Archival application of digital forensics methods for authenticity, description and access provision

Christopher A. Lee

When acquiring born-digital materials, archivists must often extract digital materials from media in ways that reflect the rich metadata associated with records and ensure records' integrity. They must also allow users to make sense of materials and understand their context, while preventing inadvertent disclosure of sensitive data. There are a variety of methods and strategies from the field of digital -forensics that can aid this work. This paper discusses the development and application of digital forensics tools to improve the acquisition, management and access functions of archives. It reports on the BitCurator project, which is identifying current and desirable workflows of several archival institutions, as well as developing and testing tools to support the workflows. There are a variety of potential changes within the archival profession that are associated with adopting digital forensics tools and practices.

Introduction*

Materials with archival value are now predominantly 'born digital,' and archivists have unprecedented opportunities to acquire and preserve traces of human and associated machine activity. In order to seize these opportunities, archivists must be able to extract digital materials from their storage or transfer media in ways that reflect the metadata and ensure the integrity of the materials. They must also support and mediate appropriate access: allowing users to make sense of materials and understand their context, while also preventing inadvertent disclosure of sensitive data. There are a variety of methods, strategies and applications from the field of digital forensics that can aid this work.

Applying archival principles to born-digital acquisitions

Any new application of information and communication technology to archival practices ultimately should be driven by archival principles and values. There are three fundamental archival concepts that can be advanced through the adoption of digital forensics tools and methods: provenance, original order and chain of custody.

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches graduate and continuing education courses in archival administration, records management, digital curation, and information technology for managing digital collections. His research focuses on curation of digital collections and stewardship of personal digital archives. Cal is Principal Investigator for the BitCurator project and editor of *I, Digital: Personal Collections in the Digital Era*.

* This article has been peer reviewed by members of the International Council on Archives' Section on Education and Training (ICA–SAE).

doi:10.3828/comma.2012.2.14

The provenance of a record is its 'life history.' For purposes of describing archival collections, one of the most important aspects of provenance is the identification of one or more origins or sources of a record (for example, the person who wrote a diary entry or the specific business transaction that generated a receipt). However, provenance more broadly 'consists of the social and technical processes of the records' inscription, transmission, contextualization, and interpretation which account for its existence, characteristics, and continuing history.'[1] According to the principle of provenance, records from a common origin or source should be managed together as an aggregate unit and should not be arbitrarily intermingled with records from other origins or sources.

There are many different interactions with records that are important to document, in order to understand the records' origins and 'life history' (e.g., those who influenced the creation of the records, those who received them, custodians who transformed them over time), not simply one isolated moment of creation. These considerations illustrate the importance of provenance not only as the source of a record but also as a 'history of the ownership … used as a guide to authenticity or quality' and 'a documented record of this [history]'.[2] Given the complex and evolving relationships between entities (e.g., people, agencies) and records, provenance is not simply a matter of identifying the one person who created a record at a point in time but instead 'relate[s] a multitude of contextual entities to a multitude of recordkeeping entities in a multitude of ways.'[3] In digital environments, it can be important to consider provenance at levels of granularity finer than an entire record, such as why a specific data element appears within a dataset and where specifically the data element was generated;[4] and to include additional technical components in one's notion of provenance, such as system configuration information.[5]

Closely related to provenance is the principle of original order, which indicates that archivists should organize and manage records in ways that reflect their arrangement within the creation environment. For personal records, the principle of original order implies that archivists should carry forward (either by perpetuating or attempting to reconstruct) the peculiar ways in which individuals label and organize their own records. A compelling argument for retaining original order in a digital environment is that - even if that order is messy and idiosyncratic - it conveys meaningful information about the recordkeeping context, and additional layers of description can be laid on top of that order to facilitate various forms of navigation and access.[6] However, rather than simply 'freezing or restoring

1 NESMITH Tom, 'Still fuzzy, but more accurate: Some thoughts on the 'ghosts' of archival theory' in *Archivaria*, 47, 1999, p. 146.

2 *Oxford English Dictionary*, s.v. 'provenance'.

3 HURLEY Chris, 'Problems with provenance' in *Archives and Manuscripts*, 23:2, 1995, pp. 256–257.

4 BUNEMAN Peter, KHANNA Sanjeev and TAN Wang-Chiew, 'Why and where: A Characterization of data provenance', in VAN DEN BUSSCHE Jan and VIANU Victor (eds), *Database Theory – ICDT 2001: 8th International Conference, London, UK, January 2001. Proceedings*, Springer, Berlin, 2001, pp. 316–330.

5 GUERCIO Maria, 'Archival theory and the principle of provenance for current records: Their Impact on arranging and inventorying electronic records' in ABUKHANFUSA Kerstin and SYDBECK Jan (eds), *The Principle of Provenance: Report from the First Stockholm Conference on Archival Theory and the Principle of Provenance, 2–3 September 1993*, Swedish National Archives, Stockholm, 1994, p. 82.

6 HORSMAN Peter, 'Dirty hands: A new perspective on the original order' in *Archives and Manuscript*, 27:1, 1999, pp. 42–53.

one particular past arrangement as "the" original order,"[7] original order is most usefully understood within the context of a larger, ongoing chain of custody.

The chain of custody is the 'succession of offices or persons who have held materials from the moment they were created.'[8] For purposes of legal compliance, authenticity, evidential integrity, and legal admissibility, the ideal recordkeeping system would provide 'an unblemished line of responsible custody'[9] through control, documentation, and accounting for all states of a record and changes of state (e.g., movement from one storage environment to another, transformation from one file format to another) throughout its existence - from the point of creation to each instance of use and (when appropriate) destruction.

The reality of contemporary information management is rarely consistent with the recordkeeping ideal. In most cases, the best that an information professional can do is to capture or create limited documentation of the portion of the chain of custody that occurred before he/she first encountered the records, and then attempt to provide much more detailed chain of custody control and documentation from that point forward. For example, an archivist acquiring a floppy disk containing records from a donor often will not know with certainty what the states and transitions of the records were before they were last saved onto that disk, but she can use various forms of information (e.g., other records, discussions with the donor) to make inferences about earlier points in the 'life' of the records. Tom Nesmith points out that archivists' knowledge about various aspects of the 'origins of a record' are 'bathed in hypothesis.'[10]

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

The literature on digital archives tends to place a great emphasis on the 'virtual' (i.e. intangible) nature of electronic resources. Computer systems have 'an illusion of immateriality by detecting error and correcting it,'[11] but it is essential to recognize that digital objects are created and perpetuated through physical things (e.g. charged magnetic particles, pulses of light, indentations on disks). This materiality brings challenges, because data must be read from specific artifacts, which can become damaged or obsolete. However, the materiality of digital objects also brings unprecedented opportunities for description, interpretation and use.[12]

7 HORSMAN Peter, 'The Last dance of the phoenix, or the de-discovery of the archival fonds' in *Archivaria*, 54, 2002, p. 19.

8 PEARCE-MOSES Richard, *Glossary of Archival and Records Terminology,* Society of American Archivists, Chicago, p. 67.

9 JENKINSON Hilary, *A Manual of Archive Administration: Including the Problems of War Archives and Archive Making,* Clarendon Press, Oxford, 1922, p. 11.

10 NESMITH, op. cit., p. 141.

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

12 LAVAGNINO John, 'The Analytical bibliography of electronic texts', Paper presented at the Joint Annual Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Bergen, Norway, 1996.

If records are 'persistent representations of activities or other occurrents,' it is important to recognize that one 'can expect to find representations at many different levels.'[13] These are not just levels in the functional hierarchy of records but also levels of representation. Digital records can be considered and encountered at multiple levels of representation, ranging from aggregations of records down to bits as physically inscribed on a storage medium; each level of representation can provide distinct contributions to the information and evidential value of records.[14] There is a substantial body of information within the underlying data structures of computer systems that can often be discovered or recovered, revealing new types of records or essential metadata associated with existing record types.

Recovery of data from physical media has been a topic of discussion in the professional library and archives literature for several years. More than a decade ago, a report by Seamus Ross and Ann Gow discussed the potential relevance of advances in data recovery and digital forensics to collecting institutions.[15] More recently, there has been an active stream of literature related to the use of forensic tools and methods for acquiring and managing digital collections. Many of the recent and ongoing activities in this space are discussed in a white paper produced by the BitCurator project.[16]

Forensic tools and methods to support archival functions

Access to data from a storage device normally involves mounting a volume and then copying or opening files through the filesystem. There must be hardware to detect signals on the medium, hardware and software to translate the signals into bitstreams, and hardware and software to move the bitstreams into the current working computer environment. One can then interact with data as entire files or components of files. The filesystem usually plays a mediating role between the user and the underlying data, and it is designed to facilitate interaction at the file level (e.g. file naming, viewing timestamps, access controls). The filesystem serves to 'hide' complicated information from the user about 'where and how it stores information.'[17] For most purposes, the filesystem is a very valuable abstraction mechanism, because it does not require users to understand or directly access the underlying data.

Those who are interested in the underlying data that are hidden by the filesystem can instead generate and interact with disk images, which are low-level, sector-by-sector copies of all the data that reside on the storage medium. Inspection of the disk image can reveal a significant amount of information that users of the drive did not consciously or intentionally leave there but can serve as traces of valuable contextual information. Forensic workflows

13 YEO Geoffrey, 'Concepts of record (2): Prototypes and boundary objects' in *American Archivist*, 71:1, 2008, pp. 118–143.

14 LEE Christopher A., 'Digital curation as communication mediation' in MEHLER Alexander, ROMARY Laurent, and GIBBON Dafydd (eds), *Handbook of Technical Communication*, Mouton De Gruyter, Berlin, 2012, pp. 507–530.

15 ROSS Seamus and GOW Ann, 'Digital archaeology: Rescuing neglected and damaged data resources', British Library, London, 1999.

16 LEE Christopher A., WOODS Kam, KIRSCHENBAUM Matthew, and CHASSANOFF Alexandra, *From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions*, 30 September 2013, http://www.bitcurator.net/docs/bitstreams-to-heritage.pdf [accessed 16 Dec. 2013].

17 FARMER Dan and VENEMA Wietse, *Forensic Discovery,* Addison-Wesley, Upper Saddle River, NJ, 2005.

often involve creation of a disk image to serve as a baseline copy of the data from the disk, upon which many further extraction and analysis tasks can be performed. Digital forensics professionals use hardware write blockers to ensure that no data on the disk – including essential metadata such as timestamps – are altered or overwritten during the process of copying the disk's contents.

Archives can incorporate a variety of forensics practices and methods by treating disk images, rather than individual files or packaged directories, as basic units of acquisition.[18] Using write blockers, creating full disk images and extracting data associated with files is essential to ensuring provenance, original order and chain of custody. Incorporation of digital forensics methods also will be essential to the sustainability of archives as stewards of personally identifying information; the same tools that are used to expose sensitive information can be used to identify, flag and redact or restrict access to it.

Emerging emphasis on personally controlled records

Much of the recent innovation in the application of digital forensics to archives has been undertaken within the context of acquiring records that were within the control of individuals, as opposed to records that come from formal enterprise recordkeeping systems. This includes personal papers and other non-institutional records that have traditionally been associated with the 'manuscripts tradition,' and there has been a recent influx of publications in the archival literature related to personal archives, with much of the focus being on born-digital records created by individuals. It is also important to recognize and address the numerous records within the responsibility of institutional archives (e.g. government records, corporate records) that are not managed within enterprise recordkeeping systems but are instead stored and managed by individuals on personal computers, mobile devices and external storage media. The acquisition and processing of born-digital records received on removable media is creating promising bridges between institutional archives and collecting archives or manuscript repositories.[19]

BitCurator

The BitCurator project[20] is a joint effort – led by the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill and Maryland Institute for Technology in the Humanities (MITH), and involving contributors from several other institutions - to develop a system for librarians and archivists to incorporate the functionality of many open-source digital forensics tools into their work practices.[21]

18 WOODS Kam, LEE Christopher A., and GARFINKEL Simson, 'Extending digital repository architectures to support disk image preservation and access' in *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries,* ACM Press, New York, 2011, pp. 57–66.

19 This is not a distinction that holds consistently across languages, nations or recordkeeping traditions. However, it is a division that has had significant professional implications in many countries. For a discussion of differences in terminology around this issue, see LEE Christopher A., 'Introduction' in *I, Digital: Personal Collections in the Digital Era*, edited by LEE Christopher A., Society of American Archivists, Chicago, 2011, pp. 1–26.

20 See http://bitcurator.net [accessed 16 Dec. 2013]. The BitCurator project is supported by a grant from the Andrew W. Mellon Foundation.

21 LEE Christo pher A., KIRSCHENBAUM Matthew, CHASSANOFF Alexandra, OLSEN Porter and

Digital forensics offers valuable methods that can advance the archival goals of maintaining authenticity, describing born-digital records and providing responsible access.[22] However, most digital forensics tools were not designed with archival objectives in mind. The BitCurator project is attempting to bridge this gap through engagement with digital forensics, library and archives professionals, as well as dissemination of tools and documentation that are appropriate to the needs of memory institutions. The BitCurator software is all open-source and freely available to download and install.[23]

Much BitCurator activity is translation and adaptation work, based on the belief that archivists will benefit from tools that are presented in ways that use familiar language and run on platforms that archivists can support. BitCurator – and the efforts of many of the project partners – also aim to address two fundamental needs of archives that are not priorities for digital forensics industry software developers:

1. Incorporation into the workflows of archives and libraries, e.g. supporting metadata conventions, connections to existing content management system (CMS) environments. This includes exporting forensic data in ways that can then be imported into archival descriptive systems, as well as modifying forensics triage techniques to better meet the needs of archivists.
2. Provision of public access to the data. The typical digital forensics scenario is a civil lawsuit or criminal investigation in which the public never gets direct access to the evidence. By contrast, archives that are creating disk images face issues of how to provide access to the data. This includes not only providing access interfaces, but also redacting or restricting access to components of the image, based on confidentiality, intellectual property or other sensitivities.

Two groups of external partners are contributing to BitCurator: a Professional Expert Panel (PEP) of individuals who are at various stages of implementing digital forensics tools and methods in their collecting institution contexts, and a Development Advisory Group (DAG) of individuals who have significant experience with development of software.

The project is packaging, adapting and disseminating a variety of open-source applications. Rather than developing everything from scratch, BitCurator is able to benefit from numerous existing open-source tools, many of which are now quite mature.[24] The goal is to provide a set of tools that can be used together to perform archival tasks but can also be used in combination with many other existing and emerging applications.

WOODS Kam, 'BitCurator: Tools and techniques for digital forensics in collecting institutions' in *D-Lib Magazine*, 18: 5/6, May/June 2012.

22 WOODS Kam and LEE Christopher A., 'Acquisition and processing of disk images to further archival goals' in *Proceedings of Archiving 2012,* Society for Imaging Science and Technology, Springfield, VA, 2012, pp. 147–152.

23 For BitCurator software, installation instructions and various forms of documentation, including instructional videos, see http://wiki.bitcurator.net [accessed 16 Dec. 2013].

24 Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files.

Conclusion

As archivists take on the curation of born-digital materials such as floppy disks found in boxes and new acquisitions on media such as hard drives and flash drives, they are now learning and applying many methods that have been used within digital forensics for many years. Digital forensics tools and methods hold great promise for enhancing and improving the work practices of archivists who are responsible for digital records.

There are a variety of changes within the archival profession that are implied by the above trend. First, the professional vocabulary of archivists is evolving to now include terms such as disk image, hex[adecimal] viewer, cryptographic hash, and filesystem. Second, archivists are gaining access to new professional communities and sources of guidance, e.g. papers from the annual Digital Forensics Research Workshop and instructions from gaming enthusiasts about how to create, read and mount disk images of old storage media. The first and second points are closely related; having the right vocabulary can open up many new mechanisms for learning and sharing information.

A third change in the archival profession comes from the use of tools that are designed to treat data at a very low level – as raw bitstreams off media – rather than treating data at the file level. Archivists have long argued that the essential content, structure and context elements of an electronic record can reside in multiple data sources and not just in a single file.[25] Digital forensics greatly enables such thinking; for example, it allows archivists to bypass the filesystem and read data as a raw stream to be decomposed into records as appropriate.

Finally, the introduction of digital forensics into archives has the potential to shift the 'centre of gravity' about electronic records in the archival literature from the design of institutional recordkeeping systems toward the acquisition and management of records from a much more diverse and unpredictable set of sources. Building recordkeeping functionality into live systems is as important as ever, but it is joined by concerns and activities related to records of continuing value that have not been subjected to systematic recordkeeping control.

The intersection between digital forensics and archives can be characterized as a 'trading zone' that resides between different streams of activity.[26] Individuals and groups can agree to use a common set of terms, concepts and methods in order to share ideas and coordinate their work, even if they still hold dramatically different worldviews, values or assumptions of their own responsibilities. It is likely that fundamental elements of digital forensics language and practice will ultimately become so embedded in the archival enterprise that archivists no longer perceive them as being borrowed from elsewhere; they will simply be part of what archivists do. As archivists develop new methods and tools that are based on forensics building blocks, hopefully they will also make contributions to the field of digital forensics that it can ultimately adopt as established practice. However, it is also likely that the frontiers of digital forensics and archival research will continue to develop independently, based on distinct values, mandates and constraints. There is the potential for creative and well-informed translation work across the two streams for many years ahead.

25 See e.g., BEARMAN David, 'Record-keeping systems' in *Archivaria*, 36, 1993, pp. 16–36; McDONALD John, 'Towards automated record keeping, interfaces for the capture of records of business processes' in *Archives and Museum Informatics*, 11, 1997, pp. 277–285.

26 GALISON Peter Louis, *Image and Logic: A Material Culture of Microphysics,* University of Chicago Press, Chicago, 1997.