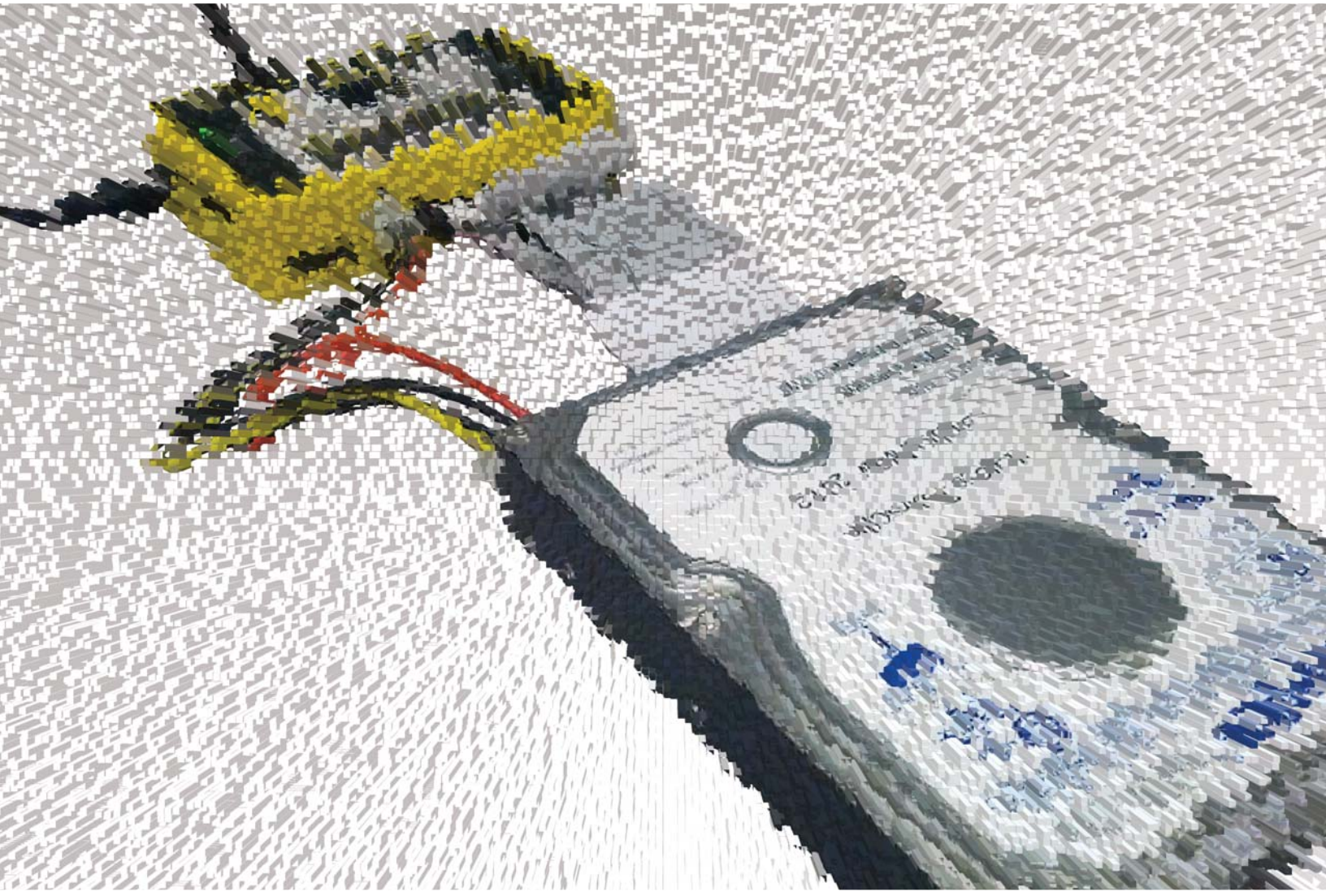
From Bitstreams to Heritage:

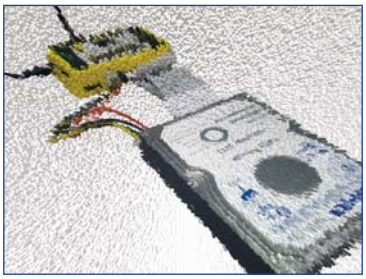
Putting Digital Forensics into Practice in Collecting Institutions



Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff

A Product of the BitCurator Project

September 30, 2013

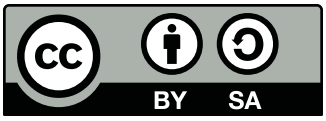


From Bitstreams to Heritage:

Putting Digital Forensics into Practice in Collecting Institutions

Christopher A. Lee, Kam Woods,
Matthew Kirschenbaum,
and Alexandra Chassanoff

A Product of the BitCurator Project
September 30, 2013



"[M]uch will be lost, but even when disks become unreadable, they may well contain information which is ultimately recoverable. Within the next ten years, a small and elite band of e-paleographers will emerge who will recover data signal by signal."

--R.J. Morris, 1998[1]

1. Introduction

This paper examines the application of digital forensics methods to materials in collecting institutions – particularly libraries, archives and museums. It discusses motivations, challenges, and emerging strategies for the use of these technologies and workflows. It is a product of the BitCurator project.

The BitCurator project began on October 1, 2011, through funding from the Andrew W. Mellon Foundation. BitCurator is an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions. It is led by the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill and the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland, and involves contributors from several other institutions. Two groups of external partners are contributing to this process: a Professional Expert Panel (PEP) of individuals who are at various levels of implementing digital forensics tools and methods in their collecting institution contexts, and a Development Advisory Group (DAG) of individuals who have significant experience with software development.[2]

This paper is a product of phase one of BitCurator (October 1, 2011 – September 30, 2013). The second phase of the project (October 1, 2013 – September 29, 2014) continues the development of the BitCurator environment, along with expanded professional engagement and community outreach activities.

1 R.J. Morris, "Electronic Documents and the History of the Late Twentieth Century: Black Holes or Warehouses?" in *History and Electronic Artefacts*, ed. Edward Higgs (Oxford: Clarendon Press, 1998), 33.

2 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.

2. Motivation

Materials with cultural, administrative, scholarly and personal value are increasingly "born digital." Collecting institutions—libraries, archives and museums (LAMs)—have unprecedented opportunities to acquire and preserve traces of human and associated machine activity through access to both consciously created electronic records (e.g. word processing files, databases, spreadsheets, email, multimedia productions, social media) and various other inscriptions that are the result of interactions with a computer (e.g. system logs, configuration files, filesystem metadata). Likewise, researchers have unprecedented opportunities to discover and learn from those traces. In order to fully realize these opportunities, LAMs must be able to extract digital materials from their storage or transfer media in ways that reflect the metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data.

LAMs are increasingly called upon to transfer born-digital materials stored on removable media into more sustainable preservation environments. This can involve media already in their holdings (e.g. disks stored in boxes along with paper materials), as well as materials that institutions are acquiring from individual donors or other producers—sometimes including entire computers.

Computers are human artifacts, built and engineered to implement a mathematically-based environment for creating, manipulating, disseminating, and storing information in symbolic form. The literature on digital collections thus tends to place a great emphasis on the "virtual" (i.e. intangible) nature of such data. Though computer systems maintain "an illusion of immateriality by detecting error and correcting it,"[3] it is essential to recognize that digital objects are created and perpetuated through physical things (e.g. charged magnetic particles, pulses of light). This distinct mode of materiality brings challenges, because data must be read from specific artifacts, which can become damaged or obsolete. However, the materiality of digital objects also brings unprecedented opportunities for description, interpretation and use.[4]

Digital materials can be considered and encountered at multiple levels of representation, ranging from aggregations of records down to bits as physically inscribed on a storage medium; each level of representation can provide distinct contributions to the informational and evidential value of the materials.[5] There is a substantial body of information within the underlying data structures of computer systems that can often be discovered or recovered, revealing new types of records or essential metadata associated with existing record types.

Digital forensics has its origins in law enforcement, both criminalistics and legal discovery. It is associated with the branch of forensic science known as "trace evidence," which owes its origins to the pioneering work of the French police investigator Edmond Locard.[6]

3 Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge, MA: MIT Press, 2008).

4 John Lavagnino, "The Analytical Bibliography of Electronic Texts" (paper presented at the Joint Annual Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Bergen, Norway, 1996).

5 Christopher A. Lee, "Digital Curation as Communication Mediation," in *Handbook of Technical Communication*, ed. Alexander Mehler, Laurent Romary, and Dafydd Gibbon (Berlin: Mouton De Gruyter, 2012), 507-530.

6 Kirschenbaum, *Mechanisms*, 48.

Figure 1: Legacy removable and fixed digital media.

Locard's "exchange principle"—often summarized as "every contact leaves a trace"—has become a basic precept of the forensic sciences, the consequence of just the kind of material interactions we acknowledge above. Procedures and tools for acquiring and validating data from physical media are well established in the field of digital forensics. Their recognition and adoption within LAMs is a more recent phenomenon. Two particular streams of activity show great promise for informing the practices of collecting institutions. First, the number of collecting institutions exploring the application of digital forensics to the acquisition of digital materials is growing rapidly. Second, there is a rich and growing body of open source tools that can be used to process, manage and disseminate forensically acquired data. While the primary target for many of these tools and methods is the law enforcement community, there is great potential for connecting these two streams of activity in order to support the work of collecting institutions.

Many digital forensics applications and strategies can aid LAMs in their work, particularly by advancing three fundamental archival principles: provenance, original order and chain of custody.[7] Provenance "consists of the social and technical processes of the records' inscription, transmission, contextualization, and interpretation which account for its existence, characteristics, and continuing history."[8] According to the principle of provenance, records from a common origin or source should be managed together as an aggregate unit and should not be arbitrarily intermingled with records from other origins or sources. In digital environments, it can be important to consider provenance at levels of granularity finer than an entire record, such as why a specific data element appears within a dataset and where specifically the data

7 Christopher A. Lee, "Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision," *Comma* (forthcoming).

8 Tom Nesmith, "Still Fuzzy, but More Accurate: Some Thoughts on the 'Ghosts' of Archival Theory," *Archivaria* 47 (1999): 146.

element was generated.[9] It is also important to include technical components in one's notion of provenance, such as system configuration information.[10]

Closely related to provenance is the principle of original order, which indicates that archivists should organize and manage records in ways that reflect their arrangement within the creation environment. There are compelling arguments for retaining original order in a digital environment. Even if this order is messy and idiosyncratic, it conveys meaningful information about the recordkeeping context.

The chain of custody is the "succession of offices or persons who have held materials from the moment they were created."[11] Ideal recordkeeping systems would provide "an unblemished line of responsible custody"[12] through control, documentation, and accounting for all states of a record and changes of state e.g., movement from one storage environment to another, or transformation from one file format to another, throughout its existence. Such a system would therefore cover the point of creation, important interactions with the record, and (when appropriate) destruction.

Professionals in collecting institutions must increasingly apply their professional principles to collections composed—in whole or in part—of born-digital materials. Among other activities, this includes moving materials that are stored on removable media into more sustainable preservation environments. This can involve media that are already in their holdings (e.g. disks stored in boxes along with paper materials), as well as being acquired for the first time from individual donors or other producers.

Forensic methods identify, capture and retain various forms of contextual information, which can be vital for users making meaningful use of digital materials.[13] Two important types of metadata that forensics tools can extract from digital materials are metadata related to document creation activities, including versioning, embedded media, software dependencies, and rights information; and logs of user activity recorded by the operating system and other software.

The authenticity of a handwritten document often depends upon the ink that was used to write it, the paper it was written on, and other artifactual characteristics. For digital documents, the environmental context from the original device or platform can be used to more thoroughly establish ownership and methods of production and use. Access to the full range of data on a disk can ensure that the record of provenance includes not only curatorial events within a repository but also events, actions and processes that have occurred between initial creation of a particular digital object and the changes or user/system activities related to the object over its lifetime within a particular digital ecosystem.[14]

Extraction of basic technical metadata (such as timestamps) from file systems can provide a foundation on which one can establish a "ground truth" for content and structure on a device.

---

9 Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan, "Why and Where: A Characterization of Data Provenance," in *Database Theory - ICDT 2001: 8th International Conference, London, UK, January 2001. Proceedings*, ed. Jan van den Bussche and Victor Vianu (Berlin: Springer, 2001), 316–30.

10 Maria Guercio, "Archival Theory and the Principle of Provenance for Current Records: Their Impact on Arranging and Inventorying Electronic Records," in *The Principle of Provenance: Report from the First Stockholm Conference on Archival Theory and the Principle of Provenance, 2-3 September 1993*, ed. Kerstin Abukhanfusa and Jan Sydbeck (Stockholm: Swedish National Archives, 1994), 82.

11 Richard Pearce-Moses, *Glossary of Archival and Records Terminology* (Chicago, IL: Society of American Archivists, 2005), 67.

12 Hilary Jenkinson, *A Manual of Archive Administration: Including the Problems of War Archives and Archive Making* (Oxford: Clarendon Press, 1922), 11.

13 Christopher A. Lee, "A Framework for Contextual Information in Digital Collections," *Journal of Documentation* 67, no.1 (2011): 95-143.

14 Kam Woods and Geoffrey Brown, "From Imaging to Access - Effective Preservation of Legacy Removable Media," in *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations: Final Program and Proceedings* (Springfield, VA: Society for Imaging Science and Technology, 2009), 213-218; Kam Woods, Christopher A. Lee, and Simson Garfinkel, "Extending Digital Repository Architectures to Support Disk Image Preservation and Access," in *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (New York, NY: ACM Press, 2011), 57-66.

They can provide significant support for assessment and preservation activities. File names may include evidence of the hierarchical structure of data within the original file system, and can often include information that should be considered private and assessed accordingly. Timestamps may help identify files that have been copied from alternate locations. Although collecting institutions often record timestamps as part of the provenance record, they often omit access and modification timestamps. Likewise, file system permissions on the host system (along with a record of users and groups with access to particular files) can assist in creating more complete documentation of creation and use contexts.

Metadata associated with user logins or user accounts associated with specific data objects also can be valuable. Examples include the following:

- System information located in common storage areas such as Microsoft Windows registry hives. Such information can assist in verifying unique user identifiers, removable media that have been used on the host computer, and application metadata.
- Extraction of user information from operating system and application-specific data stores such as SQLite databases, which are typically used to record and maintain user activity, cache information from network access, and store private data.

3. Activities to Date

Recovery of data from physical media has been a concern of LAM professionals for several decades. The application of digital forensics tools and methods is a more recent phenomenon, but one that has grown rapidly in the past decade.

Research, Development and Professional Literature

Margaret Hedstrom argued in 1984 that archivists could face "potential obliteration of significant portions of the historical records" if they did not learn more about how computers, including storage media, function; and she proceeded to explain computer systems of the time to an archival audience.[15] There have been various publications in the LAM and records management literature about the viability of particular storage media (e.g. laser disk, magnetic tape) for preservation purposes. However, until quite recently, there had been relatively little writing about the recovery of data from the media in order to move them into preservation environments.

An important contribution came in 1999, in a report by Seamus Ross and Ann Gow that discussed the potential relevance of advances in data recovery and digital forensics to collecting institutions.[16] This was followed by a few papers and articles about the skills and methods used to recover data from old media, including reports on experiences from the National Library of Australia[17] and Cornell University's File Format and Media Migration Service (initiated in 2004 but no longer active).[18] Lucie Paquet published an article in 2000 about working with personal electronic records, in which she stated, "When I visit my donors, I bring an external disk drive with me in order to copy electronic records of historical value so that I can take them back to the National Archives of Canada."[19] Paquet's activities were laudable and valuable; she was reporting on actions that relatively few archivists had yet broached in their own work. However, there was no indication at that point that she was using forensic methods to ensure the integrity of the materials.

The professional landscape has changed quite dramatically since then. In the past eight years, a number of authors have investigated the use of forensic tools and techniques to care for digital collections in libraries and archives.[20]

15 Margaret Hedstrom, *Archives and Manuscripts: Machine-Readable Records* (Chicago, IL: Society of American Archivists, 1984), 7.

16 Seamus Ross and Ann Gow, "Digital Archaeology: Rescuing Neglected and Damaged Data Resources" (London: British Library, 1999).

17 See e.g. Deborah Woodyard, "Farewell My Floppy: A Strategy for Migration of Digital Information," National Library of Australia, 1997; Deborah Woodyard, "Data Recovery and Providing Access to Digital Manuscripts" (paper presented at the Information Online 2001 Conference, Sydney, Australia, January 16-18, 2001).

18 Richard Entlich and Ellie Buckley, "Digging up Bits of the Past: Hands-on with Obsolescence," *RLG DigiNews* 10, no. 5 (2006).

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

20 A thought leader in this regard has been Jeremy Leighton John at the British Library. See e.g. Jeremy Leighton John, "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools" (paper presented at iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29-30, 2008).

The Prometheus[21] and PERPOS[22] projects have developed software for data extraction, focusing on needs of specific collecting contexts. The Personal Archives Accessible in Digital Media (PARADIGM) project investigated "issues involved in preserving digital private papers through gaining practical experience in accessioning and ingesting digital private papers into digital repositories, and processing these in line with archival and digital preservation requirements." PARADIGM's most visible outcome was the "Workbook on Digital Private Papers."[23] The Digital Lives project, led by Jeremy Leighton John at the British Library, investigated "personal digital collections and their relationship with research repositories,"[24] and it generated a major report that included discussions of digital forensics.

In 2008, MITH initiated a collaboration with the Ransom Center through an NEH-funded Digital Humanities Start-Up grant directed by Matthew Kirschenbaum that brought together practitioners at the University of Maryland, the Ransom Center, and Emory University for site visits and knowledge exchange. They generated a white paper based on a series of site visits and meetings of those working with the born-digital components of three significant literary collections.[25] A project funded by the Andrew W. Mellon Foundation called "Computer Forensics and Born-Digital Content in Cultural Heritage Collections" hosted a symposium and generated a report published by the Council on Library and Information Resources (CLIR), both of which were major contributions and milestones in the application of digital forensics in LAMs.[26]

Christopher (Cal) Lee and Kam Woods administered "Curation of a Forensic Data Collection for Education," a sub-grant of an NSF-funded project led by Simson Garfinkel of the Naval Postgraduate School. They enhanced, packaged, and distributed a collection of data that represents a realistic scenario (e.g. traces of computer use across a given timespan, multiple disk images relevant to a particular set of historical activities, and in which numerous end-user applications are installed and used), while also being appropriate for students to use in support of digital forensics education.[27] Primary focus areas were annotation, scenarios, exercises, answer keys and other forms of data that can further enhance access and use of the disk images. They also investigated strategies for ensuring that the data sets would remain available and useful beyond the life of the project.

In 2011, the Forensic Information in Digital Objects (FIDO) project addressed "the application of digital forensics to support the curation and preservation of digital information held on computer systems and digital media." FIDO investigated and documented the use of various digital forensics tools, with a particular focus on university archives.[28]

---

21 Douglas Elford, 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).

22 William E. Underwood and Sandra L. Laib, "PERPOS: An Electronic Records Repository and Archival Processing System" (paper presented at the International Symposium on Digital Curation (DigCCurr 2007), Chapel Hill, NC, April 18-20, 2007); William Underwood, Marlit Hayslett, Sheila Isbell, Sandra Laib, Scott Sherrill, and Matthew Underwood, "Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report," Technical Report ITTL/CSITD 09-05 (October 2009).

23 Susan Thomas, Renhart Gittens, Janette Martin, and Fran Baker, "Workbook on Digital Private Papers" (Paradigm Project, 2007), http://www.paradigm.ac.uk/workbook/introduction/index.html; 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 (2006): 29-56.

24 Pete Williams, Katrina Dean, Ian Rowlands, and Jeremy Leighton John, "Digital Lives: Report of Interviews with the Creators of Personal Digital Collections," *Ariadne* 55 (2008); Jeremy Leighton John, Ian Rowlands, Peter Williams, and Katrina Dean, *Digital Lives: Personal Digital Archives for the 21st Century >> An Initial Synthesis*, Version 0.2, March 3, 2010, http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf.

25 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).

26 Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine, *Digital Forensics and Born-Digital Content in Cultural Heritage Collections* (Washington, DC: Council on Library and Information Resources, 2010).

27 Woods, Lee, and Garfinkel, "Extending Digital Repository Architectures," 57-66.

28 Gareth Knight, "The Forensic Curator: Digital Forensics as a Solution to Addressing the Curatorial Challenges Posed by Personal Digital Archives," *International Journal of Digital Curation* 7, no. 2 (2012): 40-63.

The Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) project developed a framework for the stewardship of born-digital materials that includes the incorporation of digital forensics methods.[29] The Digital Records Forensics project at the University of British Columbia also articulated a variety of connections between the concepts of digital forensics and archival science.[30]

Finally, there has been a significant strain of new research on digital media. This has originated largely in the humanities under the broad rubrics of "platform studies," "software studies," "critical code studies," and "media archaeology."[31] While the particulars of these labels and affiliations vary and should not be understood as merely mutually interchangeable, work identified in these ways has important shared characteristics that bear directly on our understanding of the collecting activities now being undertaken by LAMs, and forensic methods in particular. The research contributing to these fields generally:

- assumes the computer and computational processes are material in nature, and thus subject to documentary and historical forms of understanding;
- is technically rigorous and acknowledges the material particulars of media and computation as worthy of critical investigation;
- understands the particular constraints of software, code, and platform as generative for studying the processes and products of digital culture;
- cultivates and actively seeks to refine an archival record for digital culture; and
- understands the activity of archiving itself in new and capacious ways, that include such techniques as crowd-sourcing, hacktivism, restoration and retro-computing, and citizen archivists.

Matthew Kirschenbaum's *Mechanisms: New Media and the Forensic Imagination* is emblematic in this regard: the book integrates techniques from digital forensics with long-standing practices and precepts from the humanistic fields of bibliography and textual scholarship to examine individual digital objects and reveal new histories of their production. While Kirschenbaum's methods are recognizable to practitioners versed in the digital forensics techniques we discuss here, his conclusions and analyses are aimed at scholars with the intent of demonstrating that digital objects are always embedded in material histories that are recoverable through specific tools and technological procedures.

Building Institutional Capacity

In recent years, a variety of collecting institutions have made efforts to incorporate digital forensics activities into their workflows. While the current list of such institutions is rapidly changing and expanding, some of the early leaders of note were the Bodleian Library (Oxford), the British Library, Emory University, King's College London, the National Library of Australia, the New York Public Library, Stanford University, and Yale University.

Not all of this capacity has been developed in traditional collecting institutions. For example, at the Maryland Institute for Technology in the Humanities, Matthew Kirschenbaum and Doug Reside have helped to build a collection of vintage and antiquarian hardware. This was largely drawn from their personal collections and cast-offs from the university community. MITH now owns a broad array of obsolete systems used for research and instruction. In May 2007, MITH

29 AIMS Working Group, *AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship*, 2012.

30 Luciana Duranti and Barbara Endicott-Popovsky, "Digital Records Forensics: A New Science and Academic Program for Forensic Readiness," *Journal of Digital Forensics, Security and Law* 5, no. 2 (2010).

31 See Nick Montfort and Ian Bogost, *Racing the Beam: The Atari Video Computer System* (Cambridge, MA: MIT Press, 2009); Wendy Chun, *Programmed Visions: Software and Memory* (Cambridge, MA: MIT Press, 2011); Erkki Huhtamo and Jussi Parikka, ed., *Media Archaeology: Approaches, Applications and Implications* (Berkeley, CA: University of California Press, 2011); Jussi Parikka, *What Is Media Archaeology?* (Cambridge, UK: Polity Press, 2012); Nick Montfort, Patsy Baudoin, John Bell, Ian Bogost, Jeremy Douglass, Mark C. Marino, Michael Mateas, et al. *10 Print Chr$(205.5+Rnd(1));:Goto 10* (Cambridge, MA: MIT Press, 2013); Wolfgang Ernst and Jussi Parikka, *Digital Memory and the Archive: Electronic Mediations* (Minneapolis, MN: University of Minnesota Press, 2013); Lev Manovich, *Software Takes Command* (New York: Bloomsbury, 2013).

Figure 2: The MITH facility at the University of Maryland, pictured during the Personal Digital Archiving 2013 conference.

acquired a large collection of vintage hardware, software and other archival material from Deena Larsen, an author who has been an active member of the creative electronic writing community since its inception in the mid-1980s.

Two other related initiatives have been SWAT and Jump In. The SWAT (software and workstations for antiquated technology) initiative has been led by Ricky Erway at OCLC.[32] It aims to match up institutions that do not have the capability to recover data from particular media with institutions that do have such capabilities. For example, Archive A with a collection of legacy Iomega Zip disks could send them to Archive B that has the capacity to read and copy data from the disks. Archive B could create disk images and send them (along with basic metadata and documentation of the process) to Archive A. The Jump In initiative has taken place within the Manuscript Repositories Section of the Society of American Archivists (SAA). It has been a catalyst for several archives to take initial steps in addressing their born-digital holdings. Specifically, participants promised to:

- Locate computer media in any physical form.
- Record the location, inventory number, type of physical medium, and any identifying information found on labels or media such as creator, title, description of contents, and dates. If no identifying information exists, indicate this.
- Record anything that is known about the hardware, operating system, and software used to create the files.
- Count the number of each media type, calculate the total maximum amount of data stored in each medium, and then calculate the overall total for the collection.[33]

Participants submitted essays describing what they had done, along with photographs of themselves and the media that they surveyed. Twenty-three institutions participated in Jump In, and their essays can be found on the Manuscript Repositories Section web site. Several of the participants reported on their work at the Manuscript Repositories meeting at the SAA Annual Meeting in New Orleans on August 16, 2013.

Professional Events

On February 9-11, 2009, the Digital Lives project hosted the first Digital Lives conference at the British Library in London. At that event, Simson Garfinkel made the case for the relevance of digital forensics methods to recover traces of individuals' online activities.[34] Just a few weeks later, on March 31, 2009, in association with DigCCurr 2009, Lee organized a symposium with Richard Szary (UNC-CH) and Tom Hyry (then at Yale University) called "Stewardship of E-Manuscripts: Advancing a Shared Agenda." In this symposium, an invited set of leaders in this field (coming from Australia, Austria, the United Kingdom, and the United States)

32 Ricky Erway, "Swatting the Long Tail of Digital Media: A Call for Collaboration" (Dublin, OH: OCLC Research, 2012).

33 Jump in Initiative, http://www2.archivists.org/groups/manuscript-repositories-section/jump-in-initiative. This set of steps was inspired by Ricky Erway, "You've Got to Walk Before you Can Run: First Steps for Managing Born-Digital Content Received on Physical Media" (Dublin, OH: OCLC Research, 2012).

34 Simson Garfinkel and David Cox, "Finding and Archiving the Internet Footprint" (paper presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, UK, February 9-11, 2009).

shared strategies for and experience with the selection, acquisition, arrangement, description, preservation, and access to personal materials in digital form.[35] This included participation of leading experts on the application of digital forensics techniques to the acquisition of digital collections.

On May 14-15, 2010, MITH hosted an invitational meeting funded by the Andrew W. Mellon Foundation in support of the project and report discussed earlier, entitled *Computer Forensics and Born-Digital Content in Cultural Heritage Collections*, which was published by the Council on Library and Information Resources in December 2010. The co-authors of the report were also the co-organizers of the meeting: Matthew Kirschenbaum (Maryland), Richard Ovenden (Bodleian), and Gabriela Redwine (Harry Ransom Center). The agenda included talks from LAM professionals, scholars and digital forensics experts. The event and its discussions significantly advanced the connections between digital forensics and the activities of LAMs.

On June 28, 2011, the Digital Preservation Coalition (DPC) in the UK sponsored an event called Digital Forensics for Preservation in Oxford. Speakers addressed the nature of the problem, e-discovery and sense-making, digital forensics tools, mobile forensics, lab setup experiences, and the "practical and reasonable limits of forensics." The DPC also commissioned Jeremy Leighton John to write a Technology Watch Report on "Digital Forensics and Preservation," which was published in November 2012.[36]

A conference called "The Memory of the World in the Digital Age: Digitization and Preservation" was held on September 26-28, 2012 in Vancouver, British Columbia, Canada, by the Memory of the World Program of the United Nations Educational, Scientific and Cultural Organization (UNESCO) and the University of British Columbia in collaboration with the University of Toronto. There were more than 500 participants from a wide diversity of countries. The conference included numerous talks and sessions related to digital forensics,[37] including a sub-conference that constituted the Seventh International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE). This allowed for stimulating interactions between digital forensics experts and LAM professionals. SADFE had been held six times previously, but never in coordination with an event focused on cultural heritage. The call for papers for the next year's SADFE—to be held in Hong Kong on November 21-22, 2013—included even more explicit references to archives, records management and preservation considerations.[38] Lee is serving as the program co-chair for SADFE 2013, along with Adel S. Elmaghraby from the University of Louisville.

More recently, there have been two hackathons focused on applying forensics to digital collections: one in Copenhagen on May 15-17, 2013 (pictured in Figure 3) and one in Chapel Hill on June 3-5, 2013 (shown in Figure 4). The hackathon model has proved to be quite successful in bringing together practitioners and software developers, identifying existing technical gaps through real-world issues encountered within collecting institutions. These events produce a wide range of "serendipitous coding," in which existing tools are adapted to address novel use cases, and the intellectual seeds of longer-term development projects and intellectual collaborations are planted.

Both hackathons were conducted in partnership with the Open Planets Foundation. As part of its commitment to supporting the digital preservation community and the emerging preservation and forensics community both within the United States and internationally, the BitCurator project introduced new software technologies at both of these events and benefitted from feedback from participating members.

35 Stewardship of E-Manuscripts: Advancing a Shared Agenda, March 31, 2009, http://ils.unc.edu/callee/emanuscripts-stewardship/.

36 Jeremy Leighton John, "Digital Forensics and Preservation," DPC Technology Watch Report 12-03 (Digital Preservation Coalition, November 2012), http://www.dpconline.org/component/docman/doc_download/810-dpctw12-03.pdf.

37 The full proceedings are available at: http://www.unesco.org/new/en/communication-and-information/events/calendar-of-events/events-websites/the-memory-of-the-world-in-the-digital-age-digitization-and-preservation/

38 http://conf.ncku.edu.tw/sadfe/sadfe13/cfp.html

Figure 3: 2013 OPF Hackathon in Copenhagen.

Figure 4 - Hackathon in Chapel Hill

Affirmation of the Importance of Forensics to LAMs

A product of the Memory of the World in the Digital Age conference discussed above was the "UNESCO/UBC Vancouver Declaration." The document urges the UNESCO Secretariat to "encourage engagement of cultural heritage professionals knowledgeable about digital forensics concepts, methods and tools in order to ensure capture and reliable preservation of authentic, contextualized and meaningful information, and appropriate mediation of access to the information."[39]

On July 23, 2013, the National Digital Stewardship Alliance (NDSA) announced the National Agenda for Digital Stewardship 2014, which is designed to "to provide funders and executive decision-makers insight into emerging technological trends, gaps in digital stewardship capacity, and key areas for funding, research and development to ensure that today's valuable digital content remains accessible and comprehensible in the future." The agenda includes the following:

> As more digital materials are selected for long-term digital preservation, the need to integrate digital forensics tools into production workflows for collections becomes increasingly important. This will require identifying the boundaries between technical infrastructure development and organizational policies, and where there is tension that creates issues for providing access or pursuing work that reduces tension whether it be new or refined policies or services and tools development. Integration of these tools can build on exploratory work using digital forensics. Tools Currently under development can be leveraged and workflows can be implemented. Aside from the need for tools and workflow developments, there are also important opportunities for organizations to share resources in order to tackle these issues. In this respect, pioneering new organizational models for centers of stewardship, such as SWAT sites, can help to support the development of centers of excellence that help to scale up this kind of activity.
>
> There is a clear need to move the basic research in digital forensics tools from research to implementation in production workflows for organizations. This would require investment in scaling up tools and creating collaborative models for sharing resources to make this work possible. The digital preservation community would also benefit from a shared space for exchanging knowledge around how forensics tools are being integrated into production preservation activities.[40]

The *Signal* is a newsletter-style blog concerning digital preservation. It is produced by the Library of Congress. It receives considerable attention from professionals working in the digital preservation arena, and has given significant treatment to digital forensics issues over the past three years.[41]

---

39 "UNESCO/UBC Vancouver Declaration: The Memory of the World in the Digital Age: Digitization and Preservation," Vancouver, British Columbia, Canada, September 28, 2012.

40 http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf

41 See e.g. Leslie Johnston, "Digital Forensics and Digital Preservation," June 7, 2011, http://blogs.loc.gov/digitalpreservation/2011/06/digital-forensics-and-digital-preservation/; Martha Anderson, "F is for Forensics," November 30, 2011, http://blogs.loc.gov/digitalpreservation/2011/11/f-is-for-forensics/; Bradley Daigle, Matthew Kirschenbaum, and Christopher Lee, "Bit by Bit: Recent Projects on Digital Forensics for Collecting Institutions," January 24, 2012, http://blogs.loc.gov/digitalpreservation/2012/01/bit-by-bit-recent-projects-on-digital-forensics-for-collecting-institutions/; Bill LeFurgy, "Floppy Disks are Dead, Long Live Floppy Disks," April 11, 2012, http://blogs.loc.gov/digitalpreservation/2012/04/floppy-disks-are-dead-long-live-floppy-disks/; Trevor Owens, "Life-Saving: The National Software Reference Library," May 4, 2012, http://blogs.loc.gov/digitalpreservation/2012/05/life-saving-the-national-software-reference-library/; Susan Manus, "Digging Up the Recent Past: An Interview with Doug Reside," August 14, 2012, http://blogs.loc.gov/digitalpreservation/2012/08/digging-up-the-recent-past-an-interview-with-doug-reside/; Susan Manus, "The Born Digital In the Archives: One Curator's Experience," August 29, 2012, http://blogs.loc.gov/digitalpreservation/2012/08/the-born-digital-in-the-archives-one-curators-experience/; Jose Padilla, "Digital Forensic Perspective Helps Cultural Heritage Institutions Meet Deep Challenges," February 7, 2013, http://blogs.loc.gov/digitalpreservation/2013/02/digital-forensic-perspective-helps-cultural-heritage-institutions-meet-deep-challenges/; Trevor Owens, "Born Digital Archival Materials at NYPL: An Interview with Donald Mennerich," April 22, 2013; Trevor Owens, "What's a Nice English Professor Like You Doing in a Place Like This: An Interview with Matthew Kirschenbaum," August 12, 2013, http://blogs.loc.gov/digitalpreservation/2013/08/whats-a-nice-english-professor-like-you-doing-in-a-place-like-this-an-interview-matthew-kirschenbaum/.

Educational Offerings

Lee developed an educational offering called "Applying Digital Forensics Techniques to Materials Acquired on Physical Media," which he taught as a half-day workshop at the IS&T Archiving Conference in May 2009 and a full-day workshop for staff of UNC Libraries in August of the same year.

From June 2010 to June 2011, Lee served as the Principal Investigator for the Digital Acquisition Learning Laboratory (DALL) project, which was funded by the Andrew W. Mellon Foundation.[42] He oversaw the installation and setup of the digital forensics hardware and software to be used in support of both course work (offered to undergraduate and graduate students) at SILS and continuing professional education offerings. In early 2010, Lee and the SILS computing staff worked closely to build expertise and infrastructure within the school to administer novel educational exercises integrating digital forensics into LAM acquisition tasks. In addition to building significant hardware, software and procedural capacity for future digital forensics teaching and research, the DALL project provided SILS personnel with practical experience concerning opportunities and challenges for LAM professionals using digital forensics tools. It also allowed Lee and a member of the SILS information technology staff to take a series of online courses on digital forensics, offered by AccessData, as well as an on-site course offered by Digital Intelligence.

The FIDO project discussed above ran a training event called "Applying Digital Forensic techniques to AIM [Archives and Information Management]" on August 16, 2011 at King's College London.[43] There were three talks followed by a breakout discussion of specific professional and ethical scenarios. There was then a hands-on session, in which participants used OSForensics[44] to analyze a disk image from a laptop computer.

In Spring 2011, Lee administered an integrated set of exercises within the Electronic Records Management course (INLS 525), which he taught with fellow SILS faculty member Richard Marciano. Lee and Kam Woods then developed a new special topics course at SILS, Acquiring Information from Digital Storage Media (INLS 490-141), which they offered in Spring 2011. They substantially expanded the course to become Digital Forensics for Curation of Digital Collections (INLS 690-141) in Fall 2013. These courses have introduced students to digital forensics concepts and methods using both commercial and open-source software. Another SILS course that covers several core digital forensics concepts and methods is entitled Understanding Information Technology for Managing Digital Collections (INLS 465), which was introduced in Fall 2008 as an outgrowth of the DigCCurr (Digital Curation Curriculum) project funded by the Institute for Museum and Library Services (IMLS). Lee has also incorporated core digital forensics exercises into the DigCCurr II Professional Institute, which is a week-long continuing education course that has been offered annually in Chapel Hill (and once in Copenhagen) since 2009, as well as the State Electronic Records Initiative (SERI) Institute which was administered in Indianapolis, Indiana on July 8-12, 2013.

In 2012, Lee began administering a series of Digital Forensics for Archivists classes for the Society of American Archivists as part of the DAS (Digital Archives Specialization) certification.[45] Dozens of practitioners in locations across the United States have participated in these workshops, which focus on "bootstrapping" practical applications of digital forensics within institutions handling born-digital materials. Kam Woods also began serving as instructor for the classes in December 2012; Lee and Woods have recently expanded it into a two-day event, which includes significant hands-on exercises components, using both commercial and open-source software.

42 Christopher A. Lee and Kam Woods, *Digital Acquisition Learning Laboratory: A White Paper* (November 2011), http://www.ils.unc.edu/callee/dall-white-paper.pdf.

43 https://fido.cerch.kcl.ac.uk/digital-forensics-for-archivists-training-event/

44 http://www.osforensics.com/

45 http://www2.archivists.org/prof-education/das

Since 2010, Matthew Kirschenbaum and Naomi Nelson (a member of the BitCurator Professional Experts Panel) have co-taught annually a newly designed course at the University of Virginia's Rare Book School (RBS). While the RBS is renowned for its offerings in areas such as book history, paleography, illustration and bookmaking techniques, and typography, it has also had a strong record of offerings in electronic text, digital librarianship—notably Encoded Archival Description (EAD) and the Text Encoding Initiative (TEI)—and digital humanities. Kirschenbaum and Nelson's offering on Born-Digital Materials: Theory and Practice is a week-long course, which typically enrolls a dozen students for some 30 contact hours, covers the life-cycle of a digital object, from initial conversations with a donor to accessioning, arrangement and description, preservation and metadata, and patron access to born-digital materials.

Kirschenbaum teaches several sessions dedicated to digital forensics as part of this curriculum. In the first iteration of the course attempts were made, with at best partial success, to have students download and install existing open source applications such as the Sleuth Kit for instruction. The vagaries of individual operating systems and environments (for example, students lacking administrative access to computers owned and maintained by their institutions) as well as varying levels of technical skill and familiarity with the UNIX command line proved prohibitive for a productive instructional experience. With the advent of BitCurator, most students in the class can download and install a complete digital forensics processing environment and begin working within an instructional context in a very short period of time; exercises walk the students through activities ranging from the imaging of media to the generation of human- and machine-readable reports. In addition to these hands-on activities, Kirschenbaum also covers the theoretical contexts for digital forensics, and closely related issues such as privacy and ethics.[46]

At the iSchool at the University of Texas, Patricia Galloway has also led an effort over the past several years to provide students with hands-on experience in recovering data from media. In 2009, she created a specialized Digital Archaeology Lab, which teams of students have used to work with a variety of born-digital materials. This is an outgrowth of groundwork for the lab that began in 2005 and more than a decade of courses at the iSchool in which Galloway has had students work in teams to address challenges associated with real-world digital collections.

Conclusion

The professional vocabulary of those working in collecting institutions is evolving to now include terms such as disk image, hex (hexadecimal) viewer, cryptographic hash, and file system. LAM professionals are also gaining access to new communities and sources of guidance, e.g. papers from the annual Digital Forensics Research Workshop and instructions from gaming enthusiasts about how to create, read and mount disk images of old storage media. The first and second points are closely related; having the right vocabulary can open up many new mechanisms for learning and sharing information.

46 Course materials, including a reading list and student evaluations, are available at: http://www.rarebookschool.org/courses/libraries/l95/

4. Technical Opportunities and Methods for Applying Digital Forensics

A range of free and powerful open source digital forensics tools exist to assist in processing large (and legacy) data collections. There are numerous benefits to the use of these technologies:

- *Automation and processing efficiency:* Access to tools that allow users—particularly those who have not received extensive technical training—to rapidly and accurately extract and report on the contents of a given file system, those items most likely to require further human intervention, and export technical and preservation-specific metadata on file items.
- *Accuracy in data triage, and reducing temporal footprints:* Ensuring that *coverage* of the contents of a given storage medium is as complete as possible, i.e. that a particular tool (or set of tools) identifies—accurately and using an algorithmic method than can be reviewed and explained—specific data items that are relevant within a particular preservation context.
- *Assurance of data integrity:* Using software and hardware technologies that are well-documented (and freely documented), and which have been sufficiently tested to ensure their performance in real-world scenarios.
- *Identifying personally identifying and sensitive information (PII):* For example, Simson Garfinkel's bulk_extractor extracts "features" such as email addresses, credit card numbers, and URLs (among others) from disk images and directories of files.
- *Establishing environmental and technical context:* Disk images retain *all* of the data on a source medium, supporting answers to a range of potential future questions, including "What kinds of programs were used to create this data?" and "What did the daily workflow of a particular user look like?" Digital forensics tools may be used to extract information corresponding to installed software, user activities (for example, times and dates that a particular user was logged on), and create high-level timelines of activity.

Information located in disk images can assist in linking digital objects to other data sources and activities:

- *Versioning information*: The version of a software package used to create a given document may have specific bearing on the future renderability of that document. For example, although many current and legacy office document formats are "cross-platform," documents created on a given platform may exhibit rendering idiosyncrasies on another. Metadata about the production software may not always be embedded in the document, so access to a disk image that contains a complete copy of the production environment may improve preservation outcomes.
- *System Logs*: Many types of data transfer (such as plugging a USB flash drive into a Windows machine, accessing a network drive, or uploading data to an online service) leave traces on disk. Knowing that such traces exist can assist both in recovering lost data and in organization of the raw materials received.
- *Local and network user activity*: Log files and local databases recording user activity may be queried to address questions of document provenance, or to assist in producing a narrative about how a particular producer worked on a given system. Alternately, such information

may be redacted to protect private and sensitive data, or assigned a higher or lower priority during data triage depending on the contents. Materials collected (particularly those from the past decade) also are likely to include traces of network activity, such as use of online email services, social networks, and document-sharing mechanisms. Such traces may be used to create a more complete picture of the work habits of a particular donor, document links to other electronic resources, or trigger redaction protocols as required.

Forensic Disk Images

Accessing data on a storage device normally involves mounting a volume and then copying or opening files by interacting with the file system. There must be hardware to detect signals on the medium, hardware and software to translate the signals into bitstreams, and hardware and software to move the bitstreams into the current working computer environment. One can then interact with data as entire files or components of files. The file system mediates between the user and the underlying data, and it is designed to facilitate interaction at the file level (e.g. file naming, viewing timestamps, access controls). The file system serves to "hide" complicated information from the user about "where and how it stores information."[47] For most purposes, the file system is a valuable high-level abstraction, because it does not require users to understand or directly access the underlying data.

Those who are interested in the underlying data that are hidden by the file system can instead generate and interact with disk images, or sector-by-sector copies of all the data that reside on the storage medium. Inspection of the disk image can reveal a significant amount of information that users of the drive did not consciously or intentionally leave there,[48] but can also lead to traces of valuable contextual information. These may include information about other users who have interacted with the device or storage medium, indications of data transfer to online storage or social media platforms, and information a user may have considered "lost" or unrecoverable.

Forensic workflows often involve creation of a disk image to serve as a baseline copy of the data from the disk, upon which many further extraction and analysis tasks can be performed. Digital forensics professionals use hardware write blockers to ensure that no data on the disk – including essential metadata such as timestamps – are altered or overwritten during the process of copying the disk's contents. Device failures and the presence of unforeseen firmware or software bugs cannot be eliminated, but established workflows in forensics practice reduce the probability that these will occur undetected. The National Institute of Standards and Technology (NIST) tests and reports on hardware- and software-based write-blockers.[49]

LAMs can incorporate a variety of forensics practices and methods by treating disk images, rather than individual files or packaged directories, as basic units of acquisition.[50] Using write blockers, creating full disk images and extracting data associated with files is essential to ensuring provenance, original order and chain of custody.[51] There are a wide range of hardware write blockers available today from companies such as Tableau and WiebeTech, but—performance and packaging considerations aside—they all essentially perform the same task: preventing the host system (computer used for acquisition) from writing any data back to connected source media. In Figure 5, a legacy PATA (IDE) hard disk drive is connected to a WiebeTech device that provides write blocked access to PATA and SATA hard drives. This particular device allows a host with USB 2.0, FireWire, or eSATA ports to read these media.

Even the "modern" interfaces that allow these devices to interact with legacy media will ultimately become obsolete, and there are ongoing discussions concerning how and when the

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

48 Simson L. Garfinkel and Abhi Shelat, "Remembrance of Data Passed: A Study of Disk Sanitization Practices," *IEEE Security and Privacy* 1 (2003): 17-27.

49 http://www.cftt.nist.gov/hardware_write_block.htm

50 Woods, Lee, and Garfinkel, "Extending Digital Repository Architectures," 57-66.

51 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.

Figure 5: A legacy IDE hard disk connected to a modern forensic write blocker.

forensic hardware itself should be preserved and documented. This is especially important for institutions that elect to purchase hardware as a hedge against acquisitions problems long into the future.

In combination with a disk imaging software tool (such as the Guymager tool shown acquiring a USB flash device in Figure 6), the disk imaging process is both simple (from the standpoint of the end user) and highly reliable. Disk imaging tools will typically produce metadata about the acquisition process in a human-readable text file; technical provenance may therefore be retained even if the user elects to produce a raw, rather than a forensically-packaged disk image. This metadata typically includes the time and method of acquisition, technical details on the device or media carrier from which the image was extracted, any errors encountered (such as bad blocks on a disk), and notes created by the person administering the process. In the following section, we describe both this and other forms of forensic metadata in additional detail.

Creating and Extracting Forensic Metadata

Forensic metadata incorporated into packaging such as Guidance Software's Expert Witness Format and the Advanced Forensic Format (v3) incorporate low-level technical metadata about the disk, including the block size, the size of the physical medium in blocks, any compression method used, and one or more digest hashes. This is useful for tracking provenance and verification of integrity, but the true power of forensic imaging and analysis lies in the metadata that can be extracted from the file system(s) contained within a disk image.

The core facilities used by an operating system to interact with a file system can be replicated by forensic software designed to parse the contents of a disk image without mounting that image; a low-level approach to extracting large amounts of potentially relevant information. From the standpoint of forensics tools, the most basic of these views is to provide a listing of the contents of the file system in question – the volumes, directories, files, and other data contained on a particular medium.

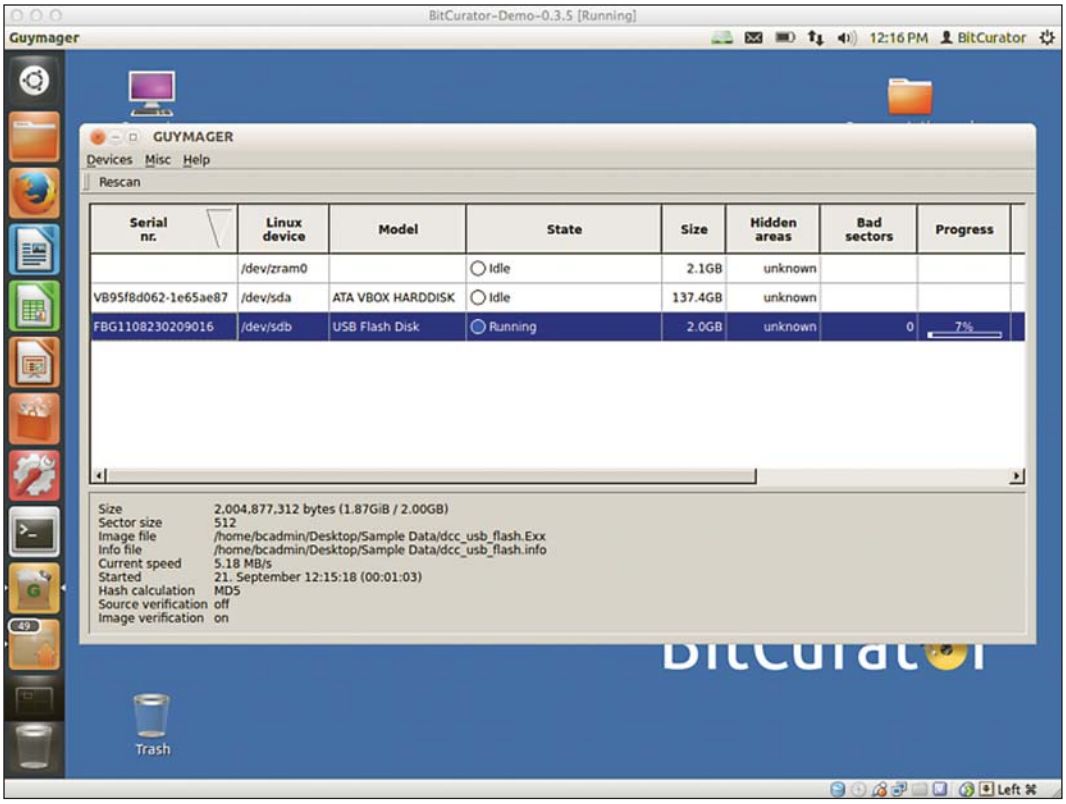


Figure 6: Guymager acquiring an image of a removable USB flash device in the BitCurator environment.

In the past decade, there have been several pushes towards the creation of interoperable standards for digital forensics metadata. Among the most visible of these has been Digital Forensics XML (DFXML), originally designed and developed by Simson Garfinkel. DFXML is intended to serve as "a standardized set of tags and representations for current XML tools," along with "a DTD and schema to allow XML validation."[52] At the time of writing, the schema has not been finalized (although a formal release is currently being reviewed by the DFXML Working Group[53]), but a range of relatively mature tools exist to produce and process core DFXML output describing file systems.

The same tools that generate DFXML can also incorporate metadata from other tools used in repositories. For example, fiwalk includes a plugin mechanism which can be used by LAM professionals to accommodate the output of preservation-specific tools and resources such as the PRONOM registry.[54]

Identification and Redaction of Sensitive Information

Forensic tools can be used to identify, flag and redact, or restrict access to sensitive information.[55] LAMs may use these tools to target gaps in existing workflows, or to shift away from time-consuming, expensive manual analysis toward automated and reliable procedures that can free LAM professionals to focus their energy on the tasks that require human intervention.

The BitCurator project has constructed an environment that incorporates bulk_extractor, a tool capable of identifying "features of interest" within a disk image, bitstream, or location (folder containing files) within a live file system. The features identified by bulk_extractor are typically

52 http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML
53 https://github.com/dfxml-working-group
54 https://github.com/anarchivist/fiwalk-dgi
55 Christopher A. Lee and Kam Woods, "Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage," in *Proceedings of Memory of the World in the Digital Age: Digitization and Preservation: An International Conference on Permanent Access to Digital Documentary Heritage, 26-28 September 2012, Vancouver, British Columbia, Canada*, ed. Luciana Duranti and Elizabeth Shaffer (United Nations Educational, Scientific and Cultural Organization, 2013), 298-313.

textual patterns with a specific semantic interpretation (e.g. Social Security numbers, email headers, image and geolocation metadata, and user activity history such as URLs corresponding to search results). These features may not all constitute PII (Private and Individually Identifying) information for a given device, but their extraction can assist in performing triage on large-scale digital containers (or large collections of small-scale digital containers), and in allowing those materials deemed most critical for human review to filter up to the top.

Because it operates on the raw underlying bitstream, bulk_extractor can be run on disks that contain any file system. It identifies the locations of features in terms of byte offsets (number of bytes one would have to read from the beginning of the disk to reach them), rather than indicating where they reside within the folder and file structure within the file system. For example, in Figure 7, the bulk_extractor viewer shows a domain name that appears at offset 338347207, rather than indicating what specific directory path on the disk would be associated with that location. This approach has major performance advantages, because bulk_extractor does not use the file system to read data. It also means that bulk_extractor—like many other forensics tools—can find and present data that resides in corrupted or unallocated areas of a disk (e.g. traces of deleted files), which would not be visible through normal user interaction through the file system, such as when one uses Windows Explorer to navigate to folders and files.

Other programmatic methods allow one to match features that bulk_extractor has found to specific files and locations within the file system(s) identified on a particular device (in this case, for those file systems that are understood by The Sleuth Kit). Following this process, it is possible to generate simple, human-readable lists of pertinent information: "This Social Security number was found in this file", or "Evidence of email activity was found, even though this area on the disk is no longer seen by the file system." Software developed for the BitCurator project outputs these lists in forms—.xlsx spreadsheets or PDF files—that can be read using commonly available applications.

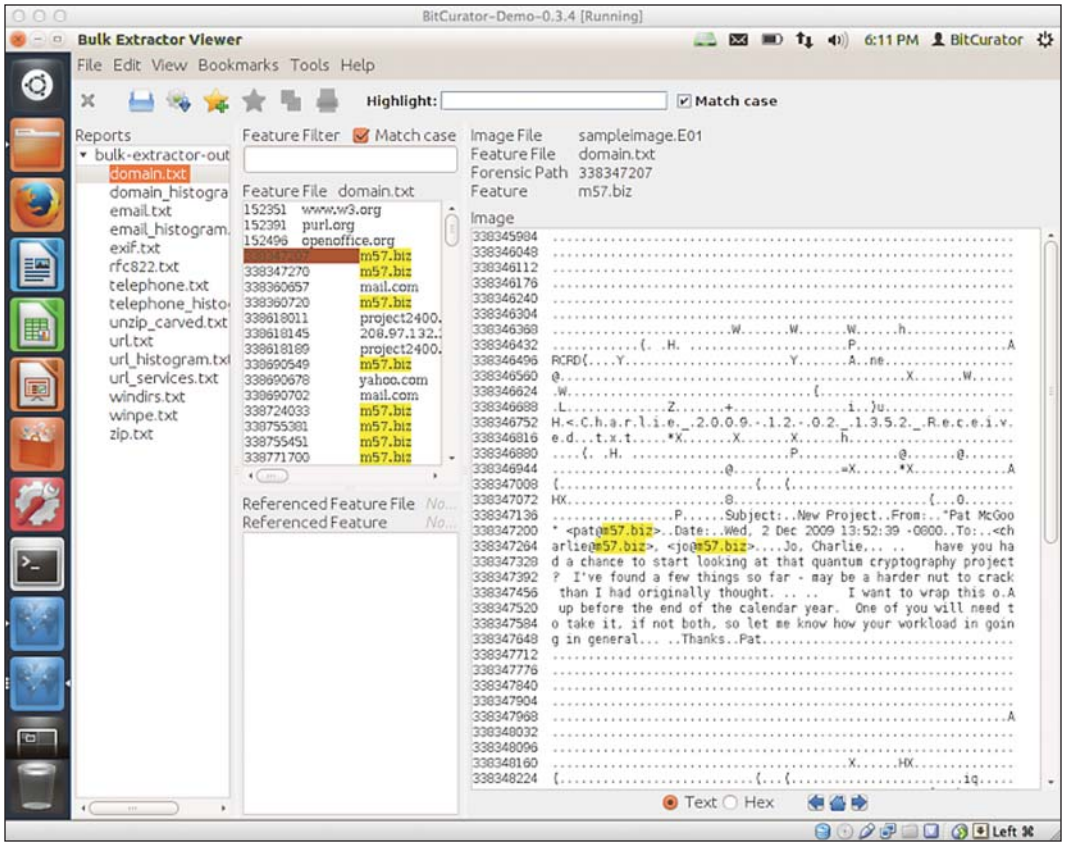


Figure 7: The bulk_extractor GUI allows users to examine feature reports, individual features, and hex and raw text views of those features within the bitstream.

Once these two forms of information are available—the raw features of interest that might correspond to private and/or sensitive information and a mapping of those features to their locations on disk—it is possible to automate the redaction of these items. A relatively simple Python script (iredact.py) distributed with the DFXML tools produced by Simson Garfinkel illustrates this possibility. Feature contents identified as potentially sensitive or private may be overwritten at the byte level—at the discretion of the user—with randomized data or a sequence of specific characters. Ongoing work is being performed in this area to build programs capable of selective encryption of these contents, laying the foundation for access scenarios in which white-lists governed by specific encryption keys may be used to moderate access to the raw disk image.

5. Challenges

Digital forensics offers valuable methods that can advance the goals of maintaining authenticity, describing and preserving born-digital records and providing responsible access.[56] However, most digital forensics tools were not designed with these objectives in mind. The BitCurator project is attempting to bridge this gap through engagement with digital forensics experts, library and archives professionals, as well as dissemination of tools and documentation that are appropriate to the needs of memory institutions. Much BitCurator activity is translation and adaptation work, based on the belief that LAM professionals will benefit from tools that are presented in ways that use familiar language and run on platforms that LAMs can support.

Challenge #1: Incorporation into the workflows of LAMs, e.g. supporting metadata conventions, connections to existing content management system (CMS) environments.

It is desirable to not just run forensics tools over disks, but to also export forensic data in ways that can then be imported into LAM collection management and descriptive systems, as well as modifying forensics triage techniques to better meet the needs of LAMs. Many institutions have now incorporated digital forensics into their workflows.[57] However, there must still be considerably more work before forensics tools and methods are tightly integrated into their overall systems and practices. For many LAMs, born-digital workflows are undergoing significant evolution, as their collections grow, they gain further expertise, and new practices continue to emerge from the field. Moreover, the metadata being routinely captured and stored in CMS environments varies widely across institutions. Integration with LAM workflows and systems will, therefore, be an ongoing endeavor.

Challenge #2: Obsolete Storage Media and File Systems

Most current forensics software is designed to work with a relatively small set of common file systems, such as FAT and NTFS (Windows), ext (Unix) and HFS+ (Mac). Not only will many tools not be able to read older and less common file systems, but the tools often will not even be able to identify those file systems. Likewise, specialized physical interfaces to a particular hardware platform are unlikely to be addressed by the current manufacturers of forensic acquisition hardware, and even support for common interfaces is likely to be discontinued within a timeframe of one to two decades (e.g. the recent removal of SCSI connectors from the latest generation of Digital Intelligence's UltraBay write blockers). An example commonly encountered by LAMs is the 5.25" floppy diskette, which is no longer supported on current computers. It is unlikely to be a target for most forensics practitioners, because it has not been actively used for many years and is thus not likely to be used as evidence in an investigation.

56 Woods and Lee, "Acquisition and Processing of Disk Images," 147-152.

57 For examples of such workflows, see 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; and AIMS Working Group, *AIMS Born-Digital Collections*. .

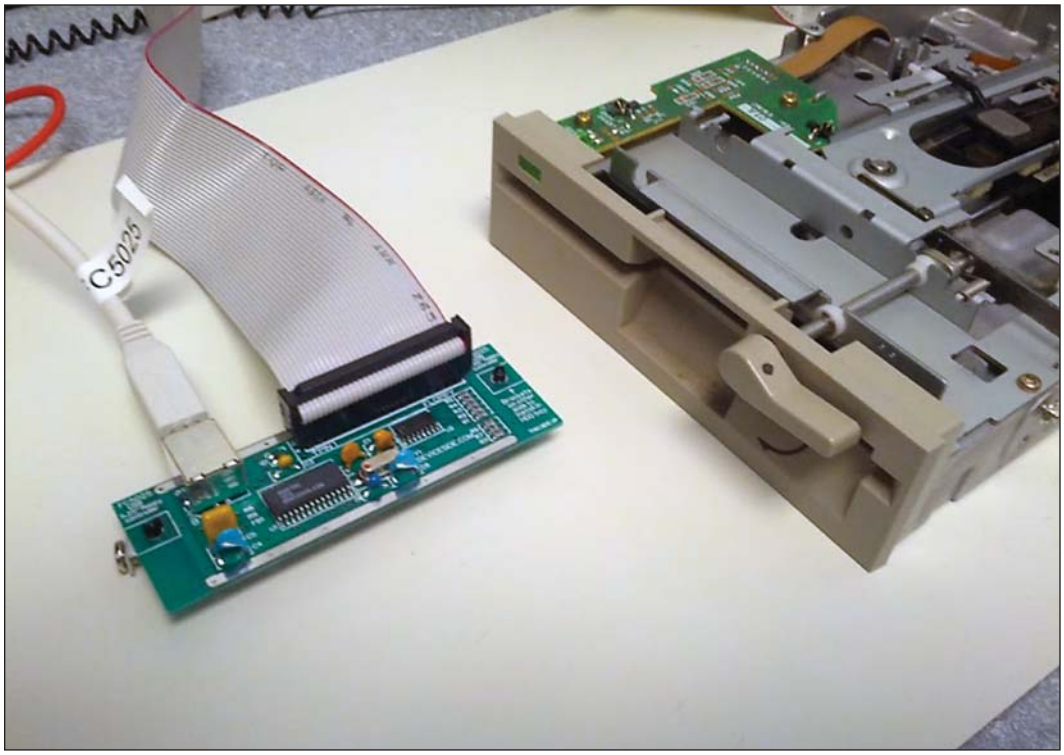


Figure 8: FC5025 being used to copy data from a 5.25" floppy disk.

However, LAMs often have many of these disks in their collections. Specialized devices such as the FC5025 shown in Figure 8 can help.[58]

Challenge #3: Dealing with large, internally complex data files.

The digital collections in many LAMs have now grown to contain millions of digital objects. This has resulted in many discussions about "big data" and the scalability of their technical architectures. These discussions have taken place in a variety of professional venues including the International Conference on Digital Curation (IDCC), the Preservation and Archiving Special Interest Group (PASIG), and series of meetings from 2007 to 2013 on Designing Storage Architectures for Preservation Collections organized by the National Digital Information Infrastructure and Preservation Program (NDIIPP).[59] The iPRES 2013 conference in Lisbon, Portugal also included a workshop on September 5-6, 2013, devoted to exploring "Digital Preservation at Scale." However, LAMs still have relatively limited experience in caring for individual items that are as large (gigabytes or even terabytes) and internally complex as disk images. Such files often require news tools, but also new arrangements for storing and transferring data, as well as workflows that can accommodate long delays as large data files are processed.

Challenge #4: Provision of public access to the data.

The typical digital forensics scenario is a criminal investigation in which the public never gets direct access to the evidence that was seized. By contrast, LAMs that are creating disk images face issues of how to provide access to the data. This includes not only providing access interfaces, but also redacting or restricting access to components of the image, based on confidentiality, intellectual property or other sensitivities.

58 See also "Use Guide for the FC5025 Floppy Disk Controller," Maryland Institute for Technology in the Humanities, http://mith.umd.edu/vintage-computers/fc5025-operation-instructions.

59 http://www.digitalpreservation.gov/meetings/

Challenge #5: Defining and Implementing Ethical Commitments

In addition to the logistical issues, there are also institutional and ethical issues,[60] which will be explored and clarified in ways that would not be possible without direct experience with the technology. LAM professionals can benefit from new ethical frameworks that address long-standing principles and values but are attentive to the many levels of representation that are present in digital environments.[61] It will also be important to clarify donor expectations and agreements regarding the retention, treatment and exposure of various traces of data (e.g. deleted files, user logs, configuration files, tracked change data within office documents).

60 Christopher A. Lee, "Bringing Values to the Bitstream: A Framework for Digitally-Aware Professional Ethics of Curation" (presentation to the Society of American Archivists (SAA) Research Forum, Austin, TX, August 11, 2009).

61 Christopher A. Lee, "Computer-Supported Elicitation of Curatorial Intent," in *Dagstuhl Seminar Proceedings 10291, Automation in Digital Preservation*, ed. Andreas Rauber, Jean-Pierre Chanod, Seamus Ross, and Milena Dobreva, 2010.

6. Lessons and Insights

Experiences from the BitCurator project—including monitoring and engagement with various other contemporary activities—have yielded a number of lessons and insights.

Digital forensics has arrived for archival processing.

As evidenced by the discussion above, there have been dramatic changes in the status of digital forensics within LAMs in just a few years. Many institutions now acknowledge that procedures and practices for the curation of born-digital materials should involve forensic tools and methods. There is growing recognition, for example, of the value of creating forensic disk images.

There is still much work to done, however. In a 2012 survey of libraries belong to the Association of Research Libraries (ARL), 78% of respondents indicated that a strategy they employ regarding born-digital records stored on legacy media is "storing legacy media as is (without transfer to new media or server storage and/or keeping it with [an] analog collection)."[62]

The introduction of digital forensics to LAMs does not dictate specific policies or practices.

One can adapt and adopt tools and procedures to suit the circumstances of an individual institution and environment. For example, some institutions may retain significant amounts of forensic metadata as integral to the Archival Information Packages in their collections, while others may be much more selective about the metadata that they retain. One situation may call for forensic imaging of a full set of disks as the first step in processing them, while another situation may call for active triage and selection before determining whether a full disk image is necessary or desirable. The emergence of forensics tools and methods within LAMs should be seen as an additional set of options and capabilities, rather than a new set of rigid prescriptions for professionals to follow.

The disk image is a cornerstone of many forensics methods.[63]

LAM professionals should become comfortable with capturing and processing disk images as digital objects. Digital repository structures (and associated metadata schemas) should be adapted to accommodate disk images as baseline digital objects. A disk image provides the most complete representation of the information that was stored on a drive.

Modern digital forensics hardware and software simplifies the process of extracting disk images, but knowledge of the structure and capabilities of common media types should always inform related workflows. For example, neither raw nor forensically packaged disk images can be created more quickly than the maximum read speed supported by a particular disk or disk

62 Naomi L. Nelson et al, *Managing Born-Digital Special Collections and Archival Materials*, SPEC Kit 329 (Washington, DC: Association of Research Libraries, 2012), 35.

63 Note that are many contemporary digital forensics activities do not focus on disk images, e.g. live memory analysis and analysis of network packets. However, such "live" forensics is not as directly relevant to digital curation work within LAMs, so it is largely outside the scope of our discussion.

drive, but many practitioners are unaware of the typical read speeds supported by many types of common media. Familiarity with such factors (e.g. read speed of a USB 3 vs. USB 2 devices) can greatly facilitate decision making and planning within LAMs.

"Taking bitstreams seriously" can have major advantages.

A dramatic change in the work of LAMs can come from the use of tools that are designed to treat data at a very low level—as raw bitstreams off media—rather than treating data at the file level. Archivists have long argued that the essential content, structure and context elements of an electronic record can reside in multiple data sources and not just in a single file.[64] Digital forensics greatly enables such thinking; for example, it allows LAMs to bypass the file system and read data as a raw stream. This can reveal the most complete set of information, and it can also have significant performance advantages (not needing to mediate all actions through the file system).

Treating disk images as objects of preservation is not only beneficial to LAM professionals, it also can enable many new access scenarios.

Virtualization and Emulation

One such scenario involves the use of virtualization or emulation. A disk image of an individual piece of media allows that "disk" to be virtually "booted" in an emulator or virtual machine. One can then browse the file system and open individual documents much as an original user might have. This can be based on a disk image that has been created from an original, physical storage medium. However, an increasingly common scenario in the field of digital forensics is investigation of materials that were stored within a virtual machine in the original creation environment.[65]

Emulation has been used in the computer industry since the 1960s,[66] and it has been discussed as a long-term digital preservation strategy since the 1990s. In 1995, Jeff Rothenberg made the first argument for the use of emulation in digital preservation.[67] He generated several more publications and reports on the issue in the following five years.[68] From 1999-2003, the Creative Archiving in Michigan and Leeds, Emulated the Old on the New (CAMiLEON) project, funded jointly by the UK Joint Information Systems Committee (JISC) and the US National Science Foundation (NSF), investigated and compared migration and emulation strategies, through a combination of technical implementation, analysis of significant properties, and empirical user testing.[69] CAMiLEON drew upon work carried about by the CURL Exemplars in Digital ARchiveS (CEDARS) project, which was funded by JISC in the UK from 1998 to 2002, and also investigated emulation among other strategies.[70]

Another significant initiative involved personnel from IBM, Delft University of Technology,

64 See e.g., David Bearman, "Record-Keeping Systems," *Archivaria* 36 (1993): 16-36; John McDonald, "Towards Automated Record Keeping, Interfaces for the Capture of Records of Business Processes," *Archives and Museum Informatics* 11 (1997): 277-85.

65 Diane Barrett and Greg Kipper, *Virtualization and Forensics: A Digital Forensic Investigator's Guide to Virtual Environments* (Amsterdam: Syngress, 2010).

66 M.A. McCormack, T. T. Schansman, and K. K. Womack, "1401 Compatibility Feature on the IBM System/360 Model 30," *Communications of the ACM* 8, no. 12 (1965): 773-76; Stuart G. Tucker, "Emulation of Large Systems," *Communications of the ACM* 8, no. 12 (1965): 753-61.

67 Jeff Rothenberg, "Ensuring the Longevity of Digital Documents," *Scientific American* 272, no. 1 (January 1995): 42-47.

68 See e.g. Jeff Rothenberg, "Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation" (Washington, DC: Council on Library and Information Resources, 1999); Jeff Rothenberg, "An Experiment in Using Emulation to Preserve Digital Publications" (National Library of the Netherlands, 2000).

69 Margaret Hedstrom and Clifford Lampe, "Emulation vs. Migration: Do Users Care?" *RLG DigiNews* 5, no. 6 (December 15 2001); Phil Mellor, "CAMiLEON: Emulation and BBC Domesday," *RLG DigiNews* 7, no. 2 (April 15 2003); Margaret L. Hedstrom, Christopher A. Lee, Judith S. Olson, and Clifford A. Lampe, "'The Old Version Flickers More': Digital Preservation from the User's Perspective," *American Archivist* 69, no. 1 (Spring/Summer 2006): 159-87.

70 See e.g. David Holdsworth and Paul Wheatley, "Emulation, Preservation, and Abstraction," *RLG DigiNews* 5, no. 4 (2001).

Department of Computer, Science (EEMCS), Antwerp University, Department of Information and Library Science (IBW), and the National Library of the Netherlands (Koninklijke Bibliotheek - KB), who specified and built a prototype of a Universal Virtual Computer (UVC),71 an idea first articulated and investigated by Raymond Lorie at IBM.72 In 2004, the KB and Nationaal Archief of the Netherlands began investigating emulation strategies together; this work included development and testing of an emulator, called Dioscuri that was specifically designed to support preservation through a modular architecture and was publicly released in 2007.73

Several other recent projects have advanced the research and development of emulation environments for preservation purposes. The Keeping Emulation Environments Portable (KEEP) project was funded by the European Commission from 2009 to 2012. [74] KEEP generated an emulation framework, a transfer tools framework, a prototype virtual machine, as well as guidance on metadata and legal issues. The Preservation and Long-term Access through Networked Services (PLANETS) project, which ran 2006-2010 through funding by the European Commission, also provided important advances in the investigation of architectures for preservation and access through emulation. [75] The Preserving Virtual Worlds project, which ran 2008-2010 with funding from the National Digital Information Infrastructure Preservation Program (NDIIPP), investigated implications and technical strategies for using emulation to experience obsolete and complex born-digital objects. Other work has investigated the identification of software dependencies to support emulation.[76] More recently, the Baden-Wuerttemberg Functional Longterm Archiving and Access (bwFLA) project at the University of Freiberg in Germany, has explored various emulation strategies, including Emulation as a Service (EaaS).[77] The International Conference on the Preservation of Digital Objects (iPRES) 2013 was the first year in which there was a full-day workshop dedicated to emulation tools and strategies.[78]

There have been many more specific cases of emulation being demonstrated as a way to recover creative and artistic works that depend upon obsolete equipment.[79] An extreme case is illustrated by William Gibson's "Agrippa," a piece of electronic literature encoded to encrypt itself after a single reading: with a disk image of the original poem (and program), however, one can re-run it endlessly in an emulator, experiencing unique programmed behaviors (such as sound effects). Though the virtual copy of the diskette duly encrypts itself and is thus unusable after a single use, a user can simply replicate additional copies ad infinitum from an unexpended master.[80] Similarly, a ROM of a game cartridge allows the game to be played in an emulated

71 Jeffrey van der Hoeven, R. J. van Diessen, and K. van der Meer, "Development of a Universal Virtual Computer (UVC) for Long-Term Preservation of Digital Objects," *Journal of Information Science* 31, no. 3 (June 1 2005): 196-208.

72 Raymond A. Lorie, "Long Term Preservation of Digital Information," In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, ed. Edward Fox and Christine Borgman (New York, NY: ACM Press, 2001), 346-52.

73 Jeffrey van der Hoeven, Bram Lohman, and Remco Verdegem, "Emulation for Digital Preservation in Practice: The Results," *International Journal of Digital Curation* 2, no.2 (2008): 123-132.

74 http://www.keep-project.eu/

75 Dirk von Suchodoletz and Jeffrey van der Hoeven, "Emulation: From Digital Artefact to Remotely Rendered Environments," *International Journal of Digital Curation* 3, no. 4 (2009): 146-55.

76 Thomas Reichherzer and Geoffrey Brown, "Quantifying Software Requirements for Supporting Archived Office Documents Using Emulation," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY: ACM Press, 2006), 86-94.

77 Isgandar Valizada, Klaus Rechert, Konrad Meier, Dennis Wehrle, Dirk Von Suchodoletz and Leander Sabel, "Cloudy Emulation – Efficient and Scaleable Emulation-based Services," in *Proceedings of iPRES 2013*.

78 Dirk von Suchodoletz, Mark Guttenbrunner, and Klaus Rechert, "*Report on the first iPres Workshop on Practical Emulation Tools and Strategies," D-Lib Magazine* 19, (3/4), 2013, http://www.dlib.org/dlib/march13/vonsuchodoletz/03vonsuchodoletz.html.

79 For an example involving electronic music, see e.g. Jaime Bullock and Lamberto Coccioli, "Modernising Live Electronics Technology in the Works of Jonathan Harvey," in *Proceedings of the International Computer Music Conference* (Barcelona, Spain, 2005).

80 Matthew Kirschenbaum, Doug Reside, and Alan Liu, "'No Round Trip': Two New Primary Sources for *Agrippa*": http://agrippa.english.ucsb.edu/kirschenbaum-matthew-g-with-doug-reside-and-alan-liu-no-round-trip-two-new-primary-sources-for-agrippa

environment.[81] An image of a complete hard drive, meanwhile, not only affords a user access to a file system but also to a complete computing environment, including the range of applications installed on the original machine, so-called ambient data such as temporary or cached copies of files, Web browser histories, and seemingly incidental features which nonetheless might prove insightful to a researcher, such as the desktop wallpaper or system preferences. Access to such an environment entails a scenario that is potentially akin to walking in to a donor's virtual "house," complete with opening the drawers in the office, browsing the family photographs on the mantel, looking at what is playing on the stereo, and sorting through what is been left behind in the trash.[82]

Of course, a donor may not always want to expose all these digital traces to the world. A prominent example of using emulation to provide rich, but selected access to materials is the Salman Rushdie collection at Emory University's Manuscript, Archives, and Rare Book Library (MARBL). Emory staff used SheepShaver emulation software to view Rushdie's files in their original software environments. They then assigned one of four possible access categories to each file:

> As is – the file can be released "as is" for both the emulated environment and the searchable database;
>
> Redacted – the file will need to be redacted for access; it will not be available for emulation, but it will be available in the searchable database;
>
> Restricted – the file will be restricted and not be accessible in either environment;
>
> Virtual only – these files will only appear in the emulated environment; they will not be in the searchable database.[83]

To provide end-user access to the Rushdie materials, MARBL adopted a strategy that includes a searchable database, an emulation of Rushdie's Performa 5400 computing environment, and an archival finding aid. The emulation environment was created by removing all user-generated files and "then re-populating it with approved data in order to ensure that no restricted data would remain."[84] As explained by the MARBL team:

> The directory structure, desktop, user preferences, and file naming conventions established by Rushdie are also still intact and await exploration. Though it can be startling to users, the emulation allows researchers to fully interact with Rushdie's digital content: files can be modified, directories can be deleted, and games can be played. It seems that changes are actually made to the data itself, but each time the emulated environment is launched it refreshes the disk image, resetting to the original images. No changes are saved and no modifications are kept.[85]

Jeremy Leighton John at the British Library has written thoughtfully and experimented practically with the use of emulation and virtualization. He describes a compelling case:

> As reported at the Digital Lives Research Seminar 2010, SheepShaver has been adopted at the British Library in order to boot a disk image of one of the hard drives (G3 PowerMac with Macintosh System 8) from the evolutionary biologist WD Hamilton. Each time the computer disk is booted, one of several potential desktop pictures is revealed; for example, personally taken aerial photos of the river system in Amazonia.

81 For a discussion of both the potential and limitations of emulation in this regard, see Jerome McDonough, Robert Olendorf, Matthew Kirschenbaum, et al, "Preserving Virtual Worlds Final Report," Aug. 31, 2010, http://www.ideals.illinois.edu/handle/2142/17097.

82 Brian Carrier has developed the implications of analogizing a computer to a physical crime scene; see http://www.digital-evidence.org for his writings and papers on the subject.

83 Laura Carroll, Erika Farr, Peter Hornsby, and Ben Ranker, "A Comprehensive Approach to Born-Digital Archives," *Archivaria* 72 (Fall 2011): 61-92, 74.

84 Ibid., 84.

85 Ibid., 85.

> After a while, a screensaver appears. In addition to the usual applications such as Microsoft Word, Acrobat Reader and Photoshop 4.0, it has been possible to open CodeWarrior and run C++ programs residing on the original disk, displaying dynamic graphics.[86]

Mounting the Original Filesystem

A second access scenario involves accessing a disk image as if it were a physical disk connected to the user's computer. The user's operating system may allow him/her to mount some types of disk images and file systems, but this will often require specialized software. For example, the Linux-based BitCurator environment provides access to disk images via several methods. The user may mount a copy of a raw or forensically packaged disk image encoding a range of file systems, providing a simple avenue of access into the original file system without fear of contaminating the original bitstream. Current versions of Microsoft and Apple operating systems (Windows 7/8, OS X) have limited native support for reading disk images and can only mount a select set of file systems.

Accessing (But Not Mounting) Disk Images Using Forensics Software

A third access scenario involves use of digital forensics tools—e.g. The Sleuth Kit, FTK, FTK Imager, EnCase—to navigate (but not mount) the contents of a disk image. These tools allow the user to navigate the file tree, view contents in a hex editor and examine various metadata elements associate with each volume, folder and file. In some cases, users in a reading room could be provided with disk images and a computer running one or more of the above applications. They could then examine the contents of the disks without having to worry about mounting the file system(s) on the disks.

Remote, Dynamic Access to Disk Image Contents

A fourth access scenario involves serving out the disk image to the user through a server upon request. The disk images are stored on a server, and users can click down through the folder structure of each disk without ever having to download the disk image to their local machines. The technologies required to support this method have been demonstrated[87] and are largely in place, although not well distributed.

Cross-Drive Analysis

Current forensics software suites—including EnCase, FTK and The Sleuth Kit—allow one to load multiple disk images into a common environment called a "case." Users can then apply a variety of searches, filters and scripts to all of those images at the same time. For example, one can search for instances of a given email address, generating results that appeared on any of the disks. This can enable many useful tasks, but it also requires considerable pre-processing before the user can perform such an investigation. For example, the Stanford University Libraries have recently allowed two researchers to conduct research on one of their collections by using FTK in their forensics lab. The researchers were able to conduct searches across images of floppy disks and CDs (more than 200) and two hard drives simultaneously. This required the Stanford staff to first load all of the images into an FTK case, which took more than eight hours.[88]

Querying data does not need to be limited, however, to the bounds of a pre-defined forensics "case." Simson Garfinkel at the Naval Postgraduate School has pioneered research and

86 Jeremy Leighton John, *Digital Forensics and Preservation*, DPC Technology Watch Report 12-03 (Digital Preservation Coalition, 2012).

87 Woods and Brown, "From Imaging to Access," 213-18.

88 Personal correspondence with Peter Chan, Digital Archivist, Stanford University Libraries, September 30, 2013.

development in cross-drive analysis.[89] For example, one can identify all of the unique email addresses and all of the unique cryptographic hashes of files that appear on any disk image within a repository. One can then make comparisons across the drives, to support possible inferences about the relationships between those drives. In our example, if one email address (or file hash) appears on only two different drives in the whole repository, this could warrant further investigation to see if there is an interesting relationship between those disk images. They may have come from the same individual/organization or from two different individuals/organizations that were sharing information with each other. In Garfinkel's work, statistical techniques are used to produce and score correlations lists for large data collections. This can allow an analyst not only to identify features that are shared between disks, but also which of those disks may have originated from a particular producer. Such exploration of features identified on the disk and named entities (e.g. names, places) across images of drives could open up many exciting possibilities for description of collections within and across LAMs, as well as facilitating many new forms of research.

Reflections on New Access Scenarios

The above approaches are compatible with long-standing scholarly practices which place a premium on direct access to primary sources in the analog world. Disciplines such as philology, diplomatics, descriptive and analytical bibliography, and textual criticism, as well as more recent scholarly developments such as the platform studies and media archaeology discussed above all attend to the materiality of primary source documents and artifacts through the direct study of the objects themselves, or else through surrogates of the highest integrity and authenticity. Digital forensics provides the methodological bridge for ensuring that such approaches, which are compatible with long-standing humanistic and scholarly precepts—all the way back to the primal tableaux proferred by Carlo Ginzburg as our first reader of signs, the hunter crouched on the forest floor studying the tracks of his prey[90]—are respected and sustained as the scholarly record expands to encompass new born-digital objects of study. In that sense, digital forensics is not only an aid for professionals processing collections, but also a service to a future in which we are unable to anticipate the needs and desires of the patrons of those collections. By ensuring the survivability of complete bitstreams supported by robust metadata, guarantors of authenticity, and ongoing fixity and integrity checks one can ensure the continuance of what the literary critic Van Wyck Brooks once termed "a usable past."

As discussed above, the digital environment affords multiple opportunities for interacting with information at various levels of representation. Different layers of description facilitate various forms of navigation and access.[91] Rather than seeing any of the above access scenarios as being canonical or mutually exclusive, LAMs and the researchers who use them can explore a variety of access methods in order to best meet their needs and interests.

89 See e.g. Simson L. Garfinkel, "Forensic Feature Extraction and Cross-Drive Analysis," *Digital Investigation 3* (September 2006): S71-S81.

90 Carlo Ginzburg, "Clues: The Roots of an Evidential Paradigm," in *Clues, Myths, and Historical Methods*, trans. John and Anne C. Tedeschi (Baltimore, MA: Johns Hopkins University Press, 1989), 96-125.

91 Peter Horsman, "Dirty Hands: A New Perspective on the Original Order," *Archives and Manuscripts* 27, no. 1 (1999): 42–53.

7. Recommendations for Future Activities

This section includes both recommendations for what LAM professionals can do and what research and development remains for tools and methods. [92]

Shared Documentation of Storage Media

One promising area of development is in the provision of information about digital storage media. The first time that one acquires or examines an unfamiliar or obsolete type of storage medium, there can be a steep learning curve in becoming familiar with the medium. This can involve an understanding of physical connectors, power plugs, drives and enclosures, but also likely failure modes and dependencies on specific hardware, software, and firmware in order to read data off the device. There have been efforts to pool information related to media to serve as a common professional resource, including Mediapedia at the National Library of Australia[93] and the Trustworthy Online Technical Environment Metadata (TOTEM) Registry work of the KEEP project.[94] The Computer Product Manuals Collection at the Charles Babbage Institute can also serve as a useful resource. The work of LAM institutions could benefit substantially from further efforts to organize, manage and disseminate such information.

Develop and Share Corpora for Education, Research and Tool Development

An ongoing issue in digital forensics has been the need to develop and distribute corpora that accurately reflect the day-to-day issues faced by forensic examiners and can be used in both research and education. These issues are closely mirrored in LAMs. Students in professional and continuing education programs benefit from exposure to the types of data they will encounter in the real world, and research on digital curation tools and strategies similarly benefits from test data that reflect actual challenges. However, realistic datasets can be complicated and time-consuming to produce. Consequently, many educational or research efforts are based on either (1) "toy data" that were purposefully designed to have a single, identifiable solution or test a specific feature,[95] (2) reuse of data that are not well suited to the task at hand, or (3) extensive labor on the part of the educator or researcher to custom-build data for a particular purpose. LAM professionals could benefit significantly from more systematic development and sharing of realistic corpora for education, research and tool development.

Real-world collections often present more complex challenges. The fundamental tension in developing corpora is to present situations in which the data are sufficiently complex to avoid trivial (or circumvention-based) solutions, but eliminate enough of the "noise" encountered in real world data to facilitate identification of the solution in a reasonable amount of time.

92 Several of the following recommendations were previously articulated in Jeremy Leighton John, Matthew Kirschenbaum, Mark Matienzo, Don Mennerich, Christopher A. Lee, Porter Olsen, and Kam Woods, "Crossing the Bitstreams: A Call for Collaborative Application of Forensics to Digital Curation Work" (paper presented at the Aligning Digital Preservation across Nations Workshop, Eighth International Digital Curation Conference, Amsterdam, The Netherlands, January 14, 2013).

93 http://mediapedia.nla.gov.au/home.php

94 http://keep-totem.co.uk/

95 See, for example, Brian Carrier's valuable, but highly targeted Digital Forensics Tool Testing Images, http://dftt.sourceforge.net/, and the Computer Forensic Reference Data Sets (CFReDS) from the National Institute for Standards and Technology (NIST), http://www.cfreds.nist.gov/

Previous projects have generated realistic corpora to be used in digital forensics education, including the M57-Patents data set, which was created through funding by the National Science Foundation.[96] The M57-Patents data set has been used in LAM education and tool testing.[97] Simson Garfinkel also maintains various other corpora at the site digitalcorpora.org, including govdocs1, which contains approximately one million documents downloaded from web sites in the .gov domain.[98] A range of test images and challenges can also be found on the Forensic Focus site.[99]

Although they are extremely valuable for various purposes, these corpora currently address only a limited set of scenarios, and they are not ones that specifically reflect the sorts of data and challenges that are most likely to be encountered in LAM settings.

Some other related activities have been:

- DigitalPreservationEurope ran three digital preservation challenges[100] from 2007 to 2009 geared towards students and researchers. These challenges included a small number of self-contained problems focused on preservation, none of which included disk images.
- Through the leadership of Paul Wheatley from the University of Leads, the Open Planets Foundation, maintains a site that includes "Datasets, preservation and curation Issues with those Datasets, and Solutions to those Issues." The experiences of solving specific Issues are written up on Solution pages, which then link to pages in the OPF Tool Registry. In many cases, this leads to "actual code that can be downloaded and re-used." [101]
- The Preservation and Access Virtual Education Lab (PAVEL) project at the University of Michigan School of Information, funded by the National Endowment for the Humanities, has developed a "virtual education laboratory featuring digital access and preservation tools." This currently includes four data sets:
  - a small set of six files for testing preservation tools
  - Elena Kagan's email (approximately 19,000 messages) from her tenure in the White House
  - the Enron email collection (approximately 600,000 messages)
  - a set of documents from the University of Michigan College of Literature, Science, and the Arts IT department (approximately 450 megabytes of Microsoft Office files)[102]

The above PAVEL data sets have the advantage of being selected for use in archival education, and they contain important supplementary information (e.g. organizational charts) for understanding their contexts of creation. However, they do not allow for the replication of many forensics tasks. For example, all of the email messages are stored as separate ASCII text files, rather than being embedded in a disk image or within a wrapper format such as .pst, a scenario likely to be encountered during acquisition of a new collection.

There is still a pressing need to generate and share corpora that:

- have a significant temporal element, such as data from a range of producers over a period of years;
- reflect significant technical challenges, such as disk images from a wide range of media or computing platforms used by a single producer;

96 See Simson Garfinkel, Paul Farrell, Vassil Roussev, and George Dinolt, "Bringing Science to Digital Forensics with Standardized Forensic Corpora," *Digital Investigation* 6 (2009): S2-S11; Kam Woods, Christopher A. Lee, Simson Garfinkel, David Dittrich, Adam Russell, and Kris Kearton, "Creating Realistic Corpora for Forensic and Security Education," in *Proceedings of the ADFSL Conference on Digital Forensics, Security and Law* (2011),123-24.

97 Lee and Woods have used the M57-Patents data set in a variety of educational settings, and Mark Matienzo at Yale University has used them in the testing of tools that he has developed.

98 See http://digitalcorpora.org. Don Mennerich at the New York Public Library has made use of the NPS-Canon images in testing of tools he has developed.

99 http://www.forensicfocus.com/images-and-challenges

100 Digital Preservation Challenge, DigitalPreservationEurope, http://www.digitalpreservationeurope.eu/challenge/.

101 Digital Preservation and Data Curation Requirements and Solutions, Open Planets Foundation, http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

102 http://www.virtualarchiveslab.org/view/datasets

- synthesize or mimic private and sensitive information likely to be targets for LAMs, such as private correspondence, human subjects research data, student records, and health records;
- can be used to benchmark the coverage and performance of existing LAM and digital forensics software environment, or can be used for training and education. Such corpora require gold-standard annotations that specify each item of interest, along with appropriate documentation and exercises.

Creating corpora that meet one or more of these needs and can be freely redistributed is a potentially complex and expensive task that calls out for collaboration.

Better Support for Pattern Detection Relevant to Curation Tasks

As mentioned earlier, most current forensics software is designed to work with a relatively small set of common file systems, such as FAT and NTFS (Windows), ext (Unix) and HFS+ (Mac). A great contribution would be a body of shared information about how to detect various encoding schemes and file systems on disks.[103] For example, what distinct pattern of signals or bits appears at the beginning of Commodore 64 disk (CBMFS file system) or older Macintosh disks (HFS file system)? By sharing this information, LAM professionals would be better able to conduct basic triage on their collections and develop strategies for what do with the media in their care.

A much wider set of considerations relate to pattern detection more generally. There are a variety of algorithms and regular expressions that one can use to identify file types at the file level (e.g. using headers and extensions) or sector level (e.g. using end-of-line markers), as well as other features in the data. These may include credit card numbers, Social Security numbers, dates of birth, and other potentially personally identifying information that warrants redaction or restriction. Rather than reinventing such algorithms and expressions each time they are needed, LAM professionals will benefit from sharing and collaborative development of pattern detection methods related to common curatorial tasks.

Further Integration of Forensics and LAM Metadata

LAM professionals want not only to extract information from disks, but also to incorporate the information into their collection management and access environments. Several open-source digital forensics tools share a common set of metadata elements — Digital Forensics XML (DFXML) — and LAMs take advantage of those conventions when incorporating metadata into their systems.[104] However, mappings from DFXML to existing LAM metadata schemes are still in early development, and there is great potential for further work on cross-walks, transformations and application profiles for given settings and situations. LAMs would be further served by a standardized DFXML schema.

Further Connections between Open Source Efforts

The model for software development and community engagement observed by the BitCurator project emphasizes transparency, clarity, and reusability. Open source software and freely available documentation and training materials support both future development efforts and educational programs. However, software tools are only as useful as the mechanisms that allow people to find them and understand what they are designed to accomplish. UNC SILS has partnered with the Open Planets Foundation to support the construction of mechanisms that emphasize discovery and extension, such as the Open Planets Foundation Knowledge

103 John, Kirschenbaum, Matienzo, Mennerich, Lee, Olsen, and Woods, "Crossing the Bitstreams".
104 Kam Woods, Alexandra Chassanoff, and Christopher A. Lee, "Managing and Transforming Digital Forensics Metadata for Digital Collections," in *Proceedings of iPRES 2013*.

Base.[105] BitCurator team members have worked to package software used to support these methods in ways that are extensible and can be integrated into existing platforms. Many of the core software libraries and dependencies for BitCurator are shared by Archivematica, an open source environment developed by Artefactual Systems. We are also in regular communication with the developers of ArchivesSpace, another important open-source product for archives.

Increased Collaboration with Digital Humanities

The thriving international Digital Humanities community is an obvious yet thus far underutilized collaborator for LAMs with born-digital collections. This is perhaps partly because Digital Humanities research tends to gravitate toward pre-20th century subjects and corpora where copyright restrictions do not obtain. Born-digital materials, by contrast, are tied to more contemporary subjects, placing them outside the typical purview of DH. Yet there is enormous potential for cross-transfer of technology and further collaboration. Digital Humanities, for its part, can benefit from applications of digital forensics to improve its strategies for data handling, curation, and preservation. LAMs, meanwhile, could use the new tools and techniques that are commonplace in Digital Humanities—ranging from large-scale text analysis with techniques such as topic modeling to GIS modeling and visualization—to analyze and explore collections of born-digital materials. If digital forensics is indeed the modern-day incarnation of such centuries-old techniques as diplomatics, philology, and bibliography, then such connections ought to be within reach.[106]

105 http://wiki.opf-labs.org/display/KB/Home

106 For a more complete discussion of the opportunities and challenges around such collaborations, see Matthew Kirschenbaum, "The .txtual Condition: Digital Humanities, Born-Digital Archives, and the Future Literary," *Digital Humanities Quarterly* 7, no. 1 (2013), http://www.digitalhumanities.org/dhq/vol/7/1/000151/000151.html

8. Conclusions

The application of forensics tools and methods to the curation of born-digital collections in LAMs has advanced significantly over the past several years. It would have been quite surprising, for example, to hear an archivist talking about write blockers or disk images ten years ago, but such terms are now used frequently at archival conferences and increasingly in the professional literature. The BitCurator project is actively working to construct both the tools and the necessary documentation to help LAMs integrate digital forensics into their workflows. The BitCurator software environment is freely available for download and installation, and we continue to add associated documentation.[107] The next phase of BitCurator will focus on continued software development, professional engagement activities, and further uptake of the software as we work to resolve real-world challenges facing LAMs.

We expect that many of the methods described in this paper will become standard practice in collecting institutions in the years ahead. While "applying forensics to the curation of digital collections" is currently an active topic of discussion among LAM professionals, we expect this framing of the issues to change dramatically as the methods become more commonplace. For example, creating a disk image or using a library of cryptographic hashes to identify known files are currently characterized as borrowing methods from forensics. However, as LAMs widely incorporate such activities into their regular repertoire of professional practices, it is likely that they will no longer be seen as being borrowed at all. Instead, they will part of the regular course of doing business, much as the use of Web-based online access systems in LAMs was considered novel in the 1990s but is now taken for granted.

We look forward to many exciting advances in the coming years, as both LAM professionals and scholars who use digital collections further enhance their capabilities to tell authentic and compelling stories.

Acknowledgements

BitCurator is supported by a grant from the Andrew W. Mellon Foundation. Members of the BitCurator project team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, Amanda Visconti, and Kam Woods. We would like to acknowledge the members of the BitCurator Development Advisory Group (DAG) and Professional Expert Panel (PEP) for the for their valuable support and input: Bradley Daigle, Erika Farr, Geoffrey Brown (2011-2012), Barbara Guttman, Jeremy Leighton John, Leslie Johnston, Jennie Levine Knies, Jerome McDonough, Mark Matienzo, Courtney Mumma, Naomi Nelson, Erin O'Meara, Michael Olson, David Pearson, Gabriela Redwine, Doug Reside, Seth Shaw, Susan Thomas, William Underwood, and Peter Van Garderen (2011-2012). We would also like to thank the numerous LAM professionals who have provided us individual feedback on the project's products, approaches and strategies.

107 http://wiki.bitcurator.net

From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions

Christopher A. Lee, Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff