ISO, AFF, and Encase, and are linked to metadata in activity logs and spreadsheets. Staff generate checksums from the disk images, which are then transferred onto external drives to await processing.[115]

Optical disks are loaded to a Linux workstation to begin imaging.[116] Disks that are readable are imaged using dd, and a checksum is generated for later validation. If there is an error in accessing the optical disk, processors attempt to recover data using ddrescue, a command-line based data copying and recovery program. If there is still a problem accessing data, the disk is set aside and a note is made in spreadsheet documentation. Successfully recovered media have a checksum generated, and content is transferred to

---

[114]For more information on write blockers, see Appendix D, "Write-blockers." It is also important to note that not all of the drives necessarily have to be formatted for Linux. For example, for working with Apple products, the drives could be formatted as HFS+ non-journaled disks.

[115]This is an update of earlier procedures. Initially, when working with materials from the VANOC fonds, Mumma and her cohort chose to use rsync, which does not create a disk image but a copy of the disk contents with metadata logging and auditing. When digital forensic practices developed around the use of disk images, CVA shifted their internal practices. For more information on the process that had been previously used, see Mumma, Dingwall, and Bigelow, "A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds," 112. Additionally, there is no policy in place for the disposition of original media carriers; they are currently maintained in storage, though no further preservation action has been taken upon them.

[116]Write blockers are unnecessary for imaging optical discs, as there is no danger of writing onto the disk in the process of reading its contents.

external drives to await processing.

Once material has been successfully copied off its original media and stabilized on external drives, the workflow for managing born-digital content is largely driven by Archivematica, the CVA's digital preservation system.[117] Staff load disk images protected by a write-blocker onto a workstation with Archivematica, which extracts files designated for preservation into an automated pipeline that manages preservation actions and provides opportunities for the processor to supply additional metadata and appraise records.[118] A high-level overview of the Archivematica pipeline follows.

First, Unique Universal Identifiers (UUIDs) are assigned to individual file objects, and checksums are generated, verified, and stored as associated metadata in a file using the XML-based Metadata Encoding and Transmission Standard (METS).[119] ClamAV is used to check files for viruses, and file directory and transfer path names are verified. Files are characterized using the File Identification Tool Set (FITS), and bundled into a Submission Information Package (SIP).[120] Staff populate fields with metadata from spreadsheets, provide ingest submission documentation for the SIP, and verify checksums to insure the integrity of individual file objects through the preservation pipeline. The METS file is populated with additional metadata from the Preservation Metadata: Implementation Strategies (PREMIS) metadata schema, and a Dissemination Information Package (DIP) is generated, along with an Archival Information Package (AIP) that is

---

[117]For more information on Archivematica, see Appendix D, "Archivematica."

[118]Detailed information on the preservation activities managed by Archivematica is available at the Archivematica Wiki, at: https://www.archivematica.org/wiki/Main_Page (accessed June 2012).

[119]See Appendix D, "Metadata Encoding and Transmission Standard (METS)."

[120]For more information on FITS, see Appendix D, "File Identification Tool Set (FITS)." Archivematica can support multiple types of ingest file packages, including DSpace Simple Archive Format and BagIt. See "Transfer and SIP Creation," *Archivematica* https://www.archivematica.org/wiki/Transfer_and_SIP_creation (accessed June 2012).

bundled with the BagIt specification.[121] The AIP is preserved in archival storage, while the DIP is transferred to access storage. Descriptive metadata is entered through ICA-Atom, the archival description and access software developed by Artefactual Systems, Inc.[122] ICA-Atom is also the access system for publicly available born-digital content.

One step not reflected in the workflow involves the multiple levels of appraisal that are applied to born-digital content at the CVA. Appraisal in the CVA system occurs in three stages: "1) Selection for Acquisition; 2) Selection for Submission; and 3) Selection for Preservation."[123] Appraisal is successively narrowed in scope, from the acquisition of material to be transferred to the archive, to selecting material to be included in the creation of SIPS, to which specific files in a SIP ought to be included for final preservation within the archival repository. This is an important component of the digital preservation activity undertaken at CVA, that unfortunately falls outside of the scope of the present work. However, the application of this appraisal framework to the CVA born-digital content workflow is described in great detail in published materials on the acquisition of the Vancouver Olympic Commission digital materials.[124] The City of Vancouver Archives born-digital content workflow model is provided below.[125]

---

[121]Information on PREMIS can be found in Appendix D, "Preservation Metadata: Implementation Strategies (PREMIS)." SIP, AIP, and DIP are all terms associated with the Open Archival Information System Reference Model (OAIS), a high-level framework for developing repositories for preserving digital information. See Christopher A. Lee, "Open Archival Information System (OAIS) Reference Model," *Encyclopedia of Library and Information Sciences, Third Edition* http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120044377 (Taylor & Francis, 2010), 4020-4030.

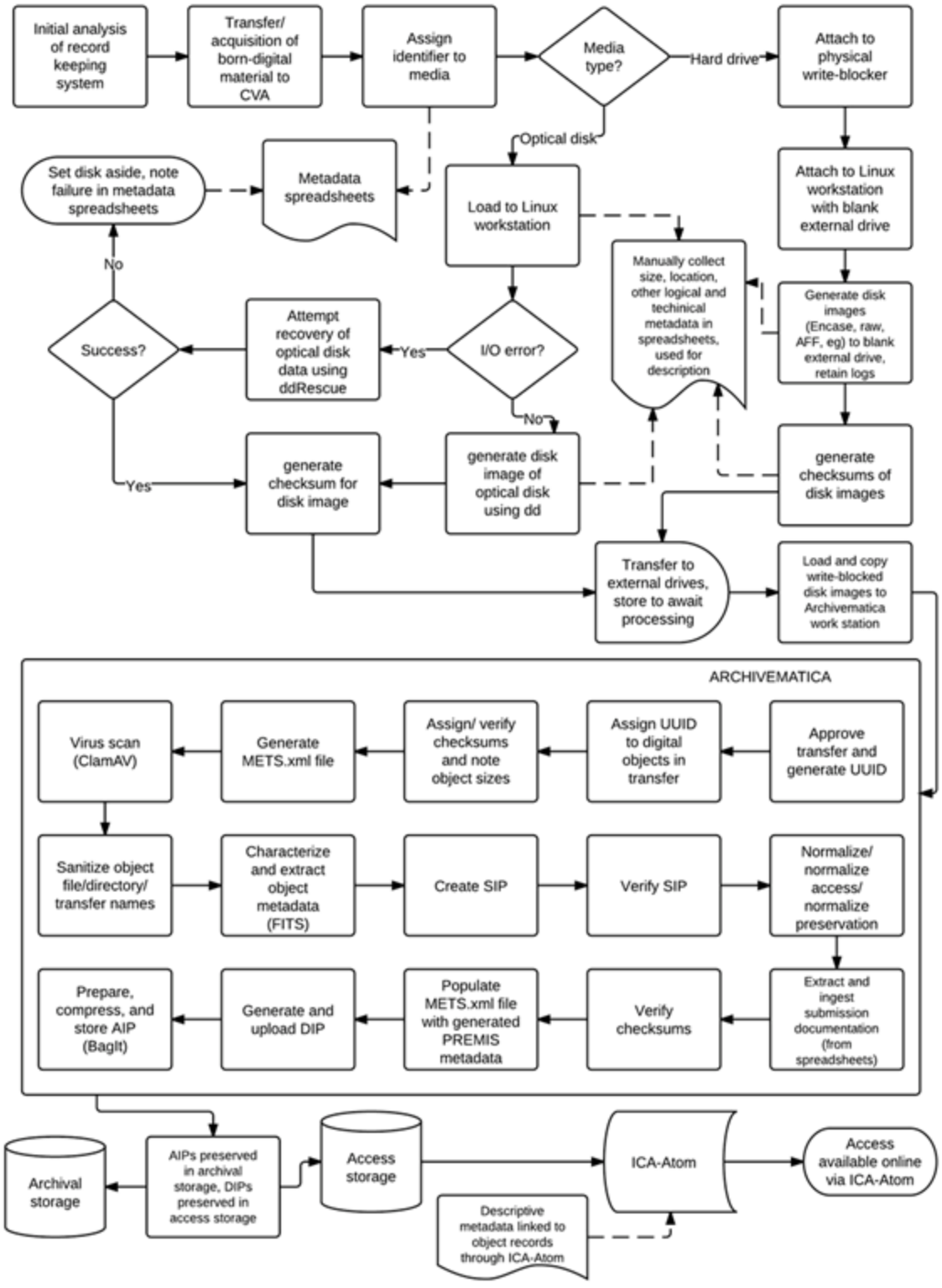
[122]For more informaton on ICA-Atom, see Appendix D, "ICA-Atom."

[123]Mumma, Dingwall, And Bigelow,111.

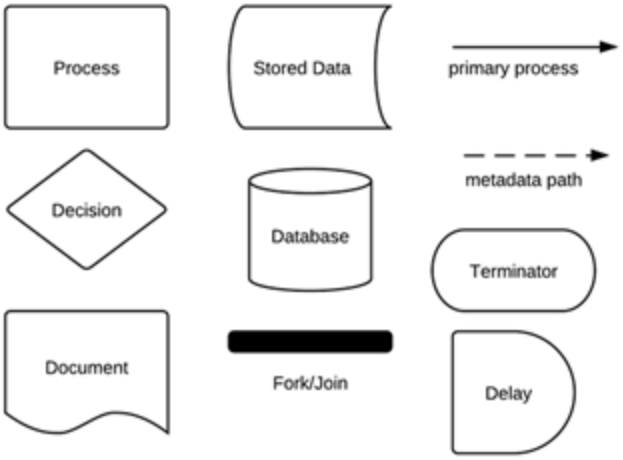
[124]Ibid., 111-120.

[125]Steps in this workflow have been developed from documentation available on the Vancouver Digital Archives wiki, at: http://artefactual.com/wiki/index.php?title=Requirements_Analysis#Ingest (accessed June 2012); and the Archivematica wiki, at: https://www.archivematica.org/wiki/Archivematica_0.8_Micro-services (accessed June 2012).

Figure 3. City of Vancouver Archives (CVA)



Workflow Map Legend



Duke University Archives

The Duke University Archives was established in 1977 as "the official repository for University records of enduring value."[126] In addition to this collections mandate, the University Archives collections include campus publications, audio-visual materials, faculty papers, student and employee organizational records, theses, dissertations, and other select student and scholarly output.[127] As other interviews revealed, born-digital material—much of it floppy disks, optical disks, and other removable media—had long been present the Duke University Archives, mixed within accessioned analog collections. This born-digital content began to appear in collections as many as fifteen years ago, or even earlier. In 2007 Duke University Archives hired Seth Shaw to further develop initial steps that had been taken to manage born-digital content in the department.[128]

When a donor has born-digital material that they wish to donate to the University Archives, the Electronic Records Archivist is often brought in by curators to help facilitate that process. When digital material is successfully delivered to the archives, it is associated with an accession number, regardless of type of media. The accession number connects the digital objects or media carrier to a record in the University Archives' management tool, Archivists' Toolkit (AT). Some accessions, such as files donated via email or a secure FTP (file transfer protocol) dropbox, are copied directly to a networked processing station, where checksums can be generated. For removable media that is interfiled with paper records, a separation sheet is inserted where the media previously

---

[126]"About the Duke University Archives," *University Archives in the David M. Rubenstein Rare Book and Manuscript Library* http://library.duke.edu/uarchives/about/index.html (accessed July 2012).

[127]Ibid.

[128]Interview with Seth Shaw conducted on May 7, 2012. Though he is organizationally a part of University Archives, Shaw's role has grown to include processing and managing born-digital content in manuscript collections, as well.

resided among the paper records, to maintain its context within the collection. Both separation sheet and media are affixed with barcodes that serve as a unique media identifier (non-interfiled media just receive one barcode). The media is then photographed; the photographs and other physical and logical metadata are tracked by media identifier and accession number in a secondary media database, which also tracks the progress of the digital material at various points through the workflow process.

After the initial metadata collection and media separation stages, the tools used for imaging digital media differ by type: for floppy disks, University Archives uses the KryoFlux USB floppy controller; for optical disks such as CDs or DVDs, or external hard drives and thumb drives, an AFF disk image is created using FTK Imager.[129] The AFF (Advanced Forensic Format) disk image format allows lossless compression of the disk image file, as well as the ability to imbed disk image metadata within the AFF file.[130] Imaging media with FTK Imager can also automatically generate and verify checksums. For those media carriers which University Archives does not have a way to read, the media is tracked in a database. Otherwise, once a disk image has been created, University Archives generates a checksum for the disk image that will be used to verify in the future that no alteration or degradation has occurred. The disk image is scanned for viruses and a proprietary tool is used to scan for sensitive electronic information (SEI) such as social security or credit card numbers.

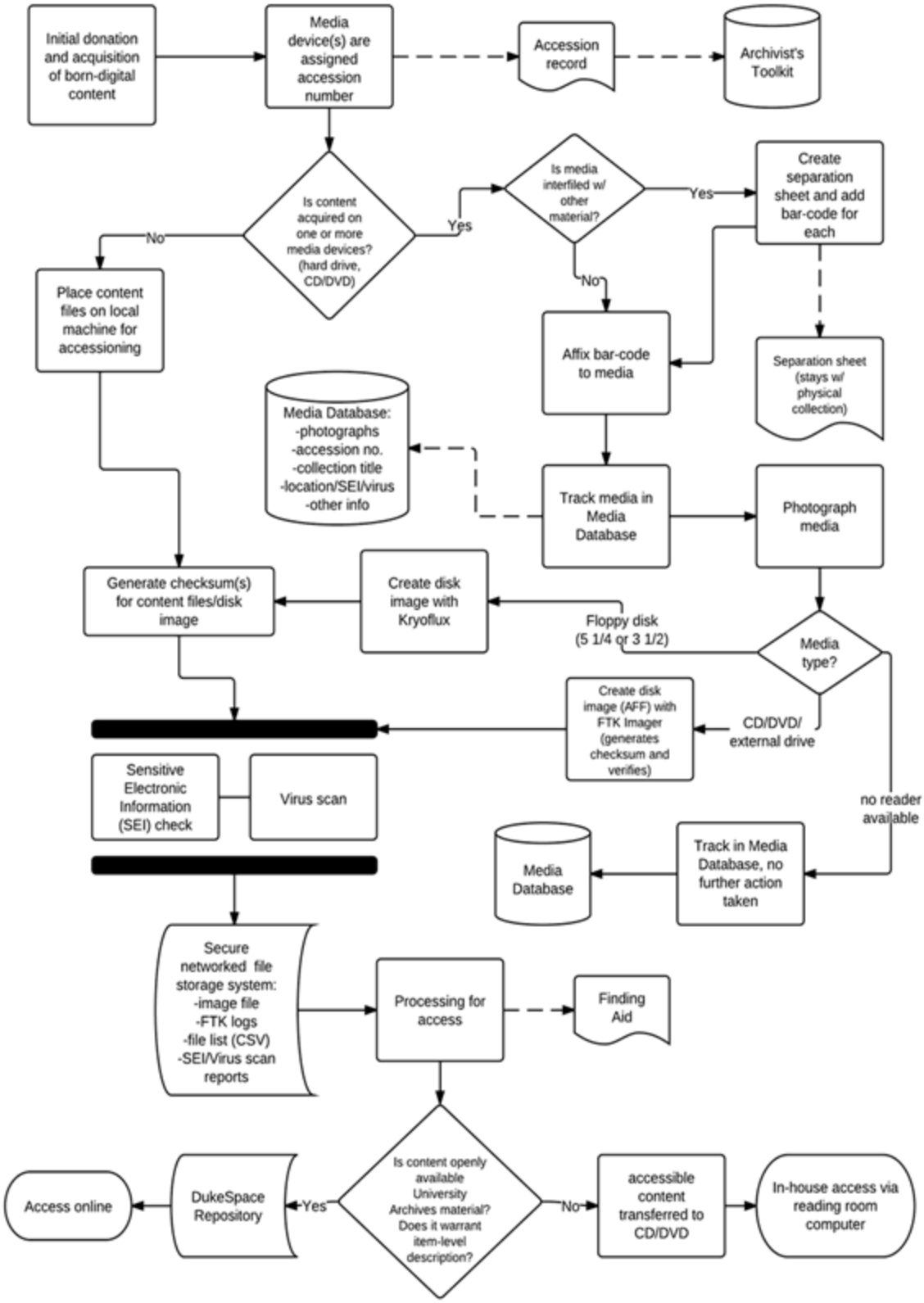
Following the virus and SEI scans, the content is ready to be moved to secure network storage. The disk image, a comma separated value format (CSV) file list

[129]More information on these tools can be found in Appendix D: Digital Preservation Tools and Technologies.

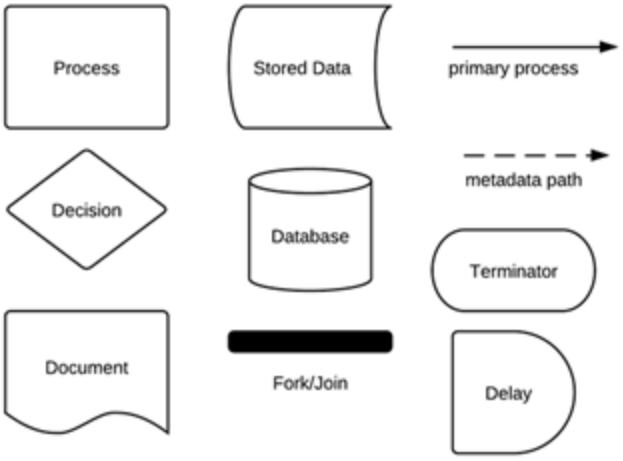
[130]More information on the Advanced Forensic Format can be found in Appendix D, "Advanced Forensic Format (AFF)."

extracted by FTK Imager, and logs generated as a result of the imaging process as well as any additional reports documenting the SEI and virus scans, are transferred to storage on a secure network managed within the library. There, digital content awaits processing for access. Finding aids for collections note the presence of digital content where available. For Duke University publications that are publicly available, access is provided via the DukeSpace repository, a local implementation of the DSpace repository system. Other content is made available by transferring files onto optical disks, which are provided to users on a non-network access computer station in the archive's reading room. The Duke University Archives born-digital content workflow model is provided below.

Figure 4. Duke University Archives



Workflow Map Legend



Maryland Institute for Technology in the Humanities, University of Maryland

The Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, College Park, was founded in 1999 with joint support from the University of Maryland College of Arts and Humanities and the University of Maryland Libraries.[131] It is not a collecting institution in the traditional sense. Rather, it is a digital humanities research center, specializing in "text and image analytics for cultural heritage collections, data curation, digital preservation, linked data applications, and data publishing."[132] It is in MITH's capacity as a research center that it has taken on several collections of born-digital content.

In 2007 MITH acquired the Deena Larsen Collection, including personal and professional papers and a significant amount of born-digital material, as well as several Macintosh computers Larsen used to exhibit her work.[133] MITH staff developed a donor agreement and deed of gift specifically for the acquisition, as Larsen had a prior working relationship with Kirshenbaum and MITH staff, and had offered MITH the collection. Upon arrival, it was evident that original order would not be significant in the processing of the collection.[134]

Following the acquisition of the collection, the first step was to create an initial

---

[131]"About | Maryland Institute for Technology in the Humanities," *Maryland Institute for Technology in the Humanities* http://mith.umd.edu/about/ (accessed July 2012).

[132]Ibid.

[133]Interview with Matthew Kirschenbaum, conducted May 15, 2012. Kirschenbaum has written extensively on the acquisition and processing of this collection. See Kirschenbaum, et al., "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use," (May 2009), and Kirschenbaum et al., *Digital Materiality: Preserving Access to Computers as Complete Environments* (2009). Following the accession of the Larsen collection, MITH acquired the collection of William Bly, a contemporary of Larsen. MITH staff are working with staff at Special Collections at the University of Maryland to develop processes for preserving the Bly Collection, and so this interview primarily focused on the procedures undertaken in work with the Larsen collection.

[134]Interview, Matthew Kirschenbaum. "The car pulled up to the loading dock here at the library, and the door opened, and...boxes and bags full of diskettes, papers, and hardware came spilling out. I deeply regret not having had the foresight to have had a camera with me."

inventory of its contents, both physical and digital. This first pass did not include any item-level inventory, only registering the presence of digital media when it was discovered. The inventory was maintained on an Excel spreadsheet, which became the finding aid for the collection. Physical materials (papers, folders, and other manuscript items) were processed according to traditional methods of arrangement and description.

The collection included, in addition to the computer systems mentioned above, nearly 800 3.5 inch diskettes—mostly formatted for Macintosh, with either 1.4-megabyte (MB) or 800-kilobyte (KB) capacity. These were assigned unique identifiers, which were added to the inventory spreadsheet. The diskettes themselves had write-blocking enabled through manipulation of a small tab on the underside of the diskette. When "flipped," this tab prevents material to be written to the diskette. Additional write-blocking protection is provided by loading the diskettes onto a computer with a different operating system, in this instance a Linux workstation. Disk images were generated for each of the diskettes using dd, and were transferred to network storage, where they are maintained for preservation.[135] The diskettes themselves have been returned to the collection.

The Macintosh computers acquired with the collection have been retained in order to provide access to the content on the diskettes according to the intentions of their creator. This is a significant area in which the MITH workflow has differed from that of more traditional collecting institutions. Both the original diskettes and the computers donated to MITH are available for patron use. Kirschenbaum states: "Deena herself was very clear on the point that she wanted this to be a living collection that people could have as much unfettered access to as possible."[136] Access is facilitated by staff at MITH,

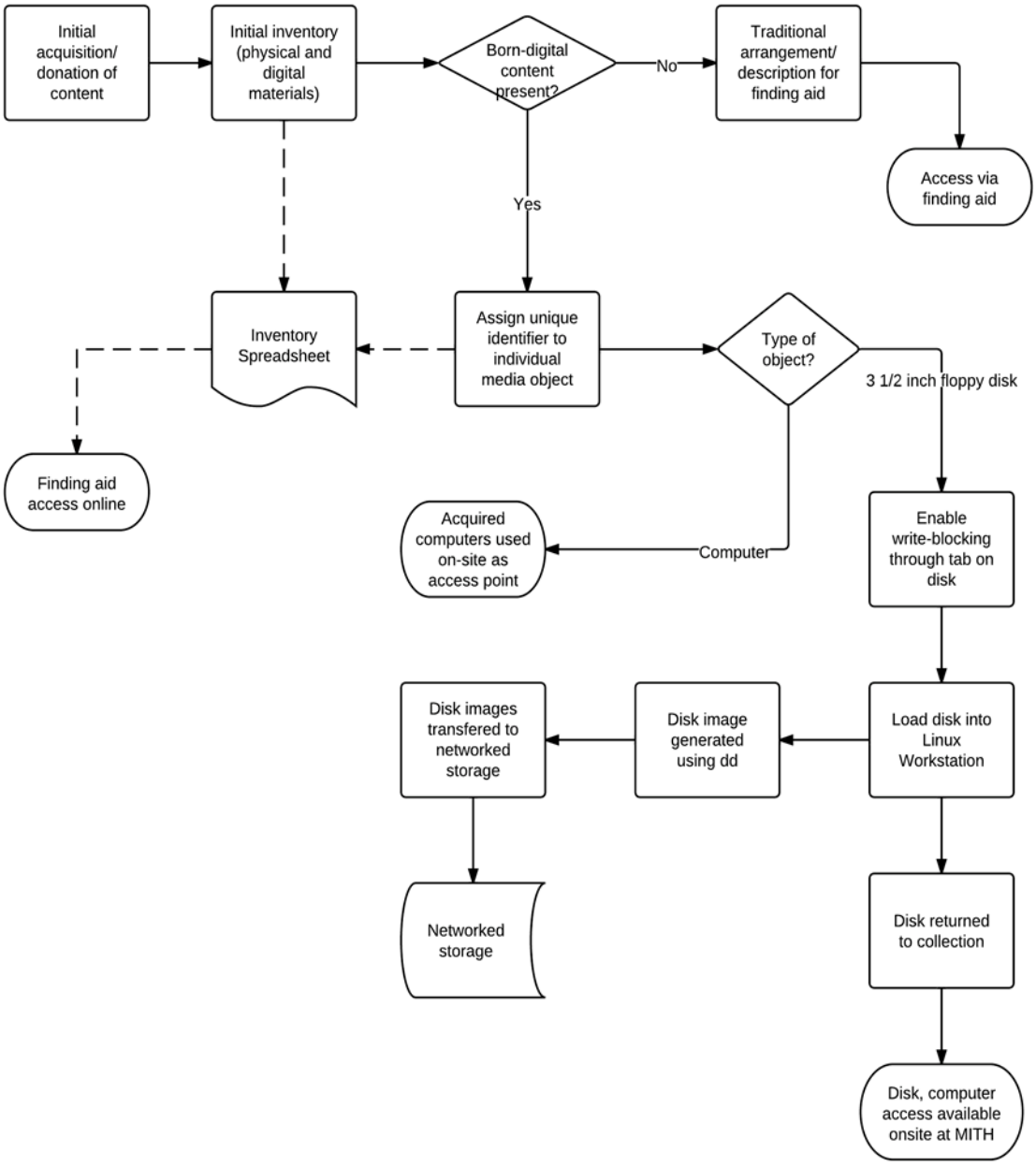
[135]For more information on dd, see Appendix D, "dd."
[136]Interview, Matthew Kirschenbaum.

who can provide either an un-redacted disk image of a diskette from the collection, or the physical diskette itself for use on one of the collection's computers. More recently, an Omeka-based online collection has been created to facilitate online access.[137]

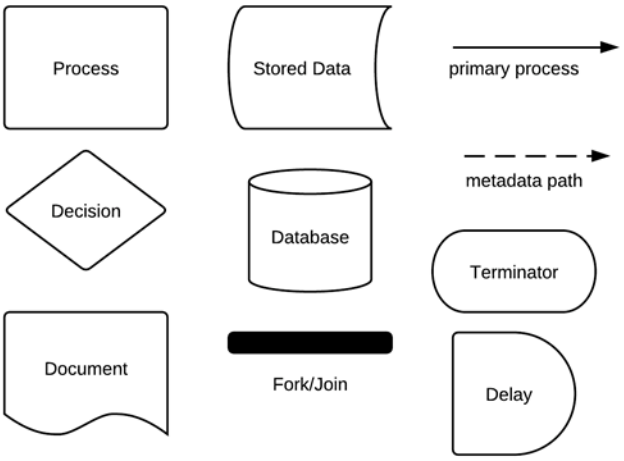
Subsequent work on MITH collections will be a collaborative process involving archivists in Special Collections at the University of Maryland Libraries, as well as the University of Maryland College of Information Studies. Kirschenbaum anticipates the development of finding aids using Encoded Archival Description (EAD) and other tools to facilitate scholarly use of the collections, as well as digital repository storage for born-digital content in other collections.[138] The workflow model developed for born-digital materials at the Maryland Institute for Technology in the Humanities is provided below.

[137]"The Deena Larsen Collection at the Maryland Institute for Technology in the Humanities," *Maryland Institute for Technology and the Humanities* http://mith.umd.edu/larsen/ (accessed July 2012).
[138]Interview, Matthew Kirschenbaum.

Figure 5. Maryland Institute for Technology in the Humanities, University of Maryland



Workflow Map Legend



National Library of Australia, Digital Preservation Section

The National Library of Australia (NLA) has existed since early after the Australian Federation in 1901, with the NLA established as an entity separate from the Parliamentary Library in 1960.[139] The Library operates a range of different collections, with each collection responsible for its own acquisitions and collection development.[140] Since the 1980s, across all collections there has been a slow accumulation of born-digital material, primarily contained on media carriers including floppy disks, CDs, and DVDs.[141] Initiatives in the 1990s and early 2000s assisted the NLA in identifying digital media in collections and assessing the level of risk inherent in the media present.[142]

The challenges faced by the NLA were considerable: media was accessioned without significant control over the incoming types and formats, and a decentralized system meant that any workflow for born-digital content required buy-in from many different stakeholders.[143] In 2008, the NLA implemented the Prometheus Digital Preservation Workbench, a "semi-automated, scalable process for transferring data from physical carriers to preservation digital mass storage."[144] The open-source tool, developed within the NLA, continues to be the centerpiece of workflows for managing born-digital content on removable media.

---

[139]"History of the Library," *National Library of Australia* http://www.nla.gov.au/history-of-the-library (accessed July 2012).

[140]"Our Collections," *National Library of Australia* http://www.nla.gov.au/our-collections (accessed July 2012).

[141]Interview with David Pearson, conducted May 23, 2012.

[142]Douglas Elford, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, Colin Webb, "Media matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia." Paper presented at the 74th IFLA General Conference and Council, Québec, Canada (August 10-14, 2008), 2-3, http://www.nla.gov.au/digital-preservation/related-staff-papers (accessed July 2012).

[143]Elford et al., "Media matters," 4-6.

[144]"Prometheus – About," *Prometheus Digital Preservation Workbench* http://mediapedia.nla.gov.au/prometheus/about.html (accessed July 2012). For more information on Prometheus, see Appendix D: Digital Preservation Tools and Technologies.

After the acquisition of removable media as part of an NLA collection, the item is catalogued according to the internal policies of that collection, which may include associating the item with a specific accession. Once identified by the collection for preservation, the media is attached to the Prometheus "mini-jukebox"—a portable bank of drives and open ports that can accept a variety of different media types, including CD/DVD, external hard drives, USB drives, and floppy disks, either 3.5-inch or 5.25-inch.[145] The web-based Prometheus workflow software then takes the processor of the collection through subsequent file and metadata extraction steps. In development, the digital preservation workflow had been imagined as a centralized service, however the use of the mini-jukebox enabled the majority of legacy formats to be managed in the course of regular cataloging processes. This decentralized the more straightforward steps in the capture of born-digital content, with members of the Digital Preservation section handling trouble-shooting and irregular formats.[146]

Within the Prometheus web interface, the processor can begin a new job and connect to an existing catalog record, associating the media and captured born-digital content with a collection. Prometheus assigns a persistent identifier for the media, and generates a disk image and a checksum against that image. Failed imaging notifies the Digital Preservation section, who can then step in to provide technical expertise. The disk image is copied to the processing workstation, and successful transfer is verified against the initial checksum. The image is mounted and unpacked, making the file system and files accessible for the processor to inspect and add descriptive metadata. A structure map is generated in a METS file, documenting all files in the file system, after which a virus

[145]Interview, David Pearson; and "Prometheus – Workflow," *Prometheus Digital Preservation Workbench* http://mediapedia.nla.gov.au/prometheus/workflow.html (accessed July 2012).
[146]Interview, David Pearson, and Elford et al., "Media Matters," 6.

scan runs to inspect for any malware.[147] Additional analysis through DROID and JHOVE provides information on file types, and generates additional file-level metadata with the New Zealand Metadata Extractor.[148] The processor can continue to image additional media associated with that catalog record in a single job. After all media have been imaged and analyzed, Prometheus provides the opportunity for quality assurance on the collection metadata, ensuring that all associated information is correctly entered.

After quality assurance, the processor can finish the job and ingest the extracted born-digital content into the Digital Object Storage System (DOSS) Repository. The DOSS acquires both the disk image and the extracted file system and files, so DOSS actually maintains two copies of the content. The media can be returned to the collection with a sticker denoting that it has undergone a successful capture. In order to access the content, NLA staff locate the desired material by its persistent identifier. Accessible content files are copied to a networked computer station in the NLA reading room for use. The workflow model developed for born-digital materials at the National Library of Australia, Digital Preservation Section, is provided below.[149]

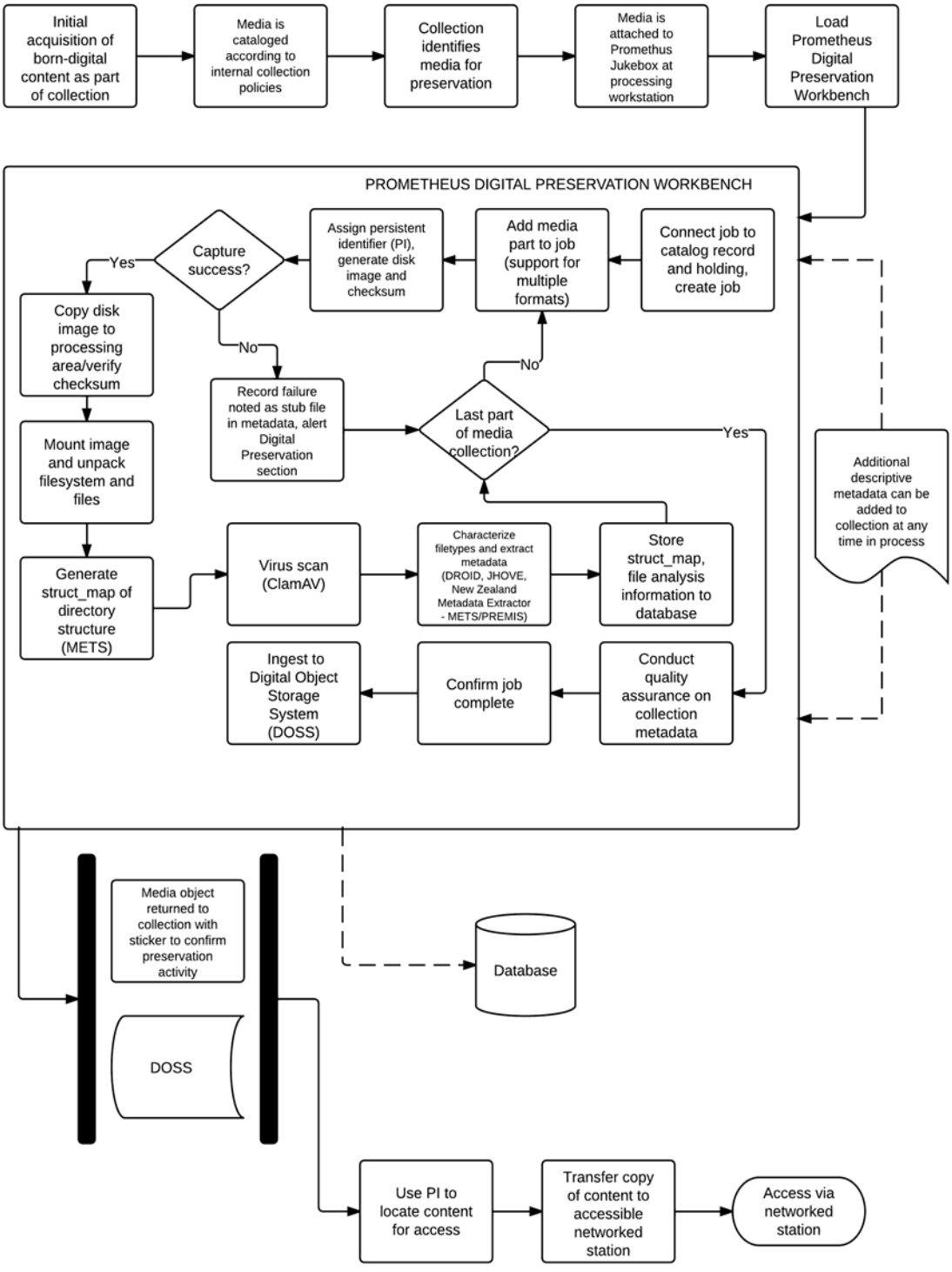
---

[147] For more information on METS and the ClamAV virus checker, see Appendix D: Digital Preservation Tools and Technologies.

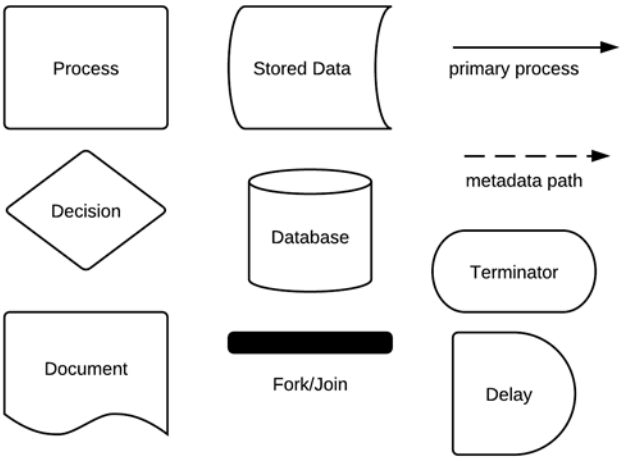
[148] For more information on these tools, see Appendix D: Digital Preservation Tools and Technologies.

[149] Workflow steps for Prometheus developed in consultation with "Prometheus Capture Walkthrough," http://prometheus-digi.sourceforge.net/prometheus-capture-walkthrough.html; and "Digital Collecting and Prometheus: Update 2012," http://mediapedia.nla.gov.au/prometheus/pdfs/Digital%20collecting%20and%20Prometheus%20Update-2012_web.pdf See also: Elford, et al., "Media matters."

Figure 6. National Library of Australia



Workflow Map Legend



University Archives, University of North Carolina at Chapel Hill

University Archives and Records Management at the University of North Carolina were not recognized as a separate entity within the University until 1978, though they possess material dating to 1789.[150] Over time the University Archives formed part of the core collections of archival material within Wilson Special Collections Library at UNC-Chapel Hill, with the Southern Historical Collection, the North Carolina Collection, the Rare Book Collection, and the Southern Folklife Collection. Although there have been previous attempts to develop practices for working with born digital content in the individual collections, there had not been an effort to devise a Wilson-wide digital preservation workflow until Erin O'Meara was hired as Electronic Records Archivist in 2009.[151] O'Meara agreed to provide descriptions of born-digital content workflows that she assisted in implementing through early 2012, when she left University Archives.[152]

Born-digital content workflows in University Archives call for curators to involve the Electronic Records Archivist early in the acquisition process to determine the best methods for capturing and transferring born-digital content included in newly acquired collections. Curators conduct a survey of material during their initial appraisal of the collection; at this time they also document any existing born-digital content in an electronic records appraisal form maintained with collection documentation. Basic technical and logical information including the extent, size, file formats and hardware

---

[150]"About the University Archives," *University Archives and Records Management Services at The Wilson Library* http://www.lib.unc.edu/mss/uars/uabout.html (accessed July 2012).

[151]Previous efforts at developing born-digital workflows are documented in John A. Blythe, "Digital Dixie: Processing Born Digital Materials in the Southern Historical Collection." A Master's paper for the M.S. in Information Science degree at the University of North Carolina at Chapel Hill. Advisor: Katherine M. Wisser (July 2009), http://ils.unc.edu/MSpapers/3541.pdf (accessed July 2012).

[152]Interview with Erin O'Meara, conducted May 9, 2012. O'Meara is currently Archivist at the Gates Family Archive. Also see Erin O'Meara and Meg Tuomala, "Finding Balance Between Archival Principles and Real-Life Practices in an Institutional Repository," *Archivaria* 73 (2012), 81-103, http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13385 (accessed August 2012).

represented in the acquisition is entered into a Media Log maintained in Microsoft Excel.

The steps taken to capture born-digital material are dependent upon the type and format of the digital media carrier that has been acquired. For hard drives, thumb drives, or floppy disks, a forensic bridge (a type of physical write-blocker) is used to ensure the integrity of the content on the media before a disk image is created.[153] An image is created using FTK Imager and transferred to local storage for quarantine as a precaution against malware. In cases when born-digital content entrusted to University Archives arrives via email, the files are transferred directly to the quarantine storage. In some cases virus checks are deemed unnecessary, owing to the nature of the material being processed, but in most cases a virus check is conducted.

At this point, the digital content—either disk image or, in the case of emailed acquisitions, the content files—is transferred to networked storage. Curators and processors may then appraise the born-digital content in a stabilized environment, and search for sensitive or unwanted information—such as personally identifiable information or duplicate files. This appraisal opportunity allows curators to determine a priority level for ingesting the content into the Carolina Digital Repository (CDR). If the decision is made to remove or redact any content, following these changes a secondary image consisting of the desired material is created, while maintaining appropriate documentation on actions of appraisal and selection. Prior to ingest, a disk image checksum is created, and individual files are pulled from the disk image for CDR ingest.

Content is ingested into the CDR via the Curator's Workbench, an open-source tool developed at the University of North Carolina at Chapel Hill. The Curator's

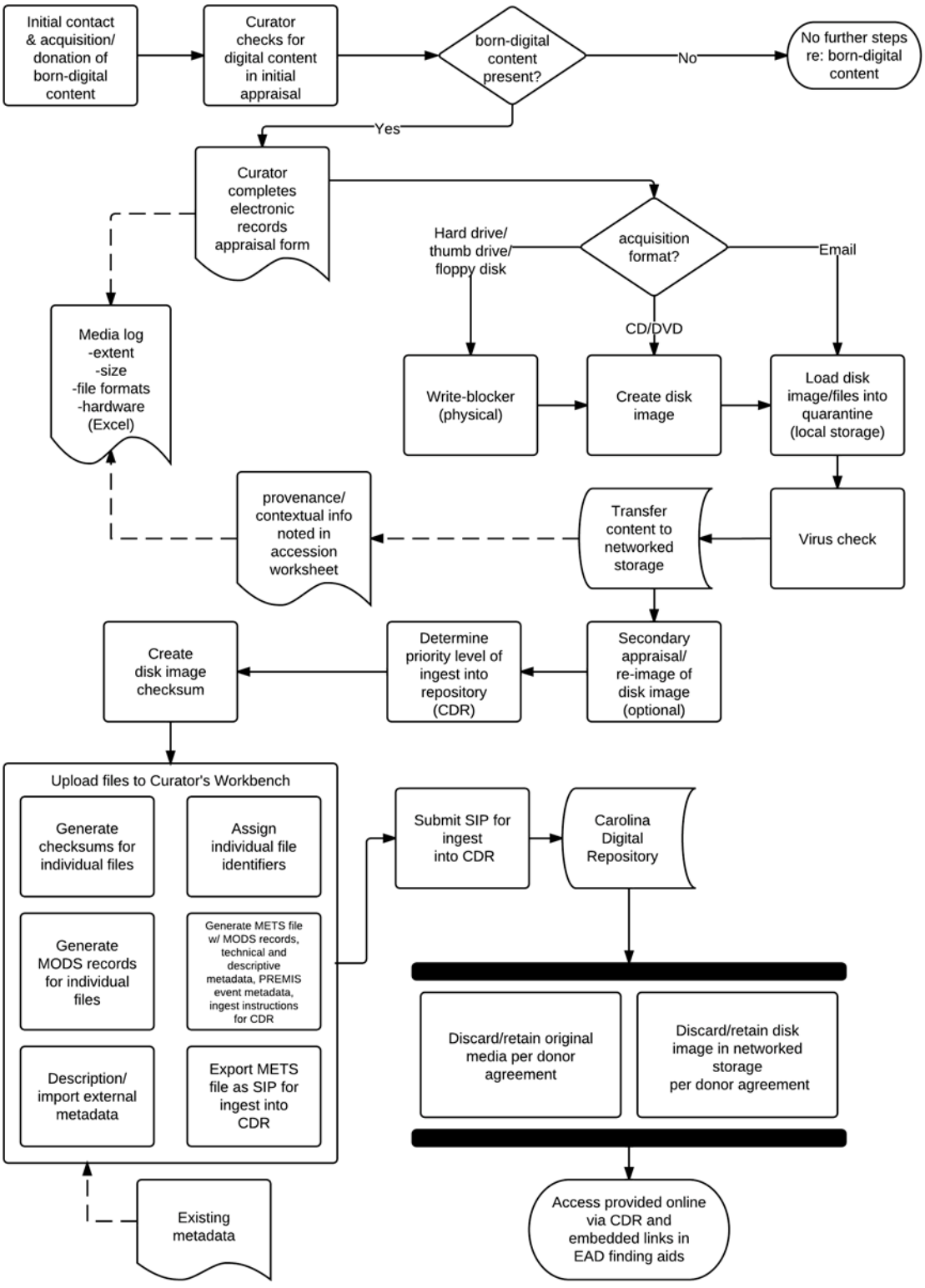
[153]Information on write-blockers can be found in Appendix D. Write-blocking is not necessary for optical discs such as CDs and DVDs, as they are primarily read-only digital media carriers.

workbench automates the extraction of file-level metadata, generates checksums for individual file objects, assigns identifiers, maintains PREMIS event metadata, and uses the Metadata Object Description Schema (MODS) to characterize file objects within a Metadata Encoding and Transmission Standard (METS) file wrapper.[154] Users can input existing metadata in the form of a delimited text file, and using a graphical metadata crosswalk, map the supplied metadata fields to MODS metadata elements, facilitating batch processing for large collections of digital objects. The Curator's Workbench outputs a METS XML file that functions as the Submission Information Package (SIP) for ingest into the CDR.

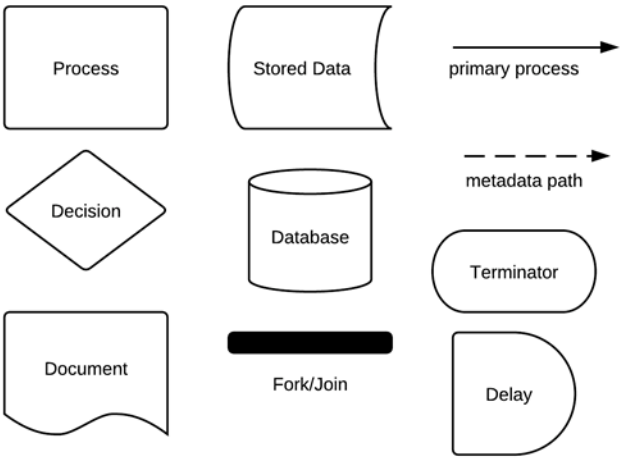
Following ingest of born-digital content into the CDR, disposition of the original media and disk image is carried out according to the terms agreed upon in the initial donor agreement. The disk image may be retained exterior to the CDR in network storage or deleted, while the original media may also be retained or destroyed. Much of the born-digital content that has been acquired and ingested into the CDR is not yet publicly available, meaning that the repository functions primarily as a dark archive. However, some access to digital content in the CDR is provided online, through embedded links within Encoded Archival Description (EAD) finding aids. The workflow model developed for born-digital materials at the University Archives, University of North Carolina, is provided below.

[154]Information on METS and MODS can be found in Appendix D.

Figure 7. University Archives, University of North Carolina at Chapel Hill



Workflow Map Legend



University of Virginia Libraries

The Albert and Shirley Small Special Collections Library is the primary repository of archival and manuscript material in the University of Virginia (UVa) Libraries. Collections include more than 13 million manuscripts, as well as 3.6 million items in the University Archives, in addition to numerous holdings in maps, broadsides, photographs, microfilm, and audio-visual materials.[155] UVa also maintains a growing quantity of digitized analog content, as well as born-digital material that has been acquired through collecting activities. In 2009, UVa took a major step forward in their management of born-digital collections through participation in *AIMS Born-Digital Collections*, a collaborative research project funded by the Andrew W. Mellon Foundation, involving Yale University Libraries, Stanford University Libraries and Academic Resources, and the University of Hull Library.[156] The AIMS project "was developed to define good practice in terms of archival tasks and objectives necessary for success," through a framework that could be flexibly applied to a number of different institutional contexts.[157] The project ended in early 2012 with the publication of a final report, *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*.[158]

Prior to their participation in the AIMS project, UVa had not taken significant action to gain control of born-digital content within their collections.[159] UVa is now building from the lessons learned in AIMS to develop a program for managing born-

---

[155] "About Small Special Collections Library," *Albert and Shirley Small Special Collections Library, University of Virginia Library* http://www2.lib.virginia.edu/small/about/about_small_lib.html(accessed July 2012). Additional information about the Small Special Collections Library can be found in AIMS Work Group, "Appendix D. Institutional Summaries and Collection Descriptions," *AIMS Born-Digital Collections*, 91-92.
[156] AIMS Work Group, "Acknowledgments," *AIMS Born-Digital Collections.*
[157] AIMS Work Group, "Forward," *AIMS Born-Digital Collections*, ii-iii.
[158]Ibid.
[159]Interview with Bradley Daigle, conducted May 22, 2012.

digital content. UVa staff have undertaken a comprehensive inventory of born-digital content in their collections, and focused on developing workflows to stabilize and preserve born-digital material, while exploring options for the provision of access at a later date. To learn about born-digital content management at UVa, I interviewed Bradley Daigle, Director of Digital Curation Services and Digital Strategist for Special Collections. Like other participants, he stressed the evolving nature of the workflows in development at UVa.

Born-digital content is identified after the donation or acquisition of a collection, or located within an existing collection. Individual media receive identifiers connecting them to a specific collection and accession, and logical and technical metadata is collected on the type, format, and condition of digital media carriers. This information is maintained in a media control record database. The digital media carrier is photographed and loaded into a digital forensic workstation, the Forensic Recovery of Evidence Device (FRED) from Digital Intelligence.[160] The FRED combines a number of forensic acquisition features, including write-blocking, the automation of some forensic processes, and powerful analysis through the use of Forensic Toolkit (FTK). Through the FRED, the processor can create a disk images of media, which go into quarantine to ensure there is no presence of malware. Disk images are then loaded into FTK for de-duplication, and to mine for erased data or partial files that may exist on the disk image. FTK is also used to search for personally identifiable information (PII) that may require redaction, and a virus scan verifies that the image is free of malware. Finally, file formats are normalized into preservation standards, and FTK is used to generate a preservation copy of the disk image.

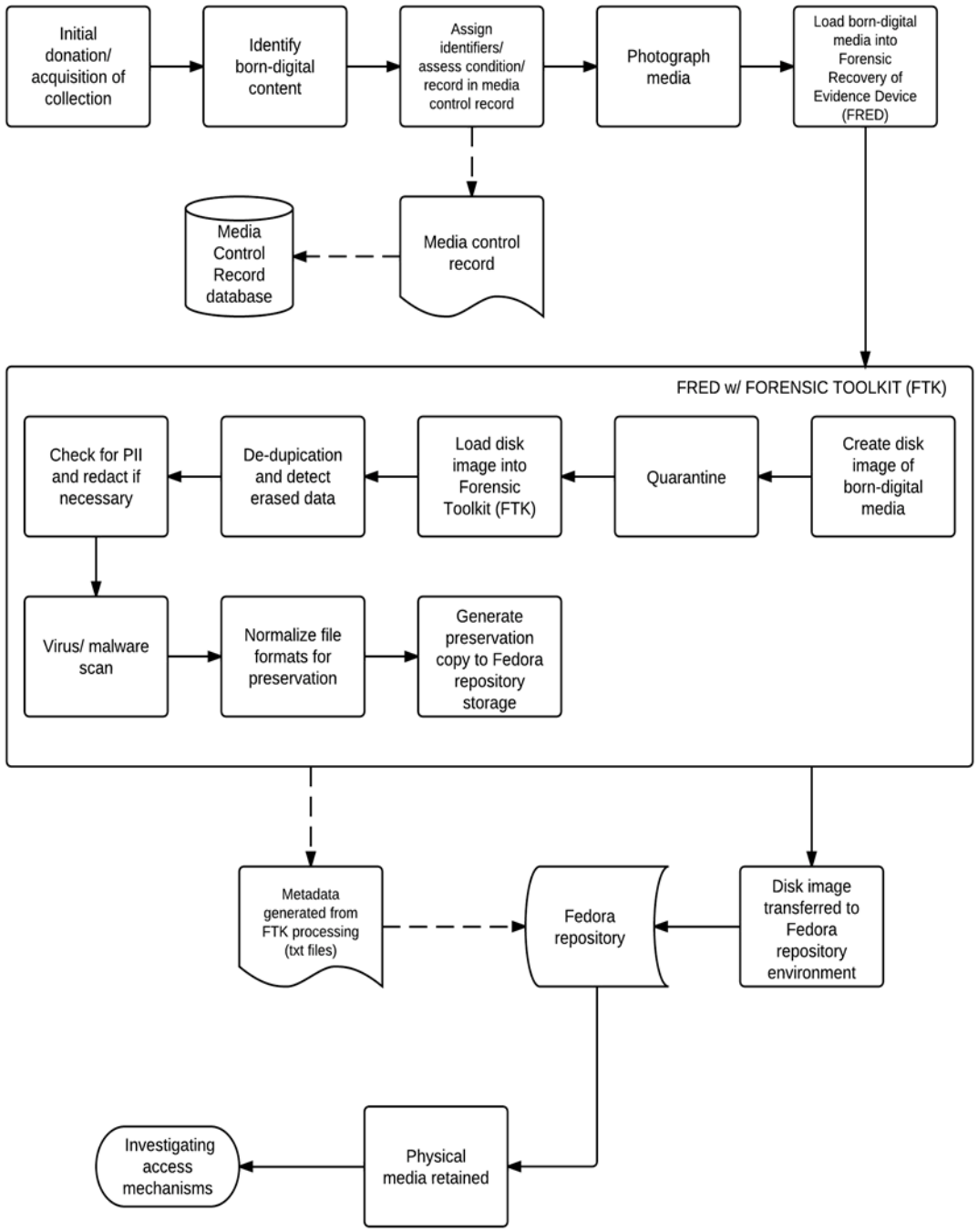
[160]For information on the FRED, see Appendix D, "Forensic Recovery of Evidence Device (FRED)."

This preservation disk image, along with all the metadata generated through the forensic processing with FTK on the FRED station, are associated in a Fedora-based preservation repository. Fedora can oversee many of the functions of archival storage and data management, and is designed to support additional functionality through Application Programming Interfaces (APIs).[161] While the physical media is maintained at present, UVa is still developing policies for the ultimate disposition of original media and disk images not intended for preservation. Similarly, they are exploring mechanisms for providing access at different levels of authorization. The workflow map developed for born-digital materials at the University of Virginia is provided below.[162]

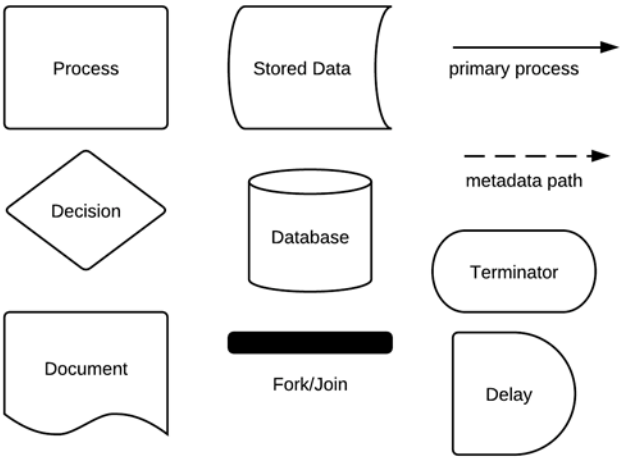
[161]For more information on Fedora, see Appendix D, "Fedora Repository Software."

[162]In addition to the interview conducted with Bradley Daigle, the workflow map includes information found in Gretchen Gueguen, "Born-Digital @UVa: The AIMS Framework Approach to Born-Digital Archives at UVa," Presentation to the Mid-Atlantic Regional Archives Conference (MARAC) 2012 Spring Conference (Cape May, NJ, April 13, 2012), gretchengueguen.com/professional/GueguenMARACs12.ppt (accessed June 2012).

Figure 8. University of Virginia Libraries



Workflow Map Legend



Library of Congress Manuscript Division

The Manuscript Division at the Library of Congress was established in 1897, and became the repository for personal papers acquired by the federal government following a 1903 executive order. Today its holdings include over sixty million items in more than eleven thousand collections.[163] In addition to this vast amount of physical material, the Manuscript Division has been receiving born-digital materials for well over twenty years, generally within hybrid collections.[164] It is estimated that about 90 different collections currently in the Manuscript Division include born-digital materials, contained in a wide variety of different media: "CPUs, 5 1/4-inch disks, 3 1/2-inch disks, flash drives, CDs, DVDs, Zip [disks], Jazz [disks], external USB drives, and hard drives."[165] It is only in the past 2-3 years that the Manuscript Division has begun to develop standardized workflows for capturing and preserving born-digital content, both in its accessioned collections and in new acquisitions. Because of time constraints it was not possible to include a workflow from the Manuscripts Division, Library of Congress; however, a recent blog post on the Library of Congress Digital Preservation blog, *The Signal*, provides insight into existing practices for imaging optical media.[166]

---

[163]"About the Manuscript Division," *Manuscript Reading Room, Library of Congress* http://www.loc.gov/rr/mss/mss_abt.html (accessed July 2012).

[164]Interview with Leslie Johnston, conducted May 18, 2012. Though the Manuscript Division is not the sole collecting unit of the Library of Congress, this interview focused on their activities to emphasize specific practices. The Library of Congress currently estimates the total size of its digital holdings to be somewhere around 3.5 petabytes of information, with a much greater total storage capacity. See Leslie Johnston, "A 'Library of Congress' Worth of Data: It's All In How You Define It," April 25th, 2012, *The Signal: Digital Preservation* http://blogs.loc.gov/digitalpreservation/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/ (accessed July 2012).

[165]Interview, Leslie Johnston, conducted May 18, 2012.

[166]Butch Lazorchak, "Rescuing the Tangible from the Intangible," July 2, 2012, *The Signal: Digital Preservation* http://blogs.loc.gov/digitalpreservation/2012/07/rescuing-the-tangible-from-the-intangible/ (accessed July 2012).

Discussion

The nine participants interviewed for this project provided a wealth of information about workflows for managing born-digital content that have been developed within their institutions. They also readily discussed challenges and successes they had experienced in developing the current processes, as well as additional work they would like to undertake. The following section addresses the workflow maps that were developed, as well as common threads in the comments and experiences described by project participants through the interview process.

Every born-digital content workflow must start somewhere.

Of the nine interviews conducted, six participants indicated that previous efforts to manage born-digital content had been ad hoc, project based, or otherwise not a sustained part of institutional practice.[167] For some, there had been no concerted effort to manage born-digital content prior to their arrival at the institution.[168] When asked to describe early practices for managing born-digital content, many responded with humor or sympathetic responses that indicated the state of digital collecting activities:

- "When I first got there, people were just trying not to accept born-digital materials. People were kind of scared about taking stuff in."[169]
- "We have certainly been getting digital materials as part of our collections for probably 20 years, like most other organizations. And what we did with those was mostly ignore them, to be completely honest. We made sure that the

[167]Sometimes referred to as "boutique" projects. These comments were generally made in response to Question 6 from the interview protocol which related to prior born-digital preservation workflows (see Appendix C: Interview Protocol). This may have been difficult for participants to answer, as employees are not always fully aware of the situation prior to their arrival at a collecting institution. Interview A; Interview B; Interview C; Interview D; Interview F; Interview G, Interview H, Interview J.
[168]Interview C.
[169]Interview H.

media was housed in proper housing, in acid-free folders, […] but we didn't start actively doing anything with the media we had until 2-3 years ago."[170]

- "We knew at the time that we had digital media that had born-digital content, but we weren't actively, or even aggressively, or even passively, really, looking for born-digital materials at that point."[171]
- About half the participants reported that they had been brought in specifically to take on the responsibility of managing some aspect of the institution's digital content. [172] For those who were not hired for this purpose, a position that evolved to encompass those responsibilities; others were brought in for other reasons, or did not indicate a response.[173]

A workflow is never complete or finalized.

All participants expressed that their workflows were under revision, incomplete, or still being tested in some way.[174] Some participants indicated they were pleased with the present iteration of the workflow, and the changes reflected tinkering to achieve the most successful strategy for working with born-digital content: "The way we've developed our accessioning process in particular is pretty iterative. We've done a fair amount of work thinking about how we want it to work, but we're still refining the process."[175] Many participants described workflow creation as an iterative process, gauging not when a workflow was finished, but when it was usable for their institution, with the understanding that it would continue to develop.[176] For others workflows are still in an early stage, as in the case of the participant who stated: "I wouldn't say that we have

[170]Interview D.
[171]Interview C.
[172]Interview B; Interview E; Interview G; Interview H; Interview J.
[173]Interview A; Interview C; Interview D; Interview F.
[174]Interview A; Interview B; Interview C; Interview D; Interview E; Interview F; Interview G; Interview H; Interview J. Note that there is no Interview I.
[175]Interview G.
[176]Interview B; Interview D; Interview E; Interview J.

a digital archives program, because it's still very much loosely defined. We don't have a ton of policies."[177] A documented, robust born-digital content workflow signals a well-formed overall digital archives program in the collecting institution.

In fact, there is a recognition that the present workflows will have to change, in order to enable the preservation of digital content for the long-term. One participant noted that "we're doing things this way now, but we know that it's not the right way, and let's just wait and see what happens, because maybe they'll come up with something better."[178] Another archivist, discussing the use of modular preservation tools, noted a benefit in their ability to accommodate change: "where you can define the start and stop points for each of your services, the better position your successors will be in to analyze what's working and what's not, and what you can supersede with other things."[179] This sentiment echoes that of the AIMS project, which listed the iterative nature of project member workflows among its findings: "workflows would have to be iterative both within one archival function and between functions."[180] The development of standardized, documented generic workflows for capturing and preserving born-digital content helps to minimize unforeseen changes that force archivists to be reactive, rather than pro-active.[181]

---

[177]Interview G.
[178]Interview E.
[179]Interview C.
[180]See AIMS Project Group, "Foreward," *AIMS Born-digital Collections*, vii.
[181] The Paradigm final project report notes that organizations "must develop business rules for the treatment of data types and data formats so that digital preservation can be as consistent, predictable and economical as possible." Thomas, *Paradigm: a practical approach to the preservation of personal digital archives*, 26.

Workflows develop within a variety of interrelated contexts.

Rather than outline a series of hard-and-fast requirements, the *Paradigm Workbook* offers a variety of options and approaches for curators to choose from and weigh from their particular contexts. Concluding remarks recognize that "the approach will need to fit with the preferences, personality, and skills of the individual creating the archive."[182] This can be extended to recognize that there are a variety of factors that affect the implementation of a workflow for managing born-digital content.

Interviews reveal that a digital preservation workflow is a moving target, constantly undergoing re-evaluation in light of new tools and technologies, changes in institutional policy, funding fluctuations, staff turnover and reorganization, and the publication of new findings that affect best practices for managing born-digital content. Technological factors in the implementation of participant born-digital content workflow included an institution-wide software upgrade and an impending library infrastructure program.[183] Larger institutions may adjust to existing needs with more caution and deliberation: a participant reflected that "there is quite a lengthy lead time, from analyzing and evaluating software, being able to make a justification, and being able to run it in our environment.[...] It's complex, and it can be lengthy."[184] Organizational shifts, such as a library departmental consolidation or reorganization may also play a disruptive role in the continuity of born-digital workflows.[185] A workflow is always changing, as is the environment in which it is implemented.

Ultimately this suggests that there is no single structure or arrangement to support

---

[182]Paradigm Project, "Conclusions," *Workbook on Digital Private Papers*.
[183]Interview J; Interview B.
[184]Interview D.
[185]Interview H.

a digital preservation program. Said one participant, "I don't think there's a single staffing model that we need to do this kind of work. It kind of depends where you have the staff in place and where the expertise comes from."[186] Born-digital content exists within an "ecosystem of responsibilities and projects," in which digital forensic practices may play a relatively small role.[187] The most suitable deployment of staff and resources within this ecosystem can take on a variety of different forms.

Collaboration is a key to implementing a born-digital content workflow.

Many of the participants raised the point that collaboration was an important factor in the implementation of their present workflow, and that continued collaboration was important to future development.[188] This collaboration occurred within the institution, as archivists reached out to curators of collections and archivists in other units in their organization; and across institutions, as archivists participated in multi-party grant-funded projects to develop a better understanding of born-digital content management.

Interviews agreed that collaboration within the collecting institutions was the best way to implement a workflow across a variety of different units acting under different curators, with different collecting focuses and internal practices: "accessioning practices are so idiosyncratic that it's hard to say, 'Ok, this is going to work for everybody.'"[189] In order to accommodate these idiosyncrasies, archives may adopt a centralized approach to born-digital content management, where content from multiple collections comes through

[186]Interview A.
[187]Interview J.
[188]Interview A; Interview B; Interview C; Interview F; Interview G; Interview H.
[189]Interview C.

a central pipeline: "I was designated the born-digital person, but my strategy was to incorporate people within all the different activities […] I was the hub, and then I tried to start developing spokes."[190] An alternative is to develop an approach in which relatively basic forensic processes are spread out and made the responsibility of each collecting unit: "What we've decided to do is decentralize the process so that catalogers can deal with it. Part of the reason why is that they know what the material is better than we know what the material is."[191] Both approaches have strengthened communication across traditional organizational lines of authority. As one participant put it, "we don't have an obligation to work together, but we've certainly found that it's easier for us to work together."[192]

Across institutions, collaborative projects with a strong digital forensics component have led to better sharing of information and experiences, pooling time and resources to tackle issues of concern to all collecting institutions.[193] Several participants suggested that external collaboration further stimulated their collaborative activities at home.[194] In addition, the variety of differing viewpoints offered through such projects can provide new perspectives on thorny issues within their own institutions, as demonstrated in Matthew G. Kirschenbaum's perspective on the participation of MITH (Maryland Institute for Technology in the Humanities) in the BitCurator Project: "We're probably a little closer to the needs of the future scholars who will be the patrons for these kinds of collections, and we might have more of an on-the-ground sense of what scholars are

---

[190]Interview H.
[191]Interview B.
[192]Interview G.
[193]Examples of include the recently concluded *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*, The National Endowment for the Humanities-funded *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*, and the *BitCurator Project*.
[194]Interview A; Interview C; Interview F; Interview G. Said one participant: "In particular, I think that a big part of the reason we've been able to work together […] is that we were part of the AIMS Project."

going to want to be doing with these materials."[195] Collaboration—internal, external, and interdisciplinary—has thus been vital not only to the development of born-digital content workflows within institutions, but also as a driver for research into a common base of best practices in the field at large, and in anticipation of the potential needs of users of digital content.

The creation of a forensic disk image is central to digital forensic workflows.

While there is a great deal of variety in the implementation of born-digital content workflows discussed by participants, one common thread among workflows is the need to create a forensic disk images of a digital media carrier, containing a bit-level copy of the media's contents.[196] Participants cited the importance of stabilizing the image and separating it from the medium upon which it arrived, maintaining an "archival snap shot" of the content to be ingested into preservation storage.[197] Other common features include the use of checksums to verify the integrity of imaged media, as well as any files that may be generated from the image. Institutions may choose to maintain multiple copies of a disk image, with one remaining isolated as a preservation master and another used to generate access copies for researchers.[198]

There are also variations within the practice of generating disk images. If participants specified the file format his/her institution was using to create a disk image,

---

[195]Interview, Matthew Kirschenbaum.
[196]Interviews A-J.
[197]Interview G; Interview C; Interview F. For eight out of the nine participants consulted, some type of write-blocker was also engaged: either through the use of a physical product; the use of incompatible file/operating systems which made interaction between the media and imaging stations ineffective; or physically switching the "tab" on floppy disks, enabling write-protection. Write-blocking is unnecessary when imaging CDs and DVDs.
[198]Interview A. This is also an early strategy for working with born-digital content advocated in Goldman, "Bridging the Gap," 17.

it was generally to note that they were creating raw forensic images—uncompressed files that are the same size as the media from which they are imaged.[199] Three participants discussed using or considering the use of other forensic formats, either proprietary formats for use with powerful forensic software packages, or open-source forensic formats such as the AFF (Advanced Forensic Format), citing their smaller storage footprint and enhanced ability to associate metadata with the disk image.[200]

Though there was widespread agreement among participants on the use of disk images for the capture and forensic processing of born-digital content, participants also were willing to explore options that did not involve the use of disk images. One participant described a scenario in which donors email individual files to the collecting institution; in such cases a disk image is not necessarily an appropriate step, as much of the metadata intended to be preserved by a disk image had already been lost at that point.[201] Another participant argued that the decision to create a disk image needs to be made within the context of a larger institutional acquisition policy; that creating and maintaining a disk image is essentially an appraisal decision:

> in your acquisition strategy as a repository, you should have some sort of tiered method of deciding how much do we trust this acquisition, this recordkeeping system, this method of transfer, and how important is it to this particular records set that we copy it, or image it, or whatever."[202]

Another participant made a similar suggestion indirectly, acknowledging that although they create disk images in their institution, it is not the disk image that is ultimately

---

[199] Interview A; Interview B; Interview F; Interview G. A "raw" disk image could mean one of several different formats, depending upon the tool used to create the disk image. See "Disk image," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Disk_image (accessed July 2012).

[200] Interview E, Interview G, Interview J. In interview G, the participant noted that "So far we've only been using raw images. I'd like to consider using AFF. The biggest thing right now would be the compression, but that's certainly not the end of the world, given that we're mostly accessioning media and not hard drives, so it hasn't been a huge priority."

[201] Interview H.

[202] Interview E.

preserved but the files from the disk image that are found to have enduring value through appraisal.[203] Disk imaging is thus a practice accepted by participants, but also in need of further examination as digital forensic tools advance and become more widely used within the digital preservation community.

Born-digital content workflows that implement digital forensic tools and processes can capitalize on modularity and increase workflow automation.

Most participants expressed at least a desire for, if not active movement toward, a more automated born-digital content workflow.[204] Participants described their workflows as unfinished, and many participants expressed a desire to eventually reach a solution based in the automated digital forensic extraction of metadata necessary for preservation. However, this assertion must be made with a caveat: while automation is a goal, participants also recognized that the state of available technology makes that goal difficult to achieve.[205] One participant suggested that "[i]t's increasingly apparent that there's no single piece of software or hardware that solves all of your needs. It's more a matter of finding the right suite of tools that allows you to do what you need to do, and what can streamline your process without it being too complicated."[206] This perspective is

---

[203]Interview H.

[204]Interview A-J. A representative example of this view is articulated by the participant in Interview E: "We also know that we need to automate more of our processes, and that's something we will do. But we needed to do the proof of concept that this workflow is the right one, and now we can start looking toward automating processes."

[205]Interview D. There are of course exceptions to this; for examples of tools in active development that are geared toward automating digital forensic processes, see "Archivematica," "BitCurator", and the "Prometheus Digital Preservation Workbench," in Appendix D. The *BitCurator Project* website notes that while there do exist Linux-based software packages for conducting forensic analysis of digital content, "they are not very approachable to library/archives professionals in terms of interface and documentation." "About | Bitcurator," *BitCurator* http://www.bitcurator.net/aboutbc/ (accessed August 2012).

[206]Interview A. Another example from Interview C: "I'm really doubtful that there's going to be one piece of software that is going to be able to manage every piece of born-digital media or content that has ever

also articulated by Jeremy Leighton John, who notes that "[t]here is no single product that will meet all requirements of the forensic examiner or for that matter the digital curator or preservation expert, which explains why there is a flourishing diversity of specialist products."[207]

The central premise, one participant explained, is that "in all cases, but especially in cases where you have something that is highly complex, with many moving parts, [such as] workflows, it's that the more modular, the better."[208] Thus some participants described a pragmatic approach: breaking the workflow into modular component parts for execution by paraprofessionals, volunteers, or student workers.[209] The adoption of digital forensics tools that can generate specific metadata helps to facilitate these efforts, as does it enable the alteration of the workflow in response to shifts in technology if necessary.[210] Having student workers or volunteers to track spreadsheets of metadata, photograph media carriers, and execute modular digital preservation tools also has the benefit of freeing the archivist to do other work, as one participant confessed: "because assigning a digital archivist to that kind of work is not cost-effective."[211]

This last point illustrates a hurdle to implementation of forensic workflows for born-digital content. Developing a workflow requires awareness of available tools and best practices for digital preservation, and at least some technical knowledge in order to use those tools. Many of the most powerful open-source digital forensic tools use a

---

been or will ever be created."

[207] John, "Adapting Existing Technologies for Digitally Archiving Personal Lives," 2.

[208] Interview C.

[209] Interview C; Interview D; Interview E; Interview F; Interview G; Interview J.

[210] Interview C noted that developing born-digital content workflows draws on similar themes and concepts as developing a digitization workflow: "[Just like] in any digitization workflow, you're going to have content, you know, the descriptive metadata is going to be created by human, and a lot of the technical/administrative [metadata] is going to be created by machine."

[211] Interview E. Or as suggested in Interview J, "because it's menial."

command-line interface, for example.[212] Once the workflow has been implemented with sufficient documentation, however, the work would ideally be redistributed in a more cost-effective fashion. Even then a born-digital content workflow can be difficult to carry out: tracking metadata in spreadsheets and databases and piecing together information from a variety of tools can potentially add to the difficulty in packaging and storing metadata in a digital preservation workflow. One participant noted that "[p]art of the reason why it's been hard to get it to a point to where we can hand it off is that all of the software that we use for imaging in particular is so different, and there's enough that could potentially go wrong."[213] Another confessed that they did not photograph digital media that had been collected as it presented an issue of added complexity for the hired paraprofessional technician: "rather than add a step and another digital object to our process that is already quite laborious, he's just typing metadata and using the item number in the title and tracking in the spreadsheet."[214] One participant gave this description as a more cyclical view of the process of workflow development:

> You see this development of the program where you have a methodology and once it changes it constricts on that one individual who created the new workflow, and they're able to train others to be able to fill that responsibility as well. But then if you make a change again, then it shrinks back to that individual with their responsibility until they can provide additional training. And you have this expansion and contraction of responsibility over time as you develop new processes.[215]

Thus the iterative development of born-digital workflows utilizing digital forensic tools can potentially constrain the digital archivist, leaving them unable to address other issues and responsibilities—and with the limited resources commanded by many collecting

---

[212]See "fiwalk," in Appendix D. Software developed through *BitCurator* is intended make digital forensics tools with user-friendly interfaces. See "BitCurator," also in Appendix D.
[213]Interview G.
[214]Interview E.
[215]Interview J.

institutions, there are always other issues and responsibilities.[216]

Many digital archivists are struggling to gain recognition as an integral part of their institutions, and even when recognized archivists are still struggling to gain support—primarily financial—from higher levels of institutional administration.

While undoubtedly a thorny issue, several participants expressed frustration at the difficulty of gaining traction and significant institutional support for the preservation of born-digital content.[217] "I think any digital archivist is a change agent," reflected one participant.[218] However, participants argued that change must be recognized and supported at higher levels of the institution, by curators and administrative bodies, as well as supported with education and training at other levels of the institution.[219] Said one participant: "Administrators, whose buy-in is required to fund and scale this, […] they don't understand the content, they don't understand the complexities of the content, they don't understand how and why it is so different."[220] Several participants argued that lack of financial support and administrative openness to institutional change hindered progress in the successful management of born-digital content, either specific to their institution or more generally, as a problem experienced throughout the field.

Several participants suggested that part of the key to making born-digital content an institutional priority was to acquire collections with a significant born-digital

---

[216]In Interview A the participant noted, "There's a lot of work to get done, and this is one of many things that needs to get done."
[217]Interview A; Interview C; Interview G; Interview H.
[218]Interview H.
[219]The necessary systemic change was characterized during Interview E: if reference staff have not received adequate preparation in receiving born-digital materials, it can compound problems later on in the capture process: "It's generally the reference staff who are not digital archivists typically, and don't have any digital archives experience, who are being approached by donors about digital acquisitions, and just didn't know the questions to ask."
[220]Interview C. Similar views were also expressed in Interview A; Interview G; and Interview H.

component. "I think it's sort of been a struggle for some of the staff here that work on collection development to be able to recognize when there might be born-digital content of value," said one participant.[221] Other participants were more direct in their comments on the issue: "What really needs to happen is that curators need to fund this work, and they need to change their collecting and their approach to collecting, but without that I feel like you can't really enact change."[222] Several others noted that without adequate resources it was difficult to make significant progress: "collectively, certainly we're making an effort to come up with common best practices, but resources vary from one place to the next."[223] Lack of resources, compounded with lack of administrative support for institutional change, can be counterproductive to digital preservation efforts not only within the institution, but also within the wider field. "This is why we talk and talk and talk in the literature but nothing really happens, because no one has dedicated very specific resources, and no one has done it in a way to change the organization."[224]

There is still much work to be done around arrangement and description of born digital materials, as well as the provision of access to those materials.

Frustrations about institutional support for the collection and management of born-digital content are merited, in light of the future work that participants argue will be necessary to make such content accessible to users. I asked participants about challenges faced by digital archivists, both in implementing current workflows and in the overall

[221]Interview G.
[222]Interview H.
[223]Interview A. Also, Interview F: "I think the biggest problem would simply be one of resources."
[224]Interview H.

management of born-digital content.[225] These questions elicited a variety of responses. Some were specific to the participant's institutions, focusing on particular needs in technical infrastructure, digital preservation tools, and content management systems.[226] However, two particular overarching themes emerged from participant responses as more general challenges facing digital archivists: the arrangement and description of born-digital content, and the provision of access to born-digital materials.

In several interviews, participants argued that today's descriptive practices are not well-suited to the kind of information that is generated while processing digital collections.[227] Participants cited current descriptive standards and the way in which users and archivists are accustomed to interacting with collections, as well as the nature of born-digital content itself, as difficult to characterize through current practices. This was phrased as a difference of interpretation between those who actively deal with digital collections and the catalogers and processors who deal with more traditional collections:

> One of the major problems intellectually is the way [individuals who work in the] collections [unit] see the materials and the way we see the materials. So [individuals who work in the] collections [unit] see things in terms of intellectual entities and item-level description, series level descriptions, whatever. We see things in terms of files and formats. This is a very big problem, I think.[228]

Other participants agreed: "The traditional levels of description that archivists have been using for years are not appropriate for describing the very deep hierarchical structures we see in digital collections," stated one participant.[229] Born-digital collections can be overwhelmingly large, heavily arranged, and have complex internal relationships, thanks

---

[225]See Appendix C: Interview Protocol.
[226]Interview C; Interview D; Interview G.
[227]Interview B; Interview C; Interview E.
[228]Interview B.
[229]Interview E. At another point in the interview, the participant was more specific about how this affected workload: "Because I couldn't find a good directory printer at first, I was physically typing in what the directory structure looked like, and that's really tedious."

to constantly increasing storage, the ease of creating additional hierarchical layers in a digital file system structure, and the development of complex digital objects. Traditional processing—geared toward aggregate-level discovery—may not be able to sufficiently reflect some of these qualities of born-digital collections. This in turn affects usability, as suggested by one participant: "[b]y and large, archivists are still kind of hung up with the idea of EAD, or collection-level discovery, as the only way, or the best way, to make this content discoverable."[230]

Adequate description of born-digital content is tied to its accessibility and discoverability to users. Recent projects using digital forensic approaches have crafted unique, highly sophisticated mechanisms for online and on-site access.[231] However, it is likely that this level of individual attention will not be possible or desirable for all born-digital content. This makes the development of access methods for collecting institutions all the more important in the implementation of a workflow for born-digital content.

Providing access to born-digital content still poses a significant hurdle for many interview participants. In its current workflow for born-digital content, the University of Virginia has emphasized acquisition, stabilization, and preservation of born-digital content while it explores potential access solutions.[232] Six other interview participants indicated that their institution currently provides access to born-digital content on-site, at

---

[230]Interview C.

[231]Examples include *Salman Rushdie's Digital Life* at Emory University's Manuscript, Archives, and Rare Book Library (MARBL); available at http://marbl.library.emory.edu/innovations/salman-rushdie; the *Deena Larsen Collection* at the Maryland Institute for Technology in the Humanities at the University of Maryland, available at: http://mith.umd.edu/larsen/; and the *Michael Joyce Papers* at the Harry Ransom Center at the University of Texas at Austin. While *Salman Rushdie's Digital Life* provides on-site access to emulated Rushdie materials, there is an online browse function associated with the *Deena Larsen Collection*, and the *Michael Joyce Papers* are available onsite through UT Austin's DSpace repository service. These collections are discussed in more detail in Kirschenbaum, et al., *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use.* Also see Catherine Stollar Peters, "When Not All Papers are Paper: A Case Study in Digital Archivy," *Provenance* vol. 24 (2006): 23-35, https://pacer.ischool.utexas.edu/bitstream/2081/2226/1/023-035.pdf (accessed July 2012).

[232] Interview with Bradley Daigle.

a computer station in the reading room.[233] The remaining two have developed web access portals for publicly available digital content, though for each of these, not all content in their digital repository is publicly available.[234] Participants listed a number of issues in their institution's ability to provide access to born-digital content: copyright and intellectual property,[235] access management and personally identifiable information,[236] and the ability to integrate digital and physical holdings in the same access system.[237]

For several participants, the lack of a robust access mechanism was tied to the lack of requests for access to born-digital content.[238] For accessing born-digital content, one participant noted: "We do have a locked-down station within the reading room, but whenever someone requests it, we have to load that material on manually. […] that probably needs to change, but we haven't really gotten enough demand to force us to change that yet."[239] Thus progress can stall because of a cyclical effect: born-digital content is not collected, because that content is not accessed, because access mechanisms are not well-developed, because it is not a funding priority, and thus born-digital content

---

[233] The institutions are: Duke University, Yale University Manuscripts and Archives, Maryland Institute for Technology in the Humanities, National Library of Australia, Library of Congress Manuscripts Collection, and Beinecke Rare Book and Manuscript Library. Duke University does have an online access portal for University Publications through its DukeSpace repository (available at http://dukespace.lib.duke.edu/dspace/); and the Maryland Institute for Technology in the Humanities maintains a project website and online exhibition for the *Deena Larsen Collection*, created after the accession and processing of the collection (available at: http://mith.umd.edu/larsen/).

[234] Interviews, Erin O'Meara and Courtney Mumma. The University of North Carolina at Chapel Hill links digital content in the Carolina Digital Repository (CDR) to finding aids for online access when available and provides online search and browse at https://cdr.lib.unc.edu/, but much of the content in the CDR is otherwise currently inaccessible. The City of Vancouver Archives makes digital content available through an ICA-Atom online search portal, available at: http://searcharchives.vancouver.ca/. For more information about the ICA-Atom search portal, see: Heather Gordon, "Our New Online Search," April 11, 2012, *AuthentiCity: The City of Vancouver Archives Blog.* http://www.vancouverarchives.ca/2012/04/our-new-online-search/ (accessed July 2012).

[235] Interview C; Interview J.

[236] Interview C; Interview D.

[237] Interview B; Interview C; Interview El Interview F.

[238] Interview G. Another example is found in Kirschenbaum et al., *Approaches to Managing and Collecting Born-Digital Literary Materials*, 11. Staff at the Harry Ransom Center reported three requests for use of a born-digital collection of literary materials in two years.

[239] Interview G.

is not collected, and on and on.[240] While acknowledging the difficulties posed by born-digital arrangement and description, participants also recognized this as an ongoing transition as such materials become more prevalent in institutional holdings.[241]

Conclusions – "Sometimes we just have to try."[242]

This paper has explored born-digital content workflows that implement digital forensic tools and practices from nine different institutions, across three countries and two continents. Through workflow mapping, it has demonstrated the types of digital forensic tools used by practitioners to manage born-digital content in collecting institutions, and the placement of those tools within an institution's born-digital preservation workflow. By approaching these workflows through semi-structured interviews with digital preservation and digital curation professionals working in the field, the most prominent revelation is that of an area of scholarship and practical application that is still rapidly developing. As one participant put it, "you know, it has to be admitted that we're making this stuff up, and sometimes we just have to try, and nothing is going to be perfect."[243] The majority of the interview participants consulted for this work are practitioners who are implementing digital forensic tools in the field, and

---

[240] This is supported by the Blue Ribbon Task Force on Sustainable Preservation and Access, *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information.* Francine Berman and Brian Lavoie, co-chairs. (La Jolla, CA: Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010), 19, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (accessed July 2012). The task force notes that "the first challenge to preservation arises when demand is diffuse or weakly articulated."

[241] This also connected back to a more general transition of the arrangement and description of born-digital content. From Interview C: "It's just that the tools are going to vary, just like getting people to stop typing up their finding aids, giving them a client like Oxygen to create it in native XML. It's similar. You're going to have archivists who can make that leap from hand-written to Oxygen, you're going to have the same archivists who can make that leap from FTK or BitCurator to creating a digital object out of this material natively, with all of their ancestral relationships embedded in that process."

[242] Interview B.

[243] Interview B.

who have developed their knowledge through participation in research projects and rigorous engagement with both the scholarly and professional archival communities. "From a professional perspective, you have work you need to do. We need to do something, [or] you run some risk if you just let this stuff sit indefinitely. So you have this need to be good custodians of this material."[244] While this study has been exploratory, it is hoped that it can add to the body of scholarship on digital preservation programs in collecting institutions.

This study also points to some clear areas for further study, in the development of methods for arrangement and description of digital content; and for the provision of access to that content in born-digital collections applying forensic tools and practices. One interview participant noted: "I am a digital forensics person, not that I'm good at any of it, but I know it's the way to go for us, for archives in particular, and I know that we need a better understanding of it all."[245] With the growing body of knowledge on best practices provided by recent publications such as the *AIMS Born-Digital Collections* report, and ongoing work on digital forensic tools through the BitCurator Project, the use of digital forensic tools in digital curation and digital preservation is an area that has garnered significant interest. Participant responses also indicated that there are opportunities for study on the practical application of these tools: "I don't know of any institution that has a broadly publicized or scalable soup to nuts solution for born-digital. I know people who have strengths in certain areas, whether it's the accessioning or the delivery piece; but no one has soup to nuts, here is the whole workflow."[246] In mapping

[244]Interview A. An example of this experimentation is provided in Interview B: "When you're dealing with billions of files and you're dealing with batch processing, you're going to break stuff."
[245]Interview E.
[246]Interview C.

the current practices of a number of different types of archives and collecting institutions currently engaging born-digital preservation, this work has attempted to elucidate where some of those strengths and weaknesses can be found in existing practices.

References

"About ClamAV." *Clam Anti Virus.* http://www.clamav.net/lang/en/about/ (accessed July 2012).

"About the ICA-Atom Project." *ICA-Atom: Open Source Archival Description Software.* https://www.ica-atom.org/about/ (accessed July 2012).

"About Manuscripts and Archives: Introduction." *Manuscripts and Archives, Yale University Library.* http://www.library.yale.edu/mssa/about_intro.html (accessed July 2012).

"About the Manuscript Division." *Manuscript Reading Room, Library of Congress.* http://www.loc.gov/rr/mss/mss_abt.html (accessed July 2012).

"About | Maryland Institute for Technology in the Humanities." *Maryland Institute for Technology in the Humanities.* http://mith.umd.edu/about/ (accessed July 2012).

"About Small Special Collections Library." *Albert and Shirley Small Special Collections Library, University of Virginia Library.* http://www2.lib.virginia.edu/small/about/about_small_lib.html (accessed July 2012).

"About the University Archives." *University Archives and Records Management Services at The Wilson Library.* http://www.lib.unc.edu/mss/uars/uabout.html (accessed July 2012).

Abrams, Stephen, John Kunze, and David Loy. "An Emergent Micro-Services Approach to Digital Curation Infrastructure," *The International Journal of Digital Curation* 5, no. 1 (2010): 172-186. http://www.ijdc.net/index.php/ijdc/article/view/154 (accessed June 2012).

Abrams, Stephen. Sheila Morrissey, and Tom Cramer. "What? So What?": The Next-Generation JHOVE2 Architecture for Format-Aware Characterization." *Paper presented at the Fifth International Conference on Preservation of Digital Objects, British Library, 29-30 September 2008.* https://bytebucket.org/jhove2/main/wiki/documents/Abrams_a70_pdf.pdf (accessed July 2012).

"AFF." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/AFF (accessed July 2012).

AIMS Project Group. *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship.* January 2012. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf (accessed June 2012).

"Archivematica Main Page." *Archivematica.* https://www.archivematica.org/wiki/Main_Page (accessed June 2012).

"ArchivesSpace: building a next generation archives management tool." *ArchivesSpace.* http://www.archivesspace.org/ (accessed June 2012).

"Archivists' Toolkit | For Archivists By Archivists." *Archivists' Toolkit.*

http://www.archiviststoolkit.org/ (accessed June 2012).

"Archon: The Simple Archival Information System." *Archon.* http://www.archon.org/about.php (accessed June 2012).

Awre, Chris. "Fedora and digital preservation." *Tackling the Preservation Challenge. Presentation given December 12, 2008 at the University of Hull, UK.* http://www.dpconline.org/events/previous-events/426-practical-steps-for-repository-managers (accessed July 2012).

"BagIt." *Wikipedia – The Free Encyclopedia.* http://en.wikipedia.org/wiki/BagIt (accessed June 2012).

Bantin, Philip C. "Strategies for Managing Electronic Records: A New Archival Paradigm? An Affirmation of our Archival Traditions?" *Archival Issues* 23, no. 1 (1998):17-34.

Beagrie, Neil. "Plenty of Room at the Bottom? Personal Digital Libraries and Collections." *D-Lib Magazine* 11, no. 6 (June 2005). http://www.dlib.org/dlib/june05/beagrie/06beagrie.html (accessed August 2012).

Bearman, David. "An Indefensible Bastion: Archives as a Repository in the Digital Age." in David Bearman (ed.) *Archival Management of Electronic Records, Archives and Museum Informatics Technical Report* 13 (1991): 14-24.

Beebe, Nicole. "Digital Forensic Research: The Good, The Bad, and the Unaddressed." *Advances in Digital Forensics V*, Gilbert Peterson and Sujeet Shenoi, eds., International Federation for Information Processing 306 (2009): 17-36. http://dx.doi.org/10.1007/978-3-642-04155-6_2 (accessed August 2012).

"Beinecke Rare Book and Manuscript Library: About the Collections." *Yale University: Beineicke Rare Book and Manuscript Library.* http://www.library.yale.edu/beinecke/brblinfo/brblguide.html (accessed June 2012).

Berman, Francine, and Brian Lavoie, co-chairs, Blue Ribbon Task Force on Sustainable Preservation and Access. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information.* La Jolla, CA: Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (accessed July 2012).

"BitCurator Project," *BitCurator*. http://www.bitcurator.net/ (accessed August 2012).

"BitCurator Wiki." *The BitCurator Wiki.* http://wiki.bitcurator.net/index.php?title=Main_Page (accessed July 2012).

Blythe, John A. "Digital Dixie: Processing Born Digital Materials in the Southern Historical Collection." A Master's paper for the M.S. in Information Science degree at the University of North Carolina at Chapel Hill. Advisor: Katherine M. Wisser. July 2009. http://ils.unc.edu/MSpapers/3541.pdf (accessed July 2012).

Caplan, Priscilla. *Understanding PREMIS*. Washington, D.C.: Library of Congress Network Development and MARC Standards Office, 2009. http://www.loc.gov/standards/premis/understanding-premis.pdf (accessed July 2012).

Carrier, Brian. "Defining Digital Forensic Examination and Analysis Tools Using Abstraction Layers." *International Journal of Digital Evidence* 1, no. 4 (Winter 2003): 1-12. http://www.informatik.uni-trier.de/~ley/db/journals/ijde/ijde1.html (accessed July 2012).

Carrier, Brian. "Description," *Autopsy Forensic Browser.* http://www.sleuthkit.org/autopsy/desc.php/index.php (accessed July 2012).

Carrier, Brian. *File System Forensic Analysis*. Boston, MA: Addison-Wesley, 2005.

Carrier, Brian. "Overview," *The Sleuth Kit.* http://www.sleuthkit.org/sleuthkit/ (accessed July 2012).

Carrier, Brian, and Eugene H. Spafford. "Getting Physical with the Digital Investigation Process." *International Journal of Digital Evidence* 2, no. 2 (Fall 2003): 1-20. https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2003-29.pdf (accessed August 2012).

"City of Vancouver Archives – Home Page." *City of Vancouver Archives.* http://vancouver.ca/ctyclerk/archives/ (accessed July 2012).

Cullen, Charles T., Peter B. Hirtle, David Levy, Clifford A. Lynch, and Jeff Rothenberg. *Authenticity in a Digital Environment.* Washington, D.C.: Council on Library and Information Resources, 2000. http://www.clir.org/pubs/abstract//reports/pub92 (accessed August 2012).

Cunningham, Adrian. "The Archival Management of Personal Records in Electronic Form: Some Suggestions." *Archives and Management* 22, no. 1 (1994): 94-105.

Cunningham, Adrian. "Ghosts in the Machine: Towards a Principles-Based Approach to Making and Keeping Digital Personal Records." in *I, Digital: Personal Collections in the Digital Era*. Christopher A. Lee, ed., 78-89. Chicago, IL: Society of American Archivists, 2011.

Cunningham, Adrian. "Waiting for the Ghost Train: Strategies for managing electronic personal records before it is too late." Pre-publication version of a paper delivered to the Society of American Archivists Annual Meeting, August 23-29, 1999, Pittsburgh, PA, USA. Published in *Archival Issues: Journal of the Midwest Archives Conference* 24, no. 1 (1999): 55-64. http://www.mybestdocs.com/cunningham-waiting2.htm (accessed June 2012).

Damelio, Robert. *The Basics of Process Mapping*. Portland, OR: Productivity, Inc., 1996.

Davis, Susan E. "Electronic Records Planning in 'Collecting' Repositories." *American Archivist* 71 (Spring/Summer 2008): 167-189, http://archivists.metapress.com/content/024q2020828t7332/fulltext.pdf (accessed July 2012).

"dd." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/Dd (accessed on July 2012).

"dd (UNIX)." *Wikipedia – The Free Encyclopedia.* http://en.wikipedia.org/wiki/Dd_(Unix) (accessed July 2012).

"Ddrescue." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/Ddrescue (accessed July 2012).

"The Deena Larsen Collection at the Maryland Institute for Technology in the Humanities." *Maryland Institute for Technology and the Humanities.* http://mith.umd.edu/larsen/ (accessed July 2012).

"Delta encoding." *Wikipedia – the free online encyclopedia.* http://en.wikipedia.org/wiki/Delta_encoding (accessed July 2012).

"Disk image." *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Disk_image (accessed July 2012).

Dooley, Jackie M., and Katherine Luce, *Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives.* Dublin, OH: OCLC Research, 2010. http://www.oclc.org/research/publications/library/2010/2010-11.pdf (accessed July 2012).

Dow, Elizabeth. *Electronic Records in the Manuscript Repository.* Lanham, Maryland: The Scarecrow Press, 2009.

"DROID fact sheet: Using DROID to profile your file formats." *The National Archives.* http://www.nationalarchives.gov.uk/documents/information-management/droid-factsheet.pdf (accessed July 2012).

DROID: How to Use it and How to Interpret your Results. London, UK: The National Archives, 2011. http://www.nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf (accessed July 2012).

Duranti, Luciana. "Diplomatics: New Uses for an Old Science." *Archivaria* 28 (1989): 7-27. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/11567/12513 (accessed August 2012).

Duranti, Luciana. "From Digital Diplomatics to Digital Records Forensics." *Archivaria* 68 (2009): 39-66. http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13229 (accessed July 2012).

Duranti, Luciana. "Reliability and Authenticity: The Concepts and their Implications." *Archivaria* 39 (1995): 5-10. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12063/13035 (accessed August 2012).

Duranti, Luciana, and Heather MacNeil. "The Protection of the Integrity of Electronic Records: An Overview of the Findings of the UBC-MAS Research Project." *Archivaria* 42 (Fall 1996): 46-67. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/12153 (accessed July 2012).

Elford, Douglas, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, and Colin Webb. "Media matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia." Paper presented at the 74th IFLA General Conference and Council, Québec, Canada, August 10-14, 2008. http://www.nla.gov.au/digital-preservation/related-staff-papers (accessed July 2012).

Farmer, Dan, and Wietse Venema. *Forensic Discovery*. Upper Saddle River, NJ: Addison-Wesley, 2005. http://www.porcupine.org/forensics/forensic-discovery/ (accessed July 2012).

"FC5025 USB 5.25 Floppy Controller – Device Side Data." *Deviceside.com.* http://www.deviceside.com/fc5025.html (accessed July 2012).

"FC5025 Floppy Disk Controller :Use Guide" *MITH's Vintage Computers* http://mith.umd.edu/vintage-computers/fc5025-operation-instructions (accessed July 2012).

"Fedora Repository – About." *Fedora Commons.* http://www.fedora-commons.org/about (accessed July 2012).

"fits – File Information Tool Set – Google Project Hosting." *fits – File Information Tool Set.* http://code.google.com/p/fits/ (accessed July 2012).

"Fiwalk – Forensics Wiki." *Digital Forensics Wiki*. http://www.forensicswiki.org/wiki/Fiwalk (accessed July 2012).

"Floppy disk controller." *Wikipedia – The Free Encyclopedia.* http://en.wikipedia.org/wiki/Floppy_disk_controller (accesssed June 2012).

"Forensic Hashing." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/Hashing (accessed July 2012).

"Forensic Toolkit." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/FTK (accesed July 2012).

Forstrom, Michael. "Managing Electronic Records in Manuscript Collections: A Case Study from the Beinecke Rare Book and Manuscript Library." *American Archivist* 72 (Fall/Winter 2009): 460-477. http://archivists.metapress.com/content/b82533tvr7713471/fulltext.pdf (accessed July 2012).

"FRED." *Digital Intelligence: mastering the science of digital forensics.* http://www.digitalintelligence.com/products/fred/ (accessed July 2012).

"FTK 4 Data Sheet." *Accessdata.com.* http://accessdata.com/downloads/media/FTK_DataSheet_web.pdf (accessed July 2012).

"FTK Imager." *Digital Forensics Wiki.* http://www.forensicswiki.org/wiki/FTK_Imager (accessed July 2012).

Garfinkel, Simson. "AFF: A New Format for Storing Hard Drive Images." *Communications of the ACM* 49, no. 2 (February 2006): 85-87. http://simson.net/clips/academic/2006.CACM.AFF.pdf (accessed June 2012).

Garfinkel, Simson. "Digital forensics research: The next 10 years." *Digital Investigation* 7 (2010): S64-S73. http://dfrws.org/2010/proceedings/2010-308.pdf (accessed July 2012).

Garfinkel, Simson. "Digital Forensics XML and the DFXML Toolset." *Digital Investigation* 8 (2012): 161-74. http://simson.net/clips/academic/2012.DI.dfxml.pdf (accessed June 2012).

Garfinkel, Simson. "Fiwalk." *DEEP: Digital Evaluation and Exploitation: Department of Computer Science, Naval Postgraduate School, Monterey, CA.* https://domex.nps.edu/deep/Fiwalk.html (accessed July 2012).

Garfinkel, Simson L., David J. Malan, Karl-Alexander Dubec, Christopher C. Stevens, and Cecile Pham. "Advanced forensic format: An open, extensible format for disk imaging." In *Advances in Digital Forensics II: FIP International Conference on Digital Forensics, National Center for Forensic Science, Orlando, Florida, January 29-February 1, 2006*, ed. Martin Olivier and Sujeet Shenoi, 17-31. New York: Springer, 2006. http://nrs.harvard.edu/urn-3:HUL.InstRepos:2829932 (accessed June 2012).

Garfinkel, Simson, Paul Farrell, Vassil Roussev, and George Dinolt. "Bringing Science to Digital Forensics with Standardized Forensic Corpora." *Digital Investigation* 6 (2009): S2-S11. http://www.dfrws.org/2009/proceedings/p2-garfinkel.pdf (accessed June 2012).

Garfinkel, Simson, and Abhi Shelat. “Remembrance of Data Passed: A Study of Disk Sanitation Practices.” *IEEE Security and Privacy* (January/February 2003): 17-27, http://cdn.computerscience1.net/2005/fall/lectures/8/articles8.pdf (accessed August 2012).

Gengenbach, Martin. “About the Curator's Workbench.” July 14, 2011. *Carolina Digital Repository Blog.* http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench/ (accessed June 2012).

Glisson, William Bradley. “Use of Computer Forensics in the Digital Curation of Removable Media.” In *Digital Curation: Practice, Promise and Prospects: Proceedings of DigCCurr 2009*, April 1-3, 2009, Chapel Hill, NC, USA, edited by H.R. Tibbo, 110-111. School of Information and Library Science, University of North Carolina at Chapel Hill, 2009. http://eprints.gla.ac.uk/33687/1/33687.pdf (accessed August 2012).

Goldman, Ben. “Bridging the Gap: Taking Practical Steps Towards Managing Born-Digital Collections in Manuscript Repositories.” *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 12, no. 1 (2011): 11-24, http://rbm.acrl.org/content/12/1/11.full.pdf+html (accessed August 3, 2012).

Gordon, Heather. “Our New Online Search.” April 11, 2012. *AuthentiCity: The City of Vancouver Archives Blog.* http://www.vancouverarchives.ca/2012/04/our-new-online-search/ (accessed July 2012).

Gueguen, Gretchen. “Born-Digital @UVa: The AIMS Framework Approach to Born-Digital Archives at UVa,” Presentation to the Mid-Atlantic Regional Archives Conference (MARAC) 2012 Spring Conference, Cape May, NJ. April 13, 2012. gretchengueguen.com/professional/GueguenMARACs12.ppt (accessed June 2012).

Harvey, Ross. “So Where's the Black Hole in Our Collective Memory? A Provocative Position Paper (PPP).” Digital Preservation Europe (DPE), 18 January 2008. http://www.digitalpreservationeurope.eu/publications/position/Ross_Harvey_black_hole_PPP.pdf (accessed June 2012).

Hedstrom, Margaret. “Digital Preservation: A Time Bomb for Digital Libraries.” *Computers and the Humanities* 31, no. 3 (1997): 189-202. http://hdl.handle.net/2027.42/42573 (accessed July 2012).

Henry, Linda J. “Schellenberg in Cyberspace.” *American Archivist* 61, no. 2 (Fall 1998): 309-327. http://www.jstor.org/stable/40294090 (accessed July 2012).

“History of the Library.” *National Library of Australia.* http://www.nla.gov.au/history-of-the-library (accessed July 2012).

“Home.” *JHOVE-JSTOR/Harvard Object Validation Environment.* http://hul.harvard.edu/jhove/ (accessed July 2012).

Hyry, Tom, and Rachel Onuf, “The personality of electronic records: the impact of new information technology on personal papers,” *Archival Issues* 22, no. 1 (1997): 37-44. http://minds.wisconsin.edu/handle/1793/45828 (accessed June 2012).

Jansen, Greg. “Announcing the Curator's Workbench.” December 1, 2010. *Carolina Digital Repository Blog.* http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/ (accessed June 2012).

Jarocki, John. "Forensics 101: Acquiring an Image with FTK Imager." June 18, 2009, *SANS Computer Forensics and Incident Response.* http://computer-forensics.sans.org/blog/2009/06/18/forensics-101-acquiring-an-image-with-ftk-imager/ (accessed July 2012).

John, Jeremy Leighton. "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools." Paper presented at the iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK. 2008. http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf (accessed August 2012).

John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. "Digital Lives: Personal digital archives for the 21st century >> an initial synthesis." A Digital Lives Research Paper. Beta Version 0.2, March 3, 2010. http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf (accessed June 2012).

Johnston, Leslie. "A 'Library of Congress' Worth of Data: It's All In How You Define It." April 25th, 2012. *The Signal: Digital Preservation.* http://blogs.loc.gov/digitalpreservation/2012/04/a-library-of-congress-worth-of-data-its-all-in-how-you-define-it/ (accessed July 2012).

Kirschenbaum, Matthew G. Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use.* College Park, MD: University of Maryland, 2009. http://mith.umd.edu/wp-content/uploads/whitepaper_HD-50346.Kirschenbaum.WP.pdf (accessed July 2012).

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriella 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/abstract/reports/pub149 (accessed July 2012).

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." in *Proceedings, iPRES 2009: the Sixth International Conference on Preservation of Digital Objects* (2009): 105-112. http://escholarship.org/uc/item/7d3465vg (accessed July 2012).

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination.* Cambridge, Mass: The MIT Press, 2008.

"KryoFlux." *KryoFlux – USB Floppy Controller.* http://www.kryoflux.com/ (accessed July 2012).

Kuny, Terry. "A Digital Dark Ages? Challenges in the Preservation of Electronic Information." *63rd International Federation of Library Associations (IFLA) Council and General Conference.* September 4, 1997. http://archive.ifla.org/IV/ifla63/63kuny1.pdf (accessed June 2012).

Lazorchak, Butch. "Rescuing the Tangible from the Intangible." July 2, 2012. *The Signal: Digital Preservation.* http://blogs.loc.gov/digitalpreservation/2012/07/rescuing-the-tangible-from-the-intangible/ (accessed July 2012).

Lee, Christopher A. "Donor Agreements." in *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*, 57. Washington, D.C.: Council on Library and Information Resources, 2010. http://www.clir.org/pubs/abstract/reports/pub149 (accessed July 2012).

Lee, Christopher A., ed. *I, Digital: Personal Collections in the Digital Era*. Chicago, IL: Society of American Archivists, 2011.

Lee, Christopher A. "Open Archival Information System (OAIS) Reference Model." *Encyclopedia of Library and Information Sciences, Third Edition.* Taylor & Francis, 2010: 4020-4030. http://www.tandfonline.com/doi/abs/10.1081/E-ELIS3-120044377 (accessed August 2012).

Lee, Christopher A., and Robert Capra. "And Now the Twain Shall Meet: Exploring the Connections between PIM and Archives." in *I, Digital: Personal Collections in the Digital Era,* edited by Christopher A. Lee, 29-77. Chicago, IL: Society of American Archivists, 2011.

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, no. 5/6 (May/June 2012). http://www.dlib.org/dlib/may12/lee/05lee.html (accessed August 2012).

Lefurgy, Bill. "Visualizing Digital Preservation Workflows." March 8, 2012. *The Signal: Digital Preservation* http://blogs.loc.gov/digitalpreservation/2012/03/visualizing-digital-preservation-workflows/ (accessed July 2012).

Levy, David. *Scrolling Forward: Making Sense of Documents in the Digital World.* New York, NY: Arcade Publishing, 2001.

Lynch, Clifford. "The Integrity of Digital Information: Mechanics and Definitional Issues." *Journal of the American Society for Information Science* 45, no. 10 (1994): 737-744.

Marshall, Cathy. "Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field." *D-Lib Magazine* 14, no. 3/4 (March/April 2008). http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html (accessed August 2012).

McKemmish, Rodney."When Is Digital Evidence Forensically Sound?" *Advances in Digital Forensics IV.* Indrajit Ray and Sujeet Shenoi, eds. IFIP International Federation for Information Processing 285. Boston: Springer, 2008: 3-15. http://www.springerlink.com/content/048j747850234355/ (accessed August 2012).

"Metadata Basics." *Dublin Core Metadata Initiative. http://dublincore.org/metadata-basics/* (accessed July 2012).

<METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual. Version 1.6 Revised. Washington, D.C.: Digital Library Federation, 2010. http://www.loc.gov/standards/mets/METSPrimerRevised.pdf (accessed July 2012).

"Metadata Extraction Tool." *National Library of New Zealand.* http://meta-extractor.sourceforge.net/ (accessed July 2012).

"Metadata Extraction Tool Information Sheet." *National Library of New Zealand.* http://meta-extractor.sourceforge.net/meta-extractor-info-sheet.pdf (accessed July 2012).

"Metadata Object Description Schema: MODS (Library of Congress)." *Library of Congress.* http://www.loc.gov/standards/mods/ (accessed July 2012).

Mocas, Sarah. "Building theoretical underpinnings for digital forensics research." *Digital Investigation* 1, no. 1 (2004): 61-68. http://www.dblp.org/db/journals/di/di1.html (accessed August 2012).

Mumma, Courtney, Glen Dingwall, and Sue Bigelow. "A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value="true";." *Archivaria* 72 (Fall 2011): 93-122. http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13361 (accessed July 2012).

"NDIIP Partner Tools and Services Inventory." *Digital Preservation (Library of Congress).* http://www.digitalpreservation.gov/tools/ (accessed July 2012).

Nelson, Naomi, chair. "Managing Electronic Records and Assets: A Pilot Study on Identifying Best Practices." Chicago, IL: Society of American Archivists Technology Best Practices Task Force, 2009. http://www.archivists.org/governance/taskforces/MERA-PilotStudy.pdf (accessed July 2012).

O'Meara, Erin, and Meg Tuomala. "Finding Balance Between Archival Principles and Real-Life Practices in an Institutional Repository." *Archivaria* 73 (2012): 81-103. http://journals.sfu.ca/archivar/index.php/archivaria/article/viewArticle/13385 (accessed August 2012).

Onuf, Rachel, and Tom Hyry. "Take It Personally: The Implications of Personal Records in Electronic Form." in *I, Digital: Personal Collections in the Digital Era*. Christopher A. Lee, ed., 241-256. Chicago, IL: Society of American Archivists, 2011.

"The Open Archives Initiative Protocol for Metadata Harvesting v 2.0." *Open Archives.* http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed July 2012).

The Open Group. "dd: IEEE Standard 1003.1-2008." *The Open Group Base Specifications* 7 (2008). http://pubs.opengroup.org/onlinepubs/9699919799/utilities/dd.html (accessed July 2012).

Palmer, Gary. *A Road Map for Digital Forensic Research. Technical Report DTR- T0010-01*. Report from the First Digital Forensic Research Workshop (DFRWS), November 2001. http://www.dfrws.org/2001/dfrws-rm-final.pdf (accessed August 2012).

Paquet, Lucie. "Appraisal, Acquisition and Control of Personal Electronic Records: From Myth to Reality." *Archives and Manuscripts* 28, no. 2 (2000): 71-91.

"paradigm | workbook on digital private papers." *Paradigm Project.* 2007. http://www.paradigm.ac.uk/workbook/ (accessed July 2012).

Peters, Catherine Stollar. "When Not All Papers are Paper: A Case Study in Digital Archivy." *Provenance* 24 (2006): 23-35. https://pacer.ischool.utexas.edu/bitstream/2081/2226/1/023-035.pdf (accessed July 2012).

Pollitt, Mark M. "An Ad Hoc Review of Digital Forensic Models." *Proceeding of the Second International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE'07)*, Washington, D.C., USA, 2007. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4155349 (accessed August 2012).

del Pozo, Nicholas, Andrew Stawowczyk Long and David Pearson. "'Land of the lost:' a discussion of what can be preserved through digital preservation." *Library Hi Tech* 28, no. 2 (2010): 290-300. doi: *10.1108/07378831011047686* (accessed June 2012).

PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata.* Version 2.2 Washington, D.C.: Library of Congress, July 2012. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf (accessed July 2012).

"Prometheus – About." *Prometheus Digital Preservation Workbench.* http://mediapedia.nla.gov.au/prometheus/about.html (accessed July 2012).

"Prometheus Digi Pres Workbench." *SourceForge.net.* http://sourceforge.net/projects/prometheus-digi/ (accessed July 2012).

Raimas, Alan J., and Richard Rummler, "The Evolution of the Effective Process Framework: A Model for Redesigning Business processes." *Performance Improvement* 48, no. 10 (November/December 2009): 25-32. http://onlinelibrary.wiley.com/doi/10.1002/pfi.20112/abstract (accessed May 2012).

Reith, Mark, Clint Carr, and Gregg Gunsch. "An Examination of Digital Forensic Models." *International Journal of Digital Evidence* 1, no. 3 (Fall 2002): 1-12. http://people.emich.edu/pstephen/other_papers/Digital_Forensic_Models.pdf (accessed August 2012).

Ross, Seamus, and Ann Gow. *Digital Archaeology: Rescuing Neglected and Damaged Data Resources. London: British Library, 1999.* http://www.ukoln.ac.uk/services/elib/papers/supporting/pdf/p2.pdf (accessed June 2012).

Rothenberg, Jeff. *Ensuring the Longevity of Digital Information.* Washington, D.C.: Council on Library and Information Resources, 1999. http://www.clir.org/pubs/archives/ensuring.pdf (accessed July 2012).

"rsync." *Wikipedia – the free online encyclopedia.* http://en.wikipedia.org/wiki/Rsync (accessed July 2012).

Schoenfeld, Jens. "Individual Computers: CatWeasel," *Individual Computers.* http://www.jschoenfeld.com/home/indexe.htm (accessed June 2012).

Theimer, Kate. "Intro to the Archivists' Toolkit." September 17, 2009. *ArchivesNext* http://www.archivesnext.com/?p=370 (accessed June 2012).

Thibodeau, Kenneth. "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years." in *The State of Digital Preservation: An International Perspective*. Washington, D.C.: Council on Library and Information Resources, 2002. http://www.clir.org/pubs/reports/pub107/thibodeau.html/ (accessed July 2012).

Thomas, Susan. "Curating the I, Digital: Experiences at the Bodleian Library," in *I, Digital: Personal Collections in the Digital Era,* edited by Christopher A. Lee, 280-305. Chicago, IL: Society of American Archivists, 2011.

Thomas, Susan. *Paradigm: A practical approach to the preservation of personal digital archives.* Oxford, UK: Bodleian Library, 2007. http://www.paradigm.ac.uk/projectdocs/jiscreports/index.html (accessed July 2012).

Thomas, Susan, and Janette Martin. "Using the papers of contemporary British politicians as a testbed for the preservation of digital personal archives." *Journal of the Society of Archivists* 27, no. 1 (April 2006): 29-56. http://eprints.hud.ac.uk/7893/ (accessed July 2012).

Waters, Donald, and John Garrett. *Preserving Digital Information: Report of the Task Force on Archiving Digital Information.* Washington, D.C.: Committee on Preservation and Access, 1996. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf (accessed July 2012).

White, Ron. *How Computers Work.* Timothy Edward Downs, illus. 9th edition. Indianapolis, IN: Que Publishing, 2008.

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

Woods, Kam, Christopher A. Lee, and Simson Garfinkel. "Extending Digital Repository Architectures to Support Disk Image Preservation and Access." *JCDL 11, Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries,* June 13-17, 2011, Ottawa, Ontario, Canada (2011): 57-66. http://ils.unc.edu/callee/p57-woods.pdf (accessed July 2012).

Woods, Kam, and Geoffrey Brown, "From Imaging to Access – Effective Preservation of Legacy Removable Media," In *Proceedings of Archiving 2009,* 213-218. Springfield, VA: Society for Imaging Science and Technology, 2009. http://www.digpres.com/publications/woodsbrownarch09.pdf (accessed August 2012).

"Workflow." *Wikipedia – The Free Encyclopedia. http://en.wikipedia.org/wiki/Workflow* (accessed June 2012).

"Write Blockers – Forensics Wiki." *The Forensics Wiki.* http://www.forensicswiki.org/wiki/Write_Blockers (accessed July 2012).

"Yale University Rescue Repository: About the Rescue Repository." *Yale University Library: Integrated Systems and Programming Group.* http://www.library.yale.edu/ito/RRweb/AboutRescueRepository.html (accessed July 2012).

Yusoff, Yunus, Roslan Ismail and Zainuddin Hassan. "Common Phases of Computer Forensics Investigation Models." *International Journal of Computer Science and Information Technology (IJCSIT)* 3, no. 3 (June 2011): 17-31. http://airccse.org/journal/jcsit/0611csit02.pdf (accessed August 2012).

Appendix A: Participant Solicitation Email

Email Recruitment Revision Date: 10 April 2012

Title of Study: Modeling Digital Preservation Workflows in Collecting Institutions

Principle Investigator: Martin Gengenbach, MLS Candidate
UNC Chapel Hill Department: School of Information and Library Science
Phone: (253) 224-9648 Email: gengenmj@live.unc.edu

Faculty Advisor: Christopher (Cal) Lee, Assistant Professor
UNC Chapel Hill Department: School of Information and Library Science
Phone: (919) 962-7024 Email: callee@email.unc.edu

Dear _____,

My name is Marty Gengenbach. I am currently a Master's student at the School of Information and Library Science at the University of North Carolina at Chapel Hill.

I am writing to you to request your participation in a research project on the implementation of digital forensic tools and practices in workflows for the preservation of born-digital content in collecting institutions. You have written of your own active use of digital forensic tools in the recent publication (*name and date of publication*). This work has added to a growing body of scholarship in this area, and identified you as a significant contributor to the field.

In my Master's paper I hope to explore how institutional workflows change by implementing digital forensic tools and practices to manage born-digital content. Because of your previous experience in this area, I would greatly appreciate the opportunity to interview you, to learn more about your experiences implementing such tools in your institutional setting. My primary interests are in your particular workflows for handling born-digital content, and the challenges, opportunities, and strategies involved in implementing that workflow.

The interview will last approximately one hour, and your participation is entirely voluntary. Pay will not be provided, but your participation will further contribute to this exciting and rapidly developing research area. Attached you will find a statement of informed consent, which contains more detailed information on the study and your role in it. Continuing your participation indicates your understanding and acceptance of the contents of the statement of informed consent, so please look it over. Note that you can withdraw from the study at any time, and that your participation is completely voluntary. If you are interested in participating, please contact me at gengenmj@live.unc.edu. You can also contact my faculty advisor, Dr. Christopher (Cal) Lee, at callee@email.unc.edu, or (919) 962-7024.

I hope to hear from you!

Sincerely,

Martin J. Gengenbach
MSLS '12, University of North Carolina at Chapel Hill
gengenmj@live.unc.edu
(253) 224-9648

Appendix B: Statement of Informed Consent

University of North Carolina-Chapel Hill
Consent to Participate in a Research Study
Adult Participants Social Behavioral Form

Consent Form Version Date: April 2012

Title of Study: Modeling Digital Forensic Workflows in Collecting Institutions

Principal Investigator: Martin Gengenbach
UNC-Chapel Hill Department: School of Information and Library Science
UNC-Chapel Hill Phone number: (253) 224 - 9648
Email Address: gengenmj@live.unc.edu
Faculty Advisor: Christopher A. (Cal) Lee, School of Information and Library Science
Faculty Advisor Phone Number: (919) 962-7024
Faculty Advisor email: callee@email.unc.edu

Study Contact telephone number: (253) 224-9648
Study Contact email: gengenmj@live.unc.edu

What are some general things you should know about research studies?

You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, without penalty. For University of North Carolina at Chapel Hill employees, participation in the study will not affect your employment status or benefits at UNC-CH.

Research studies are designed to obtain new knowledge. This new information may help people in the future. You may not receive any direct benefit from being in the research study. There also may be risks to being in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study. You will be given a copy of this consent form. You should ask the researchers named above, or staff members who may assist them, any questions you have about this study at any time.

What is the purpose of this study?

The purpose of this research study is to explore how digital forensic tools and technologies are implemented in the workflows of archives and institutions charged with collecting and preserving born-digital content. This study will consult with professionals active in the field of digital preservation through semi-structured interviews to examine the how digital forensic tools are used in collecting institutions, and how workflows

change with the implementation of such tools and practices in different collecting institutions.

You are being asked to be in the study because of your leadership in the field as an archivist/manager of electronic records and born-digital content.

How many people will take part in this study?

If you decide to be in this study, you will be one of approximately 10-12 people in this research study.

How long will your part in this study last?

Your time commitment for the purpose of this study will consist of an interview lasting approximately one hour, conducted via telephone or Skype. You can stop the interview at any time. Other possible time commitments relate to answering follow-up emails to clarify or elaborate upon points discussed in the original interview.

What will happen if you take part in the study?

1. If you take part in this study, the researcher will contact you via email to determine a preferred time to engage in an individual interview, the topic of which will be the workflows through which your institution acquires, ingests, processes, and provides access to born-digital content.
2. The interviewer will provide you a copy of the workflow documentation that has been compiled from existing sources. The interview will be based on an interview protocol in addition to questions specific to existing workflow documentation (when available).
3. This interview will last one hour at the most, and will be conducted via Skype or telephone.
4. This interview will be audio-recorded and transcribed for reference by the researcher. The interview data will be de-identified, and you will not be attributed for any quotes or statements made in the interview.
5. Workflow documentation generated from the interviews will be sent to you for verification and comment, clarification, and correction.
6. Following the interview, it is possible that you will be contacted via email and asked to clarify or further elaborate upon some portion of the interview.
7. De-identified statements from interviews and workflow documentation created from those interviews will be used by the researcher in publications. Workflow documentation will be identified by institution.

What are the possible benefits from being in this study?

Research is designed to benefit society by gaining new knowledge. You may also expect to benefit by participating in this study by gaining a better understand and documentation of your institution's own procedures and workflows for acquiring and accessioning digital

content.

What are the possible risks or discomforts involved from being in this study?

The interview and any follow-up communication will have minimal potential for immediate or long-term physical, psychological, or social risks/discomforts. The topics covered will be primarily related to workplace practices and procedures in working with digital content. There is a chance of deductive disclosure of your identity, due to the small population of archivists implementing digital forensic tools and practices. You can decide at what level of granularity you wish to provide information on your institution's practices, according to your own level of comfort. However, there may be other uncommon or previously unknown risks. You should report any problems to the researcher.

How will your privacy be protected?

1. The only person with access to identifiable interview data will be the researcher.
2. It will be necessary for the researcher to collect personal contact information including your name, email address, and telephone number, in order to communicate during the duration of the study. To the greatest extent possible, personal information will be kept separate from the general body of data. A coded system will link participants to their information in a password-protected document that will be kept away from the general body of research data.
3. Workflow documentation will be identifiable by institution only.
4. During the course of the study, all audio recordings in digital form will be retained in password-protected files on the researcher's personal computer requiring password access upon startup.
5. Audio recordings in digital form will be permanently deleted after transcription has occurred.
6. The interview transcript will be permanently deleted at the end of the study; print copies will be destroyed in a secure and confidential manner at that time, as well.

Although every effort will be made to keep research records private, there may be times when federal or state law requires the disclosure of such records, including personal information. This is very unlikely, but if disclosure is ever required, UNC-Chapel Hill will take steps allowable by law to protect the privacy of personal information. In some cases, your information in this research study could be reviewed by representatives of the University, research sponsors, or government agencies for purposes such as quality control or safety.

What if you want to stop before your part in the study is complete?

You can withdraw from this study at any time, without penalty. The investigators also have the right to stop your participation at any time. If you withdraw from this study, the interview recording, as well as the interview transcript and any related notes based on that recording or transcript, will be destroyed in a secure and confidential manner.

Will you receive anything for being in this study?

You will not receive anything for taking part in this study.

What if you have questions about this study?

You have the right to ask, and have answered, any questions you may have about this research. If you have questions, complaints, concerns, or if a research-related injury occurs, you should contact the researchers listed on the first page of this form.

What if you have questions about your rights as a research participant?

All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject, or if you would like to obtain information or offer input, you may contact the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.

Thank you for helping me with this study.

- -

Title of Study: Modeling Digital Forensic Workflows in Collecting Institutions

Principal Investigator: Martin Gengenbach

Participant's Agreement:

I have read the information provided above. I have asked all the questions I have at this time. I voluntarily agree to participate in this research study.

_____ _____

Signature of Research Participant Date

Appendix C: Interview Protocol

Martin Gengenbach
Master's Paper: Modeling Digital Preservation Workflows in Collecting Institutions
4/10/2012
Application for IRB Approval: Interview Protocol

Interview Protocol Revision Date: 10 April 2012

Title of Study: Modeling Digital Preservation Workflows in Collecting Institutions

Principle Investigator: Martin Gengenbach, MLS Candidate
UNC Chapel Hill Department: School of Information and Library Science
Phone: (253) 224-9648 Email: gengenmj@live.unc.edu

Faculty Advisor: Christopher (Cal) Lee, Assistant Professor
UNC Chapel Hill Department: School of Information and Library Science
Phone: (919) 962-7024 Email: callee@email.unc.edu

Interview Protocol

1. What is your position title, and could you speak a little about your role within your institution?

2. Are you the only person in your institution who works with born-digital content?

3. How long have you been accepting born-digital content at your institution?

4. Do your donor/collection development policies specifically address the donation and accession of digital content?

5. Could you talk me through your institution's current workflow for acquiring born-digital materials?

6. Is this different from how you used to acquire materials?

7. What changes in your workflow have you found to be the most helpful or important?

8. Which changes in your workflows have been most difficult to implement? Why have these changes been difficult?

9. Are there other changes that you'd like to make to your workflow? If so, what kind of changes?

10. Are you taking any steps to ingest born-digital content from hybrid collections that

have already been accessioned? (legacy materials, disks-in-boxes)

11. What do you see as your biggest problems in working with born-digital material?

12. How do you prioritize high value born-digital collections?

Appendix D: Digital Preservation Tools and Technologies Referenced

1. Advanced Forensic Format (AFF)

The Advanced Forensic Format (AFF), developed by Simson Garfinkel with support from Basis Technology Corp, is an openly documented file format for storing disk images that can be integrated into digital forensics tools.[247] The AFF disk image differs from raw disk images in several respects. Whereas raw disk images cannot retain associated metadata within the same file, AFF files can retain disk image metadata in the same file as the disk image itself, or as an associated companion file.[248] The disk image file is separated into two layers: a disk representation layer, which "defines a schema that is used for storing disk images and associated metadata[;]" and a disk storage layer defining the storage of disk-image segments within the AFF file.[249] Additionally, a raw disk image must be uncompressed to be accessible to digital forensic tools, as reading a disk image requires the same random access that is required of a regular hard drive.[250] AFF files use segmentation and header information that allow for disk image compression without sacrificing accessibility to the information in the image.[251]

---

[247] "AFF," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/AFF (accessed July 2012).

[248]Simson Garfinkel, "AFF: A New Format for Storing Hard Drive Images," *Communications of the ACM* 49 no. 2 (February 2006): 86, http://simson.net/clips/academic/2006.CACM.AFF.pdf (accessed June 2012).

[249]Simson Garfinkel et al., "Advanced Forensic Format: An Open, Extensible Format for Disk Imaging," *Advances in Digital Forensics II,* ed. Martin Olivier and Sujeet Shenoi. FIP International Conference on Digital Forensics, National Center for Forensic Science, Orlando, Florida, January 29-February 1, 2006, (New York: Springer, 2006), 23, http://nrs.harvard.edu/urn-3:HUL.InstRepos:2829932 (accessed June 2012).

[250] Garfinkel, "AFF: A New Format for Storing Hard Drive Images," 85.

[251]Garfinkel, "Advanced Forensic Format," 26-27.

2. Archivematica

Archivematica is a free, open-source, digital preservation software system, managed by Artefactual Systems, Inc. and developed in collaboration with a wide range of partners including the City of Vancouver, British Columbia, Canada; the University of British Columbia, Vancouver; and the UNESCO Memory of the World's Subcommittee on Technology.[252] The software is built to function within an Open Archival Information System (OAIS) compliant digital preservation repository, and is developed around a suite of micro-services—"granular system tasks which operate on a conceptual entity that is equivalent to an OAIS information package."[253] Information packages are moved through the micro-services pipeline, triggering digital preservation actions arranged around OAIS processes such as ingest, storage, data management, administration, and access.[254] These micro-services are open-source and single-purpose, meaning if a new tool or technology supersedes an existing micro-service, that tool can be "swapped out."[255] There is a growing community of active Archivematica users, though as of this writing it is only in the pre-release 0.8 alpha development phase.[256] The first beta release, 0.9, is expected in the near future.[257]

---

[252]"Archivematica Main Page" *Archivematica* https://www.archivematica.org/wiki/Main_Page (accessed June 2012).

[253]"Micro-services," *Archivematica* https://www.archivematica.org/wiki/Micro-services (accessed June 2012).

[254]There is an extensive amount of documentation available on how Archivematica maps to conceptual OAIS stages. See "UML Activity Diagrams – Archivematica," *Archivematica* https://www.archivematica.org/wiki/UML_Activity_Diagrams (accessed June 2012).

[255]"Micro-services," *Archivematica* https://www.archivematica.org/wiki/Micro-services (accessed June 2012). For more information on the concept of micro-services see Stephen Abrams, John Kunze, and David Loy, "An Emergent Micro-Services Approach to Digital Curation Infrastructure," *The International Journal of Digital Curation* 5, no. 1 (2010): 172-186, http://www.ijdc.net/index.php/ijdc/article/view/154 (accessed June 2012).

[256]Users and developers maintain an active discussion list that is open to the public at: http://groups.google.ca/group/archivematica (accessed June 2012).

[257]More information on Archivematica is available at: https://www.archivematica.org/wiki/Main_Page.

3. Archivists' Toolkit

The Archivists' Toolkit is "the first open source archival data management system to provide broad, integrated support for the management of archives."[258] The project, initiated in 2006, is a collaboration between the University of California San Diego Libraries, New York University Libraries, and Five Colleges, Inc., Libraries, funded by the Andrew W. Mellon Foundation, though it has since grown to include more than 100 different institutions as users.[259] Project goals include "to support archival processing and production of access instruments, promote data standardization, promote efficiency, and lower training costs."[260] It is not a digital preservation tool specifically, but an archives management tool that allows the tracking and collection of data about archival collections in a centralized database. Recently, Archivists' Toolkit announced its merger with another open access archives management tool, Archon.[261] The combined tool, ArchivesSpace, is still in development and is hosted by Lyrasis.[262]

4. BagIt Specification

BagIt is a file packaging specification developed by the Library of Congress, the California Digital Library, and Stanford University through the National Digital

---

[258]"Archivists' Toolkit | For Archivists By Archivists," *Archivists' Toolkit* http://www.archiviststoolkit.org/ (accessed June 2012).
[259]"List of AT Users | Archivists' Toolkit," *Archivists' Toolkit* http://archiviststoolkit.org/support/ListofATUsers (accessed June 2012).
[260]"Frequently Asked Questions | Archivists' Toolkit," *Archivists' Toolkit* http://archiviststoolkit.org/faq (accessed June 2012).
[261]"Archon: The Simple Archival Information System," *Archon* http://www.archon.org/about.php (accessed June 2012).
[262]Information on the ArchivesSpace project can be found at: "ArchivesSpace: building a next generation archives management tool," *ArchivesSpace* http://www.archivesspace.org/ (accessed June 2012). For an introduction to Archivists' Toolkit, see Kate Theimer, "Intro to the Archivists' Toolkit," September 17, 2009, *ArchivesNext* http://www.archivesnext.com/?p=370 (accessed June 2012).

Information Infrastructure and Preservation Program (NDIIPP) in 2007.[263] BagIt is designed to allow the transfer of digital content with verification that the content reaches its point of arrival unchanged. The basic specification consists of a "bag" containing digital content and a "tag," a simple text (.txt) file listing the files included in the bag with a hash value generated for each file packaged in the bag.[264] "A slightly more sophisticated bag lists URLs [Universal Resource Locators] instead of simple directory paths. A script consults the tag, detects the URLs and retrieves the files over the Internet, ten or more at a time. This type of simultaneous multiple transfer reduces overall data-transfer times. In another optional file, users can supply metadata that describes the bag."[265] NDIIPP has also developed various software tools to retrieve, verify, and validate bags to ensure bag transfers are completed successfully and efficiently.[266]

5. BitCurator

BitCurator is an ongoing, multi-institutional "effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions."[267] The project highlights two particular needs in the archival community, "incorporation into the workflow of archives/library ingest and collection management environments, and provision of public access to the data[,]" which it aims to meet through open-source software products currently in development.[268] BitCurator includes a Professional Experts Panel (PEP) with experience implementing forensic tools

---

[263]"BagIt," *Wikipedia – The Free Encyclopedia* http://en.wikipedia.org/wiki/BagIt (accessed June 2012).
[264]"NDIIP Partner Tools and Services Inventory" *Digital Preservation (Library of Congress)* http://www.digitalpreservation.gov/tools/ (accessed July 2012). A description of hashing can be found on page 15-16 of this work.
[265]Ibid.
[266]Information on BagIt is available at: http://www.digitalpreservation.gov/tools/ (accessed June 2012).
[267]"About | BitCurator," *The BitCurator Project* http://www.bitcurator.net/aboutbc/ (accessed July 2012).
[268]Ibid.

in institutional settings to advise on functionality requirements, and a Development Advisory Group (DAG) to provide guidance on the software development process.[269] BitCurator aims to provide integrated access to a number of digital forensic tools, facilitating disk imaging, data triage, metadata extraction, and redaction and access support, in an archivist-friendly interface.[270]The project is funded by the Andrew W. Mellon Foundation, and is a collaboration between the School of Information and Library Science at the University of North Carolina at Chapel Hill, and the Maryland Institute for Technology in the Humanities at the University of Maryland.

6. CatWeasel

Individual Computers' CatWeasel is a universal floppy disk controller, a piece of hardware that is used to connect and translate between a computer and an otherwise obsolete 3.5 or 5.25 inch floppy drive.[271] The disk controller contains the necessary chips and circuitry to facilitate the transfer of information between the disk drive and computer system.[272] This makes it possible for data from largely obsolete media formats, such as floppy disks, to be read and accessed by modern computer systems. It should be noted that the CatWeasel is no longer in production.[273]

---

[269]"People | BitCurator," *The BitCurator Project* http://www.bitcurator.net/people/ (accessed July 2012).
[270]"BitCurator," *The BitCurator Wiki* http://wiki.bitcurator.net/index.php?title=Main_Page (accessed July 2012).
[271]Jens Schoenfeld, "Individual Computers: CatWeasel," *Individual Computers* http://www.jschoenfeld.com/home/indexe.htm (accessed June 2012).
[272]"Floppy disk controller," *Wikipedia – The Free Encyclopedia* http://en.wikipedia.org/wiki/Floppy_disk_controller (accesssed June 2012).
[273]"FAQ – BitCurator," *BitCurator* http://wiki.bitcurator.net/index.php?title=FAQ (accessed July 2012).

7. ClamAV

ClamAV is a free, open-source anti-virus toolkit for UNIX-based operating systems, developed in 2002.[274] It is based on an anti-virus engine available as a shared library that is updated multiple times a day, and is designed primarily for email gateway scanning. ClamAV is capable of scanning for viruses in a variety of compressed data formats, so it can scan compressed files within a repository ingest package.[275] ClamAV serves as the engine at the heart of anti-virus tools for Mac, Windows, and Linux.[276] Because it is open-source, it can be integrated as a module in micro-service based digital preservation systems, such as Archivematica.[277]

8. Curator's Workbench

The Curator's Workbench is a pre-ingest collection management and workflow tool developed in 2010 at the University of North Carolina at Chapel Hill. The tool was developed for use within the Carolina Digital Repository (CDR), but is open-source and free to download.[278] "The Curator's Workbench is designed to facilitate the staging of large batches of objects with custom supplied metadata."[279] One innovative feature of the Workbench is the metadata crosswalk. As users work within the tool to map custom

---

[274]"About ClamAV," *Clam Anti Virus* http://www.clamav.net/lang/en/about/ (accessed July 2012).
[275]Ibid.
[276]For Mac, see *ClamXav: The Free Anti-Virus Solution for Mac OS* http://www.clamxav.com/ (accessed July 2012). For Linux distributions, see: *Clam Anti Virus* http://www.clamav.net/lang/en/download/packages/packages-linux/ (accessed July 2012). For Windows, see "Windows Antivirus," *Clam Anti Virus* http://www.clamav.net/lang/en/about/win32/ (accessed July 2012).
[277]See Appendix D, "Archivematica."
[278]Download is available at: http://www.lib.unc.edu/software/.
[279]Martin Gengenbach, "About the Curator's Workbench," July 14, 2011, *Carolina Digital Repository Blog* http://www.lib.unc.edu/blogs/cdr/index.php/about-the-curators-workbench/ (accessed June 2012).

metadata fields to Metadata Object Description Schema (MODS) elements, the Workbench builds out an extensible markup language (XML)-based Metadata Encoding and Transmission Standard (METS) file manifest incorporating that MODS metadata for each object in the ingest package, along with Preservation Metadata: Implementation Strategies (PREMIS) preservation event data, including checksums, UUIDs, and other information.[280] The final product is a Submission Information Package (SIP) that can be ingested into a digital repository.[281]

9. dd

dd is a UNIX command that outputs a bit-level copy of the target file, device, or drive.[282] A major benefit of dd is that it is available in any UNIX-based distribution, as it is a basic command-line instruction. By inputting a "dd" command with the appropriate parameters one can create a disk image of a target drive that disregards the filesystem and instead copies each block of underlying data, though the specific, limited capability of dd may make it a less desirable choice for creating bit-level copies.[283] It is one of the oldest tools used for imaging, and does not implement any of the metadata collection, error correction, or more advanced features used by other imaging tools.[284] For example, unless special instructions are input, the program will stop at the first error it

---

[280]Greg Jansen, "Announcing the Curator's Workbench," December 1, 2010, *Carolina Digital Repository Blog* http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/ (accessed June 2012).

[281]For more information on the Curator's Workbench, see the project website at: http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/ (accessed June 2012).

[282]"dd (UNIX)," *Wikipedia – The Free Encyclopedia* http://en.wikipedia.org/wiki/Dd_(Unix) (accessed July 2012).

[283]dd does not always take a full bit-level representation of a target medium. For example, when dd is used upon a CD, it will only copy the disc as represented by the filesystem. See the Prometheus FAQ, question 4, at: http://prometheus-digi.sourceforge.net/faq.html (accessed July 2012).

[284]"dd" *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Dd (accessed on July 2012).

encounters.[285]

10. ddrescue

ddrescue is a data recovery tool that attempts to minimize data loss and provides documentation on bad data blocks through a ddrescue log file. Whereas the dd command reads data sequentially, ddrescue attempts to efficiently read and copy the good data on a volume, going back and carving bad data blocks down to the sector level to maximize data recovery.[286] The ddrescue log file can be used to interrupt and resume a ddrescue recovery, and if multiple copies of a damaged file are each scanned with ddrescue, their logfiles can be merged to attempt to recreate the complete file.[287] Note that there is another program, dd_rescue, which is not related to ddrescue but shares similar functionality, with both adding to the basic block-level copying functionality of dd.[288]

11. Digital Forensics Extensible Markup Language (DFXML)

Digital Forensics Extensible Markup Language (DFXML) is an XML-based set of metadata conventions for representing information acquired through forensic processing and analysis.[289] Simson Garfinkel developed DFXML and has been using it in various forms since 2007; a number of users in the open-source digital forensics community have also modified and implemented the schema.[290] DFXML can be used to represent forensic

---

[285]The Open Group, "dd: IEEE Standard 1003.1-2008," *The Open Group Base Specifications* 7 (2008), http://pubs.opengroup.org/onlinepubs/9699919799/utilities/dd.html (accessed July 2012).
[286] "GNU ddrescue manual," *Ddrescue: Data Recovery Tool* http://www.gnu.org/software/ddrescue/manual/ddrescue_manual.html (accessed July 2012).
[287]Ibid.
[288]See "Ddrescue," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Ddrescue; and "Dd Rescue," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Dd_rescue (accessed July 2012).
[289]Simson Garfinkel, "Digital Forensics XML and the DFXML Toolset," 161.
[290]A list of tools that create or use DFXML is available at: "Category:Digital Forensics XML," *Digital*

processes, work products, and metadata, and to facilitate the interchange of information among forensic tools.[291] Used in a tool such as fiwalk, DFXML has five primary elements: <metadata>, containing header information which defines metadata used in the schema such as namespace and schema declarations; <creator>, containing information about the computer system and programs in which the forensic analysis is conducted; <source>, which details the computing environment of the target of forensic analysis; <volume>, the largest portion of the DFXML file, contains information compiled through analysis of the individual files on the target drive; and <runstats>, which contains additional information detailing the process of the analysis.[292] The BitCurator Project tools are also developed around DFXML outputs, and the project is working with Garfinkel to standardize the schema.[293]

12. Digital Record Object Identification (DROID)

DROID is an open-source, BSD-licensed file format identification tool developed by the The National Archives (UK).[294] DROID identifies information on file type, file type version, its size and date of creation, when it was last altered, and other information including its location (file path).[295] The current version of DROID (DROID 6) can

---

Forensics Wiki http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML#Tools (accessed July 2012).

[291]Garfinkel, "Digital Forensics XML and the DFXML Toolset," 162.

[292]See Garfinkel, "Digital Forensics XML," 166.

[293]"FAQ – BitCurator," *BitCurator Wiki* http://wiki.bitcurator.net/index.php?title=FAQ (accessed July 2012).

[294] "The BSD 2-Clause License," *The Open Source Initiative* http://opensource.org/licenses/bsd-license.php (accessed July 2012).

[295]"Using DROID to profile your file formats," *The National Archives* http://www.nationalarchives.gov.uk/documents/information-management/droid-factsheet.pdf (accessed July 2012).

identify 250 different file types, and generate MD5 checksums.[296] DROID was developed to check file types against the PRONOM file format registry, and so it also records a PRONOM Unique Identifier associated with a particular format.[297] It is sometimes used with other file format identification services such as JHOVE.

13. FC5025 Floppy Drive Controller

The Device Side Data FC5025 is a universal floppy drive controller. Drive controllers allow external drives to communicate with computer systems; in this case the FC5025 enables communication between a 5.25-inch floppy drive and a modern computer through a USB 1.1 or 2.0 port, in order to read data from obsolete floppy disks.[298] It includes software for accessing the drive via Mac (OS X), Linux, and Windows. The FC5025 is also capable of creating disk images of floppy disks, and if directed can create an error log to note data recovery issues.[299]

14. Fedora Repository Software

Fedora (Flexible Extensible Digital Object Repository Architecture) is an open-source digital object repository management framework.[300] It is not a full digital preservation system; rather it is a digital preservation platform developed to be integrated with digital

---

[296] "DROID: How to Use it and How to Interpret your Results," (London, UK: The National Archives, 2011): 6, http://www.nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf (accessed July 2012).
[297] "DROID: How to Use it and How to Interpret your Results," 16.
[298] "FC5025 USB 5.25" Floppy Controller – Device Side Data," *Deviceside.com* http://www.deviceside.com/fc5025.html (accessed July 2012).
[299] "Use Guide for the FC5025 Floppy Disk Controller," *MITH's Vintage Computers* http://mith.umd.edu/vintage-computers/fc5025-operation-instructions (accessed July 2012).
[300] "About – Fedora Repository," *Fedora Commons* http://www.fedora-commons.org/about (accessed July 2012).

preservation services through an application-programming interface (API).[301] Fedora provides digital object management services, can define access and security policies to classes of digital objects, link objects through semantic relationships, and is compliant with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) standard.[302] It is an open-access tool distributed under the Apache License, Version 2.0.[303] Fedora was developed at Cornell University in the late 1990s, with further development funded through an Andrew W. Mellon Foundation grant.[304] Fedora is the repository engine at the heart of a number of different initiatives, including the Hydra Project, a multi-institution repository development collaboration, Hypatia, a project associated with the AIMS project, and the Carolina Digital Repository, the digital repository for the University of North Carolina at Chapel Hill.[305]

15. File Information Tool Set (FITS)

The File Information Tool Set (FITS) is a wrapper for multiple open source tools used to identify, validate, and extract technical metadata for a variety of different file formats. It was developed by the Harvard University Library Office for Information Systems, and is

---

[301] Ibid.

[302] For more information on OIA-PMH, see: "The Open Archives Initiative Protocol for Metadata Harvesting v 2.0," *Open Archives* http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed July 2012).

[303] For information on the Apache License, Version 2.0, see: http://www.apache.org/licenses/LICENSE-2.0.html.

[304] Chris Awre, "Fedora and digital preservation," *Tackling the Preservation Challenge.* Presentation given December 12, 2008 at the University of Hull, UK. http://www.dpconline.org/events/previous-events/426-practical-steps-for-repository-managers (accessed July 2012).

[305] A more comprehensive list of projects using the Fedora platform can be found on the *Fedora Commons Community Registry* http://www.fedora-commons.org/community/comreg (accessed July 2012). Information on Hypatia can be found at: https://wiki.duraspace.org/display/HYPAT/Home (accessed August 2012). Information on the Hydra Project can be found at: http://projecthydra.org/ (accessed July 2012).

available on a GNU Lesser General Public License.[306] FITS allows the outputs from several open-source format identification and analysis tools to be standardized and wrapped within a single extensible markup language (XML) output file, enhancing interoperability and consolidating metadata. It can be operated as a standalone tool with a command-line interface, or it can be integrated into other systems through an API.[307] The current tools included in FITS are: JHOVE, ExifTool, New Zealand Metadata Extractor, DROID, FFident, and File Utility.[308]

16. fiwalk

fiwalk is an open-source, command-line software tool used to conduct batch forensic analysis on a target disk image, live system, or raw device.[309] It automates a number of forensic processes, and outputs a DFXML file containing the results of these processes, including a list of the disk's file system, individual file metadata, and information about the hosting and target systems.[310] fiwalk has been incorporated into the SleuthKit project, and comes with a Python module to facilitate the creation of customized tools for forensic analysis.[311]

---

[306] "fits – File Information Tool Set – Google Project Hosting," *fits – File Information Tool Set* http://code.google.com/p/fits/ (accessed July 2012). For more information on the GNU Lesser Public License, see: "GNU Lesser Public License," *GNU Operating System* http://www.gnu.org/licenses/lgpl.html (accessed July 2012).

[307] "general – fits – Introduction to FITS – File Object Tool Set – Google Project Hosting" *fits – File Information Tool Set* http://code.google.com/p/fits/wiki/general (accessed July 2012).

[308] More information on several of these tools (JHOVE, DROID, and the New Zealand Metadata Extractor) can be found within this Appendix, and at: "tools – fits – Tools Overview – File Information Tool Set – Google Project Hosting," *fits – File Information Tool Set* http://code.google.com/p/fits/wiki/tools (accessed July 2012).

[309] Simson Garfinkel, "Fiwalk," *DEEP: Digital Evaluation and Exploitation: Department of Computer Science, Naval Postgraduate School, Monterey, CA* https://domex.nps.edu/deep/Fiwalk.html (accessed July 2012).

[310] "Fiwalk – Forensics Wiki," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/Fiwalk (accessed July 2012).

[311] "Fiwalk," *DEEP: Digital Evaluation and Exploitation*.

17. Forensic Recovery of Evidence Device (FRED)

The Forensic Recovery of Evidence Device is not a single tool, but a line of computer systems designed specifically for the retrieval of forensic information and constructed by Digital Intelligence.[312] The FRED workstation features a wide variety of connection ports to allow the acquisition of data from different types of drives and devices, configured to read-only status to avoid any accidental information transfer, and a number of other special enhancements geared toward forensic processing. The FRED is only a hardware system, however; it needs to have installed a forensic software package, such as FTK or Encase, in order to conduct forensic analysis.

18. The Forensic Toolkit (FTK)

The Forensic Toolkit (FTK) is a commercial forensic analysis software package developed by AccessData. It enables a number of forensic processes, including the identification of duplicate and similar files, email analysis, volatile memory analysis, and multi-threaded forensic processing (making use of multiple processors in computers to speed forensic analysis).[313] FTK can process different kinds of disk image formats, including the Advanced Forensic Format (AFF) and several proprietary formats, and can parse filesystems including various NTFS, FAT, and EXT filesystems.[314]

19. FTK Imager

FTK Imager is a commercially licensed, free disk imaging software tool, developed for

---

[312] "FRED," *Digital Intelligence: mastering the science of digital forensics* http://www.digitalintelligence.com/products/fred/ (accessed July 2012).
[313] "FTK 4 Data Sheet," *Accessdata.com* http://accessdata.com/downloads/media/FTK_DataSheet_web.pdf (accessed July 2012).
[314] "Forensic Toolkit," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/FTK (accesed July 2012).

Windows OS by AccessData.[315] FTK Imager can be used to create a disk image of a target drive or volume. In addition, it will create an MD-5 or SHA-1 hash of the disk image, retained in a separate log file. This file also retains other technical metadata about the target drive and the imaging process.[316] It can also read disk images created in raw, SMART, ISO, AFF, and Encase formats, providing image and document previews, and hexadecimal editing for file carving and viewing drive slack space.[317]

20. ICA-Atom (International Council on Archives – Access to Memory)

ICA-Atom is an open-source, web-based software tool for archival description and access. Though the primary contractor and developer of the tool is Artefactual Systems, Inc., the ICA-Atom project is overseen by the International Council on Archives and has received funding from a variety of international organizations.[318] It is based on ICA archival description standards, making it potentially useful to archives around the world, while also allowing for the application of more specific national and regional descriptive standards. It can also serve as a federated search node, with multiple repositories feeding entries into a single interface.[319] ICA-Atom is free to use under an AGPL-3 software license.[320]

---

[315] "FTK Imager," *Digital Forensics Wiki* http://www.forensicswiki.org/wiki/FTK_Imager (accessed July 2012).

[316] John Jarocki, "Forensics 101: Acquiring an Image with FTK Imager," June 18, 2009, *SANS Computer Forensics and Incident Response* http://computer-forensics.sans.org/blog/2009/06/18/forensics-101-acquiring-an-image-with-ftk-imager/ (accessed July 2012).

[317] "FTK Imager," *Digital Forensics Wiki.*

[318] "About the ICA-Atom Project," *ICA-Atom: Open Source Archival Description Software* https://www.ica-atom.org/about/ (accessed July 2012).

[319] "What is ICA-Atom?," *ICA-Atom: Open Source Archival Description Software* https://www.ica-atom.org/doc/What_is_ICA-AtoM%3F (accessed July 2012).

[320] For more information on the AGPL-3 license, see: "GNU Affero General Public License," *GNU Operating System* http://www.gnu.org/licenses/agpl.html (accessed July 2012).

21. J-STOR/Harvard Object Validation Environment (JHOVE/JHOVE2)

The J-STOR/Harvard Object Validation Environment (JHOVE) "provides functions to perform format-specific identification, validation, and characterization of digital objects."[321] The tool is open source and distributed under a GNU Lesser General Public License, and while it can be used both through a command-line and GUI interfaces, it can also be integrated with other tools into automated services (such as FITS).[322] JHOVE was funded by a grant to JSTOR from the Andrew W. Mellon Foundation for the Electronic Archiving Initiative (now Portico).[323] More recently a new iteration of JHOVE has been developed with funding from the National Digital Information Infrastructure and Preservation Program (NDIIPP), as a collaboration between Portico, the California Digital Library, and Stanford University.[324] JHOVE2 improves on the performance of JHOVE: where JHOVE used iterative attempts to "match" a file against its knowledge of format types, JHOVE2 uses signature-based identification (via DROID and PRONOM) to more quickly and efficiently match format types to files, among a number of other enhancements.[325] JHOVE2 is available under an open-source BSD license.[326]

---

[321] "Home," *JHOVE-JSTOR/Harvard Object Validation Environment* http://hul.harvard.edu/jhove/ (accessed July 2012).

[322] "Tutorial: Using JHOVE," *JHOVE – JSTOR/Harvard Object Validation Environment* http://hul.harvard.edu/jhove/using.html (accessed July 2012). For information on the GNU Lesser General Public License, see: "GNU Lesser General Public License," *GNU Operating System* http://www.gnu.org/licenses/lgpl.html (accessed July 2012).

[323] "Home," *JHOVE – JSTOR/Harvard Object Validation Environment*. For information on Portico, see: "Our Organization – Portico," *Portico* http://www.portico.org/digital-preservation/about-us/our-organization (accessed July 2012).

[324] Stephen Abrams, Sheila Morrissey, and Tom Cramer, "What? So What?": The Next-Generation JHOVE2 Architecture for Format-Aware Characterization," Paper presented at the Fifth International Conference on Preservation of Digital Objects, British Library, 29-30 September 2008, https://bytebucket.org/jhove2/main/wiki/documents/Abrams_a70_pdf.pdf (accessed July 2012).

[325] "JHOVE2 Frequently Asked Question (FAQ)s," *jhove2 / main / wiki – Bitbucket* https://bitbucket.org/jhove2/main/wiki/JHOVE2_Frequently_Asked_Questions_(FAQ) (accessed July 2012).

[326] "The BSD 2-Clause License," *The Open Source Initiative* http://opensource.org/licenses/bsd-license.php (accessed July 2012).

22. KryoFlux

KryoFlux is a USB-based floppy disk controller, developed by the Software Preservation Society. Like other floppy disk controllers (such as CatWeasel), KryoFlux is designed to allow data to be read from otherwise obsolete floppy disks. However unlike controllers like the FC5025, which images at the sector level, the Kryoflux is designed to capture the raw magnetic flux transitions from the disk, advertised as "the lowest level possible."[327] The KryoFlux is format-agnostic, allowing it to copy a wide variety of different early media formats, exporting the information as a raw bitstream or into a number of common sector formats.[328] It also generates a summary log that tracks any errors in the imaging process. KryoFlux is operated through a GUI that is available for Mac OS, Windows, and Linux distributions.[329]

23. Metadata Encoding and Transmission Standard (METS)

The Metadata Encoding and Transmission Standard (METS) is an XML standard for packaging digital preservation metadata to facilitate the storage and exchange of complex digital objects between repositories.[330] The basic design consists of seven high-level elements with an extensive set of sub-elements, allowing for documentation of descriptive, technical, and administrative metadata, as well as documenting the relationships between associated digital objects through the collection of structural metadata.[331] This can be beneficial in the case of large collections of related digital

---

[327] "KryoFlux," *KryoFlux – USB Floppy Controller* http://www.kryoflux.com/ (accessed July 2012).
[328] Ibid.
[329] Ibid.
[330] *<METS> Metadata Encoding and Transmission Standard: Primer and Reference Manual*. Version 1.6 Revised (Washington, D.C.: Digital Library Federation, 2010), 15, http://www.loc.gov/standards/mets/METSPrimerRevised.pdf (accessed July 2012).
[331] Ibid., 18-20.

objects, or complex digital objects with many components. Documentation on the current METS schema is maintained by the Library of Congress.

24. Metadata Object Description Standard (MODS)

The Metadata Object Description Standard (MODS) is "a schema for a bibliographic element set that may be used for a variety of purposes, and particularly for library applications."[332] MODS is an 18-element schema, developed within the Library of Congress in 2002 as an alternative digital object cataloging element set to the Dublin Core element set. As such, it can be used to facilitate the representation of bibliographic records previously maintained in Machine-Readable Cataloging (MARC) form.[333] MODS can also be used to represent digital object metadata within a transmission schema such as METS, where a METS package contains individually defined MODS digital objects.[334]

25. New Zealand Metadata Extractor

The New Zealand Metadata Extraction Tool was developed by the National Library of New Zealand in 2003, and released as an open source tool in 2007. The Metadata Extractor collects information from a variety of different common file formats, and for those that it cannot read it will extract common fields such as size, date created, and

---

[332] "Metadata Object Description Schema: MODS (Library of Congress)," *Library of Congress* http://www.loc.gov/standards/mods/ (accessed July 2012).

[333] The Dublin Core metadata element set was born out of a joint meeting of the National Center for Supercomputing Applications (NCSA) and OCLC in Dublin, Ohio, in 1995. The 15-element set is a recognized standard by many national and international organizations, and is a part of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). "Metadata Basics," *Dublin Core Metadata Initiative* http://dublincore.org/metadata-basics/ (accessed July 2012).

[334] This is how MODS and METS are used together in SIPS prepared for the Carolina Digital Repository within the Curator's Workbench. See Appendix D, "Curator's Workbench," for more information.

filename.[335] It runs on a series of adapters, which read information from different parts of the file in order to extract different pieces of metadata. This metadata is compiled into an XML output file.[336] The Metadata extractor can accommodate single-file or batch processing, and can run on Windows or Linux, or as a component in an integrated software package such as FITS or Prometheus.[337]

26. Preservation Metadata: Implementation Strategies (PREMIS)

Preservation Metadata: Implementation Strategies (PREMIS) is a standard for representing digital preservation metadata in repository environments. PREMIS emerged in the form of a data dictionary, as one product of a working group sponsored by OCLC and the Research Libraries Group (RLG) from 2003-2005.[338] Other products include a PREMIS XML schema and ongoing activities around PREMIS maintained by the Library of Congress, with the data dictionary being most commonly referenced.[339] "The Data Dictionary defines preservation metadata that:

- Supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;
- Represents the information most preservation repositories need to know to preserve digital materials over the long-term;
- Emphasizes 'implementable metadata': rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated

[335] "Metadata Extraction Tool," *National Library of New Zealand* http://meta-extractor.sourceforge.net/ (accessed July 2012).

[336] "Metadata Extraction Tool Information Sheet," *National Library of New Zealand* http://meta-extractor.sourceforge.net/meta-extractor-info-sheet.pdf (accessed July 2012).

[337] For more information, see Appendix D, "File Identification Tool Set (FITS)," and Appendix D, "Prometheus Digital Preservation Workbench."

[338] Priscilla Caplan, *Understanding PREMIS*. (Washington, D.C.: Library of Congress Network Development and MARC Standards Office, 2009): 4, http://www.loc.gov/standards/premis/understanding-premis.pdf (accessed July 2012).

[339] The current version (2.2) of the *PREMIS Data Dictionary for Preservation Metadata* is available on the Library of Congress website at: http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

workflows; and

- Embodies technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc."[340]

The data model defined in the PREMIS Data Dictionary is comprised of five entities: *Rights, Objects, Agents, Events, and Intellectual Entities*.[341] One difficulty posed in implementing PREMIS is that it is rigorous to the degree that there are currently few tools which export directly to the PREMIS XML schema.[342] Some institutions, recognizing the value of PREMIS metadata elements to digital preservation, represent PREMIS elements in other schema where parallels may exist, such as METS.[343]

27. Prometheus Digital Preservation Workflow and Workbench

The Prometheus Digital Preservation Workflow can be considered as two separate components: the mini-jukebox and the Digital Preservation Workbench (DPW) software application. The mini-jukebox is a piece of hardware consisting of storage and a variety of different drives and connection ports. This is the physical capture station for born-digital content that is imaged from media carriers such as hard drives, floppy disks, or CDs and DVDs.[344] The Digital Preservation Workbench is a web-based software

---

[340] PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata.* Version 2.2 (Washington, D.C.: Library of Congress, July 2012): 1, http://www.loc.gov/standards/premis/v2/premis-2-2.pdf (accessed July 2012).

[341] Caplan, *Understanding PREMIS*, 8-9.

[342] A list of tools that to create PREMIS XML is available at: http://www.loc.gov/standards/premis/tools_for_premis.php.

[343] For example, the Carolina Digital Repository represents PREMIS event metadata within Curator's Workbench METS XML files. See University Archives, University of North Carolina at Chapel Hill born-digital preservation workflow on page 64 of this work.

[344] Elford et al., "Media Matters," 6, http://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf (accessed June 2012). On the technical specifications of the mini-jukebox, see the *Component Installation Guide* http://sourceforge.net/projects/prometheus-digi/files/Documentation/Prometheus_Component_Installation_Guide_SF-v1.pdf/download (accessed July 2012).

application for the cataloging and description of born-digital content, developed by the National Library of Australia and introduced in 2008.[345] Through the DPW, a fully developed metadata record is created for the digital object, associating it with a holding in institution's catalog. The DPW also facilitates the ingest of digital objects into the NLA digital repository Digital Object Storage System (DOSS), and includes built-in, semi-automated tools for media imaging and file format identification, verification, and validation.[346] The tools used are largely open source, and build into a Java-based web architecture. The Prometheus DPW is available for download and is subject to a GNU General Public License.[347]

28. rsync

rsync is a data copying software application for Windows and Linux, developed in 1996.[348] Where other software applications such as dd or FTK Imager create a bit-level disk image of the target drive or volume, rsync does not. Rather, rsync is able to consistently transfer data from one location to another using a transfer and synchronization method called delta encoding.[349] Delta encoding allows for accurate and efficient transfer of data and minimizes storage needs, but it does not reflect the same bit-level replication that disk imaging does. Given the capability of forensic imaging formats

---

[345]"Prometheus – About," *Prometheus Digital Preservation Workbench* http://mediapedia.nla.gov.au/prometheus/about.html (accessed July 2012).

[346]Tools include JHOVE, DROID, and the New Zealand Metadata Extraction Tool. For information on these tools, see Appendix D: Digital Preservation Tools and Technologies.

[347]"Prometheus Digi Pres Workbench," *SourceForge.net* http://sourceforge.net/projects/prometheus-digi/ (accessed July 2012). For more information on the GNU General Public License, see: "General Public License," *GNU Operating System* http://www.gnu.org/copyleft/gpl.html (accessed July 2012).

[348]"rsync," *Wikipedia – the free online encyclopedia* http://en.wikipedia.org/wiki/Rsync (accessed July 2012).

[349]Ibid. "Delta encoding," *Wikipedia – the free online encyclopedia* http://en.wikipedia.org/wiki/Delta_encoding (accessed July 2012).

such as AFF to compress disk images to more manageable size, use of the rsync tool is limited in forensic applications.

29. The Sleuth Kit (TSK)

The Sleuth Kit (TSK) "is a library and collection of command line tools that allow you to investigate disk images."[350] The library allows the collection of tools to be integrated into other digital forensic software packages, while the tools can be used individually through the command line and a graphical interface, the Autopsy Forensic Browser.[351] The software tools are configured for use with Windows, Linux, and other Unix distributions, and have been used on a variety of different target operating systems and computing platforms.[352] The primary uses of the tools included in TSK are for file system analysis—including creating timelines of filesystem activity and recovery of deleted or hidden files and directories—and a plug-in framework that allows for the creation of more customized tools for forensic analysis.[353] TSK was developed by Brian Carrier. Different portions of the source code are distributed under several different licenses.[354]

30. Write-blockers

A write blocker allows information to be read from a drive while preventing a host computer used for processing or forensic analysis from writing data to that drive. If that

---

[350] Brian Carrier, "Overview," *The Sleuth Kit* http://www.sleuthkit.org/sleuthkit/ (accessed July 2012).
[351] Carrier, "Description," *Autopsy Forensic Browser* http://www.sleuthkit.org/autopsy/desc.php/index.php (accessed July 2012).
[352] Brian Carrier, "Documents," *The Sleuth Kit* http://www.sleuthkit.org/sleuthkit/docs.php (accessed July 2012).
[353]Brian Carrier, "File Systems," *The Sleuth Kit* http://www.sleuthkit.org/sleuthkit/desc.php; "Plug-In Frameworks," *The Sleuth Kit* http://www.sleuthkit.org/sleuthkit/framework.php (accessed July 2012).
[354]Brian Carrier, "Licenses," *The Sleuth Kit* http://www.sleuthkit.org/sleuthkit/licenses.php (accessed July 2012).

drive is the target of forensic analysis, accessing data to be read can actually overwrite or alter contextual information on the drive.[355] Write blockers can be either hardware or software based. Software-based write blockers may be operating system dependent; hardware-based write blockers (which must be physically interposed between the target drive and the host computer) are designed to be software-agnostic.[356] Other forms of write-blocking are internal to a particular type of device: CDs and DVDs are generally "write-once, read-many" (WORM) storage devices that do not require write-blocking; also most 3.5 inch floppy disks were developed with a built-in write-blocking mechanism that is engaged through a small tab on the back of the disk.

[355] "Write Blockers – Forensics Wiki," *The Forensics Wiki* http://www.forensicswiki.org/wiki/Write_Blockers (accessed July 2012).
[356] Ibid.