Acquisition and Processing of Disk Images to Further Archival Goals

Kam Woods and Christopher A. Lee; School of Information and Library Science, University of North Carolina at Chapel Hill; Chapel Hill, NC, USA

Abstract

Disk imaging can provide significant data processing and information extraction benefits in archival ingest and preservation workflows, including more efficient automation, increased accuracy in data triage, assurance of data integrity, identifying personally identifying and sensitive information, and establishing environmental and technical context. Information located within disk images can also assist in linking digital objects to other data sources and activities such as versioning information, backups, related local and network user activity, and system logs. We examine each of these benefits and discuss the incorporation of modern digital forensics technologies into archival workflows.

Introduction

Acquiring and processing information from raw digital sources such as hard disks and removable digital storage media is a common task for archives and other collecting institutions. These media often contain significant amounts of contextual information along with potentially private and sensitive information in both created content and file and system metadata. Identification and management of this supporting information can be critical to ensure compliance with donor or submission agreements, establish provenance, and enable future access. There is currently little standardization in the methods used to process and analyze digital media. Disk imaging – the process of creating a bit-identical copy of the source media – along with automated disk image analysis can assist in addressing these issues.

Integrating digital disk imaging into archival workflows can help collecting institutions to ensure the authenticity, integrity, and provenance of digital materials. More than a decade ago, a report by Seamus Ross and Ann Gow [10] discussed the potential relevance of advances in data recovery and digital forensics to collecting institutions. More recently, there has been an active stream of literature related to the use of disk imaging and associated forensic tools and methods for acquiring and managing digital collections [1, 3, 6, 8, 9, 12, 13, 14].

In concert with raw bitstream acquisition, packaging mechanisms derived from the field of digital forensics incorporate extensive metadata about the capture process and the format, organization, and other characteristics of an original physical device. Interoperable metadata formats, including Digital Forensics Extensible Markup Language (DFXML), can be shared among systems and applications [4]. Disk imaging and image analysis techniques can lower risk in the acquisition process, provide expanded opportunities for recovery of damaged and "lost" data, and facilitate various access and analysis goals.

In this paper we describe methods for and potential benefits of creating disk images. We examine packaging formats used in digital forensics, focusing on those that embed or include support for interoperable metadata formats. We examine how metadata can be shared among systems and applications to provide archivists with simplified mechanisms and workflows for filesystem analysis. We summarize our experiences in developing tools and methods to automate disk image analysis: lowering the risk of information loss and future release of private or personally identifying information; triage of problematic documents and software items; and establishing baseline parameters to support data integrity and access. We discuss application of these techniques in BitCurator, a project to integrate digital forensics procedures and software into archival workflows.

Disk Imaging

Disk imaging – extracting unaltered bitstreams from digital storage media (magnetic, optical, or solid-state) – is used for a variety of purposes. These include data rescue and recovery, full backup, cloning of drives to provision new hardware, and the creation of images that can be mounted and used as virtual drives by an operating system. There are many applications that can create raw disk images, including the UNIX dd utility (and derivatives), open source graphic user interface (GUI) applications such as Clonezilla,, and commercial products such as EASUS disk copy.

The imaging process is generally agnostic with respect to the organization of the underlying filesystem(s), as it copies data sector-by-sector from the raw device. All organizational characteristics of the original data store are retained, including (for a given physical or logical volume) partitions containing active file system(s) and unused portions of the drive, sector contents marked as deleted, and "slack" space from partially-filled sectors.

A complete copy of the sectors on the underlying physical device affords the imaging software user a great deal of flexibility. A cryptographic hash of the raw image file can be used to verify whether any changes have been incurred on the physical media in future handling or access events. Filesystems that are damaged or contain recovery logs can be manipulated (using a copy of the master image) without risk of data loss.

More generally, parsing created image contents (either programmatically or interactively via virtualization) *in their original context* provides the user with valuable information about how the device was organized, who used it, and which users had access to particular contents on the device.

Raw disk images have some limitations. As sector-by-sector copies of the drive contents, they do not retain additional metadata about the capture process or supporting actions performed during acquisition. In the digital forensics community, practitioners must demonstrate chain-of-custody for seized evidence. Consequently, companies and researchers have developed several binary packaging and wrapper formats to encapsulate both disk images and metadata generated during imaging. Such metadata often includes information about the user performing the imaging, the

physical storage medium, the system on which the imaging was performed, and various cryptographic checksums and timestamps.

In the following sections we describe how access to bit-identical disk images and metadata provided by forensic packaging can be useful in an archival context.

Benefits of Imaging

Disk imaging can significantly lower various risks in the acquisition process. Foremost among these is operational risk in handling of the original media.

First, a single pass over the contents of a physical device reduces the ongoing risk of physical failure. Such failure may include drive head malfunction, bearing failure, and platter separation on older hard disks; read failures due to excessive handling of fragile magnetic and optical media; failure or degradation of flash memory on removable flash media, and mitigation of operations performed automatically by the flash translation layer on solid state drives (SSDs).

Second, in conjunction with the use of write-blocking hardware (which prevents the host from writing any data to the storage medium), disk images can be reliably extracted while ensuring that no changes are made deliberately, inadvertently, or by automated processes on the host system. Examples of such changes include automated recovery of damage to journaling filesystems, automated indexing data written to improve performance (notably Mac OS X's Spotlight), and batch operations for drive cleanup (such as TRIM commands for SSDs).

Another benefit of disk imaging is that *all* metadata are retained from the original filesystem. This is significant both for its technical consequences and for provenance implications. As an example, consider a file which is copied from an NTFS-based Windows system onto an HFS+-based Macintosh drive. Even if this file is renamed, the cryptographic checksum remains the same. However, the file will now have new creation and modification timestamps, and complex metadata concerning local and network users (contained in the NTFS Security Descriptor) are lost. Without the original filesystem, it is impossible to create an accurate permissions record. Disk imaging alleviates this issue by allowing the acquiring party to identify and export permissions metadata, which can be retained as part of an Archival Information Package (AIP).

Disk images that contain complete operating systems capture significant information about the "digital ecosystem" in which documents and media were created. Retention of the complete disk image allows for the documentation of applications (and application versions) used to produce documents and media. Data on a disk image may identify users and groups associated with a particular device and identify traces of network activity and online services that were used to import, transfer, or produce content. Disk images can be parsed to identify supporting software mechanisms used to produce documents (including software libraries, fonts, and linked objects). Finally, disk image contents can be used to identify "lost," deleted, and potentially private information.

Image Analysis

Disk images provide flexible analysis paths for data ingested into an archival repository. These include parsing filesystems directly, analyzing the image as a single data stream, and exporting statistical information on image contents.

Disk images often provide performance and security benefits. Speed of access to images will generally be higher than removable (or legacy fixed) media, and eliminates the requirement for continuous access to specialized hardware. Common filesystems usually can be parsed using software libraries without mounting the image, and the risk of malware transmission is reduced.

Disk images facilitate automation of data triage and data integrity tasks. Partition maps provide a simple high-level overview of the source media. Filesystem consistency checks can be run without risking alteration of the bitstream. Cryptographic filesystems and encrypted partitions can be identified via entropy analysis and system artifacts. Finally, disk images can be used to produce unified maps or hierarchies of both allocated and unallocated space on the original device.

These types of general, baseline reporting can improve efficiency in the data triage process. Hashes can be used to efficiently map both the degree of replication on a device and permissions associated with replicated objects, and prioritize those which may pose preservation issues. Versioning and backup tools originally in place can be identified, allowing recovery of both digital objects and records of those events (e.g. Time Machine backups). Metadata produced by the operating system can assist in determining whether individual file objects were transferred to or from external media, or distributed over a network.

Stream-based analysis of a disk image (e.g. by reading a fixed-byte window of data from the bitstream, disregarding the organization of the filesystem) allows for efficient identification of many instances of potentially private and sensitive data contained within the image.

Each of these modes of analysis can help both archivists and users of archival collections to identify contextual information for making sense of digital objects.

Archival Concerns

Generation and management of disk images remains relatively rare in current repositories. This could be due in part to a general lack of familiarity with the benefits of the process, and in part to the assumption that the creation of disk images dictates a "save everything" approach that is not appropriate for many materials. Collecting institutions may additionally have concerns over whether the creation of disk image violates institutional protocols or donor agreements. In this section we examine each of these issues and provide suggestions for addressing and mitigating them.

Privacy and Confidentiality

Repositories should establish policies and guidelines that indicate circumstances in which disk imaging is appropriate for digital materials acquired from individuals and organizations. While the applicability of this practice is likely to depend significantly on specific situations, institutional guidelines, and donor agreements, we have documented situations in which forensic discovery are likely to have a major and immediate impact on long-term preservation and accessibility.

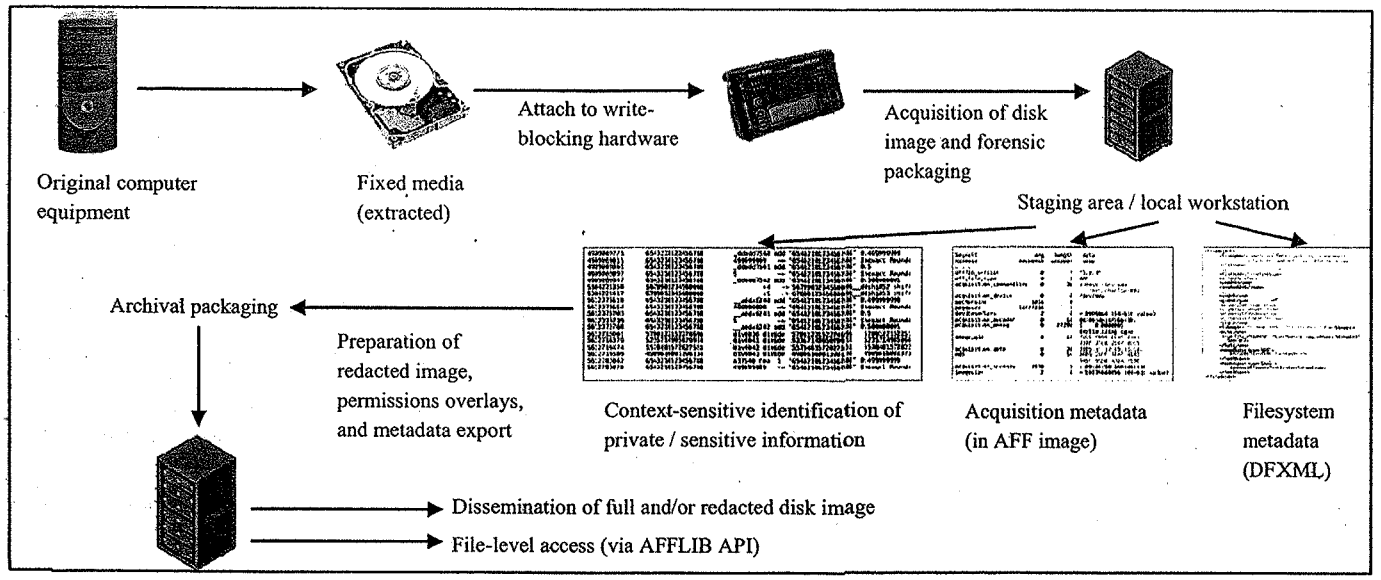

Figure 1: Example of forensic acquisition and information extraction from fixed media (hard disk).

The first of these concerns encryption and password protection. Documents created by individuals and organizations are frequently encrypted or password protected because of short-term data protection mandates or concerns. The requirement for item-level encryption or password protection may no longer be applicable when items are stored in a repository that applies its own security measures, or the time horizon for necessary restrictions has passed. Individuals may have also encrypted or protected files inadvertently, or without any institutional mandate for doing so (a government official who encrypts a file that should be considered a public record or company employee who encrypts a file that should be accessible by the company as an institutional record).

Materials that contain important societal, cultural, and historic value can arrive at collecting institutions with no supporting documentation concerning passwords or encryption keys used to protect them *in situ*. In such cases, environmental context from a disk image may be advantageous – in identifying likely locations of password and encryption key stores, using word lists based on drive contents to reduce the search space required by password recovery utilities, or circumventing protection directly.

In some cases, producers may believe that drive or system failure has caused unrecoverable data loss (or may lack the technical expertise to perform the recovery themselves). Similarly, producers may incur data losses due to inadvertent or malicious user activities that compromise or render the original filesystem unusable. Imaging and forensic analysis at the sector level can assist in both recovery and event analysis.

In order to determine appropriate levels of access to data from an acquire disk, archivists will ideally be able to consult individual producers, representatives of creating organizations, detailed donor agreements, and (when appropriate) applicable laws that dictate who is entitled to access data. However, such information is often not available, and archivists must make their best professional judgments. The information embedded in a disk image can often help to make such determinations. For example, information about when and how a file was encrypted, what user accounts are associated with it, whether someone has created a decrypted version, what other actions have been taken on the file, and where it is located in the filesystem hierarchy can all provide important evidence of the likely intentions of the file's creator and the context of its creation.

Imaging Practices and Associated Workflows

In this section we provide a description of methods for creating disk images, and associated benefits. We focus on techniques that are agnostic with respect to underlying filesystems. We examine packaging mechanisms derived from the field of digital forensics that provide metadata about both the capture process and the format, organization, and other salient characteristics of the original physical device, and support interoperable metadata that can be shared among systems and applications to provide archivists with simplified mechanisms and workflows for filesystem analysis.

Write Blocking

Without the use of a hardware write-blocker, it is extremely difficult to ensure that the host operating system and hardware will not issue commands to alter the contents of connected, write-enabled devices (including traditional spinning-platter hard disks and SSDs, flash-based media and portable devices, magnetic, and writable optical media). Hardware write-blockers are desirable for the following reasons:

1. Additional feedback on data processing (via displays and warning lights) which can be monitored manually.
2. If one ever hopes to compare a disk image to the contents of the original disk (or compare two different disk images from the same disk) by using cryptographic hashes, the bitstreams must be exactly the same. Even a single bit written accidentally to the disk will result in a completely different hash value.

3. Use of a write-blocker provides an operational and legal foundation on which an institution can rely if questions arise about technical competency and due diligence when handling high-value materials.

Write-blockers are not a panacea; institutions should be aware (both in practice and in established written guidelines) that like any other device, write-blockers can be misused, fail, and suffer from manufacturing defects. Software controls and appropriate protocols for cryptographic checks can serve to mitigate the likelihood of data damage.

Image Packaging Format and Image Metadata

Disk image packaging formats, whether used only for initial data analysis or for long-term archival preservation, should comply with several specific technical requirements for storage, description, and access:

1. The image packaging format should be open. Open-description image packaging formats should include a complete technical description of the format structure, and a freely accessible application programming interface (API) to access data contained in an instance of the format.
2. Image packaging formats should provide support for extensible metadata to describe the acquisition process and organization of the acquired device.
3. Image packaging formats should balance efficiency with respect to space (storage requirements) and time (processing required to extract and access the contents). The space requirement, therefore, has the further requirement of a reliable, open compression format. Access to data items at a specific offset within the compressed disk image should not require decompressing the entire image.

The Advanced Forensics Format (AFF) meets all of the above criteria. In the following section, we discuss AFF and associated tools that can be used to process it.

Formats, Tools and Workflow Integration

As noted earlier, most low-level disk imaging utilities used to produce sector-by-sector copies of the physical media do not record acquisition metadata that is desirable in long-term archival contexts.

A number of forensic packaging and disk imaging formats are in use today.[1] Disregarding commercial binary formats which have been reverse-engineered, those which are platform-limited, and formats which add specific limited features such as cryptographic signing and seekable access to compressed bitstreams, AFF is currently the most complete open standard for filesystem-independent image packaging which incorporates extensible metadata [2, 5]. AFF provides additional flexibility in that an existing open source library (AFFLIB) provides various facilities for image manipulation, including extraction of the original raw disk image from the zip64-compressed packaging format.

[1] A list of many available formats can be found on the Forensics Wiki, http://www.forensicswiki.org/wiki/Forensic_file_formats.

Several open source and commercial applications now support AFF creation. These include the command-line tool aimage, Guymager, and AccessData's Forensic Toolkit (FTK) and FTK Imager. Each of these applications successfully produces compliant AFF images, although only aimage can be easily incorporated into batch-processing mechanisms.

Archival institutions often perform pre-ingest processing of digital materials to verify, validate, organize and record metadata about the contents of a device. This includes identification of and validation of digital object file formats; identification and assessment of select materials to be preserved within the archive; extraction of existing metadata and recording of metadata to be incorporated into a submission information package (SIP) or archival information package (AIP).

AFF objects containing images of modern filesystems can be browsed interactively using both open source and commercial forensics tools such as The Sleuth Kit (TSK) and FTK. The open source tool fiwalk can generate reports of all files on a drive, along with their associated filesystem metadata and locations within the filesystem hierarchy (file paths) [5, 7].

The bulk extractor tool, developed by Simson Garfinkel, provides an efficient method for performing stream-based analysis of disk images, identifying and re porting on private and sensitive data that includes (at the time of writing):

- Email, email addresses, and email header information
- Phone numbers and credit card numbers
- Search terms, visited URLs, and search history
- GPS coordinates (geo-location data)
- Exchangeable Image File Format (EXIF) metadata
- Pagefile and registry data on Windows systems
- Word lists from all text streams on a device

Bulk extractor provides facilities for contextual stop-lists and regular expression search to narrow the scope of the reported data, allowing the user to target those areas of private information that are most relevant to a given task.

When applied as part of an integrated workflow, use of AFF for image acquisition – along with fiwalk, bulk extractor, and related tools for filesystem reporting and data analytics – can lower the technical risk of information loss and inadvertent release of private and personally identifying information. Use of these tools can assist in rapidly identifying problematic documents, establishing parameters successful for future access, and reducing reliance on tools that lack shared mechanisms for import and export of metadata.

In the following section we describe current work to build a software toolkit for archives and other collecting institutions based on these technologies.

Development and Packaging of Applications - BitCurator

BitCurator is a project funded by the Andrew W. Mellon Foundation to integrate digital forensics procedures and open-source digital forensics software into archival workflows. It is a joint effort led by the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill and the Maryland Institute for Technology in the Humanities (MITH) at the University of Maryland. We are drawing from the knowledge of a variety of experts on two advisory boards – a Professional Experts Panel (PEP) and Development Advisory Group (DAG) –

as well as experiences with incorporating digital forensics techniques into courses at SILS.

As a software development initiative, BitCurator has identified fast, reliable open source digital forensics technologies that can be adapted to the needs and requirements of collecting institutions. We are executing this in several ways:

1. Development of new software to support image acquisition and processing workflows.
2. Use of existing plugin architectures and APIs to adapt imaging and data analytics software (focusing on bulk extractor and fiwalk) for use in data acquisition in archives and libraries
3. Improvements and modifications to existing GUI interfaces for users who have minimal technical expertise

BitCurator software dependencies are primarily those that support use of AFF for disk imaging (and related libraries for manipulation AFF data) and DFXML for metadata production and interchange.

Members of the BitCurator PEP and DAG are playing a vital role in providing feedback on the project's decisions and products. They have provided input on the BitCurator requirements documents, and they will be testing features of the software as it is released. The PEP and DAG bring a significant body of experience with incorporating digital forensics methods into software and archival workflows.

BitCurator intends to serve professionals in collecting institutions that have limited technical expertise and information technology infrastructure. One mechanism for reaching this wider professional audience will be "BitCurator-in-a-Box." This will include a bootable Ubuntu environment on a USB flash drive (prepared with precompiled, executable versions of the imaging and information extraction software described in the previous section), plug-and-play write-blocking hardware, and access to support materials providing step-by-step guides for disk imaging, metadata creation and export, and data analytics.

Figure 2 provides a high-level overview of the BitCurator software architecture, along with intended methods of metadata export. BitCurator depends directly on software and services that support AFF for disk imaging. Secondary support for raw images and certain commercial formats is available in some tools. Whenever possible, reporting on filesystem contents and advanced data analytics – including identification of private and individually identifying information (PII), Windows registry exports, and file similarity analysis – are performed by tools that support DFXML.

This figure mirrors a pre-ingest disk image acquisition and analysis workflow, parts of which may be conducted in parallel. Using a batch-capable acquisition tool (aimage) or GUI-driven tool (Guymager), the raw image is extracted from the disk and packaged as an AFF file. Fiwalk is used to produce an XML filesystem report. Bulk extractor is run to identify private and individually identifying information (with offsets into the disk image noted for each feature instance). A script then builds a file-to-disk-block map that associates each instance of a PII feature with a file or unallocated block(s) on the disk. Specialized tools are used independently to report on operating system characteristics, find files that are similar but not identical [11], and export information from the DFXML reports produced by fiwalk and bulk extractor to be incorporated into archival descriptions and collection management systems.

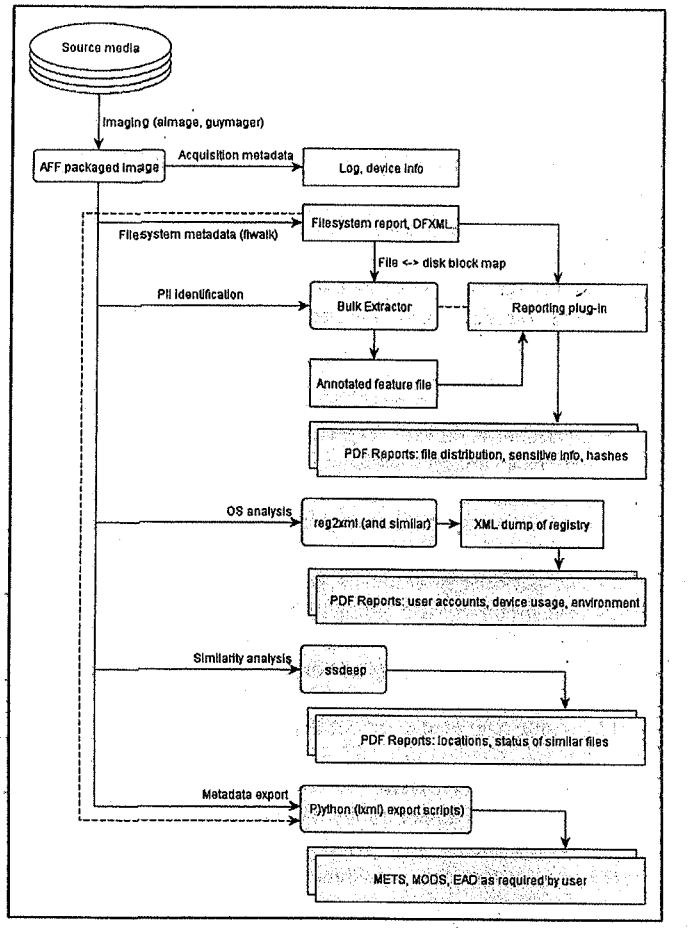

Figure 2: BitCurator architecture.

We are providing ongoing access to our software development repository, community outreach efforts, and supporting information (including FAQs and suggestions for professional practices) at http://www.bitcurator.net/.

Alignment with Professional Needs

In December 2011, PEP members gathered with the BitCurator team for at MITH in College Park, Maryland to discuss BitCurator design documents, identify needs not currently met by existing open source digital forensics tools, and discuss how BitCurator might complement, support and enhance existing digital curation workflows. PEP recommendations included the need for clear, approachable documentation; modular, cross-platform software tools; software and guidance for collecting born-digital materials remotely at a donor's facility or residence; an easily navigable graphical user interface, an API for integration with existing software platforms; command line tools that support batch processing; and data triage functions to automate repetitive or technically challenging tasks. We have been working with PEP members to outline and document their existing acquisition and archival processing workflows. We are refining a master map that identifies parallel areas in the workflow groups, identifies process limitations (or gaps), and links to novel functionality provided by the tools.

The first BitCurator DAG meeting was held in January 2012 at SILS in Chapel Hill, North Carolina. DAG members addressed a variety of issues including BitCurator's role in the broader ecology of digital archives tools; project scope, objectives and planned deliverables; defining intended user groups for BitCurator software; the need for both GUI and command line interfaces, facilitating interactive and batch processing; the need to identify and assess private and sensitive data during multiple stages of the curation process; education and documentation requirements; opportunities for collaboration with and among members of the DAG; and outreach and long-term support and for project deliverables.

Discussion

Significant technical expertise is currently required to implement disk image acquisition, data triage, and processing procedures that incorporate forensic disk imaging, data analytics, and metadata cross-walks to archival and library metadata standards. Both commercial and open source software packages capable of performing some of these tasks have steep learning curves. BitCurator is an ongoing attempt to address these issues, by developing software and interfaces for these communities based on mature, open source digital forensics products, and through outreach to potential users in institutions that require low-cost, reliable, and scalable solutions for handling digital media.

Acknowledgements

This work is supported by a grant from the Andrew W. Mellon Foundation. We would like to acknowledge the contributions of the other members of the BitCurator project team: Alexandra Chassanoff, Matthew Kirschenbaum (Co-PI), and Porter Olson, as well as the members of the project's Professional Expert Panel (Bradley Daigle, Erika Farr, Jeremy Leighton John, Leslie Johnston, Courtney Mumma, Naomi Nelson, Erin O'Meara, Michael Olson, Gabriela Redwine, and Susan Thomas) and Development Advisory Group (Geoffrey Brown, Barbara Guttman, Jerome McDonough, Mark Matienzo, David Pearson, Doug Reside, Seth Shaw, William Underwood, and Peter Van Garderen).

References

[1] AIMS Working Group. AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. 2012.

[2] M.I. Cohen, S. Garfinkel, and B. Schatz, Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow, Proc. DFRWS (2009).

[3] D. Elford, N.D. Pozo, S. Mihajlovic, D. Pearson, G. Clifton, and C. Webb, Media matters: Developing processes for preserving digital objects on physical carriers at the National Library of Australia, 74th IFLA General Conference and Council (2008).

[4] S. Garfinkel, Digital Forensics XML and the DFXML Toolkit, http://simson.net/ref/2011/dfxml.pdf.

[5] S. L. Garfinkel, "Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools," International Journal of Digital Crime and Forensics, 1, 1 (2009) pg. 1-28.

[6] S. Garfinkel and D. Cox, Finding and Archiving the Internet Footprint, First Digital Lives Research Conference (2009).

[7] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt, Bringing Science to Digital Forensics with Standardized Forensic Corpora, 9th Annual Digital Forensic Research Workshop (2009).

[8] J.L. John, Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools, Proc. iPRES (2008).

[9] M.G. Kirschenbaum, R. Ovenden, and G. Redwine, Digital Forensics and Born-Digital Content in Cultural Heritage Collections (Council on Library and Information Resources, Washington, DC, 2010).

[10] S. Ross and A. Gow, Digital Archaeology: Rescuing Neglected and Damaged Data Resources (Library Information Technology Centre, London, 1999).

[11] V. Roussev, An Evaluation of Forensic Similarity Hashes, Proc. DFRWS (2011).

[12] W. Underwood, M. Hayslett, S. Isbell, S. Laib, S. Sherrill, and M. Underwood, Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific And Technical Report (2009).

[13] K. Woods and G. Brown, From Imaging to Access - Effective Preservation of Legacy Removable Media, Proc. Archiving, pg. 213-18. (2009).

[14] K. Woods, C.A. Lee, and S. Garfinkel, Extending Digital Repository Architectures to Support Disk Image Preservation and Access, Proc. JCDL, pg. 57-66. (2011).

Author Biographies

Kam Woods is a Postdoctoral Research Associate in the School of Information and Library Science at the University of North Carolina at Chapel Hill. His research focuses on long-term digital preservation, data forensics, and file system analysis. He holds a Ph.D. in Computer Science from Indiana University Bloomington and a B.A. with a special major in Computer Science from Swarthmore College.

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina at Chapel Hill. His primary area of research is the long-term curation of digital collections. He is particularly interested in the professionalization of this work and the diffusion of existing tools and methods into professional practice. He has a Master of Science and PhD in Information from the University of Michigan.