Automated Redaction of Private and Personal Data in Collections: Toward Responsible Stewardship of Digital Heritage

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

216 Lenoir Drive, CB #3360
100 Manning Hall, Chapel Hill, NC 27599-3360
Phone: +1 919 962 7024
Fax: +1 919 962 8071
Email: callee@ils.unc.edu, kamwoods@email.unc.edu

Abstract

In order to support digital heritage, collecting institutions must serve as trustworthy and responsible stewards of digital information. This requires not only selecting, acquiring and retaining valuable collections, but also providing appropriate access to their contents. Access provision involves data mediation to provide useful access points, to convey contextual information, and to ensure that private information is protected. Identification and redacting of private information is a significant challenge for collecting institutions providing access to born-digital collections. We describe work in the BitCurator project to provide collecting institutions with reliable, automated software and reporting procedures to address the above issues.

1. Introduction

In order to support digital heritage, collecting institutions must serve as trustworthy and competent stewards of digital information. Responsible stewardship requires not only selecting, acquiring and retaining valuable collections, but also providing appropriate access to contents of collections. Access provision involves various forms of mediation, in order to provide useful access points, convey contextual information, and ensure that sensitive information is not inappropriately disclosed.

Modern computing devices often contain a significant amount of private and sensitive information. The information may appear embedded in document content and document metadata, within system files obscured by operating system arcana, and in traces of content held within local storage systems after interaction with services over a network. It may even inadvertently be retained on disk or in static memory after a device has been retired or formatted without being overwritten.

If collecting institutions do not improve their processes for discovery, identification and redaction of sensitive information, there are three major risks to digital heritage. First, collecting institutions could fail to be perceived as trusted actors who can responsibly care for digital collections. As a result, producers of digital content may be unwilling to transfer their materials to institutions for long-term care. Second, if the costs of processing collections are prohibitively high, then institutions are likely to acquire fewer collections than they would otherwise. Finally, reliance on labor-intensive manual procedures will ultimately result in rapidly growing backlogs of unprocessed material that are stored but not available for use. Lack of use is one of the most serious threats to long-term preservation of digital collections. Many of the tasks associated with identification and redaction of private and sensitive information can be simplified and performed with improved coverage and accuracy through the use of open source digital forensics tools that have been designed to work efficiently with large data streams, scale effectively on multiprocessor and multi-core hardware, support a common metadata schema for information transfer, and emphasize extensibility through the use of developer-friendly plug-in architectures.

2. Private Data in Digital Collections

For the purposes of automated analysis, triage, and redaction, we present the following definition of private data: any data that are personally identifying, could be used to establish the identity of the producer, establish the identity or personal details of individuals known to the producer (e.g., friends, family, and clients) or are associated with a private record (e.g. medical, employment, and education). This definition is specific to features one can extract from digital materials that link them directly to the donor or content producer. A donor could consider to be "private" a collection of documents or media which contain no personally-identifying information, but which he/she would not wish to disseminate publicly. Such collections of digital objects do not fit our working definition of private, but they may be considered *personal* or *personally identifying.*

We can further posit instances of features within born-digital objects that fall within our stated criteria, but do not require protection or redaction. However, when working with large (multi-terabyte or larger) collections of heterogeneous objects, it can be desirable to flag all potential instances of such data and perform filtering and triage after the fact.

2.1 Distinguishing between Private and Non-Private Data

Not all private data requires redaction, and the majority (in terms of bytes) of raw data held on media procured from private donors is likely to be non-private. Donor media obtained as disks or disk images of a modern bootable operating system, for example, will likely contain tens or hundreds of gigabytes of operating system and program executables, shared libraries, and documentation. Furthermore, common locations of private content – user-specific directories, storage volumes, and removable media – are just as likely to contain large collections of non-private content (e.g., PDFs downloaded from the web, commercial music and video content, and installation files).

Private data, under the definition provided above, can be associated directly with an individual. This can includes, but is not limited to, social security numbers, credit card numbers, financial records, medical records, employment information, education records, passwords and cryptographic keys, and local and online account records. Private data may also be embedded in data carriers that are not inherently private, such as personal emails and other forms of electronic correspondence.

Collecting institutions deal with large volumes of information that do not meet this strict definition of private, but still requires identification, management, and protection. *Personal* and *personally* *identifying information* is not inherently private, and includes information that belongs to a specific individual (or group of individuals). Users are typically aware that contact names, telephone numbers, emails, and email addresses are personal and contain personally identifying (and in some cases private) information. Users may be unaware of other types of personally identifying information; for example, EXIF metadata within jpeg image containing geolocation data (GPS tracking information), and logs of user activity stored by an operating system over time.

The operating system on modern computing platforms will often store private and personal data in storage locations alongside non-private data. Depending on configuration and use, a wide variety of structures stored within the filesystem may include data that may be considered personal, personally identifying, and/or private. Logs created by installed applications can include usernames, full legal names, and other personally-identifying data such as addresses, birthdates, and passwords. Usernames and passwords and user account information may be stored in plaintext or using cryptographic techniques that have known attack vectors and readily-available exploits. Reconstruction of network traffic cached by certain applications can yield further vulnerabilities. Log information stored both by the operating system (for example, records of types of external storage devices along with serial numbers and timestamps), by applications (chat and social network logs and cached data), and even as a result of efforts by the user to redact or clear events on the system may also be used to reconstruct traces of personal and private data. Form data and cookies stored by web browsers are a well-known vector for such information. Less well understood (even, in many cases, by technically proficient users) is the fact that personal and private data can be recorded 'inadvertently' in hibernation and restore files used by modern operating systems to ensure stateful recovery from suspend events (such as putting hardware to sleep) and system crashes.

Not all traces of interactive use and other activities are private. We would contend that the following will generally contain non-private data:

- Filesystem type and organization (including partitioning info and existence of cryptographically-protected volumes)
- File modification times
- File names (except when those names contain personally identifying information)
- Names and versions of software used to produce digital documents
- Metadata about the hardware, operating system version, and security updates

Accurate identification of non-private data can be a critical aspect of preparing born-digital materials for long-term archival management and access. Specifically, it is often more efficient to cull non-private and non-unique data using known file hash sets (such as those provided in the National Software Reference Library[1]) prior to performing deeper analysis on file contents. Identification of this type of data can also assist in building profiles describing and recording the environment in which documents were produced over the lifetime of the filesystem.

2.2 Identifying and Managing Private and Personally Identifying Data

Finding and categorizing private and personally identifying data in raw, heterogeneous data sources can be an arduous, labor-intensive, and error-prone task. As an example, string and regular expression searches[2] are limited by prior knowledge of what is being sought, fail in the presence of extended character sets, and are often implemented in programs that do not attempt to parse known binary formats, compressed files or filesystems, or file metadata. Procedures that parse only the currently allocated areas of a filesystem may fail to find private and personal data that remains in 'deleted' (deallocated but not overwritten) files, damaged areas of the filesystem, and 'slack space' - data existing

[1] http://www.nsrl.nist.gov/new.html

[2] String searches look for specific runs of characters within the data. Regular expressions are a more complicated way to identify particular partners, allowing one to find e.g. given combinations of characters even if they appear in different sequences or are separated by other characters.

in blocks associated with allocated files that have not been completely overwritten (Garfinkel, 2010). Many tools that provide interactive low-level access to raw devices or disk images depend on extensive post-processing to generate useful reports on what is being observed, or depend on significant expertise and training on the part of the user.

Addressing the requirements of accurately identifying specific features in diverse filesystems and data formats can result in a patchwork of procedures and software mechanisms accumulated over years of effort and experimentation. This patchwork must then itself be curated, along with its associated training documentation and administrative overhead. This is a fundamental barrier to sustainability; knowledge of the system can become increasingly fragmented as training costs and complexity increase. These approaches also scale poorly with increased data volume, as each tool must typically be run independently over the objects in question.

These issues are not unique to collecting institutions, and have many parallels to issues encountered by digital forensics investigators. Garfinkel (Lessons Learned 2012) summarizes findings by Hibshi et al. (2011) that describe fundamental issues that law enforcement professionals have when dealing with raw, heterogeneous data collections:

> *They are general deadline-driven and overworked*. Examiners that have substantial knowledge in one area ... will routinely encounter problems requiring knowledge of other areas ... for which they have no training. Certifications and masters' degrees are helpful, but cannot fundamentally address the diversity problem as *any examiner might reasonably be expected to analyze any information that their organization might possibly encounter on digital media* (S82, emphasis added).

One of the underlying problems – that of working with heterogeneous forms of data from many sources – can be mitigated through the use of high performance, multipurpose software tools that consolidate functionality. Ideally, these tools should provide mechanisms for reporting on data in ways that are neutral with respect to a use case or workflow, allowing institutions to customize the technology for unique aspects of their respective institutional environment.

3. Applying Digital Forensics to Digital Collections

More than a decade ago, a report by Seamus Ross and Ann Gow (1999) 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 forensic tools and methods for acquiring and managing digital collections. This has included activities at the British Library (John 2008), National Library of Australia (Elford et al 2008), and Indiana University (Woods and Brown 2008; Woods and Brown, 2009). The PERPOS project at Georgia Tech has also applied data capture and extraction to US presidential materials (Underwood and Laib 2007; Underwood et al 2009). A project called "Computer Forensics and Born-Digital Content in Cultural Heritage Collections" hosted a symposium and generated a report (Kirschenbaum, Ovenden and Redwine 2010), which provided significant contributions to this discussion. The Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) project developed a framework for the stewardship of born-digital materials that includes the incorporation of digital forensics methods (AIMS Working Group 2012). The Digital Records Forensics project has also articulated a variety of connections between the concepts of digital forensics and archival science (Duranti 2009; Duranti and Endicott-Popovsky 2010; Xie 2011).

3.1 Acquiring Content from Born-Digital Media

When information is obtained from an individual or organization on fixed or removable digital media (such as hard disks, CD-ROMs, and USB drives), the collecting institution often does not know the state of the – whether they are physically damaged, whether the filesystems were left in consistent states by the producer or program which last interacted with them, or whether there are traces of previous (unrelated to the acquisition) filesystem activity remaining on the device.

Simply mounting a device on an appropriately equipped workstation and examining the contents of the filesystem in a desktop environment can pose significant risks to the consistency and authenticity of writeable filesystems (even when the device itself is mounted read-only). Likewise, many modern operating systems will ignore multiple disk volumes (mounting only the first volume), or fail to mount filesystems that are common on other modern platforms (or were common on a legacy platform).

In order to ensure that data from an acquired device remain unchanged (but can still be analyzed thoroughly in their original state), the media can be *forensically imaged.* A bit-identical copy of the data on the media (known as a disk image) can be extracted using *write-blocking hardware*, which is connected as a physical proxy between the media-reading device and the host system and prevents the host system from changing any data on the original media. Typically, this disk image will be packaged with supporting metadata that describes the time, date, and other relevant information about the imaging process, using forensic imaging software that can write to an open imaging format such as the Advanced Forensic Format (AFF) (Garfinkel 2006, 2009), or well-documented commercial formats such as Guidance Software's EnCase Forensic Image. Examples of such software include the open source tool Guymager and the commercial tool Access Data's FTK Imager.

Bit-identical images of source media are useful for many reasons other than identifying and redacting private data. A donor may have forgotten where certain materials are on disk, whether they are password-protected or encrypted, or may have accidentally deleted important records. Access to the complete disk image provides a greater chance of recovery in such situations. Programs and documents located on the original disk may depend on fixed paths. The original filesystem may have damage that can be identified and repaired with access to logs and other system recovery data. Once a disk image has been created, one can automatically extract both general information about the filesystem, and instances of features corresponding to the types of private data outlined earlier.

Digital forensics workflows usually incorporate disk image packing formats, which include not only the raw data from a disk (including all original filesystem metadata from the disk) but also rich metadata about the capture process. Consider the following selected sample of metadata encapsulated in an AFF disk image extracted from a circa-1999 4.2GB external USB hard disk:

Table 1 - Metadata Associated with a Disk Image as Part of an AFF Package

Segment	**Data Length**	**Data**
description	48	Quantum Fireball 4.2GB External
notes	16	000ECC21000831B9
badsectors	8	0 (64-bit value)
md5	16	C553 4AEA 0440 C75E 1B0A 508D...
sha256	32	954B A2B8 30F9 19C5 81F9 F546...
acquisition_seconds	0	= 00:09:46 (hh:mm:ss)

The complete AFF metadata provides a significantly more nuanced view of the capture process, but the information in Table 1 alone provides valuable reference points for future analysis; capture time, multiple cryptographic checksums for the raw image, the serial number of the original hardware (when available), and the number of bad sectors encountered.[3]

Both the filesystem metadata within the disk image and the supplementary metadata within the disk image package can be used to document provenance[4] and chain of custody.[5]   Forensic formats typically “chunk” data, associating cryptographic checksums with each chunk. When compression is

[3] Note that “Data Length” refers to the structure of the output item, not the relevant data value – in this case, no bad sectors were found.

[4] Provenance “consists of the social and technical processes of the records’ inscription, transmission, contextualization, and interpretation which account for its existence, characteristics, and continuing history” (Nesmith 1999, 146).

[5] Chain of custody is the “succession of offices or persons who have held materials from the moment they were created” (Pearce-Moses, 2005, 67). It can be ensured through control, documentation, and accounting for the properties of a digital object 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.

used, it is generally at the chunk level rather than at the level of the file; because of this, if partial data loss occurs (due to physical or logical failure on the storage medium), the undamaged parts of the disk image can be recovered with relative ease.

Note that analysis of disk images does not depend on the use of a custom forensic format; the majority of modern forensics tools will operate just as effectively on raw bitstreams (and even raw devices, should the need arise). In the following section, we discuss software tools and approaches being integrated into BitCurator.

3.2 BitCurator

The BitCurator project is a joint effort—led by the School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) and Maryland Institute for Technology in the Humanities (MITH), and involving contributors from several other institutions—to develop a system for librarians and archivists that incorporates the functionality of many digital forensics tools (Lee et al 2012).

Digital forensics offers methods that can advance the archival goals of maintaining authenticity, describing born-digital records and providing responsible access (Woods and Lee 2012). 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. Much BitCurator activity is translation and adaptation work, based on the belief that professionals in collecting institutions will benefit from tools that are presented in ways that use familiar language and platforms.

BitCurator – and the efforts of many of the project partners – also aim to address two fundamental needs of collecting institutions 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 descriptive systems, as well as modifying forensics triage techniques to better meet the needs of collecting institutions.

(2) Provision of public access to the data. The typical digital forensics scenario is a criminal investigation in which the public never gets access to the evidence that was seized. By contrast, collecting institutions that are creating disk images face issues of how to provide access to the data. This includes not only access interface issues, but also how to redact or restrict 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 core project team met with the PEP in December of 2011 and the DAG in January of 2012 to discuss the design assumptions and goals of the project. We have also received comments and suggestions from individuals in a variety of organizational settings. These various forms of input have helped us to refine the project's requirements and clarify the goals and expectations of working professionals.

The project is packaging, adapting and disseminating a variety of open-source applications. BitCurator is able to benefit from numerous existing open-source tools.[6] The goal is to provide a set of tools that can be used together to perform curatorial tasks but can also be used in combination with many other existing and emerging applications.

In the following sections we briefly discuss risk-reducing practices for acquisition of content from born-digital media, and examine in detail how modern digital forensics tools can be used to ensure that private and personally-identifying data is quickly and accurately identified and reported.

4. Applying Digital Forensics to Identify and Redact Private and Personally Identifying Data in Born-Digital Collections

A primary goal of BitCurator is to enable professionals at collecting institutions to rapidly and accurately identify private and personally identifying data. We facilitate this by using a toolset that is intended to be simple to operate, constructs well-formatted human- and machine-readable reports on those data, and interoperates effectively with existing archival data management systems.

The operational aspects of research efforts conducted as part of BitCurator depend on a set of mature open source software technologies originally developed for digital forensics and law enforcement investigations. The technologies we depend on include bulk extractor and fiwalk (developed by Simson Garfinkel), The Sleuth Kit (developed by Brian Carrier and Basis Technology), and sdhash (developed by Vassil Roussev). These projects, their functions, and our adaptations are described in the following sections.

[6] In addition to those discussed in the following text, BitCurator is also incorporating Guymager, a program for capturing disk images, and Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser.

4.1 Filesystem Analysis

While it can be useful to mount and explore a filesystem interactively on a host machine, it is often undesirable to rely solely on information gained this way. First, this is a risky activity if a hardware write-blocker is not used. Second, this will not yield information on deleted contents, unallocated space, and (often, without specialized software) secondary or alternate filesystems contained on the device. Finally, this form of investigation is time-consuming and prone to human error.

To report on the contents of the filesystem, we use fiwalk, a program originally developed by Simson Garfinkel and now integrated into The Sleuth Kit. Fiwalk processes disk images using the filesystem processing libraries in The Sleuth Kit, and generates results either in Digital Forensics XML (DFXML)[7] or as human-readable plaintext. A typical run of fiwalk will produce some technical metadata on the operation of the program and the host environment, along with a complete walk of the file hierarchy – including volume information, directories, regular files (including lengths and block offsets within the disk images), and files which are no longer allocated (deleted)[8]. Additional uses and examples can be found at http://afflib.org/software/fiwalk.

As an example, consider the following sample of XML output of fiwalk being run against the legacy 4.2GB external USB drive referenced in Section 3.1 (a disk that includes real-world data and is part of the BitCurator Disk Image Test Corpus discussed in Section 4.6). This sample includes a single "fileobject" entry for a file that is orphaned – that is, a file that originally supported the operation of an installed program and is no longer used or referenced by that program. Such files are typically not visible to users interacting with a filesystem:

[7] For a discussion of DFXML, see our section on Data Reporting.

[8] Fiwalk currently recognizes those disk formats which are supported by The Sleuth Kit, including FAT16 and FAT 32, NTFS, HFS+, ext2/3/4, and ISO9660.

```
<fileobject>
    <filename>$OrphanFiles/INDEX.DAT</filename>
    <partition>1</partition>
    <id>90</id>
    <name_type>-</name_type>
    <filesize>49152</filesize>
    <unalloc>1</unalloc>
    <used>1</used>
    <inode>45990</inode>
    <meta_type>1</meta_type>
    <mode>511</mode>
    <nlink>1</nlink>
    <uid>0</uid>
    <gid>0</gid>
    <mtime>2001-07-03T06:19:54Z</mtime>
    <atime>2002-11-21T05:00:00Z</atime>
    <crtime>2001-07-02T18:41:45Z</crtime>
    <byte_runs>
      <byte_run file_offset='0' fs_offset='9715712' img_offset='9747968' len='49152'/>
    </byte_runs>
    <hashdigest type='md5'>59b4d7bede8501c3ac70cd89c6184f56</hashdigest>
    <hashdigest type='sha1'>d53b5dd171246766cca64c8038d323277ffc3fee</hashdigest>
</fileobject>
```

Note that this file is almost 50KB in size. In section 4.3, we provide a further example of how information extracted by forensic data analytics tools can be used to link private and personally identifying information to specific data objects (and specific byte offsets within those objects), particularly when they are “hidden” in this manner.

Using an existing Python module and supporting code distributed with fiwalk, the BitCurator project is preparing programs that can be run against collections of disk images to generate human-readable reports on information useful for triage and data sanitization tasks: for example, the number and type of filesystems recognized, and distribution and location of file types.

4.2 Bulk Data Analysis and Stream-Based Forensics

In order to identify instances of potentially private and personally identifying data, we use bulk extractor, a bulk data analysis tool also developed by Garfinkel. While there are many existing tools that can be used to parse filesystem structures and individual file formats, bulk extractor ignores file system structure and instead processes information from a disk image as a sequence of 16MiByte pages. As a practical consequence, bulk extractor can process different sections of a disk in parallel, greatly increasing performance on modern multicore hardware. Furthermore, bulk extractor can identify, decompress, and recursively reprocess compressed data (for example, modern zipped, XML-based document formats such as Office Open XML).

Bulk extractor is capable of identifying numerous data features using a collection of software modules designed to target feature types, particularly those likely to include private and personally identifying information. These include:

- Private accounts information (credit-card numbers)
- Unique private identifiers (social security numbers)
- Hexadecimal- and Base64-encoded text
- Internet domain names
- Email addresses, email messages and headers
- Ethernet MAC addresses
- EXIF metadata from JPEG images
- Telephone numbers
- URLs and search histories
- Compressed files (zip and gzip files, which BE can proactively decompress and analyze)
- Windows hibernation file fragments
- User-defined keywords

A more detailed listing can be found at https://github.com/simsong/bulk_extractor/wiki, and within the technical documentation distributed with the bulk extractor source code. A typical run of bulk extractor will produce a directory containing text files for each feature type, populated with feature instances and associated offsets into the related disk image. Histogram data for feature instances (e.g. how many times a given email address appeared on a given medium) is generated in separate files.

4.3 Data Triage

Using the filesystem information provided by fiwalk, and the feature data produced by bulk extractor, one can perform a wide variety of triage and reporting tasks working directly from an unmounted disk image. Bulk extractor includes a Python script which maps each feature instance back to a specific location on disk, allowing one to build a list of which feature instances (specifically potentially private and personally-identifying data) correspond to allocated files and which are located in deleted files or currently unallocated space.

Continuing our previous example using data from a legacy 4.3GB external USB hard disk, Table 2 illustrates that features identified by bulk extractor (in this case, URLs visited and searches performed by the user) can often be linked to extant file objects on disk even when primary application caches have been cleared.

Table 2 - Example of Feature Data from Bulk Extractor

Position	16844161
Feature	http://www.yahoo.com
Context	[REDACTED@http://www.yahoo.com\x00\xAD\x0B\...[REDACTED]
File Name	$OrphanFiles/INDEX.DAT
File MD5	39c52b472ec890cc29f71419d6aba999

This example has been selected (and partially redacted) because it is innocuous, and cannot be linked to a specific user. However, even on this relatively small disk (notably, a disk that was never used as a primary boot medium) hundreds of examples of personally identifying and potentially sensitive feature instances were found in the linked output.

One can further identify areas where user activity and user-created data are most prevalent, by isolating collections of relevant features. For example, on a modern Windows 7 operating system this would likely include the user's home folder, the user's registry hive (NTUSER.DAT), common or shared document and media directories, hibernation files produced by the operating system, and the Recycle Bin. Bulk extractor proactively decompresses and reprocesses several common compressed formats, further reducing the need for manual intervention.

An additional method for performing data triage is fuzzy hashing. Comparing standard cryptographic hashes for two files simply indicates whether the two files are exactly the same (are the same bitstreams). Fuzzy hashes, on the other hand, can indicate whether two files are approximately similar. This can help with the managing of private or personal data in at least two ways. First, if a user identifies an item (File A) that is relevant to his/her search goals, but File A contains private data (and thus cannot be viewed by him/her), then a repository could provide a different item (File B) that is quite similar to file A but does not contain the private data in question. Conversely, if a repository contains a set of files that are known to contain private data (based on use of a tools such as bulk extractor), then fuzzy hashing could be used to identify other similar files that might benefit from further human review to determine if they might also contain other forms of private data that were not identified in the initial investigation. BitCurator is incorporating a fuzzy hashing tool called sdhash (Roussev 2011).

4.4 Data Reporting

Using the output of fiwalk, bulk extractor, and additional open source digital forensics tools, BitCurator facilitates the creation of machine-readable and human-readable reports. These reports can be constructed in a modular fashion by reading and reprocessing Digital Forensics XML (DFXML)[9] output and other tool-produced metadata (Garfinkel, Digital Forensics, 2012). DFXML is an evolving schema designed to simplify and standardize the interoperation of tools that produce digital forensics output. Designed to be simple to generate and reprocess, DFXML can be used to describe filesystems and the objects they contain (including low-level information such as permissions, timestamps, location on disk, and cryptographic hashes). Disk image metadata generated using DFXML-producing tools can readily be annotated with Dublin Core tags.

BitCurator will generate reports from disk image contents and DFXML output as PDFs, slide-show presentations, and machine-readable metadata. These can fill a variety of roles: highlighting distribution of data on disk (as simple charts or treemaps); listing areas likely to contain substantial amounts of private data; generating timelines of email activity; identifying use of external devices; and building a difference map between a source disk and a copy.

More sophisticated reports can be produced depending on the options one specifies for (and information one provides to) the bulk data processing tools. For example, bulk extractor supports both *stop lists* and *context-sensitive stop lists* and can be directed to suppress reporting on particular feature instances always or within certain contexts. This can be useful for organizations that are working with large volumes of material and it is known ahead of time that certain data that might appear private is, in fact, not private.

[9] http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML

As an example, any set of feature instances that is output by the bulk extractor tool (such as a list of email addresses from a particular domain) is initially flagged with the byte offset in the disk image where it appears. Assuming the filesystem(s) on the disk have been recognized, one can run a secondary utility distributed with bulk extractor (identify_filenames.py) to link each feature instance either to a file currently allocated within the filesystem, a deleted file, slack within the filesystem, or an unallocated area of the disk. Subsequently, one can organize this output in a way that clearly shows the user where items of concern may appear.

4.5 Redaction, Whitelisting, and Access

A significant issue with many modern digital collections is that while they may contain mostly non-private data, lack of appropriate private-data protection and distribution models often means that they remain "dark," effectively inaccessible.

Redacted disk images can be produced by using the mapping between private data feature instances and disk blocks within the forensically-packaged disk image. Overwriting the relevant parts of the bitstream with randomized or zeroed-out data on a copy of the forensically-packaged image produces a disk image that no longer contains the private data. The redacted image can then be used in various ways - as a backing store for a web front end that provides access to contents of the filesystem, or as a complete distributable object.

Alternatively, whitelisting can be used to produce a "permissions overlay" for the original image, which may be hosted in a sandboxed virtual environment where the user (either at a local workstation or via a remote terminal) interacts with the live environment but is disallowed access to certain files or subtrees within the filesystem. This approach is particularly appropriate when the acquired medium includes a bootable operating system.

If the goal is to generate a collection known to have been produced by a single person or organization, and a donor has confirmed that all identities and data within that collection are to be made public, only a redaction profile for private data from other sources may be required. In every case, however, all private data should be clearly and systematically identified.

4.6 The BitCurator Disk Image Test Corpus

A major challenge to improving the consistency and coverage of private data handling across collecting institutions is the lack of shared corpora on which to test software designed to identify such data. It is unlikely that such a corpus could be created from existing collections materials (or unprocessed materials within backlogs) due to existing donor agreements, legal guidelines, and institutional mandates. Although it is possible to construct large corpora of data from private materials inadvertently or maliciously released onto the Web, this can be ethically problematic to further disseminate and such corpora also will not generally include the complex data structures and interrelationships found on media originally belonging to individual users and corporate entities.

Simson Garfinkel has addressed this problem in the forensics community by constructing the Real Data Corpus (RDC), approximately 30TB of data corresponding to disk images extracted from media and devices purchased on the secondary market. Researchers can obtain access to the RDC with institutional review board (IRB) approval.[10] While the RDC is an extremely useful research and forensic tool development resource, it contains materials from many sources which collecting institutions would identify as having little or no archival value (for example, hard disk drives from retired ATM machines).

[10] http://digitalcorpora.org/corpora/disk-images/real-data-corpus

The National Institute of Standards and Technology has prepared a publicly available corpus, entitled the Computer Forensic Reference Data Set (CFreDS). At the time of writing, this set of disk images and exercises is largely geared towards law enforcement, and includes detailed but relatively anodyne synthesized exercises geared towards basic training and tool testing.

For the BitCurator project, we have been constructing a test corpus in order to replicate common data analysis and triage workflow steps in digital archives. This consists of a non-public corpus of disk images extracted from fixed and removable media identified as containing data likely to have archival value (or requiring long-term preservation) by researchers and practitioners from the project's two advisory groups. In the first year of the project, we requested data from the ten project advisors on our Professional Experts Panel, and nine on our Development Advisory Group. We received data in the form of raw and forensically packaged disk images from the City of Vancouver Archives, the National Institute of Standards and Technology, Duke University, and the National Library of Australia. The transfer is based on a data transfer agreement in which the project team agrees to use the data only for purposes of research and testing within the context of the project.

We have subsequently added to this corpus approximately ten years of disk images from retired workstations and legacy external media provided by iBiblio at the University of North Carolina at Chapel Hill. Additionally, we have included approximately 100,000 government documents in common office file formats crawled from the Web for the purposes of sampling document metadata and content. The corpus currently includes approximately 7.5TB of data, with coverage of major disk formats including FAT16 and FAT32, NTFS, HFS and HFS+, ext3/4, and various double- and high-density floppy images.

In an upcoming phase of the project we will be using this corpus as a testbed for our triage, reporting, and redaction tools. This will allow us to provide statistics on consistency and coverage of our procedures, and isolate problem cases corresponding to damaged or unrecognized filesystems, rare data encodings, and any inconsistencies identified in tool output.

5. Discussion and Future Work

The ability to rapidly and accurately identify the location and nature of private data on a device has many potential applications in libraries and archives. These include:

1. **Formal accounting of best practices - focusing on technical challenges in identifying and handling private and sensitive data in new and existing collections.**

Collecting institutions have written mandates, formal procedures, and legal guidelines in place for handling private data, but these guidelines can result in both conceptual and technical pitfalls in implementation.

2. **Integration of existing digital forensics tools into workflows of collecting institutions.**

Tools such as bulk extractor and The Sleuth Kit provide high-quality coverage of a wide range of private information in many common disk formats, and can be extended to address the needs of those outside police investigation contexts. Ongoing and future work in this area will address the following:

- Providing cumulative statistics on what has been identified in both existing collections and raw donor materials.
- Establishing more complete chains-of-custody and technical provenance metadata in order to support records of authenticity with increased coverage and accuracy.
- Integration of existing digital forensics metadata currently used for tool interoperability (including but not limited to Digital Forensics XML) into extensible metadata schemas and standards supported by the wider community and maintained by a recognized authority.

3. **Preparing customized collections for specific audiences.**

Many institutions wish to provide alternate “views” of raw (unredacted) data sources based on credentials, location, and other forms of authentication. Redaction profiles that limit access to specific areas of a bitstream (or provide a complete copy of the bitstream with sensitive areas overwritten with junk data) are a natural solution to this. Providing simple links between such profiles and authentication mechanisms already in use could provide a powerful mechanism for improved public access to large backlogs of digital materials.

6. Conclusion

We have discussed basic approaches and tools for addressing private and non-private data on digital media. As part of this analysis, we have identified specific circumstances under which they can be identified and (or) redacted. We have examined some current digital forensics technologies that handle this problem efficiently with a low incidence of unwanted results. We have discussed how existing open source digital forensics tools and platforms, including The Sleuth Kit, bulk extractor, and fiwalk can be incorporated into an efficient, automated workflow in order to produce machine- and human-readable reports and metadata. We have addressed these points in the context of BitCurator, an ongoing research initiative at UNC Chapel Hill and the Maryland Institute of Technology and the Humanities to build, test, and analyze software and systems for incorporating digital forensics methods and technologies into the workflows of collecting institutions.

As stated earlier, it is important for collecting institutions to address issues of private and sensitive data for a variety of reasons: in order to serve as trusted actors who can responsibly care for digital collections, to keep the costs of processing collections from being prohibitively high, and to avoid growing backlogs of unprocessed material that are stored but not available for use. Perpetuating humanity’s heritage will increasingly involve the curation of digital traces generated by individuals

(Lee 2011). Caring for and perpetuating this heritage will require responsible curation of content, attending to access restrictions and protection of individuals. Such work will require a great deal of creativity and judgment, supported by efficient and reliable tools.

7. Acknowledgements

This work has been supported by a grant from the Andrew W. Mellon Foundation.

8. References

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

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

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

Elford, Douglas, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, and Colin Webb. "Media Matters: Developing Processes for Preserving Digital Objects on Physical Carriers at the National Library of Australia." Paper presented at the 74th IFLA General Conference and Council, Québec, Canada, August 10-14, 2008

Garfinkel, Simson L. "AFF: A New Format for Storing Hard Drive Images." *Communications of the ACM* 49, no. 2 (2006): 85-87.

Garfinkel, Simson. "Digital Forensics XML and the DFXML Toolset." *Digital Investigation* 8 (2012): 161-174

Garfinkel, Simson L. "Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools." *International Journal of Digital Crime and Forensics* 1, no. 1 (2009): 1-28.

Garfinkel, Simson. "Lessons Learned Writing Digital Forensics Tools and Managing a 30TB Digital Evidence Corpus." *Digital Investigation* 9 (2012): S80-S89.

Garfinkel, Simson, "Digital Forensics Research: The Next 10 Years." DFRWS 2010, Portland, OR, August 2010

Hibshi, Hanan, Timothy Vidas, and Lorrie Cranor. "Usability of Forensics Tools: A User Study." In *IMF 2011: 6th International Conference on It Security Incident Management & IT Forensics:*

Proceedings, 10-12 May 2011, Stuttgart, Germany, 81-91. Los Alamitos, CA: IEEE Computer Society, 2011.

John, Jeremy Leighton. "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools." Paper presented at iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29-30, 2008

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

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

Lee, Christopher A., Matthew Kirschenbaum, Alexandra Chassanoff, Porter Olsen, and Kam Woods. "BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions." *D-Lib Magazine* 18, No. 5/6 (May/June 2012). http://www.dlib.org/dlib/may12/lee/05lee.html

Nesmith, Tom. "Still Fuzzy, but More Accurate: Some Thoughts on the 'Ghosts' of Archival Theory." *Archivaria* 47 (1999): 136-50.

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

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

Roussev, Vassil. “An Evaluation of Forensic Similarity Hashes.” *Digital Investigation* 8 (2011): S34-S41.

Underwood, William, Marlit Hayslett, Sheila Isbell, Sandra Laib, Scott Sherrill, and Matthew Underwood. “Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report.” Technical Report ITTL/CSITD 09-05. October 2009.

Underwood, William, E. and Sandra L. Laib. "PERPOS: An Electronic Records Repository and Archival Processing System." Paper presented at the International Symposium on Digital Curation (DigCCurr 2007), Chapel Hill, NC, April 18-20, 2007.

Woods, Kam and Geoffrey Brown. “From Imaging to Access - Effective Preservation of Legacy Removable Media.” In *Proceedings of Archiving 2009*, 213-18. Springfield, VA: Society for Imaging Science and Technology.

Woods, Kam, and Geoffrey Brown. “Migration Performance for Legacy Data Access.” *International Journal of Digital Curation* 3, no. 2 (2008): 74-88;

Woods, Kam and Christopher A. Lee. “Acquisition and Processing of Disk Images to Further Archival Goals." In *Proceedings of Archiving 2012*, 147-152. Springfield, VA: Society for Imaging Science and Technology, 2012.

Woods, Kam, Christopher A. Lee, and Simson Garfinkel. “Extending Digital Repository Architectures to Support Disk Image Preservation and Access.” In *JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*, 57-66. New York, NY: ACM Press, 2011.

Xie, Sherry L. "Building Foundations for Digital Records Forensics: A Comparative Study of the Concept of Reproduction in Digital Records Management and Digital Forensics." *American Archivist* 74, no. 2 (2011): 576-99.

Biographies:

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, 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. Lee edited and provided several chapters to *I, Digital: Personal Collections in the Digital Era*. He is Principal Investigator of the BitCurator project, which is developing and disseminating open-source digital forensics tools for use by archivists and librarians.

Kam Woods is Postdoctoral Research Associate at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He is currently the Technical Lead on the BitCurator project. Woods received his Ph.D. in Computer Science from Indiana University, Bloomington and his Bachelors with a special major in Computer Science from Swarthmore College. His research focuses on long-term preservation of born-digital materials; he is interested in interdisciplinary approaches that combine expertise in the areas of archiving, computer science, and digital forensics in enabling and maintaining access to digital objects that are at risk due to obsolescence.