Digital Forensics and Born-Digital Content in Cultural Heritage Collections

by Matthew G. Kirschenbaum
Richard Ovenden
Gabriela Redwine
with research assistance from Rachel Donahue

December 2010

Council on Library and Information Resources
Washington, D.C.

ISBN 978-1-932326-37-6
CLIR Publication No. 149
Published by:

Council on Library and Information Resources
1752 N Street, NW, Suite 800
Washington, DC 20036
Web site at http://www.clir.org

Additional copies are available for $25 each. Orders must be placed through CLIR's Web site.
This publication is also available online at http://www.clir.org/pubs/abstract/pub149abst.html.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Cover photo collage: Inside view of a hard drive, by SPBer, licensed under Creative Commons; On The Road Manuscript #3, by Thomas Hawk, licensed under Creative Commons.

Library of Congress Cataloging-in-Publication Data

Kirschenbaum, Matthew G.
Digital forensics and born-digital content in cultural heritage collections / by Matthew G. Kirschenbaum, Richard Ovenden, Gabriela Redwine ; with research assistance from Rachel Donahue.
p. cm. -- (CLIR publication ; no. 149)
Includes bibliographical references.
ISBN 978-1-932326-37-6 (alk. paper)
1. Electronic records--Management. 2. Archives--Administration. 3. Digital preservation. 4. Archives--Data processing. 5. Archives--Administration--Technological innovations. 6. Forensic sciences. 7. Humanities--Data processing. I. Ovenden, Richard. II. Redwine, Gabriela. III. Donahue, Rachel. IV. Title. V. Series.

CD974.4.K57 2010
070.5'797--dc22

2010048734

Contents

Figures

Sidebars

About the Authors

Matthew G. Kirschenbaum is associate professor in the Department of English at the University of Maryland and associate director of the Maryland Institute for Technology in the Humanities (MITH). Much of his work now focuses on the intersection between literary scholarship and born-digital cultural heritage. His first book, *Mechanisms: New Media and the Forensic Imagination*, was published by the MIT Press in 2008 and won the 16th annual Prize for a First Book from the Modern Language Association. Kirschenbaum was the principal investigator for the National Endowment for the Humanities project "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use" (2008), and is a co-principal investigator for the Preserving Virtual Worlds project, funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program and the Institute of Museum and Library Services.

Richard Ovenden is associate director and keeper of special collections of the Bodleian Libraries, University of Oxford, and a professorial fellow at St Hugh's College, Oxford. He has worked at Durham University Library, the House of Lords Library, the National Library of Scotland, and the University of Edinburgh. He has been in his present role at Oxford since 2003. He is the author of *John Thomson (1837–1920): Photographer* (1997) and *A Radical's Books* (1999). He is director of the futureArch Project at the Bodleian, and chair of the Digital Preservation Coalition.

Gabriela Redwine is archivist and electronic records/metadata specialist at the Harry Ransom Center, where she is responsible for developing and implementing digital preservation policies and procedures, processing paper-based archives, and reviewing EAD. She earned her B.A. in English from Yale University and her M.S. in Information Science and M.A. in Women's and Gender Studies from The University of Texas at Austin.

Rachel Donahue is a doctoral student at the University of Maryland's iSchool, researching the preservation of complex, interactive digital objects, especially video games; she is also a research assistant at the Maryland Institute for Technology in the Humanities (MITH). Donahue received a B.A. in English and Illustration from Juniata College in 2004, and an M.L.S. with a specialization in archival science from the University of Maryland in 2009. In 2009, she was elected for a three-year term to the Society of American Archivists' (SAA) Electronic Records Section steering committee.

Consultants

Luciana Duranti, University of British Columbia
W. Bradley Glisson, University of Glasgow
Cal Lee, University of North Carolina at Chapel Hill
Rob Maxwell, University of Maryland
Doug Reside, University of Maryland
Susan Thomas, Bodleian Libraries

Acknowledgments

The research and writing of this report, as well as the May 2010 symposium at the University of Maryland, were made possible by an award from The Andrew W. Mellon Foundation. The authors are deeply grateful for this support, and for the advice and assistance of foundation officers Helen Cullyer and Donald J. Waters. Likewise, the authors are grateful to Christa Williford, our program officer at CLIR, and to Kathlin Smith at CLIR, who expertly oversaw the copyediting and production of the report.

Rachel Donahue, an archives doctoral student at the University of Maryland's iSchool, provided research and editorial assistance throughout the project, was instrumental in organizing the May symposium, and assumed primary responsibility for compiling Appendixes A and B. Her contributions have been essential. Chris Grogan at the Maryland Institute for Technology in the Humanities oversaw our accounting. The Harry Ransom Center graciously supported our work through contributions of Gabriela Redwine's time.

Several paragraphs in sections 1.3 and 2.5 of this report first appeared in slightly different form in Kirschenbaum's *Mechanisms: New Media and the Forensic Imagination* (2008). We are grateful to the MIT Press for permission to reuse them.

We are deeply indebted to our consultants, who read and commented on our drafts, wrote sidebars, and saved us from at least some potential pratfalls: Luciana Duranti, Brad Glisson, Cal Lee, Rob Maxwell, Doug Reside, and Susan Thomas.

We are also indebted to other individuals who commented on our drafts or otherwise assisted, including Cynthia Biggers, Paul Conway, Neil Fraistat, Patricia Galloway, Simson Garfinkel, Jeremy Leighton John, Kari M. Kraus, Jerome McDonough, Michael Olson (who also authored one of the sidebars), Catherine Stollar Peters, Andrew Prescott, Virginia Raymond, and Seamus Ross.

The authors alone assume full responsibility for any errors or misstatements.

Foreword

Digital Forensics and Born-Digital Content in Cultural Heritage Collections examines digital forensics and its relevance for contemporary research. The applicability of digital forensics to archivists, curators, and others working within our cultural heritage is not necessarily intuitive. When the shared interests of digital forensics and responsibilities associated with securing and maintaining our cultural legacy are identified—preservation, extraction, documentation, and interpretation, as this report details—the correspondence between these fields of study becomes logical and compelling.

There is a palpable urgency to better understanding digital forensics as an important resource for the humanities. About 90 percent of our records today are born digital; with a similar surge in digital-based documentation in the humanities and digitally produced and versioned primary sources, interpreting, preserving, tracing, and authenticating these sources requires the greatest degree of sophistication.

This report makes many noteworthy observations. One is the porosity of our digital environment: there is little demarcation between various storage methods, delivery mechanisms, and the machines with which we access, read, and interpret our sources. There is similarly a very thin line, if any, between the kind of digital information subject to forensic analysis and that of, for example, literary or historical studies. The data, the machines, and the methods are almost aggressively agnostic, which in turn allows for such extraordinary and unprecedented interdisciplinarity.

As this report notes, whether executing a forensic analysis of a suspected criminal's hard drive or organizing and interpreting a Nobel laureate's "papers," we are tunneling through layer upon layer of abstraction. The more we can appreciate and respond to this new world of information, the more effective we will become in sustaining it and discovering new knowledge within it. This requires not only a broader recognition of complementary work in what were once considered disparate or tangential fields of study, but also building new communities of shared interest and wider discourse.

Charles Henry
President
Council on Library and Information Resources

1. Introduction

Digital forensics is an applied field originating in law enforcement, computer security, and national defense. It is concerned with discovering, authenticating, and analyzing data in digital formats to the standard of admissibility in a legal setting. While its purview was once narrow and specialized (catching black-hat hackers or white-collar cybercriminals), the increasing ubiquity of computers and electronic devices means that digital forensics is now employed in a wide variety of cases and circumstances. The floppy disk used to pinpoint the identity of the "BTK Killer" and the GPS device carried by the Washington, DC, sniper duo—both of which yielded critical trial evidence—are two high-profile examples. Digital forensics is also now routinely used in counter-terrorism and military intelligence.

While such activities may seem happily removed from the concerns of the cultural heritage sector, the methods and tools developed by forensics experts represent a novel approach to key issues and challenges in the archives and curatorial community. Libraries, special collections, and other collecting institutions increasingly receive computer storage media (and sometimes entire computers) as part of their acquisition of "papers" from contemporary artists, writers, musicians, government officials, politicians, scholars, scientists,

Fig. 1.1: An assortment of disks from the Ransom Center's collection. Photographer: Pete Smith, Harry Ransom Center, The University of Texas at Austin.

and other public figures. Smart phones, e-book readers, and other data-rich devices will surely follow. For governmental, corporate, and organizational repositories, meanwhile, the stakes are similar: ARMA International estimates that upwards of 90 percent of the records being created today are born digital (Dow 2009, xi).

The same forensics software that indexes a criminal suspect's hard drive allows the archivist to prepare a comprehensive manifest of the electronic files a donor has turned over for accession; the same hardware that allows the forensics investigator to create an algorithmically authenticated "image" of a file system allows the archivist to ensure the integrity of digital content once captured from its source media; the same data-recovery procedures that allow the specialist to discover, recover, and present as trial evidence an "erased" file may allow a scholar to reconstruct a lost or inadvertently deleted version of an electronic manuscript—and do so with enough confidence to stake reputation and career.

Digital forensics therefore offers archivists, as well as an archive's patrons, new tools, new methodologies, and new capabilities. Yet as even this brief description must suggest, digital forensics does not affect archivists' practices solely at the level of procedures and tools. Its methods and outcomes raise important legal, ethical, and hermeneutical questions about the nature of the cultural record, the boundaries between public and private knowledge, and the roles and responsibilities of donor, archivist, and the public in a new technological era.

1.1. Purpose and Audience

The purpose of this report is twofold: first, to introduce the field of digital forensics to professionals in the cultural heritage sector; and second, to explore some particular points of convergence between the interests of those charged with collecting and maintaining born-digital cultural heritage materials and those charged with collecting and maintaining legal evidence. A third purpose is implicit in the first two; namely, to serve as a catalyst for increased contact between expert personnel from these two seemingly disparate fields, thereby helping create more opportunities for knowledge exchange as well as, where appropriate, the development of shared research agendas.

Given these objectives, the primary audience for this report is professionals in the cultural heritage sector charged with preserving and providing access to born-digital content in their collections, especially in manuscript collections and in archives. We also hope that the report will be of some interest to those in legal or industry settings, not least in terms of building awareness of additional constituencies for their methods and tools. In fact, the distance between the two fields may be overstated. There are deep historical connections between the emergence of archival science and the Roman law of antiquity, founded on concepts such as chain of custody. (The forensics of modern evidentiary standards is etymologically rooted in the forensics of verbal disputation—"forensics" comes from the Latin *forensis*, "before the forum.")

Other possible audiences for this report include funders (who may be called upon to help implement the recommendations in section 4.1), depositors, and dealers, who will likely play an increasing role in valuating and brokering born-digital materials. The role of the latter in particular should not be overlooked, since it seems likely that until there is a recognized marketplace for born-digital content, archives and collections will continue to acquire it in a more or less haphazard manner.

Finally, the report ought to be of interest to scholars whose research necessitates the use of born-digital collections, and especially to textual scholars or to anyone interested in the technologies of documents or records and their storage and transmission. As high-profile examples such as the Salman Rushdie digital papers at Emory University Libraries or the Stephen Jay Gould collection at Stanford University Libraries illustrate, any scholar working on topics in literary studies, cultural studies, art, music, film, theater, history, politics, or science from the 1980s forward will likely confront born-digital materials among her primary sources. Those scholars who lack well-grounded knowledge of the technical makeup of these materials will risk unknowingly compromising or truncating their investigations.

While portions of this report are necessarily technical, the archivist who wishes to become a capable forensics practitioner will need to look elsewhere for formal education and training. We make no claim of having written a how-to guide or field manual. Under no circumstances should this report be regarded as sufficient preparation for anyone seeking to conduct a digital forensic investigation. Publications and resources for further study are listed in Appendix C.

1.2. Terminology and Scope

As Eoghan Casey notes, the term *computer forensics* is a "syntactical mess" that "uses the noun *computer* as an adjective and the adjective *forensic* as a noun" (2004, 31). *Digital forensics*, our term of choice, fares no better with regard to syntax but has become increasingly common and enjoys wider scope, encompassing devices that are not, strictly speaking, computers. *Forensic computing* is also sometimes proffered, but there the gerund presents its own issues for usage. *Digital heritage forensics* and *digital records forensics* have been suggested by Duranti (2009). Casey himself favors *digital evidence examination*, but this seems too narrowly legalistic for our purposes. We have thus opted for *digital forensics* for the sake of its inclusivity and increasingly widespread recognition. (*E-discovery* is a neighboring term that refers to locating electronic evidence in civil litigation.)

Digital forensics breaks down into several subfields. *Incident response* is the branch of computer security and forensics that deals with the first responder on the scene of an actual crime or incident. This kind of fieldwork does have some relevance to the archivist, who may be charged with collecting computers and other hardware or media from a remote site. Certain routine practices for the crime scene investigator, such as obtaining still-image and video

documentation, are useful in an archival context, where aspects of the computer's original setting (e.g., Did the user work with a tandem display?) might be relevant to later inquiries. *Intrusion detection*, meanwhile, is primarily the domain of systems administrators and security experts who work to counter active threats and collect evidence from compromised systems. Investigators working in intrusion detection are used to operating on "live" computers, meaning machines that are still turned on or connected to a network at the time of the expert's intervention. This seems an unlikely scenario for an archivist, though in the future perhaps not too far afield for a records manager, and of course archives with online content must themselves guard against hostile network-based attacks. For the most part, however, the file system will be the premier locus of activity for a practitioner employing digital forensics in a cultural heritage setting. If a complete computer (as opposed to removable media) is involved, the machine can be assumed to be turned off when it comes into the archivist's possession. *File system forensics*, as opposed to intrusion detection and incident response, will thus be our focus here.

Finally, there are the emerging domains of Web and mobile forensics, driven by the recent and rapid rise of cloud computing and Web 2.0 services and mobile devices like smart phones and personal digital assistants (PDAs). Many high-profile individuals (writers, politicians, and others likely to become donors of personal papers) lead active online lives, participating in communities like Facebook, MySpace, Flickr, Google (and using applications like Google Docs), Twitter, and even virtual worlds like Second Life. E-mail may be stored locally, in the cloud, or both. The challenges here are legal as well as technical: different Web services are governed by different end-user license agreements, and too often these do not include provisions for access even by family members or next of kin, let alone archivists. Remote backup providers like iDisk or Carbonite present the same issues. It is not difficult to foresee a time when hands-on access to a physical piece of media containing the data of interest will be the rarity for the archivist. Similarly, the growing popularity of smart phones, PDAs, tablet computers, and other devices with the potential to store all manner of information, including e-mail, text, video, voice messages, contacts, Web-browsing activity, and more, will present new challenges for the archivist in the not-too-distant future. Indeed, mobile forensics is already a major growth area in the commercial forensics industry and even in the consumer market, where readily available subscriber identity module (SIM) card readers facilitate the recovery of deleted contacts and text messages.

There are no absolute boundaries between the cloud and a local file system, or between mobile devices and a file system. Browser caches may reveal evidence of online activity, passwords for Web services may be discovered on local systems (or even on notes in the desk drawer next to them), and mobile devices may back up to a desktop or laptop computer—or the cloud. Future archivists will clearly need to contend with a fluid information ecology spanning all

current classes of devices and services. For the time being, however, especially as archivists contend with the legacy of the first several decades of personal computing, local file systems and removable media are likely to remain the primary venue for their work. Hence our focus here.

1.3. Background and Assumptions

Any field that concerns itself with the "preservation, identification, extraction, documentation, and interpretation" of recorded events would seem to require no special pleading for the attention of the archivist, scholar, or other steward of cultural heritage (Kruse and Heiser 2002, 2). Only the object of these activities—namely, digital data, which are seemingly abstract, numeric, or symbolic as opposed to embodied and material—could possibly raise questions of relevance for the cultural heritage professional. In fact, however, digital forensics forces its practitioners to confront precisely the dual identity of digital data both as an abstract, symbolic entity *and* as material marks or traces indelibly inscribed in a medium.

In the forensic sciences, the most relevant precedent for digital forensics is the field of questioned document examination, which dates to the end of the nineteenth century. Questioned document examination concerns itself with the physical evidence related to written and printed documents, especially handwriting attribution and the identification of forgeries. While digital data may seem volatile and ephemeral, gone forever at the flip of a switch or maddeningly out of reach even if the device is in the palm of one's hand, in fact stored data have a measurable physical presence in the world. Stored data are possessed of length and breadth, a fact that accounts for what is known as the *areal density* of a given piece of storage media—literally, how closely bits can be packed together on a discrete surface. (Advances in areal density are what explain the astonishing rise in the capacity of hard drives, outstripping even Moore's law, which projects that the speed of microprocessors doubles every two years.) Currently, areal density on hard drives is upwards of 100 billion bits per square inch. Some scientists argue that we are approaching the superparamagnetic limit, which is the point on the nanoscale at which the physical properties of magnetic material break down—in other words, bits can only be made so small while retaining their physical properties. While digital forensics rarely descends to this microscopic level (despite the ubiquity of magnifying glasses hovering over keyboards and hard drives in the field's iconography) the inevitable physical residue of data, known as *remanence*, is the scientific basis of all digital forensics techniques (see section 2.5.1). Even the contents of RAM memory may be subject to forensic recovery under the proper conditions. In short, there is rarely any computation without some corresponding representation in a physical medium.

Digital forensics therefore belongs to the branch of forensic science known as *trace evidence*, which owes its existence to the work of the French investigator Edmond Locard, whose famous exchange

principle may be glossed as follows: "A cross-transfer of evidence takes place whenever a criminal comes into contact with a victim, an object, or a crime scene" (Nickell and Fischer 1999, 10). Locard, a professed admirer of Sir Arthur Conan Doyle who worked out of a police laboratory in Lyons until his death in 1966, pioneered the study of hair, fibers, soil, glass, paint, and other crime scene ephemera, primarily through microscopic means. His life's work is the cornerstone of the dictum that underlies contemporary forensic science: "Every contact leaves a trace." As many malefactors have discovered, this is more, not less, true in the supposedly virtual confines of computer systems. Much hacker and cracker lore is given over to the problem of covering one's "footsteps" when operating on a system uninvited; conversely, computer security often involves uncovering traces of suspicious activity inadvertently left behind in logs and system records. The 75-cent accounting error that starts off Clifford Stoll's *The Cuckoo's Egg* (1990), a best-selling account of true computer espionage, is a classic example of Locard's exchange principle in a digital setting.

Grasping the nature of the interaction between the physical and symbolic dimensions of computation is therefore essential to understanding digital data as trace evidence. A skilled investigator is able to leverage the features of the software operating system (OS) along with the physical properties of the machine's storage media. But a comparison of digital evidence to hair, fibers, and paint chips will take us only so far. Specialists recognize that the characteristics of digital data are different from those of other forms of physical evidence, and these differences are significant for the archival practitioner as well. As probative evidence, data are clearly vulnerable to being tampered with and manipulated. Chain of custody is therefore just as important as it is in the physical world, but investigators also employ cryptographic measures to guarantee the integrity of trial data. Here then is one of the central paradoxes of information in a digital form: the same symbolic regimen that makes it susceptible to undetectable manipulation also provides the means for mathematically ensuring its integrity.

Moreover, digital evidence is almost always partial or incomplete. An investigator may be able to recover only fragments of a file; a server log might capture some aspects of an event, but not others. This, too, is not unlike the nature of evidence in the physical world, but here we must remember that there is, finally, no direct access to data without mediation through complex instrumentation or layers of interpretative software. An investigator must constantly make sure that his or her data are not changed in the mere act of collection and analysis. Brian Carrier compares gaining access to a suspect's computer with surveying a physical crime scene, and develops a comprehensive investigative model along just those lines. Crucially, he describes a computer as a doorway to a new room, or a "house where an investigator must look at thousands of objects" (Carrier and Spafford 2003, 2). The analogies seem particularly apt in the case of a magnetic hard disk, which is the default storage technology for

most contemporary systems: all manner of events, both monumental and mundane, are routinely committed to the hard disk, often without a user's knowledge or intervention. Computers today function as personal environments and extensions of self—we inhabit and customize our computers, and their desktops are the reflecting pool of our digital lives. The digital archivist, therefore, has much to learn from techniques that model the computer as a physical environment replete with potential evidence.

In preparing this report, we were struck again and again by the extent of the crossover between the archivist's world and that of the modern forensic investigator. The same concepts appear—chain of custody, for example, or "de-duping" (removing duplicate items from a collection). Specific techniques in digital forensics such as digital stratigraphy, which entails reconstructing the layers and sequence of data deposited on a particular segment of media, often manifest explicit parallels to long-standing practices in bibliography and archival description. We maintain that such parallels are not coincidental, but rather evidence of something fundamental about the study of the material past, in whatever medium or form. As early as 1985, D. F. McKenzie, in his Panizzi lectures, explicitly placed electronic content within the purview of bibliography and textual criticism, saying, "I define 'texts' to include verbal, visual, oral, and numeric data, in the form of maps, prints, and music, of archives of recorded sound, of films, videos, and *any* computer-stored information, everything in fact from epigraphy to the latest forms of discography" (1999, 13). The significance of this formulation is not just its inclusivity or specific mention of digital data. The intellectual foundation of McKenzie's entire career as a student of books in their physical form was a ruthless peeling away of the abstractions inherent in bibliographical conjecture—mere "printers of the mind," as the title of his most famous essay, an attack on key assumptions concerning what was known about the printing of certain Shakespearean texts, has it—to the material particulars of what is essentially forensic inquiry (McKenzie 1969).

This peeling away of abstractions is the modus operandi of any digital forensics investigator. There is a fiction that computing is all about numbers, specifically ones and zeros. But there are no actual ones and zeros inside the case. We have, instead, layers of abstraction, from the pixels on the screen to the magnetic traces on the disk. Just because a particular user is identified as the owner of a certain file in its metadata, for example, is no guarantee that he or she is the individual who physically laid hands on keyboard to create it. To locate and leverage—artfully, but equitably—the tipping point at which evidence extrapolated from internal states of a computer operating system becomes associated *beyond a reasonable doubt* with actions and agents in the real, physical world is the essence of the forensic investigator's challenge in the digital realm. Dan Farmer and Wietse Venema, two authorities in the field, put it this way: "As we peel away layer after layer of illusions, information becomes more and more accurate because it has undergone less and less processing.

But as we descend closer and closer toward the level of raw bits the information becomes less meaningful, because we know less and less about its purpose" (2005, 9).

In practical terms, this means we must learn to access and evaluate multiple levels of the system in order to draw reliable conclusions about the data on a given piece of media. An incorrect system clock, for example, can render a file system's date- and time-stamps unreliable. A knowledgeable observer could sometimes detect tampering on an old-fashioned automobile odometer on the basis of tell-tale signs such as a tendency for digits to "stick" at certain places; there is, however, nothing tangible to suggest that a computer's internal clock has been rolled back or reset. This does not mean that an investigator with the proper training cannot evaluate evidence from the clock effectively, either to rule out or rule in the possibility of error or tampering. On UNIX-based systems, including the Mac OS, when a file is created it is assigned a unique identifier known as an inode number. File systems assign their inode numbers sequentially. Examining the inode numbers associated with a group of files—an activity performed from the UNIX command line—can reveal whether the numbers match the creation sequence suggested by the system's date- and time-stamps. The point in this context is not the details of the procedure, but rather that peeling away one layer of abstraction (or "illusion" in Farmer and Venema's more colorful language) brings us not to absolute truth but to a further layer of computational abstraction that we can leverage against the first in order to reach a more informed evaluation about the state of the digital materials in question. Both the forensic investigator and the cultural heritage professional bear an important responsibility to avoid conjuring "users of the mind," as it were.

The practice of digital forensics is a kind of four-way modulation between abstraction and individualization, and between volatility and stability. These are not merely intersecting oppositions: collectively, they are the enabling conditions for computation in the tradition of a universal Turing machine. Farmer and Venema put it this way: "Volatility is an artifact of the abstractions that make computer systems useful" (2005, 12). To this we would add an observation about inscription and legible signs more generally: the alphabet, for example, by consolidating and abstracting earlier writing systems into a collection of some two dozen arbitrary symbols, simultaneously served to amplify the power of writing beyond measure and to open the door for error in many new guises. Whatever differences might exist in terms of the professional goals or societal function of an archivist or a scholar and a legal forensic specialist, they have in common the nature of their relationship to the unique inscriptive environment we call a computer.

1.4. Prior Work

The professional literature on digital forensics is vast (see Appendix C), as is the literature on digital preservation and manuscript

archives.[1] A comprehensive survey of either is beyond the scope of this report, so we limit ourselves here to reviewing only those prior efforts that specifically address points of convergence between the two fields.

The starting place for any cultural heritage professional interested in matters of forensics, data recovery, and storage formats is a 1999 JISC/NIPO study coauthored by Seamus Ross and Ann Gow and entitled *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*. Although more than a decade old, the report remains invaluable. In particular, the emphasis on recovery of data from obsolescent media is a welcome complement to much of the professional digital forensics literature, where the emphasis tends to be on contemporary systems and platforms (often the more cutting edge the better, as rival publishers vie to outdo one another for a share of the market). An archivist is as likely to be working with a Wang word processor as a Netbook or iPhone. Ross and Gow provide considerable detail on the physical properties of magnetic and optical storage media; they discuss emulation as a primary strategy for preserving access to migrated data as well as the experimental technique known as retargetable binary translation (RBT), an automated process for translating binary code from one platform, file format, and operating system to another; and they develop a number of case studies to demonstrate particular techniques in real-world situations. The report makes a sharp distinction between data recovery and data intelligibility; while it may be technically possible to recover patterns of bits from magnetic media, by itself this is no guarantee of their legibility or usability. Ross and Gow also rightfully insist that "archivists, librarians, and information scientists need to extend their investigations of media and studies of its durability to the scientific journals where this material is published" (Ross and Gow 1999, 6).

Perhaps the first individual to recognize the deep linkage between the archival mind-set and digital forensics methodology was Elizabeth Diamond, writing in 1994. Diamond argues persuasively for the relevance of archival training to the work of historians, constructing an analogy to the role of forensic scientists in legal settings. Yet Diamond realizes that the relationship is more than just analogy. She places particular emphasis in this regard on electronic records as an emerging class of archival object in which descriptors such as "original" and "trustworthy" are problematic: "Archivists, like forensic scientists, become expert witnesses, testifying to the nature of documents. More and more often with electronic records . . . the archivist must 'translate' the records and be able to testify that they have not been tampered with or falsified" (Diamond 1994, 142).

This research agenda has since been taken up by Luciana Duranti and others who are developing new models for combining traditional diplomatics—the centuries-old practice of evaluating the fixity, integrity, and accuracy of analog and now digital records (see the sidebar on "Diplomatics")—with digital forensics, resulting in

[1] Elizabeth H. Dow's *Electronic Records in the Manuscript Repository* (2009) is a recent, convenient introduction to the latter subject.

Diplomatics

Diplomatics is a science that was developed in France in the seventeenth century by the Benedictine monk Dom Jean Mabillon in a treatise entitled *De Re Diplomatica Libri VI* (1681) for the purpose of ascertaining the provenance and authenticity of records that attested to patrimonial rights. It later grew into a legal, historical, and philological discipline as it came to be used by lawyers to resolve disputes, by historians to interpret records, and by editors to publish medieval deeds and charters. Its name comes from the Latin word *diploma*, which was used in ancient Rome to refer to documents written on two tablets attached with a hinge, and later to any recorded deed, and it means "about records." However, over the centuries, the focus of diplomatics has expanded from its original concern with medieval deeds to an all-encompassing study of any document produced in the ordinary course of activity as a means for it and a residue of it.

It is useful to distinguish "classic diplomatics" from "modern diplomatics," because these two branches of the discipline do not represent a natural evolution of the latter from the former, but exist in parallel and focus on different objects of study. Classic diplomatics uses the concepts and methodologies developed by diplomatists living between the seventeenth and the twentieth centuries, and studies medieval charters, instruments, and deeds. Modern diplomatics has adapted, elaborated, and developed the core concepts and methodology of classic diplomatics to study modern and contemporary records of all types. Classic diplomatics studies only documents that are meant to have legal consequences and therefore requires specific documentary forms; it is defined as the knowledge of the formal rules that apply to legal records. Modern diplomatics has a broader scope; it is concerned with all documents that are created in the course of affairs of any kind, and is defined as "the discipline which studies the genesis, forms, and transmission" of records, and "their relationship with the facts represented in them and with their creator, in order to identify, evaluate, and communicate their true nature" (Duranti 1998, 45).

The primary focus of both classic and modern diplomatics is to assess the trustworthiness of records; however, the former establishes it retrospectively, looking at records issued several centuries ago, while the latter is concerned not only with establishing the trustworthiness of existing records but also with ensuring the trustworthiness of records that have yet to be created. Additionally, classic diplomatics identifies trustworthiness solely with authenticity, while modern diplomatics distinguishes several aspects of trustworthiness. For classic diplomatics, "trustworthy" records are authentic records, that is, documents written according to the practice of the time and place indicated in the text, and signed with the name(s) of the person(s) competent to create them. Modern diplomatics concerns itself with four aspects of trustworthiness: reliability, authenticity, accuracy, and authentication.

Diplomatics regards the documentary world as a system and uses a parallel system to understand and explain it. Classic diplomatists rationalized, formalized, and universalized the creation of a document identifying its relevant elements, extending their relevance in time and space, eliminating their particularities, and relating those elements to each other and to their ultimate purpose. These elements are building blocks that have an inherent order and can be analyzed in sequence from the general to the specific, following a natural method of inquiry. The building blocks used by classic diplomatists were: (1) the juridical system, which is the context of records creation; (2) the act, which is the reason for records creation; (3) the persons, which are the agents; (4) the procedures, which guide the actions and determine their documentary residue; and (5) the documentary form, which reflects the act and allows it to reach its purpose. To these five blocks, modern diplomatics has added a sixth: the archival bond. The concept of archival bond is unknown to classic diplomatics because of its focus on medieval records, the main characteristic of which was the fact that each incorporated the entire act as carried out through the acting procedure and the subsequent documentary procedure. The focus of modern diplomatics on modern records meant that one of its main concerns had to be the interrelationship that each modern record has with the previous and subsequent records that participate in the same act and/or integrated business and documentary procedure. This interrelationship, following archival theory, was called the *archival bond* by modern diplomatists, and was configured as an incremental network of relationships that links all the records of the same file and/or same series, and the same archival *fonds*.

This system of building blocks is used to carry out the analysis of the records under examination. The structure of diplomatic analysis, or criticism, as it is called by classic diplomatists, is rigorous and systematic, and may proceed from the general to the specific or vice versa, depending on the available information. The early diplomatists first separated the record from the world and

continued on next page

what Duranti terms a "digital records forensics." She offers an overview in a recent article "From Digital Diplomatics to Digital Records Forensics" (2009), emphasizing that the classification of a digital object as a "record" has implications for its admissibility as courtroom evidence. The piece has value beyond this technical discussion, however, particularly insofar as it serves as an introduction both to diplomatics and to digital forensics more generally, and makes a number of points about the special nature of records, as well as of other kinds of documents, in digital settings. This work is developed and extended at both the theoretical and practical levels in the research of the InterPARES (International Research on Permanent Authentic Records in Electronic Systems) Project, which has been funded by the Social Sciences and Humanities Research Council of Canada's Community-University Research Alliances under Duranti's direction in three phases since 1999. Case studies for the research have ranged from government records to the visual and performing arts. (The third phase of InterPARES, set to conclude in 2012, focuses on the implementation of findings from the first two, paving the way for a comprehensive legal, archival, and technical framework for the management and evaluation of electronic records.) Meanwhile, Duranti's Digital Records Forensics Project involves researchers at the University of British Columbia in a collaboration with the Vancouver Police Department, taking as one of its principal objectives development of "the theoretical and methodological content of a new discipline, called 'Digital Records Forensics,' resulting from an integration of Archival Diplomatics, Computer Forensics and the Law of Evidence with the project's newly developed knowledge."[2]

Many who have worked with born-digital materials in library and archival settings are familiar with the pioneering efforts of Jeremy Leighton John and the Digital Lives project at the British Library.[3] John was among the first to transfer techniques from digital forensics to his work recovering and archiving personal papers in a variety of computer formats and media. He has given numerous presentations

[2] See http://www.digitalrecordsforensics.org/.

[3] See http://www.bl.uk/digital-lives/.

Diplomatics *continued from prior page*

then put the two into relation, trying to understand the world through the record. Thus, they began analyzing the formal elements of the records and, from the results of such analysis, reached conclusions about procedures, persons, acts, and contexts. They firmly believed in the possibility of discovering a consistent, underlying truth about the nature of a record and of the act producing it through the use of a scientific method for analyzing its various components.

Indeed, diplomatics enables record professionals to work with a heuristic device, a diagnostic tool for establishing the meaning of the phenomenon under investigation, thereby making possible the understanding of unprecedented manifestations of records, the assessment of the trustworthiness of records that come to us at the end of several reproduction processes, and the identification of what needs to be protected and of how to ensure that a trace of our actions will be carried into the future. Thus, it can be considered the oldest form of records forensics.

—Luciana Duranti, University of British Columbia

on the topic, and the Digital Lives project's recently published final report offers extensive coverage of issues around personal digital archives and records, including several sections describing the role of forensics in their acquisition and management (John et al. 2010). The report concludes that authentication of electronic records and objects is a key application for digital forensics in archives, specifically with regard to the interpretation of date- and time-stamps, the capacity to capture authentic digital copies of the materials, and the ability to extract significant metadata from the original file system. John acknowledges the importance of informed consent by the donor as a prerequisite for forensic processing, and suggests the potential value of forensic tools to scholarly research through their ability to ascertain revision histories and other details about a document's composition. Finally, John underscores the role of forensic methods and tools in identifying forgeries, a seemingly inevitable fact of digital life.

The Bodleian Libraries, meanwhile, have been doing what are likely the most comprehensive studies to date on workflow for acquiring, processing, and making available personal papers in a variety of digital formats. The *Workbook on Digital Private Papers* produced by the Bodleian's Paradigm project remains the closest thing the archives community has to a textbook on the subject. The Paradigm *Workbook*, however, addresses digital forensics only in passing. Forensics is within the scope of the Bodleian's futureArch (Future of Archives) project (more detail is available in the sidebar on "Digital Forensics at the Bodleian Libraries"). The Digital Preservation Workflow Project (Prometheus) at the National Library of Australia is similarly engaged, with particular emphasis on creating scalable and reliable practices for the transfer of data from legacy storage media to contemporary repository systems. Stanford University Libraries, a partner (with the University of Virginia, Yale University, and Hull University) in the Mellon-funded AIMS (An Inter-institutional Model for Stewardship) project on digital papers, has acquired two forensic computing workstations for use with its collection processing, and maintains an active blog on the subject (more detail is available in the sidebar on p.30).[4] As of this writing, AIMS is still in an early stage. Finally, the PERPOS project, led by Bill Underwood at Georgia Tech, has been investigating issues related to electronic records management in the specific domain of the Presidential Records Act, and has leveraged approaches from computational linguistics and digital forensics, the latter in the area of file-format identification.

The file system and format researcher who has had the most contact to date with the cultural heritage community is Simson Garfinkel of the Naval Postgraduate School in Monterey, California, who has published a number of papers of relevance to archives and digital personal papers.[5]

---

[4] See https://lib.stanford.edu/digital-forensics for the Stanford University Libraries forensics blog and http://born-digital-archives.blogspot.com/ for the AIMS project blog.

[5] Many of these are available from Garfinkel's home page at http://simson.net/page/Main_Page.

Matthew Kirschenbaum, a coauthor of this report, has commented on digital forensics, textual scholarship, and the materiality of born-digital objects in his monograph *Mechanisms: New Media and the Forensic Imagination* (2008). In particular, Kirschenbaum argues that insights from digital forensics serve as a counterweight to many commonplace assumptions about electronic data, namely, their unqualified ephemerality, volatility, and malleability. Kirschenbaum et al. also note the promise of forensics in the white paper "Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use" (2009), prepared with support from the National Endowment for the Humanities.

Finally, *History and Electronic Artefacts* is a prescient book edited by Edward Higgs (1998) containing several contributions (Seamus Ross, R. J. Morris, Ronald Zweig, Doron Swade) that seemingly set the stage for the application of forensics in electronic cultural records and archives—such as when R. J. Morris predicts in his chapter that "much 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" (33). For an epigraph, we could do worse than this last.

1.5. About This Report

The authors undertook research and writing for this report in 2009–2010, with advice and assistance from Duranti, Glisson, Lee, Maxwell, Reside, and Thomas. In May 2010, a symposium was convened at the University of Maryland to solicit feedback and comment on a first draft of the report from a community of practitioners. Details related to the meeting's agenda and attendees, as well as a recap of its proceedings, can be found in Appendix D. Following the meeting, the authors and consultants produced a final draft of the report, which they submitted in September to the Council on Library and Information Resources (CLIR) for copyediting and publication. The authors presented overviews of the report at the Digital Lives project seminar at the British Library and at the annual partners meeting of the National Digital Information Infrastructure and Preservation Program, both in July 2010. These presentations constituted further occasions for feedback.

Section 1 of the report describes its purpose and audience, explains decisions regarding terminology and scope, provides details on the process by which this document was researched and written, and acknowledges our sources of support. It also selectively reviews relevant literature and articulates some of the issues and ideas that form the assumptions for the work that follows.

Section 2 is organized topically. It covers challenges such as legacy formats, unique and irreplaceable data, trustworthiness, authenticity, data recovery, and costing forensic work.

Section 3 considers the ethical issues that arise with forensics and their effect on archivists' relationships with current and potential donors.

Section 4 offers recommendations to the scholarly and archives communities in terms of their current and near-future engagement with digital forensics, as well as suggestions for establishing and maintaining communication between the cultural heritage sector and legal or government practitioners.

Independently authored sidebars throughout serve to amplify and extend selected topics apart from the main body of the report.

Appendixes A and B offer surveys of forensic software and hardware, respectively. Appendix C offers recommendations for further reading and study, and Appendix D summarizes the proceedings of the May 2010 meeting at the University of Maryland.

Mention of specific products or vendors, either in the body of this report or its appendixes, does not constitute endorsement by the authors or consultants, their institutions, The Andrew W. Mellon Foundation, or CLIR, and none of the preceding individuals and organizations may be held accountable for damages caused by the use of products and procedures discussed herein.

2. Challenges

Born-digital materials present challenges as multifarious as the items themselves. Issues ranging from how to identify and capture digital cultural heritage (and the related ethical concerns); to technical questions related to data integrity, accessibility, and recovery; to concerns about the cost of digital preservation projects are among the challenges that archivists, curators, and others concerned with preserving born-digital cultural heritage materials must confront. The following sections examine these and other issues in detail and discuss the benefits and drawbacks of inserting digital forensics methods into an archival workflow.

2.1. Legacy Formats

The digital media received by archival repositories often contain a combination of legacy and contemporary formats.[6] Because computers and external data-storage devices obsolesce at several levels (file format, file system, operating system, application, and hardware and media), an archivist must consider a variety of factors when developing strategies to preserve and provide access to the files on these media. Finding the hardware necessary to access older media is among the first steps, followed closely by identifying the wide range of operating and file systems these media contain and deciding on the best way to make the files accessible to researchers. This section focuses on historical, or legacy, media and the challenges they pose for digital preservation, as well as on the ways in which incorporating forensic techniques at certain points in the archival workflow can

[6] The *Oxford English Dictionary* defines "legacy" in the context of computing as "designating software or hardware which, although outdated or limiting, is an integral part of a computer system and difficult to replace." Available at http://dictionary.oed.com/ (accessed 28 January 2010).

help make the capture and identification of legacy materials more efficient and secure.[7]

2.1.1. File System

The file system controls how files are organized, named, described, and retrieved, which means that it is important not only in relation to the files themselves but also to their metadata.[8] Like hardware and operating systems, file systems continue to evolve. Because file systems dictate different file parameters, the files created in one system often differ in substantive ways from those created in another. For example, file names in some of the earlier Microsoft file systems (e.g., File Allocation Table [FAT] 12 and 16) were limited to eight characters, whereas later systems have limits between 254 and 256 characters. Another difference is the type of characters allowed in directory and file names. The Macintosh Hierarchical File System (HFS), for example, allows everything except : whereas the Windows New Technology File System (NTFS) restricts the characters / \ and : in addition to others. Similarly, some operating systems restrict the use of certain characters across all file systems: for example, DOS, Windows, and OS/2 prohibit the characters \ / : ? " > < and * among others, in file and directory names.

These differences between file systems underscore the interplay between personal practice and the parameters dictated by any particular computing system. In other words, the limitations and affordances of a particular file system have an effect on how a creator organizes and names the files—establishes a personal filing system—on her computer. Creators operate within the confines of their computing systems, yet make important and personal choices from within these imposed structures. As important expressions of a creator's naming and organizational conventions, and as reflections of the computing environment within which they were created, file and directory names and the characters that constitute them should be preserved unaltered.

File-system differences can become problematic for archivists working to capture files from original media. For example, an archivist will get an error message if she tries to copy an older Mac file with / in the file name from an original disk or computer to a Windows-formatted external hard drive that does not allow that particular character. File systems also have parameters dictating what size file can be copied. For example, an external hard drive formatted as FAT 32 only accepts files smaller than 4 gigabytes (GB). Consider the following scenario: an archivist uses the dd ("disk dump") utility to create a disk image of an entire hard drive from a modern computer.

---

[7] Some forensic software packages include functions that can be performed just as easily by stand-alone tools. For example, a freeware hex editor could be used to identify file type and glean other sorts of information. For more on the uses of hex editors, see section 2.5.

[8] For an informative overview and links to additional resources, see the Wikipedia entry for "File system" at http://en.wikipedia.org/wiki/File_system (accessed 29 January 2010). For a more in-depth explanation of file systems, see Carrier 2005, especially chapters 8 through 17.

A Digital Forensics Workflow

A generic digital forensics workflow consists of the following decisions and actions (Glisson 2009). First, one must decide where to store the information. To ensure that data remanence does not contaminate the information stored on the target drive, the target drive needs to be forensically cleaned. This entails wiping the target drive by writing all zeros or ones to it. However, the 2006 National Industry Security Program Operating Manual (also referred to as the DOD 5220.22-M) does not specify the number of passes required to achieve sanitation (Department of Defense 2006). Even though there is some disagreement regarding the effectiveness of overwriting for sanitation purposes, it is a good idea from a forensic practice perspective.

The second step is to document the hardware, including serial numbers and manufacturer information. The third step is to start the chain of custody and to transport the device to a secure lab for processing.

At this point, a bit stream copy of the removable media should be made by creating either a clone or a forensic image of the device. Write-blocking hardware or software should be employed to prevent inadvertent alteration of the original media during the copying. All write-blocking solutions should be tested and documented prior to implementation. A bit stream copy of the removable media copies every bit on the source drive (Nelson et al. 2008). Once a bit stream copy has been saved to another drive, i.e., the target drive, so that the target drive is bootable, it is commonly referred to as a clone. This is generally done using a drive that is physically identical to the source. When the bit stream copy is saved to an image file, it is commonly referred to as a forensic image. It is possible to take a forensic image and restore the image to a drive, making a clone of the source drive. At this point, the forensic copy of the removable media needs to be authenticated. This is typically done through the execution of a one-way hash on both devices to verify that they are identical.

The next issue to address is the file system. It can be argued that the file system is part of the application layer, the presentation layer, and the session layer as defined in the Open Systems Interconnection (OSI) seven-layer model (SearchNetworking.com). The file system is responsible for the organization of the files, i.e., it is responsible for the logical placement of the files on the storage drive. Hence, the file system is manipulating the sectors on a drive so that they are treated as clusters. These clusters are then linked, as needed, so that they can be treated as a file with associated metadata. The size of the clusters will vary depending on the size of the hard disk drive and the file system (Nelson et al. 2008). Understanding this interaction is critical to the retrieval of data that have been accidentally or intentionally deleted on various types of files systems like the File Allocation Table (FAT) system, New Technology File System (NTFS), High-Performance File System (HPFS), or Hierarchical File System (HFS).

The next step is to analyze the drive to identify active files and inactive files. Active files are readily identifiable and can be accessed with the appropriate software and, in some cases, the required security information. Inactive files can be located by carving the unallocated space and slack space off of the drive. Unallocated space is space that has not been used by the file system. It can contain deleted files as well. Information can also be found in two types of unallocated slack space: file slack and RAM slack (sometimes both are referred to as drive slack) (Nelson et al. 2008). Any anomalies that are identified, such as encrypted information, proprietary software formats, and missing partitions, are noted and examined individually. All information found is documented appropriately.

This detailed documentation includes all the issues that were encountered and the evidence that was discovered in the process. It also includes the methods used in the investigation, along with citations supporting the analysts' stated opinions. The detailed reports are then passed to the appropriate legal parties or agencies for examination.

–Brad Glisson, University of Glasgow, and Rob Maxwell, University of Maryland

The resulting image is 9 GB. The next step in the archivist's procedure is to use a flash drive to transfer that 9 GB file to the external hard drive used to house the repository's preservation master copies. She connects the flash drive, copies the file, and attempts to paste it into the flash drive's window, but an error message notifies her that the file is too large to be copied. The flash drive has a capacity of 32 GB, which is more than enough to accommodate the image file, so size should not be an issue; however, because the flash drive's file system is FAT 32, it only accepts files smaller than 4 GB.

These and related systems challenges will persist as new devices and strategies for storing data—for example, mobile devices, flash drives, and solid-state drives—emerge with technology to manage their contents. The file systems mentioned above were developed primarily for use on hard drives, although, like the flash drive in the previous example, there are also FAT-formatted media. Several other file systems have been developed for specific uses or media, such as ISO 9660 (including an extension for multisession CDs) and Universal Disk Format (UDF) for optical media; and ZFS, NTFS with Encrypting File System (Windows), and eCryptfs (Linux) for encrypted file systems. Each has unique characteristics that may need to be taken into account when capturing the contents of media and making choices about storage configuration.

The use of forensic technology to capture original bit copies has the potential to lessen the impact of file-system differences, at least in the initial stages of long-term preservation. To a certain degree, the disk image format may serve as a buffer between the file system of the storage environment in which the image is saved and the individual files within the image. For example, the individual files on a FAT-12 disk will be named according to the idiosyncrasies of that file system, which might not be compatible with the file system of a modern flash drive, external hard drive, or server (i.e., a repository's storage environment). But when a repository images that disk, the contents become part of a more complex directory structure. The outer layer of the structure consists of the disk image format; inside are the original FAT-12-formatted files. Because these files are contained within an image file, the file system of the storage device will interact with that image file rather than with the FAT-12-formatted files within. Ideally, this image file will be named according to a repository's conventions and will not include potentially problematic characters. As individual files and groupings of files are carved from disk images for processing (see section 2.5.3), the impact of file-system specifications on naming and organizational practices will likely resurface and influence the methods archivists use to discern and preserve them, and to store these files.

2.1.2. Operating System and Application

Legacy software, including operating systems, presents preservation challenges similar to those described above; namely, how to identify the application used to create a particular file, and then formulate a preservation strategy that does not risk fundamentally altering the

file's characteristics. A computer's OS facilitates interaction between the user and the underlying chip set as well as peripheral devices, and is also the basic environment, or host, for software applications. Software is often OS-specific; in other words, a version of a program designed for Mac OS cannot be successfully installed on a Windows machine, and vice versa. Similarly, software designed for an older operating system may not run its contemporary counterpart, which in turn means that files created using the software native to these older systems might not be accessible on current computers. For example, a word processing document created in Windows 3.1 or Mac System 7.5 might not open with a modern office suite installed on Windows 7 or OSX. And even if software is designed to be backwardly compatible, the final consumer product may not fulfill this promise. These compatibility problems arise, in part, from the different file systems supported by each OS. To access individual files and groupings of files (e.g., database, container) in their native formats, it is necessary to have a machine with an OS and application capable of reading the data the medium holds.[9]

Metadata harvesters (e.g., National Library of New Zealand [NLNZ] Metadata Extraction Tool) and batch-identification tools (e.g., Digital Record Object Identification [DROID]) can be used in conjunction with file registries such as PRONOM and the Global Digital Format Registry (GDFR) Project to identify file formats and learn more about their specifications.[10] Some tools, such as the JSTOR/Harvard Object Validation Environment (JHOVE), include automatic file-format-identification capabilities.[11] Forensic software such as the Forensic ToolKit (FTK), EnCase Forensic, and open-source alternatives such as The Sleuth Kit (see Appendix A for more detail) can also help automate the analysis of born-digital materials. They can extract and record metadata about file type, file dates, file size, and the relationships among files in a hierarchy, as well as other information. The ability of these tools to analyze data throughout a disk image will make it easier for archivists to locate all the files in a given format. For example, if analysis indicates that all the text on

[9] Alternatively, if a repository does not have access to legacy software or the means or technical knowledge to run emulated platforms, a conversion tool (e.g., ABC Amber Text Converter) could be used to transfer certain file types into other, more broadly legible file types that could be searched or skimmed to ascertain the content. OpenOffice, an open-source, freeware alternative to Microsoft Word, is also able to read files created in a wide range of legacy proprietary software formats. For a list of the formats OpenOffice can open, see the File Formats page of the OpenOffice.org Wiki, available at http://wiki.services.openoffice.org/wiki/Documentation/OOo3_User_Guides/Getting_Started/File_formats (accessed 18 August 2010).

[10] To learn more about the NLNZ Metadata Extractor, see http://www.natlib.govt.nz/services/get-advice/digital-libraries/metadata-extraction-tool (accessed 24 April 2010). To find out more about DROID, see http://freshmeat.net/projects/droid (accessed 24 April 2010). And for more about PRONOM and the Global Digital Format Registry Project (GDFR), which are in the process of combining to form the Unified Digital Formats Registry (UDFR), see http://www.nationalarchives.gov.uk/aboutapps/PRONOM/tools.htm (PRONOM, accessed 30 January 2010); http://www.gdfr.info (GDFR, accessed 11 August 2010); and http://www.udfr.org (UDFR, accessed 11 August 2010).

[11] For more about JHOVE, see http://hul.harvard.edu/jhove/ (accessed 30 January 2010).

a disk was created using WordPerfect 7, and the repository already has that particular software, it might be more cost-effective and efficient for the archivist to process that disk rather than one with files that would require purchasing additional software to access. Digital archivists can use information generated by forensic tools to make informed decisions about how best to preserve files for the long term and what time frame is realistic for providing patrons with access to the materials.

2.1.3. Hardware

Hardware can arrive at a repository in a variety of ways, and acquisitions increasingly include intact computers as well as external data-storage devices such as disks, cartridges, compact discs, memory cards, and flash drives. To capture files from legacy disks and other storage media, an archivist needs access to a workstation with compatible drives and ports (e.g., 5.25-inch floppy drive, DB-9 or DB-25 connectors—see the sidebar on "Rosetta Computers"). Several companies and organizations have developed external floppy drives, adapters, and controllers that can be connected to a modern computer via a USB port or that plug directly into existing floppy disk connectors.[12] These may provide a cheaper access alternative for repositories without the resources to invest in a full forensic workstation or those that want to give priority to capturing files from

[12] Examples include D Bit's FDADAP board, which adapts 8-inch floppy drives to work with 3.5- and 5.25-inch connectors (www.dbit.com/fdadap.html), Device Side Data's USB 5.25-inch Floppy Controller (http://shop.deviceside.com/), and the external USB 3.5-inch floppy drives offered by a variety of companies online (all accessed 11 August 2010). Jeremy Leighton John mentions additional tools in his article "Adapting Existing Technologies for Digital Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools" (2008).

Fig. 2.1: Laptops in the Ransom Center's collection. Photographer: Gabriela Redwine, Harry Ransom Center, The University of Texas at Austin.

Rosetta Computers

Migration of data from obsolete media formats is one of the most difficult problems in digital forensics. Although a clever developer can write emulators to migrate data from a disk image, physically connecting a device capable of reading an obsolete media format requires not only rare software expertise (to write drivers) but also expertise in electrical engineering and access to materials that may be difficult to obtain. It is, for instance, somewhat difficult to migrate data from a 5.25-inch Commodore 64 disk, but because the media fit physically into drives that were used by most major computer manufacturers and that operated in roughly the same way, there are now several ways to migrate data from these disks to modern PCs. A Commodore 64 data cartridge, on the other hand, is much harder to image, largely because making a physical connection between the cartridge and, say, a 2010 MacBook Pro would require an array of custom-built hardware.

In the future, there may emerge a class of archival technologists whose role it is to construct such hardware. Enterprising hobbyists have already built devices (such as the unfortunately named Catweasel or the even less mellifluous FD5025 card) for reading 5.25-inch floppy drives with twenty-first-century machines. Similar efforts will likely keep USB devices and the magnetic, rotating hard drive usable long after they vanish from consumer machines. However, historically there has been a significant lag between the time that a device becomes difficult to find and the commercial availability of custom-built bridging devices. In the interim, some of the most useful tools for migrating data from an obsolete to a modern (or at least slightly less obsolete) format are those computers that were manufactured at a moment when a popular new media format or transfer protocol had just emerged. Such computers often have ports or drives, along with associated drivers, capable of using older, and in their time more common, technologies as well as new ones. I call these liminal computers "Rosetta machines" because, like their namesake, the Rosetta Stone, they provide a translation aid for those wishing to transfer information from one encoding to another.

Examples of recent Rosetta machines are those that include readers for the multitude of flash media cards that were developed between 2000 and 2010 (Compact Flash, Sony Memory Sticks, Secure Digital, etc.) and machines that have DB-25 parallel ports and RS-232 serial ports in addition to USB ports. Earlier, and now very valuable, examples include machines that can read both 5.25-inch and 3.5-inch floppies, and Macintosh computers with "super disk drives" that can read both 800K and 1.4Mb floppies. The Rosetta machine par excellence, however, is the Macintosh Wallstreet Powerbook G3. The laptop, manufactured between May and September 1998[1], came with swappable CD, DVD,

and floppy drives capable of reading 800K and 400K disks. A swappable Zip drive could be purchased for the machine, an Ethernet port allowed data to be transferred from the computer using standard networking protocols, and PCMCIA slots permitted the addition of USB ports through a third-party card to which an external hard drive, or even flash media, could be attached. The hardware is capable of supporting older versions of Linux, and with it many contemporary open-source software packages. The machine does not natively support 5.25-inch floppies or other more archaic formats, but it does serve as an example of the sorts of machines that may prove valuable to digital preservation laboratories in the future.

Obtaining and maintaining Rosetta machines such as the Macintosh Wallstreet Powerbook G3 will be a challenge for future archivists. Today, such machines are most easily found on eBay, Craigslist, and other online advertising and auction sites; these sites and their future analogs will likely continue to be invaluable to archivists.

Once obtained, these aging machines must be kept in working order. For this reason, it is probably wise for major repositories to employ electrical engineers capable of servicing a wide range of devices (just as chemists and mechanics are regularly employed to preserve paper and magnetic media). However, since in most cases Rosetta machines are a stopgap measure—a relatively inexpensive way of accessing old media until replacement technology (such as the Kryoflux) is developed—long-term investment in any one Rosetta device is probably unnecessary. In most cases, it may be cheaper to turn to eBay for a replacement rather than to devote vast resources to maintaining idiosyncratic hardware.

—Doug Reside, University of Maryland

[1] http://lowendmac.com/pb2/wallstreet-powerbook-g3-i.html (accessed 8 September 2010).

only one media format (e.g., 3.5-inch disks). Even a preconfigured forensic workstation (e.g., Forensic Recovery of Evidence Device [FRED] by Digital Intelligence) may need to be customized to include older drives. Regardless of whether a processing workstation is constructed locally or purchased preconfigured, write protection is a necessary element. This can be as simple as flipping the write-protect tab on a 3.5-inch disk, using the command line to configure the workstation's floppy drive as read-only, or purchasing a write blocker, a device engineered to prevent data transfer to a given piece of source media.

With a computer, a repository potentially receives a complete physical environment: data files, at least some of the software necessary to read them, and contextual information at the systems level that can be helpful in learning more about the contents, the person who created them, and her working practices. In one sense, the environment is "complete," in that by the time the machine reaches a repository, the creator has finished with it. In another sense, however, it is no more possible to capture a complete computing environment than it is to transfer or acquire a complete paper-based archive. The materials a repository receives tell only a partial story. Included among each shoebox of letters, sheaf of manuscript pages, or gigabyte of computer files are the traces of absent materials—a letter that mentions an enclosed photograph long since misplaced; an editorial comment about a missing earlier draft; a reference to a labeled disk not found in the accession. A computer is a working environment that contains tantalizing traces and reminders that any single machine is part of a much larger material and virtual network and has relationships with a variety of other computers, devices, and servers not transferred to the archives.

Turning on a computer to determine whether it is functional risks writing data to the hard disk and altering the registry (see section 2.5.3). Capturing a forensic image of the hard disk, using either a version of the dd utility or imaging software, is a less invasive approach that will ensure the safety of the collection materials.[13] In the case of legacy machines, collecting older connectors, drives, and other equipment may enable archivists—individually or in collaboration with technologists—to devise strategies for capturing images of older media in the event that the methods and technology developed for use with more contemporary machines are inadequate.

2.1.4. Conclusions

The challenges presented by legacy formats are ongoing and will continue to change as technology evolves. Forensic techniques and tools will not eliminate the problems presented by older media, but they can make certain parts of the preservation process more efficient and more secure. Forensic and other tools can help archivists image

[13] For more information about the dd utility, see http://wiki.linuxquestions.org/wiki/Dd (accessed 11 August 2010). For more information about dcfldd, an updated version of dd with "features useful for forensics and security," see http://dcfldd.sourceforge.net/ (accessed 11 August 2010).

born-digital materials and determine their native formats, but how to proceed beyond that point is less clear and will likely be determined by a variety of factors, not least of which are educational opportunities, and some of which (e.g., funding, staff, equipment, institutional support) are beyond an archivist's control.

One option is to invest resources in migrating files to contemporary formats, preserving both the original bit copy and the newer representations, with the understanding that some of the formatting may be lost. Another is to use legacy media and software to make files available in their native formats so that researchers can experience the look and feel of the original materials as the creator may have last seen them. Emulation, another option, would enable archivists to run an older system using a current machine so that researchers could experience files in their native environments or, in the case of a hard disk image, interact with an emulated version of a creator's computer. The Koninklijke Bibliotheek and the Nationaal Archief of the Netherlands have pursued emulation as both a preservation and access strategy, as has the team at Emory University responsible for the Salman Rushdie Papers (van der Hoeven et al. 2007, 2:2; Loftus 2010a). The CAMiLEON project (1999–2003), undertaken by the Universities of Michigan and Leeds, also explores the issues that arise with using emulation as a preservation strategy.[14]

These and other projects raise questions about what archivists and curators need to know about legacy formats, and technology more broadly, in order to preserve born-digital materials and make them available to researchers. Do archivists and curators need information technology training to understand the hardware, software, and other details of the digital objects in their collections? Or is a collaborative model involving a variety of stakeholders with different skill sets—for example, archivists, technologists, and forensic experts—a more realistic approach? Researchers who access born-digital archival materials in repositories will also need to be equipped with certain skill sets and tools to make full use of the materials, but it remains to be determined whether the onus for supplying these resources will be on the researcher or the repository and its staff. What tools and access mechanisms (e.g., hex editor, emulated platforms, legacy OS and applications) is it reasonable for a repository to provide, and which should a researcher bring? These questions are not unique to legacy computing systems. It is not unusual, for example, for a patron to bring a portable collator in to a research collection, but should that patron also be expected to have a suite of text-analysis software installed on her laptop? Beyond access to particular skills or tools, researchers will need to be educated in the ethical boundaries of their inquiries. Access to a disk image, even one thought to be properly redacted, may inadvertently expose systems data, temporary files, or the kind of "hidden" information characteristic of files created with the Microsoft Office suite (see section 2.5.3). Without diminishing the responsibility archivists have to ensure appropriate

[14] See http://www2.si.umich.edu/CAMILEON/index.html.

redaction, it seems likely that there will be instances when scholars must exercise professional and ethical judgment as to the appropriateness of using some of the born-digital evidence to which they have access, especially when materials have been processed in batch.

2.2. Unique and Irreplaceable

The United Nations Educational, Scientific and Cultural Organization (UNESCO) defines culture as "a set of distinctive spiritual, material, intellectual and emotional features of society or a social group [that] encompasses, in addition to art and literature, lifestyles, ways of living together, values systems, traditions and beliefs" (UNESCO 2008). Historically, governments, organizations, communities, families, and individuals have identified as important different aspects of the varied traditions that comprise the cultural record, and have worked to preserve them. To a certain extent, culture arises from the patterns according to which people interact. Such relationships are not unique to sentient beings; computer files also exist as part of a complex system that defines how they relate to one another. Preserving born-digital materials means preserving not only the object itself but also its relationship to other objects, or its position as part of a larger process. Those relationships—how a file fits into a particular system, whether that system is actually the file system, a personal organizational strategy, or a much larger network—are what make each file unique and irreplaceable.

2.2.1. Materials at Risk

During the 2009 election protests in Iran, protestors and others used Twitter and YouTube to share information about the military presence on the streets and photos and videos documenting the violence as it unfolded.[15] One particularly powerful video was a YouTube clip showing Neda Agha-Soltan bleeding to death from a gunshot wound in the streets of Tehran (Fathi 2009).[16] Although this political example may seem far removed from the safe walls of some modern archival repositories, the protests in Iran generated born-digital documentation of a moment that has already proved to be of great historical importance, not only in terms of the country's political situation but also because of the unprecedented role social media and digital technology played in documenting the protests and instantaneously disseminating the information worldwide. The ability of the Internet to facilitate the spread of born-digital files, whether in textual, video, or audio form, has direct bearing on the question of what types of digital cultural heritage materials exist and are in danger of falling by the wayside. Failing to preserve ephemeral born-digital cultural artifacts—the original digital videos and photos, the tweets, the YouTube content—would mean the loss of a large swath of the primary

[15] See http://twitter.com/iranelection09 (accessed 17 March 2010).

[16] The YouTube video, formerly available at http://www.youtube.com/verify_age?&next_url=/watch%3Fv%3DOjQxq5N--Kc, "Basij shots [sic] to death a young woman June 20th," is now available only to subscribers over the age of 18.

source materials documenting the 2009 elections in Iran.[17]

Part of the challenge is that the historical and cultural value of an item, including its relationship to other events or items, is often not obvious. Failure to preserve these digital objects could result in the loss of materials whose cultural significance is not immediately apparent. Many may represent the germ of an important idea—a fragment of text, a snippet of video, or an image that inspires the development of a current or future project. The Michael Joyce Papers at the Harry Ransom Center include a newspaper clipping from the *Jackson Citizen Patriot*, dated 28 January 1978, with a black-and-white photo of snowmobilers watching their vehicle burn. It is the direct antecedent for a passage that appears nearly 10 years later in the "winter" node of Joyce's seminal hypertext work *afternoon, a story* (1987), born-digital versions of which also reside in the Joyce papers: "They stood, as if posed, all begoggled, all in helmets, nylon jumpsuits and foam injected boots, watching helplessly as a snowmobile burned in the snow before them" (Joyce 1990). This prose passage in a hypertext work that exists only in digital form not only illustrates the hybrid nature of the contemporary archives being created today but also underscores that relationships exist among different media types in the same holding. One of the primary challenges archivists and others face is figuring out how to preserve these connections—across media types as well as within a shared environment—and then represent that information to users.

Preserving relationships at the file level may become somewhat easier when the digital object is a personal computer: a contained *fonds*,[18] or record group, with file system, organizational structure, and interrelationships intact. The computers in the Salman Rushdie Papers at Emory University are an example not only of the type of acquisitions archivists and others can expect to receive more of in the near future but also of the potential for technology to transform and embody certain aspects of a creator's life. One outcome of the furor surrounding the publication of *The Satanic Verses* in 1988 and the subsequent *fatwa* was a substantial shift in Rushdie's writing practices. Speaking to Amrit Dhillon in 1995, Rushdie commented that "one of the effects of [the *fatwa*] is that it taught me to write on a computer since I had to have a way of moving my office" (Dhillon 2000, 172). As both a writing tool and an artifact, the computer itself, as well as the manuscripts, drawings, correspondence, and personalized features contained within its environment, reveals important information about Rushdie, his work, and its cultural impact. An emulated version of one of Rushdie's computers, a Macintosh Performa 5400,

[17] Nor are scholars necessarily waiting for archivists. In this instance, the HyperCities project at the University of California, Los Angeles, has launched a geodistributed, crowd-curated "collection" of images, Twitter feeds, and YouTube videos from the election and its aftermath. See http://hypercities.com/blog/2009/12/08/new-featured-collection-election-protests-in-iran/ for more details.

[18] The Society of American Archivists' Glossary of Archival and Records Terminology defines *fonds* as "the entire body of records of an organization, family, or individual that have been created and accumulated as the result of an organic process reflecting the functions of the creator." Available at http://www.archivists.org/glossary/term_details.asp?DefinitionKey=756 (accessed 17 August 2010).

has been made available to users in the reading room at Emory, in addition to a full-text-searchable database containing born-digital files and related metadata (Loftus 2010a).[19]

In the case of the Rushdie materials, the relationships among the files within the Performa's system are presented to researchers in situ rather than as file paths apparent only by looking retrospectively at the structural metadata. In many cases, however, the data on the computer or other media are only one part of a much larger organism, consisting of files, people, external storage media, and machines, that perhaps must be reconstituted from the parts rather than saved whole. Forensics can provide archivists and other information professionals with a methodology and techniques to capture as much information as possible from a piece of digital media and to properly document the initial stages of the preservation process, but many of the questions arising from the three previous examples remain open and unresolved.

2.2.2. Forensics

Forensic techniques, coupled with an acquisition such as the Rushdie computers, give archivists the ability to capture a significant portion of a creator's digital working environment and to begin to suss out the relationships of the materials contained within. Using disk imaging, it is possible to capture a bit-for-bit copy, or image, of an entire machine, including aesthetic details like desktop wallpaper and screen saver settings, organizational elements such as directory structures, metadata about individual files, and the contents of files. Additional recoverable information includes data that a creator may have left on the machine unknowingly, such as the Internet-browsing history, recycle bin contents, and hidden or temporary files, as well as items documenting the machine's relationship with other personal digital devices (e.g., cell phone, iPod, flash drive), networks, and cloud-based information. (The ethical issues raised by forensic methods of capture and analysis are addressed in section 3.)

Capturing bit-for-bit images of digital media ensures that the contents of the original media, including hidden and deleted files, will be copied in such a way that all available data are preserved intact. Files on digital media can range from the relatively simple—for example, a single-page text document with no special formatting—to the more complex, such as a hypertext manuscript of Michael Joyce's *afternoon*, a Web site or database, or, as with the Rushdie materials, an entire personal computer. But even the most "simple" documents may contain personalized elements or hidden data, both of which can have implications for long-term preservation and access. Features particular to certain types of software enable creators to customize their files. The British playwright Arnold Wesker, for example, used Microsoft Word field codes to insert date information at the top of many of his letters. Every time one of those Word files is opened, the date at the top of the letter automatically changes to

[19] For a broader view of the Emory project, see Cohen 2010.

the current date (Dong et al. 2007). Wesker's field codes illustrate but one way a file could inadvertently be changed at the moment of initial access and make a strong case for a forensics-based acquisition strategy that focuses on capturing original bit copies of born-digital archival materials before making any attempt to access the contents. In both situations, the image file acts as a container of sorts, capturing and packaging the contents so that they are not modified, until some future date when the archivist is ready to work with individual file formats or has procedures in place regarding how to handle hidden data.

Once a disk has been imaged, checksums can be used to verify that the information in the disk image matches that on the original medium. Forensic techniques ranging from image capture to complex data analysis will give archivists the ability to capture and preserve as much information as possible, and to do it more efficiently than if they were working with individually copied files. Capturing a single image file of a disk containing 100 individual files organized in a complex hierarchy is much easier and less time-consuming than copying each of the 100 files individually and then documenting a process that might well vary for each file. In addition, devising a naming convention and assigning preservation metadata to a single disk image, or even a hundred disk images of the same format, is much easier than naming and generating metadata for an assortment of individually copied files of different formats.[20] Forensic methodologies will help archivists simplify the initial stages of capture, preservation metadata, and storage so that they can capture data from digital media sooner rather than later, and consequently be able to devote more time to the later, more complex activities associated with long-term preservation. Nonetheless, it is important to remember that even a disk image is an abstraction, or more properly an interpretation, of physical phenomena on an original piece of media. The disk image is still a surrogate for the artifact.

2.3. Trustworthiness

The concept of trust, or trustworthiness, with regard to archival materials can be traced to the emergence in the sixth century of a set of criteria for distinguishing forged documents from authentic originals, which by the seventeenth century had developed into a field of study called diplomatics (Duranti 1998, 36; see the sidebar on pp. 10–11). In *Trusting Records*, Heather MacNeil breaks

[20] Although this focus on efficiency bears some resemblance to the "more product, less process" approach advocated by Mark Greene and Dennis Meissner in their 2005 *American Archivist* article, it is important to note that here we are discussing capture and storage, not processing. The potential security concerns presented by born-digital materials are serious, and we are neither proposing that repositories provide public access to forensic images without a creator's permission nor suggesting that disk images that have not been examined for sensitive information and cleared be handed over to researchers for use. Although some repositories have processed born-digital collection materials, the amount of time processing takes (or even what processing entails in the digital realm) is too variable for there to be any reliable data about average processing time for these materials.

trustworthiness down into two components: authenticity and reliability. "Reliability," she explains, "means that the record is capable of standing for the facts to which it attests, while authenticity means that the record is what it claims to be" (MacNeil 2000, xi).[21] But an authentic source may be deceptive or unreliable, and although reliability is an important component of trustworthiness, the veracity of a document's content is often not the concern of archivists working with cultural heritage materials. Rather, the provenance of both analog and digital materials, as well as documentation about their storage environment, what has been done to them, and by whom, are the key aspects of establishing and maintaining trust. Trustworthiness—of an institution, a custodian, or a document—plays an important role in the acquisition and maintenance of born-digital materials. How best to determine and document that quality in a digital environment and with regard to the stewardship of born-digital materials is a question that remains under consideration.[22] This section addresses the broad issues related to trust, or trustworthiness, with regard to born-digital materials, and in particular the role forensics can play in defining and establishing this trust. (A more detailed consideration of authenticity is undertaken in section 2.4.)

2.3.1. Tracking Trust

Trustworthiness is a concept and an obligation that spans the life of a document, whether it is a sheaf of paper or a WordPerfect file. The needs of born-digital objects shift as files move through the stages of the preservation process, from initial capture and metadata extraction to longer-term strategies such as migration and rights management. Born-digital *fonds* are similarly mobile as they pass from the creator, to an intermediary such as a dealer or other agent (human or technological), to staff at an archival repository, and, finally, to storage and, perhaps, ingest into a digital repository. The stages of that journey constitute the chain of custody for a digital object, and each stage has important implications for the trustworthiness of the born-digital materials in a given accession.

Clifford Lynch remarks in "Authenticity and Integrity in the Digital Environment" that "it is important to recognize that trust is not necessarily an absolute, but often a subjective probability that we assign case by case" (Lynch 2000, 46). This subjectivity seems particularly important with regard to cultural heritage materials, many of which are personal files created by individuals rather than records generated by the employees of an institution, and most of which pass through several hands before arriving at a repository.

---

[21] The definitions in the Society of American Archivists' glossary are slightly different. See http://www.archivists.org/glossary/.

[22] The InterPARES projects have done important work in this area. In particular, Domain 2 of the second project considered whether and in what ways concepts of reliability and authenticity are applicable across artistic, scientific, and government activities. See the InterPARES Web site for information about all three projects: http://www.interpares.org/ and the Domain 2 Task Force Report in the InterPARES II book at http://www.interpares.org/ip2/book.cfm (accessed April 2010). Also see MacNeil 2000 and Lynch 2000.

This trajectory is not all that different from that of paper materials; however, with born digital, there is a greater potential for changing digital objects—in other words, for disrupting the metadata that form one component of trustworthiness—by the very act of access. On the other hand, it may be possible to use forensic techniques to determine what has been altered and when, thus not only allowing archivists, repositories, or dealers to reestablish provenance but perhaps also enabling archivists to document the absence, as well as presence, of certain materials. What does trust look like in the digital landscape, and what is the role of the creator, or even the dealer, in establishing and transferring that trust?

2.3.2. Intermediaries

Unless a creator delivers born-digital items directly to a repository, there are intermediaries involved in the transfer process. These can include family members, rare-book and manuscript dealers, moving companies, networks and servers (if the files are transferred virtually), external hard drives or flash drives (in the case of snapshot accessions or similar capture arrangements), and others.[23] In the digital realm, the question of what trustworthy stewardship means is complicated by the potential for the mere act of opening a file or booting up a computer to alter the archival materials in fundamental ways. For example, if a dealer or a family member accesses a floppy disk after a creator's death to determine the contents, the date- and time-stamps for the opened files may reflect when that person accessed a file, rather than when the creator last read or manipulated it. When the born-digital object in question is a computer, simply turning on the machine can result in data being written to the hard drive. In other words, born-digital materials can be compromised not only physically (e.g., broken or exposed to adverse conditions), but also at the logical level (e.g., altered files and metadata). The time between when born-digital materials leave a creator's possession and when they arrive at the repository is marked by particular vulnerablity.[24]

In order for the materials to travel safely from creator to archival repository and to be documented properly, dealers and others will need to assume some level of responsibility for the trustworthiness of the digital files.[25] As digital items make up an ever-larger portion

---

[23] In "The Archival Management of Personal Records in Electronic Form: Some Suggestions" (*Archives and Manuscripts* 22 [May 1994]: 94-105), Adrian Cunningham uses the term "pre-custodial intervention" to argue for the responsible creation, management, and documentation of personal records before they arrive at a repository.

[24] Cathy Marshall notes that "changing institutional or professional affiliation is a consistent source of vulnerability for personal archives, trumping many expected problems with formats and media." In many ways, the situation Marshall describes is analogous to a transfer of digital materials from a creator to a repository. "Change makes digital belongings more vulnerable," she concludes (Marshall 2008c).

[25] At the 2009 First Digital Lives Research Conference at the British Library, the three dealers who served on the panel entitled "On the Monetary Value of Personal Digital Objects" acknowledged that at that time (February 2009) they had no formal procedures in place for valuing or handling the born-digital materials in collections. See the conference Web site at http://www.bl.uk/digital-lives/conference.html (accessed 13 April 2010).

of archival collections, dealers in particular will need to find noninvasive ways to assess the contents of digital media for representation in collection inventories and the like. In addition, once a dealer or creator has found a home for a born-digital collection, it would be ideal for the materials to reach their final destination with a documented chain of custody (perhaps even including access history) and authentication information that can be verified upon arrival.[26]

2.3.3. Repositories

Duranti notes that in the ancient world "authenticity [was] not an intrinsic character of documents but [was] accorded to them by the fact of their preservation in a designated place, a temple, public office, treasury, or archives" (Duranti 1998, 36). The location of the originals, the repository, conferred authenticity. Although other methods now exist by which to authenticate documents, physical location still confers a certain amount of weight. However, an institution that has a proven track record with regard to conserving, processing, and making available paper manuscripts—in other words, is trusted to handle traditional archival materials—is not necessarily a trustworthy custodian of digital objects.[27] Archival repositories must earn the trust of current and future digital creators. Developing a robust infrastructure and long-term preservation plan are necessary steps toward demonstrating that an archival repository and its staff are trustworthy stewards of the born-digital materials in their care.

Digital repositories within an archives or another organization should also conform to agreed-upon models or standards. In 2002, the Consultative Committee for Space Data Systems (a standards body composed of representatives from the world's major space agencies) published the *Reference Model for an Open Archival Information System (OAIS),* which suggests standards for the submission, identification, search and retrieval, migration, and more, of digital materials. The OAIS model "provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access, and for describing and comparing architectures and operations of existing and future archives."[28] It became an approved International Organization for Standardization (ISO) standard in 2003 and has been adopted by a variety of groups and institutions.[29] In 2007, the Center for Research Libraries (CRL), in collaboration with the Research Libraries

[26] The potential for forgeries in the digital age has direct bearing on many of the issues related to trust and different types of value addressed in this and other sections. The concept of the "original" (or even the "original copy") is very different and differently determined when the materials in question are born digital. It remains to be seen how the trade in (and detection of) forgeries will evolve to fit the digital landscape, and the ramifications for collecting archives, scholars, and other stakeholders.

[27] In 1996, the Task Force on Archiving of Digital Information, a joint effort of the Commission on Preservation and Access and the Research Libraries Group (RLG), concluded that "a process of certification for digital archives is needed to create an overall climate of trust about the prospects of preserving digital information" (Garrett and Waters 1996, 40).

[28] See http://public.ccsds.org/publications/RefModel.aspx (accessed 22 August 2010).

[29] See the "OAIS in Practice: Some Examples" section of the Paradigm *Workbook*. Available at http://www.paradigm.ac.uk/workbook/introduction/oais-examples.html (accessed 14 April 2010).

Digital Forensics at Stanford University Libraries

In the fall of 2008, Stanford University Libraries undertook a survey to identify the digital archival materials (handheld media) in its collections. We defined "handheld media" as materials stored offline on digital carriers of various forms and ages. The goal of the survey was to quantify the volume, distribution, and age of these materials and to identify collections at risk of loss owing to bit rot and format obsolescence. The survey identified more than 18,000 unique items of handheld media widely distributed across all bibliographic collecting areas.

The scope and scale of the challenge presented by born-digital materials on handheld media are growing. A review of statistics gathered from the accession logs of special and archival collections reveals that the percentage of Stanford collections with digital materials has increased nearly fivefold in the past five years. These materials are at great risk of loss, and without near-term action are likely to disappear from the corpus of primary source materials.

In 2009, survey results in hand, our library staff met with Jeremy Leighton John of the British Library and Susan Thomas at the Bodleian Libraries. Both graciously shared their forensic knowledge and offered recommendations about hardware and software. In the summer of 2009, Stanford University Libraries began building its own digital forensics lab (http://lib.stanford.edu/digital-forensics). Two Forensic Recovery of Evidence Devices (FREDs) were purchased, along with a copy stand and a digital SLR camera to photograph the handheld media. Licenses to commercial forensic software (Access Data's Forensic Toolkit and Guidance Software's EnCase Forensic) were purchased, and special collections staff were trained in the use of this hardware and software. The hardware was locally modified by installing a wide range of legacy drives. With these modifications, the digital forensics lab is capable of forensically imaging floppy disks of many densities and sizes, magnetic hard drives, optical discs, flash memory devices, and Iomega Zip Disks.

In fall 2009, Stanford University Libraries became a member of the AIMS Project. As part of this project, staff members are planning and testing a working model for dealing with handheld media at Stanford University Libraries & Academic Information Resources (SULAIR). The AIMS funding allowed Stanford to hire a digital archivist to staff the digital forensics lab and to begin processing born-digital manuscript collections.

To date, the computer media in the Stephen Jay Gould Papers and the Robert Creeley Papers have been forensically preserved and described using Access Data's Forensic Toolkit. These two collections contain more than 200 pieces of handheld media. We have been unable to forensically image approximately 5 percent of the handheld media in these collections because of physical damage, format incompatibilities, and bit rot. Data about the forensic imaging process are being tracked in a database with the goal of using such data to better target future preservation efforts. Another goal is to preserve all handheld media in newly acquired collections. We are developing a workflow that we hope will make this goal feasible.

—Michael Olson, Stanford University Libraries

Group (RLG) and the National Archives and Records Administration (NARA), published Trustworthy Repositories Audit & Certification (TRAC), a set of criteria by which to judge the trustworthiness of a digital repository.[30] Two recent projects in the United Kingdom, the Data Audit Framework (DAF) led by the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow in collaboration with the Digital Curation Centre, and the DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) project developed by the Digital Curation Centre and DigitalPreservationEurope, present audit methodologies, as well as other information, to help organizations better manage and curate their digital objects.[31]

Implementing models based on shared standards is one step toward becoming a trustworthy repository for born-digital materials. Adopting forensic practices geared toward establishing a chain of custody and implementing a series of checks and balances to ensure that when digital objects arrive at an archival repository they are transferred intact and with appropriate documentation are two other important steps. This level of information management is closely linked to the role of transparency in establishing an archival repository, as well as the repository it uses to manage digital objects, as a trustworthy custodian of the born-digital materials in its collection. Forensic techniques can aid archivists in the processes of capture and preliminary analysis that precede ingest into storage (e.g., external hard drive, server) or a digital repository, as well as with further analysis, file recovery, and archival processing.

2.3.4. Forensics

At the most basic level, forensic practices are geared toward establishing the authenticity of files, conducting analysis to discern their characteristics, and generating documentation about what has been done and when. Forensic methods of capture (e.g., creating disk images), authentication (comparing checksums and other metadata to verify both physical and intellectual integrity), and documentation can ensure that information is acquired from a born-digital object in a way that can be proved not to alter the original bit streams. If creators or dealers are willing to create disk images of materials themselves or allow archivists to do so, the image format will provide a protective container of sorts (in that the operating system will interact with the image file rather than its contents) that will be easier to transfer from creator to intermediaries to archival repository. Checksums generated in conjunction with the capture process can be compared by creators, intermediaries, and repositories at different stages of the transfer process to verify that the disk images and other files are exact copies of the original bit streams. In addition,

[30] See http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying (accessed January 2010).

[31] See DAF project Web site at http://www.data-audit.eu/; and Jones, Ross, and Ruusalepp 2008. See DRAMBORA project Web site at http://www.repositoryaudit.eu/; and McHugh et al. 2008.

documentation such as a list of files and their relationships within the original disk image could be verified by a repository to ensure that all the files have not only arrived with their integrity intact (i.e., hashes) but also retain their native contextual information.

Forensic tools can also be used to recover deleted or hidden files, as well as to conduct text and image searches in order to discover particular content or types of files. These analytic capabilities raise serious ethical questions. For example, how should archives handle the discovery of data—deleted files, browsing history, residual temporary files—that a creator might not have intended to include with the accession? (For an in-depth consideration of ethics, see section 3.) On the other hand, archivists can use the same forensic techniques to locate and redact information that creators have specified as "restricted" in their contracts with the repository. As more repositories move toward nontraditional acquisition strategies, such as snapshot accessions or even self-archiving, forensic tools may give archivists the ability to explain to a creator the different types of data in her born-digital archives and come to an agreement, prior to formal acquisition, about what she does and does not want to transfer to the repository. Ideally, these tools and techniques will not only help archivists establish the trustworthiness of the materials but will also help repositories build informed relationships with the creators whose digital materials are in their care.

2.4. Authenticity

One of the key challenges facing archivists and scholars who work with digital materials at any level of complexity relates to the authenticity of the digital object. Questions about authenticity have been at the heart of the scholarly process since Renaissance scholars invented the discipline of historical enquiry in its modern sense. The expectations of scholars with regard to the reliability of sources have evolved over the centuries, from the assumption that librarians and archivists would present researchers with evidence that could be relied upon to be verifiable, to more modern understandings that dispense with the ideal of the reliable source and consider all texts as potentially deceptive and richly ambiguous. Ideally, the methods of operation and processes developed by repositories over years of working with scholars and other patrons enable staff to provide researchers with documentation about the provenance and acquisition of the items in their care. This type of contextual information supports the scholarly process by providing evidence, as it were, about the documents in question. The process of assigning library or archive reference numbers to materials allows other scholars to investigate these same documents and scrutinize them anew. The conclusions that scholars draw from undertaking such studies have therefore developed a legitimacy that is intimately bound up not only with the legitimacy of the source materials that formed the basis of the initial scholarly investigation but also with the reliability of the internal systems by which the repository documents how and when

the items were acquired, their provenance, and the circumstances of their storage and organization once on-site.

To support this nexus, scholars, librarians, and archivists have developed a set of tools that have themselves spawned a number of subdisciplines: palaeography (the study of handwriting), codicology (the study of the codex or book), papyrology (the study of papyrus as a material for documentation), diplomatics (the study of documents; see the sidebar on diplomatics), and descriptive and analytical bibliography. The growth and development of these subdisciplines have encouraged scholars to specialize in them as offshoots of major subject areas such as literature and history. Librarians, archivists, and other staff, all of whom have the opportunity to interact with a broader range of documents than most scholars do, have developed related skill sets based on the utility of these subdisciplines for their own work. In recent years, librarians and archivists have worked increasingly closely with conservation professionals, who have brought yet another set of skills and experience to bear on these issues. Conservators have a more nuanced view of the materiality of books and manuscripts and have increasingly forged links with scientists who have been able to bring techniques such as PIXE analysis and Raman spectroscopy to bear in the analysis of pigments and ink, for example.

With the emergence of primary sources in digital form, the demand on the librarian and archivist to continue to support scholars by presenting them with trusted primary sources has reached a level of complexity undreamed of by the palaeographers of previous generations. The technological, ethical, conceptual, and procedural issues driving this complexity are relatively new to the humanities and information studies, and so far lack the weight of scholarly legitimacy that surrounds more traditional subdisciplines such as codicology.[32] Even in the relatively short span of their existence, the document types and file formats present in the era of the digital archive have changed with unsettling rapidity.

Scholars, librarians, and archivists tasked with preserving, describing, or analyzing born-digital materials face an additional challenge: whether and how to leverage the fundamental tools and concepts derived from scholarly traditions developed over the past 150 years, and how to incorporate those tools into methodologies designed to tackle new media formats. Since the mid-nineteenth century, such tools as catalogs of dated and datable manuscripts, guidance manuals defining and classifying scripts (handwriting), repertories of watermarks, and catalogs of book bindings, among others, have

[32] Nonetheless, "digital humanities," as it is increasingly widely known, has been practiced in some form or other for decades, going back to the text-processing experiments of Father Roberto Busa, SJ, who used early IBM mainframes to assist in the compilation of his majestic concordance to Thomas Aquinas. Many younger scholars increasingly self-identify as "digital humanists," and the Alliance of Digital Humanities Organizations runs a well-attended international conference every year (see http://digitalhumanities.org/). The digital humanities community offers much promise for future collaborations between scholars and archivists around born-digital materials.

served as reference weapons in the armory of the scholar, librarian, and archivist. The scholarly community has passed these tools, approaches, and traditions to the book trade, where reputable dealers have been able to use citations to the literature of reference works of the types described above to confirm the authenticity of the books and manuscripts that make up part of their commercial stock. In the digital domain, with a bibliography that is both younger and of necessity much smaller, the scholar, librarian, and archivist have potentially more to fear from the trade than from any time since the days of the notorious but highly skillful forger Thomas James Wise.

The scholar in the digital era needs to able to ascertain whether the digital records she is using are authentic, reliable, and usable—in short, whether they are trustworthy. The archivist is most concerned with verifying the authenticity of the born-digital materials in her care; generating documentation that other staff and scholars can easily use to trace an object's chain of custody, determine its integrity, and accurately represent its context within the larger body of materials; and, finally, making the materials available to researchers. Advances in digital watermarking, digital signatures, and other forms of electronic-document security are likely to proceed apace. But while technological tools can be deployed to authenticate a digital object, extract metadata, and make the object available to scholars, some of the most important techniques needed to establish the trustworthiness of digital objects are organizational, revolving around human behaviors such as encouraging professional ethics, recruiting staff with the appropriate skills and attitudes, and monitoring their work. (See section 3 for a more detailed consideration of these organizational challenges.) Final judgments must continue to reside with the instinct and expertise of the individual scholar.

2.4.1. Origination and Identification

The scholar, librarian, and archivist have often relied upon understanding the provenance of an item as a fundamental tool in assessing the trustworthiness of a book or manuscript as a carrier of information. In 1985, Roger Stoddard, then curator of rare books at Harvard University's Houghton Library, wrote a seminal essay in which he stated that "as anthropologists have discovered, traces of wear can tell us how artifacts were used by human beings. Books," he continued, "no less than tools, apparel and habitats can show signs of wear, but their markings can be far more eloquent of manufacturing processes, specific of provenance, telling of human relations and suggestive of human thought" (Stoddard 1985, 1). This understanding of provenance has been adopted in the trade in books and manuscripts as a critical factor in assessing price.

In the digital world, provenance is much harder to establish. Very often, the most that can be done is to document the circumstances in which the digital files were acquired; for this reason, it is often crucial that visits and interviews by an archivist from an acquiring institution be made in order to judge the context and environment from which data are being acquired. The same questions

addressed by the archivist undertaking an acquisition exercise can be used subsequently by scholars examining digital records stored in an archive in order to establish information about their veracity:

- Have the files been acquired directly from the data creator? If so, the supporting documentation should be completed so that scholars working with the files can see that there have been no opportunities to interfere with the files in an unauthorized or unidentified manner.
- What was the computing environment in which the digital records were created and stored? Was it managed in a way that minimized the risk of unauthorized intervention? Have the files been created in an offline computing environment? If so, what are the implications of this?

When it is not possible to answer these questions, physical examination of the hardware and media that carry the files may indicate whether they have been in compromising circumstances.

2.4.2. Data Integrity and Fixity

The advent of born-digital materials in archival repositories adds another layer of complexity to what has long been a primary concern for scholars, archivists, and others working with manuscript materials: the physical and intellectual integrity of the items in a collection. Although generating fixity information such as hash functions (e.g., MD5, SHA-1) is relatively easy, managing that information over time is harder as documentation of the authenticity and integrity of digital objects poses administrative challenges with direct bearing on the needs of scholars and other patrons. The scholar analyzing a primary source must know whether it has been modified since its creation and completion as a carrier of information. In the case of many documents, the addition of subsequent layers of information is unproblematic, as revision by the creator and others indicates subsequent use, consideration, or elaboration. Take, for example, the Harry Ransom Center's copy of *The Georgiad*, South African writer Roy Campbell's poetic attack on Vita Sackville-West and her circle, which he wrote after she seduced his wife, Mary. The Center's printed copy of the poem contains Campbell's handwritten elaboration of his accusations—a secondary layer of text that provides additional meaning. Such additions can often be identified through provenance analysis, therefore attributed to known individuals, and assessed appropriately by scholars.

In the case of digital archives, however, it is easier both to identify and conceal surreptitious interventions to the original text. On the one hand, this facility potentially creates severe (some would say interesting) problems for future researchers; on the other hand, because digital files often conceal as much information as they present, forensic analysis might make it possible to recover additional information that would allow researchers to attribute certain changes to one person rather than another. The Ransom Center's Thomas Zigal Papers include Microsoft Word page proofs of Zigal's novel

Digital Forensics at the Bodleian Libraries

The Bodleian Libraries collect archives from a wide range of individuals and organizations. Like other research libraries, we find that accessions of modern manuscripts contain an increasing quantity of digital material; our policy has been to encourage this in the hope of obtaining the best-possible record for the scholarly community.

The Clutag Press archive is a good example of a modern archive collection. From the beginning, its contents have been hybrid, consisting of paper and digital materials in equal measure. Tranches of the archive have been collected regularly since 2007; these include disks and digital transfers as well as the papers that are more familiar to archivists. Working with hybrid collections such as Clutag has been a means of gaining valuable experience in digital curation, and we are building on this experience as we develop the Bodleian Electronic Archives and Manuscripts (BEAM) service to curate digital archives.

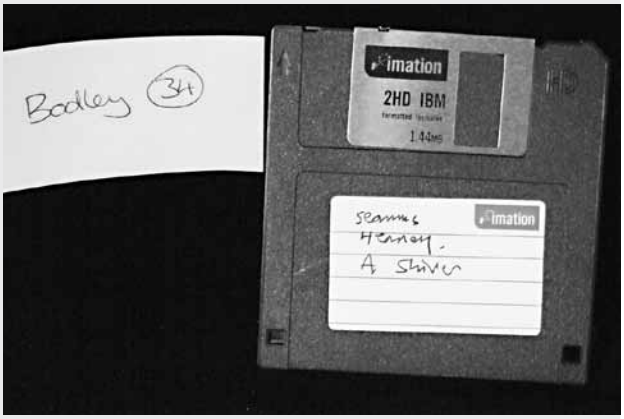
Digital forensics tools are one part of BEAM's armory. We use these tools as archivists rather than as forensic examiners, and our purpose in using them is to support activities in the archival workflow. Digital forensics tools are of interest to us because they are designed to maintain the authenticity of source data and are capable of dealing with large volumes of heterogeneous data, such as we find in personal or corporate digital archives. To illustrate our use of digital forensics tools, the following paragraphs consider some steps of the workflow typically applied to a hybrid archive.

Separation

Since remodeling the Western Manuscripts Department's accessioning workflow, we now separate digital storage media from related papers and feed them through to BEAM. This ensures that their contents can be captured and ingested to the BEAM preservation store at the earliest possible opportunity. BEAM provides integrity monitoring of the original submissions, and improvements in tools to combat technical obsolescence are planned. Digital and traditional materials remain linked via a collections-management database and will be further reunited through the archive's finding aid once cataloging of the hybrid archive is complete.

Capture

The capture process for data residing on the creator's own original disks involves assigning each disk an inventory number, which ensures that we can track it and related metadata over time. At this stage we also photograph disks to create a visual record of annotations for future catalogers and researchers.



3.5-inch disk submitted to the Clutag Press by Seamus Heaney.

Next, we use forensic hardware and software tools to create a disk image of each item. The process varies slightly according to the type of medium, but in each case we use a write-blocking device to ensure that no data are written back to the source disk. The image below shows a hard disk being imaged.

A hard disk is imaged using one of BEAM's "capture" workstations.

Finally, we verify the disk image to ensure we have an accurate representation of the data inscribed on the original media.

continued on next page

The White League (2005) with tracked changes showing not only the editor's comments and edits (with date- and time-stamps) but also Zigal's responses to her changes. Born-digital documents are as unstable and as fecund as any historic text, and part of the researcher's challenge will be to reconcile the affordances with the ambiguities and to figure out how to continue to uphold scholarly traditions, such as the inclusion of verifiable references, while making room for new practices. Martin Joseph Routh (1755–1854), when president of Magdalen College Oxford, encouraged his students to "always verify your references." This scholarly process remains a cornerstone of academic and professional life, but takes on new meaning in an age of online databases and other digital resources (Burgon, 1888–9, 1:73).

If the references cannot be trusted, the conclusions drawn from an analysis of them are rendered similarly questionable. In recognition of this fundamental element of scholarship, core aspects of digital library architecture, such as the OAIS Reference Model (CCSDS 2002), have been designed to incorporate "fixity information," or information documenting authentication mechanisms such as hashes

Bodleian *continued from prior page*

In sum, this capture process gives us:

- a reliable copy of each disk's contents wrapped in a single digital object—the disk image—and a unique hash value for that object
- an inventory of the disk's contents, including hashes and format metadata for each object
- metadata about the imaging process
- photographs of the disks.

This material is packaged in a specific directory structure and ingested to our preservation store with some basic collection and accession metadata.

Materials Captured On-site

Some archives include digital material that has been captured on-site by the digital archivist. This often requires more-selective disk imaging, with the depositor outlining which materials she wishes to add to the archive. In these cases, targeted data are acquired using a USB write-blocking device and a disk-imaging utility, with an external hard disk serving as a storage medium for the data in transit. After the data have been returned to BEAM, they are processed in a similar way as are items received on removable media.

Digital Forensics for Analysis

BEAM also uses digital forensics analysis tools. These tools are used to gain an initial overview of formats in the archive, to identify problem items, and to prepare the archive for use by a cataloger. All the disk images associated with a given collection can be added to a "case," permitting the material to be interrogated in many useful ways. The user can browse the creator's structure, browse by format, sort by various file system attributes, filter data according to numerous criteria, label or bookmark data, and perform full-text searches. The tool can render many items without the need for separate application software, which can be useful for previewing data in legacy formats.

To prepare the digital material for a cataloger, the digital archivist marks up material that the cataloger can safely ignore (such as clip art or software) and notes duplicate items, etc. The digital archivist can then export copies of files suitable for arrangement and description, and migrate them to more modern formats when necessary. The analysis tool is also used to generate a table of metadata that the cataloger will augment with descriptive and other information. Important work done by the cataloger includes allocating items to archival series within the finding aid, marking items for disposal, assigning identifiers for citation, and providing metadata concerning restrictions. If the cataloging process identifies material for disposal, then the data in the BEAM preservation store must later be updated to reflect this change.

Digital Forensics for Researchers?

BEAM's use of forensics tools ends here for the time being. The library has yet to encounter a depositor who is willing to permit researchers access to a reinstantiated version of his or her entire computing environment or to its indexed full contents. This remains a possibility for the future with the right donor.

—Susan Thomas, Bodleian Libraries

and digital signatures, in order to facilitate verification of the integrity of digital materials. This approach forms part of the core elements of the OAIS model, which has guided the architectural development of many digital repositories, and in particular the Submission Information Package (SIP) and Archival Information Package (AIP). However, other techniques, such as digital stratigraphy and the analysis of mutual referencing, can also assist in the verification of digital information. The frequency of citation and quotation of a digital document will help establish that information as a trusted source in the scholarly world and render subsequent unauthorized changes harder to enact without identification.

Managing fixity information and monitoring data integrity is an ongoing process that accompanies a digital object as it moves from creator, to an archival repository (acquisition, accession), and then to a digital repository (ingest) or other storage. One way of approaching issues relating to data integrity is to examine documents twice: prior to and after accessioning into a formal archive.

2.4.3. Preaccession

The transfer process for digital materials needs to be managed carefully, and with rigorous adherence to documented procedures incorporating standard elements of archival accessioning that have been adapted to the needs of digital objects. Transferring archives from creator to repository in a managed way will minimize the risk of interference with the material, and will allow the archivist to ensure that what is accessioned into an archival collection can be monitored and kept intact. This process, as described in the Paradigm *Workbook* (2008), involves documentation through the completion of a transfer list—which collects details about ownership, context, permissions, and technical characteristics—and technical processing, including, most importantly, generation of checksums for comparison in future integrity checks. Such documentation is no guarantee that the digital records have not been interfered with prior to their arrival at the repository, but it provides a terminus ad quem that will enable the archivist responsible for the materials to offer documentation about their status at the point of accession. Scholars, in turn, can use this documentation to determine whether to consider as trustworthy the resources in question. The role of metadata that record fixity information can extend to data creators, who could, for example, set encryption keys to provide security controls. An encryption key would provide a means of ensuring the integrity of a digital record as it passes from the creator to the recipient archive; however, this added layer would be complex and expensive to administer, and thus may be feasible only for organizations rather than individuals.

2.4.4. Postaccession

Ensuring the integrity of data after accession requires adherence to good archival practice, the rigorous application of technical procedures to ensure that the digital records are not interfered with or altered once in the custody of the repository, and the active

intervention of digital archivists at various stages in the archival life cycle. These archival processes include rigorous maintenance of the metadata required to manage both the use and administration of the digital records. Two of the most widely used metadata schemas for digital libraries, PREMIS and METS, include provision for recording fixity information: hash functions, checksums, cryptographic hash functions, cyclical redundancy checks, and digital signatures. (Such information is also a mandatory component of the Data Audit Framework suggested by Jones, Ross, and Ruusalepp [2009].) The elements suggested by these schemas can be used as the basis for a simple database to help an archivist manage the metadata for digital objects in a repository's collection, or they can be incorporated into the intellectual framework underlying a digital repository bought or built to manage and deliver digital objects. The Paradigm *Workbook* provides useful guidance on the use of metadata (see especially section 5 of the workbook).

Some archival institutions maintain "dark" archives. As a rule, such archives are not regularly and routinely accessed, and may be maintained under strict security by staff. Dark archives may be created because depositors wish for the contents to be kept hidden from view until after a certain period of time has expired (such as a period of time following the death of the depositor or of other individuals). Sometimes dark archives are kept offline in order to create an additional security mechanism. In these situations, physical security becomes a prime means of ensuring that the materials are not interfered with in an unauthorized manner, sometimes to the extent of keeping hardware such as servers locked in a strong room or vault.

2.5. Data Recovery

Recovering lost, eradicated, or obfuscated data from a computer system is without a doubt the most dramatic act a forensic investigator can perform—the stuff of suspense thrillers and film or television drama. It is also the most invasive and ethically fraught act, since the procedure may run counter to a donor's wishes and intentions. And it is potentially the most misleading aspect of digital forensics, since recovering deleted or hidden data can promote the illusion that the evidentiary record is now complete. In fact, a seasoned forensic investigator knows the same thing that any good scholar or archivist knows: there is always more to discover, and what survives of the past is only at best a partial representation of reality.

Unlike a forensic investigation undertaken in a legal setting, where individual wishes and intentions are subsumed by jurisprudence, in a manuscript collection or other cultural heritage institution any forensic process is voluntary at some level. If an archivist and a forensic investigator find themselves equipped with the same tools and capabilities, the former is in a much more precarious ethical position. Sometimes, of course, a donor may sanction a repository's efforts to recover deleted material, either because it was deleted inadvertently or because the donor did not appreciate its value

at the time. But what if the donor's wishes are not known? Does an archivist have a de facto responsibility to attempt to perform data recovery in order to provide the most complete assessment possible of the content on a given piece of media? What if a scholar requests that a hard drive be examined for deleted versions of a manuscript he is studying? To address such questions, we must first provide an overview of how data recovery is done, both general principles and specific practices, while working to separate fact from folklore.

2.5.1. Remanence

A defining characteristic of digital storage is the seeming paradox of its volatility coupled with near-miraculous stories of successful data recovery under the most extraordinary circumstances. The Internet abounds with tales of data rescued from computers submerged underwater, scorched in house fires, or run over by trucks. Some of the most colorful of these tales come from the Web site of a company called DriveSavers, which includes celebrity testimonials from the likes of Keith Richards, Sean Connery, Sarah Jessica Parker, Sting, and many more. DriveSavers was responsible for the recovery of the scripts for a dozen then-unproduced episodes of the hit TV show *The Simpsons*, including the season finale "Who Shot Mr. Burns," which has since become a fan favorite.[33] In a more sober vein, Ross and Gow (1999) chronicle the successful efforts to recover data from the *Challenger*'s flight recorders, while Kirschenbaum (2008) narrates the restoration of data from storage media salvaged from the ruins of the World Trade Center.

The remarkable staying power of data stored in digital form is a function of the physical property of magnetic media known as *hysteresis*, or its capacity to retain a charge over time. Magnetic storage media have been the mainstay of the computer industry from early experiments with magnetic drums and ringlets in the 1940s, to magnetic tape, to the introduction of floppy disks and the current ubiquity of hard drives. Most storage media that an archivist is likely to encounter will be magnetic, though optical devices such as CD-ROMs and solid-state (flash) memory will also be present as part of collections. These, too, are often surprisingly resilient in the face of fire, flood, and other disasters that would have spelled doom for their paper precursors.

Despite their supposedly pure symbolic nature (bits are often described as "switches" or "ones and zeros") all digital data are stored as physical traces or inscriptions. This means that given sufficient time and resources, data can often be recovered even if the supporting medium has been traumatized. Governments and other institutions with secrets to keep have recognized this phenomenon, and procedures for securely erasing ("wiping") magnetic media have long been pursued. In *A Guide to Understanding Data Remanence in Automated Information Systems*, first published by the National Computer Security Center in 1991, *data remanence* is defined as "the

[33] See http://www.drivesaversdatarecovery.com/company-info/hall-of-fame/.

residual physical representation of data that has been in some way erased." The report states, "After storage media is erased there may be some physical characteristics that allow data to be reconstructed." It indicates that the "problem" of remanence has been understood since the early 1960s. The Department of Defense's Clearing and Sanitization Matrix (DoD 5220.22-M) articulates a range of solutions for various magnetic and optical storage media, including degaussing and overwriting the media with random patterns. For the most sensitive materials, the recommendation is to "disintegrate, incinerate, pulverize, shred, or melt." (Cathode-ray tube displays are also to be inspected for burn-in, and scraps from paper shredders are to be destroyed.) In the civilian sector, by contrast, computer storage media are often discarded without even basic steps toward sanitization. Garfinkel and Shelat (2003) surveyed 158 hard drives acquired on the secondary market and found that the vast majority contained at least some sensitive data accessible through minimal effort.

In 1996, Peter Gutmann, a researcher at the University of Auckland, disseminated a widely read paper entitled "Secure Deletion of Data from Magnetic and Solid-State Memory." He detailed a then-obscure data-recovery process known as magnetic force microscopy (MFM), which uses a variation on the scanning tunneling electron microscope to image patterns of magnetic fluctuation on the surfaces of computer storage media.

Using MFM technology, Gutmann suggested, an investigator could exploit both the hysteretic properties of magnetic media and registration problems resulting in palimpsest-like deposits of magnetic data (a function of the drive head never writing in exactly the same place twice) to reconstruct files that had been subjected to

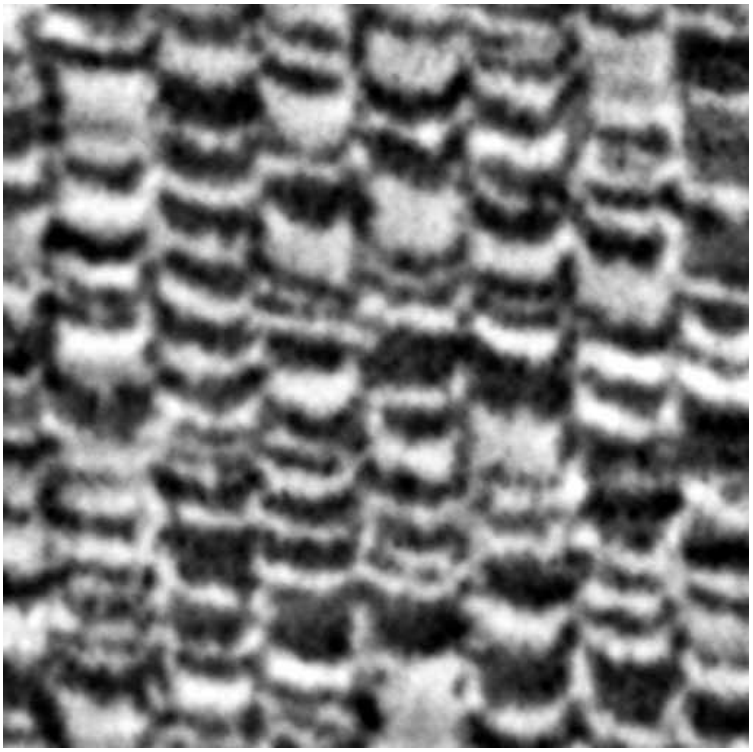


Fig. 2.2: Magnetic Force Microscopy image of data on the surface of a hard disk; the area depicted is one micron square. Wikimedia Commons.

many generations of overwrites. The dramatic implication was that data might be recovered from any magnetic medium that had not been subjected to total physical destruction, and that, moreover, the practice might be far more common than acknowledged: "Even for a relatively inexperienced user the time to start getting images of the data on a drive platter is about 5 minutes. To start getting useful images of a particular track requires more than a passing knowledge of disk formats, but these are well-documented, and once the correct location on the platter is found a single image would take approximately 2–10 minutes depending on the skill of the operator and the resolution required." To counter this supposed capability, Gutmann detailed 35 different sequences of random data designed to systematically overcome the hysteretic tendencies of magnetic media and reset their polarity back to neutral, thereby rendering MFM ineffective. These 35 patterns have since become canonical, and many disk-erasure utilities include them as an option.

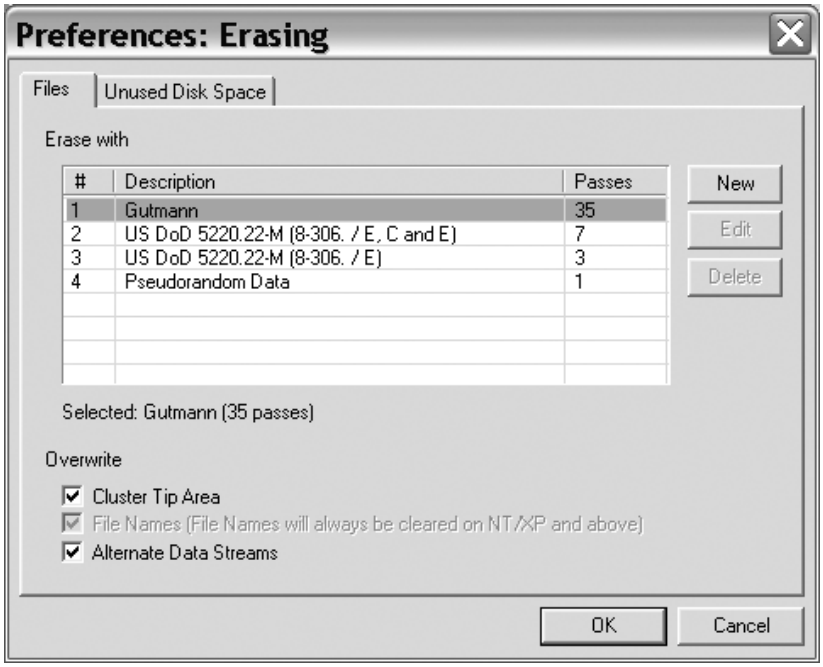


Fig. 2.3: Available settings in a common Windows file erase utility, showing options for various Department of Defense standards-compliant overwrites, as well as the 35 Gutmann patterns.

The benefits of going to such extremes, however, are dubious, and the exact capabilities of MFM are still not widely understood. The technique probably has more application in industry (for study of the integrity of magnetic media) than in forensic investigation. The simple truth is that no documented instances of a complete file retrieved via MFM sampling are known to exist. A recent paper by Wright, Kleiman, and Sundhar (2008) argues that Gutmann's conclusions are overblown, and demonstrates that in practice it seldom takes more than one pass ("wipe") to sanitize a magnetic hard disk. While reading this recondite literature may make the archivist feel a little like a character on *CSI*, a basic understanding of the physical properties of magnetic and other storage media is essential for anyone responsible for born-digital collections. It also serves to

demystify computer storage media, since even information that is not visible to the naked eye (nor legible to us if it were) is still a part of the physical, material world and thereby subject to its laws.

2.5.2. File Systems

Hysteresis and the physical properties of computer storage media are the foundation for understanding what data recovery can and cannot do, but the archivist also requires a working knowledge of computer file systems and their relationship to different operating systems. Many users know that when they delete a file from their trash or recycle bin it is not immediately expunged from their hard drive. The "delete" command simply tells the file system to make the clusters associated with a given file available again for future use (a special hex character [E5h]) is affixed to the beginning of the file name), but the data stay intact on the platter. This explains the standard injunction to allow as little time as possible to elapse before attempting to restore a lost file. File-recovery utilities work by removing the special character and restoring a file's status as allocated clusters. Because a file's physical storage location will change each time it is opened and modified, its earlier incarnations will also persist until such time as those data may be overwritten. It is not the case that as hard drive capacities continue to increase, information will persist for longer and longer periods of time on the surface of the disk: the writing and rewriting of data to storage is not evenly distributed. Even on a very large drive, a great deal of data may be overwritten very quickly if the operations performed on the machine over time follow consistent patterns. For example, the data on a Web server may be overwritten very quickly, even if the hard drive is huge, because the HTTP operations use similar spots on the drive.

The layperson's view of a file system—accessed from the desktop through standard directory structures and tree menus—is both optimized and impoverished, a partial and simplistic window onto the diverse data deposits that have accumulated on the media. Creating a file and saving it to a hard drive does not yield a simple, one-to-one correspondence between the file and its record on the disk. This fact gives rise to what is sometimes known as *ambient data*. For example, word processors and other productivity software routinely include an auto-save function that writes a temporary snapshot of an open file to the disk at set intervals. Most computers also use a portion of their hard disk as an extension of their RAM, a type of storage known as virtual memory or *swap space*. Forensic investigators recover all manner of otherwise "ephemeral" matter, including passwords and encryption keys, from the swap space. So-called *slack space* (not to be confused with swap space) presents yet another opportunity for extracting remnants of supposedly long-discarded files: data on a magnetic hard drive are stored in clusters of a fixed length (4096 bytes is typical for NTFS and FAT file systems). This is what accounts for the discrepancy between the actual size of a file and its "size on disk," as revealed by Windows Properties; even a 1-byte file—a single ASCII character—will require the allocation of a full

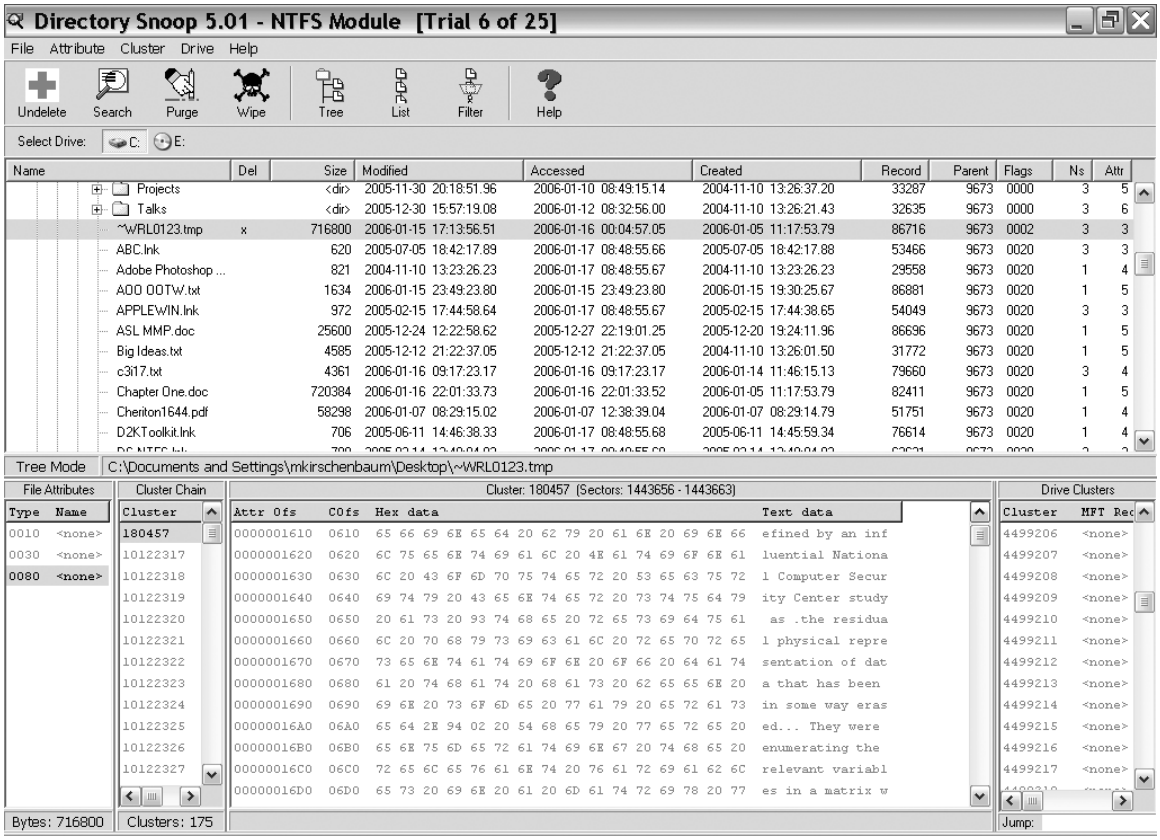


Fig. 2.4: A hex utility revealing the text of a "deleted" document on a Windows file system.

cluster, 4096 bytes, to store. But since files themselves are rarely the exact same size, and hence occupy variable numbers of clusters, it is also frequently possible to find the partial remains of earlier files at the end of a cluster chain, a phenomenon known as *slack*.

A skilled investigator develops instincts for where slack space is to be found. In addition to temporary copies and other multiples of the actual file, metadata (the name of the file, the file type, date- and time-stamps) proliferate even more aggressively through the operating system. Therefore, even if the content of a file has been completely erased it is still possible to recover evidence that testifies to its past presence.

The interactions of modern productivity software and mature physical storage media such as a hard drive may finally resemble a cybernetic pinball machine, with a single, simple input by the user sending files careening through the internal mechanisms of the operating system, these files leaving persistent versions of themselves behind at every point they touch—like afterimages that only gradually fade—and the persistent versions themselves creating versions that multiply in a like manner through the system. There is, in short, no simple way to know how many instances of a "single" file are residing in how many states in how many locations at any given moment in the operating system. Likewise, there is no simple way to know how many metadata records of a file (or any of its ambient versions) exist. For these reasons, it is not hard to see why one expert concludes that "secure file deletion on Windows platforms is a major

exercise, and can only be part of a secure 'wipe' of one's entire hard disk. Anything less than that is likely to leave discoverable electronic evidence behind" (Caloyannides 2001, 28). Other platforms, including Linux and the Mac OS, present their own challenges and idiosyncrasies in this regard.

2.5.3. Forensics

A number of software tools and utilities are now available that allow a trained forensic investigator to exploit these conditions. Precise features and capabilities are covered in Appendix A. One common method, known as data carving, treats the data deposited on the physical media as an entity that can be searched and examined regardless of abstractions imposed by the file system. Data stored in slack space (as described above), for example, become visible and accessible to the forensic software's search-and-discovery functions. An archivist will need to become familiar with the conventions of hexadecimal notation, which provides a mapping of data's physical location on the media in question. Once we are no longer accessing and retrieving files via the operating system, we can examine fragments of files that are missing their identifying headers or other key structural elements. Thus, while some portions of a file may have been overwritten, other pieces may persist elsewhere on the media. These would not be discoverable through the normal operating system. Likewise, dispersed fragments of image files may be reconstructed through the viewers in various forensics packages, even if the file in its entirety cannot be retrieved. Hex itself is nothing more than a shorthand convenience, a concession to the human visual system as a way of notating binary numbers (which are themselves symbolic or shorthand conveniences for voltage differentials in the computer's circuitry). With practice, one can use a hex editor to pinpoint and inspect the bytes recorded at any location on the virtual surface of the disk image. Likewise, strings of character data can be searched and mined through automated retrieval. A skilled investigator can also draw conclusions about the time when particular files were created, modified, and perhaps deleted on the basis of an evaluation of the contents of the disk's slack space. This procedure is known as digital stratigraphy, and dramatically makes the point that digital inscription is ultimately a physical and material process, not so different from the manner in which a conservator can reconstruct a palimpsest of writing from a physical piece of paper or parchment.

One can also discover a wide array of information that is embedded in an individual file but not always visible when rendering the file through its default application. Such information includes comments within the code, stored rules and styles, change-tracking information, metadata (often stored in file headers), executable code (including viruses), and other information that a user intended to remove ("crop") from a file. The Microsoft Office suite offers a case in point. Forms of data that may be hidden from users include information about the application used to create a document; authors, user

names, organizational affiliations, and author history; comments; custom properties; database queries; embedded objects (OLE); Fast Save—that is, change history appended to the end of a file, rather than applied to the body of a document; the GUID (Globally Unique Identifier) for the originating computer; hidden cells, slides, and text (purposely hidden but then possibly forgotten); Outlook (e-mail) properties and routing slips; path information, including audio and video paths, author history, linked objects, printers, hyperlinks, and included fields; presentation notes; printer driver information; RSID (Revision Save ID), which differentiates changes from different editing sessions); tracked changes (added to PowerPoint and Excel in Office XP) versions; Visual Basic code, including macros and viruses (and the identity of the code's creators); Web server information; and white text (on white background).[34] Thus, regardless of how an archivist chooses to handle the recovery of slack data associated with disk images, she will likely still have an ethical obligation to address many forms of hidden data within the files themselves.

Beyond files and file systems, an archivist may need to acquire at least a basic familiarity with the registry. On Windows machines, the registry is a database that maintains configuration information about applications and devices attached to the computer. (Other systems lack the equivalent of a registry and store their configuration information in different formats, often plain text.) Specific holdings within the registry are stored in binary arrangements known as *keys*. Examples of information that can be gleaned from an exploration of the registry keys (using tools such as "regedit") include paths to external drives that are or have at one time been mounted on the system, settings for program applications, and lists of recently opened or saved files. Mastering the cryptic confines of the registry is a subspecialty all its own, and one should never access the registry of a system that has not been properly duplicated for risk of doing irreparable damage. Carvey (2009) has written the best guide to date to forensic analysis of the Windows registry.

2.5.4. Conclusions

The erasure and recovery of data go to the heart of some of our most cherished considerations of what it means to have a cultural record. Presence and absence of information is not simply an ambient or accidental condition, but rather one that may be managed and imposed through the selective use of procedures and tools. The sheer volume of data available on computer storage media, coupled with the extent to which computers function as the locus for many different aspects of our lives, promotes these considerations to new levels of urgency. Data recovery also collides directly with questions of intentionality, and, therefore, with issues of personal privacy and archival ethics. It seems likely, therefore, that policies regarding data recovery and forensic exploration will have to become part of future donor agreements, so that an originator's intentions are rendered as unambiguously as possible.

[34] Our thanks to Cal Lee for this list of potential hidden information in MS Office files.

Clearly there will also be situations in which even a donor's stated intentions may not be the final determining factor. The Freedom of Information Act, subpoenas, or other legal instruments may impinge on a donor's intentions. Societal attitudes may change, such that given sufficient passage of time the potential value of information may outweigh whatever restrictions were once placed upon its use. (Who among us could say that if we discovered one of Shakespeare's missing manuscripts today we would forgo the opportunity to resurrect the lost text of a play, even if we somehow knew it was against the Bard's intentions?) It is also worth understanding that not all potential for data recovery lies with archivists: a patron with access to bit stream disk images, for example, may be able to undertake a forensic exploration without the knowledge of the archivist. This, too, must be considered as repositories decide about what kinds of access to born-digital materials to provide scholars.

2.6. Costing

The costs of providing an infrastructure for digital forensics are not yet well understood; however, several projects and institutions have produced reports describing their investigations into the costs of digital preservation. The CEDARS Project, the Digital Preservation Coalition, the British Library, the Koninklijke Bibliotheek, the National Archives, and the LIFE (Life Cycle Information for E-Literature) Project have all made contributions to our understanding of the financial commitment needed to support the long-term preservation of digital content.

The LIFE[2] Model (2008) delineates life cycle and non–life cycle costs. Life cycle costs are those associated with the processes necessary to preserve digital objects and can be summarized as follows:

- acquisition
- ingest
- bit stream preservation
- content preservation
- access

Each of these main headings can be broken into subareas, all of which carry cost elements and can be used to generate a formulaic approach to estimating the preservation costs of digital objects.

Non–life cycle costs are those that relate to the management and maintenance of the repository: management, inflation, hardware, and software.

Main heading	Subareas
Acquisition	Selection, intellectual property rights, licensing, ordering and invoicing, obtaining, check-in
Ingest	Quality assurance, metadata, deposit, holdings update, reference linking
Bit Stream Preservation	Repository administration, storage provision, refreshment, backup, inspection
Content Preservation	Preservation watch, preservation planning, preservation action, reingest, disposal
Access	Access provision, access control, user support

The cost headings developed for the LIFE and LIFE² models are applicable across the broad array of materials held in digital repositories, ranging from Web archives to collections of digitized images.[35] The models are sensitive enough to cope with complex digital collections, especially personal digital materials, which are likely to contain items that an archivist might need to submit to forensic analysis in order to render a file's inherent information in an accessible form or to determine whether other modes of intervention have taken place on the materials.

The work done on costs to date has largely ignored the specific expenditures associated with digital forensics, although many of the generic headings in the LIFE costing model can subsume digital forensic actions. For example, subareas under acquisition, the first main heading, may include arrangements over intellectual property rights (IPR) and licensing that may be need to be worked out before forensic analysis can take place. The actions associated with check-in, another subarea under acquisition, include verification actions, but under the LIFE² model these processes are fairly simple, often including only a manual verification of the content, and could not be categorized as actions that are specific to the field of digital forensics.

Under the metadata subarea, which is part of the ingest main heading, are file-format validation and integrity-check actions intended to provide automated matching of the content with the specifications of the format the content purports to be. These actions may include verification that the content is valid and well formed. They may also include (sampled) manual checking that the content renders with the access software currently provided by the organization or commonly employed by users, but again go no further to ascertain the provenance of the digital materials in a more rigorous fashion.

Under bit stream preservation, the third main heading, a series of tasks falls under the subheading of repository administration, including the maintenance of security systems to preserve the integrity of the materials stored in the repository and accessed by scholars. Again, however, there may be circumstances in which extremely sophisticated and expensive security systems may be required to ensure the integrity of preserved material. Also under the heading of bit stream preservation, LIFE² posits under the subheading of inspection a series of automated and manual actions designed to provide automated auditing of stored objects to ensure that regenerated checksums match previously stored checksums (thus verifying that the content remains unchanged) and, through manual inspection, to ensure that they can be retrieved and rendered as expected. Although these actions are necessary to ensure the integrity of material in a routine digital repository, there are likely to be circumstances in which they are insufficient to provide an appropriate level of assurance that highly sensitive and important digital content has not been interfered with in any unauthorized fashion.

[35] More information about LIFE² is available at http://www.life.ac.uk/2/.

Under the content preservation heading, the cost of maintaining systems to enable technological changes to be tracked, and to plan future preservation actions, as well as identifying occasions when reingesting digital objects may be required, is likely to be significant, even when institutions can collaborate in areas like technology watch.

Finally, under the access heading, provision of access controls is identified as a significant subheading for costs. Here, again, the model assumes that security systems can be maintained relatively simply by a systems administrator working for a limited number of hours per month. The provision of robust and trustworthy access-control systems for important and sensitive material will undoubtedly require a greater degree of effort than this, and is likely to require a greater financial commitment.

In summary, the existing methodologies that advise institutions of the costs of maintaining digital repository systems, including the most widely accepted methodology, the LIFE^2 model, are probably too generic to provide anything more than broad guidance about the costs of acquiring, capturing, managing, securing, and providing controlled access to sensitive digital information. The report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, published in February 2010 and entitled *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, provides a higher-level view of the economic factors driving the provision of digital infrastructure. The report emphasizes the need to express value propositions clearly so that necessary resources can be directed toward the curation of digital objects. The task force recommends a number of societal changes to ensure that sufficient incentives for digital curation are in place. But the report also notes that although skilled staff, the creation of economies of scale, and the development of public-private partnerships may help organizations that are involved in digital curation, technical developments designed to lower the costs of digital preservation and determine the optimal level of technical curation are also necessary.

The level of technical and organizational commitment required to develop capacity and competence in the area of digital forensics is likely to be high for the foreseeable future according to myriad reports published over the past 15 years. The number of institutions capable of delivering this level of technical competence and organizational capacity over a sustained period will continue to be small, and the financial resources required by these few will likely be greater than what most institutions need to fulfill their existing levels of commitment to maintain the basic infrastructure required for digital preservation and access.

3. Ethics

The notion that the archivist (and, by extension, the archives) is a neutral and independent participant in the scholarly process, whose function can be trusted by all who use a given repository, has been

challenged by numerous historians, philosophers, theorists, and others as scholars and archivists have come to acknowledge the subjectivity inherent in the acts of collection, appraisal, organization, and description.[36] Within the archives profession, some of the most important ethical issues in the digital era relate to an archivist's privileged access to the born-digital materials of an individual or organization. Archivists have always been in this position of power, but over the centuries the development of professional ethics, combined with the growth of the archival profession, has created structures to prevent its abuse. The arrival of born-digital materials in archives highlights the need for archivists and other professionals who work with these items to have a more nuanced understanding of professional ethics.

As born-digital materials become commonplace within libraries and archives, the librarian's and the archivist's commitment to professional ethics is being tested under a new and constantly changing set of technical circumstances. The prevalence of digital technologies has spurred a renewed interest in these professions and a need for staff from fields other than information studies. Staff whose backgrounds are in commerce, industry, technology, and other fields may not be familiar with the professional culture of the librarian and archivist. This lack of exposure to the traditionally trusted routes into the profession, which provide training regimes and cultivate a certain professional ethos, may necessitate additional on-the-job training in professional ethics specific to libraries and archives. An additional benefit of an influx of talent and insight from other disciplines may be that librarians and archivists are challenged to examine their own assumptions about the functions of privacy, information, technology, access, and ethics in the digital age.

Key to the ethical culture of librarians and archivists has been the notion of "curation," a concept that has derived from the museum community and some large library and archival institutions. The digital library and archives communities have recently adopted the term to refer to the management of digital materials through their life cycle.[37] The traditional curator of books and manuscripts has been involved in the complete archival cycle, from selection to acquisition; to appraisal, listing, and cataloging; to providing access through inquiries, Web publication, and exhibitions. Regional variations in traditional curatorial duties, as well as differences at the local level between repositories, mean that not all curators undertake the same type of work. In some repositories, for example, curators select, acquire, and interpret materials (e.g., via exhibition or answering inquiries), but the work of appraisal, cataloging, and description is left to archivists. The digital curator, regardless of job title, is the person who manages the capture and long-term preservation of the digital object and has the best sense of the provenance of a particular document, as well as the skills and knowledge needed to advise the

[36] See Light and Hyry 2002, and Steedman 2002.

[37] See http://www.dcc.ac.uk/digital-curation/what-digital-curation (accessed 7 September 2010).

scholar on issues such as authenticity and documentary integrity. Today's digital curators, and those of the future, will need to adapt their skill sets to the nature of the born-digital materials in their custody, and strive for the same level of understanding and competence as curators had in the traditional era. Among the skills to foster will be even closer working relationships with data creators and depositors, although this kind of activity will be possible only for libraries and archives with sufficient staffing resources to deal with these issues on a personal basis.

The work being done with the Salman Rushdie Papers at Emory University highlights a key ethical issue facing digital archivists: the presence of personal digital materials provides an opportunity, when staff deploy the tools of digital forensics, to access the personal digital histories of the individuals whose collections they are curating. In caring for Salman Rushdie's digital literary manuscripts, those responsible for digital collections could potentially access significant portions of his personal Web history, for example, revealing behavior that was never intended to be disclosed to any other individual. Such access could not only uncover online behaviors but also expose financial and medical information of a highly confidential nature. These ethical concerns operate at the level of both the individual digital curator and the institution. If a staff member has access to sensitive personal information, this may leave the creator at risk of personal hostile action such as violation of privacy, blackmail, or theft. At the institutional level, inadvertent exposure of such information to scholars or the public may engender serious risk of loss of trust; the owner or creator of the digital content may request that her materials be removed from the repository or embark on legal action against the institution. An understanding of these ethical concerns and established mechanisms by which to address them must be embedded in the organization's professional culture, in its employment policies and staff handbooks, in the managerial regimes that monitor the behavior of staff, and in the transparency and public availability of a repository's policies governing the transfer, preservation, management, and delivery of potentially sensitive materials.

3.1. Security Issues

Security is a major concern for the creators and owners of archives. The ease and speed with which digital materials can be accessed, copied, and communicated have created a culture of fear among many depositors and staff members. In the past, sensitive material could be separated from "open" material, often in boxes with physical seals. Sometimes these items were excluded from published catalogs or finding aids available to researchers in the reading room. As the result of serious leaks (of e-mails in particular), research has found that digital information is perceived to be more vulnerable to unwarranted access and more likely to be tampered with than its paper counterpart (see Mayer-Schönberger 2009). Furthermore, the culture of social networking has encouraged a growing public

unease surrounding notions of privacy. Archives and libraries also have to deal with evolving attitudes about freedom of information and expectations of accountability and the access to "public" information (that is, information related to activities of the state or bodies and individuals funded by the state or by taxpayers). The experience of the team at Oxford working on the Paradigm Project with active politicians revealed a fear among this group of data creators that confidential information could be hacked through insecure university networks and over inadequate firewalls and released into the public domain, potentially causing political damage to the individuals concerned and their political parties (see, for example, section 3 of the Paradigm *Workbook*). Such fears have encouraged creators of archives to destroy information rather than allow it to be kept in unsafe regimes, and even to ask that information be stored on servers that are physically unconnected to the Internet and locked in strong rooms along with rare books and medieval manuscripts. Only with such extreme physical measures will some data creators be satisfied with the security arrangements for their archives.

3.1.1. Access Controls and Oversight of Use

Depositors of archives often have control over access at the forefront of their minds when assessing the trustworthiness of a repository. For the archivist and librarian, personal relationships are just as important as technical arrangements when trying to engender a sense of trust in a repository. Staff must be able to demonstrate that the digital materials placed in their care under access controls cannot be accessed without permission. This will take the form of:

- Provision of clear information to users and of training and instructions for staff to ensure that restricted archives are managed appropriately.
- Controls to ensure that access by authorized users is automatically recorded by the repository system and can be retrieved and easily produced as documentary evidence.
- Controls on users so that records, when accessed under the operation of permissions, are not misused. This may involve requiring authorized users to sign disclaimers and not allowing them to copy or remove electronic records from the archival repository (e.g., allowing only handwritten notes to be taken when accessing digital archives in a reading room).
- Controls on collections to ensure that they cannot be accessed surreptitiously and without authorization. For digital archives, this means building a digital infrastructure that supports security mechanisms (e.g., management of firewalls) and other aspects of good practice to avoid unauthorized entry into controlled systems. Within the specialist aspects of digital repositories, this means a range of activities that include the creation, publication, and maintenance of policies making it clear that unauthorized use is prohibited, and that legal constraints (such as those that operate under data-protection laws) are taken seriously and, when appropriate, policed.

In many instances, engendering trust will be the only way to convince data creators that their information is secure. Ultimately, the reputation of individual repositories and their staff may have to carry sufficient weight to convey the assurances that are needed. Over time, archival repositories will be able to demonstrate a track record of keeping digital information safe and secure. Recent high-profile lapses in security (such as the leaking of e-mails from climate change scientists at the University of East Anglia) have been from within essentially unmanaged systems. A key test will be whether potential depositors of digital material with the University of East Anglia's archives will be dissuaded from transferring their materials because of the damage to that institution's reputation inflicted by the climate change affair (G. Marshall 2009).

Value is another aspect to consider. At present, most digital information has no commercial value in the same sense that, for example, a copy of the first folio of Shakespeare or a collection of modern literary manuscripts by a famous writer might command. This is partly because of the perception that e-mail archives and collections of word-processed documents lack the tangible appeal of a paper or parchment collection. Born-digital materials will be understood to possess a significant monetary value once their affordances become more widely known. At that point, the organizations and individuals charged with their care will have to pay even closer attention to security. As digital archival materials gain traction in the marketplace, security violations will no longer be simply a matter of reputation, but will indicate a financial loss for the custodial institution, possibly triggering insurance claims and other financial consequences. Internal management systems are likely to create tighter security in situations in which financial loss is a possible outcome of lax security.

3.2. Privacy

3.2.1. Conduct and Confidentiality

Privacy is at the heart of the ethical issues surrounding born-digital materials. The Salman Rushdie Papers at Emory University, which consist of both paper and born-digital materials, illustrate the importance of privacy and provide an example of how custodians at one repository have worked with a creator to provide a satisfactory level of security while also allowing researchers a reasonable level of access. "Privacy was the major issue for me," Rushdie explained, describing as "exhaustive" the conditions under which he allowed his born-digital materials to be made publicly available (Loftus 2010b). The professional team at Emory faced a considerable challenge in working with Rushdie to secure his privacy, while also meeting the preservation and access needs central to the library's mission. "University librarians, archivists and legal experts have worked with [Rushdie] every step of the way to determine what must be kept confidential—and for how long," Mary J. Loftus wrote in her article about the project (Loftus 2010a). "We have to find a balance between protecting his privacy and providing significant content to

researchers who would find value from it," elaborated Lisa Macklin, coordinator of Emory University Libraries' IPR office (Loftus 2010a). E-mail and born-digital manuscript materials are among the items researchers can access from an Emory reading room, either by using a database or by interacting directly with an emulated version of one of Rushdie's computers.

The security of personal digital information is covered by the term *information privacy*. The ease with which privacy can be infringed is causing some individuals to exercise a new concept, known as *digital abstinence*, defined as the avoidance of putting personal information into digital form and thereby making it subject to copying and distribution actions that place at risk the information privacy of an individual (Mayer-Schönberger 2009, 134-154). Approaches such as digital abstinence create significant problems for scholars, librarians, and archivists interested in reconstructing a person's digital life. But even if a person resolutely avoids putting personal information online, she has no control over the conditions under which other people expose her information: for example, banks may keep their clients' personal information on networked servers, or a business owner may use a networked database to manage orders she receives in the mail. The emerging social and cultural climate with respect to digital information, and especially cloud-based personal digital information, is creating a culture of fear around unauthorized disclosure at many levels. First, criminal activity increasingly targets personal digital information for identity theft, either to gain access to bank accounts and credit card accounts to steal money or, more broadly, to appropriate identities for criminal activities such as falsifying passports and other official documentation.

Members of the public in many Western societies increasingly see governments' use of cyberwarfare techniques as a threat, and there is growing unease about the way publicly accessible personal information in social networking sites such as Facebook and MySpace may be accessed and manipulated by foreign powers (see Mayer-Schönberger 2009). These concerns may be unfounded, and to some extent stimulated by the West for political purposes, but they nonetheless represent a genuine fear on behalf of some sectors of the public in Western democracies. The experience of Google in China is seen as an example of the conflict of interest between the commercially driven Western model, characterized by open communication through the Internet, and that of a Communist society that uses censorship and state regulation of communication systems as a tool for social and political control. Google originally attempted to work alongside Chinese authorities to address the issues of state censorship of information discovery systems, but Western public and political unease about this strategy led the company to adopt a more overtly moral stance (Brannigan 2010).

Another aspect of social networking that must be considered in relation to concerns over privacy relates to using aliases. Since the inception of the Internet, many individuals have preferred to use aliases to protect their identity. This practice could cause problems

for archivists, digital curators, and researchers interested in identifying the various online presences, such as blog postings or a Facebook page, of individuals whose papers are in a repository's collection. For example, if a writer dies without leaving any documentary record of her aliases, it might be difficult for the repository to create a full picture of her online communities. Likewise, an archivist taking responsibility for a particular digital collection of a living individual will need to consider the ethical issues involved in revealing the true identity of a person who uses an alias online. Another possibility poses serious ethical implications: because computers and other media can contain traces of a user's Internet activity, including usernames and passwords, it might be possible for a digital archivist to discover a creator's alias and either reveal her true identity or use the information to gain access to server-based e-mail accounts, social networking sites, and the like. Examples such as these drive home how important it is for a prospective donor to work with repository staff to craft an agreement that lays out explicitly what digital information she intends to transfer to the repository and to what uses it may be put (see the sidebar on "Donor Agreements").

3.2.2. Recruitment, Training, and Encouragement of Staff

The archival mission is one that balances immediate over long-term issues, where the demands of creators and users of archives in the short term must be weighed against the needs of the materials as well as the projected long-term demands of future users and of society as a whole. In the paper era, one manifestation of these issues has been the classic tension between preservation and access. A common strategy for balancing these concerns is evident in the decisions often made by archivists to hold information for long periods of time, even though it will not be publicly accessible for many years. In the United Kingdom, this is classically exemplified in the various stipulations of the 1958 Public Records Act, where classes of public records were designated as being subject to closure regimes, such as the "thirty-year rule," or longer periods where Official Secrets or Public Honours were concerned. Although this information would continue to hold a long-term interest, legal and other custodial pressures have enforced a rigorous ethic within the archival community that "closed," "embargoed," or "restricted" collections should remain so, and that the archivists themselves should exercise the same levels of discretion over access to this information. In the digital age, the means of controlling access to collections that fall into this category becomes more problematic as the physical barriers of the paper era, controlled by lock and key, give way to digital information housed on servers that ideally are both physically and virtually secure. In this latter case, trusted archivists may well have access through digital asset management systems to archival data that are closed to other individuals. Regimes of oversight, including automated access logs that are routinely monitored by senior staff, will be required to maintain appropriate levels of internal vigilance.

3.3. Working with Data Creators

The introduction of forensic techniques to the archives profession means that archivists can capture and preserve more information by and about creators than ever before. Current ethical practices for paper materials vary by repository and nation, but are generally geared toward ensuring that sensitive information protected either by law (e.g., Social Security numbers, financial information, education records, medical records) or by contract (e.g., restricted personal content) is not made available to the public. The same concerns apply for born-digital items, but the complex nature of digital materials presents new challenges for creators who want to prune their files of sensitive content before transferring them to an archival repository. Levels of comfort with technology vary widely, and many creators will not have the skills necessary to locate and delete certain kinds of information from their own born-digital materials prior to transfer.[38] To foster and maintain mutually beneficial working relationships with data creators, archival repositories that accept born-digital materials and use forensic methods to preserve them must make their methodologies transparent and, when possible, must work with creators to ensure that they understand what they are transferring and how it may be used.

Until fairly recently, most born-digital materials arrived at archives as part of larger accessions consisting primarily of paper materials, transferred by a creator at the culmination of his or her career.[39] In this way, many repositories have accumulated a significant volume of born-digital materials. The paper items in these collections have been processed in accordance with deeds of gift, but these contracts were likely written without born-digital materials in mind and may not specify preservation and access restrictions for sensitive digital files. The default action has been to apply the same strictures to digital content as to paper materials; however, this solution does not address the unique forms of digital information that can be gathered using forensic techniques.

Forensic methods of capture create a bit-for-bit image of a floppy disk, hard disk drive, or other digital media. This forensic image replicates exactly all areas of the disk, including files often not preserved by other means. These captured files can comprise unallocated sectors containing information from deleted files, registry files containing username and password information and a record of Internet searches, and other kinds of files. It would be possible for creators to identify and manage this information for themselves, on the original media and prior to transfer to a repository, by using the command line or forensic analysis tools (both open-source and proprietary options are available online; see Appendix A); however, it seems

[38] It is equally likely that some creators will not care what digital files they send to the repository or what is done with them.

[39] See the collection development section of the Paradigm *Workbook* for an informative breakdown and exposition of different acquisition methods. Available at http://www.paradigm.ac.uk/workbook/collection-development/index.html (accessed 19 April 2010).

Donor Agreements

When acquiring materials from individuals, collecting institutions traditionally rely on written donor agreements to clarify and document donor expectations about the management of and access to the materials. The language used in existing donor agreements is rarely specific enough to resolve the ethical issues described in this report.

Currently missing from most acquisition activities are mechanisms for eliciting the specific "curatorial intent" of donors with respect to a given set of materials—that is, what properties of the materials an archivist should be sure to reproduce over time, even if technology changes, and what forms of access to the materials the archivist should allow.

One of the primary challenges of eliciting curatorial intent stems from the fact that digital resources are not simply composed of the information visible to the donor on the screen the last time he or she viewed a set of documents. Instead, digital resources comprise interacting components that can be considered and accessed at different levels of representation:

- a bit stream (series of 1s and 0s) as it resides on a physical storage device;
- a bit stream as it is read from the storage device using a particular combination of input-output (I/O) hardware and software;
- a small data structure that can be accessed using specialized software (below the file-system level)—for example, information about the deletion of a file or information within a photograph that indicates what kind of camera was used to take it;
- a file as read through the file system (i.e., the way a user typically interacts with the system by clicking on folders and files to open them);
- a discrete digital object as experienced through a particular application, for example, a photograph or simple text document;
- a composite document that is composed of multiple files, such as a Web page with embedded images or an e-mail with an attachment; and
- a larger aggregation of content, such as an entire Web site or e-mail account.

Professionals responsible for the care of digital materials will need to expand the traditional notion of a donor agreement to address the various forms of representation that are manifested in digital objects. Fundamental questions to address in such an agreement are as follows:

- Exactly what does the donor intend to transfer to the repository? For example, does she want to transfer the whole bit stream on a floppy disk, or just the files from it? The entire Microsoft Outlook .pst file (including saved and sent messages, calendar items, draft and deleted messages, address book, and possibly viruses), or only a selection of messages from it?
- What types or levels of representation are particularly sensitive to the parties represented in the materials? Many parties other than the donor (e.g., individuals in pictures, correspondents within an e-mail thread, intellectual property rights holders) may be represented in the materials.
- How might the encoding or rendering decisions of a repository promote or violate the interests of stakeholders?

Regarding particular types of information within bit streams, one could decide to remove the information completely at the time of acquisition or to retain it but restrict access. A common mechanism for controlling sensitive information is to specify in the donor agreement that certain portions of the collection will remain closed for a given period of time or until a designated trigger event (e.g., the death of a named individual). The temporary closing of collections and subcollections from public access is likely to remain important; however, the numerous levels of representation of digital materials will require new types of closure arrangements. For example, a repository could provide public access to a page-image representation of a Word document (e.g., printed to PDF/A) for a limited period, after which users would be allowed to access the original Word bit stream, including various forms of embedded information within the file.

—Cal Lee, University of North Carolina at Chapel Hill

Related Resources

Urls are current as of November 22, 2010

Model Gift Agreement. In *Workbook on Digital Private Papers*. Paradigm Project. Available at http://www.paradigm.ac.uk/workbook/appendices/gift-agreement.html.

Submission Policy Input Form. Directory of Open Access Repositories (OpenDOAR). Available at http://www.sherpa.ac.uk/OpenDOAR/opendoarsubmissionpolicy.php.

Tufts Accessioning Program for Electronic Records (TAPER). Working documents related to Submission Agreements. Available at http://dca.tufts.edu/?pid=136&c=166.

Variable Media Questionnaire. Available at http://variablemediaquestionnaire.net/.

unlikely that many will be inclined to do so. Deleted files, Internet queries, and the like will certainly be of interest to researchers, and the ability to recover such information and make it publicly available raises questions about the ethical responsibility repositories have to the creators whose born-digital materials are in their care. What are the implications of this responsibility for collection development and acquisition strategies? What are reasonable expectations for a precustodial relationship between archivist and creator? Similar questions arise regarding access to born-digital materials. If everything can be captured from a hard drive, what should be made available to researchers? What role should repositories play in helping creators make informed decisions about which born-digital materials to restrict, if any? (See sections 3.1 and 3.2 for a more in-depth consideration of factors related to security and privacy.)

The same forensic capabilities that prompt ethics, privacy, and security concerns also provide archivists with the means by which to isolate and sequester materials specified as restricted in the purchase or donation agreement, or even to avoid accessioning certain kinds of materials altogether. Software such as EnCase Forensic and Forensic Toolkit (both proprietary) and The Sleuth Kit (open source) facilitates analysis of the data on disks, hard drives, and other storage devices (see Appendix A for a list of these and other options). With these tools, archivists can preview and analyze the contents of a disk image in order to ascertain what types of files it contains and where the likely files of interest reside. Forensic tools can also be used to search the contents of a hard drive for keywords (e.g., names, e-mail addresses, strings of numbers) and preview files containing graphics (e.g., photos). Repositories and creators could use this information prior to acquisition to develop a more precise donor agreement—for example, specifying that e-mails to and from a particular person be restricted until 20 years after the creator's death, or that deleted and hidden files not be preserved—that would establish clear expectations for both the repository and the creator prior to the official acquisition of a disk image. Another option would be to use forensic tools in collaboration with the creator to identify which directories on a disk or hard drive she wants to transfer to the repository, and then do a targeted capture and acquisition of only those materials.

The pioneering research done by Cathy Marshall and Jeremy Leighton John into the computing and digital archiving habits of individuals highlights an area of inquiry with vital implications for digital archivists and researchers (Marshall 2008a and b; John et al. 2010).[40] Learning more about the imprint of technology on people's lives will provide insight into the context of the digital archives

[40] Susan Thomas and the Paradigm project team have also argued that repositories need to be more directly involved with the creators whose records will soon be in their care. See Appendix C of the *Workbook* for a sample survey that includes questions related to the personal management and organization of digital materials. Available at http://www.paradigm.ac.uk/workbook/appendices/records-survey.html (accessed 20 April 2010).

created now and in the future. Many of these collections will be hybrids consisting of a variety of media, such as disks, papers, a hard drive, and perhaps even a Blackberry or an iPhone, some of which may be related to one another in unexpected ways. For example, in 2008, the Irish writer Sebastian Barry described sending the closing line of a play to the director via text message: "I had pulled the car off the road, on some dark backroad of Wicklow," he said, "and jotted down a last line for the play that had risen up like a trout in the river. I texted it to the director ..." (Rochester 2008). If Barry's cell phone or SIM card arrived as part of the materials he transferred to a repository, an archivist might be able to use forensic tools, such as Paraben's SIM Card Seizure, to recover Barry's text message to the director. But how meaningful would that text message be without a larger context, both in terms of content (why are the words significant) and media (how do this digital message and its metadata relate to the other materials in the collection)? And perhaps more importantly, what is a repository's ethical responsibility when something like a SIM card is discovered unexpectedly among other collection files and the creator's intentions with regard to it are unclear? In many instances, the only way a repository will be able to learn more about the relationships between the various media in a collection and ensure that a creator understands the implications of transferring his or her born-digital materials to a repository is to establish a working relationship with the creator prior to formal acquisition.

Ethical considerations will only increase in complexity as repositories begin acquiring more born-digital materials. The promise of increased capabilities for information recovery raises the stakes for both repositories and creators. To continue to abide by evolving professional ethics and maintain respectful working relationships with creators, repositories will need to figure out what information can be extracted from born-digital materials, how, and to what ends. Repositories will also need to make transparent their methodologies and policies regarding the preservation of born-digital materials and the extent to which they are being made available to researchers. Sustained interaction between repository staff and creators will go a long way toward establishing archival repositories as trustworthy custodians of born-digital materials and ensuring that ethical concerns remain an important part of the transfer process.

4. Conclusions and Recommendations

Forensics is a Janus-faced word, encompassing seemingly countervailing meanings of verbal persuasion and empirical demonstration. Historically and etymologically, this duality is circumscribed by the legal sphere in which various forms of evidence and argumentation are marshaled before the bench. The long association between law and archival science—from Roman law to the emergence of diplomatics and the subsequent appearance of legislation governing public access to government records in Sweden, France, and elsewhere in Enlightenment Europe—suggests it should not be surprising that the

tools and methods of digital forensics are creating new possibilities for archival practitioners. However, the first and most compelling conclusion we have drawn from our research and conversations is that over the long term, digital forensics should not simply be imported and adopted in toto into manuscript archives and the broader cultural heritage and scholarly communities.

Given the popular connotations of "forensics" with *CSI* and the like, the mere appearance of the term may prove problematic in certain settings—particularly if an author or originator is within earshot. (As Clifford Lynch asked of the attendees at the Maryland symposium, "How many of *you* would like to be the subject of a forensic investigation?") The Manuscripts, Archives, and Rare Books Library at Emory University is reportedly adopting the term *data analysis*, rather than *forensics*, with exactly such sensitivities in mind. While these concerns are legitimate, we believe that careful donor education can allay much of the anxiety around forensic analysis, and that in fact many donors may be reassured by the availability of this capability.

The benefits of the tools and techniques we have been describing are numerous. Incorporating forensics methodology and tools into the archival workflow will enable digital archivists and curators to capture more information about the content and makeup of digital objects; help repositories manage the data copied from disks more efficiently and in accordance with established standards; reinforce the importance of documentation to all aspects of the curation cycle; and give archivists, donors, and others the ability to preview the contents of both isolated storage media and complete computing systems to formulate acquisition and preservation strategies.

Nonetheless, scholars and archivists may well find themselves proceeding from different sets of assumptions and priorities than the legal community does, and they will almost certainly find themselves working within different sets of constraints. Time is typically of the essence in criminal or civil proceedings: the statute of limitations means that an analyst is under pressure to deliver results as expeditiously as possible. A legal forensic expert must excel at packaging and presenting a discovery in a manner that will be persuasive to the nonspecialist. Ambiguities will be resolved, rather than embraced.

In the legal and law enforcement community, much of the focus of a forensic investigation is on the proverbial "smoking gun"—some key piece of digital evidence that will seal an indictment, corroborate a witness's testimony, sow doubt in the mind of a juror, bring the opposition to the bargaining table, and so forth. The major commercial forensics software packages reflect what an increasingly competitive marketplace demands, with features supporting the automatic extraction of pornographic images, credit card numbers, phone numbers, names, dates, or other incriminating file types and data. In the archives and cultural heritage communities, however, smoking guns are usually red herrings. An archivist may well want to isolate personally identifiable information so that it can be redacted, but rarely

is a single document or image the object of a scholar's attention. Indeed, it is not yet clear what kinds of demands scholars will place on born-digital material, or what expectations they will have in terms of access, analysis, and publication of their findings. One thing, however, is certain: they are more likely to be interested in wide-ranging discovery and exploration. As the size and rate of acquisition for personal digital papers continue to accelerate, scholars and patrons are likely going to need tools for data mining, social network visualization, and discovery that are subtler and more forgiving than those now supported in the commercial sector.

Perhaps the strongest point we can make is to reinforce the distinction between tools and procedures. Not every institution or repository has the resources to adopt a complete forensic workflow, and in many cases commercial software and a full forensic workstation may exceed the needs of a repository's collection materials. For example, if a repository has only one media type (e.g., Zip disks) in its collection and limited resources, investing in a forensic workstation designed to capture images from a variety of media is not essential. More important than specific technical skills or a large budget is a willingness to figure out what knowledge and tools are necessary to get the job done, and how to go about acquiring them. Interoperable and extensible tools, procedures, and conceptual frameworks are paramount. Technology is expensive, but methodology is free: even if a repository does not have the funding or staff resources to invest in equipment, it can adopt components of a forensic methodology (see the sidebar on a "Digital Forensics Workflow") and rely on freeware or built-in utilities to perform actions such as capture and metadata extraction and basic analysis, or to find a way to share resources with other repositories in the area doing similar work. Ideally, this report, in conjunction with other resources (see Appendix C), will provide enough information about forensic tools and methodologies for archivists and others to make informed choices about the components that best suit their collection materials, funding and staff resources, and level of institutional support.

While we have focused on file-system forensics as the aspect of industry practice most directly relevant to the needs of archivists, we would offer a closing word of caution with respect to privileging the file as the default unit of forensic analysis or of scholarly attention. Files, after all, are just logical conveniences, both at the level of the computer operating system and user interaction. Files always exist within larger contexts, whether they take the form of the local file system, an individual user's complete ecology of machines and devices, or the material world around an individual computer, whose edges may be as physically proximate yet functionally removed from the digital object as a sticky note mounted on the edge of a screen display. At the Maryland symposium, Seamus Ross provocatively referred to digital forensics as a "discipline of failure," meaning that its raison d'être was to compensate for shortcomings in the custodianship of the digital cultural record to date. For Ross, forensics seemed to represent the possibility of overcoming limits to

knowledge imposed by technological obsolescence, media degradation, insufficient institutional resources, and simple human error. Other symposium attendees felt differently; they pointed out that lived human experience is always going to be messy and out of sync with seamless mappings to archival structures and workflows. In their view, forensics is not compensatory but enabling, representing an opportunity for archivists to achieve greater consistency, transparency, and flexibility in their processing of the documents and records in their care. We believe that this apparent contrast is in fact yet another manifestation of the Janus-faced nature of forensics itself, simultaneously retrospective and prescriptive, elegiac and idealistic, empirical and rhetorical. Computers and digital media clearly will provide new challenges for this age-old mix of tensions and oppositions, but the custodians of the born-digital cultural record can and will continue to build on the base of achievement that has come to them from diplomatics, archival practice, textual scholarship, and scientific method.

4.1. Next Steps

The community represented at the Maryland symposium felt strongly that the momentum from the event should not be lost, particularly with regard to contacts that had been opened between repositories and various federal agencies. We offer the following recommendations, many of which emerged out of conversations at the symposium:

1. **Develop policy frameworks and best-practice agreements for donor relations, liability, workflows, and researcher access.** There was a strong collective sense among participants that manuscript archives need to work together at the administrative level to develop policy frameworks for the issues that will define relations between donors, collecting institutions, and patrons in the born-digital era (see the sidebar on "Donor Agreements"). These issues include legal and public policy agendas; guidelines for handling sensitive or illicit materials that may come into an institution's possession as part of a born-digital collection; liability with regard to third-party, personally identifiable information in a collection; donor education; shared deeds of gift; and appropriate access controls for different classes of researchers and patrons.
2. **Develop regional networks for collaboration**. Not every institution will be able to support and maintain infrastructure for digital forensics, and indeed it is not necessary that identical expertise reside at every institution. The development of regional networks for sharing knowledge, hard-won experience, and specialized hardware or facilities seems essential to the success of the individual archivist who is faced with the prospect of processing a born-digital collection absent a robustly equipped local environment for doing so.
3. **Define requirements for and develop new tools**. As the archives and cultural heritage community grows increasingly

sophisticated in its adoption of forensic procedures, friction will likely arise between the features and frameworks of existing tools and the needs of this new constituency. Given the realities of the marketplace, those needs are unlikely to be addressed by commercial vendors. Therefore, open-source development and sponsored research in academic settings will be the way forward. Existing open-source packages such as Sleuth Kit may provide an adequate base from which to build, but requirements will need to be scoped and defined. Researchers will also need new tools, especially those focusing on data mining, visualization, and discovery. The work of the international digital humanities community in particular is likely to be a source from which to build, given that much expertise and practical experience have already been invested in tools for text analysis, text mining, and visualization. Finally, creators will require new tools. There is much ground for debate as to the appropriateness of early intervention in a creator's daily desktop computing habits, but it increasingly seems desirable that individuals should be educated about the personal-archiving tools available to them (Hoppla and Anthologize are both promising in this regard).[41]

4. **Aid in articulating a scholarly research agenda**. Born-digital archives will expand exponentially the query scope of research in traditional fields. But what kinds of questions will scholars want to ask of this new class of materials, and how can archives support those queries and investigations while maintaining their obligations to the donor? Informational sessions at professional meetings such as those of the Modern Language Association, American Historical Association, and Association of Writing Professionals represent one outreach path, as do (for a broader audience) social media outlets such as YouTube (as has been demonstrated by both Digital Preservation Europe and the Library of Congress).[42]
5. **Collect more stories and case studies**. More case studies are needed to document and examine how forensic methods and tools are being integrated into workflows within specific institutional settings. Outcomes will need to be assessed and cost-benefits established. The AIMS project (Virginia, Stanford, Yale, Hull), futureArch (Bodleian), and the Digital Records Forensics Project (University of British Columbia [UBC]) are all likely sources for such case studies, as is the ongoing work under the auspices of the Digital Lives project at the British Library.
6. **Facilitate training**. Corporate training programs such as those provided by major industry vendors are likely beyond the reach of archivists in all but a handful of settings. Therefore, the cultural heritage community needs to develop its own venues for exposure to and training in forensic tools and procedures. The

[41] See http://www.ifs.tuwien.ac.at/dp/hoppla/ and http://anthologize.org/, respectively.

[42] See, for example, http://www.youtube.com/watch?v=pbBa6Oam7-w and http://www.youtube.com/watch?v=qEmmeFFafUs&feature=channel.

School of Information and Library Science at the University of North Carolina, Chapel Hill, is undertaking such a pilot program (funded by The Mellon Foundation), and its outcomes should be followed. The Rare Book School at the University of Virginia has recently begun to offer a course entitled "Born-Digital Materials: Theory and Practice." Taught by Matthew Kirschenbaum and Naomi Nelson, it includes forensics in the curriculum. Other opportunities may become available in the form of workshops at conferences and special institutes. The community should work together to publicize these opportunities.

7. **Encourage cross-publication of research literature and cross-promotion of professional events**. As Duranti (2009) notes, "Digital forensics experts are still mostly practitioners and the discipline has only recently entered academia; thus it lacks the kind of *fora* offered to theory development by a multiplicity of scholarly journals and a well-established community of academics writing for them" (43). Duranti recommends the *International Journal of Digital Evidence*, as well as the more recently established *Journal of Digital Forensics, Security, and Law,* for this purpose. Outreach to the editorial boards of these journals might yield special issues on topics of interest to the cultural heritage community. Likewise, the major professional meetings and publication venues in the archives community might initiate on a larger scale outreach to legal and government practitioners of the kind undertaken for the Maryland symposium. The *Forensics Wiki*, an independently curated information-sharing site, is one potential resource for coordinating such efforts.[43] Efforts for cross-sharing research and expertise could also be initiated in conjunction with the regional networks for collaboration suggested above, as well as more informally, along the lines of meet-ups, hack fests, and other social-productivity events commonplace in the open-source world.
8. **Pursue terminology mapping**. Given the extensive overlap in terminology between the forensics, e-discovery, and archival communities, some symposium participants felt that a shared mapping of terminology would be an essential first step toward fostering further contacts between the fields. The Digital Records Forensics Project at the University of British Columbia offers an example of what such a mapping might look like.

[43] Available at http://www.forensicswiki.org/wiki/Main_Page.

Reference List

***Note**: URLs are current as of November 22, 2010.*

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. Available at http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Brannigan, Tania. 2010. "Google Raises Stakes in China Censorship Row." *The Guardian* (22 March).

Burgon, John William. 1888–89. *Lives of Twelve Good Men*. vol. 1. London: John Murray.

Caloyannides, Michael. 2001. *Computer Forensics and Privacy*. Norwood, MA: Artech House.

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

Carrier, Brian, and Eugene H. Spafford. 2003. Getting Physical with the Digital Investigation Process. *International Journal of Digital Evidence* 2(2).

Carvey, Harlan. 2009. *Windows Forensic Analysis DVT Toolkit*. 2nd ed. Burlington, MA: Elsevier.

Casey, Eoghan. 2004. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. 2nd ed. Amsterdam: Elsevier Academic Press.

Cohen, Patricia. 2010. "Fending Off Digital Decay, Bit by Bit." *New York Times* (15 March). Available at http://www.nytimes.com/2010/03/16/books/16archive.html.

Consultative Committee for Space Data Systems (CCSDS). 2002. "Recommendation for Space Data System Standards," in *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1, Blue Book Issue 1. Washington, DC: CCSDS Secretariat. Available at http://public.ccsds.org/publications/archive/650x0b1.pdf.

Cunningham, Adrian. 1994. The Archival Management of Personal Records in Electronic Form: Some Suggestions. *Archives and Manuscripts* 22, 94-105.

Department of Defense. 2006. *National Industry Security Program Operating Manual*. Washington, DC: Defense Technical Information Center.

Department of Defense, Department of Energy, Nuclear Regulatory Commission, and Central Intelligence Agency. 1995–97. *DoD 5220.22-M National Industrial Security Program Operating Manual*. Washington DC: US Government Printing Office.

Dhillon, Amrit. 2000. I Am Pessimistic about the Changes Occurring in India. In *Conversations with Salman Rushdie*, ed. Michael Reder. Jackson, MS: University Press of Mississippi.

Diamond, Elizabeth. 1994. The Archivist as Forensic Scientist: Seeing Ourselves in a Different Way. *Archivaria* 38, 139-154.

Dong, Lorrie, Megan Durden, and Sarah Kim. 2007. Archiving the Arnold Wesker Collection in DSpace: Creating a Batch Ingest Workflow for Digital Files at the Harry Ransom Center; or, Build, My Darlings, Build … A Digital Archive. Available at https://pacer.ischool.utexas.edu/bitstream/2081/8884/1/INF392K-Weskerproject-final_report-2007.pdf.

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

Duranti, Luciana. 1998. *Diplomatics: New Uses for an Old Science*. Lanham, MD: Scarecrow Press.

Duranti, Luciana. 2009. From Digital Diplomatics to Digital Records Forensics. *Archivaria* 68, 39-66.

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

Fathi, Nazila. 2009. "In a Death Seen Around the World, a Symbol of Iranian Protests." *New York Times* (22 June). Available at http://www.nytimes.com/2009/06/23/world/middleeast/23neda.html.

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

Garrett, John, and Donald Waters. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub63watersgarrett.pdf.

Glisson, W. B. 2009. Use of Computer Forensics in the Digital Curation of Removable Media. In H. R. Tibbo, ed., *Digital Curation: Practice, Promise and Prospects. Proceedings of DigCCurr 2009, April 1–3, 2009, Chapel Hill, NC*. School of Information and Library Science, University of North Carolina at Chapel Hill.

Greene, Mark A., and Dennis Meissner. 2005. More Product, Less Process: Revamping Traditional Archival Processing. *American Archivist* 68(2): 208-263.

Gutmann, Peter. 1996. Secure Deletion of Data from Magnetic and

Solid-State Memory. *Proceedings of the Sixth USENIX Security Symposium*, July 22-25, 1996, San Jose, CA. Available at http://www.cs.auckland.ac.nz/~pgut001/pubs/secure_del.html.

Higgs, Edward, ed. 1998. *History and Electronic Artefacts*. Oxford: Clarendon Press.

John, Jeremy Leighton. 2008. Adapting Existing Technologies for Digital Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools. iPres2008. Available at http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf.

John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. 2010. *Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis* (Beta Version 0.2). Digital Lives Research Paper (3 March). Available at http://britishlibrary.typepad.co.uk/digital_lives.

Jones, Sarah, Seamus Ross, and Raivo Ruusalepp. 2008. The Data Audit Framework: A Toolkit to Identify Research Assets and Improve Data Management in Research-led Institutions. iPres 2008. Available at http://www.data-audit.eu/docs/DAF_iPRES_paper.pdf.

Jones, Sarah, Seamus Ross, and Raivo Ruusalepp. 2009. *Data Audit Framework Methodology*. Draft for discussion, version 1.8. Available at http://www.data-audit.eu/DAF_Methodology.pdf.

Joyce, Michael. 1987. *afternoon, a story*. Eastgate Systems, 1990. Available at http://www.eastgate.com/catalog/Afternoon.html.

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

Kirschenbaum, Matthew G., Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. 2009. *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. White Paper. Washington, DC: National Endowment for the Humanities. Available at http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37.

Kruse II, Warren G., and Jay G. Heiser, 2002. *Computer Forensics: Incident Response Essentials*. Upper Saddle River, NJ: Addison-Wesley.

LIFE². 2008. LIFE: Life Cycle Information for E-Literature. LIFE² Web Site. Available at http://www.life.ac.uk/2/.

Light, Michelle, and Tom Hyry. 2002. Colophons and Annotations: New Directions for the Finding Aid. *American Archivist* 65(2): 216-230.

Loftus, Mary J. 2010a. The Author's Desktop. *Emory Magazine*. Available at http://www.emory.edu/EMORY_MAGAZINE/2010/winter/authors.html.

Loftus, Mary J. 2010b. The Revisionist: An E-Q&A with Salman Rushdie. *Emory Magazine*. Available at http://www.emory.edu/EMORY_MAGAZINE/2010/winter/rushdie.html.

Lynch, Clifford. 2000. Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust. In *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources. Available at http://www.clir.org/pubs/reports/pub92/pub92.pdf.

MacNeil, Heather. 2000. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. London: Kluwer Academic Publishers.

Marshall, Catherine C. 2008a. Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *DLib Magazine*. Available at doi:10.1045/march2008-marshall-pt1.

Marshall, Catherine C. 2008b. Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, and Institutions. *DLib Magazine*. Available at doi:10.1045/march2008-marshall-pt2.

Marshall, Catherine C. 2008c. From Writing and Analysis to the Repository: Taking the Scholars' Perspective on Scholarly Archiving. Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries, June 16-20, 2008, Pittsburgh, PA, 251-260.

Marshall, George. 2009. "Leaked Email Climate Smear Was a PR Disaster for UEA." *The Guardian* (23 November). Available at http://www.guardian.co.uk/environment/cif-green/2009/nov/23/leaked-email-climate-change.

Mayer-Schönberger, Viktor. 2009. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press.

McHugh, Andrew, Seamus Ross, Perla Innocenti, Raivo Ruusalepp, and Hans Hofman. 2008. Bringing Self-Assessment Home: Repository Profiling and Key Lines of Enquiry Within DRAMBORA.*International Journal of Data Curation* 2(3): 130-142.

McKenzie, D. F. 1969. Printers of the Mind: Some Notes on Bibliographical Theories and Printing-House Practices. In *Studies in Bibliography: Papers of the Bibliographical Society of the University of Virginia* 22, 1-75.

McKenzie, D. F. 1999. *Bibliography and the Sociology of Texts*. The Panizzi Lectures, 1985. Cambridge, U.K.: Cambridge University Press.

National Computer Security Center. 1991. A Guide to Understanding Data Remanence in Automated Information Systems. Available at http://www.fas.org/irp/nsa/rainbow/tg025-2.htm.

Nelson, Bill, Amelia Phillips, Frank Enfinger, and Christopher Steuart. 2008. *Guide to Computer Forensics and Investigations*. 3rd ed. Boston, MA: Thomson Course Technology.

Nickell, Joe, and John F. Fischer. 1999. *Crime Science: Methods of Forensic Detection*. Lexington, KY: University of Kentucky Press.

Paradigm Project. 2008. *Workbook on Digital Private Papers*. Available at http://www.paradigm.ac.uk/workbook/index.html.

PREMIS Data Dictionary for Preservation Metadata. 2008. Version 2.0. Available at http://www.loc.gov/standards/premis/v2/premis-2-0.pdf.

Rochester, Sophie. 2008. Recovering Ireland's Hidden History. The Man Booker Prizes Web Site. Available at http://www.themanbookerprize.com/perspective/articles/1137.

Ross, Seamus, and Ann Gow. 1999. *Digital Archaeology: Rescuing Neglected and Damaged Data Resources*. *A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials*. Available at http://eprints.erpanet.org/47/.

SearchNetworking.com. OSI. Available at http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci212725,00.html.

Steedman, Carolyn. 2002. *Dust: The Archive and Cultural History*. New Brunswick, NJ: Rutgers University Press.

Stoddard, Roger L. 1985. *Marks in Books, Illustrated and Explained*. Cambridge, MA: Harvard University Press.

Stoll, Clifford. 1990. *The Cuckoo's Egg*. New York: Pocket Books.

UNESCO (United Nations Educational, Scientific and Cultural Organization). 2008. What Is Cultural Diversity? Available at http://portal.unesco.org/culture/en/ev.php-URL_ID=13031&URL_DO=DO_TOPIC&URL_SECTION=201.html.

Van der Hoeven, Jeffrey, Bram Lohman, and Verdegem Remco. 2007. Emulation for Digital Preservation in Practice: The Results. *International Journal of Digital Curation* 2(2): 123-132.

Wright, Craig, Dave Kleiman, and Shyaam Sundhar R. S. 2008. Overwriting Hard Drive Data: The Great Wiping Controversy. Information Systems Security. *Lecture Notes in Computer Science*, 5352, 243-257.

APPENDIX A

Forensic Software

Introduction

The bulk of forensics work done in archives and other cultural heritage institutions uses software to read, rebuild, analyze, and secure data from acquired storage devices and media—in forensics terms, using software to conduct "dead analysis." These tools may be bundled in a full-blown forensics package or distributed as individual, specialized programs or scripts. They may have graphical user interfaces or may require familiarity with command line interactions. The functions most relevant to archival work are imaging, data recovery, and logging.

Tables in this appendix are coded according to the following convention:

Y	**Included**
P	**Partially supported**
Blank	**Support not advertised**

Methodology

The tools in the following tables were primarily discovered through syllabi for digital forensics courses, conference literature, the Forensics Wiki, and Wikipedia entries for forensic methodologies. Additional data were gathered from the National Institute of Standards and Technology (NIST) Computer Forensic Tool Testing Program (CFTT), vendor Web sites and other marketing materials, correspondence with vendors, user reviews, and forum posts.

Disclaimer

All information is accurate to the best of our knowledge at the time of publication. However, digital forensics is a large and complex market, and vendors, products, and capabilities change often. These tables are merely intended to serve as guides. Under no circumstances should inclusion of a product or vendor be taken as endorsement, nor should exclusion be taken as intentional. Individuals or institutions must assume full responsibility for their own independent verification of any information provided herein before using it as the basis of a purchasing or policy decision. Under no circumstances can the authors, consultants, CLIR, or the Mellon Foundation be held responsible for purchasing or other decisions made on the basis of the information that follows.

The authors regret any errors or omissions.

Glossary

Imaging

Imaging is the technique of making a soft copy of an entire storage medium (or partition, in the case of hard drives) rather than copying individual files. The resulting image can be manipulated in the same ways as the original, without the hardware. Images come in two general varieties: sparse and bit-exact (sometimes called a clone). A sparse image copies only the sectors that contain data, ignoring any zero-byte sectors and thereby resulting in a smaller file. Bit-exact images contain the entire disk or partition, resulting in a file the same size as the original medium's full capacity. Disk clones can typically be restored only to a partition of the same size as the original, while sparse images can be restored to any partition large enough to contain the data.

Data Recovery

In addition to imaging, data recovery covers all processes to retrieve damaged, deleted, or otherwise hidden data. The broad categories of data recovery are rebuilding, in which damaged file systems are rebuilt; data carving, which can be used to recover data even in the absence of a healthy file system; and steganalysis, or processes to find and retrieve hidden data.

Logging

As with any preservation activity, processes used to recover data must be recorded. Because forensics software has been designed for legal investigation, many of these programs have robust, automatic logging systems.

Other Uses

Many functions outside these core features may also be valuable to an archivist. Encryption software can ensure adequate protection for sensitive materials, and decryption software may aid in accessing the data of donors unable to supply their passwords. Annotation and bookmarking abilities can aid in highlighting materials of interest or flagging trouble spots. Filtering, search, and (meta)data extraction have their obvious benefits, and if the item in question is an entire computer, it is possible that, in the case of a full-workstation donation, RAM and registry analysis could be valuable. Many of the packages also offer a range of visualization and time-lining options that can aid in understanding the range of materials in a donation.

Table A-1: Forensic Software Packages

Program	Annotation	Automation—Built-in	Automation—Scriptable	Bookmarking	Built-in File Viewers	Command Line	Custom Filtering	Data Carving	Data Visualization	Dead Analysis	Cryptanalysis	Encryption	Export Data/Files	Export Metadata	Identification/Verification	GUI[1]	Hex Editor	Hex Viewer	Imaging/Cloning	Live Analysis	Major OSes/File Systems[2]	RAM Capture	RAM Editing	Registry Analysis	Reporting/Logging	Search	Steganalysis	Timelines	Virtualization	Volume/File Rebuilding	Write-blocking
Autopsy Forensic Browser /Sleuth Kit	Y		Y	P		Y		Y		Y	P		Y	Y	P	Y		Y		Y	P				Y	Y	Y	Y		P	
ByteBack								Y		Y			Y		Y	P	Y		Y		P				Y		P			Y	Y
Coroner's Toolkit[3]						Y		Y		Y	Y		Y		Y				Y	Y	Y						Y			Y	
Encase Forensic	Y	Y	Y	Y	Y		P	P		Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y
Forensic Toolkit	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y		Y	Y	Y	Y		Y		Y
Helix 3 Pro					Y					Y	P		Y	Y		Y				Y	Y	Y			Y	Y	Y			Y	Y
iLook	Y		Y		Y		Y	Y	Y	Y	Y		Y			Y		Y	Y		Y			Y	Y	Y		Y			
MacForensicsLab	Y	P		Y	Y		Y	Y		Y				Y		Y		Y	Y	[4]	Y				Y	Y	Y			Y	P
Macintosh Forensic Suite					P	P		Y		Y						Y										Y	Y				
MacMarshall			Y							Y	Y		Y	Y		Y									Y	Y	Y				Y
NTI Computer Incident Response Suite										Y	P							Y	Y							Y		Y		Y	
Nuix Desktop	Y						Y		Y					Y		Y										Y		P			
Online DFS													Y	Y		Y			Y	Y	Y	Y		Y		Y					
P3 Command Kit	Y			Y	Y	P	Y	Y		Y	Y	Y	Y	Y	Y	Y		Y	Y	[5]	P	Y		Y	Y	Y	Y			P	Y
ProDiscover Forensics		Y	Y		Y					Y				Y		Y			Y	[6]	P			Y	Y	Y	Y				Y
Responder Pro			Y						Y	Y						Y				P		Y			Y						
SMART								Y	P	Y			Y	Y	P	Y		Y	Y	P	P			Y	Y	Y					Y
X-Ways Forensics	Y			Y	Y		Y	Y	P	Y	P		Y	Y	Y	Y			Y		P	Y		Y	Y	Y				Y	Y

[1] Menu-driven text interfaces are considered to be partial GUIs.
[2] OS support is considered to be fully included only when all three major operating systems (MacOS, Linux, and Windows) and their default file systems are supported.
[3] While still available for download, Coroner's Toolkit has been superseded by Sleuth Kit.
[4] Live analysis is available with the addition of Mac ForensicsLab Field Agent.
[5] Available from Paraben with the Enterprise Shuttle tool.
[6] ProDiscover Incident Response includes Live Analysis.

Table A-2: Imaging Features by Package

Program	File Type	Open/ Closed[7]	File-System Support[8]						Compression	Clone	Sparse	HDD	Media
			NTFS	FAT	HFS	EXT	UDF	HPA/ DCO					
Autopsy Forensic Browser /Sleuth Kit[9]	AFF, E01, EWF	Open	Y	Y	HFS+	Y		Y	Y			Y	
ByteBack	Unknown		Y	Y					Y	Y		Y	
Encase Forensic	E01, EWF, Raw	Closed	Y	Y	Y	Y	Y		Y			Y	Y
Forensic Toolkit	EWF, E01, S01, SafeBack, Raw	Closed	Y	Y	Y	Y	Y		Y			Y	Y
Helix 3 Pro[9]	Raw, E01, AFF, ISO	Both	Y	Y	Y	Y		Y	Y	Y		Y	Y
iLook	ASB, IRBF, IEIF, IDIF	Closed	Y	Y	Y	Y		Y	Y	Y		Y	Y
MacForensicsLab	ISO, DMG, Sparseimage, Raw	Both	Y	Y	Y	Y			Y		Y	Y	Y
Macintosh Forensic Suite	DMG, Raw	Closed	Y	Y	Y							Y	
NTI Computer Incident Response Suite	Safeback, dd	Closed	Y	Y		Y					Y	Y	Y
Online DFS[10]	dd	Open								Y		Y	Y
P3 Command Kit[11]	PFR/PFIF, ISO	Closed	Y	Y		Y			Y			Y	Y
ProDiscover Forensics[9]	CMP, EVE, Raw, dd	Open	Y	Y		Y		Y	Y			Y	
Raptor	E01, DMG, Raw	Closed	Y	FAT32	HFS+	EXT3		Y	Y	Y		Y	
SMART	SMART default (.image), SMART Compressed (.S01) E01, EWF, RAW	Closed		Y	Y	Y		Y	Y			Y	Y
X-Ways Forensics	E01, EWF, Raw, WinHex	Closed	Y	Y	Y	Y	Y	Y	Y	Y		Y	

[7] Refers to whether the file-format specification is open or proprietary.
[8] File systems include all variants of each unless otherwise specified.
[9] Also supports UFS.
[10] File-system support for DFS is unclear, but the product information page lists support for Windows, Unix, and MacOS.
[11] It is unclear what types of images Paraben's Forensic Replicator can create, but it can work with NTFS, EXT, and FAT.

Table A-3: Package Details

Publisher/Product	Cost	Technical Requirements/Notes
Autopsy Forensic Browser/ Sleuth Kit	Free	UNIX based, but can be accessed through an HTML interface. No information available on system requirements. http://www.sleuthkit.org/contact.php Autopsy provides a front end for Sleuth Kit, which itself is heavily based on the Coroner's Toolkit.
Tools That Work ByteBack	$492.95 Free trial	MS-DOS (includes FreeDOS) Drive + media to create boot disk No information on system requirements available. http://www.toolsthatwork.com/computer-forensic.htm
Guidance Software EnCase Forensic	Standard: $3,600 Government: $2,850	Windows 2000 or newer 3.0 GHz CPU 2 GB RAM 1 USB port http://www.guidancesoftware.com/computer-forensics-digital-investigation-law-enforcement.htm
Access Data Forensic Toolkit (FTK) 2.x	Stand-alone + 1-year subscription $3,835	64-bit edition of Windows Intel i7 2.66 GHz or better CPU 8 GB DDR3 RAM (12 GB recommended) 60 GB 10,000 RPM hard disk (for OS) 20 GB 10,000 RPM hard disk (for Oracle) Supports RAID configurations http://www.accessdata.com/
e-fense Helix3 Pro	$239/year CD only (no support): $150 Free trial available	CD or DVD drive The Forensics Wiki lists some known issues with this software. http://www.e-fense.com/products.php
Perlustro iLook	Free	Windows NT or newer No further information available on system requirements. Free to qualifying institutions with law enforcement missions. Some criminal justice education programs may qualify. http://www.perlustro.com/ustreasury_website/index.html
MacForensicsLab	$1,445 Free trial available	***All versions require a HASP license dongle for license verification (supplied with purchase).*** Mac OS X 10.4 or newer; Windows XP or newer; x86-based Linux distribution with GTK+ 2.0 (or higher), glibc-2.3 (or higher), and CUPS (Common UNIX Printing System) 800 MHZ or faster (2 GHZ or better recommended) CPU 512 MB of RAM (2 GB or more recommended) DVD-ROM drive for Boot CD/DVD and Installation from DVD 1 TB or more hard disk space recommended http://www.macforensicslab.com Extensive documentation available on the Web site
BlackBag Macintosh Forensic Suite	Standard: $799 Government: $699	OS X 10.4 or later 1 GB or more RAM 500 MB available hard disk space http://www.blackbagtech.com/store/software/blackbag_macintosh_forensic_suite.html The Forensic Suite may support more features than listed; little documentation is available.

Table A-3: Package Details (cont.)

Publisher/Product	Cost	Technical Requirements/Notes
Architecture Technology Corp. **Mac Marshal**	$995	Mac OS X 10.4 or 10.5 50 MB available hard disk space http://macmarshal.atc-nycorp.com/
New Technologies, Inc. **Computer Incident Response Suite**	$1,118	No information available on system requirements. http://www.forensics-intl.com/suite1.html
Nuix Desktop	$7,500–$15,000 per user per year; price dependent on features selected.	Focuses primarily on e-mail. No information available on system requirements. http://www.nuix.com/
Cyber Security Technologies Corporation **OnLine Digital Forensic Suite (DFS)**	Standard: $9,000 Law Enforcement: $3,000	No information available on system requirements. http://www.onlinedfs.com/products_dfs.asp
Paraben Corp. **P3 Command Kit**	$3,995 Demo available Prior edition (P2) available for $1,995	Windows 2000 or newer 1.4 GHz or better processor 1 GB RAM 200 MB available hard disk space http://www.paraben-forensics.com/
ProDiscover Forensics	$2,195 Demo available	Windows 2000 or newer 1.2 GHz or higher Pentium-compatible CPU 256 MB RAM (512 MB recommended) 500 MB available hard disk space CD-ROM or DVD-ROM drive VGA or higher-resolution monitor http://www.techpathways.com/prodiscoverdft.htm
HBGary **Responder Pro**	$9,000 "Field" edition available for $979	Windows 2000 or newer No further information available on system requirements. Primarily used for malware detection and analysis; has many options for visualization. http://www.hbgary.com/
ASR **SMART**	$2,000	Linux No information available on system requirements. Available as a Live CD http://www.asrdata.com/
X-Ways Forensics	$1,118	Recommended (minimums not listed) Windows 7 64-bit Quadore CPU 4 GB or more RAM Hex viewer/editor and RAM editor available through WinHex, which can be downloaded as an add-on. http://www.x-ways.net/

Table A-4: Live Distributions

Live distributions, often referred to as "live CDs" or "live distros," are operating systems designed to run from external media (CD, DVD, or flash drive) without installation. Working in this way is useful if the target computer is malfunctioning or compromised, or simply lacks the necessary tools.

Title	Environment	Specialization	Price	Web Site
BackTrack	Slackware		Free	http://www.remote-exploit.org/backtrack.html
ByteBack DRIS	Freedos	HDD and MBR repair	Free	http://www.toolsthatwork.com/byteback.html
CAINE Live CD	Ubuntu	User-friendly GUI and automated reporting	Free	http://www.caine-live.net/
DEFT Linux	XUbuntu		Free	http://www.deftlinux.net/
Digital Forensics Live CD (DFLCD)			Free	http://www.forensiclivecd.com/ **Note:** No longer actively supported.
FCCU GNU/Linux Forensic Boot CD	Debian		Free	http://www.lnx4n6.be/
Grml	Debian			http://grml.org/
Helix3 and Helix3 Pro	Ubuntu		$0–$239	http://www.e-fense.com/helix/ **Note:** Free version no longer receiving updates.
INSERT Rescue Security Toolkit	Knoppix			http://www.inside-security.de/insert_en.html
MacQuisition		Imaging Mac devices	$400–$599	http://www.blackbagtech.com/
Masterkey	Slackware	Incident response		http://masterkeylinux.com/
Openwall (Owl)	Linux	Password recovery	Free to download, $28.86 to ship.	http://www.openwall.com/john/
Operator	Debian/Knoppix	Network security		http://www.ussysadmin.com/operator/
Professional Hacker's Linux Assault Kit (PHLAK)	Morphix	Network security	Free	http://sourceforge.net/projects/phlakproject/
Raptor	Ubuntu	Acquisition	$49.95	http://www.raptorforensics.com/
SPADA	Knoppix			
System Acquisition Forensic Environment (SAFE) Boot Disk	Windows	Read-only imaging	$399	http://www.forensicsoft.com/catalog/product_info.php?products_id=31http://www.forensicsoft.com/catalog/product_info.php?products_id=31
THE FARMER'S BOOT CD (FBCD)		Preacquisition examination	$225	http://www.forensicbootcd.com/

Stand-alone Tools

Table A-5: Imaging Tools

Title	File Type	OS	NTFS	FAT	HFS	EXT	UDF	HPA/DCO	Compression	Clone	Sparse	HDD	Media	Price	Web Site
AFFlib	AFF, AFM (raw), E0/EWF, splitraw, VMDK	Unix	Y	Y		Y			Y	Y		Y		Free	http://www.afflib.org
Arconis Backup & Recovery (formerly True Image)		Windows/ Linux	Y	Y	N	N	N		Y	Y		Y		$49.99+	http://www.acronis.com
Clonezilla		Unix	Y	Y	Y	Y		?/Y	Y	Y	Y	Y	N	Free	http://www.clonezilla.org/
Daemon Tools Pro	B5T, B6T, BWT, CCD, CDI, CUE, ISO, MDS, NRG, PDI, and ISZ	Windows	N/A	N/A	N/A	N/A	N/A						CD \ DVD	$140	http://www.daemon-tools.cc/
dcfldd	RAW	Unix/Mac	Y	Y	Y	Y	Y	?/Y	N	Y	N	Y	Y	Free	http://dcfldd.sourceforge.net
dd	RAW	Unix/Mac	Y	Y	Y	Y	Y	?/Y	N	Y	N	Y	Y	Free	Built-in command
dd_rescue	RAW	Unix	Y	Y	Y	Y	Y	?/Y	N	Y	N	Y	Y	Free	http://www.gnu.org/software/ddrescue/ddrescue.htm
Ddrescueddrescue	RAW	Unix	Y	Y	Y	Y	Y	?/Y	N	Y	N	Y	Y	Free	http://www.garloff.de/kurt/linux/ddrescue
Disk Utility	DMG, CDR	Mac		Y	Y				Y	Y	Y	Y	Y		Included in MacOS
DriveImageXML	Combination of DAT and XML	Windows	Y						Y			Y		Free	http://www.runtime.org/driveimage-xml.htm
Guymager	Rawraw, EWF, E01, AFF	Unix												Free	http://guymager.sourceforge.net/
Macrium Reflect			Y	Y		Ext2			Y			Y	Y		http://www.macrium.com/
Ntfsclone		Unix	Y							Y	Y	Y		Free	http://linux-ntfs.org/
PartImage		Unix						?/Y					N		

Table A-6: Data Carving

Program	Block Based	Characteristic Based	Header/Footer	Header/Maximum	Header/Embedded	File Structure	Semantic	Fragmented Recovery	Price	Web Site
File Extractor Pro			Y	Y	Y				$155	http://www.datalifter.comhttp://www.datalifter.com
Foremost			Y	Y	Y	Y			Free	http://foremost.sourceforge.net/
Magic Rescue	Y		Y					Y	Free	http://www.student.dtu.dk/~s042078/magicrescue/
Scalpel			Y						Free	http://www.digitalforensicssolutions.com/Scalpel/
Simple Carver Suite							Y		$95	www.simplecarver.com

Table A-7: Cryptanalysis

Program	Brute Force	Dictionary Attack	Linear	Differential	Integral	Impossible Differential	Boomerang	Mod n	Slide	Rainbow Tablestables	Cache Searchsearch	Reverse Hashinghashing	Collision	Timing	XSL	Price	Notes	Web Site
Advanced EFS Data Recovery		Y								Y	Y					$149–$299	For Microsoft Encrypting File System; demo available.	http://www.elcomsoft.com/aefsdr.html
Cain & Abel	Y	Y								Y	Y	Y				Free	Windows-centric	http://www.oxid.it/cain.html
Cryptool	Y							Y					Y			Free		http://www.cryptool.com
Decryption Collection	Y	Y														$365–$495	Demo available; limited to 3-character passwords.	http://www.paraben-forensics.com/catalog/product_info.php?cPath=25&products_id=402
Distributed Network Attack (DNA)										Y						$1,800	Like PRT, but designed to utilize networked machines for greater processing power.	http://www.accessdata.com/decryptionTool.html
John the Ripper	Y	Y										Y				$0–$185		http://www.openwall.com/john/
Lastbit	Y	Y										Y				$199	Application-specific modules available; Windows-centric.	http://www.lastbit.com
Password Recovery Toolkit										Y						$790	Focuses on recovering application passwords; generates rainbow tables based on hard drive contents.	http://www.accessdata.com/decryptionTool.html
Portable Office Rainbow Tables	Y									Y						Unknown	For MS Office	http://www.accessdata.com/decryptionTool.html
Rainbow Tables[12]	Y									Y						Unknown	MS Office, PDF, LAN passwords	http://www.accessdata.com/decryptionTool.html

[12] The ForensicWiki has a list of free rainbow tables at http://www.forensicswiki.org/wiki/Rainbow_Tables.

Table A-8: Deleted File Recovery

Program	MacOS	Linux	Windows	HDD	Media	Price	Web Site
Active UnDelete			Y	Y		$50–$100	http://www.active-undelete.com/
Active UnEraser			Y			$50	http://www.uneraser.com/
Data Rescue[13]	Y		Y	Y	Y	$100–$250	http://www.prosofteng.com/
Easy Undelete	P[14]	P	Y	Y		$23	http://www.easy-undelete.com/
eData Unerase			Y	Y	Y	$32	http://www.octanesoft.com/
File Scavenger			Y			$185	http://www.quetek.com/
Stellar Data Recovery[13]	Y	Y	Y	Y	Y	$50–$100	http://www.stellarinfo.com/
TestDisk	Y	Y	Y	Y		Free	http://www.cgsecurity.org/wiki/TestDisk
WinUndelete			Y			$50–$65	http://www.winundelete.com/
Zero Assumption Recovery		Y	Y	Y	Y	$50	http://www.z-a-recovery.com/

[13] Each OS is a stand-alone program requiring separate purchase.

[14] P indicates that the file systems are supported, but the application does not run natively in the OS.

APPENDIX B

Forensic Hardware

Glossary

There are four major categories of forensic hardware: write-blockers, cryptographic hardware, data copiers, and adapters.

Write-blockers
Floppy disks were once made with a tab that allowed them to be accessed in "write-protect mode." This manual precaution ensured that whatever was done with the data by the computer accessing it, the original disk would not change. Optical media and hard drives offer no such built-in protections, and including a hardware intermediary between the read-device and the computer provides extra assurance that the original data are unchanged.

Prices range from $150–$200 for a simple USB adapter or dock to $1,000–$2,000 for write-blocked data-duplication devices.

Cryptography devices
Hardware devices exist for both encryption and decryption. Decryption devices perform brute-force attacks that the user can hook to an encrypted device while using his or her workstation for other tasks. On the encryption side, hardware offers an extra layer of security (or barrier to entry): hardware-encrypted media cannot be decrypted without the physical key.

Because of the extreme processing power required for brute-force attacks, decryption devices cost between $5,000 and $20,000. A USB encryption key may cost as little as $10, while encrypted hard drives run between $500 and $1,000.

Data copiers
Data copiers are equipped with bays for drives or media to be copied from and to. These devices typically take bit-exact images of whatever they are copying, and are often designed with mass copying in mind.

Adapters
The number of connectors for internal and external devices is astounding: SCSI, IDE, SATA, SAS, ESDI, Firewire, a dozen varieties of USB, and more. Having every type of connection built into a machine is unlikely, especially when dealing with archival (i.e., likely obsolete) materials. In many cases, an adapter is available to convert

an acquired drive's interface into one supported by the user's system (e.g. using a SATA-to-USB cable to read a laptop hard drive with a desktop PC).

Adapters may be cables, enclosures, or dongles and range in price from $10–$100.

Vendors

In some cases, devices are sold through a vendor but developed by a third party; this tends to be true of vendors that stock complete systems. Device manufacturers are indicated by (M) in the following table.

Table B-1: Hardware Vendors

Vendor	Pre-built Systems	Write-Blockers	Decryption Devices	Encryption Devices	Detection	Data Copiers	Adapters	Web Site
Digital Intelligence	Y	Y	Y			Y	Y	http://www.digitalintelligence.com/
Forensic Computers	Y	Y	Y			Y	Y	http://www.forensic-computers.com/
Wiebetech (M)		Y			Y	Y	Y	http://www.wiebetech.com/
CRU Dataport (M)				Y			Y	http://www.cru-dataport.com/
ForensicPC	Y	Y	Y	Y	Y	Y	Y	http://www.forensicpc.com/
Paraben (M)		Y					Y	http://www.paraben-hardware.com/
Tableau (M)		Y	Y		Y	Y	Y	http://www.tableau.com/
Intelligent Computer Solutions (M)		Y		Y	Y	Y	Y	http://www.ics-iq.com
Voom Technologies (M)						Y		http://www.voomtech.com
Diskology (M)		Y				Y	Y	http://www.diskology.com
CPR Tools (M)		Y	Y[15]			Y	Y	http://www.cprtools.net/
Logicube (M)		Y				Y	Y	http://www.logicubeforensics.com

[15] CPR's DriveKey is available only to law enforcement and government agencies.

Baseline Forensic Systems

Table B-2: FRED

Vendor/Product	Specifications	
Digital Intelligence Forensic Recovery of Evidence Device (FRED) Cost: $5,999	**Processor:**	Intel i7 920 CPU (quad processor), 2.66 GHz, 8 MB cache, 4.80 GT/s Intel® QPI
	RAM:	6 GB DDR3-1333 triple channel memory
	Storage:	1 x 150 GBGB 10,000 RPM 3.0 GbGB/s SATA hard drive in shock-mounted tray 1 x 1.5 TBTB 7200 RPM 3.0 GbGB/s SATA hard drive in shock-mounted tray
	Internal Drives:	BD-R/BD-RE/DVD ± RW/CD ± RW Blu-ray burner dual-layer combo drive Digital Intelligence Integrated Forensic media card reader
	External Drives:	USB 3-1/2″ floppy drive with write-protect switch
	Port/Slots:	6 ports (6 drives) primary 3.0 GbGB/s serial ATA (SATA) controller (RAID capable) 2 ports (2 drives) SAS-serial attached SCSI controller (RAID capable) 2 ports eSATA 150/300 SATA On-the-GO (RAID capable) 1 port (2 drives) DMA 66/100/133 parallel ATA (IDE) controller 1 PS/2 combo port (keyboard & mouse) 11 USB 2.0/1.x ports: 8 back mounted, 3 front mounted (1 write blocked) 2 FireWire IEEE 1394a (400 MB/s) ports: 2 back mounted 3 FireWire IEEE 1394b (800 MB/s) ports: 1 back mounted, 2 front mounted (1 write blocked) 2 x PCI-Express (x16), 1 x PCI-Express (x1), 2xPCI-X, 1xPCI(2.2) slots
	Software:	MS-DOS 6.22 (pre-installed & configured) Microsoft Windows 98SE Standalone DOS (pre-installed & configured/installed & configured) Microsoft Windows XP Pro (pre-installed & configured/installed & configured) Suse Linux Professional (preconfigured/configured) Norton GHOST Nero DVD/CD authoring software DriveSpy, Image, PDWipe, PDBlock, PART
	Cables:	All the necessary cables, adapters, and terminators to image and process internal/external SCSI drives, 1.8-inch IDE (iPod) drives, 2.5-inch IDE (laptop) drives, and 3½-½ and 5¼-¼inch IDE drives.
	Bays:	2 x Native shock mounted SATA removable hard drive bays (IDE capable) 3 x HotSwap shock mounted universal (IDE/SATA compatible) removable hard drive bays
	Accessories:	Extendable/retractable imaging workshelf/retractable imaging workshelf with integrated ventilation Security screwdriver set integrated ventilation Security screwdriver set

Table B-3: Forensic Tower

Vendor/Product	Specifications	
Forensic Computers Forensic Tower Cost: $2,995	**Processor:**	Intel® Pentium D 940 3.2-GHz, 2X2 L2 cache, LGA 775
	RAM:	4 GB DDR2 PC2-5300 DDR2-667
	Storage:	150 GB VelociRaptor SATA II hard drive 500 GB SATA II hard drive
	Internal Drives:	1.44 floppy drive 22X DVD-RW drive 16X DVD-ROM /40X CD-ROM
	External Drives:	
	Port/Slots:	1 open PCI-X slot; 3 open PCI slots 1 front mounted and 1 back mounted FireWire 400 port 1 front mounted and 2 back mounted FireWire 800 ports 4 front mounted and 3 back mounted USB 2.0 ports 1 back mounted eSATA port
	Software:	Microsoft Windows XP Professional QuickView Plus Version 10
	Cables:	
	Bays:	Tableau T35i Forensic SATA/IDE Bridge with a DC Out Molex port, a SATA port, and an IDE port. One CRU DataPort V Plus SATA removable storage module (READ/WRITE) (Hot-Swappable). Also includes a CRU DataPort V IDE to SATA tray.
	Accessories:	30-piece security screwdriver set

Table B-4: FPC-T1

Vendor/Product	Specifications	
ForensicPC FPC-T1 Cost: $3,995	**Processor:**	Intel Core 2 Duo E7400 2.8 GHz 1066 MHz
	RAM:	2 GB DDR2
	Storage:	(2) 500 GB SATA drives (rpm unspecified)
	Internal Drives:	Dual layer DVD writer
	External Drives:	
	Port/Slots:	Write-blocked multi-format memory card reader
	Software:	
	Cables:	
	Bays:	Forensic Drive Bay Controller with multibay read/write status Shock-mounted SATA and IDE write-blocked bays
	Accessories:	Accessory drawer with adapter storage

APPENDIX C

Further Resources

We have taken a deliberately broad and catholic view of what constitutes "further resources," aiming for diversity of perspective as much as or more than completeness of coverage. Thus, textbooks and technical reports on digital forensics appear alongside works examining evidence, information, and archives across human history. We hope readers find this siting of digital forensics within a broader context useful, even as the listings provide solid guidance for further study for the serious practitioner.

We have not included any entries for articles; we refer readers instead to the section on journals offering coverage of the relevant fields. Cal Lee's bibliographies also offer excellent starting points, and are available at http://ils.unc.edu/callee/emanuscripts-stewardship/related-resources.html.

Commercial and open source software and hardware are covered in Appendixes A and B, respectively.

Books

***Note**: URLs are current as of November 22, 2010.*

Abelson, Hal, Ken Ledeen, and Harry Lewis. 2008. *Blown to Bits: Your Life, Liberty, and Happiness after the Digital Explosion*. Upper Saddle River, NJ: Addison-Wesley.

Apple Computer. 1992. *Inside Macintosh: Files*. Reading, MA: Addison-Wesley Publishing Company. Available at http://dubeiko.com/development/FileSystems/HFS/inside_macintosh/inside_macintosh.htm.

Baron, Dennis. 2009. *A Better Pencil: Readers, Writers, and the Digital Revolution*. Oxford: Oxford University Press.

Bergeron, Bryan. 2002. *Dark Ages II: When the Digital Data Die*. Upper Saddle River, NJ: Prentice Hall.

Boles, Frank. 2005. *Selecting and Appraising Archives and Manuscripts*. Chicago: Society of American Archivists.

Brown, Christopher. 2010. *Computer Evidence: Collection and Preservation*. 2nd ed. Boston, MA: Charles River Media.

Brown, John Seely, and Paul Duguid. 2000. *The Social Life of Information*. Cambridge, MA: Harvard Business School Press.

Bunting, Steve. 2008. *EnCase Computer Forensics: The Official EnCE: EnCase Certified Examiner Study Guide.* 2nd ed. Indianapolis, IN: Wiley Publishing.

Caloyannides, Michael. 2001. *Computer Forensics and Privacy*. Norwood, MA: Artech House.

Cardwell, Kevin. 2007. *The Best Damn Cybercrime and Digital Forensics Book Period*. Burlington, MA: Syngress Publishing.

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

Casey, Eoghan. 2004. *Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet*. 2nd ed. Amsterdam: Elsevier Academic Press.

Cohen, Tyler. 2007. *Alternate Data Storage Forensics*. Burlington, MA: Syngress Publishing.

Custer, Helen. 1992. *Inside Windows NT*. Redmond, WA: Microsoft Press.

Daniel, Eric D., C. Dennis Mee, and Mark H. Clark. 1999. *Magnetic Recording: The First One Hundred Years*. New York: IEEE Press.

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

Duranti, Luciana. 1998. *Diplomatics: New Uses for an Old Science*. Lanham, MD: Scarecrow Press.

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

Finn, Christina A. 2001. *Artifacts: An Archeologist's Year in Silicon Valley*. Cambridge, MA: MIT Press.

Greetham, David C. 2010. *The Pleasures of Contamination: Evidence, Text, and Voice in Textual Studies*. Bloomington, IN: Indiana University Press.

Higgs, Edward, ed. 1998. *History and Electronic Artefacts*. Oxford: Clarendon Press.

Hillis, W. Daniel. 1998. *The Pattern in the Stone: The Simple Ideas that Make Computers Work*. New York: Basic Books.

Hilton, Ordway. 1982. *Scientific Examination of Questioned Documents*. Revised edition. New York: Elsevier.

Jones, Keith J., Richard Bejtlich, and Curtis W. Rose. 2005. *Real Digital Forensics: Computer Security and Incident Response*. Upper Saddle River, NJ: Addison-Wesley.

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

Kruse II, Warren G., and Jay G. Heiser. 2002. *Computer Forensics: Incident Response Essentials*. Upper Saddle River, NJ: Addison-Wesley.

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

MacNeil, Heather. 2000. *Trusting Records: Legal, Historical and Diplomatic Perspectives*. London: Kluwer Academic Publishers.

Marcella, Albert. 2008. *Cyber Forensics: A Field Manual for Collecting, Examining, and Preserving Evidence of Computer Crimes*. 2nd ed. New York: Auerbach Publications.

Mayer-Schönberger, Viktor. 2009. *Delete: The Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press.

McGann, Jerome. 1991. *The Textual Condition*. Princeton, NJ: Princeton University Press.

McKenzie, D. F. 1999. *Bibliography and the Sociology of Texts*. The Panizzi Lectures, 1985. Cambridge, UK: Cambridge University Press.

Nelson, Bill, Amelia Phillips, and Christopher Steuart. 2010. *Guide to Computer Forensics and Investigations*. 4th ed. Course Technology Cengage Learning.

Nickell, Joe, and John F. Fischer. 1999. *Crime Science: Methods of Forensic Detection*. Lexington, KY: University of Kentucky Press.

Petzold, Charles. 2000. *Code: The Hidden Language of Computer Hardware and Software*. Redmond, WA: Microsoft Press.

Philipp, Aaron, David Cowen, and Chris Davis. 2005. *Hacking Exposed: Computer Forensics Secrets & Solutions*. 2nd ed. Emeryville, CA: McGraw Hill/Osborne Media.

Sheetz, Michael. 2007. *Computer Forensics: An Essential Guide for Accountants, Lawyers, and Managers*. Hoboken NJ: John Wiley & Sons.

Silberschatz, Abraham, Peter Baer Galvin, and Greg Gagne. 2004. *Operating System Concepts.* 7th ed. Hoboken, NJ: John Wiley & Sons.

Slade, Robert M. 2004. *Software Forensics: Collecting Evidence from the Scene of a Digital Crime*. New York: McGraw Hill.

Stille, Alexander. 2002. *The Future of the Past*. New York: Farrar, Straus and Giroux.

Stoddard, Roger L. 1985. *Marks in Books, Illustrated and Explained*. Cambridge, MA: Harvard University Press.

Tanenbaum, Andrew S. 2008. *Modern Operating Systems*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Volonino, Linda, Reynaldo Anzaldua, and Jana Godwin. 2006. *Computer Forensics: Principles and Practices*. 1st ed. Upper Saddle River, NJ: Prentice Hall.

Von Hagen, William. 2002. *Linux Filesystems*. Indianapolis, IN: Sams Publishing.

Wang, Wallace. 2001. *Steal This Computer Book: What They Won't Tell You about the Internet*. 2nd ed. San Francisco: No Starch Press.

Technical References and Reports

Bearman, David. 1994. *Electronic Evidence: Strategies for Managing Records in Contemporary Organizations*. Pittsburgh, PA: Archives and Museum Informatics.

Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. Available at http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf.

Byers, Fred. 2003. *Care and Handling of CDs and DVDs: A Guide for Librarians and Archivists*. Washington, DC: Council on Library and Information Resources.

CLIR. 2000. *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources.

CLIR. 2002. *The State of Digital Preservation: An International Perspective*. Washington, DC: Council on Library and Information Resources.

CLIR. 2004. *Access in the Future Tense*. Washington, DC: Council on Library and Information Resources.

Department of Defense, Department of Energy, Nuclear Regulatory Commission, and Central Intelligence Agency. 1995–97. DoD 5220.22-M National Industrial Security Program Operating Manual. Washington DC: US Government Printing Office.

Depocas, Alain, Jon Ippolito, and Caitlin Jones, eds. 2003. *The Variable Media Approach.* New York: Guggenheim Museum. Available at http://www.variablemedia.net/pdf/Permanence.pdf.

Goldston, James, and National Computer Security Center (US). 1991. *A Guide to Understanding Data Remanence in Automated Information Systems*. 2nd ed. Fort George G. Meade, MD: National Computer Security Center. Available at http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA393188.

John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. 2010. *Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis* (Beta Version 0.2). Digital Lives Research Paper (3 March). Available at http://britishlibrary.typepad.co.uk/digital_lives.

Kahn, Miriam B. 2003. *Disaster Response and Planning for Libraries*. 2nd ed. Chicago, IL: ALA Editions.

Kirschenbaum, Matthew G., Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. 2009. *Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use*. White Paper. Washington, DC: National Endowment for the Humanities. Available at http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37.

Lord, Philip, and Alison Macdonald. 2003. *e-Science Curation Report. Data Curation for e-Science in the UK: An Audit to Establish*

Requirements for Future Curation and Provision. Twickenham: JISC Committee for the Support of Research. Available at http://www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf.

McDonough, Jerome P., et al. 2010. *Preserving Virtual Worlds: Final Report*. Available at https://www.ideals.illinois.edu/handle/2142/17097.

McPherson, Andrew. 2004. *Law Enforcement Tools and Technologies for Investigating Cyber Attacks: Gap Analysis Report*. Hanover, NH: Dartmouth College Institute for Security, Technology, and Society. Available at http://www.ists.dartmouth.edu/projects/archives/gar.html.

National Library of Australia. n.d. Digital Preservation—Recovering and Converting Data from Manuscripts Collection Discs. Available at http://www.nla.gov.au/preserve/digipres/recovering.html.

Paradigm Project. 2008. *Workbook on Digital Private Papers*. Available at http://www.paradigm.ac.uk/workbook/index.html.

Pollitt, Mark, and Sujeet Shenoi, eds. 2006. *Advances in Digital Forensics.* IFIP International Conference on Digital Forensics, National Center for Forensic Science, Orlando, Florida, February 13-16, 2005. New York: Springer.

Ross, Seamus, and Ann Gow. 1999. *Digital Archaeology: Rescuing Neglected and Damaged Data Resources. A JISC/NPO Study within the Electronic Libraries (eLib) Programme on the Preservation of Electronic Materials.* Available at http://eprints.erpanet.org/47/.

Rothenberg, Jeff. 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Washington, DC: Council on Library and Information Resources.

Science and Technology Council. 2007. *The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials*. Academy of Motion Picture Arts and Sciences.

Waters, Donald, and John Garrett. 1996. *Preserving Digital Information, Report of the Task Force on Archiving of Digital Information*. Washington, DC: Council on Library and Information Resources.

Working Together or Apart: Promoting the Next Generation of Digital Scholarship. 2009. Washington, DC: Council on Library and Information Resources.

Organizations

Alliance of Digital Humanities Organizations: http://digitalhumanities.org/

American Academy of Forensic Sciences: http://www.aafs.org/

Association of Canadian Archivists: http://archivists.ca/

Computer Crime & Intellectual Property Section, United States Department of Justice: http://www.cybercrime.gov/

Digital Curation Centre: http://www.dcc.ac.uk/

Digital Forensics Research Conference: http://www.dfrws.org/

Digital Preservation Coalition: http://www.dpconline.org/

National Digital Information Infrastructure and Preservation Program: http://www.digitalpreservation.gov/

National Institute of Standards and Technology: http://www.nist.gov

Rare Book School: http://www.rarebookschool.org/

Rare Books and Manuscripts Section, Association of College and Research Libraries: http://www.rbms.info/index.shtml

Society of American Archivists: http://www2.archivists.org/

Software Preservation Society: http://www.softpres.org/

Selected Projects and Other Resources

AIMS: An Inter-Institutional Model for Stewardship: http://www2.lib.virginia.edu/aims/

Brian Carrier: Digital Investigation: Forensics and Evidence Research: http://www.digital-evidence.org/

Computer Forensics Reference Data Sets (CFReDS) Project: http://www.cfreds.nist.gov/

DFI News. *Digital Forensic Investigator*: http://www.dfinews.com/

DigCCurr: http://www.ils.unc.edu/digccurr/

Digital Forensics @ Stanford Libraries: http://lib.stanford.edu/digital-forensics

Digital Preservation Management Workshops and Tutorial: http://www.icpsr.umich.edu/dpm/index.html

Digital Records Forensics Project: http://www.digitalrecordsforensics.org/

E-Evidence Information Center: http://www.e-evidence.info/

Electronic Evidence Resource List: Legal, Technical and Training. Office of Justice Programs: http://www.ojp.gov/nij/topics/technology/electronic-crime/resources.htm

The Ethical Hacker Network: http://www.ethicalhacker.net/

FileFormat.Info: http://www.fileformat.info/

Forensics Wiki: http://www.forensicswiki.org/

futureArch: http://futurearchives.blogspot.com/

InterPARES Project: http://www.interpares.org/

KEEP: Keeping Emulation Environments Portable: http://www.keep-project.eu/ezpub2/index.php

MITH's Vintage Computers: http://mith.umd.edu/vintage-computers/

National Center for Forensic Science Digital Evidence Research: http://www.ncfs.org/research_digital.html

National Software Reference Library: http://www.nsrl.nist.gov/

NIST Computer Forensic Tool Testing Program: http://www.cftt.nist.gov/

Paradigm: http://www.paradigm.ac.uk/

Planets: http://www.planets-project.eu/

Preserving Virtual Worlds: http://pvw.illinois.edu/pvw/

Textfiles: http://www.textfiles.com/

Journals

2600: The Hacker Quarterly. Available at http://www.2600.com/.

Archivaria. Available at http://journals.sfu.ca/archivar/index.php/archivaria.

American Archivist. Available at http://archivists.metapress.com/home/main.mpx.

D-Lib Magazine. Available at http://www.dlib.org/.

Digital Investigation: The International Journal of Digital Forensics and Incident Response. Available at http://www.elsevier.com/wps/find/journaldescription.cws_home/702130/description#description.

Haking: IT Security Magazine. Available at http://www.hakin9.org/.

International Journal of Digital Crime and Forensics (IJDCF). Available at http://www.igi-global.com/Bookstore/TitleDetails.aspx?TitleId=1112&DetailsType=Description.

International Journal of Digital Curation. Available at http://www.ijdc.net/index.php/ijdc.

International Journal of Digital Evidence. Available at http://www.utica.edu/academic/institutes/ecii/ijde/.

International Journal of Electronic Security and Digital Forensics. Available at http://www.inderscience.com/browse/index.php?journalCODE=ijesdf.

Journal of Digital Forensics, Security, and Law. Available at http://www.jdfsl.org/index.htm.

Journal of Digital Forensic Practice. Available at http://www.tandf.co.uk/journals/titles/15567281.asp.

Journal of the Society of Archivists. Available at http://www.archives.org.uk/publications/journalofthesocietyofarchivists.html.

APPENDIX D

The Maryland Symposium

Computer Forensics and Cultural Heritage

University of Maryland, May 14–15, 2010

An integral part of the proposal for the research and writing of this report was an invitational symposium on *Computer Forensics and Cultural Heritage*, held May 14–15, 2010, and hosted by the Maryland Institute for Technology in the Humanities (MITH) on the campus of the University of Maryland in College Park—a location designed to exploit the concentration of government and industry expertise in the surrounding area. Some 60 individuals, representing archives, information and library science, computer science, the forensics industry, government agencies, and the world of scholarship, attended the meeting. To the best of our knowledge, it was the first large-scale meeting ever to be convened on the convergence of digital forensics and cultural heritage. The meeting served the dual purposes of allowing for comment on a draft version of this report, and providing a catalyst for contact between personnel from these otherwise seemingly disparate fields, with the aim of leading to more regular occasions for knowledge exchange and the development of shared research agendas. A Web site used in support of the meeting, including a complete list of attendees, is available at http://mith.info/forensics/.

Day one of the event was devoted to formal presentations clustered around such rubrics as perspectives, education, fieldwork, and government practices. The program was designed to accommodate both broad-reaching theoretical statements and detailed reports from those already engaged in hands-on work with forensics methods and tools. Speakers included Luciana Duranti (University of British Columbia), William Eber (Department of Defense Cybercrime Center), Stephen Eniss (Folger Shakespeare Library), Amy Friedlander (*Journal on Computing and Cultural Heritage*), Patricia Galloway (University of Texas), Simson Garfinkel (Naval Postgraduate School), Brad Glisson (University of Glasgow), Barbara Gutmann (National Institute of Standards and Technology), Peter Hornsby (Emory University), Jeremy Leighton John (British Library), Leslie Johnston (National Digital Information Infrastructure and Preservation Program [NDIIPP]), Cal Lee (University of North Carolina at Chapel Hill), Clifford Lynch (Coalition for Networked Information), Rob Maxwell

(University of Maryland), Michael Olson (Stanford University), Seamus Ross (University of Toronto), Leo Scanlon (National Archives), and Susan Thomas (Bodleian Libraries). Each session included ample time for questions and discussion from the audience. The agenda also included an hour for lightning talks, for which participants were able to sign up at the meeting site. These constituted eight additional presentations.

Day two opened with an hour-long presentation of the draft report by coauthors Matthew Kirschenbaum, Richard Ovenden, and Gabriela Redwine. (The draft had also been previously circulated to selected attendees.) The meeting then divided into breakout groups facilitated by each of the three coauthors, which allowed for an hour of focused and candid feedback. Many attendees also passed annotated electronic or hard copy of the report to the authors with additional notes and suggestions. The meeting concluded with a wrap-up session devoted to summarizing conclusions and articulating an agenda for next steps. (This discussion heavily informed the conclusions and recommendations in this report.) The authors spent the remainder of the second day in conference with Duranti, Glisson, Lee, Maxwell, Reside, and Thomas, assessing the impact of the meeting and developing a revision strategy for the report.

Audio from both days of the proceedings was captured and used by the authors as a reference in the course of their revisions.

Slides and audio from a number of the first day's presentations, available at http://mith.info/forensics/?page_id=120, complement the material covered in these pages. (The slides and audio are also accessible as "Presentations" from the main site link above.)

Clifford Lynch discussed the symposium in the May 2010 edition of the podcast *CNI Conversations*. The event was also written up for the Library of Congress's NDIIPP blog.[1] Twitter traffic is available under the hashtag #4n6umd.

The authors regard the meeting as an invaluable opportunity to survey representatives from relevant communities on issues covered in the report and to obtain their feedback on matters both general and particular. This report should not, however, be taken to represent a strict consensus among the attendees at the meeting, nor do the authors seek to place the burden of errors or misstatements on any persons but themselves.

[1] See http://news.cni.org/2010/06/02/cni-conversations-may-recording-available/ and http://www.digitalpreservation.gov/news/2010/20100610news_article_forensics_meeting.html, respectively.